EE/ACM 150 - Applications of Convex Optimization in Signal Processing and Communications
Lecture 5

Andre Tkacenko

Signal Processing Research Group
Jet Propulsion Laboratory

April 17, 2012

Caltech

Outline

Introduction to Vector/Matrix Differentiation

Often we will be interested in computing stationary points of scalar/vector/matrix-valued functions which are themselves functions of scalars, vectors, or matrices. This will involve calculating a set of partial derivatives. In many cases, the variation of the function with repsect to its arguments can be greatly simplified by exploiting what is known about matrix analysis.

Suppose $x, y \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^{m_1}$ with $x_k \triangleq [\mathbf{x}]_k$, $\mathbf{y} \in \mathbb{R}^{m_2}$ with $y_k \triangleq [\mathbf{y}]_k$, $\mathbf{X} \in \mathbb{R}^{n_1 \times p_1}$ with $X_{k,\ell} \triangleq [\mathbf{X}]_{k,\ell}$, and $\mathbf{Y} \in \mathbb{R}^{n_2 \times p_2}$ with $Y_{k,\ell} \triangleq [\mathbf{Y}]_{k,\ell}$. Here, x, $\mathbf{x}$, or $\mathbf{X}$ will represent the function argument while y, $\mathbf{y}$, or $\mathbf{Y}$ will denote the function under consideration. The derivative of the function under consideration will be formed from the partials as shown in the table below.

	scalar - y	vector - $\mathbf{y}$	matrix - $\mathbf{Y}$
scalar - x	$\frac{dy}{dx}$	$\frac{d\mathbf{y}}{dx} = \left[\frac{\partial y_k}{\partial x}\right]$	$\frac{d\mathbf{Y}}{dx} = \left[\frac{\partial Y_{k,\ell}}{\partial x}\right]$
vector - $\mathbf{x}$	$\frac{dy}{d\mathbf{x}} = \left[\frac{\partial y}{\partial x_\ell}\right]$	$\frac{d\mathbf{y}}{d\mathbf{x}} = \left[\frac{\partial y_k}{\partial x_\ell}\right]$	
matrix - $\mathbf{X}$	$\frac{dy}{d\mathbf{X}} = \left[\frac{\partial y}{\partial X_{\ell,k}}\right]$		

Here, the partials with respect to the numerator are laid out *according to the shape of the function*, whereas the partials with respect to the denominator are laid out *according to the transpose of the function argument*. For example, $\frac{d\mathbf{y}}{dx}$ is a *column* vector, while $\frac{dy}{d\mathbf{x}}$ is a *row* vector.

Nuances of Adopted Notational Convention

The subtleties associated with the transposed function argument convention are best explained with an example. Suppose that we have

$$\begin{aligned} y &\triangleq \mathbf{x}^T \mathbf{A} \mathbf{x} = \sum_{k_1, k_2} x_{k_1} A_{k_1, k_2} x_{k_2} \,, \\ &= \sum_{k_1 \neq \ell} \sum_{k_2 \neq \ell} x_{k_1} A_{k_1, k_2} x_{k_2} + \sum_{k_1 \neq \ell} x_{k_1} A_{k_1, \ell} x_\ell + \sum_{k_2 \neq \ell} x_\ell A_{\ell, k_2} x_{k_2} + A_{\ell, \ell} x_\ell^2 \,. \end{aligned}$$

Then we have

$$\frac{\partial y}{\partial x_\ell} = \sum_{k_1 \neq \ell} x_{k_1} A_{k_1, \ell} + \sum_{k_2 \neq \ell} A_{\ell, k_2} x_{k_2} + 2 A_{\ell, \ell} x_\ell = \sum_{k_1} x_{k_1} A_{k_1, \ell} + \sum_{k_2} A_{\ell, k_2} x_{k_2} \,.$$

From this, it follows that

$$\frac{\partial y}{\partial x_\ell} = \left[\mathbf{A}^T \mathbf{x} \right]_\ell + [\mathbf{A} \mathbf{x}]_\ell = \left[\mathbf{x}^T \mathbf{A} \right]_\ell + \left[\mathbf{x}^T \mathbf{A}^T \right]_\ell \,.$$

However, since by convention, $\frac{dy}{d\mathbf{x}}$ must be a row vector here, we have

$$\frac{dy}{d\mathbf{x}} = \mathbf{x}^T \left(\mathbf{A} + \mathbf{A}^T \right) \,.$$

A more traditional definition for the derivative would yield $\frac{dy}{d\mathbf{x}} = \left(\mathbf{A} + \mathbf{A}^T \right) \mathbf{x}$. However, this will be *more cumbersome* to work with than the convention used here.

Differentials and the Transposed Notation Convention

To justify the use of the transposed function argument notation convention, we need only look at the *differential* of the function under consideration. Suppose y is a scalar-valued function of a matrix $\mathbf{X}$ with $X_{k,\ell} = [\mathbf{X}]_{k,\ell}$. Then, the differential of y can be obtained by invoking the *chain rule* from multivariable calculus and leads to the following:

$$dy = \sum_k \sum_\ell \left[\frac{\partial y}{\partial X_{k,\ell}} \right] dX_{k,\ell} \, .$$

But recall that $dX_{k,\ell}$ is simply the (k, ℓ)-th entry of the differential of $\mathbf{X}$, denoted $d\mathbf{X}$. Hence, according to our notational convention, we have

$$dy = \sum_k \sum_\ell \left[\frac{dy}{d\mathbf{X}} \right]_{\ell,k} [d\mathbf{X}]_{k,\ell} = \mathrm{tr}\left(\left(\frac{dy}{d\mathbf{X}} \right) d\mathbf{X} \right) \, .$$

Here, we used the identity that

$$\sum_k \sum_\ell [\mathbf{A}]_{\ell,k} [\mathbf{B}]_{k,\ell} = \mathrm{tr}(\mathbf{AB}) \, .$$

This allows us to easily determine the derivative of y with respect to $\mathbf{X}$ from the differential of y as a function of the differential of $\mathbf{X}$.

Note that if we adhered to a more traditional notational convention, then we would end up with $dy = \mathrm{tr}\left(\left(\frac{dy}{d\mathbf{X}} \right)^T d\mathbf{X} \right)$, which would be more cumbersome to work with than the formula above.

Calculating Differentials

As mentioned above, it will be useful to calculate derivatives of functions by computing appropriate differentials. For a scalar/vector/matrix-valued function $\mathbf{F}$ of a scalar/vector/matrix-valued argument $\mathbf{X}$, the differential $d\mathbf{F}(\mathbf{X})$ can be obtained by taking the term of the difference $(\mathbf{F}(\mathbf{X}+d\mathbf{X})-\mathbf{F}(\mathbf{X}))$ which is *linear* in $d\mathbf{X}$. In other words,

$$\mathbf{F}(\mathbf{X}+d\mathbf{X}) = \mathbf{F}(\mathbf{X}) + \underbrace{\text{linear in } (d\mathbf{X})}_{d\mathbf{F}(\mathbf{X})} + \text{higher order terms}\,.$$

Examples:

- $d\mathbf{A} = \mathbf{0}$ if $\mathbf{A}$ is a constant.
- $d(\alpha\mathbf{X}) = \alpha\,(d\mathbf{X})$.
- $d(\mathbf{X}+\mathbf{Y}) = d\mathbf{X} + d\mathbf{Y}$.
- $d(\mathrm{tr}(\mathbf{X})) = \mathrm{tr}(d\mathbf{X})$.
- $d(\mathbf{X}\mathbf{Y}) = (d\mathbf{X})\,\mathbf{Y} + \mathbf{X}\,(d\mathbf{Y})$.
- $d(\mathbf{X}\otimes\mathbf{Y}) = (d\mathbf{X})\otimes\mathbf{Y} + \mathbf{X}\otimes(d\mathbf{Y})$.
- $d(\mathbf{X}\circ\mathbf{Y}) = (d\mathbf{X})\circ\mathbf{Y} + \mathbf{X}\circ(d\mathbf{Y})$.
- $d\left(\mathbf{X}^{-1}\right) = -\mathbf{X}^{-1}\,(d\mathbf{X})\,\mathbf{X}^{-1}$.
- $d(\det(\mathbf{X})) = \det(\mathbf{X})\cdot\mathrm{tr}\left(\mathbf{X}^{-1}d\mathbf{X}\right)$.
- $d(\log(\det(\mathbf{X}))) = \mathrm{tr}\left(\mathbf{X}^{-1}d\mathbf{X}\right)$.
- $d\left(\mathbf{X}^{T}\right) = (d\mathbf{X})^{T}$ and $d(\mathrm{vec}(\mathbf{X})) = \mathrm{vec}(d\mathbf{X})$.

Computing Differentials: Key Examples

To illustrate the use of differentials, it is worthwhile deriving a few key examples.

- **Example #1: Product Rule**
  Suppose we define the function $\mathbf{F}(\mathbf{X}, \mathbf{Y}) \triangleq \mathbf{XY}$. Then we have

$$\begin{aligned} \mathbf{F}(\mathbf{X} + d\mathbf{X}, \mathbf{Y} + d\mathbf{Y}) &= (\mathbf{X} + d\mathbf{X})(\mathbf{Y} + d\mathbf{Y}) , \\ &= \mathbf{XY} + (d\mathbf{X})\mathbf{Y} + \mathbf{X}(d\mathbf{Y}) + (d\mathbf{X})(d\mathbf{Y}) , \\ &= \mathbf{F}(\mathbf{X}, \mathbf{Y}) + \underbrace{(d\mathbf{X})\mathbf{Y} + \mathbf{X}(d\mathbf{Y})}_{\text{linear terms}} + \underbrace{(d\mathbf{X})(d\mathbf{Y})}_{\text{higher order terms}} . \end{aligned}$$

  From this, we conclude that $d(\mathbf{XY}) = (d\mathbf{X})\mathbf{Y} + \mathbf{X}(d\mathbf{Y})$.

- **Example #2: Inverse Rule**
  To derive the expression for $d(\mathbf{X}^{-1})$, we will use the rule for constant matrices as well as the product rule. Note that we have the following:

$$\mathbf{0} = d(\mathbf{I}) = d(\mathbf{X}^{-1}\mathbf{X}) = \left(d(\mathbf{X}^{-1})\right)\mathbf{X} + \mathbf{X}^{-1}(d\mathbf{X}) .$$

  Thus, we get

$$\left(d(\mathbf{X}^{-1})\right)\mathbf{X} = -\mathbf{X}^{-1}(d\mathbf{X}) \Longleftrightarrow d(\mathbf{X}^{-1}) = -\mathbf{X}^{-1}(d\mathbf{X})\mathbf{X}^{-1} .$$

Relating Differentials to Derivatives

Recall from above that a scalar-valued function y of a matrix-valued argument $\mathbf{X}$ yielded the following relation between differentials and derivatives:

$$dy = \mathrm{tr}\left(\left(\frac{dy}{d\mathbf{X}}\right) d\mathbf{X}\right) .$$

This relation followed from the *chain rule* from multivariable calculus. Analogous to the table above defining the notation to use for derivatives, we have the following one relating the differentials to the derivatives for the scalar, vector, and matrix cases.

	scalar - y	vector - $\mathbf{y}$	matrix - $\mathbf{Y}$
scalar - x	$dy = \left(\frac{dy}{dx}\right)(dx)$	$d\mathbf{y} = \left(\frac{d\mathbf{y}}{dx}\right)(dx)$	$d\mathbf{Y} = \left(\frac{d\mathbf{Y}}{dx}\right)(dx)$
vector - $\mathbf{x}$	$dy = \left(\frac{dy}{d\mathbf{x}}\right)(d\mathbf{x})$	$d\mathbf{y} = \left(\frac{d\mathbf{y}}{d\mathbf{x}}\right)(d\mathbf{x})$	
matrix - $\mathbf{X}$	$dy = \mathrm{tr}\left(\left(\frac{dy}{d\mathbf{X}}\right)(d\mathbf{X})\right)$		

From this, the derivative of expressions involving vectors/matrices can be obtained in two steps:

- Compute the differential of the function using rules such as those from above.
- Put the differential in a *canonical form* as in the above table and read off the derivative.

Calculating Derivatives: Examples

For computing derivatives of functions of a matrix variable, one particularly useful identity is $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$.

Examples:

- $y = \mathrm{tr}(\mathbf{AXB})$:
  In this case, we have
  $$dy = \mathrm{tr}(d(\mathbf{AXB})) = \mathrm{tr}(\mathbf{A}\,(d\mathbf{X})\,\mathbf{B}) = \mathrm{tr}(\mathbf{BA}\,(d\mathbf{X}))\ .$$
  Thus, $\frac{dy}{d\mathbf{X}} = \mathbf{BA}$.
- $y = \mathrm{tr}\left(\mathbf{AX}^T\mathbf{BXC}\right)$:
  Here, we have
  $$\begin{aligned} dy &= \mathrm{tr}\left(\mathbf{A}d\left(\mathbf{X}^T\mathbf{BX}\right)\mathbf{C}\right) = \mathrm{tr}\left(\mathbf{A}\left((d\mathbf{X})^T\mathbf{BX} + \mathbf{X}^T\mathbf{B}\,(d\mathbf{X})\right)\mathbf{C}\right)\ , \\ &= \mathrm{tr}\left(\left(\mathbf{A}\,(d\mathbf{X})^T\,\mathbf{BXC}\right)^T\right) + \mathrm{tr}\left(\mathbf{AX}^T\mathbf{B}\,(d\mathbf{X})\,\mathbf{C}\right)\ , \\ &= \mathrm{tr}\left(\mathbf{C}^T\mathbf{X}^T\mathbf{B}^T\,(d\mathbf{X})\,\mathbf{A}^T\right) + \mathrm{tr}\left(\mathbf{CAX}^T\mathbf{B}\,(d\mathbf{X})\right)\ , \\ &= \mathrm{tr}\left(\left(\mathbf{CAX}^T\mathbf{B} + \mathbf{A}^T\mathbf{C}^T\mathbf{X}^T\mathbf{B}^T\right)(d\mathbf{X})\right)\ . \end{aligned}$$
  Hence, $\frac{dy}{d\mathbf{X}} = \left(\mathbf{CAX}^T\mathbf{B} + \mathbf{A}^T\mathbf{C}^T\mathbf{X}^T\mathbf{B}^T\right)$.

Calculating Derivatives: Examples (Continued)

Examples: (Continued)

- $y = \mathrm{tr}\big(\mathbf{A}\mathbf{X}^{-1}\mathbf{B}\big)$:
  For this case, we have

$$dy = \mathrm{tr}\big(\mathbf{A}\left(d\big(\mathbf{X}^{-1}\big)\right)\mathbf{B}\big) = \mathrm{tr}\big(\mathbf{A}\left(-\mathbf{X}^{-1}\left(d\mathbf{X}\right)\mathbf{X}^{-1}\right)\mathbf{B}\big) = \mathrm{tr}\big(\left(-\mathbf{X}^{-1}\mathbf{B}\mathbf{A}\mathbf{X}^{-1}\right)\left(\mathrm{d}\mathbf{X}\right)\big),$$

  Thus, we have $\frac{dy}{d\mathbf{X}} = -\mathbf{X}^{-1}\mathbf{B}\mathbf{A}\mathbf{X}^{-1}$.
- $y = \det\big(\mathbf{X}^T\mathbf{X}\big)$:
  In this example, we have

$$\begin{aligned} dy &= \det\left(\mathbf{X}^T\mathbf{X}\right)\cdot\mathrm{tr}\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}d\left(\mathbf{X}^T\mathbf{X}\right)\right), \\ &= \det\left(\mathbf{X}^T\mathbf{X}\right)\cdot\mathrm{tr}\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\left(d\mathbf{X}\right)^T\mathbf{X}+\mathbf{X}^T\left(d\mathbf{X}\right)\right)\right), \\ &= \det\left(\mathbf{X}^T\mathbf{X}\right)\cdot 2\mathrm{tr}\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(d\mathbf{X}\right)\right). \end{aligned}$$

  Hence, we get $\frac{dy}{d\mathbf{X}} = 2\det\big(\mathbf{X}^T\mathbf{X}\big)\cdot\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$.
- $y = f(\mathbf{X}\mathbf{z})$:
  By definition, we have $df(\mathbf{x}) = \frac{df(\mathbf{x})}{d\mathbf{x}}d\mathbf{x}$, and so

$$dy = \left.\frac{df(\mathbf{x})}{d\mathbf{x}}\right|_{\mathbf{x}=\mathbf{X}\mathbf{z}} d(\mathbf{X}\mathbf{z}) = \left.\frac{df(\mathbf{x})}{d\mathbf{x}}\right|_{\mathbf{x}=\mathbf{X}\mathbf{z}}\left(d\mathbf{X}\right)\mathbf{z} = \mathrm{tr}\left(\left(\mathbf{z}\left(\left.\frac{df(\mathbf{x})}{d\mathbf{x}}\right|_{\mathbf{x}=\mathbf{X}\mathbf{z}}\right)\right)\left(d\mathbf{X}\right)\right).$$

  Thus, we get $\frac{dy}{d\mathbf{X}} = \mathbf{z}\left(\left.\frac{df(\mathbf{x})}{d\mathbf{x}}\right|_{\mathbf{x}=\mathbf{X}\mathbf{z}}\right)$.

Computing Derivatives with Structured Arguments

Previously, the matrix argument $\mathbf{X}$ was assumed to have no structure. However, for many practical cases, $\mathbf{X}$ may have some structure, such as being symmetric, diagonal, Toeplitz, etc.

When the matrix argument has structure, we must simply invoke the chain rule from multivariable calculus in a clever way. Applying the chain rule to the partial derivative $\frac{\partial y}{\partial X_{\ell,k}}$ leads to

$$\begin{aligned}\left[\frac{dy}{d\mathbf{X}}\right]_{k,\ell} &= \frac{\partial y}{\partial X_{\ell,k}} = \sum_{m,p} \frac{\partial y}{\partial X_{m,p}} \cdot \frac{\partial X_{m,p}}{\partial X_{\ell,k}} = \sum_{m,p} \left[\frac{dy}{d\mathbf{X}}\right]_{p,m} \left[\frac{\partial \mathbf{X}}{\partial X_{\ell,k}}\right]_{m,p} , \\ &= \mathrm{tr}\left(\left(\frac{dy}{d\mathbf{X}}\right) \frac{\partial \mathbf{X}}{\partial X_{\ell,k}}\right) = \mathrm{tr}\left(\left(\frac{dy}{d\mathbf{X}}\right) \mathbf{S}_{\ell,k}\right) .\end{aligned}$$

Here, the quantity $\mathbf{S}_{k,\ell} \triangleq \frac{d\mathbf{X}}{dX_{k,\ell}}$ is called the *structure matrix* of $\mathbf{X}$ at the (k,ℓ)-th entry. Any structure in $\mathbf{X}$ can be captured through this quantity.

If X has no structure, then $\mathbf{S}_{k,\ell} = \mathbf{J}^{k,\ell}$, where $\mathbf{J}^{k,\ell}$ is the *single-entry matrix* with a one in the (k,ℓ)-th entry and zero otherwise. On the other hand, if $\mathbf{X}$ is symmetric, then we have

$$\mathbf{S}_{k,\ell} = \mathbf{J}^{k,\ell} + \mathbf{J}^{\ell,k} - \mathbf{J}^{k,\ell}\mathbf{J}^{\ell,k} .$$

In this case, we have

$$\frac{dy}{d\mathbf{X}} = \left(\frac{dy}{d\mathbf{X}}\right) + \left(\frac{dy}{d\mathbf{X}}\right)^T - \mathrm{diag}\left(\frac{dy}{d\mathbf{X}}\right) = \left(\frac{dy}{d\mathbf{X}}\right) + \left(\frac{dy}{d\mathbf{X}}\right)^T - \left(\left(\frac{dy}{d\mathbf{X}}\right) \circ \mathbf{I}\right) ,$$

where it is understood that $\mathbf{X}$ should be regarded as symmetric on the left hand side and unstructured on the right hand side.

Derivatives with Symmetric Arguments

Examples:

- $y = \mathrm{tr}(\mathbf{AXB})$:
  Recall that when $\mathbf{X}$ is unstructured, we have $\frac{dy}{d\mathbf{X}} = \mathbf{BA}$. Thus, we get

  $$\frac{dy}{d\mathbf{X}} = \mathbf{BA} + (\mathbf{BA})^T - \mathrm{diag}(\mathbf{BA}) = \mathbf{BA} + (\mathbf{BA})^T - (\mathbf{BA} \circ \mathbf{I}) .$$

  If $\mathbf{B} = \mathbf{A}^T$, then we have

  $$\frac{d}{d\mathbf{X}} \mathrm{tr}\left(\mathbf{AXA}^T\right) = 2\mathbf{A}^T\mathbf{A} - \mathrm{diag}\left(\mathbf{A}^T\mathbf{A}\right) = 2\mathbf{A}^T\mathbf{A} - \left(\mathbf{A}^T\mathbf{A} \circ \mathbf{I}\right) .$$

- $y = \log(\det(\mathbf{X}))$:
  Recall that when $\mathbf{X}$ is unstructured, we have $\frac{dy}{d\mathbf{X}} = \mathbf{X}^{-1}$. Hence, we have

  $$\frac{dy}{d\mathbf{X}} = \frac{d}{d\mathbf{X}} \log(\det(\mathbf{X})) = 2\mathbf{X}^{-1} - \mathrm{diag}\left(\mathbf{X}^{-1}\right) = 2\mathbf{X}^{-1} - \left(\mathbf{X}^{-1} \circ \mathbf{I}\right) .$$

Introduction to Complex Differentiation

In many cases, we will want to differentiate a function of a complex variable whose output may be real or complex. For this case, we will always want to express the given function in terms of the complex variable and its conjugate. As an example, consider

$$f(z, z^*) \triangleq |z|^2 = z^* z.$$

The key to handling complex differentiation is to view the complex variable and its conjugate as *linearly independent* variables. Once this is accomplished, the differentiation can proceed in much the same way as was done for the real case. For the example above, we have here

$$\frac{\partial f}{\partial z} = z^*, \ \frac{\partial f}{\partial z^*} = z\,.$$

Defining Complex Differentials

As with the real case, the complex differential of a scalar/vector/ matrix-valued function $\mathbf{F}$ of a pair of complex scalar/vector/ matrix-valued arguments $\mathbf{Z}_1$ and $\mathbf{Z}_2$ can be obtained by taking the terms of the difference $\mathbf{F}(\mathbf{Z}_1 + d\mathbf{Z}_1, \mathbf{Z}_2 + d\mathbf{Z}_2) - \mathbf{F}(\mathbf{Z}_1, \mathbf{Z}_2)$ which are *linear* in $d\mathbf{Z}_1$ and $d\mathbf{Z}_2$. In other words,

$$\mathbf{F}(\mathbf{Z}_1 + d\mathbf{Z}_1, \mathbf{Z}_2 + d\mathbf{Z}_2) - \mathbf{F}(\mathbf{Z}_1, \mathbf{Z}_2) = \underbrace{\text{linear in } (d\mathbf{Z}_1) \text{ and } (d\mathbf{Z}_2)}_{d\mathbf{F}(\mathbf{Z}_1, \mathbf{Z}_2)} + \text{ higher order terms}.$$

This leads to the same properties for complex differentials as for real ones. Two new properties are as follows:

- $d(\mathbf{Z}^*) = (d\mathbf{Z})^*$.
- $d\left(\mathbf{Z}^\dagger\right) = (d\mathbf{Z})^\dagger$.

By setting $\mathbf{Z}_1 = \mathbf{Z}$ and $\mathbf{Z}_2 = \mathbf{Z}^*$, we can proceed in much the same way as before for the real case.

Linear Independence of Conjugate Differentials

Assuming linear independence of real and imaginary differentials

Decomposing $\mathbf{Z}$ and $\mathbf{Z}^*$ into real and imaginary parts as

$$\mathbf{Z} = \mathrm{Re}[\mathbf{Z}] + j\mathrm{Im}[\mathbf{Z}] \; , \; \mathbf{Z}^* = \mathrm{Re}[\mathbf{Z}] - j\mathrm{Im}[\mathbf{Z}] \; ,$$

then, assuming that the differentials of the real and imaginary parts $d(\mathrm{Re}[\mathbf{Z}])$ and $d(\mathrm{Im}[\mathbf{Z}])$, respectively, are linearly independent, then so too are $d\mathbf{Z}$ and $d(\mathbf{Z}^*)$.

To show this, consider the conditions under which

$$\mathbf{A}_1 \left(d(\mathrm{vec}(\mathbf{Z})) \right) + \mathbf{A}_2 \left(d(\mathrm{vec}(\mathbf{Z}^*)) \right) = \mathbf{0} \, .$$

This leads to

$$\underbrace{(\mathbf{A}_1 + \mathbf{A}_2)}_{\mathbf{B}_1} \left(d(\mathrm{vec}(\mathrm{Re}[\mathbf{Z}])) \right) + \underbrace{(j \left(\mathbf{A}_1 - \mathbf{A}_2 \right))}_{\mathbf{B}_2} \left(d(\mathrm{vec}(\mathrm{Im}[\mathbf{Z}])) \right) = \mathbf{0} \, .$$

As $d(\mathrm{vec}(\mathrm{Re}[\mathbf{Z}]))$ and $d(\mathrm{vec}(\mathrm{Im}[\mathbf{Z}]))$ are linearly independent, the above condition holds if and only if $\mathbf{B}_1 = \mathbf{B}_2 = \mathbf{0}$. But that is equivalent to saying that $\mathbf{A}_1 = \mathbf{A}_2 = \mathbf{0}$. Hence, $d(\mathrm{vec}(\mathbf{Z}))$ and $d(\mathrm{vec}(\mathbf{Z}^*))$ are linearly independent.

Relating Complex Differentials to Derivatives

As with the real case, it is useful to calculate derivatives through the use of differentials. Similar to the table above relating real differentials to derivatives, we have the following table for complex differentials and complex functions.

	scalar - f	vector - $\mathbf{f}$	matrix - $\mathbf{F}$
scalar - z, z^*	$df = \left(\frac{\partial f}{\partial z}\right)(dz) + \left(\frac{\partial f}{\partial z^*}\right)(dz^*)$	$d\mathbf{f} = \left(\frac{\partial \mathbf{f}}{\partial z}\right)(dz) + \left(\frac{\partial \mathbf{f}}{\partial z^*}\right)(dz^*)$	$d\mathbf{F} = \left(\frac{\partial \mathbf{F}}{\partial z}\right)(dz) + \left(\frac{\partial \mathbf{F}}{\partial z^*}\right)(dz^*)$
vector - $\mathbf{z}$, $\mathbf{z}^*$	$df = \left(\frac{\partial f}{\partial \mathbf{z}}\right)(d\mathbf{z}) + \left(\frac{\partial f}{\partial \mathbf{z}^*}\right)(d\mathbf{z}^*)$	$d\mathbf{f} = \left(\frac{\partial \mathbf{f}}{\partial \mathbf{z}}\right)(d\mathbf{z}) + \left(\frac{\partial \mathbf{f}}{\partial \mathbf{z}^*}\right)(d\mathbf{z}^*)$	
matrix - $\mathbf{Z}$, $\mathbf{Z}^*$	$df = \mathrm{tr}\left(\left(\frac{\partial f}{\partial \mathbf{Z}}\right)(d\mathbf{Z})\right) + \mathrm{tr}\left(\left(\frac{\partial f}{\partial \mathbf{Z}^*}\right)(d\mathbf{Z}^*)\right)$		

Stationarity of a Real Function of Complex Arguments

Suppose $\mathbf{Z}, \mathbf{Z}^* \in \mathbb{C}^{m\times n}$ and $f(\mathbf{Z}, \mathbf{Z}^*)$ is a real function, i.e., $f : \mathbb{C}^{m\times n} \times \mathbb{C}^{m\times n} \to \mathbb{R}$. Decomposing $\mathbf{Z}$ into real and imaginary parts as $\mathbf{Z} = \mathbf{X} + j\mathbf{Y}$ (so that $\mathbf{Z}^* = \mathbf{X} + j\mathbf{Y}$), allows us to express $f(\mathbf{Z}, \mathbf{Z}^*)$ as

$$f(\mathbf{Z}, \mathbf{Z}^*) = g(\mathbf{X}, \mathbf{Y}) \ ,$$

where $g : \mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n} \to \mathbb{R}$. Assuming that $\mathbf{X} = \mathrm{Re}[\mathbf{Z}] \in \mathbb{R}^{m\times n}$ and $\mathbf{Y} = \mathrm{Im}[\mathbf{Z}] \in \mathbb{R}^{m\times n}$ consist of independent variables, it can be shown that the following conditions for stationarity are equivalent.

- $\dfrac{\partial g}{\partial \mathbf{X}} = \mathbf{0}_{n\times m}$, $\dfrac{\partial g}{\partial \mathbf{Y}} = \mathbf{0}_{n\times m}$. (*two $n \times m$ real equations*)
- $\dfrac{\partial f}{\partial \mathbf{Z}} = \mathbf{0}_{n\times m}$. (*one $n \times m$ complex equation*)
- $\dfrac{\partial f}{\partial \mathbf{Z}^*} = \mathbf{0}_{n\times m}$. (*one $n \times m$ complex equation*)

This justifies a *conjugate gradient* approach for finding stationary points.

Second-Order Approximations of Real Functions

Many functions which appear in optimization consist of real-valued scalar functions of real-valued column vectors. As an example, we might be interested in analyzing the behavior of $f(\mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^n$ and $f(\mathbf{x}) \in \mathbb{R}$. For this reason, it is convenient to define a column-based *gradient vector* and a column/row-based *Hessian matrix* as follows.

- **Gradient Vector:**
$$\nabla f(\mathbf{x}) \triangleq \left(\frac{df}{d\mathbf{x}}\right)^T = \frac{df}{d\mathbf{x}^T} \in \mathbb{R}^n \,.$$

- **Hessian Matrix:**
$$\nabla^2 f(\mathbf{x}) \triangleq \left(\frac{d^2 f}{d\mathbf{x}^T d\mathbf{x}}\right)^T = \frac{d^2 f}{d\mathbf{x}\, d\mathbf{x}^T} = \frac{d}{d\mathbf{x}}\left(\frac{df}{d\mathbf{x}^T}\right) \in \mathbb{R}^{n \times n} \,.$$

Then, the second-order *Taylor* approximation of f at $\mathbf{x}$ is given by $\widehat{f}(\mathbf{x}+\mathbf{v})$, where

$$\widehat{f}(\mathbf{x}+\mathbf{v}) = f(\mathbf{x}) + \left(\nabla f(\mathbf{x})\right)^T \mathbf{v} + \frac{1}{2}\mathbf{v}^T \left(\nabla^2 f(\mathbf{x})\right) \mathbf{v} \,.$$

Note that the approximation is exact for $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$, where $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{b} \in \mathbb{R}^n$, and $c \in \mathbb{R}$.

Second-Order Approximations of Complex Functions

In addition to focusing on real-valued vector arguments, it is also worthwhile considering real-valued scalar functions of complex-valued column vectors. As an example, we might be interested in analyzing the behavior of $f(\mathbf{z})$, where $\mathbf{z} \in \mathbb{C}^n$ and $f(\mathbf{z}) \in \mathbb{R}$. Analogous to the real-valued argument case, it is convenient to define a complex-valued column-based *gradient vector* and a complex-valued column/row-based *Hessian matrix* as follows.

- **Gradient Vector:**

$$\nabla f(\mathbf{z}) \triangleq \left(\frac{\partial f}{\partial \mathbf{z}^*} \right)^T = \frac{\partial f}{\partial \mathbf{z}^\dagger} \in \mathbb{C}^n .$$

- **Hessian Matrix:**

$$\nabla^2 f(\mathbf{z}) \triangleq \left(\frac{\partial^2 f}{\partial \mathbf{z}^T \partial \mathbf{z}^*} \right)^T = \frac{\partial^2 f}{\partial \mathbf{z}\, \partial \mathbf{z}^\dagger} = \frac{\partial}{\partial \mathbf{z}} \left(\frac{\partial f}{\partial \mathbf{z}^\dagger} \right) \in \mathbb{C}^{n \times n} .$$

Then, the second-order *Taylor* approximation of f at $\mathbf{z}$ is given by $\widehat{f}(\mathbf{z} + \mathbf{v})$, where

$$\widehat{f}(\mathbf{z} + \mathbf{v}) = f(\mathbf{z}) + 2\mathrm{Re}\left[(\nabla f(\mathbf{z}))^\dagger \, \mathbf{v} \right] + \mathbf{v}^\dagger \left(\nabla^2 f(\mathbf{z}) \right) \mathbf{v} .$$

Note that the approximation is exact for $f(\mathbf{z}) = \mathbf{z}^\dagger \mathbf{A} \mathbf{z} + 2\mathrm{Re}\left[\mathbf{b}^\dagger \mathbf{z} \right] + c$, where $\mathbf{A} \in \mathbb{H}^n$, $\mathbf{b} \in \mathbb{C}^n$, and $c \in \mathbb{R}$. To make the Taylor approximation look more like the real case, sometimes the gradient and Hessian are defined to be twice the value given above.