1

Mathematical Preliminaries

1.1. Introduction

In this chapter, we first introduce the basic tools that are used in the remainder of the book: the statistical characterization and the optimization in the complex domain. For the statistical characterization, we emphasize the importance of taking full statistical information including potential non-circularity of the signals into account, and for the optimization, we review Wirtinger calculus that enables us to perform optimization in the complex domain in a manner very similar to the real-valued case, hence significantly simplifying algorithm derivation and analysis.

1.2. Linear mixing model

We consider the following multidimensional signal mixing model

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \qquad [1.1]$$

where t is the time index, $\mathbf{x}(t) \in \mathbb{C}^{N_o}$ is the set of observations, $\mathbf{s}(t) \in \mathbb{C}^{N_s}$ is the source (component) vector, $\mathbf{n}(t) \in \mathbb{C}^{N_o}$ is the additive random noise vector and

$\mathbf{A} \in \mathbb{C}^{N_o \times N_s}$ is the mixing matrix. This mixing matrix characterizes all the physical propagation transformations between the source signals, and the observations when the sources correspond to actual physical quantities. Furthermore, this model can be exploratory in nature where the mixing matrix summarizes the contribution of each underlying component.

The model in [1.1] corresponds to the so-called linear memoryless mixing model that is typically found in a wide range of applications, including biomedicine, communications, finance and remote sensing [COM 10]. Since the sources $\mathbf{s}(t)$ are not observable, the problem is their *blind* identification given only the mixture, observations $\mathbf{x}(t)$.

Since both $\mathbf{A}$ and $\mathbf{s}(t)$ are unknown, the model in [1.1] is not unique. Indeed, considering an invertible matrix $\mathbf{C}$ to replace $\mathbf{A}$ and $\mathbf{s}(t)$ with $\mathbf{AC}$ and $\mathbf{C}^{-1}\mathbf{s}(t)$, respectively, leaves $\mathbf{x}(t)$ unchanged. Hence, it is important to impose additional assumptions in order to have a tractable model that helps with the main indeterminacies. Throughout the following – unless specified otherwise – we consider that:

A1. matrix $\mathbf{A}$ has full column rank;

A2. source signals are *zero-mean*, unit power and statistically independent stochastic signals;

A3. noise is a zero-mean (white) Gaussian random process;

A4. source signals and noise are statistically mutually independent.

Assumption A1 ensures that a pseudo-inverse $\mathbf{W} = \mathbf{A}^{\dagger}$ of $\mathbf{A}$ exists. Assumption A2 is the key assumption that will make the blind identification or the source separation possible. Assumptions A3 and A4 are related to the nature of noise, and will be important when we consider this more general model. In the rest of this book, we focus on the noiseless case.

1.3. Problem definition

We consider the problems of blind identification and source separation. Blind identification consists of directly estimating the mixing matrix $\mathbf{A}$ whereas source separation has the goal of estimating the source signals $\mathrm{s}(t)$. Here, we consider the latter problem through the direct estimation of an inverse of matrix $\mathbf{A}$, which is accomplished by directly estimating an inverse matrix, called the *demixing matrix* – subject to the ambiguities – denoted by $\mathbf{W}$.

We have seen that given an invertible matrix $\mathbf{C}$, if we change both $\mathbf{A}$ to $\mathbf{AC}$ and $\mathbf{C}^{-1}\mathrm{s}(t)$ in relation to [1.1], then the observations remain unchanged. In practice, this is not acceptable and it is the reason why additional assumptions are imposed. The most important assumption is the statistical independence of source signals. Hence, if $\mathrm{s}(t)$ has statistically independent components, then we no longer have $\mathbf{C}^{-1}\mathrm{s}(t)$ for any non-singular matrix $\mathbf{C}$, but now $\mathbf{C}$ is constrained so as to be $\mathbf{C} = \mathbf{DP}$, where $\mathbf{D}$ is an invertible diagonal matrix and $\mathbf{P}$ is a permutation matrix.

These two indeterminacies corresponding to $\mathbf{D}$ and $\mathbf{P}$ are often acceptable in practice. Indeed, they correspond to an arbitrary ordering and an arbitrary power of the source signals. They also directly correspond to an arbitrary ordering and an arbitrary norm for the columns of the mixing matrix $\mathbf{A}$.

Hence, we estimate the mixing matrix – or rather an inverse of the mixing matrix – up to these two indeterminacies. For the source separation problem, this corresponds to the estimation of matrix $\mathbf{W}$ called a separating or demixing matrix such that we have

$$\mathbf{WA} = \mathbf{DP} \qquad [1.2]$$

where $\mathbf{D}$ is an invertible diagonal matrix and $\mathbf{P}$ is a permutation matrix.

1.4. Statistics

In this section, we provide an overview of the basic concepts relevant to the discussion in the rest of the book. For a more detailed treatment on statistics, readers are referred to [ADA 11, ADA 13, SHC 10].

1.4.1. *Statistics of random variables and random vectors*

A complex-valued random variable $X = X_r + jX_i$ is defined through the joint probability density function (pdf) $p_X(x) \triangleq p_{X_r X_i}(x_r, x_i)$ provided that it exists. The joint pdf for a complex random vector $\mathbf{X} = \mathbf{X}_r + j\mathbf{X}_i \in \mathbb{C}^N$ is written similarly as $p_{\mathbf{X}_r\mathbf{X}_i}(\mathbf{x}_r, \mathbf{x}_i)$. In the subsequent discussions, whenever there is no reason for confusion, we drop the variable subscripts in the definitions of pdfs and statistics to simplify the notation. We use the notation $p(\mathbf{x}) = p(\mathbf{x}_r + j\mathbf{x}_i) \triangleq p(\mathbf{x}_r, \mathbf{x}_i)$ and define the expectations with respect to the corresponding joint pdf. In addition, we assume that all the variables are zero-mean except in few expressions where we specifically include the mean.

Second-order statistics of a complex random vector $\mathbf{X}$ are completely defined through two (auto) covariance matrices: the covariance matrix

$$\mathbf{C}_x = E\{\mathbf{X}\mathbf{X}^H\}$$

that is commonly used and, in addition, the complementary covariance matrix [SCH 03] – also called the *pseudo-*

covariance [NEE 93] or the relation matrix [PIC 96] – given by

$$\widetilde{\mathbf{C}}_x = E\{\mathbf{X}\mathbf{X}^T\}$$

Through their definitions, the covariance matrix is Hermitian and the complementary covariance matrix is complex symmetric. The non-negative eigenvalues of the covariance matrix – which is non-negative definite, and in practice typically positive definite – can be identified using eigenvalue decomposition. However, for the complementary covariance matrix, we need to use Takagi factorization [HOR 99] to obtain the spectral representation. Assuming that $\mathbf{C}$ has full rank, we write the coherence matrix [SCH 06]

$$\mathbf{R} = \mathbf{C}^{-1/2}\widetilde{\mathbf{C}}(\mathbf{C}^*)^{-H/2} = \mathbf{C}^{-1/2}\widetilde{\mathbf{C}}\mathbf{C}^{-T/2}. \qquad [1.3]$$

Since $\mathbf{R}$ is complex symmetric, $\mathbf{R} = \mathbf{R}^T$, not Hermitian symmetric, i.e. $\mathbf{R} \neq \mathbf{R}^H$, there exists a special singular value decomposition (SVD), called the *Takagi factorization*: [HOR 99]

$$\mathbf{R} = \mathbf{F}\mathbf{K}\mathbf{F}^T. \qquad [1.4]$$

The complex matrix $\mathbf{F}$ is unitary and $\mathbf{K} = \text{diag}(k_1, k_2, ..., k_N)$ contains the canonical correlations between $\mathbf{x}$ and $\mathbf{x}^*$, given by $1 \geq k_1 \geq k_2 \geq \cdots \geq k_N \geq 0$ on its diagonal. The squared canonical correlations k_n^2 are the eigenvalues of the squared coherence matrix $\mathbf{R}\mathbf{R}^H = \mathbf{C}^{-1/2}\widetilde{\mathbf{C}}\mathbf{C}^{-*}\widetilde{\mathbf{C}}^*\mathbf{C}^{-H/2}$, or equivalently, of the matrix $\mathbf{C}^{-1}\widetilde{\mathbf{C}}\mathbf{C}^{-*}\widetilde{\mathbf{C}}^*$ [SCH 06, SCH 10]. As in [ERI 06], we refer to these canonical correlations as *circularity coefficients* though the name impropriety coefficients would be more appropriate given the fact that they are a measure of second-order non-circularity [ADA 11].

Second-order circularity properties of complex-valued random variables and vectors are defined in terms of their complementary covariances. A zero-mean complex random variable is called proper [SCH 03, NEE 93] or second-order circular [PIC 94] when its complementary covariance is zero, i.e.

$$E\{X^2\} = 0$$

which implies that $\sigma_{X_r} = \sigma_{X_i}$ and $E\{X_r X_i\} = 0$ where σ_{X_r} and σ_{X_i} are the standard deviations of the real and imaginary parts of the variable. For a random vector $\mathbf{X}$, the condition for propriety or second-order circularity is given by $\widetilde{\mathbf{C}} = \mathbf{0}$, which implies that $E\{\mathbf{X}_r\mathbf{X}_r^T\} = E\{\mathbf{X}_i\mathbf{X}_i^T\}$ and $E\{\mathbf{X}_r\mathbf{X}_i^T\} = -E\{\mathbf{X}_i\mathbf{X}_r^T\}$.

A stronger condition for circularity is based on the pdf of the random variable. A random variable X is called *circular in the strict sense*, or simply *circular*, if X and $Xe^{j\theta}$ have the same pdf, i.e., the pdf is invariant to phase rotations [PIC 94]. In this case, the phase is non-informative and the pdf is a function of only the magnitude. A direct consequence of this property is that $E\{X^p(X^*)^q\} = 0$, for all $p \neq q$, if X is circular. Circularity is a strong property, preserved under linear transformations, and since it implies non-informative phase, a real-valued approach and a complex-valued approach for this case are usually equivalent [VAK 96].

As we would expect, circularity implies second-order circularity, and only for a Gaussian-distributed random variable, second-order circularity implies (strict sense) circularity. Otherwise, the reverse is not true.

1.4.2. *Differential entropy of complex random vectors*

The differential entropy of a zero-mean random vector $\mathbf{X} \in \mathbb{C}^N$ is given by the joint entropy $H(\mathbf{X}_r, \mathbf{X}_i)$, and satisfies [NEE 93]

$$H(\mathbf{X}) \leq \log\left[(\pi e)^N \det(\mathbf{C})\right] \qquad [1.5]$$

with equality, if and only if, $\mathbf{X}$ is second-order circular and Gaussian with zero-mean. Thus, it is a *circular* Gaussian random variable that maximizes the entropy for the complex case. It is also worthwhile to note that orthogonality and Gaussianity, together do not imply independence for complex Gaussian random variables, unless the variable is circular.

For a non-circular Gaussian random vector, we have [ERI 06, SCH 08]

$$H_{\text{noncirc}} = \underbrace{\log\left[(\pi e)^N \det(\mathbf{R})\right]}_{H_{\text{circ}}} + \frac{1}{2}\log\prod_{n=1}^{N}(1-k_n^2)$$

where $_n$ are the circularity coefficients as defined and $k_n = 0$, $n = 0, \ldots, N-1$, when the random vector is circular. Hence, the circularity coefficients provide an attractive measure for quantifying circularity and a number of those measures are studied in [SCH 08]. Since $k_n \leq 1$ for all n, the second term is negative for non-circular random variables decreasing the overall differential entropy as a function of the circularity coefficients.

1.4.3. *Statistics of random processes*

References [AMB 96b, NEE 93, PIC 93, COM 94a, PIC 94] give a detailed account of the statistical characterization and properties of complex random processes. Here, we mainly concentrate on second- and fourth-order statistical properties.

First, the covariance – equivalently correlation since all quantities are assumed to be zero-mean – matrix $\mathbf{R}_x(t)$ of a stochastic signal $\mathbf{x}(t)$ is defined as

$$\mathbf{R}_x(t) = \mathsf{E}\{\mathbf{x}(t)\mathbf{x}^H(t)\} \quad [1.6]$$

and the auto-correlation matrix $\mathbf{R}_x(t,\tau)$ as

$$\mathbf{R}_x(t,\tau) = \mathsf{E}\{\mathbf{x}(t)\mathbf{x}^H(t-\tau)\}. \quad [1.7]$$

Obviously, we have $\mathbf{R}_x(t) = \mathbf{R}_x(t,0)$.

In cases where the random signal vector $\mathbf{x}(t)$ is assumed (broad sense) stationary, then the above two matrices do not depend on t anymore. If the dependence on variable t is periodic, then the stochastic vectorial signal is called cyclostationary.

Similarly, for complex random processes as well, we can introduce the complementary auto-correlation matrix $\widetilde{\mathbf{R}}_x(t)$ of the random vector $\mathbf{x}(t)$ as

$$\widetilde{\mathbf{R}}_x(t) = \mathsf{E}\{\mathbf{x}(t)\mathbf{x}^T(t)\} \quad [1.8]$$

and the complementary auto-correlation matrix as

$$\widetilde{\mathbf{R}}_x(t,\tau) = \mathsf{E}\{\mathbf{x}(t)\mathbf{x}^T(t-\tau)\} \quad [1.9]$$

Again, we have $\widetilde{\mathbf{R}}_x(t) = \widetilde{\mathbf{R}}_x(t,0)$.

As will be seen in the following, second-order statistics can be sufficient for blind identification or source separation. A more general case requires the use of statistics of an order higher than two. These statistics could be moments but cumulants are preferred because of their useful properties w.r.t. statistical independence. For simplicity, we only consider fourth-order cumulants, but all the following

derivations can be very easily extended to cumulants of any order greater than or equal to three.

Depending on the number of complex conjugates, three fourth-order cumulants can be defined in the complex case, they are

$$C_{x,ijkl} = \mathsf{Cum}\{x_i, x_j^*, x_k, x_l^*\} \quad [1.10]$$

$$C'_{x,ijkl} = \mathsf{Cum}\{x_i, x_j, x_k, x_l^*\} \quad [1.11]$$

and

$$C''_{x,ijkl} = \mathsf{Cum}\{x_i, x_j, x_k, x_l\}. \quad [1.12]$$

Note that for simplicity, we only consider the stationary case with no delay. Hence, the cumulants no longer depend on the time index t. This is the reason why we drop it in all of the above definitions. The natural block representation of these high-order cumulants is a tensor of order equal to the cumulant order. However, this tensor can also be well described by sets of matrices corresponding to particular tensor slices.

The simplest way consists of considering the following matrices

$$\big(\mathbf{C}_x(k,l)\big)_{ij} = C_{x,ijkl} \quad [1.13]$$

$$\big(\mathbf{C}'_x(k,l)\big)_{ij} = C'_{x,ijkl} \quad [1.14]$$

and

$$\big(\mathbf{C}''_x(k,l)\big)_{ij} = C''_{x,ijkl}. \quad [1.15]$$

Since $\mathbf{x}$ is an N_o dimensional vector, in each of the above cases, we have N_o^2 matrices of size (N_o, N_o). Thus, in each case, all the N_o^4 cumulants are found in the above matrices.

Sometimes it can be useful to reduce the number of matrices. The following matrices can be considered

$$\left(\mathbf{C}_x(k)\right)_{ij} = C_{x,ijkk} \quad [1.16]$$

$$\left(\mathbf{C}'_x(k)\right)_{ij} = C'_{x,ijkk} \quad [1.17]$$

and

$$\left(\mathbf{C}''_x(k)\right)_{ij} = C''_{x,ijkk}. \quad [1.18]$$

In each of the above three cases, we now only have N_o matrices of size (N_o, N_o). In the first and third cases, it is not a problem since the *a priori* "missing" cumulants can be found within the N_o matrices because of cumulant symmetries. However, in the second case, not all the "missing" cumulants can be found within the N_o matrices. We will see later that this missing statistical information will not be a true problem.

For the reduction of the number of matrices, we can also consider sums of matrices as

$$\left(\mathbf{C}_x(\mathbf{S})\right)_{ij} = \sum_{k,l=1}^{N_o} C_{x,ijkl} S_{kl} \quad [1.19]$$

$$\left(\mathbf{C}'_x(\mathbf{S})\right)_{ij} = \sum_{k,l=1}^{N_o} C'_{x,ijkl} S_{kl} \quad [1.20]$$

and

$$\left(\mathbf{C}''_x(\mathbf{S})\right)_{ij} = \sum_{k,l=1}^{N_o} C''_{x,ijkl} S_{kl} \quad [1.21]$$

where $\mathbf{S}$ is a fixed matrix corresponding to the coefficients of the sum. In each of the above cases, we now only have one matrix of size (N_o, N_o).

Note that the N_o^2 matrices in [1.13] (respectively [1.14] and [1.15]) are special cases of matrix of the form [1.19] (respectively [1.20] and [1.21]) when considering the N_o^2 matrices for $\mathbf{S}$ in the set

$$\left\{\mathbf{E}_{k,\ell} = \mathbf{e}_k \mathbf{e}_\ell^T \quad | \quad 1 \leq k, \ell \leq N_o\right\} \qquad [1.22]$$

where $\mathbf{e}_k$ is the N_o dimensional column vector with 1 in position k and 0 elsewhere.

Note that the N_o matrices in [1.16] (respectively [1.17] and [1.18]) are also special cases of matrix of the form [1.19] (respectively [1.20] and [1.21]) when considering the N_o matrices for $\mathbf{S}$ in the set

$$\left\{\mathbf{E}_{k,k} \quad | \quad 1 \leq k \leq N_o\right\}. \qquad [1.23]$$

1.4.4. *Complex matrix decompositions*

In the noiseless case, using [1.1] in [1.7], [1.9] and [1.19], [1.20] and [1.21], it is rather straightforward to see that

$$\mathbf{R}_x(t,\tau) = \mathbf{A}\mathbf{R}_s(t,\tau)\mathbf{A}^H \qquad [1.24]$$

$$\widetilde{\mathbf{R}}_x(t,\tau) = \mathbf{A}\widetilde{\mathbf{R}}_s(t,\tau)\mathbf{A}^T \qquad [1.25]$$

$$\mathbf{C}_x(\mathbf{S}) = \mathbf{A}\mathbf{C}_{s,x}(\mathbf{S})\mathbf{A}^H \qquad [1.26]$$

$$\mathbf{C}'_x(\mathbf{S}) = \mathbf{A}\mathbf{C}'_{s,x}(\mathbf{S})\mathbf{A}^T \qquad [1.27]$$

and

$$\mathbf{C}''_x(\mathbf{S}) = \mathbf{A}\mathbf{C}''_{s,x}(\mathbf{S})\mathbf{A}^T \qquad [1.28]$$

where by definition

$$\left(\mathbf{C}_{s,x}(\mathbf{S})\right)_{ij} = \sum_{k,l=1}^{N_o} \mathsf{Cum}\{s_i, s_j^*, x_k, x_l^*\} S_{kl} \qquad [1.29]$$

$$\left(\mathbf{C}'_{s,x}(\mathbf{S})\right)_{ij} = \sum_{k,l=1}^{N_o} \mathsf{Cum}\{s_i, s_j, x_k, x_l^*\} S_{kl} \qquad [1.30]$$

and

$$\left(\mathbf{C}''_{s,x}(\mathbf{S})\right)_{ij} = \sum_{k,l=1}^{N_o} \mathsf{Cum}\{s_i, s_j, x_k, x_l\} S_{kl}. \qquad [1.31]$$

Now because of assumption A2 ensuring statistically independent sources, all the matrices $\mathbf{R}_s(t,\tau)$, $\mathbf{R}'_s(t,\tau)$, $\mathbf{C}_{s,x}(\mathbf{S})$, $\mathbf{C}'_{s,x}(\mathbf{S})$ and $\mathbf{C}''_{s,x}(\mathbf{S})$ are *diagonal*. Hence, depending on considered statistics, we can find two kinds of matrix decomposition. They are written as

$$\mathbf{M}(n) = \mathbf{A}\mathbf{D}(n)\mathbf{A}^H \qquad [1.32]$$

$$\mathbf{M}'(n) = \mathbf{A}\mathbf{D}'(n)\mathbf{A}^T \qquad [1.33]$$

where both matrices $\mathbf{D}(n)$ and $\mathbf{D}'(n)$ are diagonal.

Thus, matrices $\mathbf{M}'(n)$ are always *complex symmetric*, while matrices $\mathbf{M}(n)$ are Hermitian when all matrices $\mathbf{D}(n)$ are real. If this is not the case, we can always consider the Hermitian part of $\mathbf{M}(n)$. Hence, we will talk about matrices $\mathbf{M}(n)$ as *Hermitian* in the following even if the diagonal matrices $\mathbf{D}(n)$ are not real.

We denote the set of Hermitian matrices with $\mathcal{M}_h$

$$\mathcal{M}_h = \left\{\mathbf{M}(n) \in \mathbb{C}^{N_o \times N_o}\ ,\ n = 1, \ldots N_H\right\} \qquad [1.34]$$

and the set of complex symmetric matrices with $\mathcal{M}_s$

$$\mathcal{M}_s = \left\{ \mathbf{M}'(n) \in \mathbb{C}^{N_o \times N_o} \ , \ n = 1, \ldots N_T \right\} \qquad [1.35]$$

Important remark

Given matrix $\mathbf{A}$, the only degrees of freedom for matrices $\mathbf{M}(n)$ in [1.32] correspond to the diagonal components of $\mathbf{D}(n)$. Since there are N_s diagonal components, matrices $\mathbf{M}(n)$ belong to a linear space of dimension N_s. Exactly the same remark holds for matrices $\mathbf{M}'(n)$ in [1.33]. They also belong to a linear space of dimension N_s.

Hence, the set $\mathcal{M}_h$ of Hermitian matrices is said to be *complete* when it contains a basis of the corresponding N_s-dimensional linear space. The set $\mathcal{M}_s$ of complex symmetric matrices is also said to be *complete* when it contains a basis of the corresponding N_s-dimensional linear space.

1.5. Optimization: Wirtinger calculus

In the derivation of independent component analysis (ICA) algorithms and their analyses, we often have to compute gradients and Hessians of the cost functions. Since cost functions are real valued, i.e., are scalar quantities in the complex vector space, they are not analytic, and hence not differentiable in a given open set. To overcome this basic limitation, a number of approaches have been traditionally adopted in the signal processing literature, the most common of which is the evaluation of separate derivatives with respect to the real and complex parts of the non-analytic function. Another approach has been to define augmented vectors by stacking the real and imaginary parts in a vector of twice the original dimension and performing all the evaluations in the real domain, and finally, converting the

solution back to the complex domain. Needless to say, both approaches are cumbersome, and might also lead to the need to make additional assumptions such as circularity to simplify the expressions.

The framework based on Wirtinger calculus [WIR 27, ADA 10] – also called $\mathbb{CR}$ calculus [KRE 07] – provides a simple and straightforward approach to calculate derivatives in the complex plane, in particular for the important case we mention above, for non-analytic functions. Wirtinger calculus allows us to perform all the derivations and analyses in the complex domain without having to consider the real and imaginary parts separately, or without doubling the dimensionality, the approach taken by [VAN 94]. Hence, all the computations can be carried out *in a manner very similar to the real-valued case*, and they become quite straightforward making many tools and methods developed for the real case readily available for the complex case.

In this section, we introduce the main idea behind Wirtinger calculus for scalar, vector, and matrix optimization, and give examples to demonstrate its application. We note that besides keeping the expressions and evaluations simple, a key advantage is that assumptions that have become common practice in complex-valued signal processing – most notably the assumption of circularity of signals – can thus be avoided since evaluations do not become unnecessarily complex.

1.5.1. *Scalar case*

We first consider a complex-valued function $f(z) = u(z_r, z_i) + jv(z_r, z_i)$, where $z = z_r + jz_i$. The classical definition of differentiability, which is identified as *complex*

differentiability in Wirtinger calculus, requires that the derivatives defined as the limit

$$f'(z_0) = \lim_{\Delta z \to 0} \frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z} \qquad [1.36]$$

be independent of the direction in which Δz approaches 0 in the complex plane. This requires that the Cauchy–Riemann equations [ABL 03, ADA 10]

$$\frac{\partial u}{\partial z_r} = \frac{\partial v}{\partial z_i} \text{ and } \frac{\partial u}{\partial z_i} = -\frac{\partial v}{\partial z_r} \qquad [1.37]$$

be satisfied. These conditions are *necessary* for $f(z)$ to be complex-differentiable. If the partial derivatives of $u(z_r, z_i)$ and $v(z_r, z_i)$ are continuous, then they are *sufficient* as well. A function that is complex-differentiable on its entire domain is called *holomorphic* or *analytic*. Obviously, since real-valued cost functions have $v(z_r, z_i) = 0$, the Cauchy–Riemann conditions do not hold, and hence cost functions are not analytic. Indeed, the Cauchy–Riemann equations impose a rigid structure on $u(z_r, z_i)$ and $v(z_r, z_i)$ and thus $f(z)$. A simple demonstration of this fact is that either $u(z_r, z_i)$ or $v(z_r, z_i)$ alone suffices to express the derivatives of an analytic function.

Wirtinger calculus provides a general framework for differentiating non-analytic functions, and is general in the sense that it includes analytic functions as a special case. It only requires that $f(z)$ be differentiable when expressed as a function $f : \mathbb{R}^2 \to \mathbb{R}^2$. Such a function is called real-differentiable. If $u(z_r, z_i)$ and $v(z_r, z_i)$ have continuous partial derivatives with respect to z_r and z_i, then f is real-differentiable. For such a function, we can write:

$$\frac{\partial f}{\partial z} \triangleq \frac{1}{2}\left(\frac{\partial f}{\partial z_r} - j\frac{\partial f}{\partial z_i}\right) \quad \text{and} \quad \frac{\partial f}{\partial z^*} \triangleq \frac{1}{2}\left(\frac{\partial f}{\partial z_r} + j\frac{\partial f}{\partial z_i}\right), \qquad [1.38]$$

which can be derived by writing $z_r = (z + z^*)/2$ and $z_i = (z - z^*)/2j$ and then using the chain rule [REM 91]. The key point is that rather than formally implementing [1.38] as separate derivatives with respect to real and imaginary parts, we can simply consider f to be a bivariate function $f(z, z^*)$ and treat z and z^* as *independent variables*. That is, when applying $\partial f/\partial z$, we take the derivative with respect to z, while formally treating z^* as a constant. Similarly, $\partial f/\partial z^*$ yields the derivative with respect to z^*, formally regarding z as a constant. Thus, there is no need to develop new differentiation rules. This was shown in [BRA 83] in 1983 without a specific reference to Wirtinger's earlier work [WIR 27]. Interestingly, many of the references that refer to [BRA 83] and use the generalized derivatives [1.38] evaluate them by computing derivatives with respect to z_r and z_i separately, instead of directly considering the function in the form $f(z, z^*)$ and directly taking the derivative with respect to z or z^*. This leads to unnecessarily complicated derivations.

When we consider the function in the form $f(z, z^*)$, the Cauchy–Riemann equations can simply be stated as $\partial f/\partial z^* = 0$. In other words, an analytic function cannot depend on z^*. If f is analytic, then the usual complex derivative in [1.36] and $\partial f/\partial z$ in [1.38] coincide. Hence, Wirtinger calculus contains standard complex calculus as a special case.

For *real-valued* $f(z)$, we have $(\partial f/\partial z)^* = \partial f/\partial z^*$, i.e., the derivative and the conjugate derivative are complex conjugates of each other. Because they are related through conjugation, we only need to compute one or the other. As a result, a necessary and sufficient condition for real-valued f to have a stationary point is $\partial f/\partial z = 0$. An equivalent, necessary and sufficient condition is $\partial f/\partial z^* = 0$ [BRA 83].

EXAMPLE 1.1.– Consider the real-valued function $f(z) = |z|^4 = z_r^4 + 2z_r^2 z_i^2 + z_i^4$. We can evaluate $\partial f/\partial z$ by differentiating separately with respect to z_r and z_i,

$$\frac{\partial f}{\partial z} = \frac{1}{2}\left(\frac{\partial f}{\partial z_r} - j\frac{\partial f}{\partial z_i}\right) = 2z_r^3 + 2z_r z_i^2 - 2j(z_r^2 z_i + z_i^3) \quad [1.39]$$

or we can write the function as $f(z) = f(z, z^*) = z^2(z^*)^2$ and differentiate by treating z^* as a constant,

$$\frac{\partial f}{\partial z} = 2z(z^*)^2. \quad [1.40]$$

The second approach is clearly simpler. It can be easily shown that the two expressions, [1.39] and [1.40], are equal. However, while the expression in [1.39] can easily be derived from [1.40], it is not quite as straightforward the other way round. Because $f(z)$ is real-valued, there is no need to compute $\partial f/\partial z^*$: it is simply the conjugate of $\partial f/\partial z$.

Series expansions are a valuable tool in the study of nonlinear functions. For analytic, i.e., complex differentiable functions, the Taylor series expansion assumes the same form as in the real case:

$$f(z) = \sum_{k=0}^{\infty} \frac{f^{(k)}(z_0)}{k!}(z - z_0)^k, \quad [1.41]$$

where $f^{(k)}(z_0)$ denotes the kth order derivative of f evaluated at z_0. If $f(z)$ is analytic for $|z| \leq R$, then the Taylor series given in [1.41] converges uniformly in $|z| \leq R_1 < R$.

As in the case of Taylor expansions, the desire to have the complex domain representation follow the real-valued case closely has also been the main motivation for defining differentiability in the complex domain using [1.36]. However, the class of functions that admit such a representation is

limited, excluding the important group of cost functions. For functions that are *real differentiable*, Wirtinger calculus can be used to write the Taylor series of a non-analytic function as an expansion in z and z^*. We discuss this approach in more detail in section 1.5.2 when we introduce vector optimization using Wirtinger calculus. This simple but useful idea for Taylor series expansions of real-differentiable functions has been introduced in [AMB 96a] and formalized in [ERI 10] using the duality between $\mathbb{R}^{2N}$ and $\mathbb{C}^N$.

1.5.2. *Vector case*

1.5.2.1. *Second-order expansions*

In the development of adaptive signal processing algorithms, i.e., in iterative optimization of a selected cost function and in performance analysis, the first- and second-order expansions prove to be most useful. For an *analytic* function $f(\mathbf{z}) : \mathbb{C}^N \mapsto \mathbb{C}$, we define $\Delta f = f(\mathbf{z}) - f(\mathbf{z}_0)$ and $\Delta\mathbf{z} = \mathbf{z} - \mathbf{z}_0$ to write the second-order approximation to the function in the neighborhood of $\mathbf{z}_0$ as

$$\begin{aligned}\Delta f &\approx \Delta\mathbf{z}^T \nabla_{\mathbf{z}} f + \frac{1}{2}\Delta\mathbf{z}^T \mathbf{H}(\mathbf{z})\, \Delta\mathbf{z} \\ &= \langle \nabla_{\mathbf{z}} f\,, \Delta\mathbf{z}^* \rangle + \frac{1}{2}\langle \mathbf{H}(\mathbf{z})\, \Delta\mathbf{z}\,, \Delta\mathbf{z}^* \rangle \end{aligned} \qquad [1.42]$$

where

$$\nabla_{\mathbf{z}} f = \left.\frac{\partial f(\mathbf{z})}{\partial \mathbf{z}}\right|_{\mathbf{z}_0}$$

is the gradient evaluated at $\mathbf{z}_0$ and

$$\nabla^2_{\mathbf{z}} f \triangleq \mathbf{H}(\mathbf{z}) = \left.\frac{\partial^2 f(\mathbf{z})}{\partial \mathbf{z}\, \partial \mathbf{z}^T}\right|_{\mathbf{z}_0}$$

is the Hessian matrix evaluated at $\mathbf{z}_0$. As in the real-valued case, the Hessian matrix is symmetric and it is constant if the function is quadratic.

For a cost function, on the other hand, $f(\mathbf{z}) : \mathbb{C}^N \mapsto \mathbb{R}$, which is non-analytic, we can use Wirtinger calculus to expand $f(\mathbf{z})$ in two variables $\mathbf{z}$ and $\mathbf{z}^*$, which are treated as independent:

$$\Delta f(\mathbf{z}, \mathbf{z}^*) \approx \langle \nabla_{\mathbf{z}} f, \Delta \mathbf{z}^* \rangle + \langle \nabla_{\mathbf{z}^*} f, \Delta \mathbf{z} \rangle + \frac{1}{2} \left\langle \frac{\partial f^2}{\partial \mathbf{z} \partial \mathbf{z}^T} \Delta \mathbf{z}, \Delta \mathbf{z}^* \right\rangle + \left\langle \frac{\partial f^2}{\partial \mathbf{z} \partial \mathbf{z}^H} \Delta \mathbf{z}^*, \Delta \mathbf{z}^* \right\rangle + \frac{1}{2} \left\langle \frac{\partial f^2}{\partial \mathbf{z}^* \partial \mathbf{z}^H} \Delta \mathbf{z}^*, \Delta \mathbf{z} \right\rangle. \quad [1.43]$$

Thus, the series expansion has the same form as for a real-valued function of two variables, except that these are replaced by $\mathbf{z}$ and $\mathbf{z}^*$. Note that when $f(\mathbf{z}, \mathbf{z}^*)$ is real valued, we have

$$\langle \nabla_{\mathbf{z}} f, \Delta \mathbf{z}^* \rangle + \langle \nabla_{\mathbf{z}^*} f, \Delta \mathbf{z} \rangle = 2\mathrm{Re} \left\{ \langle \nabla_{\mathbf{z}^*} f, \Delta \mathbf{z} \rangle \right\} \quad [1.44]$$

since in this case $\nabla f_{\mathbf{z}^*} = (\nabla f_{\mathbf{z}})^*$. Using the Cauchy–Bunyakovskii–Schwarz inequality, see e.g. [MEY 00], we have

$$|\Delta \mathbf{z}^H \nabla f_{\mathbf{z}^*}| \leq \|\Delta \mathbf{z}\| \|\nabla f_{\mathbf{z}^*}\|$$

which holds with equality when $\Delta \mathbf{z}$ is in the same direction as $\nabla f_{\mathbf{z}^*}$. Hence, it is the gradient *with respect to the complex conjugate of the variable* $\nabla f(\mathbf{z}^*)$ that yields the maximum change Δf.

It is also important to note that when $f(\mathbf{z}, \mathbf{z}^*) = f(\mathbf{z})$, i.e., the function is analytic (complex differentiable), all derivatives with respect to $\mathbf{z}^*$ in [1.43] vanish and [1.43] thus

coincides with [1.42]. As noted earlier, the Wirtinger framework includes analytic functions, and when the function is analytic, all the expressions reduce to those for analytic functions.

1.5.2.2. *Linear transformations between* $\mathbb{C}$ *and* $\mathbb{R}$

We now look at different ways that linear transformations can be described in the real and complex domains. In order to do so, we construct three closely related vectors from two real vectors $\mathbf{w}_r \in \mathbb{R}^N$ and $\mathbf{w}_i \in \mathbb{R}^N$. The first vector is the *complex* vector $\mathbf{w} = \mathbf{w}_r + j\mathbf{w}_i \in \mathbb{C}^N$, and the second vector is the *real composite* $2N$-dimensional vector $\mathbf{w}_\mathbb{R} = [\mathbf{w}_r^T, \mathbf{w}_i^T]^T \in \mathbb{R}^{2N}$, obtained by stacking $\mathbf{w}_r$ on top of $\mathbf{w}_i$. Finally, the third vector is the *complex augmented* vector $\underline{\mathbf{w}} = [\mathbf{w}^T, \mathbf{w}^H]^T \in \mathbb{C}^{2N}$, obtained by stacking $\mathbf{w}$ on top of its complex conjugate $\mathbf{w}^*$. Augmented vectors are always underlined.

Consider a function $f(\mathbf{w})$: $\mathbb{C}^N \mapsto \mathbb{R}$ that is real differentiable up to second order. If we write the function as $f(\mathbf{w}_\mathbb{R}) : \mathbb{R}^{2N} \mapsto \mathbb{R}$ using the augmented vector definition given above, we can easily establish the following two relationships [ADA 10, SCH 10]:

$$\frac{\partial f}{\partial \mathbf{w}_\mathbb{R}} = \mathbf{U}_N^H \frac{\partial f}{\partial \underline{\mathbf{w}}^*} \qquad [1.45]$$

$$\frac{\partial^2 f}{\partial \mathbf{w}_\mathbb{R} \partial {\mathbf{w}_\mathbb{R}}^T} = \mathbf{U}_N^H \frac{\partial^2 f}{\partial \underline{\mathbf{w}}^* \partial \underline{\mathbf{w}}^T} \mathbf{U}_N \qquad [1.46]$$

where

$$\mathbf{U}_N = \begin{bmatrix} \mathbf{I} & j\mathbf{I} \\ \mathbf{I} & -j\mathbf{I} \end{bmatrix} \in \mathbb{C}^{2N \times 2N}. \qquad [1.47]$$

The real-to-complex transformation $\mathbf{U}_N$ is unitary up to a factor of 2, i.e. $\mathbf{U}_N \mathbf{U}_N^H = \mathbf{U}_N^H \mathbf{U}_N = 2\mathbf{I}$. The complex

augmented vector $\underline{\mathbf{w}}$ is obviously a redundant, but convenient, representation of $\mathbf{w}_{\mathbb{R}}$.

1.5.2.3. *Complex gradient updates*

We can use the linear transformation defined above to derive the expressions for gradient descent and Newton updates for iterative optimization in the complex domain. From the real gradient update rule $\Delta\mathbf{w}_{\mathbb{R}} = -\mu\dfrac{\partial f}{\partial \mathbf{w}_{\mathbb{R}}}$, we obtain the complex update relationship

$$\Delta\underline{\mathbf{w}} = \mathbf{U}_N\Delta\mathbf{w}_{\mathbb{R}} = -\mu\mathbf{U}_N\frac{\partial f}{\partial \mathbf{w}_{\mathbb{R}}} = -2\mu\frac{\partial f}{\partial \underline{\mathbf{w}}}.$$

The dimension of the update equation can be further reduced as follows:

$$\begin{bmatrix} \Delta\mathbf{w} \\ \Delta\mathbf{w}^* \end{bmatrix} = -2\mu \begin{bmatrix} \dfrac{\partial f}{\partial \mathbf{w}^*} \\ \dfrac{\partial f}{\partial \mathbf{w}} \end{bmatrix} \Longrightarrow \Delta\mathbf{w} = -2\mu\frac{\partial f}{\partial \mathbf{w}^*}.$$

Again, we note that the gradient with respect to the conjugate of the parameter gives the direction for maximal first-order change, derived here using the representation equivalent to the real-valued case in $\mathbb{R}^{2N}$.

1.5.2.4. *Complex Newton updates*

Given the relationships in [1.45] and [1.46], the Newton update in $\mathbb{R}^{2N}$ given by

$$\frac{\partial^2 f}{\partial \mathbf{w}_{\mathbb{R}}\partial \mathbf{w}_{\mathbb{R}}^T}\Delta\mathbf{w}_{\mathbb{R}} = -\frac{\partial f}{\partial \mathbf{w}_{\mathbb{R}}} \qquad [1.48]$$

can be shown to be equivalent to [ADA 10]

$$\Delta \mathbf{w} = -(\mathbf{H}_2^* - \mathbf{H}_1^* \mathbf{H}_2^{-1} \mathbf{H}_1)^{-1} \left(\frac{\partial f}{\partial \mathbf{w}^*} - \mathbf{H}_1^* \mathbf{H}_2^{-1} \frac{\partial f}{\partial \mathbf{w}} \right) \quad [1.49]$$

in $\mathbb{C}^N$, where

$$\mathbf{H}_1 \triangleq \frac{\partial^2 f}{\partial \mathbf{w} \partial \mathbf{w}^T} \quad \text{and} \quad \mathbf{H}_2 \triangleq \frac{\partial^2 f}{\partial \mathbf{w} \partial \mathbf{w}^H}. \quad [1.50]$$

To establish this relationship, we can use [1.45] and [1.46] to express the real domain Newton update in [1.48] as

$$\frac{\partial^2 f}{\partial \underline{\mathbf{w}}^* \partial \underline{\mathbf{w}}^T} \Delta \underline{\mathbf{w}} = -\frac{\partial f}{\partial \underline{\mathbf{w}}^*},$$

which can then be rewritten as

$$\begin{bmatrix} \mathbf{H}_2^* & \mathbf{H}_1^* \\ \mathbf{H}_1 & \mathbf{H}_2 \end{bmatrix} \begin{bmatrix} \Delta \mathbf{w} \\ \Delta \mathbf{w}^* \end{bmatrix} = - \begin{bmatrix} \dfrac{\partial f}{\partial \mathbf{w}^*} \\ \dfrac{\partial f}{\partial \mathbf{w}} \end{bmatrix}$$

where $\mathbf{H}_1$ and $\mathbf{H}_2$ are defined in [1.50]. We can use the formula for the inverse of a partitioned positive definite matrix ([HOR 99], p. 472), provided that the non-negative definite matrix $\dfrac{\partial^2 f}{\partial \underline{\mathbf{w}}^* \partial \underline{\mathbf{w}}^T}$ is full rank, to write

$$\begin{bmatrix} \Delta \mathbf{w} \\ \Delta \mathbf{w}^* \end{bmatrix} = - \begin{bmatrix} \mathbf{T}^{-1} & -\mathbf{H}_2^{-*} \mathbf{H}_1^* \mathbf{T}^{-*} \\ -\mathbf{T}^{-*} \mathbf{H}_1 \mathbf{H}_2^{-*} & \mathbf{T}^{-*} \end{bmatrix} \begin{bmatrix} \dfrac{\partial f}{\partial \mathbf{w}^*} \\ \dfrac{\partial f}{\partial \mathbf{w}} \end{bmatrix} \quad [1.51]$$

where $\mathbf{T} \triangleq \mathbf{H}_2^* - \mathbf{H}_1^* \mathbf{H}_2^{-1} \mathbf{H}_1$ and $(\cdot)^{-*}$ denotes $[(\cdot)^*]^{-1}$. Since $\dfrac{\partial^2 f}{\partial \underline{\mathbf{w}}^* \partial \underline{\mathbf{w}}^T}$ is Hermitian, we finally obtain the complex Newton

update given in [1.49]. The expression for $\Delta \mathbf{w}^*$ is the conjugate of [1.49].

In [MOR 04], it was shown that the Newton algorithm for N complex variables cannot be written in a form similar to the real-valued case. However, as established here, it can be written as in [1.51] using the augmented form, which is equivalent to the Newton method in $\mathbb{R}^{2N}$. In $\mathbb{C}^N$, it can be expressed as in [1.49]. An equivalent form in $\mathbb{C}^{2N}$ is given in [VON 94] by using the 2×2 real-to-complex mapping $\underline{\mathbf{w}} = \mathbf{U}_1 \mathbf{w}_{\mathbb{R}}$ for each entry of the vector $\mathbf{w} \in \mathbb{C}^N$.

1.5.3. *Matrix case*

Wirtinger calculus extends straightforwardly to functions $f : \mathbb{C}^N \rightarrow \mathbb{C}^M$ or $f : \mathbb{C}^{N \times M} \rightarrow \mathbb{C}$. Similarly, for the matrix gradient defined for $g(\mathbf{W}, \mathbf{W}^*) : \mathbb{C}^{M \times N} \times \mathbb{C}^{M \times N} \rightarrow \mathbb{R}$, we can write

$$\begin{aligned} \triangle g &= \langle \triangle \mathbf{W}, \nabla_{\mathbf{W}^*} g \rangle + \langle \triangle \mathbf{W}^*, \nabla_{\mathbf{W}} g \rangle \\ &= 2\mathrm{Re}\{\langle \triangle \mathbf{W}, \nabla_{\mathbf{W}^*} g \rangle\} \end{aligned} \qquad [1.52]$$

where $\nabla_{\mathbf{W}^*} g = \partial g / \partial \mathbf{W}^*$ is an $M \times N$ matrix whose (k, l)th entry is the partial derivative of g with respect to w_{kl}^*. It is also important to note that, in both cases, the gradient $\nabla_{\mathbf{w}^*} g$ defines the direction of the maximum rate of change in $g(\cdot, \cdot)$ with respect to $\mathbf{w}$, not $\nabla_{\mathbf{w}} g$, as sometimes incorrectly noted. It can be easily verified by using the Cauchy–Schwarz–Bunyakovski inequality [MEY 00] that the term $\nabla_{\mathbf{w}^*} g$ leads to increments that are guaranteed to be non-positive when minimizing a given function. Hence, all the expressions from the real-valued case given, for example

in [PET 08], can be straightforwardly applied to the complex case. For instance, for $g(\mathbf{Z}, \mathbf{Z}^*) = \mathrm{Trace}(\mathbf{Z}\mathbf{Z}^H)$, we obtain

$$\frac{\partial g}{\partial \mathbf{Z}} = \frac{\partial \mathrm{Trace}(\mathbf{Z}(\mathbf{Z}^*)^T)}{\partial \mathbf{Z}} = \mathbf{Z}^* \quad \text{and} \quad \frac{\partial g}{\partial \mathbf{Z}^*} = \mathbf{Z}.$$

Also, when deriving gradient update rules for ICA, Wirtinger calculus has again proven very useful, both for the derivation of the algorithm and in stability and performance analysis [ADA 08, LI 10a, LOE 13]. Next, we demonstrate the derivation of the relative gradient updates [CAR 96a] – as well as equivalently natural gradient updates [AMA 96] – which provides significant gains in gradient optimization of the maximum likelihood (ML) cost.

EXAMPLE 1.2.– To write the relative gradient rule, consider an update of the parameter matrix $\mathbf{W}$ in the invariant form $(\triangle\mathbf{W})\mathbf{W}$ [CAR 96a]. We then write the first-order Taylor series expansion given in [1.52] for the perturbation $(\triangle\mathbf{W})\mathbf{W}$ as

$$\begin{aligned}\triangle g &= \left\langle (\triangle\mathbf{W})\mathbf{W}, \frac{\partial g}{\partial \mathbf{W}^*} \right\rangle + \left\langle (\triangle\mathbf{W}^*)\mathbf{W}^*, \frac{\partial g}{\partial \mathbf{W}} \right\rangle \\ &= 2\mathrm{Re}\left\{ \left\langle \triangle\mathbf{W}, \frac{\partial g}{\partial \mathbf{W}^*}\mathbf{W}^H \right\rangle \right\}\end{aligned}$$

to determine the quantity that maximizes the rate of change in the function. The complex relative gradient of g at $\mathbf{W}$ is then written as $(\partial g/\partial \mathbf{W}^*)\mathbf{W}^H$ leading to the relative gradient update term

$$\Delta\mathbf{W} = -\mu \frac{\partial g}{\partial \mathbf{W}^*}\mathbf{W}^H\mathbf{W}. \qquad [1.53]$$

Upon substitution of $\Delta\mathbf{W}$ into [1.52], we observe that

$$\triangle g = -2\mu \|(\partial g/\partial \mathbf{W}^*)\mathbf{W}^H\|_{\mathrm{Fro}}^2,$$

i.e., a non-positive quantity, thus a proper update term. This result also follows from the observation that the update in [1.53] is nothing but a multiplication of the gradient with a positive definite matrix, $\mathbf{W}^H\mathbf{W}$, provided that $\mathbf{W}$ is full rank.

1.5.4. *Summary*

This chapter presents the main ICA problem and an overview of the basic statistical and optimization tools important for the development in the reminder of the book. A comprehensive statistical characterization of complex-valued random variables is given in [SCH 10] along with their estimation and detection. Wirtinger calculus is presented in more detail in [ADA 10, KRE 07], and [ADA 11, ADA 13] are recent overviews on the topic and include more detailed treatment of topics such as tests of circularity.