DEEP UNFOLDING FOR MULTICHANNEL SOURCE SEPARATION SUPPLEMENTARY MATERIAL

Scott Wisdom[1]*, John Hershey*[2]*, Jonathan Le Roux*[2]*, and Shinji Watanabe*[2]

[1]Department of Electrical Engineering, University of Washington, Seattle, WA, USA
[2]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA

This report details the derivations for the paper by Wisdom, Hershey, Le Roux, and Watanabe [1]. Section 1 gives a practical review of complex gradients and Wirtinger caluclus, and section 2 details the gradient derivations for the deep MCGMM.

1. COMPLEX GRADIENTS AND WIRTINGER CALCULUS

Given a complex-valued function f of complex-valued data $x = u + \mathbb{i}v$, what is the gradient $\nabla_x f$? What is the chain rule if x is a function of an intermediate variable t? This section describes the complex gradient in practice.

Common practice is to use a "composite-real" representation of complex quantities where every complex number is transformed into a two-dimensional real-valued vector:

$$z := \begin{bmatrix} u \\ v \end{bmatrix}. \tag{1}$$

However, taking the gradient of such a representation can lead to incorrect quantities in some situations [2]. Other times, algebra using the real and imaginary parts directly can be arduous, especially for functions that have a lot of interactions between the real and imaginary parts.

An alternative (and equivalent) representation is "augmented-complex" that represents x as a two-dimensional complex vector:

$$\underline{x} := \begin{bmatrix} x \\ x^* \end{bmatrix}, \tag{2}$$

and considers x and x^* to be separate, independent variables.

We will see that the composite-real and augmented-complex representations are useful in different contexts.

1.1. Real-imaginary gradients

The general case for $f(z(t)) \in \mathbb{R}$, with $z = x + \mathbb{i}y$ and $t = r + \mathbb{i}s$:

$$\begin{aligned} \nabla_z f =& \frac{\delta f}{\delta x} + \mathbb{i}\frac{\delta f}{\delta y} \qquad (3) \\ \nabla_t f =& \frac{\delta f}{\delta r} + \mathbb{i}\frac{\delta f}{\delta r} \\ =& \frac{\delta f}{\delta x} \odot \frac{\delta x}{\delta r} + \frac{\delta f}{\delta y} \odot \frac{\delta y}{\delta r} + \mathbb{i}\frac{\delta f}{\delta x} \odot \frac{\delta x}{\delta s} + \mathbb{i}\frac{\delta f}{\delta y} \odot \frac{\delta y}{\delta s} \qquad (4) \\ =& (\Re \nabla_z f) \odot \left(\frac{\delta x}{\delta r} + \mathbb{i}\frac{\delta x}{\delta s} \right) + (\Im \nabla_z f) \odot \left(\frac{\delta y}{\delta r} + \mathbb{i}\frac{\delta y}{\delta s} \right) \qquad (5) \end{aligned}$$

We will now list some special cases that are useful identities.

Case 1: real f, real z, complex t

$$\nabla_t f = (\Re \nabla_z f) \odot \left(\frac{\delta x}{\delta r} + \mathbb{i}\frac{\delta x}{\delta s} \right) \tag{6}$$

Case 2: real f, complex z, real t

$$\nabla_t f = \frac{\delta f}{\delta x} \odot \frac{\delta x}{\delta r} + \frac{\delta f}{\delta y} \odot \frac{\delta y}{\delta r} \tag{7}$$

Case 3: real f, complex z, complex t, $z(t)$ a holomorphic function of t

Because of the Cauchy-Riemann conditions,

$$\begin{aligned} \frac{\delta x}{\delta r} =& \frac{\delta y}{\delta s} \\ \frac{\delta x}{\delta s} =& -\frac{\delta y}{\delta r}, \end{aligned} \tag{8}$$

then

$$\nabla_t f = \nabla_z f \left(\frac{\delta x}{\delta r} + \mathbb{i} \frac{\delta x}{\delta s} \right) = -\mathbb{i} \nabla_z f \left(\frac{\delta y}{\delta r} + \mathbb{i} \frac{\delta y}{\delta s} \right) \tag{9}$$

and

$$\nabla_t f = (\Re \nabla_z f) \odot \left(\frac{\delta x}{\delta r} + \mathbb{i} \frac{\delta x}{\delta s} \right) + (\Im \nabla_z f) \odot \left(\frac{\delta x}{\delta r} + \mathbb{i} \frac{\delta x}{\delta s} \right) \tag{10}$$

1.2. Wirtinger calculus

Real-imaginary gradients are completely sufficient for taking derivatives of complex data. However, for some functions, especially nonholomorphic functions of complex data, the real and imaginary parts can be tedious to derive algebraically. To tackle such derivatives, Wirtinger calculus (also called $\mathbb{CR}$-calculus because of the frequent conversions between complex and real domains) becomes useful.

The Wirtinger derivatives treat z and z^* as separate, independent variables. These derivatives are defined as

$$\frac{\delta}{\delta z} := \frac{1}{2} \left(\frac{\delta}{\delta x} - \mathbb{i} \frac{\delta}{\delta y} \right) \tag{11}$$

$$\frac{\delta}{\delta z^*} := \frac{1}{2} \left(\frac{\delta}{\delta x} + \mathbb{i} \frac{\delta}{\delta y} \right) \tag{12}$$

If f is a holomorphic function (i.e., is only a function of z and not of z^*), then $\frac{\delta f}{\delta z^*} = 0$. If f is an arbitrary (potentially nonholomorphic) complex function of z, the conjugate Wirtinger derivative satisfies the identity

$$\frac{\delta f}{\delta z^*} = \left(\frac{\delta f^*}{\delta z} \right)^* . \tag{13}$$

If f is a scalar real-valued function of z, we have the identity

$$\frac{\delta f}{\delta z^*} = \left(\frac{\delta f}{\delta z} \right)^* , \tag{14}$$

The general case for $f(z(t)) \in \mathbb{R}$, with $z = x + \mathbb{i}y$ and $t = r + \mathbb{i}s$ is

$$\nabla_t f = \frac{\delta f}{\delta z} \frac{\delta z}{\delta t} + \frac{\delta f}{\delta z^*} \frac{\delta z^*}{\delta t} . \tag{15}$$

Using the identity (13), we have

$$\nabla_t f = \frac{\delta f}{\delta z} \frac{\delta z}{\delta t} + \frac{\delta f}{\delta z^*} \left(\frac{\delta z}{\delta t^*} \right)^* . \tag{16}$$

Using the identity (14), we have

$$\nabla_t f = \frac{\delta f}{\delta z} \frac{\delta z}{\delta t} + \left(\frac{\delta f}{\delta z} \right)^* \left(\frac{\delta z}{\delta t^*} \right)^* . \tag{17}$$

Thus, for this case, we can see one of the advantages of Wirtinger calculus versus real-imaginary composite gradients: here, only three derivatives, $\frac{\delta f}{\delta z}$, $\frac{\delta z}{\delta t}$, and $\frac{\delta z}{\delta t^*}$, are required, instead of the four required by the real-imaginary composite gradient chain rule (5).

1.2.1. Gradient checking

Using (11), we can check the derivatives $\frac{\delta f}{\delta x}$ and $\frac{\delta f}{\delta x^*}$ by the following procedure, where ϵ is chosen to be a small constant, usually on the order of 10^{-6}:

$$\begin{aligned}
x^{(+r)} &= x + \epsilon \\
x^{(-r)} &= x - \epsilon \\
x^{(+\mathbb{i})} &= x + \mathbb{i}\epsilon \\
x^{(-\mathbb{i})} &= x - \mathbb{i}\epsilon \\
\frac{\delta f}{\delta x} &\approx \frac{1}{2} \left(\frac{f(x^{(+r)}) - f(x^{(-r)})}{2\epsilon} - \mathbb{i} \frac{f(x^{(+\mathbb{i})}) - f(x^{(-\mathbb{i})})}{2\epsilon} \right) \\
\frac{\delta f}{\delta x^*} &\approx \frac{1}{2} \left(\frac{f(x^{(+r)}) - f(x^{(-r)})}{2\epsilon} + \mathbb{i} \frac{f(x^{(+\mathbb{i})}) - f(x^{(-\mathbb{i})})}{2\epsilon} \right)
\end{aligned}$$

2. GRADIENTS FOR THE DEEP MCGMM

The gradients for backpropagation are as follows. All gradients are Wirtinger gradients, as described in [2], and thus use the Wirtinger chain rule. In the following, the operators $+$ and $\odot$ indicate "broadcasted" addition, and marginalizing multiplication, respectively. For example, broadcasted addition of two tensors of size $N \times 1 \times P$ and $1 \times M \times P$ results in a tensor of size $N \times M \times P$, where any dimensions of 1 are "broadcasted" over the dimension of the other tensor.

Marginalizing multiplication $\odot$ for Wirtinger gradients is defined as follows. If $\delta y/\delta x$ is a gradient of a $N_y \times M_y \times P_y$ tensor y with respect to a $N_x \times M_x \times P_x$ tensor x, and if $\delta x/\delta t$ is a gradient of x with respect to a $N_t \times M_t \times P_t$ tensor t, then we will denote $\delta y/\delta x$ as having size $\frac{N_y \times M_y \times P_y}{N_x \times M_x \times P_x}$ and $\delta x/\delta t$ as having size $\frac{N_x \times M_x \times P_x}{N_t \times M_t \times P_t}$. Then marginalizing multiplication between the two Wirtinger gradients is

$$
\begin{aligned}
\frac{\delta y}{\delta t} =& \frac{\delta y}{\delta x} \odot \frac{\delta x}{\delta t} \\
=& \sum_{n=1}^{N_x} \sum_{m=1}^{M_x} \sum_{p=1}^{P_x} \left[\frac{\delta y}{\delta x}\right]_{\frac{(:,:,:)}{(m,n,p)}} \left[\frac{\delta x}{\delta t}\right]_{\frac{(m,n,p)}{(:,:,:)}} + \sum_{n=1}^{N_x} \sum_{m=1}^{M_x} \sum_{p=1}^{P_x} \left[\frac{\delta y}{\delta x^*}\right]_{\frac{(:,:,:)}{(m,n,p)}} \left[\frac{\delta x^*}{\delta t}\right]_{\frac{(m,n,p)}{(:,:,:)}} .
\end{aligned}
\tag{18}
$$

In terms of implementation for the multichannel GMM, variables and gradients are stored as tensors that all have the same maximum size, $I \times J \times Z \times F \times T$. If a particular gradient does not have a particular dimension, then that dimension is simply set to 1. For example, the $I \times J \times F$ channel model B_f is stored as a tensor of size $I \times J \times 1 \times F \times 1$. In rare cases when more dimensions are needed (for example, the sample spatial covariance $\hat{\Sigma}_f^{xx}$ which is $J \times J \times F$), the unused dimensions can be used. For example, $\hat{\Sigma}_f^{xx}$ is stored as a $J \times J \times 1 \times F \times 1$ tensor.

We will now describe the gradients involved in backpropagation. For reference, [1, figure 2] is helpful. To find all computational paths, start at the cost function $\mathcal{D}$ and follow arrows in the opposite direction.

2.1. Last layer (K)

In the last layer, the following paths exist to $\lambda^{(K)} \equiv \log \gamma^{(K)}$:

$$
\begin{array}{ccccccc}
\mathcal{D} & \leftarrow \hat{X}^{(K)} & & & \leftarrow \bar{\mu}^{(K)} & \leftarrow \bar{\gamma}^{(K)} & \leftarrow \lambda^{(K)} \\
& & \nwarrow \bar{\pi}^{(K)} & \leftarrow L^{(K)} & & & \swarrow \\
& & \nwarrow \bar{\pi}^{(K)} & \leftarrow L^{(K)} & \leftarrow \bar{\mu}^{(K)} & \leftarrow \bar{\gamma}^{(K)} & \swarrow \\
& & \nwarrow \bar{\pi}^{(K)} & \leftarrow L^{(K)} & & \leftarrow \bar{\gamma}^{(K)} & \swarrow
\end{array} ,
\tag{19}
$$

and the paths to $A^{(K)}$ and $b^{(K)}$ are

$$
\mathcal{D} \quad \leftarrow \hat{X}^{(K)} \quad \leftarrow \bar{\pi}^{(K)} \quad \leftarrow L^{(K)} \quad \leftarrow A^{(K)}, b^{(K)}
\tag{20}
$$

These paths require the following gradients:

$$
\frac{\delta \mathcal{D}_{ESR}}{\delta \hat{X}_{f,t}^{j,(K)}} = \frac{2\left(\hat{X}_{f,t}^{j,(K)} - X_{f,t}^{j}\right)^*}{\sum_{f,t} \left|X_{f,t}^{j}\right|^2}
\tag{21}
$$

$$
\frac{\delta \mathcal{D}_{ESR}}{\delta \left(\hat{X}_{f,t}^{j,(K)}\right)^*} = \frac{2\left(\hat{X}_{f,t}^{j,(K)} - X_{f,t}^{j}\right)}{\sum_{f,t} \left|X_{f,t}^{j}\right|^2}
\tag{22}
$$

$$
\frac{\delta \hat{X}_{f,t}^{j,(K)}}{\delta \bar{\mu}_{f,t}^{j,z,(K)}} = \bar{\pi}_t^{j,z,(K)}
\tag{23}
$$

$$
\frac{\delta \bar{\mu}_{f,t}^{j,z,(K)}}{\delta \bar{\gamma}_f^{j,z,(K)}} = -\frac{\bar{\mu}_{f,t}^{j,z,(K)}}{\bar{\gamma}_f^{j,z,(K)}}
\tag{24}
$$

$$
\frac{\delta \hat{X}_{f,t}^{j,(K)}}{\delta \bar{\pi}_t^{j,z,(K)}} = \bar{\mu}_{f,t}^{j,z,(K)}
\tag{25}
$$

$$
\frac{\delta \bar{\pi}_t^{j,z,(K)}}{\delta L_t^{j,z,(K)}} = \text{softmax}'\left(\bar{\pi}_t^{j,z,(K)}\right)
\tag{26}
$$

$$
\frac{\delta L_t^{j,z,(K)}}{\delta \bar{\mu}_{f,t}^{j,z,(K)}} = \left(\frac{\delta L_t^{j,z,(K)}}{\delta (\bar{\mu}_{f,t}^{j,z,(K)})^*}\right)^* = 2\bar{\gamma}_f^{j,z,(K)} (\bar{\mu}_{f,t}^{j,z,(K)})^*
\tag{27}
$$

$$
\frac{\delta L_t^{j,z,(K)}}{\delta \bar{\gamma}_f^{j,z,(K)}} = -\frac{1}{\bar{\gamma}_f^{j,z,(K)}} + \left|\bar{\mu}_{f,t}^{j,z,(K)}\right|^2
\tag{28}
$$

$$\frac{\delta\bar{\gamma}_f^{j,z,(K)}}{\delta\lambda_f^{j,z,(K)}} = \exp\lambda_f^{j,z,(K)} \tag{29}$$

$$\frac{\delta L_t^{j,z,(K)}}{\delta\lambda_f^{j,z,(K)}} = 1 \tag{30}$$

$$\frac{\delta L_t^{j,z,(K)}}{A^{j,(K)}} = \bar{\pi}_t^{j,z,(K)} \tag{31}$$

$$\frac{\delta L_t^{j,z,(K)}}{b^{j,(K)}} = 1 \tag{32}$$

where $\mathrm{softmax}'(\cdot)$ is the derivative of the softmax function.

2.2. Lower layers ($k < K$)

To proceed downward through the network from layer k to layer $k-1$, there are two main "choke-points": $\bar{\mu}^{(k)}$ and $B^{(k)}$. The path to $\bar{\mu}^{(k)}$ is

$$\begin{array}{cccc} \mathcal{D}\ldots & \leftarrow \hat{X}^{(k)} & & \leftarrow \bar{\mu}^{(k)} \\ & & \leftarrow L^{(k)} \quad \swarrow & . \end{array} \tag{33}$$

On the way to $B^{(k)}$, paths through $\bar{\gamma}^{(k)}$ are required:

$$\begin{array}{cccc} \mathcal{D}\ldots & & \leftarrow \bar{\mu}^{(k)} & \leftarrow \bar{\gamma}^{(k)} \\ & \leftarrow L^{(k)} \quad \swarrow & & \end{array} \tag{34}$$

Finally, the paths to $B^{(k)}$ are

$$\begin{array}{cccc} \mathcal{D}\ldots & \leftarrow \bar{\mu}^{(k)} & & \leftarrow B^{(k)} \\ & & \leftarrow \bar{\gamma}^{(k)} \quad \swarrow & \end{array} \tag{35}$$

with Wirtinger gradients

$$\frac{\delta\bar{\mu}_{f,t}^{j',z,(k)}}{\delta\left[B_f^{(k)}\right]_{:,j}} = -\frac{1}{\bar{\gamma}_f^{j',z,(k)}}\left[B_f^{(k)}\right]_{:,j'}^H \psi_f \begin{cases} 0, & j = j' \\ \hat{X}_{f,t}^{j,(k)}, & j \neq j', \end{cases} \tag{36}$$

$$\frac{\delta\bar{\mu}_{f,t}^{j',(k)}}{\delta\left[B_f^{(k)}\right]_{:,j}^*} = \frac{1}{\bar{\gamma}_f^{j',z,(k)}} \begin{cases} y^T\psi_f - \left(\hat{X}_{f,t}^{\setminus j,(k)}\right)^T \left[B_f^{(k)}\right]_{:,\setminus j}^T \psi_f, & j = j' \\ 0, & j \neq j', \end{cases} \tag{37}$$

$$\frac{\delta\bar{\gamma}_f^{j',z,(k)}}{\delta\left[B_f^{(k)}\right]_{:,j}} = \left(\frac{\delta\bar{\gamma}_f^{j'(k)}}{\delta\left[B_f^{(k)}\right]_{:,j}^*}\right)^* = \begin{cases} \left[B_f^{(k)}\right]_{:,j'}^H \psi_f, & j = j' \\ 0, & j \neq j'. \end{cases} \tag{38}$$

Continuing downward, we go through $\hat{X}^{(k-1)}$:

$$\begin{array}{cccc} \mathcal{D}\ldots & \leftarrow \bar{\mu}^{(k)} & & \leftarrow \hat{X}^{(k-1)} \\ & & \leftarrow B^{(k)} \quad \swarrow & \end{array} \tag{39}$$

The top path uses gradient

$$\frac{\delta\bar{\mu}_{f,t}^{j,z,(k)}}{\delta\hat{X}_{f,t}^{j,(k-1)}} = \left(-\frac{1}{\bar{\gamma}_f^{j,z,(k)}}\left[B_f^{(k)}\right]_{:,j}^H \psi_f \left[B_f^{(k)}\right]_{:,\setminus j}\right). \tag{40}$$

The bottom path is a good example of a Wirtinger gradient chain rule, and is given by

$$\begin{aligned} \frac{\delta\mathcal{D}}{\delta\hat{X}_{f,t}^{j,(k-1)}} =& \frac{\delta\mathcal{D}}{\delta B_f^{(k)}} \odot \frac{\delta B_f^{(k)}}{\delta\hat{X}_{f,t}^{j,(k-1)}} + \frac{\delta\mathcal{D}}{\delta B_f^{(k)*}} \odot \frac{\delta B_f^{(k)*}}{\delta\hat{X}_{f,t}^{j,(k-1)}} \\ =& \frac{\delta\mathcal{D}}{\delta B_f^{(k)}} \odot \frac{\delta B_f^{(k)}}{\delta\hat{X}_{f,t}^{j,(k-1)}} + \left(\frac{\delta\mathcal{D}}{\delta B_f^{(k)}}\right)^* \odot \left(\frac{\delta B_f^{(k)}}{\delta\hat{X}_{f,t}^{j,(k-1)*}}\right)^* \\ =& 0 + \left(\frac{\delta\mathcal{D}}{\delta B_f^{(k)}}\right)^* \odot \left(y_{f,t}\left(\mathrm{diag}\hat{\Sigma}_f^{\hat{X}\hat{X},(k)}\right)^{-1}\right) \end{aligned} \tag{41}$$

Next we proceed to $\bar{\pi}^{(k-1)}$:

$$\begin{array}{ccccc} \mathcal{D}... & & \leftarrow \hat{X}^{(k-1)} & \leftarrow \bar{\pi}^{(k-1)} \\ & \leftarrow B^{(k)} & & \swarrow \end{array} \tag{42}$$

with

$$\frac{\delta B_f^{(k)}}{\delta \bar{\pi}_t^{j,z,(k-1)}} = -\sum_f \hat{\Sigma}_f^{Y\hat{X}(k)} \left[\hat{\Sigma}_f^{\hat{X}\hat{X},(k)}\right]^{-2} \left(\frac{1}{\bar{\gamma}_f^{j,z,(k-1)}} + \left| \bar{\mu}_{f,t}^{j,z,(k-1)} \right|^2 \right) \tag{43}$$

The gradient from $\bar{\pi}^{(k-1)}$ to $L^{(k-1)}$ is, as before, the derivative of the sigmoid function. Now we proceed to $\bar{\mu}^{(k-1)}$:

$$\begin{array}{ccccc} \mathcal{D}... & & \leftarrow \hat{X}^{(k-1)} & & \leftarrow \bar{\mu}^{(k-1)} \\ & & & \leftarrow L^{(k-1)} & \swarrow \\ & \leftarrow B^{(k)} & & & \swarrow \end{array} \tag{44}$$

with

$$\frac{\delta B_f^{(k)}}{\delta \bar{\mu}_{f,t}^{j,z,(k-1)}} = \left(\frac{\delta B_f^{(k)}}{\delta \left(\bar{\mu}_{f,t}^{j,z,(k-1)}\right)^*} \right)^* = -\hat{\Sigma}_f^{Y\hat{X},(k)} \left[\hat{\Sigma}_f^{\hat{X}\hat{X}(k)}\right]^{-2} \bar{\pi}_t^{j,z,(k-1)} \left(\bar{\mu}_{f,t}^{j,z,(k-1)}\right)^* \tag{45}$$

As a last step, we proceed to $B^{(k-1)}$, which requires intermediate paths through $\bar{\gamma}^{(k-1)}$:

$$\begin{array}{ccccc} \mathcal{D}... & & & \leftarrow \bar{\mu}^{(k-1)} & \leftarrow \bar{\gamma}^{(k-1)} \\ & & \leftarrow L^{(k-1)} & & \swarrow \\ & \leftarrow B^{(k)} & & & \swarrow \end{array} \tag{46}$$

with

$$\frac{\delta B_f^{(k)}}{\delta \bar{\gamma}_f^{j,z,(k-1)}} = -\sum_t \hat{\Sigma}_f^{Y\hat{X}(k)} \left[\hat{\Sigma}^{\hat{X}\hat{X}(k)}\right]^{-2} \left(-\frac{\bar{\pi}_t^{j,z,(k-1)}}{\left(\bar{\gamma}_f^{j,z,(k-1)}\right)^2} \right). \tag{47}$$

Finally, we reach $B^{(k-1)}$, which concludes the required gradients down through layer $k-1$.

$$\begin{array}{cccc} \mathcal{D}... & \leftarrow \bar{\mu}^{(k-1)} & & \leftarrow B^{(k-1)} \\ & & \leftarrow \bar{\gamma}^{(k-1)} & \swarrow \qquad . \end{array} \tag{48}$$

The gradients with respect to the trainable parameters $\lambda^{(k-1)}$, $A^{(k-1)}$, and b^{k-1} are the same as in layer K, and use gradients (29), (30), (31), and (32).

3. REFERENCES

[1] S. Wisdom, J. Hershey, J. Le Roux, and S. Watanabe, "Deep Unfolding for Multichannel Source Separation," in submission to *ICASSP*, 2016.

[2] K. Kreutz-Delgado, "The Complex Gradient Operator and the CR-Calculus," *arXiv:0906.4835 [math]*, June 2009, arXiv: 0906.4835.