Chapter 1

Mathematical Proofs

1.1 Complex Weights Initialization [?]

For complex-valued deep learning traditional approaches of Glorot [**?**] and He [**?**], are no more suitable to be used for weights initialization. So we need to re-derive, or at least adapt, those efficient methods also for the complex domain.
Given a generic neural network layer, let's call $\mathbf{w} \in \mathbb{C}$ its set of complex-valued weights, that we prefer written in polar form:

$$\mathbf{w} = \|\mathbf{w}\| e^{i\theta}$$

The variance of this set is defined as

$$\mathrm{Var}(\mathbf{w}) = \mathbb{E}\left[\mathbf{w}\bar{\mathbf{w}}\right] - \left(\mathbb{E}\left[\mathbf{w}\right]\right)^2 = \mathbb{E}\left[\|\mathbf{w}\|^2\right] - \left(\mathbb{E}\left[\mathbf{w}\right]\right)^2$$

that, if the parameters are symmetrically distributed around 0, reduces to $\mathbb{E}\left[\|\mathbf{w}\|^2\right]$.
In order to compute these quantities, we rely on the fact that the magnitude of a standard complex normally distributed variable follows the Rayleigh distribution [1], i.e. basically a Chi-Square with two degrees of freedom. Just for knowledge, its probability density function depends only on a single parameter and writes

$$f(x;\sigma) = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)}$$

So we can find a relation among the variances of $\mathbf{w}$ and its magnitude:

$$\mathrm{Var}(\|\mathbf{w}\|) = \mathrm{Var}\,\mathbf{w} - \left(\mathbb{E}\left[\|\mathbf{w}\|\right]\right)^2 \qquad \longrightarrow \qquad \mathrm{Var}(\mathbf{w}) = \mathrm{Var}(\|\mathbf{w}\|) + \left(\mathbb{E}\left[\|\mathbf{w}\|\right]\right)^2$$

But now, that we know the analytical distribution followed by $\|\mathbf{w}\|$, we can derive also the two addends of the sum above:

$$\mathbb{E}\left[\|\mathbf{w}\|\right] = \sigma\sqrt{\frac{\pi}{2}}, \qquad \mathrm{Var}\,\|\mathbf{w}\| = \frac{4-\pi}{2}\sigma^2$$

The variance of $\mathbf{w}$ can thus be expressed in terms of its generating Rayleigh distribution's single parameter σ:

$$\mathrm{Var}(\mathbf{w}) = \frac{4-\pi}{2}\sigma^2 + \left(\sigma\sqrt{\frac{\pi}{2}}\right)^2 = 2\sigma^2$$

How can we determine σ?

- Following the Xavier initialization [**?**], we would exploit a normal distribution (or a truncated-normal) with variance $\mathrm{Var}(\mathbf{w}) = 2/(n_{in} + n_{out})$, with n_{in} and n_{out} being the number of input and output units, respectively. For continuity, we have now to set

$$\sigma = 1/\sqrt{n_{in} + n_{out}}$$

[1] source

- With the He initialization [**?**], instead, we used still a normal distribution, but with a variance depending only on input units, i.e. $\text{Var}(\mathbf{w}) = 2/n_{in}$, for which we have to set correspondingly

$$\sigma = 1/\sqrt{n_{in}}$$

The magnitude of the complex parameters is then initialized using a Rayleigh distribution with an appropriate σ, while their phase (that never appeared in the equations) can be set uniformly in $[-\pi, \pi]$.

1.2 Stationary points of a real-valued function of a complex variable [?]

Let $f(z) : \mathbb{C} \to \mathbb{R}$ be a real-valued function of a complex variable z, and let's say that we want to find the extreme points of f (i.e. the values of z for which f is maximum or minimum) exploiting the differential calculus.
As explained also in **??**, differentiability in the complex plane is a quite strong assumption, and there are many function for which the stationarity condition in a point z_0,

$$\left.\frac{\partial f}{\partial z}\right|_{z=z_0}$$

cannot even be computed, because the limit in **??** is not the same from any direction.

A first approach to avoid the complex differentiability is to reformulate the problem in terms of two real variables, i.e. the real and imaginary parts of $z = x+iy$. Writing (with a minor abuse of notation) $f(z) = f(x, y)$, now the stationarity condition for a point z_0 becomes:

$$\left.\frac{\partial f}{\partial x}\right|_{z=z_0} = \left.\frac{\partial f}{\partial y}\right|_{z=z_0} = 0$$

While this method works, it is cumbersome because it involves the extra step of substituting $x + iy$ for z. Thus, we seek an equivalent method that works directly with the complex variable.

Assume now that f can be represented as $f(z) \equiv g(z, \bar{z})$, where g is an analytic function of two complex variables z and $\bar{z}$ that can be freely differentiated with respect to its arguments (because of the assumption of analyticity). Now, substituting $g(z, \bar{z})$ for $f(z)$, we can apply the chain rule of differential calculus to the stationarity conditions:

$$\frac{\partial f}{\partial x} = \frac{\partial g}{\partial z}\frac{\partial z}{\partial x} + \frac{\partial g}{\partial \bar{z}}\frac{\partial \bar{z}}{\partial x} = 0 \qquad \frac{\partial f}{\partial y} = \frac{\partial g}{\partial z}\frac{\partial z}{\partial y} + \frac{\partial g}{\partial \bar{z}}\frac{\partial \bar{z}}{\partial y} = 0$$

Evaluating the four derivatives, this becomes

$$\frac{\partial f}{\partial x} = \frac{\partial g}{\partial z} + \frac{\partial g}{\partial \bar{z}} = 0 \qquad \frac{\partial f}{\partial y} = i\frac{\partial g}{\partial z} - i\frac{\partial g}{\partial \bar{z}} = 0$$

which has a unique solution

$$\frac{\partial f}{\partial z} = \frac{\partial f}{\partial \bar{z}} = 0$$

We have uncovered that the stationary point can be found by taking the partial derivatives of $f(z)$ with respect to both z and $\bar{z}$, considering them as independent variables, and setting those derivatives to zero.

Notice that, when f is real-valued, the complex condition yield redundant information. In this case, in fact, both $\partial f/\partial x$ and $\partial f/\partial y$ must be real-valued, and so

$$\frac{\partial f}{\partial x} = \overline{\left(\frac{\partial f}{\partial x}\right)} \qquad \frac{\partial f}{\partial y} = \overline{\left(\frac{\partial f}{\partial y}\right)}$$

Plugging in the earlier expressions for these derivatives, we can solve them arriving at the conclusion that $\partial f/\partial z = \partial f/\partial \bar{z}$. This establish the redundancy and allows to rewrite the stationarity condition for a real-valued complex function:

$$\frac{\partial f}{\partial \bar{z}} = 0$$

These results are easily extended to the case of vectors of complex variables. Let $\mathbf{z} = [z_1, z_2, \dots, z_n]^T$ and assume that $g(\mathbf{z}, \bar{\mathbf{z}})$ is an analytic function of the complex vector $\mathbf{z}$ as well as its conjugate. Then the condition fo a stationary points becomes

$$\nabla_{\mathbf{z}} g = \mathbf{0} \qquad \nabla_{\bar{\mathbf{z}}} g = \mathbf{0}$$

or only the latter if g is real-valued.

1.3 Steepest complex gradient descent [?]

Once derived necessary and sufficient conditions for a certain $z_0 \in \mathbb{C}$ to be a stationary point of a real-valued function $f(z) : \mathbb{C} \to \mathbb{R}$, it is time to derive also the best optimization procedure. Using Wirtinger calculus we managed to overcome the strong requirements necessary to complex differentiability, but now we have twice the derivatives to compute ($\partial f/\partial z$ and $\partial f/\partial \bar{z}$). So, which direction should our gradient descent algorithm follow, in order to properly optimize f?

Let's start defining the gradient vector $\nabla_{\mathbf{z}} = [\partial/\partial z_1 \, , \partial/\partial z_2 \, , \dots, \partial/\partial z_n]$ for the vector $\mathbf{z} = [z_1, z_2, \dots, z_n]$ with $z_k = z_{k,Re} + i z_{k,Im}$ in order to write the first order Taylor series expansion for a function $g(\mathbf{z}, \bar{\mathbf{z}}) : \mathbb{C}^n \times \mathbb{C}^n \to \mathbb{R}$,

$$\Delta g = \langle \Delta \mathbf{z}, \nabla_{\bar{\mathbf{z}}} g \rangle + \langle \Delta \bar{\mathbf{z}}, \nabla_{\mathbf{z}} g \rangle = 2 \operatorname{Re} \{ \langle \Delta \mathbf{z}, \nabla_{\bar{\mathbf{z}}} g \rangle \}$$

where $\langle \cdot, \cdot \rangle$ is the canonical *inner product* in $\mathbb{C}^n$, and the last equality holds because g is real-valued. Using the Cauchy-Schwarz inequality, it is easy to show that the first-order change in g will be maximized when $\Delta \mathbf{z}$ and the cogradient $\nabla_{\bar{\mathbf{z}}}$ are collinear. Hence, it is the gradient with respect to the conjugate of the variable that defines the direction of the maximum rate of change in the function with respect to $\mathbf{z}$, and not the ordinary gradient $\nabla_{\mathbf{z}}$.
Thus, the gradient optimization of g should use the update rule

$$\Delta \mathbf{z} = -\alpha \nabla_{\bar{\mathbf{z}}} g$$

as this form leads to a non-positive increment given by $\Delta g = -2\alpha \|\nabla_{\bar{\mathbf{z}}}\|^2$, while the same rule but exploiting the other gradient would result in an update of $\Delta g = -2\alpha \operatorname{Re} \{ \langle \nabla_{\bar{\mathbf{z}}} g, \nabla_{\mathbf{z}} g \}$, which are not guaranteed to be non-positive.