Oja's rule: Derivation, Properties *

October 2019

1 Derivation of Oja's rule

The update for each weight of the weight vector $\boldsymbol{w} = [w_1, \ldots, w_D]^T \in \mathbb{R}^D$ for Oja's rule [1, 2] reads:

$$w_i^{n+1} = \frac{w_i^n + \beta y^n x_i^n}{\sqrt{\sum_{i=0}^{D-1}(w_i^n + \beta y^n x_i^n)^2}} \tag{1}$$

where the index n denotes the iteration number, while D is the dimension of the data vector, β is the learning rate, and i is the neuron number. In vector notation

$$\boldsymbol{w}^{n+1} = \frac{\boldsymbol{w}^n + \beta y^n \boldsymbol{x}^n}{||\boldsymbol{w}^n + \beta y^n \boldsymbol{x}^n||_2} \tag{2}$$

The gradient of expression (1) with respect to the learning rate β is given by:

$$\frac{\partial w_i^{n+1}}{\partial \beta} = \frac{y^n x_i^n \sqrt{\sum_{i=0}^{D-1}(w_i^n + \beta y^n x_i^n)^2} - \frac{1}{2}\Big(\sum_{i=0}^{D-1}(w_i^n + \beta y^n x_i^n)^2\Big)^{-\frac{1}{2}}\Big(\sum_{i=0}^{D-1} 2y^n x_i^n (w_i^n + \beta y^n x_i^n)\Big)\Big(w_i^n + \beta y^n x_i^n\Big)}{\sum_{i=0}^{D-1}(w_i^n + \beta y^n x_i^n)^2} \tag{3}$$

where we used the quotient rule $(f/g)' = (f'g - g'f)/g^2$ and that $(f(g(x))' = f'(g(x))g'(x)$. By simplifying the expression 3, we get:

$$\frac{\partial w_i^{n+1}}{\partial \beta} = \frac{y^n x_i^n}{\sqrt{\sum_{i=0}^{D-1}(w_i^n + \beta y^n x_i^n)^2}} - \frac{\Big(\sum_{i=0}^{D-1} y^n x_i^n (w_i^n + \beta y^n x_i^n)\Big)\Big(w_i^n + \beta y^n x_i^n\Big)}{\Big(\sum_{i=0}^{D-1}(w_i^n + \beta y^n x_i^n)\Big)^{\frac{3}{2}}} \tag{4}$$

Next, we evaluate the derivative at the linearization point $\beta^\star = 0$ to get:

$$\frac{\partial w_i^{n+1}}{\partial \beta}\bigg|_{\beta^\star=0} = \frac{y^n x_i^n}{\sqrt{\sum_{i=0}^{D-1}(w_i^n)^2}} - \frac{\Big(\sum_{i=0}^{D-1} y^n x_i^n (w_i^n)\Big)\Big(w_i^n\Big)}{\Big(\sum_{i=0}^{D-1}(w_i^n)\Big)^{\frac{3}{2}}}. \tag{5}$$

However, since we assume that the weights are normalized to one, we get $||\boldsymbol{w}^n||_w = \sum_{i=0}^{D-1}(w_i^n)^2$, that simplifies equation (5) to:

$$\frac{\partial w_i^{n+1}}{\partial \beta}\bigg|_{\beta^\star=0} = y^n x_i^n - w_i^n \sum_{i=0}^{D-1} y^n x_i^n w_i^n = y^n\Big(x_i^n - w_i^n \sum_{i=0}^{D-1} x_i^n w_i^n\Big) = y^n\Big(x_i^n - w_i^n y^n\Big) \tag{6}$$

In vector notation:

$$\frac{\partial \boldsymbol{w}^{n+1}}{\partial \beta}\bigg|_{\beta^\star=0} = y^n\Big(\boldsymbol{x}^n - \boldsymbol{w}^n y^n\Big) \tag{7}$$

This expression is the same as Oja's rule. Now, if we write down the Taylor series of $\boldsymbol{w}^{n+1}$, we get:

$$\boldsymbol{w}^{n+1} = f(\beta) = f(\beta^\star) + \frac{f'(\beta^\star)}{1!}(\beta - \beta^\star) + \frac{f''(\beta^\star)}{2!}(\beta - \beta^\star)^2 + \ldots \tag{8}$$

Since we are interested in a local approximation of f, terms in the order $\mathcal{O}(\beta^2)$ can be ignored, since:

$$(\beta - \beta^\star) \approx 10^{-5} \implies (\beta - \beta^\star)^2 \approx 10^{-10}. \tag{9}$$

*For typos/questions please contact: pvlachas@ethz.ch

As a consequence, we can write:

$$\boldsymbol{w}^{n+1} = f(\beta^\star) + f'(\beta^\star)\beta + \mathcal{O}(\beta^2) \tag{10}$$

with $\beta^\star = 0$ and ignore the $\mathcal{O}(\beta^2)$ terms. Since $f(\beta^\star) = \boldsymbol{w}^n$, we get:

$$\boldsymbol{w}^{n+1} \approx \boldsymbol{w}^n + \beta y^n \Big(\boldsymbol{x}^n - \boldsymbol{w}^n y^n\Big) \tag{11}$$

deriving Oja's rule as a linearization of the normalized Hebb's rule around $\beta^\star = 0$.

2 Proof of Oja's rule properties

In this section, we are going to prove some interesting properties of Oja's rule. Namely, if $\lim_{n\to infty} \boldsymbol{w} = \boldsymbol{w}^\star$:

1. If C is the covariance matrix of the data $C = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]$, with $C \in \mathbb{R}^{D\times D}$, it holds that $\boldsymbol{w}^\star$ is an eigenvalue of C, i.e. $C\boldsymbol{w}^\star = \lambda^\star \boldsymbol{w}^\star$. (this shows that $\boldsymbol{w}^\star$ is a PCA component).
2. $|\boldsymbol{w}^\star|_2 = 1$
3. $\boldsymbol{w}^\star$ is the principal component of the data, i.e. $\lambda^\star \geq \lambda_i$, $\forall$ other PCA components i.
4. $\boldsymbol{w}^\star$ maximizes the variance $\mathbb{E}[y^2]$ at the output of the perceptron over all $\boldsymbol{w}$ with $|\boldsymbol{w}|_2 = 1$.

3 Item 1: Eigenvector of C (PCA component of data)

We are going to prove these statements, one by one. Firstly, we need to show that $|\boldsymbol{w}^\star|_2 = 1$. Rewriting Oja's update rule (12), we have that:

$$\boldsymbol{w}^{n+1} = \boldsymbol{w}^n + \beta y^n \Big(\boldsymbol{x}^n - \boldsymbol{w}^n y^n\Big) \tag{12}$$

At equilibrium $\mathbb{E}[\Delta \boldsymbol{w}] = \mathbb{E}[\boldsymbol{w}^{n+1} - \boldsymbol{w}^n] = 0$ and $\lim_{n\to\infty} \boldsymbol{w}^n = \boldsymbol{w}^\star$. As a consequence, we get:

$$\mathbb{E}\Big[y^n \Big(\boldsymbol{x}^n - \boldsymbol{w}^\star y^n\Big)\Big] = 0 \implies \mathbb{E}\Big[y^n \boldsymbol{x}^n\Big] - \mathbb{E}\Big[\boldsymbol{w}^\star (y^n)^2\Big] = 0 \implies \mathbb{E}\Big[y^n \boldsymbol{x}^n\Big] - \boldsymbol{w}^\star \mathbb{E}\Big[(y^n)^2\Big] = 0. \tag{13}$$

Note that:

$$\mathbb{E}[y^n\, x_i^n] = \mathbb{E}\Big[\big(\sum_j w_j^\star x_j^n\big)\, x_i^n\Big] = \sum_j w_j^\star \underbrace{\mathbb{E}[x_j^n\, x_i^n]}_{C_{j,i}} \implies \mathbb{E}[y^n\, \boldsymbol{x}^n] = C\boldsymbol{w}^\star. \tag{14}$$

So equation (13) by plugging in the expression $\mathbb{E}[y^n\, \boldsymbol{x}^n] = C\boldsymbol{w}^\star$ and denoting $\mathbb{E}\Big[(y^n)^2\Big] = \sigma_y^2$ becomes:

$$C\boldsymbol{w}^\star - \boldsymbol{w}^\star \sigma_y^2 = 0 \implies C\boldsymbol{w}^\star = \sigma_y^2 \boldsymbol{w}^\star, \tag{15}$$

which proves item 1. Moreover, we have proven that the eigenvalue is equal to:

$$\lambda^\star = \sigma_y^2. \tag{16}$$

4 Item 2: Norm

In order to prove that the norm of $\boldsymbol{w}^\star$ is one, we first start by equation (15) and right-multiply with the vector $\boldsymbol{w}^{\star T}$ to get:

$$C\boldsymbol{w}^\star = \sigma_y^2 \boldsymbol{w}^\star \implies \boldsymbol{w}^{\star T} C \boldsymbol{w}^\star = \sigma_y^2 \boldsymbol{w}^{\star T}\boldsymbol{w}^\star \implies \boldsymbol{w}^{\star T} C \boldsymbol{w}^\star = \sigma_y^2 |\boldsymbol{w}^\star|_2^2. \tag{17}$$

However, by definition, we have that

$$\sigma_y^2 = \mathbb{E}[(y^n)^2] = \mathbb{E}[(\boldsymbol{w}^{\star T}\boldsymbol{x})(\boldsymbol{w}^{\star T}\boldsymbol{x})^T] = \mathbb{E}[\boldsymbol{w}^{\star T}(\boldsymbol{x}\boldsymbol{x}^T)\boldsymbol{w}^\star] = \boldsymbol{w}^{\star T}\mathbb{E}[(\boldsymbol{x}\boldsymbol{x}^T)]\boldsymbol{w}^\star = \boldsymbol{w}^{\star T} C \boldsymbol{w}^\star. \tag{18}$$

By plugging (17) in equation (18), we get:

$$\sigma_y^2 = \boldsymbol{w}^{\star T} C \boldsymbol{w}^\star \overbrace{=}^{eq.\,(17)} \sigma_y^2 |\boldsymbol{w}^\star|_2^2 \implies |\boldsymbol{w}^\star|_2^2 = 1, \tag{19}$$

proving item 2.

5 Item 3: Principal Component

In order to prove that $\boldsymbol{w}^\star$ is a principal component, we will perform a linear stability analysis around the convergence point $\boldsymbol{w}^\star$. We make the Ansatz

$$\boldsymbol{w}^{\star n} = \boldsymbol{w}^\star + \boldsymbol{\epsilon}^n, \tag{20}$$

where $\boldsymbol{\epsilon}^n$ is a small perturbation. This is expected to cause an update

$$\boldsymbol{w}^{\star n+1} = \text{Oja}(\boldsymbol{w}^{\star n}) = \text{Oja}(\boldsymbol{w}^\star) + \boldsymbol{\epsilon}^{n+1}, \tag{21}$$

where $\text{Oja}(\cdot)$ is the classical application of the Oja's rule (12). The question we ask is how will an initial perturbation propagate due to the update rule ? In other words, how does $\mathbb{E}[\delta\boldsymbol{\epsilon}] = \mathbb{E}[\boldsymbol{\epsilon}^{n+1} - \boldsymbol{\epsilon}^n]$ look like?

The expected update of Oja's rule around $\boldsymbol{w}$ reads:

$$\mathbb{E}[\Delta\boldsymbol{w}] = \beta\,\mathbb{E}[y\boldsymbol{x} - y^2\boldsymbol{w}] = \beta\left(C\boldsymbol{w} - \underbrace{(\boldsymbol{w}^T C\boldsymbol{w})}_{\sigma_y^2 = \mathbb{E}[y^2]}\boldsymbol{w}\right), \tag{22}$$

where we used previous results, from equations (14) and equation (18). Now, let's plug $\boldsymbol{w} = \boldsymbol{w}^\star + \boldsymbol{\epsilon}$ in the update rule. As we are simnplyfing the expression, we are ignoring $\mathcal{O}(\boldsymbol{\epsilon}^2)$ terms:

$$\frac{\mathbb{E}[\Delta(\boldsymbol{w}^\star + \boldsymbol{\epsilon}^n)]}{\beta} = C(\boldsymbol{w}^\star + \boldsymbol{\epsilon}^n) - ((\boldsymbol{w}^\star + \boldsymbol{\epsilon}^n)^T C(\boldsymbol{w}^\star + \boldsymbol{\epsilon}^n))(\boldsymbol{w}^\star + \boldsymbol{\epsilon}^n) = \tag{23}$$

$$= C\boldsymbol{w}^\star + C\boldsymbol{\epsilon}^n - (\boldsymbol{w}^{\star T} C\boldsymbol{w}^\star)\boldsymbol{w}^\star - 2(\boldsymbol{\epsilon}^{nT} C\boldsymbol{w}^\star)\boldsymbol{w}^\star \underbrace{-(\boldsymbol{\epsilon}^{nT} C\boldsymbol{\epsilon}^n)\boldsymbol{w}^\star}_{\mathcal{O}(\boldsymbol{\epsilon}^2)\approx 0} - (\boldsymbol{w}^{\star T} C\boldsymbol{w}^\star)\boldsymbol{\epsilon}^n - \underbrace{2(\boldsymbol{\epsilon}^{nT} C\boldsymbol{w}^\star)\boldsymbol{\epsilon}^n}_{\mathcal{O}(\boldsymbol{\epsilon}^2)\approx 0} - \underbrace{(\boldsymbol{\epsilon}^{nT} C\boldsymbol{\epsilon}^n)\boldsymbol{\epsilon}^n}_{\mathcal{O}(\boldsymbol{\epsilon}^3)\approx 0}. \tag{24}$$

Now, taking into account expression (24) and separating the terms that depend on $\boldsymbol{\epsilon}$ from those that do not, we get:

$$\frac{\mathbb{E}[\Delta(\boldsymbol{w}^\star + \boldsymbol{\epsilon}^n)]}{\beta} = \underbrace{C\boldsymbol{w}^\star - (\boldsymbol{w}^{\star T} C\boldsymbol{w}^\star)\boldsymbol{w}^\star}_{\text{Oja}(\mathbf{w}^\star)/\beta} + \underbrace{C\boldsymbol{\epsilon}^n - 2(\boldsymbol{\epsilon}^{nT} C\boldsymbol{w}^\star)\boldsymbol{w}^\star - (\boldsymbol{w}^{\star T} C\boldsymbol{w}^\star)\boldsymbol{\epsilon}^n}_{\boldsymbol{\epsilon}^{n+1}/\beta}. \tag{25}$$

We are interested in the error term $\boldsymbol{\epsilon}^{n+1}$ that shows how an initial perturbation to $\mathbf{w}^\star$ evolves. By rewriting the error and simplifying, we get:

$$\boldsymbol{\epsilon}^{n+1}/\beta = C\boldsymbol{\epsilon}^n - 2(\boldsymbol{\epsilon}^{nT} C\boldsymbol{w}^\star)\boldsymbol{w}^\star - (\boldsymbol{w}^{\star T} C\boldsymbol{w}^\star)\boldsymbol{\epsilon}^n = C\boldsymbol{\epsilon}^n - 2\lambda^\star(\boldsymbol{\epsilon}^{nT}\boldsymbol{w}^\star)\boldsymbol{w}^\star - \lambda^\star(\boldsymbol{w}^{\star T}\boldsymbol{w}^\star)\boldsymbol{\epsilon}^n \implies \tag{26}$$

$$\boldsymbol{\epsilon}^{n+1}/\beta = C\boldsymbol{\epsilon}^n - 2\lambda^\star(\boldsymbol{\epsilon}^{nT}\boldsymbol{w}^\star)\boldsymbol{w}^\star - \lambda^\star\boldsymbol{\epsilon}^n \tag{27}$$

where we used the fact that $\boldsymbol{w}^\star$ is an eigenvector with eigenvalue $\lambda^\star$. Now let's say $\boldsymbol{w}^\star$ is the eigenvector $\boldsymbol{v}^\star$. We need to proove that $\boldsymbol{v}^\star = \boldsymbol{v}_1$. Equation (27) becomes:

$$\boldsymbol{\epsilon}^{n+1}/\beta = C\boldsymbol{\epsilon}^n - 2\lambda^\star \underbrace{(\boldsymbol{\epsilon}^{nT}\boldsymbol{v}^\star)}_{\in\mathbb{R}}\boldsymbol{v}^\star - \lambda^\star\boldsymbol{\epsilon}^n. \tag{28}$$

If we left multiply with any other eigenvector $\boldsymbol{v}_j^T$ with $C\boldsymbol{v}_j = \lambda_j\boldsymbol{v}_j$ and $\lambda_j \neq \lambda^\star$, we get:

$$\frac{\boldsymbol{v}_j^T\boldsymbol{\epsilon}^{n+1}}{\beta} = \boldsymbol{v}_j^T C\boldsymbol{\epsilon}^n - 2\lambda^\star \underbrace{(\boldsymbol{\epsilon}^{nT}\boldsymbol{v}^\star)}_{\in\mathbb{R}}\underbrace{\boldsymbol{v}_j^T\boldsymbol{v}^\star}_{=0} - \lambda^\star\boldsymbol{v}_j^T\boldsymbol{\epsilon}^n, \tag{29}$$

where $\boldsymbol{v}_j^T\boldsymbol{v}^\star = 0$ due to orthogonality of the eigenvectors. Further, we know that

$$C\boldsymbol{v}_j = \lambda_j\boldsymbol{v}_j \implies (C\boldsymbol{v}_j)^T = \lambda_j\boldsymbol{v}_j^T \implies \boldsymbol{v}_j^T C^T = \lambda_j\boldsymbol{v}_j^T \implies \boldsymbol{v}_j^T C = \lambda_j\boldsymbol{v}_j^T \tag{30}$$

where we used that C is a symmetric real matrix (covariance matrix of independent variables). Then equation (29) becomes:

$$\frac{\boldsymbol{v}_j^T\boldsymbol{\epsilon}^{n+1}}{\beta} = \lambda_j\boldsymbol{v}_j^T\boldsymbol{\epsilon}^n - \lambda^\star\boldsymbol{v}_j^T\boldsymbol{\epsilon}^n = \left(\lambda_j - \lambda^\star\right)\boldsymbol{v}_j^T\boldsymbol{\epsilon}^n. \tag{31}$$

Expression (31) shows that convergence to $\boldsymbol{w}^\star$ can only be stable if $\lambda_j - \lambda^\star < 0$, so that the expected increase towards any other principal direction $\boldsymbol{v}_j$ is suppressed as $\frac{\boldsymbol{v}_j^T\boldsymbol{\epsilon}^{n+1}}{\beta} < 0$. If there is a principal component with $\lambda_j > \lambda^\star$, then $\frac{\boldsymbol{v}_j^T\boldsymbol{\epsilon}^{n+1}}{\beta} > 0$ will be positive and $\boldsymbol{w}^\star$ is not a stable convergence point of Oja's rule, and the updates will diverge towards the principal component. This shows that if Oja's rule converges, $\boldsymbol{w}^\star = \boldsymbol{v}_1$ is the principal component, with $\lambda^\star > \lambda_j$. Cases where the geometric multiplicity of the principal eigenvalue/eigenvector is greater than one (e.g. $\lambda_1 = \lambda_2$) are not considered in this analysis.

6 Item 4: Maximum variance

The variance σ_y^2 is (given in equation (17)):

$$\sigma_y^2 = \boldsymbol{w}^{\star T} C \boldsymbol{w}^\star. \tag{32}$$

Assume that the eigenvalue decomposition of C reads $C = V\Lambda V^{-1}$, where Λ is the diagonal matrix with the eigenvalues sorted in descending order ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$) in its diagonal:

$$\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_D \end{bmatrix}, \tag{33}$$

and $V = [\boldsymbol{v_1}, \ldots, \boldsymbol{v}_D]$ is the orthogonal matrix containing the eigenvectors of C (at the same time principal components of data matrix $X \in \mathbb{R}^{N \times D}$). We have proven that $\boldsymbol{w}^\star = \boldsymbol{v}_1$ is the principal component. Moreover, we know that the eigenvectors are orthogonal to each other, i.e. $\boldsymbol{v}_j^T \boldsymbol{v}_i = 0, \forall i \neq j$. **In the following we slightly abuse the notation for clarity and we denote the variance with $\sigma_y^2 = \sigma_{y,\boldsymbol{w}^\star}^2$** to illustrate that it is the variance at the output explained by the principal component. Expression (32) becomes:

$$\sigma_y^2 = \sigma_{y,\boldsymbol{w}^\star}^2 = \boldsymbol{w}^{\star T} V \Lambda V^{-1} \boldsymbol{w}^\star = \boldsymbol{v}_1^T V \Lambda V^{-1} \boldsymbol{v}_1 = \boldsymbol{v}_1^T [\boldsymbol{v_1}, \ldots, \boldsymbol{v}_D] \Lambda [\boldsymbol{v_1}, \ldots, \boldsymbol{v}_D]^T \boldsymbol{v}_1 = [1, 0, \ldots, 0] \Lambda [1, 0, \ldots, 0]^T \tag{34}$$

This leads to:

$$\sigma_{y,\boldsymbol{w}^\star}^2 = [1, 0, \ldots, 0] \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_D \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \implies \sigma_{y,\boldsymbol{w}^\star}^2 = \lambda_1. \tag{35}$$

Using the fact that $\sigma_{y,\boldsymbol{w}^\star}^2 = \lambda^\star$ given from equation (16), we end up with:

$$\sigma_{y,\boldsymbol{w}^\star}^2 = \lambda^\star = \lambda_1, \tag{36}$$

which shows that the variance "explained" in the output of the perceptron with weights $\boldsymbol{w}^\star$ is equal to the maximum eigenvalue of C. Is this the maximum variance that can be retained? Assume any other weight $\mathbf{w}$, which is **not** the principal component, and $|\boldsymbol{w}|_2 = 1$. V is orthonormal (C is symmetric so $V^T = V^{-1}$) and $\boldsymbol{z} = V^T \boldsymbol{w}$ also satisfies $|\boldsymbol{z}|_2 = 1$. The variance explained by any other weight $\mathbf{w}$ is given by

$$\sigma_{y,\boldsymbol{w}}^2 = \boldsymbol{w}^T V \Lambda V^{-1} \boldsymbol{w} = \boldsymbol{z}^T \Lambda \boldsymbol{z} = [z_1, z_2, \ldots, z_D] \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_D \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_D \end{bmatrix} = \sum_{i=1}^{D} \lambda_i z_i^2 \leq \lambda_1 \sum_{i=1}^{D} z_i^2 = \lambda_1 = \sigma_{y,\boldsymbol{w}^\star}^2. \tag{37}$$

As a consequence, the variance explained in the output by the principal component $\boldsymbol{w}^\star$ is the maximum.

References

[1] Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273, Nov 1982.

[2] Erkki Oja. Neural networks, principal components, and subspaces. *Int. J. Neural Syst.*, 1:61–68, 1989.