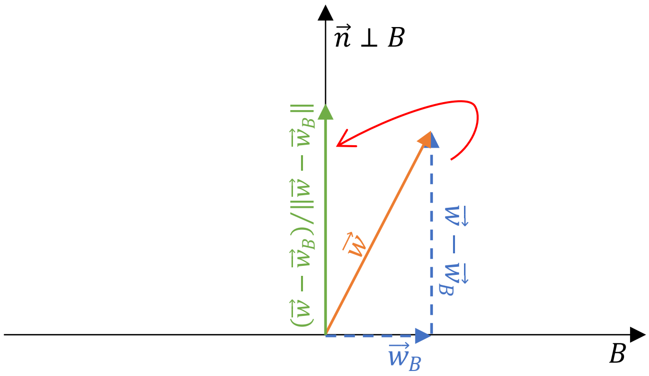
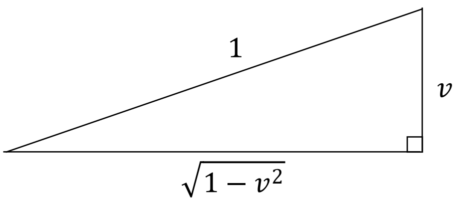
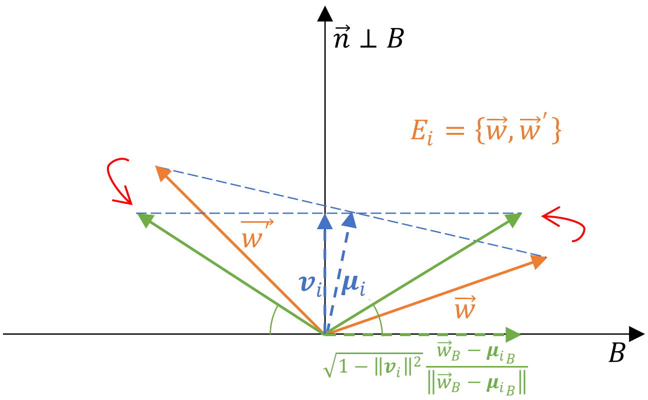
Hard Debiasing

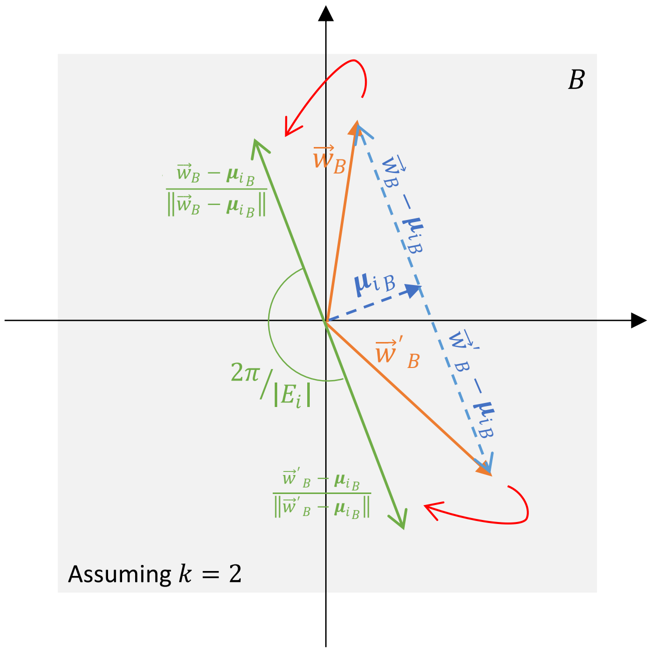
a) Neutralize: $\vec{w} \coloneqq (\vec{w} - \vec{w}_B)/\|\vec{w} - \vec{w}_B\| \quad , \forall \ w \in N$



b) Equalize: $\vec{w} \coloneqq \boldsymbol{v}_i + \sqrt{1 - \|\boldsymbol{v}_i\|^2} \frac{\vec{w}_B - \boldsymbol{\mu}_{i_B}}{\|\vec{w}_B - \boldsymbol{\mu}_{i_B}\|}$







Identify Bias Subspace

a) $B :=$ first k rows of $\mathbb{SVD}(C)$

Let the data matrix $\mathbf{X}$ be of $n \times p$ size, where n is the number of samples and p is the number of variables. Let us assume that it is *centered*, i.e. column means have been subtracted and are now equal to zero.

Then the $p \times p$ covariance matrix $\mathbf{C}$ is given by $\mathbf{C} = \mathbf{X}^\top \mathbf{X}/(n-1)$. It is a symmetric matrix and so it can be diagonalized:

$$\mathbf{C} = \mathbf{V}\mathbf{L}\mathbf{V}^\top,$$

where $\mathbf{V}$ is a matrix of eigenvectors (each column is an eigenvector) and $\mathbf{L}$ is a diagonal matrix with eigenvalues λ_i in the decreasing order on the diagonal. The eigenvectors are called *principal axes* or *principal directions* of the data. Projections of the data on the principal axes are called *principal components*, also known as *PC scores*; these can be seen as new, transformed, variables. The j-th principal component is given by j-th column of $\mathbf{XV}$. The coordinates of the i-th data point in the new PC space are given by the i-th row of $\mathbf{XV}$.

If we now perform singular value decomposition of $\mathbf{X}$, we obtain a decomposition

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\top,$$

where $\mathbf{U}$ is a unitary matrix and $\mathbf{S}$ is the diagonal matrix of singular values s_i. From here one can easily see that

$$\mathbf{C} = \mathbf{V}\mathbf{S}\mathbf{U}^\top\mathbf{U}\mathbf{S}\mathbf{V}^\top/(n-1) = \mathbf{V}\frac{\mathbf{S}^2}{n-1}\mathbf{V}^\top,$$

meaning that right singular vectors $\mathbf{V}$ are principal directions and that singular values are related to the eigenvalues of covariance matrix via $\lambda_i = s_i^2/(n-1)$. Principal components are given by $\mathbf{XV} = \mathbf{U}\mathbf{S}\mathbf{V}^\top\mathbf{V} = \mathbf{U}\mathbf{S}$.

Reference: https://bit.ly/2IJeAap

b) $C := \sum_{i=1}^{n} \sum_{w \in D_i} (\vec{w} - \boldsymbol{\mu}_i)^T (\vec{w} - \boldsymbol{\mu}_i)/|D_i|$

Assume all the sub-sample have the same sample size. If you had g sub-samples of size k (for a total of gk samples), then the variance of the combined sample depends on the mean E_j and variance V_j of each sub-sample:

$$Var(X_1, \ldots, X_{gk}) = \frac{k-1}{gk-1}(\sum_{j=1}^{g} V_j + \frac{k(g-1)}{k-1} Var(E_j)),$$

where by $Var(E_j)$ means the variance of the sample means.

Reference: https://bit.ly/2s6xqB9

Paper assumptions: $Var(E_j) \approx 0$

Note: Removing the scalar $(k-1)/(gk-1)$ does not change the *eigenvectors* (i.e. the *principle-directions*). It only serves to scale up the *eigenvalues* equally, which does not change the order of the principle directions.