We can easily extend eq. (11) to multiple examples $\{s_1, s_2 \cdots s_b\}$ and write the gradient descent update (using learning rate η) equations as:

$$\begin{aligned} \dot{f}_\theta(s_j) = -\eta \sum_i \tilde{\alpha}_i(s_j) w_i \quad &, \quad \dot{w}_i = -\eta \sum_j \tilde{\alpha}_i(s_j) f_\theta(s_j) \\ \implies \dot{f}_\theta = -\eta A W^T \quad &, \quad \dot{W} = -\eta f_\theta^T A \end{aligned} \tag{12}$$

where

$$A_{ij} = \begin{cases} \alpha_j(s_i) - 1 & \text{if } c_i = j \\ \alpha_j(s_i) & \text{else} \end{cases} \qquad (i^{th} \text{ example, } s_i\text{, belongs to the class } j)$$

B.4. A useful matrix algebra result

Lemma B.1. *Let $W(t)$ be a time-varying matrix with singular value decomposition (SVD): $W(t) = U(t)S(t)V(t)^T$, where $U(t)$ and $V(t)$ are orthogonal matrices corresponding to the left and right singular vectors, respectively, and $S(t) = diag(\sigma_1(t), \sigma_2(t), \ldots, \sigma_k(t))$ contains the singular values along its diagonal. Let $u_k(t)$ and $v_k(t)$ denote the k^{th} column vectors of $U(t)$ and $V(t)$, respectively. Then the time derivative of the k^{th} singular value, $\sigma_k(t)$, is given by:*

$$\dot{\sigma}_k(t) = u_k(t)^T \dot{W}(t) v_k(t)$$

Proof. For sake of brevity, we will drop the explicit time-dependence of each matrix from the notations. Let us write the singular vector decomposition (SVD) of matrix, $W = USV^T$. Using the product rule of differentiation:

$$\begin{aligned} \dot{W} &= \dot{U}SV^T + U\dot{S}V^T + US\dot{V}^T \\ \implies U^T\dot{W}V &= U^T\dot{U}S + \dot{S} + S\dot{V}^T V \\ \implies \dot{S} &= U^T\dot{W}V - U^T\dot{U}S - S\dot{V}^T V \\ \implies \dot{\sigma_k} &= u_k^T \dot{W} v_k - u_k^T \dot{u_k}\sigma_k - \sigma_k \dot{v_k}^T v_k \end{aligned} \tag{13}$$

where the last line is the expression for the k^{th} diagonal element of S. By definition of orthonormal vectors, $u_k^T u_k = 1$. So, $\dot{u_k}^T u_k + u_k^T \dot{u_k} = 0$. Since $\dot{u_k}^T u_k$ is a scalar, $\dot{u_k}^T u_k = u_k^T \dot{u_k}$. Therefore, $\dot{u_k}^T u_k = 0$. Similarly, $\dot{v_k}^T v_k = 0$. Therefore,

$$\dot{\sigma}_k = u_k^T \dot{W} v_k$$

□

B.5. Formal versions of theoretical results and proofs

Theorem B.2. *Let $f_\theta = U_1 S_1 V_1^T$ and $W = U_2 S_2 V_2^T$ denote the respective singular value decompositions (SVDs) of non-degenerate matrices f_θ and W, respectively. If the system is initialized such that $f_\theta^T f_\theta = WW^T$, then it holds that:*

$$V_1 = U_2 \quad , \quad S_1^2 = S_2^2$$

Proof. Let us start from the learning dynamics imposed by gradient-descent:

$$\dot{f}_\theta = -\eta A W^T \quad , \quad \dot{W} = -\eta f_\theta^T A \tag{14}$$

Let us write f_θ and W as their respective singular value decomposed form, i.e. say $f_\theta = U_1 S_1 V_1^T$ and $W = U_2 S_2 V_2$. Consider the dynamics of $f_\theta^T f_\theta$ and WW^T:

$$\begin{aligned}\frac{d}{dt}(f_\theta^T f_\theta) &= \dot{f_\theta}^T f_\theta + f_\theta \dot{f}_\theta = (-\eta A W^T)^T f_\theta + f_\theta^T(-\eta A W^T) \\ &= -\eta W A^T f_\theta - \eta f_\theta^T A W^T \end{aligned} \tag{15}$$

$$\begin{aligned}\frac{d}{dt}(WW^T) &= \dot{W}W^T + W\dot{W}^T = (-\eta f_\theta^T A)W^T + W(-\eta f_\theta^T A)^T \\ &= -\eta f_\theta^T A W^T - \eta W A^T f_\theta \end{aligned} \tag{16}$$

From eqs. (15) and (16), it is clear that $\frac{d}{dt}(f_\theta^T f_\theta) = \frac{d}{dt}(WW^T)$, i.e. $f_\theta^T f_\theta = WW^T + C$, for some constant C. If we assume the initialization to be such that $C = 0$ and f_θ and W are non-degenerate, we have:

$$f_\theta^T f_\theta = WW^T \implies V_1 S_1^2 V_1^T = U_2 S_2^2 U_2^T$$

By uniqueness of SVD (for positive semi-definite matrices):

$$\boxed{\begin{aligned} V_1 = U_2 &\implies V_1^T U_2 = I \\ S_1^2 &= S_2^2 \end{aligned}}$$

□

Theorem B.3. *Let f_θ, W be the matrices whose dynamics are governed by the gradient-descent equations as previously defined. Given the conditions from Theorem B.2, the magnitude of the time derivatives of the i^{th} singular values of f_θ and W are proportional to their respective singular values:*

$$\|\dot{\sigma}_{1i}\| \propto \sigma_{1i}$$
$$\|\dot{\sigma}_{2i}\| \propto \sigma_{2i}$$

Furthermore, assuming uniform class prediction at initialization and that number of classes, $|\mathcal{V}| \gg 1$, the time derivatives are bounded by the dominant class size:

$$\|\dot{\sigma}_{1i}\|, \|\dot{\sigma}_{2i}\| \propto \mathcal{O}(\mathcal{N}(c^{(0)}))$$

where $\mathcal{N}(c^{(0)})$ denotes the number of instances belonging to the dominant class $c^{(0)}$.

Proof. Let us start from the results of Theorem B.2: $S_1^2 = S_2^2 \implies \sigma_{1i}^2 = \sigma_{2i}^2 \; \forall i$. So, $\sigma_{1i} = \pm\sigma_{2i}$. Using this relation, we can simplify the expression of σ_{1i} dynamics. From Theorem B.1,

$$\begin{aligned}\dot{\sigma}_{1i} &= u_{1i}^T \dot{f}_\theta v_{1i} = -\eta u_{1i}^T A W^T v_{1i} \\ &= -\eta u_{1i}^T A (U_2 S_2 V_2^T)^T v_{1i} = -\eta u_{1i}^T A V_2 S_2 U_2^T v_{1i} \\ &= -\eta u_{1i}^T A V_2 S_2 V_1^T v_{1i} \qquad \text{[Using Theorem B.2]} \\ &= -\eta \sum_j (u_{1i}^T A v_{2j}) \sigma_{2j} (v_{1j} v_{1i}) = -\eta \sum_j (u_{1i}^T A v_{2j}) \sigma_{2j} \delta_{i=j} \\ \implies \dot{\sigma}_{1i} &= -\eta (u_{1i}^T A v_{2i}) \sigma_{2i} \end{aligned} \tag{17}$$

Similarly, we can simplify the dynamics for σ_{2i}:

$$\dot{\sigma}_{2i} = -\eta(u_{1i}^T A v_{2i})\sigma_{1i} \tag{18}$$

For sake of brevity, let us denote $(u_{1i}^T A v_{2i}) = g_i$. Using the relationship between σ_{1i} and σ_{2i}, we can simplify eqs. (17) and (18) as:

$$\dot{\sigma}_{1i} = -\eta g_i(\pm\sigma_{1i}) = \mp\eta g_i \sigma_{1i} \quad , \quad \dot{\sigma}_{2i} = -\eta g_i(\pm\sigma_{2i}) = \mp\eta g_i \sigma_{2i} \tag{19}$$

$$\boxed{\implies \|\dot{\sigma}_{1i}\| \propto \sigma_{1i} \quad , \quad \|\dot{\sigma}_{2i}\| \propto \sigma_{2i}} \tag{20}$$

Also, note that $g_i = u_{1i}^T A v_{2i} = \sum_{j,k} u_{1ij} A_{jk} v_{2ik}$, where $A_{jk} = \{\alpha_k(s_j) - 1, \alpha_k(s_j)\}$. Therefore, $A_{jk} \in (-1, 1)$.

At initialization, WLOG $\alpha_k(s_j) \approx \frac{1}{|\mathcal{V}|}\ \forall j, k$, i.e. uniform class prediction. Additionally, assuming $|\mathcal{V}| >> 1$, we can estimate g_i as the following:

$$\begin{aligned}
g_i &= \sum_{j,k} u_{1ij} A_{jk} v_{2ik} = \sum_k \left(\sum_{j\in\{c_j=k\}} u_{1ij}(\alpha_k(s_j) - 1)v_{2ik} + \sum_{j\in\{c_j\neq k\}} u_{1ij}\alpha_k(s_j)v_{2ik} \right) \\
\implies g_i &\approx \sum_k \left((\frac{1}{|\mathcal{V}|} - 1) \sum_{j\in\{c_j=k\}} u_{1ij}v_{2ik} + \frac{1}{|\mathcal{V}|} \sum_{j\in\{c_j\neq k\}} u_{1ij}v_{2ik} \right) \\
&\approx -(\sum_k v_{2ik})(\sum_{j\in\{c_j=k\}} u_{1ij}) = \mathcal{O}(\mathcal{N}(c^0))
\end{aligned} \tag{21}$$

where $c^{(0)}$ denotes the dominant class, i.e. the class with most number of instances. Combining eq. (21) with eq. (19), we get the desired result:

$$\boxed{\|\dot{\sigma}_{1i}\|, \|\dot{\sigma}_{2i}\| \propto \mathcal{O}(\mathcal{N}(c^0))} \tag{22}$$

$\square$