The Multiverse Loss for Robust Transfer Learning Supplementary

1. Omitted proofs

We provide proofs that were omitted from the paper for lack of space. We follow the same theorem numbering as in the paper.

Lemma 1. *The minimizers F^*, b^* of L are not unique, and it holds that for any vector $v \in \mathbb{R}^c$ and scalar s, the solutions $F^* + v\mathbf{1}_c^\top$, $b^* + s\mathbf{1}_c$ are also minimizers of L.*

Proof. denoting $V = v\mathbf{1}_c^\top$, $\mathbf{s} = s\mathbf{1}_c$,

$$\begin{aligned} L(F^* + V, b^* + \mathbf{s}, D, y) = \\ -\sum_{i=1}^{n} log(\frac{e^{d_i^\top f_{y_i} + d_i^\top v + b_{y_i} + s}}{\sum_{j=1}^{c} e^{d_i^\top f_j + d_i^\top v + b_j + s}}) \\ = -\sum_{i=1}^{n} log(\frac{e^{d_i^\top v + s} e^{d_i^\top f_{y_i} + b_{y_i}}}{\sum_{j=1}^{c} e^{d_i^\top v + s} e^{d_i^\top f_j + b_j}}) \\ = -\sum_{i=1}^{n} log(\frac{e^{d_i^\top v + s} e^{d_i^\top f_{y_i} + b_{y_i}}}{e^{d_i^\top v + s} \sum_{j=1}^{c} e^{d_i^\top f_j + b_j}}) \\ = -\sum_{i=1}^{n} log(\frac{e^{d_i^\top f_{y_i} + b_{y_i}}}{\sum_{j=1}^{c} e^{d_i^\top f_j + b_j}}) = L(F^*, b^*, D, y) \quad (1) \end{aligned}$$

□

The following simple lemma was not part of the paper. However, it is the reasoning behind the statement at the end of the proof of Thm. 1. "Since $\forall i, j \;\; p_i(j) > 0$ and since $rank(D)$ is full, $\sum_{i=1}^{n} d_i d_i^\top p_i(j) p_i(j')$ is PD."

Lemma 2. *Let $K = \sum_{i=1}^{n} d_i d_i^\top$ be a full rank $d \times d$ matrix, i.e., it is PD and not just PSD, then for all vector $q \in \mathbb{R}^n$ such that $\forall i \;\; q_i > 0$, the matrix $\hat{K} = \sum_{i=1}^{n} q_i d_i d_i^\top$ is also full rank.*

Proof. For every vector $v \in \mathbb{R}^d$, $v^\top \hat{K} v \geq (min_i q_i) v^\top K v > 0$. □

Theorem 3. *There exist a set of weights $F^1 = \left[f_1^1, f_2^1, ..., f_C^1\right], b^1, F^2 = \left[f_1^2, f_2^2, ..., f_C^2\right], b^2 ... F^m = [f_1^m, f_2^m, ..., f_C^m], b^m$ which are orthogonal $\forall jrs \;\; f_j^r \perp f_j^s$ for which the joint loss:*

$$J(F^1, b^1 ... F^m, b^m, D, y) = \sum_{r=1}^{m} L(F^r, b^r, D, y) \quad (2)$$

is bounded by:

$$\begin{aligned} mL^*(D, y) \leq J(F^1, b^1 ... F^m, b^m, D, y) \\ \leq mL^*(D, y) + \sum_{l=1}^{m-1} A_l \lambda_{d-j+1} \quad (3) \end{aligned}$$

where $[A_1 \dots A_{m-1}]$ are bounded parameters.

Proof. We again prove the theorem by constructing such a solution. Denoting by $v_{d-m+2} ... v_d$ the eigenvectors of K corresponding to $\lambda_{d-m+2} \dots \lambda_d$. Given $F^1 = F^*, b^1 = b^*$, we can construct each pair F^r, b^r as follows:

$$\begin{aligned} \forall j, r \qquad {f_j}^r = {f_1}^1 + \sum_{l=1}^{m-1} \alpha_{jlr} v_{d-l+1} \\ b^r = b^1 \quad (4) \end{aligned}$$

The tensor of parameters α_{jlr} is constructed to insure the orthogonality condition. Formally, α_{jlr} has to satisfy:

$$\forall j, r \neq s \qquad (f_j^1 + \sum_{l=1}^{m-1} \alpha_{jlr} v_{d-l+1})^\top f_j^s = 0 \quad (5)$$

Noticing that 5 constitutes a set of $\frac{1}{2} m(m-1)$ equations, it can be satisfied by the tensor α_{jlr} which contains $m(m-1)c$ parameters. Defining $\Psi^r = [\psi_1^r, \psi_2^r, \dots, \psi_c^r] = F^r -$

F^1, we have:

$$\begin{aligned}
L(F^1 + \Psi^r, b^1) &\leq L^*(D, y) \\
&+ \frac{1}{2}\sum_{j=1}^{c}\sum_{j'=j+1}^{c}(\psi_j - \psi_{j'})^\top K(\psi_j - \psi_{j'}) \\
&= L^*(D, y) \\
&+ \frac{1}{2}\sum_{j=1}^{c}\sum_{j'=j+1}^{c}\sum_{l=1}^{m-1}(\alpha_{jlr} - \alpha_{j'lr})^2 v_l^\top K v_l \\
&= L^*(D, y) \\
&+ \frac{1}{2}\sum_{j=1}^{c}\sum_{j'=j+1}^{c}\sum_{l=1}^{m-1}(\alpha_{jlr} - \alpha_{j'lr})^2 \lambda_{d-l+1} \quad (6)
\end{aligned}$$

Denoting $A_l = \frac{1}{2}\sum_{j=1}^{c}\sum_{j'=j+1}^{c}\sum_{r=1}^{m}(\alpha_{jlr} - \alpha_{j'lr})^2$ and summing over all solutions we obtain the bound:

$$J(F^1, b^1 ... F^m, b^m, D, y) \leq \sum_{l=1}^{m-1} A_l \lambda_{d-l+1} + mL^*(D, y) \quad (7)$$

We notice that if $\lambda_{d-m+2} = \lambda_{d-m+1} = \ldots \lambda_d = 0$ then $J(F^1, b^1 ... F^m, b^m, D, y) = mL^*(D, y)$. □

Below is the rigorous statement of Thm. 5 and its proof. In the proof we will be using two matrix inversion identities. The first one is the block matrix inversion identity, for a specific form of block matrices:

$$\begin{aligned}
&\begin{pmatrix} A & B \\ B & A \end{pmatrix}^{-1} \\
&= \begin{pmatrix} (A - BA^{-1}B)^{-1} & -A^{-1}B(A - BA^{-1}B)^{-1} \\ -A^{-1}B(A - BA^{-1}B)^{-1} & (A - BA^{-1}B)^{-1}) \end{pmatrix}
\end{aligned} \quad (8)$$

The second identity is the Kailath Variant of the Woodbury identity:

$$(A+BC)^{-1} = A^{-1} - A^{-1}B((I)+CA^{-1}B)^{-1}CA^{-1} \quad (9)$$

The proof of Thm. 5 will also be using the following lemma.

Lemma 3. *Let $\gamma_1 ... \gamma_d$ and $v_1 ... v_d$ be the generalized eigenvalues and eigenvectors of two positive definite matrices S_b, S_w, where S_b is invertible. The spectrum $\gamma'_1 ... \gamma'_d$ and eigenvectors $v'_1 ... v'_d$ of the generalized inverse problem $(S_b + S_w)^{-1} v'_i = \gamma'_i S_b^{-1} v'$ are given by $\gamma' = \frac{\gamma_i}{1+\gamma_i}$, $v'_i = (S_b + S_w)v_i$, and it holds that $S_b(S_b + S_w)^{-1} v'_i = \gamma'_i v'_i$.*

Proof. For the standard generalized problem we have $S_b v_i = \gamma_i S_w v_i$. Therefore, $(S_b + S_w)v_i = (1 + \frac{1}{\gamma_i})S_b v_i$. Let $v'_i = (S_b + S_w)v_i$, then $v'_i = (1 + \frac{1}{\gamma_i})S_b(S_b + S_w)^{-1} v'_i$. Multiplying both sides by $(S_b)^{-1}$ we have $S_b^{-1} v'_i = (1 + \frac{1}{\gamma_i})(S_b + S_w)^{-1} v'_i$, and finally $(S_b + S_w)^{-1} v'_i = \frac{\gamma_i}{1+\gamma_i} S_b^{-1} v'_i$. □

Theorem 5. *Given data representation D, mean μ and labels y, for any centered data point $\hat{d}_i = d_i - \mu$, we denote $d'_i = (S_b + S_w)^{-1}\hat{d}_i$. Given two centered data points $\hat{d}_1, \hat{d}_2$ such that the fisher ratios $\sigma(d'_1, S_b, S_w), \sigma(d'_2, S_b, S_w) < T$, it holds that:*

$$1 - 2T \leq \frac{\log P(d_1 - \mu, d_2 - \mu|H) + \eta_1}{\log P(d_1 - \mu, d_2 - \mu|I) + \eta_2} \leq 1 + 6T \quad (10)$$

Where $\eta 1, \eta 2$ are fixed constants.

Proof. We denote each data point $d_1 = \sum_{i=1}^{d} \alpha_i v'_i, d_2 = \sum_{i=1}^{d} \beta_i v'_i$ where $v'_1 ... v'_d$ are the eigenvectors of the generalized inverse eigen-problem $(S_b + S_w)^{-1} v'_i = \gamma'_i S_b^{-1} v'_i$. The probabilities $P(d_1, d_2|H), P(d_1, d_2|I)$ are modeled as zero mean gaussian densities with covariances:

$$\Sigma_H = \begin{pmatrix} S_b + S_w & S_b \\ S_b & S_b + S_w \end{pmatrix}, \Sigma_I = \begin{pmatrix} S_b + S_w & \mathbf{0} \\ \mathbf{0} & S_b + S_w \end{pmatrix} \quad (11)$$

Denoting $\hat{d}_1 = d_1 - \mu, \hat{d}_2 = d_2 - \mu, M = (S_b + S_w)^{-1} S_b$ using Eq. 8 we have that:

$$\begin{aligned}
\frac{\log P(\hat{d}_1, \hat{d}_2|H) + \eta_1}{\log P(\hat{d}_1, \hat{d}_2|I) + \eta_2} &= \frac{\begin{pmatrix} \hat{d}_1^\top & \hat{d}_2^\top \end{pmatrix} \Sigma_H^{-1} \begin{pmatrix} \hat{d}_1 \\ \hat{d}_2 \end{pmatrix}}{\begin{pmatrix} \hat{d}_1^\top & \hat{d}_2^\top \end{pmatrix} \Sigma_I^{-1} \begin{pmatrix} \hat{d}_1 \\ \hat{d}_2 \end{pmatrix}} \\
&= \frac{\hat{d}_1^\top (S_b + S_w - S_b M)^{-1} \hat{d}_1}{\hat{d}_1^\top (S_b + S_w)^{-1} \hat{d}_1 + \hat{d}_2^\top (S_b + S_w)^{-1} \hat{d}_2} \\
&+ \frac{\hat{d}_2^\top (S_b + S_w - S_b M)^{-1} \hat{d}_2}{\hat{d}_1^\top (S_b + S_w)^{-1} \hat{d}_1 + \hat{d}_2^\top (S_b + S_w)^{-1} \hat{d}_2} \\
&- \frac{\hat{d}_1^\top M (S_b + S_w - S_b M)^{-1} \hat{d}_2}{\hat{d}_1^\top (S_b + S_w)^{-1} \hat{d}_1 + \hat{d}_2^\top (S_b + S_w)^{-1} \hat{d}_2} \\
&- \frac{\hat{d}_2^\top M (S_b + S_w - S_b M)^{-1} \hat{d}_1}{\hat{d}_1^\top (S_b + S_w)^{-1} \hat{d}_1 + \hat{d}_2^\top (S_b + S_w)^{-1} \hat{d}_2}
\end{aligned} \quad (12)$$

where the constants of the densities have been expressed by η_1, η_2 in the left hand side of the equation. Defining $M' = (S_b + S_w)^{-1} S_b (S_b + S_w)^{-1}$ and $S = M[\mathbf{I} + M^2]^{-1}$ by using Eq. 9:

$$(S_B + S_W - S_b M)^{-1} = SM' + (S_b + S_w)^{-1} \quad (13)$$

Therefore:

$$\frac{log(d_i,\mu|H)-\eta_1}{log(d_i,\mu|I)-\eta_2}=1 + \frac{\hat{d_1}^\top SM'\hat{d_1}+\hat{d_2}^\top SM'\hat{d_2}-\hat{d_1}^\top MSM'\hat{d_2}-\hat{d_2}^\top MSM'\hat{d_1}}{\hat{d_1}^\top(S_b+S_w)^{-1}\hat{d_1}+\hat{d_2}^\top(S_b+S_w)^{-1}\hat{d_2}} - \frac{2\hat{d_1}^\top M'\hat{d_2}}{\hat{d_1}^\top(S_b+S_w)^{-1}\hat{d_1}+\hat{d_2}^\top(S_b+S_w)^{-1}\hat{d_2}} \quad (14)$$

Defining $\rho_i=\frac{\gamma_i}{1+\gamma_i}$ We notice from Lemma 3 that $v_i'^\top M=\gamma_i v_i'^\top$, $v_i'^\top S=v_i'^\top M[\mathbf{I}+M^2]^{-1}=\frac{\rho_i}{1+\rho_i^2}v_i'^\top$ And so we can expand the first term in the numerator of Eq. 19:

$$\begin{aligned}\hat{d_1}^\top SM'\hat{d_1}&=(\sum_{i=1}^k\alpha_i v_i')^\top SM'(\sum_{i=1}^k\alpha_i v_i')\\&=(\sum_{i=1}^k\alpha_i\frac{\rho_i}{1+\rho_i^2}v_i')^\top M'(\sum_{i=1}^k\alpha_i v_i')\\&=\sum_{i=1}^k\alpha_i\frac{\rho_i}{1+\rho_i^2}v_i'^\top(S_b+S_w)^{-1}S_b(S_b+S_w)^{-1}\sum_{i=1}^k\alpha_i v_i'\end{aligned} \quad (15)$$

Since $v_i'=(S_b+S_w)v_i,\ \ \forall i\neq j\ \ v_i^\top S_b v_j=0$, and $\rho>0$ we get:

$$\begin{aligned}\hat{d_1}^\top SM'\hat{d_1}&=(\sum_{i=1}^k\alpha_i\frac{\rho_i}{1+\rho_i^2}v_i)^\top S_b(\sum_{i=1}^k\alpha_i v_i)\\&=\sum_{i=1}^k\alpha_i^2\frac{\rho_i}{1+\rho_i^2}v_i^\top S_b v_i\leq\sum_{i=1}^k\alpha_i^2 v_i^\top S_b v_i=\hat{d_1}^\top S_b\hat{d_1}\end{aligned} \quad (16)$$

Since it is also true that $\forall i\neq j\ \ v_i^\top(S_b+S_w)v_j=0$, similar manipulation can be done with the denominator:

$$\begin{aligned}&\hat{d_1}^\top(S_b+S_w)^{-1}\hat{d_1}+\hat{d_2}^\top(S_b+S_w)^{-1}\hat{d_2}\\&=\sum_{i=1}^k\alpha_i v_i^\top(S_b+S_w)\sum_{i=1}^k\alpha_i v_i\\&\quad+\sum_{i=1}^k\beta_i v_i^\top(S_b+S_w)\sum_{i=1}^k\beta v_i\\&=\hat{d_1}^\top(S_b+S_w)\hat{d_1}+\hat{d_2}^\top(S_b+S_w)\hat{d_2}\end{aligned} \quad (17)$$

And so:

$$\begin{aligned}0<&\frac{\hat{d_1}^\top SM'\hat{d_1}+\hat{d_2}^\top SM'\hat{d_2}}{\hat{d_1}^\top(S_b+S_w)^{-1}\hat{d_1}+\hat{d_2}^\top(S_b+S_w)^{-1}\hat{d_2}}\\&\leq\frac{\hat{d_1}^\top S_b\hat{d_1}+\hat{d_2}^\top S_b\hat{d_2}}{\hat{d_1}^\top(S_b+S_w)\hat{d_1}+\hat{d_2}^\top(S_b+S_w)\hat{d_2}}\leq 2T\end{aligned} \quad (18)$$

The last reasoning stems from the bound on fisher scores of d_1 and d_2 and the fact that if both $\frac{a_1}{a_2}$ and $\frac{b_1}{b_2}$ are smaller than T and all terms are positive, then $\frac{a_1+b_1}{a_2+b_2}$ is smaller than both $\frac{2*\max(a_1,b_1)}{a_2}$ and $\frac{2*\max(a_1,b_1)}{b_2}$, and therefore $\frac{a_1+b_1}{a_2+b_2}<2T$.

The same manipulations to the rest of the terms in the numerator of Eq. 19 and get the following bound:

$$\left|\frac{-\hat{d_1}^\top MSM'\hat{d_2}-\hat{d_2}^\top MSM'\hat{d_1}-2\hat{d_1}^\top M'\hat{d_2}}{\hat{d_1}^\top(S_b+S_w)^{-1}\hat{d_1}+\hat{d_2}^\top(S_b+S_w)^{-1}\hat{d_2}}\right|<4T\ , \quad (19)$$

from which the theorem stems. □

Theorem 6. *Let $f^1...f^m$ be a set of m classifiers that are S_w-orthogonal for data D and labels y, and let $\gamma=[\gamma_1...\gamma_d]$ denote the Fisher spectrum. Given that $\forall 1\leq r\leq m$, for some value θ, $\sigma(f^r,S_b,S_w)\geq\theta$, it holds that $\sum_{k=1}^d\gamma_k\geq\sqrt{m}\theta$.*

Proof. The Fisher spectrum $\gamma_1...\gamma_d$ is obtained from the eigenvalues of $R=S_b^{\frac{1}{2}}S_w^{-1}S_b^{\frac{1}{2}}$. For each classifier f^r we have:

$$\forall r,\ \ \frac{f^{r\top}S_b f^r}{f^{r\top}S_w f^r}\geq\theta \quad (20)$$

Denoting $u^r=\frac{S_w^{\frac{1}{2}}f^r}{||S_w^{\frac{1}{2}}f^r||}$, we have:

$$\forall r,\ \ u^{r\top}S_w^{-\frac{1}{2}}S_b S_w^{-\frac{1}{2}}u^r=u^{r\top}\hat{R}u^r\geq\theta \quad (21)$$

Denoting by $\hat{\gamma}=[\hat{\gamma_1}...\hat{\gamma_d}]$ and $w_1...w_d$ the eigenvalues and eigenvectors of $\hat{R}$, we notice that $\sum_{k=1}^d\hat{\gamma_k}=\sum_{k=1}^d\gamma_k$, since the matrix $\hat{R}$ is a cyclic permutation of $R=S_b^{\frac{1}{2}}S_w^{-1}S_b^{\frac{1}{2}}$, and hence have equal trace. The eigenvectors of $\hat{R}$ span a d dimensional linear subspace, and so we can express each $u^r=\sum_{k=1}^d\alpha_k^r w_k, ||\alpha^r||_2=1$. From the S_w orthogonality property of the solutions $f^1...f^m$, it follows that $\forall r\neq s,\ \ u^{r\top}u^s=\alpha^{r\top}\alpha^s=0$ We therefor have:

$$\forall r,\ \ \sum_{k=1}^d\alpha_k^r w_k^\top\hat{R}\sum_{k=1}^d\alpha_k^r w_k=\sum_{k=1}^d(\alpha_k^r)^2\hat{\gamma}_k. \quad (22)$$

In matrix form:

$$\begin{pmatrix} \alpha_1^1 ... \alpha_d^1 \\ \alpha_1^2 ... \alpha_d^2 \\ \vdots \\ \alpha_1^r ... \alpha_d^r \end{pmatrix} \begin{pmatrix} \hat{\gamma_1} & 0 & ...0 \\ 0 & \hat{\gamma_2} & 0...0 \\ & \vdots & \\ 0 & 0 & ... \quad \hat{\gamma_d} \end{pmatrix} \begin{pmatrix} \alpha_1^1 ... \alpha_d^1 \\ \alpha_1^2 ... \alpha_d^2 \\ \vdots \\ \alpha_1^r ... \alpha_d^r \end{pmatrix}^\top = \Delta\Gamma\Delta^\top$$
$$\geq \begin{pmatrix} \theta \\ \theta \\ \vdots \\ \theta \end{pmatrix} \tag{23}$$

We notice that Δ consists of orthonormal vectors, and hence we can take the L_2 norm on both sides of Eq. 23:

$$||\Delta\Gamma\Delta^\top||_2 = ||\Gamma||_2 = \sum_{k=1}^{d} \sqrt{(\hat{\gamma_k})^2} \leq$$
$$\sum_{k=1}^{d} \hat{\gamma_k} = \sum_{k=1}^{d} \gamma_k \geq \sqrt{m}\theta \tag{24}$$

□

2. Network architecture

In our experiments, we employ three network architectures. For the CIFAR-100 experiments, we use the architecture of network in network [18]; for the face recognition experiments, we use an architecture similar to the scratch architecture [41] for most of our experiments (Denoted by $N1$). We also use a higher capacity network similar to [24] for further evaluation (Denoted by $N2$). The networks were trained from scratch at each experiment.

All networks are fully convolutional, and we added a hidden layer on top of the $N1$ network to apply our method on top of a vector of activations. This modification is not strictly needed and was made for implementation convenience. This top layer was used as the representation. The architectures used are fully described in the supplementary material.

Layer	Filter/Stride	#Channel	#Filter
Conv11	5×5 / 1	3	192
Conv12	1×1 / 1	192	160
Conv13	1×1 / 1	160	96
Pool1	3×3 / 2	96	–
Dropout1-0.5	–	–	–
Conv21	5×5 / 1	96	192
Conv22	1×1 / 1	192	192
Conv23	1×1 / 1	192	100
Pool2	3×3 / 2	192	–
Dropout1-0.5	–	–	–
Conv31	3×3 / 1	192	192
Conv32	1×1 / 1	192	192
Conv33	1×1 / 1	192	100
Avg Pool	7×7 / 1	100	–
FC	1×100 / 1	100	100

Table 1. The modified NIN [18] model used in the CIFAR-100 experiments. The network starts with a color input image of size $3 \times 32 \times 32$ pixels, and runs through 3 convolutional blocks interleaved with relue and max pooling layers. Following a spatial average pooling at the end of the process, a representation of size 100 is obtained. A FC layer of size 100 was added to the architecture for reasons of implementation convenience.

Layer	Filter/Stride	#Channel	#Filter
Conv11	3×3 / 1	1	32
Conv12	3×3 / 1	32	64
Max Pool	2×2 / 2	64	–
Conv21	3×3 / 1	64	64
Conv22	3×3 / 1	64	128
Max Pool	2×2 / 2	128	–
Conv31	3×3 / 1	128	96
Conv32	3×3 / 1	96	192
Max Pool	2×2 / 2	192	–
Conv41	3×3 / 1	192	128
Conv42	3×3 / 1	128	256
Max Pool	2×2 / 2	256	–
Conv51	3×3 / 1	256	160
Conv52	3×3 / 1	160	320
Avg Pool	6×6 / 1	320	–
FC	1×1 / 1	320	512
Dropout1-0.5	–	–	–

Table 2. The $N1$ network architecture. The network starts with a gray scale input image of size $1 \times 100 \times 100$ pixels, and runs through 10 convolutional layers interleaved relu and max pooling layers. Following a spatial average pooling at the end of the process, a representation of size 512 is obtained. A FC layer of size 512 was added to the scratch architecture.

Layer	Filter/Stride	#Channel	#Filter
Conv11	3×3 / 1	3	64
Conv12	3×3 / 1	64	64
Max Pool	2×2 / 2	64	–
Conv21	3×3 / 1	64	128
Conv22	3×3 / 1	128	128
Max Pool	2×2 / 2	128	–
Conv31	3×3 / 1	128	256
Conv32	3×3 / 1	256	256
Conv33	3×3 / 1	256	256
Max Pool	2×2 / 2	256	–
Conv41	3×3 / 1	256	512
Conv42	3×3 / 1	512	512
Conv43	3×3 / 1	512	512
Max Pool	2×2 / 2	512	–
Conv51	3×3 / 1	512	512
Conv52	3×3 / 1	512	512
Conv53	3×3 / 1	512	512
Max Pool	2×2 / 2	512	–
FC	4×3 / 0	512	6144
Dropout1-0.5	–	–	–
FC	1×1 / 0	6144	512

Table 3. The $N2$ network architecture. The network starts with an RGB input image of size $1 \times 162 \times 120$ pixels, and runs through 10 convolutional layers interleaved relu and max pooling layers. Following a spatial average pooling at the end of the process, a representation of size 512 is obtained.