Additional Materials

In this document we provide additional materials, to which we refer in discussions with the reviewers of the paper "Class Distribution Shifts in Zero-Shot Learning: Learning Robust Representations".

1 Optimal representation in the motivating parametric model

We follow the notation in the paper, and begin by revisiting the parametric model introduced in Section 2.

Let $z_i|c_i \sim \mathcal{N}(c_i, \Sigma_z)$, where $\Sigma_z = \nu_z I_d$, $0 < \nu_z \in \mathbb{R}$ and I_d is the d dimensional identity matrix. Classes c_i are drawn according to a Gaussian distribution $c_i \sim \mathcal{N}(0, \Sigma_a)$ corresponding to their type $a \in \{a_1, a_2\}$, where Σ_a are diagonal matrices of the form

$$\Sigma_{a_1} = \operatorname{diag}\big(\overbrace{\nu_0, \ldots, \nu_0}^{d_0}, \overbrace{\nu_1, \ldots, \nu_1}^{d_1}, \overbrace{\nu_2, \ldots, \nu_2}^{d_2}\big),$$
$$\Sigma_{a_2} = \operatorname{diag}\big(\nu_0, \ldots, \nu_0, \nu_2, \ldots, \nu_2, \nu_1, \ldots, \nu_1\big),$$

with $0 < \nu_2 < \nu_z < \nu_0 < \nu_1$. Denote the proportion of samples of type a_1 by ρ.

We consider linear representations $g(z) = Wz$, where W is a diagonal matrix with diagonal $w \in \mathbb{R}^d$, and the contrastive loss

$$\ell(z_i, z_j, y_{ij}; d_g) \coloneqq y_{ij}\, d_g^2(z_i, z_j) + (1 - y_{ij}) \max\{0, m - d_g(z_i, z_j)\}^2, \tag{1}$$

where $d_g(z_i, z_j) \coloneqq \|g(z_i - z_j)\|^2 = \|W(z_i - z_j)\|^2$ ($\|\cdot\|$ denotes the Euclidean norm[1]).

1.1 Derivation of the expected loss

Since Σ_z is of full rank, it suffices to consider a simpler version of the loss, without the hinge, that is

$$\widetilde{\ell}(z_i, z_j, y_{ij}; d_g) := y_{ij} \|W(z_i - z_j)\|^4 + (1 - y_{ij}) \left(m - \|W(z_i - z_j)\|^2\right)^2. \tag{2}$$

For a balanced sample of positive and negative examples, the expected loss is given by

$$\begin{aligned}\mathbb{E}\left[\widetilde{\ell}(z_i, z_j, y_{ij}; d_g)\right] = \frac{1}{2}\mathbb{E}_{y_{ij}=1}\left[\|W(z_i - z_j)\|^4\right] \\ + \frac{1}{2}\mathbb{E}_{y_{ij}=0}\left[m^2 - 2m\|W(z_i - z_j)\|^2 + \|W(z_i - z_j)\|^4\right].\end{aligned} \tag{3}$$

To calculate the expression above, we use the following lemma:

Lemma 1. *Let $\mu \in \mathbb{R}^d$ be a random variable and let $t|\mu \sim \mathcal{N}(\mu, \Sigma)$. If $\mu \equiv 0$ (constant), then*

1. $\mathbb{E}\|t\|^4 = 2\operatorname{tr}(\Sigma^2) + \operatorname{tr}^2(\Sigma)$.

If $\mu \sim \mathcal{N}(0, \Sigma_\mu)$, then

2. $\mathbb{E}\|t\|^2 = \operatorname{tr}(\Sigma) + \operatorname{tr}(\Sigma_\mu)$,
3. $\mathbb{E}\|t\|^4 = 2\operatorname{tr}(\Sigma^2) + 4\operatorname{tr}(\Sigma\Sigma_\mu) + \operatorname{tr}^2(\Sigma) + 2\operatorname{tr}(\Sigma)\operatorname{tr}(\Sigma_\mu) + 2\operatorname{tr}(\Sigma_\mu^2) + \operatorname{tr}^2(\Sigma_\mu)$.

[1] Squared distance is selected for its simplicity in computing the expected value of even powers of the Euclidean norm of Gaussian variables.

Proof. For any random variable $u \in \mathbb{R}^d$, such that $u \sim \mathcal{N}(\mu_u, \Sigma_u)$, and any symmetric matrix A, we have

$$\mathbb{E}_u[u^T A u] = \operatorname{tr}(A\Sigma_u) + \mu_u^T A \mu_u, \tag{4}$$

$$\mathbb{E}_u[u^T A u]^2 = 2\operatorname{tr}\left((A\Sigma_u)^2\right) + 4\mu_u^T A \Sigma_u A \mu_u + \left(\operatorname{tr}(A\Sigma_u) + \mu_u^T A \mu_u\right)^2 \tag{5}$$

(see, for example, Thm. 3.2b.2 in Mathai and Provost (1992)).

First, letting $\mu_u = 0$, $\Sigma_u = \Sigma$ and $A = I_d$ in (5) we get

$$\mathbb{E}\,\|t\|^4 = \mathbb{E}[t^T t]^2 = 2\operatorname{tr}(\Sigma^2) + \operatorname{tr}^2(\Sigma). \tag{6}$$

Now, assume that $\mu \sim \mathcal{N}(0, \Sigma_\mu)$. From (4) we get $\mathbb{E}_\mu\,\|\mu\|^2 = \mathbb{E}_\mu[\mu^T \mu] = \operatorname{tr}(\Sigma_\mu)$, and thus

$$\mathbb{E}\,\|t\|^2 = \mathbb{E}_\mu\left[\mathbb{E}_{t|\mu}[t^T t \mid \mu]\right] = \mathbb{E}_\mu\left[\operatorname{tr}(\Sigma) + \mu^T \mu\right] = \operatorname{tr}(\Sigma) + \operatorname{tr}(\Sigma_\mu). \tag{7}$$

Similarly, from (5) we have

$$\mathbb{E}\,\|t\|^4 = \mathbb{E}_\mu\left[\mathbb{E}_{t|\mu}[[t^T t]^2 \mid \mu]\right] = 2\operatorname{tr}(\Sigma^2) + 4\,\mathbb{E}_\mu[\mu^T \Sigma \mu] + \operatorname{tr}^2(\Sigma) + 2\operatorname{tr}(\Sigma)\,\mathbb{E}_\mu\,\|\mu\|^2 + \mathbb{E}_\mu\,\|\mu\|^4. \tag{8}$$

By substituting $A = \Sigma$ in (4) we get $\mathbb{E}_\mu[\mu^T \Sigma \mu] = \operatorname{tr}(\Sigma\Sigma_\mu)$, and from (5) we have $\mathbb{E}_\mu\,\|\mu\|^4 = 2\operatorname{tr}(\Sigma_\mu^2) + \operatorname{tr}^2(\Sigma_\mu)$. Therefore,

$$\mathbb{E}\,\|t\|^4 = 2\operatorname{tr}(\Sigma^2) + 4\operatorname{tr}(\Sigma\Sigma_\mu) + \operatorname{tr}^2(\Sigma) + 2\operatorname{tr}(\Sigma)\operatorname{tr}(\Sigma_\mu) + 2\operatorname{tr}(\Sigma_\mu^2) + \operatorname{tr}^2(\Sigma_\mu). \tag{9}$$

□

Now, note that $W(z_i - z_j) \sim \mathcal{N}(\mu, \Sigma)$, with $\mu = W(c_i - c_j)$ and $\Sigma = 2\nu_z W^T W$.

If $y_{ij} = 1$, then z_i and z_j are from the same class, meaning that $c_i = c_j$ and thus $\mu = 0$. Therefore, by Lemma 1.(1) we have

$$\begin{aligned}
\mathbb{E}_{y_{ij}=1}\,\|W(z_i - z_j)\|^4 &= 2\operatorname{tr}\left(\Sigma^2\right) + \operatorname{tr}^2(\Sigma) \\
&= 2 \cdot 4\nu_z^2 \operatorname{tr}\left(\left[W^T W\right]^2\right) + 4\nu_z^2 \operatorname{tr}^2\left(W^T W\right) \\
&= 8\nu_z^2 \sum_{i=1}^{d} w_i^4 + 4\nu_z^2 \left(\sum_{i=1}^{d} w_i^2\right)^2 .
\end{aligned} \tag{10}$$

However, for pairs from different classes, that is, when $y_{ij} = 0$, the mean μ is itself a Gaussian random variable distributed according to $\mathcal{N}(0, \Sigma_\mu)$, where

$$\Sigma_\mu = \begin{cases} W^T(2\Sigma_{a_1})W & c_i, c_j \text{ are both of type } a_1 \\ W^T(2\Sigma_{a_2})W & c_i, c_j \text{ are both of type } a_2 \\ W^T(\Sigma_{a_1} + \Sigma_{a_2})W & c_i, c_j \text{ are of different types .} \end{cases} \tag{11}$$

Therefore, by Lemma 1.(2) we have

$$\begin{aligned}
\mathbb{E}_{y_{ij}=0}\,\|W(z_i - z_j)\|^2 &= \mathbb{E}_{y_{ij}=0}\left[\operatorname{tr}(\Sigma_\mu) + \operatorname{tr}(\Sigma)\right] = \mathbb{E}_{y_{ij}=0}\left[\operatorname{tr}(\Sigma_\mu)\right] + \operatorname{tr}(\Sigma) \\
&= \rho^2 \operatorname{tr}\left(2W^T \Sigma_{a_1} W\right) + (1-\rho)^2 \operatorname{tr}\left(2W^T \Sigma_{a_2} W\right) \\
&\quad + 2\rho(1-\rho)\operatorname{tr}\left(W^T(\Sigma_{a_1} + \Sigma_{a_2})W\right) + \operatorname{tr}(\Sigma) \\
&= 2\left[(\nu_0 + \nu_z)\sum_{i=1}^{d_0} w_i^2 + (\alpha_1 + \nu_z)\sum_{i=d_0+1}^{d_0+d_1} w_i^2 + (\alpha_2 + \nu_z)\sum_{i=d_0+d_1+1}^{d} w_i^2\right],
\end{aligned} \tag{12}$$

where we denote for short

$$\begin{aligned}
\alpha_1 &\coloneqq \rho^2\nu_1 + (1-\rho)^2\nu_2 + \rho(1-\rho)(\nu_1 + \nu_2) = \rho\nu_1 + (1-\rho)\nu_2, \\
\alpha_2 &\coloneqq \rho^2\nu_2 + (1-\rho)^2\nu_1 + \rho(1-\rho)(\nu_1 + \nu_2) = \rho\nu_2 + (1-\rho)\nu_1, \\
\beta_1 &\coloneqq 2\rho^2\nu_1^2 + 2(1-\rho)^2\nu_2^2 + \rho(1-\rho)(\nu_1 + \nu_2)^2, \\
\beta_2 &\coloneqq 2\rho^2\nu_2^2 + 2(1-\rho)^2\nu_1^2 + \rho(1-\rho)(\nu_1 + \nu_2)^2.
\end{aligned}$$

Finally, by Lemma 1.(3) we have

$$\begin{aligned}\mathbb{E}_{y_{ij}=0}\left\|w\left(z_i-z_j\right)\right\|^4 &= \mathbb{E}_{y_{ij}=0}\left[2\operatorname{tr}\left(\Sigma^2\right)+4\operatorname{tr}\left(\Sigma\Sigma_\mu\right)+\operatorname{tr}^2(\Sigma)+2\operatorname{tr}\left(\Sigma\right)\operatorname{tr}\left(\Sigma_\mu\right)+2\operatorname{tr}\left(\Sigma_\mu^2\right)+\left(\operatorname{tr}\left(\Sigma_\mu\right)\right)^2\right]\\ &= 2\operatorname{tr}(\Sigma^2)+4\,\mathbb{E}_{y_{ij}=0}[\operatorname{tr}(\Sigma\Sigma_\mu)]+\operatorname{tr}^2(\Sigma)+2\operatorname{tr}(\Sigma)\,\mathbb{E}_{y_{ij}=0}[\operatorname{tr}(\Sigma_\mu)]\\ &\quad+2\,\mathbb{E}_{y_{ij}=0}[\operatorname{tr}(\Sigma_\mu^2)]+\mathbb{E}_{y_{ij}=0}[\operatorname{tr}^2(\Sigma_\mu)],\end{aligned} \tag{13}$$

where

$$\begin{aligned}\mathbb{E}_{y_{ij}=0}\left[\operatorname{tr}\left(\Sigma\Sigma_\mu\right)\right] &= 2\nu_z\operatorname{tr}\left(W^TW\left[2\rho^2W^T\Sigma_{a_1}W+2(1-\rho)^2W^T\Sigma_{a_2}W+2\rho(1-\rho)W^T\left(\Sigma_{a_1}+\Sigma_{a_2}\right)W\right]\right)\\ &= 4\nu_z\left[\nu_0\sum_{i=1}^{d_0}w_i^4+\alpha_1\sum_{i=d_0+1}^{d_0+d_1}w_i^4+\alpha_2\sum_{i=d_0+d_1+1}^{d}w_i^4\right];\end{aligned} \tag{14}$$

$$\begin{aligned}\mathbb{E}_{y_{ij}=0}\left[\operatorname{tr}\left(\Sigma_\mu\right)\right] &= \rho^2\operatorname{tr}\left(2W^T\Sigma_{a_1}W\right)+(1-\rho)^2\operatorname{tr}\left(2W^T\Sigma_{a_1}W\right)+2\rho(1-\rho)\operatorname{tr}\left(W^T\left(\Sigma_{a_1}+\Sigma_{a_2}\right)W\right)\\ &= 2\left[\nu_0\sum_{i=1}^{d_0}w_i^2+\alpha_1\sum_{i=d_0+1}^{d_0+d_1}w_i^2+\alpha_2\sum_{i=d_0+d_1+1}^{d}w_i^2\right],\end{aligned} \tag{15}$$

and so

$$\operatorname{tr}\left(\Sigma\right)\mathbb{E}_{y_{ij}=0}\left[\operatorname{tr}\left(\Sigma_\mu\right)\right] = 4\nu_z\left(\sum_{i=1}^{d}w_i^2\right)\left[\nu_0\sum_{i=1}^{d_0}w_i^2+\alpha_1\sum_{i=d_0+1}^{d_0+d_1}w_i^2+\alpha_2\sum_{i=d_0+d_1+1}^{d}w_i^2\right]; \tag{16}$$

$$\begin{aligned}\mathbb{E}_{y_{ij}=0}\left[\operatorname{tr}\left(\Sigma_\mu^2\right)\right] &= \rho^2\operatorname{tr}\left((2W^T\Sigma_{a_1}W)^2\right)+(1-\rho)^2\operatorname{tr}\left((2W^T\Sigma_{a_2}W)^2\right)+2\rho(1-\rho)\operatorname{tr}\left((W^T\left(\Sigma_{a_1}+\Sigma_{a_2}\right)W)^2\right)\\ &= 2\left[2\nu_0^2\sum_{i=1}^{d_0}w_i^4+\beta_1\sum_{i=d_0+1}^{d_0+d_1}w_i^4+\beta_2\sum_{i=d_0+d_1+1}^{d}w_i^4\right];\end{aligned} \tag{17}$$

and similarly

$$\begin{aligned}\mathbb{E}_{y_{ij}=0}\left[\operatorname{tr}^2\left(\Sigma_\mu\right)\right] &= 2\left[2\nu_0^2\left(\sum_{i=1}^{d_0}w_i^2\right)^2+\beta_1\left(\sum_{i=d_0+1}^{d_0+d_1}w_i^2\right)^2+\beta_2\left(\sum_{i=d_0+d_1+1}^{d}w_i^2\right)^2\right.\\ &\quad\left.+4\gamma_{0,1}\sum_{i=1}^{d_0}w_i^2\sum_{i=d_0+1}^{d_0+d_1}w_i^2+4\gamma_{0,2}\sum_{i=1}^{d_0}w_i^2\sum_{i=d_0+d_1+1}^{d}w_i^2+4\gamma_{1,2}\sum_{i=d_0+1}^{d_0+d_1}w_i^2\sum_{i=d_0+d_1+1}^{d}w_i^2\right],\end{aligned} \tag{18}$$

where we denote for short

$$\begin{aligned}\gamma_{0,1} &\coloneqq \rho^2\nu_0\nu_1+(1-\rho)^2\nu_0\nu_2+\rho(1-\rho)\nu_0\left(\nu_1+\nu_2\right),\\ \gamma_{0,2} &\coloneqq \rho^2\nu_0\nu_2+(1-\rho)^2\nu_0\nu_1+\rho(1-\rho)\nu_0\left(\nu_1+\nu_2\right),\\ \gamma_{1,2} &\coloneqq 2\rho^2\nu_1\nu_2+2(1-\rho)^2\nu_1\nu_2+\rho(1-\rho)\left(\nu_1+\nu_2\right)^2.\end{aligned}$$

1.2 Derivation of the optimal representation

In order to derive the optimal representation, we differentiate the expected loss with respect to the squared values in the diagonal of W, that is, w_i^2:

$$\frac{\partial}{\partial\left(w_i^2\right)}\operatorname{tr}\left(\Sigma^2\right)=8\nu_z^2 w_i^2 \tag{19}$$

$$\frac{\partial}{\partial\left(w_i^2\right)}\operatorname{tr}^2\left(\Sigma\right)=8\nu_z^2\sum_{j=1}^{d}w_j^2 \tag{20}$$

$$\frac{\partial}{\partial\left(w_i^2\right)}\mathbb{E}_{y=0}\left[\operatorname{tr}\left(\Sigma\Sigma_\mu\right)\right]=\begin{cases}8\nu_z\nu_0 w_i^2 & 1\le i\le d_0\\ 8\nu_z\alpha_1 w_i^2 & d_0+1\le i\le d_0+d_1\\ 8\nu_z\alpha_2 w_i^2 & d_0+d_1+1\le i\le d\end{cases} \tag{21}$$

$$\frac{\partial\left[\operatorname{tr}(\Sigma)\,\mathbb{E}_{y=0}[\operatorname{tr}\Sigma_\mu]\right]}{\partial\left(w_i^2\right)}=\begin{cases}4\nu_z\left[2\nu_0\sum\limits_{j=1}^{d_0}w_j^2+(\alpha_1+\nu_0)\sum\limits_{j=d_0+1}^{d_0+d_1}w_j^2+(\alpha_2+\nu_0)\sum\limits_{j=d_0+d_1+1}^{d}w_j^2\right] & 1\le i\le d_0\\ 4\nu_z\left[(\nu_0+\alpha_1)\sum\limits_{j=1}^{d_0}w_j^2+2\alpha_1\sum\limits_{j=d_0+1}^{d_0+d_1}w_j^2+(\alpha_2+\alpha_1)\sum\limits_{j=d_0+d_1+1}^{d}w_j^2\right] & d_0+1\le i\le d_0+d_1\\ 4\nu_z\left[(\nu_0+\alpha_2)\sum\limits_{j=1}^{d_0}w_j^2+(\alpha_1+\alpha_2)\sum\limits_{j=d_0+1}^{d_0+d_1}w_j^2+2\alpha_2\sum\limits_{j=d_0+d_1+1}^{d}w_j^2\right] & d_0+d_1+1\le i\le d\end{cases} \tag{22}$$

$$\frac{\partial}{\partial\left(w_i^2\right)}\mathbb{E}_{y=0}\left[\operatorname{tr}\left(\Sigma_\mu^2\right)\right]=\begin{cases}8\nu_0^2 w_i^2 & 1\le i\le d_0\\ 4\beta_1 w_i^2 & d_0+1\le i\le d_0+d_1\\ 4\beta_2 w_i^2 & d_0+d_1+1\le i\le d\end{cases} \tag{23}$$

$$\frac{\partial}{\partial\left(w_i^2\right)}\mathbb{E}_{y=0}\left[\operatorname{tr}^2\left(\Sigma_\mu\right)\right]=\begin{cases}8\nu_0^2\sum\limits_{j=1}^{d_0}w_j^2+8\gamma_{0,1}\sum\limits_{j=d_0+1}^{d_0+d_1}w_j^2+8\gamma_{0,2}\sum\limits_{j=d_0+d_1+1}^{d}w_j^2 & 1\le i\le d_0\\ 8\gamma_{0,1}\sum\limits_{j=1}^{d_0}w_j^2+4\beta_1\sum\limits_{j=d_0+1}^{d_0+d_1}w_j^2+8\gamma_{1,2}\sum\limits_{j=d_0+d_1+1}^{d}w_j^2 & d_0+1\le i\le d_0+d_1\\ 8\gamma_{0,2}\sum\limits_{j=1}^{d_0}w_j^2+8\gamma_{1,2}\sum\limits_{j=d_0+1}^{d_0+d_1}w_j^2+4\beta_2\sum\limits_{j=d_0+d_1+1}^{d}w_j^2 & d_0+d_1+1\le i\le d\end{cases} \tag{24}$$

Combining these results, we get for $1\le i\le d_0$

$$\begin{aligned}\partial_0 \coloneqq \frac{\partial}{\partial\left(w_i^2\right)}\widetilde{\ell}\left(z_i,z_j,y_{ij};d_g\right)=\frac{1}{2}&\left[2\cdot 8\nu_z^2 w_i^2+8\nu_z^2\sum_{j=1}^{d}w_j^2\right]-m\left[2\left(\nu_0+\nu_z\right)\right]\\ &+8\nu_z^2 w_i^2+4\nu_z^2\sum_{j=1}^{d}w_j^2+2\cdot 8\nu_z\nu_0 w_i^2\\ &+4\nu_z\left[2\nu_0\sum_{j=1}^{d_0}w_j^2+(\alpha_1+\nu_0)\sum_{j=d_0+1}^{d_0+d_1}w_j^2+(\alpha_2+\nu_0)\sum_{j=d_0+d_1+1}^{d}w_j^2\right]\\ &+8\nu_0^2 w_i^2+\frac{1}{2}\left[8\nu_0^2\sum_{j=1}^{d_0}w_j^2+8\gamma_{0,1}\sum_{j=d_0+1}^{d_0+d_1}w_j^2+8\gamma_{0,2}\sum_{j=d_0+d_1+1}^{d}w_j^2\right],\end{aligned}$$

for $d_0 + 1 \le i \le d_0 + d_1$

$$\begin{aligned}
\partial_1 \coloneqq \frac{\partial}{\partial\left(w_i^2\right)}\widetilde{\ell}\left(z_i, z_j, y_{ij}; d_g\right) = \frac{1}{2}\left[2 \cdot 8\nu_z^2 w_i^2 + 8\nu_z^2 \sum_{j=1}^{d} w_j^2\right] - m\left[2\left(\alpha_1 + \nu_z\right)\right] \\
+ 8\nu_z^2 w_i^2 + 4\nu_z^2 \sum_{j=1}^{d} w_j^2 + 2 \cdot 8\nu_z \alpha_1 w_i^2 \\
+ 4\nu_z\left[\left(\nu_0 + \alpha_1\right)\sum_{j=1}^{d_0} w_j^2 + 2\alpha_1 \sum_{j=d_0+1}^{d_0+d_1} w_j^2 + \left(\alpha_2 + \alpha_1\right)\sum_{j=d_0+d_1+1}^{d} w_j^2\right] \\
+ 4\beta_1 w_i^2 + \frac{1}{2}\left[8\gamma_{0,1}\sum_{j=1}^{d_0} w_j^2 + 4\beta_1 \sum_{j=d_0+1}^{d_0+d_1} w_j^2 + 8\gamma_{1,2}\sum_{j=d_0+d_1+1}^{d} w_j^2\right],
\end{aligned}$$

and similarly for $d_0 + d_1 + 1 \le i \le d$

$$\begin{aligned}
\partial_2 \coloneqq \frac{\partial}{\partial\left(w_i^2\right)}\widetilde{\ell}\left(z_i, z_j, y_{ij}; d_g\right) = \frac{1}{2}\left[2 \cdot 8\nu_z^2 w_i^2 + 8\nu_z^2 \sum_{j=1}^{d} w_j^2\right] - m\left[2\left(\alpha_2 + \nu_z\right)\right] \\
+ 8\nu_z^2 w_i^2 + 4\nu_z^2 \sum_{j=1}^{d} w_j^2 + 2 \cdot 8\nu_z \alpha_2 w_i^2 \\
+ 4\nu_z\left[\left(\nu_0 + \alpha_2\right)\sum_{j=1}^{d_0} w_j^2 + \left(\alpha_1 + \alpha_2\right)\sum_{j=d_0+1}^{d_0+d_1} w_j^2 + 2\alpha_2 \sum_{j=d_0+d_1+1}^{d} w_j^2\right] \\
+ 4\beta_2 w_i^2 + \frac{1}{2}\left[8\gamma_{0,2}\sum_{j=1}^{d_0} w_j^2 + 8\gamma_{1,2}\sum_{j=d_0+1}^{d_0+d_1} w_j^2 + 4\beta_2 \sum_{j=d_0+d_1+1}^{d} w_j^2\right].
\end{aligned}$$

Due to symmetry, at the optimal solution we have

$$w_i = \begin{cases} u_0 & 0 \le i \le d_0 \\ u_1 & d_0 + 1 \le i \le d_0 + d_1 \\ u_2 & d_0 + d_1 + 1 \le i \le d \end{cases} \tag{25}$$

and thus we can write

$$\partial_0 = -2m\left(\nu_0 + \nu_z\right) + u_0^2 A_{0,0} + u_1^2 A_{0,1} + u_2^2 A_{0,2}, \tag{26}$$

$$\partial_1 = -2m\left(\alpha_1 + \nu_z\right) + u_0^2 A_{1,0} + u_1^2 A_{1,1} + u_2^2 A_{1,2}, \tag{27}$$

$$\partial_2 = -2m\left(\alpha_2 + \nu_z\right) + u_0^2 A_{2,0} + u_1^2 A_{2,1} + u_2^2 A_{2,2}, \tag{28}$$

where

$$\begin{aligned}
A_{0,0} &= 16\nu_z^2 + 8\nu_z^2 d_0 + 16\nu_z\nu_0 + 8\nu_z\nu_0 d_0 + 8\nu_0^2 + 4v_0^2 d_0 \\
A_{0,1} &= 8\nu_z^2 d_1 + 4\nu_z\left(\alpha_1 + \nu_0\right) d_1 + 4\gamma_{0,1} d_1 \\
A_{0,2} &= 8\nu_z^2 d_2 + 4\nu_z\left(\alpha_2 + \nu_0\right) d_2 + 4\gamma_{0,2} d_2 \\
A_{1,0} &= 8\nu_z^2 d_0 + 4\nu_z\left(\nu_0 + \alpha_1\right) d_0 + 4\gamma_{0,1} d_0 \\
A_{1,1} &= 16\nu_z^2 + 8\nu_z^2 d_1 + 16\nu_z\alpha_1 + 8\nu_z\alpha_1 d_1 + 4\beta_1 + 4\beta_1 d_1 \\
A_{1,2} &= 8\nu_z^2 d_2 + 4\nu_z\left(\alpha_2 + \alpha_1\right) d_2 + 4\gamma_{1,2} d_2 \\
A_{2,0} &= 8\nu_z^2 d_0 + 4\nu_z\left(\nu_0 + \alpha_2\right) d_0 + 4\gamma_{0,2} d_0 \\
A_{2,1} &= 8\nu_z^2 d_1 + 4\nu_z\left(\alpha_1 + \alpha_2\right) d_1 + 4\gamma_{1,2} d_1 \\
A_{2,2} &= 16\nu_z^2 + 8\nu_z^2 d_2 + 16\nu_z\alpha_2 + 8\nu_z\alpha_2 d_2 + 4\beta_2 + 4\beta_2 d_2
\end{aligned}$$

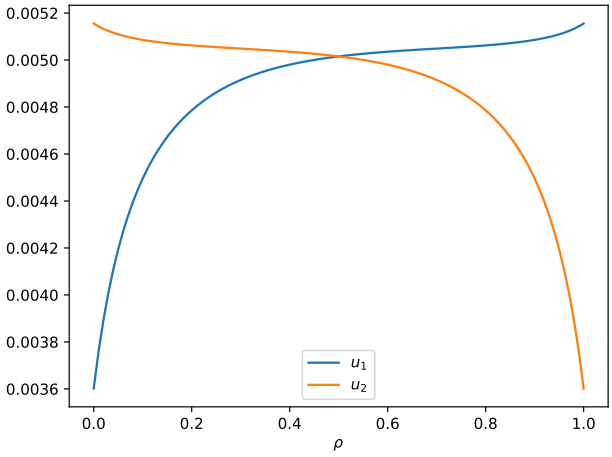

Figure 1: Optimal weights for varying values of ρ. Results obtained with $\nu_0 = 1, \nu_1 = 2, \nu_2 = 0.1, \nu_z = 0.7, m = 1$, and $d_0 = 5, d_1 = d_2 = 10$.

Therefore, the optimal representation is given by the solution to the following set of linear equations:

$$\begin{pmatrix} u_0^2 \\ u_1^2 \\ u_2^2 \end{pmatrix} = 2m\, A^{-1} \begin{pmatrix} \nu_0 + \nu_z \\ \alpha_1 + \nu_z \\ \alpha_2 + \nu_z \end{pmatrix}, \tag{29}$$

where

$$A = \begin{pmatrix} A_{0,0} & A_{0,1} & A_{0,2} \\ A_{1,0} & A_{1,1} & A_{1,2} \\ A_{2,0} & A_{2,1} & A_{2,2} \end{pmatrix}. \tag{30}$$

2 Additional simulation studies

In Section 5.1 of the paper we provided simulation results for a linear representation $g(z) = wz$ where $w \in \mathbb{R}^{d \times p}$ with $p = 16$, for varying values of $\rho \in 0.05, 0.1, 0.3$, in a setting where $\nu^+ = 2$, $\nu^- = 0.1$ ($\frac{\nu^+}{\nu^-} = 20$).

We now focus on the case of $\rho = 0.1$ and examine additional representation sizes p, and noise ratios ($\frac{\nu^+}{\nu^-} \in \{10, 40\}$). Additionally, we examine the original setting where $p = 16$ and $\nu^+ = 2$, $\nu^- = 0.1$, with varying proportions of positive and negative examples.

The results (Figure 2 of this document) show that in all the additional settings our methods provides statistically significant improvement over the baseline. FDR adjusted p-values for multiple comparisons are provided in Table 1.

3 Ablation study

Our method includes two additions over standard training: hierarchical subsampling (which was repeatedly shown to stabilize loss in meta-learning), and the balancing term over synthetic environments. Throughout the paper we compared our method against a baseline that optimizes the unpenalized ERM score over hierarchical subsamples, therefore choosing the higher burden of proof for our method. For a complete ablation study we show below the the remaining comparison with optimization of unpenalized loss over batches sampled uniformly at random.

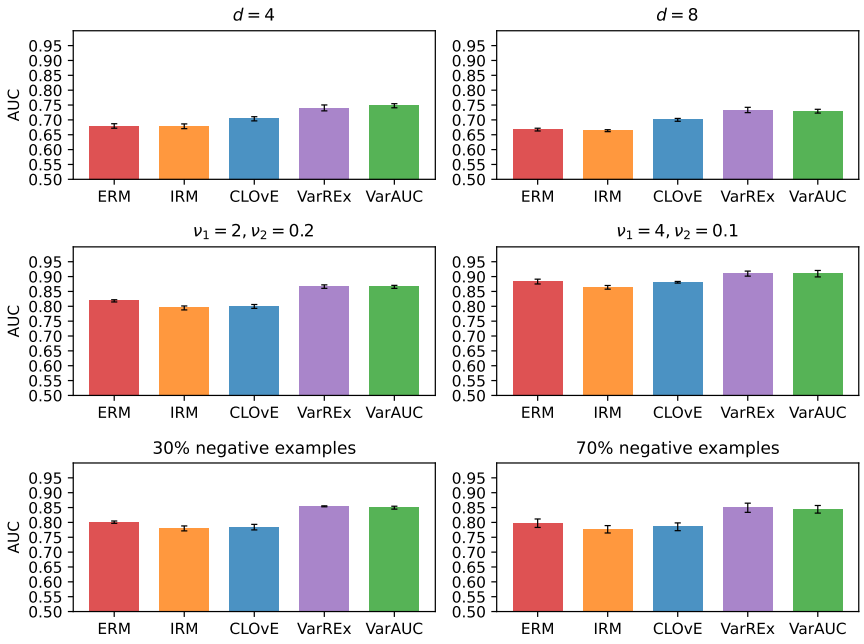

Figure 2: Additional simulation results. Top row: Additional dimensions of the representation. Middle row: additional rations of the attribute variances. Bottom row: unbalanced sets of positive and negative examples. Bars show mean AUC values on the test set across 5 repetitions of the experiment, whiskers show 2 standard deviations.

Table 1: FDR adjusted p-values for the results reported in Figure 2

Experiment	IRM	CLOvE	VarREx	VarAUC
$p = 4$	0.7339	0.0112	0.0003	0.0001
$p = 8$	0.8552	0.0005	0.0003	0.0001
$\nu_1 = 2, \nu_2 = 0.2$	0.9995	0.9995	0.0002	$<$0.0
$\nu_1 = 4, \nu_2 = 0.1$	0.9989	0.9971	0.0041	0.0041
30% negative	0.9939	0.9939	$<$0.0001	$<$0.0
70% negative	1.0	1.0	$<$0.0001	0.0002

Results for the original 3 simulations presented in the paper are shown in Figure 3. The results show that (i) hierarchical subsampling yields better results than uniform sampling. (ii) our method outperforms unpenalized loss over hierarchical samples (original baseline), and therefore unpenalized loss over uniform samples by even a larger margin.

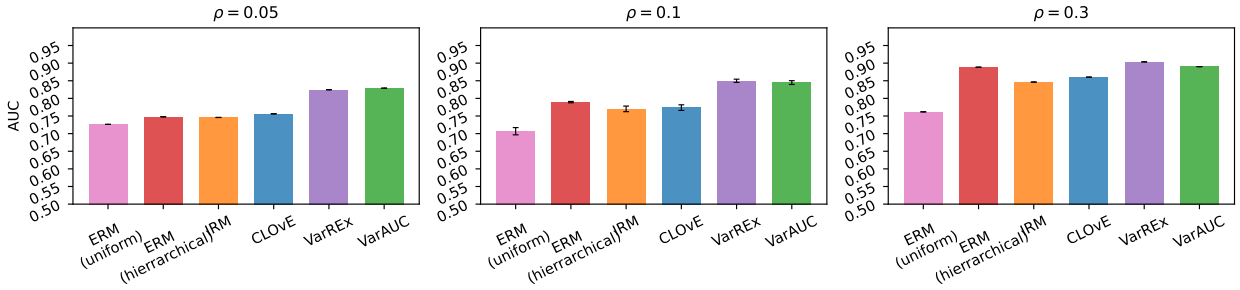

Figure 3: Ablation study. Results are shown for varying attribute proportions ρ. Bars show mean AUC values on the test set across 5 repetitions of the experiment, whiskers show 2 standard deviations.