APPENDIX

A DEFERRED THEOREMS & PROOFS

In this section, we include further theoretical results, including the proofs omitted from the main body of the paper.

On the assumption $N \leq d$. The assumption on the number of vectors allows us to convert the Lovasz theta problem into a convex one. Indeed, without this assumption, converting the problem into an SDP by setting $\mathbf{Y} = \mathbf{V}^T\mathbf{V}$ would impose constraints on the rank of the matrix $\mathbf{Y}$ (specifically, $\text{rank}(\mathbf{Y}) \leq d$). Since we know that $\text{rank}(\mathbf{Y}) \leq N$ by construction, then having $N \leq d$ makes this rank constraint trivial. If we had these rank constraints, then the problem would be non-convex, given that the domain of the problem (a set of rank constrained matrices) is non-convex.

In practice, whether this constraint is satisfied is based on our choice for model dimensionality and batch size. It can however be circumvented, by regular optimization methods operating directly on the representations $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N$ (rather than their inner products). While the problem may no longer be convex, we can still apply first order optimization methods to find a good stationary point for the loss.

Proof of Theorem 3.1. The following is an adaptation of the proof found in Gärtner & Matousek (2012).

We first show that the following problems:

$$\begin{array}{cc} \underset{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_N,\boldsymbol{c}\in\mathbb{R}^d}{\text{minimize}} & k, \\ \text{s.t.} & \|\boldsymbol{u}_1\|^2 = \|\boldsymbol{u}_2\|^2 = \cdots = \|\boldsymbol{u}_N\|^2 = \|\boldsymbol{c}\|^2 = 1, \\ & \boldsymbol{u}_i^T\boldsymbol{u}_j = 0,\ \forall (i,j) \notin E, \\ & \frac{1}{(\boldsymbol{c}^T\boldsymbol{u}_i)^2} \leq k,\ \forall i, \end{array} \tag{11}$$

and:

$$\begin{array}{cc} \underset{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_N\in\mathbb{R}^d}{\text{minimize}} & k, \\ \text{s.t.} & \|\boldsymbol{v}_1\|^2 = \|\boldsymbol{v}_2\|^2 = \cdots = \|\boldsymbol{v}_N\|^2 = 1, \\ & \boldsymbol{v}_i^T\boldsymbol{v}_j \leq -\frac{1}{k-1},\ \forall (i,j) \notin E, \end{array} \tag{12}$$

are equivalent. To do this, we shall show that an optimal solution of the first problem can be converted to a feasible solution for the second one, and vice versa.

Given an optimal solution to the first problem $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N, \boldsymbol{c}$ with optimal value k, we can define the following set of vectors:

$$\boldsymbol{z}_i = \frac{1}{\sqrt{k-1}}\left(\frac{\boldsymbol{u}_i}{\boldsymbol{c}^T\boldsymbol{u}_i} - \boldsymbol{c}\right). \tag{13}$$

For these vectors, the following hold:

- $\boldsymbol{z}_i^T\boldsymbol{z}_j = \frac{1}{k-1}\left(\frac{\boldsymbol{u}_i^T\boldsymbol{u}_j}{(\boldsymbol{c}^T\boldsymbol{u}_i)(\boldsymbol{c}^T\boldsymbol{u}_j)} - 1 - 1 + 1\right) \leq -\frac{1}{k-1}$.
- $\|\boldsymbol{z}_i\|^2 = \frac{1}{k-1}\left(\frac{\|\boldsymbol{u}_i\|^2}{(\boldsymbol{c}^T\boldsymbol{u}_i)^2} - 1 - 1 + 1\right) \leq \frac{1}{k-1}(k-1) = 1$.

Let us now define the matrix $\mathbf{Z}$, with columns the vectors $\boldsymbol{z}_i$. Thus, $\mathbf{Z}^T\mathbf{Z}$ is positive semidefinite (by construction) and has diagonal elements which are less than or equal to 1. Thus, we can define the matrix $\mathbf{Y} = \mathbf{Z}^T\mathbf{Z} + \mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix with non-negative elements, such that $y_{ii} = 1$. This matrix $\mathbf{Y}$ is also PSD, which means that we can write it as $\mathbf{Y} = \mathbf{V}^T\mathbf{V}$. The columns of $\mathbf{V}$ are precisely the vectors $\boldsymbol{v}_i$ that form a feasible solution of the second problem, since their inner products (off-diagonal elements of $\mathbf{Y}$) satisfy the required constraints, and their norms (diagonal elements of $\mathbf{Y}$) are equal to 1.

Now, given an optimal solution to the second problem, $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N$ with optimal value k, we can again define $\mathbf{Y} = \mathbf{V}^T\mathbf{V}$. Here we can make an argument that $\mathbf{Y}$ must have at least one eigenvalue

equal to 0 (in other words, the vectors $\boldsymbol{v}_i$ must be linearly dependent). Indeed, if this was not the case, we would have $\mathbf{Y} \succ 0$, so $\mathbf{Y} - \epsilon\mathbf{I} \succeq 0$, where $\epsilon > 0$ sufficiently small. This means that the matrix $\mathbf{Y}' = \frac{1}{1-\epsilon}(\mathbf{Y} - \epsilon\mathbf{I})$ is a feasible solution to the second problem (note that the off diagonal elements of $\mathbf{Y}$ are negative, so multiplying them by a positive constant decreases them) which also has lower value for the objective function, which is a contradiction. Thus $\mathbf{Y}$ is singular, which also implies that there exists a unit vector $\boldsymbol{c} \in \mathbb{R}^d$ that has $\boldsymbol{c}^T\boldsymbol{v}_i = 0$, for all i. We can then define the following vectors:

$$\boldsymbol{u}_i = \frac{1}{\sqrt{k}}(\boldsymbol{c} + \boldsymbol{v}_i\sqrt{k-1}). \tag{14}$$

For these vectors, we have:

- $\boldsymbol{u}_i^T\boldsymbol{u}_j = \frac{1}{k}(1 + 0 + 0 + (k-1)\boldsymbol{v}_i^T\boldsymbol{v}_j) \leq 0.$
- $\|\boldsymbol{u}_i\|^2 = \frac{1}{k}(1 + 0 + 0 + k - 1) = 1.$
- $\boldsymbol{c}^T\boldsymbol{u}_i = \frac{1}{\sqrt{k}} \Rightarrow \frac{1}{(\boldsymbol{c}^T\boldsymbol{u}_i)} = k.$

This means that the vectors $\boldsymbol{u}_i$ along with $\boldsymbol{c}$ form a feasible solution for the first problem, with objective value k.

For the final step, we note that by setting $t = -\frac{1}{k-1}$, then by minimizing k we also minimize t (since the latter is an increasing function of k). This means that our second problem is equivalent to:

$$\begin{array}{cc} \underset{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_N \in \mathbb{R}^d}{\text{minimize}} & t, \\ \text{s.t.} & \|\boldsymbol{v}_1\|^2 = \|\boldsymbol{v}_2\|^2 = \cdots = \|\boldsymbol{v}_N\|^2 = 1, \\ & \boldsymbol{v}_i^T\boldsymbol{v}_j \leq t,\ \forall (i,j) \notin E. \end{array} \tag{15}$$

This completes our proof.

Theorem A.1. *Formal version of Theorem 4.1 The problems:*

$$\begin{array}{cc} \text{minimize} & t, \\ \text{s.t.} & \|\boldsymbol{v}_i\|^2 = 1,\ \forall i, \\ & \boldsymbol{v}_i^T\boldsymbol{v}_j \leq t,\ \forall i \neq j, \end{array} \tag{16}$$

and:

$$\begin{array}{cc} \text{minimize} & \tau \sum_{i=1}^n \log \sum_{j \neq i} \exp\left(\frac{\boldsymbol{v}_i^T\boldsymbol{v}_j}{\tau}\right), \\ \text{s.t.} & \|\boldsymbol{v}_i\|^2 = 1,\ \forall i, \end{array} \tag{17}$$

attain their minima at the same values of the matrix $\mathbf{Y} = \mathbf{V}^T\mathbf{V}$, where $\mathbf{V}$ is the matrix which has the vectors $\boldsymbol{v}_i$ as columns.

Proof. Note that by setting $\mathbf{Y} = \mathbf{V}^T\mathbf{V}$, the first problem is converted into:

$$\begin{array}{cc} \text{minimize} & t, \\ \text{s.t.} & y_{ii} = 1,\ \forall i, \\ & y_{ij} \leq t,\ \forall i \neq j, \\ & \mathbf{Y} \succeq 0, \end{array} \tag{18}$$

while the second problem is converted into:

$$\begin{array}{cc} \text{minimize} & \tau \sum_{i=1}^n \log \sum_{j \neq i} \exp\left(\frac{y_{ij}}{\tau}\right), \\ \text{s.t.} & y_{ii} = 1,\ \forall i, \\ & \mathbf{Y} \succeq 0. \end{array} \tag{19}$$

Let us assume that we have a feasible solution $\mathbf{Y}$ for the second problem. Let $t = \frac{1}{n(n-1)} \sum_{i,j:i\neq j} y_{ij}$ (the average of the non-diagonal elements of $\mathbf{Y}$). Due to the convexity of the exponential function and the logarithm being an increasing function, we can apply Jensen's inequality in each of the sums inside the logarithms for each given i, thus giving us:

$$\begin{aligned} \tau \sum_{i=1}^n \log \sum_{j \neq i} \exp\left(\frac{y_{ij}}{\tau}\right) &\geq \tau \sum_{i=1}^n \log\left((n-1)\exp\left(\frac{\frac{1}{n-1}\sum_{j\neq i} y_{ij}}{\tau}\right)\right) \\ &= \tau n \log(n-1) + \frac{1}{n-1}\sum_{i=1}^n \sum_{j \neq i} y_{ij} \\ &= \tau n \log(n-1) + nt. \end{aligned} \tag{20}$$

The final step is precisely the value of our objective function, when all y_{ij} , $i \neq j$ are equal to t. This means that, if we replace our solution with the average of the non-diagonal elements, then we will always decrease the value of the objective. Note that the new solution will still be feasible. Indeed, since the original matrix $\mathbf{Y}$ is PSD, we have:

$$\mathbf{1}^T\mathbf{Y}\mathbf{1} = n + n(n-1)t \geq 0 \Rightarrow t \geq -\frac{1}{n-1}. \tag{21}$$

The new matrix we will use is:

$$\mathbf{Y}' = (1-t)\mathbf{I} + t\mathbf{1}, \tag{22}$$

(so the diagonal elements are 1, and the non-diagonal ones are t). This is a matrix with rank-1 difference from the identity, and it has $n-1$ eigenvalues equal to $1-t \geq 0$ (since t is the average of inner products of vectors with unit norm), and one eigenvalue equal to $1-t+tn = 1+(n-1)t \geq 0$ (due to the above). Thus the new matrix is also a feasible solution, so it is always optimal to have $y_{ij} = t$.

Thus, we can rewrite our problem as follows:

$$\begin{array}{cc} \text{minimize} & \tau \sum_{i=1}^n \log \sum_{j \neq i} \exp\left(\frac{y_{ij}}{\tau}\right), \\ \text{s.t.} & y_{ii} = 1, \ \forall i, \\ & y_{ij} = t, \ i \neq j, \\ & \mathbf{Y} \succeq 0. \end{array} \tag{23}$$

Given the condition for y_{ij}, we can rewrite our objective function as:

$$\tau \sum_{i=1}^n \log\left(n(n-1)\exp\left(\frac{t}{\tau}\right)\right) = \tau n \log\left(n(n-1)\right) + tn. \tag{24}$$

Thus, we can easily see that this problem has the same minimizing matrix $\mathbf{Y}$ as:

$$\begin{array}{cc} \text{minimize} & t, \\ \text{s.t.} & y_{ii} = 1, \ \forall i, \\ & y_{ij} = t, \ i \neq j, \\ & \mathbf{Y} \succeq 0. \end{array} \tag{25}$$

Finally, we need to argue that this problem has the same optimal solution as equation 18 (or that, in other words, having the constraints $y_{ij} = t$ be $y_{ij} \leq t$ does not change the optimal solution). This can be easily shown using the exact same argument as above - if we assume that there exists an element $y_{ij} < t$ in the optimal solution, then we can replace the non-diagonal elements of $\mathbf{Y}$ with their average $\bar{y} < t$, giving us a feasible solution $\mathbf{Y}' \succeq 0, t' = \bar{y} < t$, which is impossible. Thus, the problem in equation 18 has the same minimizer matrix $\mathbf{Y}$ as that in equation 25. This completes the proof, as having the same matrix $\mathbf{Y} = \mathbf{V}^T\mathbf{V}$ and having the vectors all be equal in norm means that the vectors chosen are unique up to rotations. □

Lemma A.1. *The following holds:*

$$\lim_{\tau \to 0^+} \tau \log \sum_{i=1}^n \exp \frac{z_i}{\tau} = \max_{i=1}^n z_i. \tag{26}$$

Proof. We have:

$$\sum_{i=1}^n \exp\frac{z_i}{\tau} \geq \exp\left(\frac{1}{\tau}\max_{i=1}^n z_i\right) \Rightarrow \max_{i=1}^n z_i \leq \tau \log \sum_{i=1}^n \exp\frac{z_i}{\tau}, \tag{27}$$

as well as:

$$\tau \log \sum_{i=1}^n \exp\frac{z_i}{\tau} \leq \tau \log\left(n \exp\left(\frac{1}{\tau}\max_{i=1}^n z_i\right)\right) = \tau \log n + \max_{i=1}^n z_i. \tag{28}$$

Thus, we get:

$$\max_{i=1}^n z_i \leq \tau \log \sum_{i=1}^n \exp\frac{z_i}{\tau} \leq \tau \log n + \max_{i=1}^n z_i, \tag{29}$$

and taking the limit as $\tau \to 0$ gives us the desired result.

We note here that the convergence is **uniform**: for a given τ, the difference between $\tau \log \sum_{i=1}^{n} \exp \frac{z_i}{\tau}$ and $\max_{i=1}^{n} z_i$ is upper bounded by $\tau \log n$, which is independent of $z_1, \ldots, z_n$. □

Lemma A.2 (Formal Version of Lemma 4.1 in the paper)**.** *For $N \leq d$, the weighted Lovasz Theta problem can be rewritten as in equation 31. In other words, the following formulations of the weighted Lovasz theta problem are equivalent:*

$$\begin{array}{cc} \underset{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_N,\boldsymbol{c}\in\mathbb{R}^d}{\text{minimize}} & \max_{i=1,\ldots,N} \frac{1}{(\boldsymbol{c}^T\boldsymbol{u}_i)^2}, \\ \text{s.t.} & \|\boldsymbol{u}_1\|^2 = \|\boldsymbol{u}_2\|^2 = \cdots = \|\boldsymbol{u}_N\|^2 = \|\boldsymbol{c}\|^2 = 1, \\ & \boldsymbol{u}_i^T\boldsymbol{u}_j \leq w_{ij}, \end{array} \tag{30}$$

and:

$$\begin{array}{cc} \underset{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_N\in\mathbb{R}^d}{\text{minimize}} & t, \\ \text{s.t.} & \|\boldsymbol{v}_1\|^2 = \|\boldsymbol{v}_2\|^2 = \cdots = \|\boldsymbol{v}_N\|^2 = 1, \\ & \boldsymbol{v}_i^T\boldsymbol{v}_j \leq w_{ij} + (1 - w_{ij})t. \end{array} \tag{31}$$

Proof. To go from the first problem to the second, we begin by reformulating it as follows:

$$\begin{array}{cc} \underset{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_N,\boldsymbol{c}\in\mathbb{R}^d}{\text{minimize}} & k, \\ \text{s.t.} & \|\boldsymbol{u}_1\|^2 = \|\boldsymbol{u}_2\|^2 = \cdots = \|\boldsymbol{u}_N\|^2 = \|\boldsymbol{c}\|^2 = 1, \\ & \boldsymbol{u}_i^T\boldsymbol{u}_j \leq w_{ij}, \\ & (\boldsymbol{c}^T\boldsymbol{u}_i)^2 \geq \frac{1}{k}. \end{array} \tag{32}$$

We can also reformulate the second problem as follows, by setting $t = -\frac{1}{k-1}$ and noticing that t is increasing as k increases:

$$\begin{array}{cc} \underset{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_N\in\mathbb{R}^d}{\text{minimize}} & k, \\ \text{s.t.} & \|\boldsymbol{v}_1\|^2 = \|\boldsymbol{v}_2\|^2 = \cdots = \|\boldsymbol{v}_N\|^2 = 1, \\ & \boldsymbol{v}_i^T\boldsymbol{v}_j \leq \frac{w_{ij}k-1}{k-1}. \end{array} \tag{33}$$

It is sufficient to show that the problems in equation 32 and equation 33 are equivalent. Following the technique used by Gärtner & Matousek (2012) the regular Lovasz theta problem, we do so by showing the two problems have the same optimal value. Let p_1 and p_2 be the optimal values of the problems in equation 32 and equation 33, respectively.

To show that $p_1 \geq p_2$, consider an optimal solution $\boldsymbol{u}_1^*, \ldots, \boldsymbol{u}_N^*, \boldsymbol{c}^*$. Let us formulate the following matrix $\mathbf{Y}$, with elements:

$$\begin{aligned} &y_{ii} = 1. \\ &y_{ij} = \frac{1}{p_1 - 1}\left(\frac{\boldsymbol{u}_i^*}{\boldsymbol{c}^{*T}\boldsymbol{u}_i^*} - \boldsymbol{c}^*\right)^T\left(\frac{\boldsymbol{u}_j^*}{\boldsymbol{c}^{*T}\boldsymbol{u}_j^*} - \boldsymbol{c}^*\right)^T = \frac{1}{p_1 - 1}\left(\frac{\boldsymbol{u}_i^{*T}\boldsymbol{u}_j^*}{(\boldsymbol{c}^{*T}\boldsymbol{u}_i^*)(\boldsymbol{c}^{*T}\boldsymbol{u}_j^*)} - 1\right) \\ &\quad \leq \frac{p_1 w_{ij} - 1}{p_1 - 1}. \end{aligned} \tag{34}$$

We can also show that this matrix is PSD; since $y_{ii} \geq y_{ij}, \forall j$, $\mathbf{Y}$ can be written as $\mathbf{Y} = \mathbf{D} + \mathbf{U}^T\mathbf{U} \succeq 0$, where $\mathbf{D}$ is a diagonal matrix with non-negative entries and $\mathbf{U}$ is a matrix with columns equal to $\frac{\boldsymbol{u}_i^*}{\boldsymbol{c}^{*T}\boldsymbol{u}_i^*} - \boldsymbol{c}^*$. Consequently, we can use this $\mathbf{Y}$, to create a feasible solution $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N$ via Cholesky factorization (i.e. $\mathbf{Y} = VV^T$, whenever $N \leq d$ (the constraints arise from the constraints placed on the matrix $\mathbf{Y}$). This feasible solution has objective value p_1 for the second problem. This means that the second problem has optimal value $p_2 \leq p_1$.

To show that $p_2 \geq p_1$, start from an optimal solution of the second problem $\boldsymbol{v}_1^*, \ldots, \boldsymbol{v}_N^*$. Let $\mathbf{Y}^*$ be the matrix with $y_{ij}^* = \boldsymbol{v}_i^{*T}\boldsymbol{v}_j^*$. We note here that this matrix must have at least one eigenvalue equal to 0. If this was not the case, then we would have $\lambda_{min}(\mathbf{Y}^*) > 0$, and we could construct the following matrix:

$$\mathbf{Y}' = \mathbf{Y}^* + \epsilon(\mathbf{I} - \mathbf{1}). \tag{35}$$

where $\mathbf{1}$ is the rank-1 matrix with all of its elements equal to 1. Note that this would strictly decrease all the non-diagonal elements of $\mathbf{Y}^*$, while leaving the diagonal ones intact. Furthermore, $\mathbf{1}$ has $N-1$ eigenvalues equal to 0, and one of them equal to N (as $\mathbf{1}\boldsymbol{u} = N\boldsymbol{u}$, where u is the all-ones vector). Thus, since $\mathbf{I} - \mathbf{1}$ is a rank-1 difference from the diagonal, $N-1$ of its eigenvalues being equal to 1 and one of them being equal to $1 - N < 0$ (given the eigenvalues of $\mathbf{I}$). Thus, we would have, for ϵ small enough:

$$\lambda_{min}(\mathbf{Y}') \geq \lambda_{min}(\mathbf{Y}^*) + \epsilon(1-N) \geq 0. \tag{36}$$

Thus, by Cholesky decomposition again, we would have a feasible solution to problem 33, with strictly smaller off-diagonal elements of y_{ij}, which is a contradiction. Thus, one of the eigenvalues of $\mathbf{Y}$ must be 0, which means that we can find a vector $\boldsymbol{c}$ which is orthogonal to all $\boldsymbol{v}_1^*, \ldots, \boldsymbol{v}_N^*$. We can then define the vectors:

$$\boldsymbol{u}_i = \frac{1}{\sqrt{p_2}}(\boldsymbol{v}_i^* \sqrt{p_2 - 1} + \boldsymbol{c}). \tag{37}$$

These vectors have:

$$\boldsymbol{u}_i^T \boldsymbol{u}_j = \frac{1}{p_2}((p_2 - 1)\boldsymbol{v}_i^{*T} \boldsymbol{v}_j^* + 1). \tag{38}$$

Thus $\|\boldsymbol{u}_i\|^2 = \frac{1}{p_2}(p_2 - 1 + 1) = 1$ and $\boldsymbol{u}_i^T \boldsymbol{u}_j \leq \frac{1}{p_2}(w_{ij} p_2 - 1 + 1) = w_{ij}$. This means that they form a feasible solution for our problem, with objective value p_2. Thus $p_1 \leq p_2$.

Combining all of the above gives us $p_1 = p_2$, making the above problems equivalent.

□

Theorem A.2 (Formal Version of Theorem 4.2)**.** *Consider the following formulation of the weighted Lovasz theta problem:*

$$\begin{array}{ll} \underset{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_N \in \mathbb{R}^d}{\text{minimize}} & t, \\ \text{s.t.} & \|\boldsymbol{v}_1\|^2 = \|\boldsymbol{v}_2\|^2 = \cdots = \|\boldsymbol{v}_N\|^2 = 1, \\ & \boldsymbol{v}_i^T \boldsymbol{v}_j \leq w_{ij} + (1 - w_{ij})t, \end{array} \tag{39}$$

as well as the minimization of the second term of the Lovasz theta contrastive loss:

$$\begin{array}{ll} \underset{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_N \in \mathbb{R}^d}{\text{minimize}} & \tau \sum_{i=1}^N \log\left(\sum_{j=1}^N \exp\left(\frac{\boldsymbol{v}_j^T \boldsymbol{v}_i - w_{ij}}{\tau(1-w_{ij})}\right)\right), \\ \text{s.t.} & \|\boldsymbol{v}_1\|^2 = \|\boldsymbol{v}_1\|^2 = \|\boldsymbol{v}_2\|^2 = \cdots = \|\boldsymbol{v}_N\|^2 = 1. \end{array} \tag{40}$$

Minimizing the limit of the objective of the second problem as $\tau \to 0$ is a relaxation of the first problem, if $w_{ij} < 1$.

Proof. Using the property of Log-Sum-Exp to converge uniformly to the maximum of its arguments as τ goes to 0, in this limit the objective of the second problem becomes:

$$\begin{array}{ll} \underset{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_N \in \mathbb{R}^d}{\text{minimize}} & \sum_{i=1}^N \max_{j \neq i} \frac{\boldsymbol{v}_j^T \boldsymbol{v}_i - w_{ij}}{1 - w_{ij}}, \\ \text{s.t.} & \|\boldsymbol{v}_1\|^2 = \|\boldsymbol{v}_1\|^2 = \|\boldsymbol{v}_2\|^2 = \cdots = \|\boldsymbol{v}_N\|^2 = 1, \end{array} \tag{41}$$

(note that we don't argue about the behavior of the minimizing solution of the problem as $\tau \to 0$, but rather just the minimization of the limit of the objective function). We now include auxiliary variables t_i, changing the problem into the following:

$$\begin{array}{ll} \underset{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_N \in \mathbb{R}^d}{\text{minimize}} & \sum_{i=1}^N t_i, \\ \text{s.t.} & \|\boldsymbol{v}_1\|^2 = \|\boldsymbol{v}_2\|^2 = \cdots = \|\boldsymbol{v}_N\|^2 = 1, \\ & \boldsymbol{v}_i^T \boldsymbol{v}_j \leq w_{ij} + (1 - w_{ij})t_i, \ \forall i \neq j. \end{array} \tag{42}$$

This is a relaxation of the following problem:

$$\begin{array}{ll} \underset{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_N \in \mathbb{R}^d}{\text{minimize}} & \sum_{i=1}^N t_i, \\ \text{s.t.} & \|\boldsymbol{v}_1\|^2 = \|\boldsymbol{v}_2\|^2 = \cdots = \|\boldsymbol{v}_N\|^2 = 1, \\ & \boldsymbol{v}_i^T \boldsymbol{v}_j \leq w_{ij} + (1 - w_{ij})t_i, \ \forall i \neq j, \\ & t_i = t, \ \forall i, \end{array} \tag{43}$$

which is equivalent to precisely the weighted Lovasz theta problem:

$$\begin{array}{ll} \underset{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_N \in \mathbb{R}^d}{\text{minimize}} & t, \\ \text{s.t.} & \|\boldsymbol{v}_1\|^2 = \|\boldsymbol{v}_2\|^2 = \cdots = \|\boldsymbol{v}_N\|^2 = 1, \\ & \boldsymbol{v}_i^T \boldsymbol{v}_j \leq w_{ij} + (1 - w_{ij})t, \; \forall i \neq j. \end{array} \tag{44}$$

In this case, the lack of symmetry of the problem does not allow us to immediately argue that all the t_i must be equal. □

B Training Details

In our CIFAR experiments, we made use of an A100 GPU to train our models. In the supervised case, our models were trained for 300 epochs, with a batch size of 512, and the same set of hyperparameters as those used in the Supervised Contrastive learning baseline (Khosla et al., 2020). The architecture used in each of the experiments was a ResNet of varying depth, and the projection head used to perform contrastive learning was a two-layer MLP, reducing the dimension of the features to 128. Evaluation was performed by training a linear classifier on top of the model for 10 epochs, and reporting the best accuracy obtained on the test set across all of the epochs. In the unsupervised case, the architecture is the same, but our models where trained using a batch size of 1024, a learning rate and temperature τ equal to 0.5, and a linear probe trained for 100 epochs over the learned representations (as done in the repository for the code of Khosla et al. (2020))

For our ImageNet-100 experiments, we made use of computing nodes with 4 RTX5000 GPUs. For models trained with cross-entropy loss, we employed a batch size of 256. Moreover, we made use of Momentum Contrast, when training these models. During MoCo training, we maintained a batch size of 256 and a memory bank of size 8192 as in Khosla et al. (2020). We trained each model for 200 epochs using the standard hyperparameter choices found in He et al. (2020).

Our baseline models were trained using the code directly provided by Khosla et al. (2020). The confusion matrix based similarities were obtained based on the predictions of a model trained with Supervised Contrastive Learning.

The overall code is provided as part of the Supplementary material for review purposes, and will be made publicly available upon acceptance. The code is based upon the publicly available code of Supervised Contrastive Learning and Momentum Contrast, and all the relevant licenses are included.

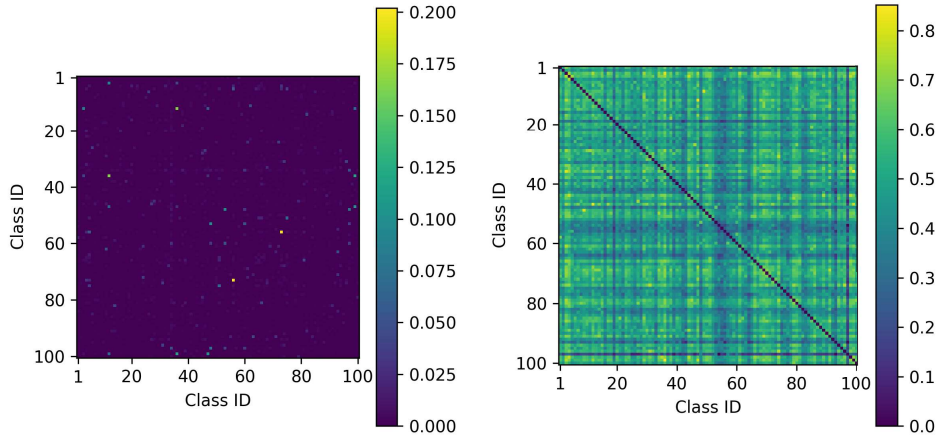
C Further Experiments

C.1 Confusion Matrix Visualization

In this section, we include a visualization of two similarity matrices used in the main paper. We can see the results in Figure 4. It is evident that the confusion matrix based approach is much more selective than the CLIP based one. Combining this with the fact that both the confusion matrix and the superclass similarities outperforms the CLIP based one, as seen in the experimental section, we can infer the result that the more selective the confusion matrix is, the better the results of our method are (which is also intuitively what we expect to happen).

C.2 Experiment on CIFAR10

In Table 5, we can see the results of our method on CIFAR10, using the similarity matrix derived via the confusion matrix of another model. We can see that while our method obtains good accuracy, we cannot get significant improvements over regular supervised contrastive learning. We believe that this is a sensible limitation for our method - due to the very small number of classes, it is highly unlikely that any are that similar to begin with. As such, the main benefit of our method, which is being able to leverage sample similarities via their classes during training does not apply. Nevertheless, since many important tasks contain a much larger number of classes, this is only a small limitation.



(a) Confusion matrix similarities.

(b) CLIP based similarities.

Figure 4: **Similarity matrices used in the main paper, supervised setting.** On the left, we see the confusion matrix similarities, and on the right the CLIP based similarities. In both cases, the diagonal has been set to 0 for the purposes of visualization. We can see that while the confusion matrix approach is much more selective in which classes are similar to each other, the CLIP based similarities assign much greater similarities to all classes. This, combined with the results of the relevant experiments, shows the importance of the chosen similarity matrix.

Table 5: **Summary of our results on CIFAR10.** In this case, our method is competitive with SupCon, but there is a limitation due to the fewer number of classes. The numbers with * are taken from Khosla et al. (2020).

	CE	SupCon	Ours
ResNet-50	95*	**96***	95.47
ResNet-34	95.15	**95.28**	95.07
ResNet-18	94.37	94.61	**94.69**

C.3 Ablation on the Superclass Similarities

As an ablation, we performed the experiments presented in the main paper regarding the superclass similarities on CIFAR, using a different value for the similarity of classes belonging to the same superclass. The results can be seen in Table 6. We can see that we get comparable results with those seen in the main paper, so tuning this hyperparameter may prove useful, in order to improve the performance of the model.

D Further Directions

Based on all of the above, it is not evident whether there is a single way to obtain the proper similarity matrix, to get the best possible results. As can be seen by our experiments, this design decision influences the performance of our models. As such, our work opens interesting research directions in identifying good similarity metrics between samples in contrastive learning.

Table 6: **Ablation on superclass similarity.** Our method has comparable results, based on how similar samples from the same superclass but different classes are considered to be (0.5 or 0.8 in this case).

	Superclass Similarity 0.5	Superclass Similarity 0.8
ResNet-50	77.60 ± 0.30	77.68 ± 0.49
ResNet-34	76.55 ± 0.40	76.35 ± 0.06
ResNet-18	74.91 ± 0.15	74.99 ± 0.32