An Information-Theoretic Understanding of Maximum Manifold Capacity Representations

Rylan Schaeffer*
Stanford University
rschaef@stanford.edu

Berivan Isik*
Stanford University
berivan.isik@stanford.edu

Victor Lecomte*
Stanford University
vlecomte@stanford.edu

Mikail Khona
MIT
mikail@mit.edu

Yann LeCun
New York University & Meta AI, FAIR

Andrey Gromov
UMD & MetaAI, FAIR
gromovand@meta.com

Ravid Shwartz-Ziv
New York University
rs8020@nyu.edu

Sanmi Koyejo
Stanford University
sanmi@cs.stanford.edu

Abstract

Maximum Manifold Capacity Representations (MMCR) is a recent multi-view self-supervised learning (MVSSL) method that matches or surpasses other leading MVSSL methods. MMCR is interesting for at least two reasons. Firstly, MMCR is an oddity in the zoo of MVSSL methods: it is not (explicitly) contrastive, applies no masking, performs no clustering, leverages no distillation, and does not (explicitly) reduce redundancy. Secondly, while many self-supervised learning (SSL) methods originate in information theory, MMCR distinguishes itself by claiming a different origin: a statistical mechanical characterization of the geometry of linear separability of data manifolds. However, given the rich connections between statistical mechanics and information theory, and given recent work showing how many SSL methods can be understood from an information-theoretic perspective, we conjecture that MMCR can be similarly understood from an information-theoretic perspective. In this paper, we leverage tools from high dimensional probability and information theory to demonstrate that an optimal solution to MMCR's nuclear norm-based objective function is the same optimal solution that maximizes a well-known lower bound on mutual information between views.

Multi-view self-supervised learning (MVSSL; also known as Joint Embedding SSL) is a powerful approach to unsupervised learning. The core idea is to create multiple transformations, or "views", of unsupervised data, then use these transformed data in a supervised-like manner to learn generally useful representations. MVSSL methods are diverse but can be loosely grouped into a number of different families: (1) contrastive, such as CPC [20], MoCo 1 [19], SimCLR [8], MoCo 2 [9], CMC [22], RPC [23] and TiCo [29]; (2) clustering such as Noise-as-Targets [4], DeepCluster [5], Self-Labeling [1], Local Aggregation [30], SwAV [6]; (3) distillation/momentum such as BYOL [18], DINO [7], SimSiam [10], TiCo [29]; (4) redundancy reduction such as Barlow Twins [28], VICReg [3], TiCo [29]. Many MVSSL methods either explicitly originate from information theory [20, 2] or can be understood from an information-theoretic perspective [25, 26, 21, 15].

Recently, [27] proposed a new MVSLL method named Maximum Manifold Capacity Representations (MMCR) that achieves similar (if not superior) performance to leading MVSSL methods. MMCR

*Denotes equal contribution and co-first authorship.

NeurIPS 2023 Workshop Information-Theoretic Principles in Cognitive Systems.

is interesting for at least two reasons. Firstly, MMCR does not fit neatly into any of the MVSSL families: it is not (explicitly) contrastive, it applies no masking, it performs no clustering, it leverages no distillation, and it does not (explicitly) reduce redundancy. Secondly, unlike many MVSSL methods that originate in information theory, MMCR distances itself, writing that estimating mutual information in high dimensions has proven difficult and that more closely approximating mutual information may not improve representations; MMCR instead claims an origin in the statistical mechanical characterization of the geometry of linear separability of data manifolds.

In this work, we seek to better understand what solutions the MMCR loss function incentivizes and how it relates to other well-known MVSSL methods. Our contributions are specifically as follows:

1. We derive a distribution of embeddings that provably minimizes the MMCR loss with high probability by leveraging tools from high dimensional probability.
2. We connect this distribution to information theory by showing it maximizes a well-known variational lower bound on the mutual information between multiple views' embeddings.

1 Background

Multi-View Self-Supervised Learning (MVSSL) Let $f_\theta : \mathcal{X} \to \mathcal{Z}$ denote our neural network with parameters θ. Suppose we have a dataset $\{\boldsymbol{x}_n\}_{n=1}^N$ and a set of transformations (sometimes called augmentations, or "views") $\mathcal{T}$ such as color jittering, Gaussian blur, solarization, etc. For each datum $\boldsymbol{x}_n$ in a batch of inputs, we sample K transformations $t^{(1)}, t^{(2)}, ..., t^{(K)} \sim \mathcal{T}$, then transform the datum: $t^{(1)}(\boldsymbol{x}_n), ..., t^{(K)}(\boldsymbol{x}_n)$. We feed these transformed data into the network and obtain *embeddings* or *representations*:

$$\boldsymbol{z}_n^{(k)} \stackrel{\text{def}}{=} f_\theta(t^{(k)}(\boldsymbol{x}_n)) \in \mathcal{Z}. \tag{1}$$

In practice, $\mathcal{Z}$ is commonly $\mathbb{R}^D$ or the D-dimensional hypersphere $\mathbb{S}^{D-1} \stackrel{\text{def}}{=} \{\boldsymbol{z} \in \mathbb{R}^D : \boldsymbol{z}^T\boldsymbol{z} = 1\}$. Given that this work will later touch on information theory, we need notation to refer to the random variables; we use $Z_n^{(k)}$ to denote the random variable for the embedding whose realization is $\boldsymbol{z}_n^{(k)}$, and $X_n^{(k)}$ to denote the random variable for the transformed datum whose realization is $t^{(k)}(\boldsymbol{x}_n)$.

Maximum Manifold Capacity Representations (MMCR) MMCR [27] originates from classical results regarding performance of linear binary classifiers [12, 16, 17]. Consider N points in dimension D, with arbitrarily assigned binary class labels; *what is the probability that a linear binary classifier will be able to successfully classify the points?* Statistical mechanical calculations reveal that in the *thermodynamic* limit ($N, D \to \infty$; $N/D \to \alpha \in (0, \infty)$), a phase transition occurs at capacity $\alpha_c = 2$. More precisely, if $\alpha < \alpha_c$, the linear binary classifier will succeed with probability 1; but if $\alpha > \alpha_c$, the classifier will succeed with probability 0. MMCR is based on an extension of this result from points to manifolds [11]. MMCR proceeds in the following manner: MMCR takes the embeddings output by the network and normalizes them to lie on the hypersphere: $\boldsymbol{z}_n^{(1)}, ..., \boldsymbol{z}_n^{(K)} \in \mathbb{S}^{D-1}$. Then, MMCR compute the average embedding per datum:

$$\boldsymbol{\mu}_n \stackrel{\text{def}}{=} \frac{1}{K} \sum_k \boldsymbol{z}_n^{(k)}. \tag{2}$$

Next, MMCR forms a $N \times D$ matrix M where the n-th row of M is $\boldsymbol{\mu}_n$ and defines the loss:

$$\mathcal{L}_{MMCR} \stackrel{\text{def}}{=} -\|M\|_* \stackrel{\text{def}}{=} -\sum_{r=1}^{rank(M)} \sigma_r(M), \tag{3}$$

where $\sigma_r(M)$ is the r-th singular value of M and $\|\cdot\|_*$ is the nuclear norm (trace norm, Schatten 1-norm). Minimizing the MMCR loss means maximizing the nuclear norm of the mean matrix M. The authors of MMCR note that no closed form solution exists for singular values of an arbitrary matrix, but when $N = 2, D = 2$, a closed form solution exists that offers intuition: $\|M\|_*$ will be maximized when (i) the norm of each mean is maximized i.e., $\|\boldsymbol{\mu}_n\|_2 = 1$ (recalling that $0 \leq \|\boldsymbol{\mu}_n\| < 1$

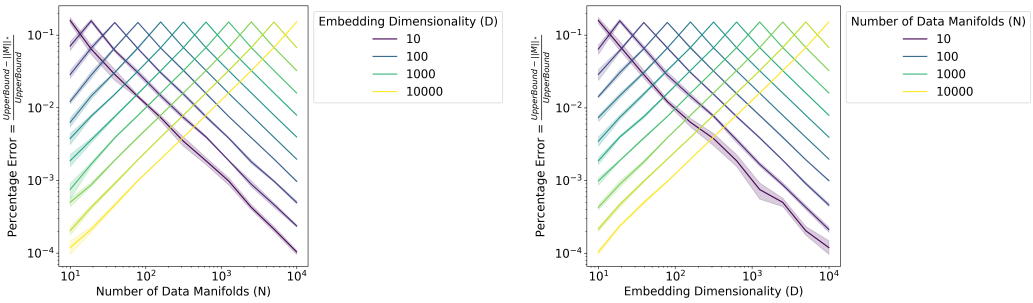

Figure 1: **Numerical simulations confirm that a network achieving perfect reconstruction and perfect uniformity achieves the lowest possible MMCR loss.** Away from the $N = D$ threshold, uniform random vectors achieve the theoretically derived upper bound on the nuclear norm (i.e. lower bound on $\mathcal{L}_{MMCR}$).

since the representations live on the hypersphere), and (ii) the means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ are orthogonal to one another. While we commend the authors for working to offer intuition, it is unclear to what extent the $N = 2, D = 2$ setting sheds light on MMCR in general, as MMCR was theoretically derived and numerically implemented in the large data and high dimension regime.

2 An Information Theoretic Understanding of MMCR

In this section, we prove and intuitively explain two properties of MMCR that shed light on it as well as relate it to other MVSSL methods. We specifically consider MMCR's regime of large dataset size N and high embedding dimension D. We contribute two results:

1. The MMCR loss $\mathcal{L}_{MMCR}$ is minimized by (a) making each mean $\boldsymbol{\mu}_n = \frac{1}{K}\sum_k \boldsymbol{z}_n^{(k)}$ lie on the surface of the hypersphere, and (b) making the distribution of means as close to uniform on the hypersphere as possible.
2. This configuration of means maximizes a well-known variational lower bound on the mutual information between embeddings [14] that was recently used to study and unify several multi-view SSL (MVSSL) families [15].

More formally, we begin by adapting two useful definitions from relevant prior works [25, 15]:

Definition 2.1 (Perfect Reconstruction)**.** *We say a network f_θ achieves perfect reconstruction if $\forall \boldsymbol{x} \in \mathcal{X}, \forall\, t^{(1)}, t^{(2)} \in \mathcal{T}$, $\boldsymbol{z}^{(1)} = f_\theta(t^{(1)}(\boldsymbol{x})) = f_\theta(t^{(2)}(\boldsymbol{x})) = \boldsymbol{z}^{(2)}$.*

Definition 2.2 (Perfect Uniformity)**.** *Let $p(Z)$ be the distribution over the network representations induced by the data sampling and transformation sampling distributions. We say a network f_θ achieves perfect uniformity if the distribution $p(Z)$ is the uniform distribution on the hypersphere.*

We will show that a network that achieves both perfect reconstruction and perfect uniformity obtains the lowest possible MMCR loss by first showing that $\mathcal{L}_{MMCR}$ has a lower bound and then showing that such a network achieves this bound.

Proposition 2.3. *Suppose that $\forall n \in [N], \boldsymbol{\mu}_n^T \boldsymbol{\mu}_n \leq 1$. Then, $0 \leq ||M||_* \leq \begin{cases} N & \text{if } N \leq D \\ \sqrt{ND} & \text{if } N \geq D \end{cases}$.*

Proof. Let $\sigma_1, \ldots, \sigma_{\min(N,D)}$ denote the singular values of M, so that $\|M\|_* = \sum_{i=1}^{\min(N,D)} \sigma_i$. The lower bound follows by the fact that singular values are nonnegative. For the upper bound, we have

$$\sum_{i=1}^{\min(N,D)} \sigma_i^2 = \mathrm{Tr}\big[MM^T\big] = \sum_{n=1}^{N} \boldsymbol{\mu}_n^T \boldsymbol{\mu}_n \leq N.$$

Then, by Cauchy-Schwarz on the sequences $(1, \ldots, 1)$ and $\left(\sigma_1, \ldots, \sigma_{\min(N,D)}\right)$, we get

$$\sum_{i=1}^{\min(N,D)} \sigma_i \leq \sqrt{\left(\sum_{i=1}^{\min(N,D)} 1\right)\left(\sum_{i=1}^{\min(N,D)} \sigma_i^2\right)} \leq \sqrt{\min(N,D)N} = \begin{cases} N & \text{if } N \leq D \\ \sqrt{ND} & \text{if } N \geq D \end{cases}.$$

□

Proposition 2.4. *Let f_θ achieve perfect reconstruction. Then, $\|\boldsymbol{\mu}_n\|_2 = 1 \; \forall n$.*

Proof. Because f_θ achieves perfect reconstruction, $\forall n, \forall t^{(1)}, t^{(2)}$, $\boldsymbol{z}_n^{(1)} = \boldsymbol{z}_n^{(2)}$. Thus $\boldsymbol{\mu}_n = (1/K)\sum_k \boldsymbol{z}_n^{(k)} = (1/K)\sum_k \boldsymbol{z}_n^{(1)} = \boldsymbol{z}_n^{(1)}$, and since $\|\boldsymbol{z}_n^{(1)}\|_2 = 1$, we have $\|\boldsymbol{\mu}_n\|_2 = 1$. □

Theorem 2.5. *Let $f_\theta : \mathcal{X} \to \mathbb{S}^D$ be a network that achieves perfect reconstruction and perfect uniformity. Then f_θ achieves the lower bound of $\mathcal{L}_{MMCR}$ with high probability. Specifically:*

$$\|M\|_* = \begin{cases} N(1 - O(N/D)) & \textit{if } N \leq D \\ \sqrt{ND}(1 - O(D/N)) & \textit{if } N \geq D \end{cases}$$

with high probability in $\min(N, D)$.

We defer the proof to Appendix A but offer intuition here. To show the inequality in Proposition 2.3 is roughly tight, we need to show the singular values σ_i are all roughly equal to each other. When $N \ll D$, since M has few rows $\boldsymbol{\mu}_n$, they are almost perfectly orthogonal to each other, so all N singular values will be $\approx \|\boldsymbol{\mu}_n\| = 1$. When $N \gg D$, since M has many rows, for any $x \in \mathbb{R}^D$ the sum $\|Mx\|_2^2 = \sum_n (\boldsymbol{\mu}_n^T x)^2$ will be concentrated, so M scales all vectors roughly equally, and therefore its D singular values are all roughly equal to each other. We confirm this via numerical simulations (Fig. 1); for code, see Appendix B.

We now turn to addressing why perfect reconstruction and perfect uniformity matter from an information theoretic perspective. The results here for MVSSL are known, e.g., [25, 15], but we repeat them to make the connection with MMCR. For input datum X, consider the mutual information between the learned embeddings of two different views $Z^{(1)} = t^{(1)}(X)$ and $Z^{(2)} = t^{(2)}(X)$; the mutual information must be at least as great as the sum of two terms: the ability of one embedding to "reconstruct" the other, plus the entropy of the embeddings:

$$I[Z^{(1)}; Z^{(2)}] \quad \geq \quad \underbrace{\mathbb{E}_{p(Z^{(1)}, Z^{(2)})}[\log q(Z^{(1)}|Z^{(2)})]}_{\text{Reconstruction}} \quad + \quad \underbrace{H[Z^{(1)}]}_{\text{Entropy}}, \tag{4}$$

where $q(Z^{(1)}|Z^{(2)})$ is a variational distribution because the true distribution $p(Z^{(1)}|Z^{(2)})$ is unknown.

Theorem 2.6. *Let $f_\theta : \mathcal{X} \to \mathbb{S}^D$ be a network, the number of views per datum be constant, and $\mathcal{Q}$ be the variational family of distributions on the hypersphere. Then f_θ maximizes the mutual information lower bound Eqn. 4 if and only if f_θ achieves perfect reconstruction and perfect uniformity.*

Proof. Perfect reconstruction maximizes reconstruction term. Perfect uniformity maximizes entropy since the maximum entropy is achieved with the uniform distribution over the support [13]. □

Theorem 2.7. *Let f_{θ^*} be a network that achieves perfect reconstruction and perfect uniformity, let the number of views per datum K be a constant, and let $\mathcal{Q}$ be the variational family of distributions on the hypersphere. Then f_{θ^*} is both a minimizer of $\mathcal{L}_{MMCR}$ and a maximizer of the variational lower bound of mutual information Eqn. 4.*

Proof. The proof follows by Theorem 2.5 and Theorem 2.6. □

3 Discussion

In this work, we leveraged tools from high dimensional probability to prove that in the large data and high dimensional regime, the MMCR loss is minimized with high probability by a network achieving perfect reconstruction and perfect uniformity. These two properties together are known to maximize a well-known variational lower bound on the mutual information between multi-view embeddings.

References

[1] YM Asano, C Rupprecht, and A Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations*, 2019.

[2] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.

[3] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

[4] Piotr Bojanowski and Armand Joulin. Unsupervised learning by predicting noise. In *International Conference on Machine Learning*, pages 517–526. PMLR, 2017.

[5] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018.

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[9] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[10] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[11] SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. Classification and geometry of general perceptual manifolds. *Physical Review X*, 8(3):031003, 2018.

[12] Thomas M Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, (3):326–334, 1965.

[13] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006.

[14] Robert G Gallager. *Information theory and reliable communication*, volume 588. Springer, 1968.

[15] Borja Rodrıguez Gálvez, Arno Blaas, Pau Rodriguez, Adam Golinski, Xavier Suau, Jason Ramapuram, Dan Busbridge, and Luca Zappella. The role of entropy and reconstruction in multi-view self-supervised learning. In *International Conference on Machine Learning*, pages 29143–29160. PMLR, 2023.

[16] Elizabeth Gardner. Maximum storage capacity in neural networks. *Europhysics letters*, 4(4):481, 1987.

[17] Elizabeth Gardner. The space of interactions in neural network models. *Journal of physics A: Mathematical and general*, 21(1):257, 1988.

[18] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[20] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[21] Ravid Shwartz-Ziv, Randall Balestriero, Kenji Kawaguchi, Tim GJ Rudner, and Yann LeCun. An information-theoretic perspective on variance-invariance-covariance regularization. *arXiv preprint arXiv:2303.00633*, 2023.

[22] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020.

[23] Yao-Hung Hubert Tsai, Martin Q. Ma, Muqiao Yang, Han Zhao, Louis-Philippe Morency, and Ruslan Salakhutdinov. Self-supervised representation learning with relative predictive coding. In *International Conference on Learning Representations*, 2021.

[24] Roman Vershynin, Y Eldar, and Gitta Kutyniok. Compressed sensing, theory and applications. In *Introduction to the non-asymptotic analysis of random matrices*, pages 210–268. Cambridge Univ. Press, 2012.

[25] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[26] Mike Wu, Chengxu Zhuang, Milan Mosse, Daniel Yamins, and Noah Goodman. On mutual information in contrastive learning for visual representations. *arXiv preprint arXiv:2005.13149*, 2020.

[27] Thomas Yerxa, Yilun Kuang, Eero Simoncelli, and SueYeon Chung. Learning efficient coding of natural images with maximum manifold capacity representations. *arXiv preprint arXiv:2303.03307*, 2023.

[28] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[29] Jiachen Zhu, Rafael M Moraes, Serkan Karakulak, Vlad Sobol, Alfredo Canziani, and Yann LeCun. Tico: Transformation invariance and covariance contrast for self-supervised visual representation learning. *arXiv preprint arXiv:2206.10698*, 2022.

[30] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6002–6012, 2019.

A Proof of Theorem 2.5

Recall that $\mathcal{L}_{MMCR} = -\|M\|_*$ is minimized when $\|M\|_*$ is maximized and that $\|M\|_*$ is upper bounded by $\sqrt{ND}$ if $N > D$ and N if $N < D$ (Proposition 2.3). We want to show a network that achieves perfect reconstruction and perfect uniformity achieves this upper bound on the nuclear norm (equivalently, lower bound on the MMCR loss).

Following the proof of Proposition 2.3, let $\sigma_1, \ldots, \sigma_{\min(N,D)}$ denote the singular values of M, so that $\|M\|_* = \sum_i \sigma_i$. By Proposition 2.4, we have

$$\sum_i \sigma_i^2 = \mathrm{Tr}\big[MM^T\big] = \sum_{n=1}^{N} \boldsymbol{\mu}_n^T \boldsymbol{\mu}_n = N.$$

Now, by the equality version of Cauchy–Schwarz on the sequences $(1, \ldots, 1)$ and $\big(\sigma_1, \ldots, \sigma_{\min(N,D)}\big)$, we have

$$\sum_i \sigma_i = \sqrt{\min(N,D)\left(\sum_i \sigma_i^2 - \sum_i \left(\sigma_i - \frac{\sum_j \sigma_j}{\min(N,D)}\right)^2\right)}. \tag{5}$$

So if we can bound this "variance" of the singular values $\sum_i \left(\sigma_i - \frac{\sum_j \sigma_j}{\min(N,D)}\right)^2$, we can show that $\|M\|_*$ closely matches the upper bound obtained in Proposition 2.3.

To do this, let us consider matrix $\sqrt{D}M$. The vectors $\boldsymbol{\mu}_n$ are uniform over the D-dimensional hypersphere $\mathbb{S}^D$, so its rows $\sqrt{D}\boldsymbol{\mu}_n$ have mean zero, are isotropic, and (by Example 5.25 in [24]) are sub-gaussian with parameter $\|\sqrt{D}\boldsymbol{\mu}_n\|_{\psi_2} = O(1)$.[2] Therefore,

- **If $\mathbf{N} \leq \mathbf{D}$**, then (using the fact that $\|\boldsymbol{\mu}_n\|_2 = 1$ for all $n \in [N]$) we can to apply Theorem 5.58 in [24] on the transpose of $\sqrt{D}M$, obtaining that for any $t \geq 0$, the singular values of $\sqrt{D}M$ are within $\sqrt{D} \pm O(\sqrt{N}) + t$ with probability at least $1 - 2\exp(-\Omega(t^2))$. Setting t to a large enough multiple of $\sqrt{N}$, they are all within $\sqrt{D} \pm O(\sqrt{N})$ with probability at least $1 - 2\exp(-N)$. Consequently, with the same probability, the singular values of M are all within $\pm O(\sqrt{N/D})$ of each other, and we get $\sum_i \left(\sigma_i - \frac{\sum_j \sigma_j}{\min(N,D)}\right)^2 \leq N \cdot O\left(\sqrt{N/D}\right)^2 = O(N^2/D)$. Plugging this into Eqn. 5, we get $\|M\|_* \leq \sqrt{N(N - O(N^2/D))} = \sqrt{N}(1 - O(N/D))$.
- **If $\mathbf{N} \geq \mathbf{D}$**, then we can apply Theorem 5.39 in [24] on $\sqrt{D}M$, obtaining that for any $t \geq 0$, the singular values of $\sqrt{D}M$ are within $\sqrt{N} \pm O(\sqrt{D}) + t$ with probability at least $1 - 2\exp(-\Omega(t^2))$. Setting t to a large enough multiple of $\sqrt{D}$, they are all within $\sqrt{N} \pm O(\sqrt{D})$ with probability at least $1 - 2\exp(-D)$. Consequently, with the same probability, the singular values of M are all within $\pm O(1)$ of each other, and we get $\sum_i \left(\sigma_i - \frac{\sum_j \sigma_j}{\min(N,D)}\right)^2 \leq D \cdot O(1)^2 = O(D)$. Plugging this into Eqn. 5, we get $\|M\|_* \leq \sqrt{D(N - O(D))} = \sqrt{ND}(1 - O(D/N))$.

[2] Here, $\|\cdot\|_{\psi_2}$ denotes the sub-gaussian norm (intuitively, the "effective standard deviation" of a sub-gaussian random variable). For a scalar random variable X, it is defined as $\|X\|_{\psi_2} := \sup_{p \geq 1} p^{-1/2} (\mathbb{E}[|X|^p])^{1/p}$ (Definition 5.7 in [24]), and for a random vector $\boldsymbol{u} \in \mathbb{R}^D$, it is defined as $\|\boldsymbol{u}\|_{\psi_2} := \sup_{\boldsymbol{v} \in \mathbb{S}^D} \|\boldsymbol{u}^T \boldsymbol{v}\|_{\psi_2}$ (Definition 5.22 in [24]).

B Python Code for Perfect Reconstruction and Perfect Uniformity Embeddings

To test our claim that networks which achieve perfect reconstruction and perfect uniformity achieve the nuclear norm upper bound, we sample a uniform distribution of embeddings and measure the nuclear norm relative to our claimed upper bound. Python code for our simulations is included below:

```
import pandas as pd
import numpy as np


N_list = np.logspace(start=1, stop=4, num=11).astype(int)
D_list = np.logspace(start=1, stop=4, num=11).astype(int)
repeats = np.arange(5).astype(int)
uniform_distribution_nuclear_norm_data_list = []

for N in N_list:
    for D in D_list:
        print(f"N: {N}\tD: {D}")
        for repeat in repeats:
            embeddings = np.random.normal(loc=0, scale=10.0, size=(N, D))
            embeddings /= np.linalg.norm(embeddings, axis=1, keepdims=True)
            row = {
                "Spectrum": "uniform",
                "Number of Data Manifolds (N)": N,
                "Embedding Dimensionality (D)": D,
                "Repeat": repeat,
                "Nuclear Norm": np.linalg.norm(embeddings, ord="nuc"),
            }
            uniform_distribution_nuclear_norm_data_list.append(row)

uniform_distribution_nuclear_norm_df = pd.DataFrame(
    uniform_distribution_nuclear_norm_data_list
)
```