Modeling Recognition Memory with Predictive Coding and Hopfield Networks

Tianjin Li
MRC Brain Network Dynamics Unit
University of Oxford, UK
tianjin.li@hertford.ox.ac.uk

Mufeng Tang, Rafal Bogacz
MRC Brain Network Dynamics Unit
University of Oxford, UK
{mufeng.tang, rafal.bogacz}@bndu.ox.ac.uk

Abstract

Associative memory (AM) and recognition memory (RM) are fundamental in human and machine cognition. RM refers to an ability to recognize if the stimulus has been seen before, or is novel. Neuroscience studies reveal that regions such as the hippocampus, known for AM, are also involved in RM. Inspired by repetition suppression in the brain, this work presents an energy-based approach to RM, where a model learns by adjusting an energy function. We employed this energy-based approach to Hopfield Networks (HNs) and Predictive Coding Networks (PCNs). Our simulations indicate that PCN outperforms HNs in RM tasks, especially with correlated patterns. In this work, we also unify the theoretical understanding of HN and PCN in RM, revealing that both perform metric learning. This theory is crucial in explaining PCN's superior performance in handling correlated data as it reveals that PCNs employ a statistical whitening step in its metric learning, which refines the distinction between familiar and novel stimuli. Overall, the superior performance of PCN, as well as the unique error neurons in its circuit implementation matching repetition suppression, provide a plausible account of how the brain performs RM, within the network architecture known to also support AM.

1 Introduction

Associative memory (AM) and **recognition memory (RM)** are vital for both human and machine intelligence. While AM retrieves detailed past patterns from incomplete queries, RM labels queries as 'novel' or 'familiar' based on previous encounters. Both types of memory hinge on storing past experiences and comparing them with current stimuli. Regions believed to manage AM, such as the hippocampus, are also involved in RM [4]. Neurobiological research highlights *novelty neurons* in various cortical areas such as infereotemporal cortex [9], perirhinal cortex and hippocampus [16]. Named for their *repetition suppression*, these neurons are most active with new stimuli, with declining activity upon repeated exposure. While their existence is well-documented [11, 13, 4, 15], the underlying computational processes remain underexplored. In particular, limited number of models have been able to demonstrate both biological plausibility and large capacity in RM tasks on real-world patterns such as natural images.

In this work, we follow an *energy-based approach* to apply AM models for RM tasks. The model adjusts its parameters to minimize an energy function when exposed to a pattern, essentially 'learning' or 'memorizing' it. Later, the energy value of a query indicates its familiarity, with a lower value signaling familiarity and vice versa, aligning with repetition suppression seen in the brain. Previous work [3] has applied this approach to Hopfield Network (HN) [8] and shows that it successfully performs RM for binary patterns. Here, we extend this energy-based approach to Predictive Coding

Associative Memory & Hopfield Networks in 2023. NeurIPS 2023 workshop.

Networks (PCNs) [7, 14] as well as Modern Continuous Hopfield Network (MCHN) [10]. We observe in experiments that PCN has a larger RM capacity than HNs, especially when the memorized patterns have a correlated structure. To explain these results, we conducted a mathematical analysis that unifies both PCN and HNs within the framework of metric learning. It reveals that PCN employs a statistical whitening step that facilitates the discrimination between familiarity and novelty. Moreover, while the energy-based approach generally offers a repetition suppression mechanism by associating it with energy minimization, PCNs uniquely have error units resembling the novelty neurons observed in neurobiology, providing a more plausible mechanism of how RM is achieved in the brain.

2 Methods & Results

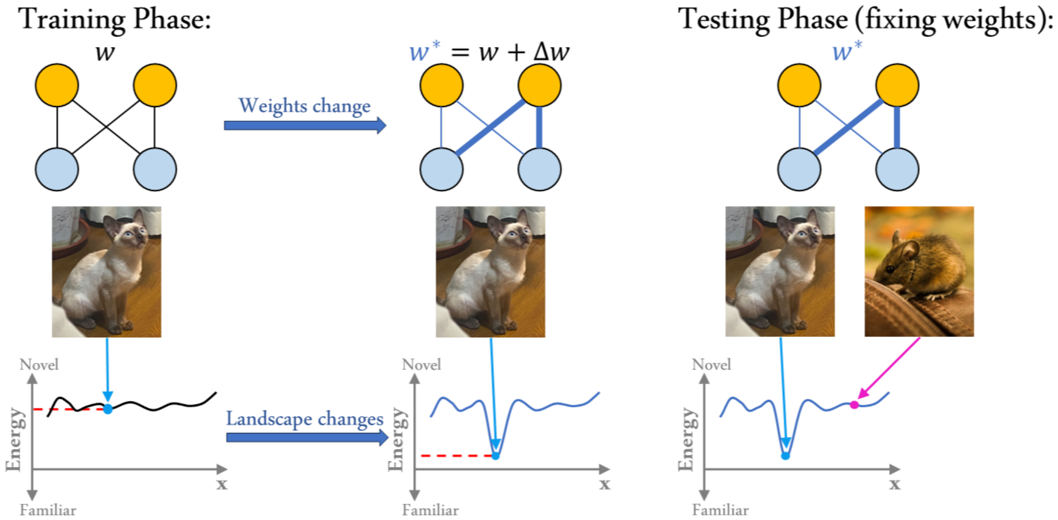

Figure 1: An illustration of the general **energy-based approach to RM**.

Formally, assume a total of N, d-dimensional patterns $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N)$ are generated i.i.d from a certain probability distribution $\mathbf{p}$. During the **training phase**, patterns to be memorized (which form the columns of the data matrix $\mathbf{X} \in \mathbb{R}^{d\times N}$) are fed to the model one by one, modifying the weights/parameters (e.g., d by d weight matrix W) in the process such that it becomes a local minimum in the altered energy landscape. Then, in the **testing phase**, a single d-dimensional query pattern $\mathbf{q}$ is provided to the network, and performing RM only requires reading out value of the energy function, which indicates the familiarity of the query. This process is demonstrated in Fig 1.

2.1 Hopfield Networks for RM

HN is a classical energy-based model for AM [8], which has also been shown to demonstrate the capacity of RM by measuring the degree of familiarity via its energy function after training [3]. In this work, we extend this approach to MCHN. Formally, the energy functions of HN and MCHN are:

$$E_{HN}(\mathbf{q}, \mathbf{X}) = -\sum\nolimits_{\mu=1}^{N} (\mathbf{q}^T \mathbf{x}^\mu)^2; E_{MCHN}(\mathbf{q}, \mathbf{X}) = -\log(\sum\nolimits_{\mu=1}^{N} \exp(\mathbf{q}^T \mathbf{x}^\mu)) + \frac{1}{2}\|\mathbf{q}\|_2^2 \quad (1)$$

2.2 Predictive Coding Networks for RM

In this work, we focus primarily on a biologically plausible variant of PCN with recurrent connections [14]. For simplicity, we refer to it as PCN here. PCN's dynamics aim to minimize the total squared prediction errors ϵ_i of all neurons predicting each other in a recurrent network, which can be considered as an energy function:

$$E_{PCN}(\mathbf{q}, W) = \frac{1}{2}\sum\nolimits_{i=1}^{d} \epsilon_i^2 = \frac{1}{2}\|\mathbf{q} - W^T\mathbf{q} - \nu\|_2^2 \quad (2)$$

where W is the weight matrix implicitly encoding covariance [14] and ν is the bias vector. Notably, this model has a biologically plausible neural implementation with local computations *and* employs neurons signaling prediction errors and thus the degree of familiarity/novelty (Fig. 2).

2.3 Experiment Results

Fig. 3 summarizes the performances of PCN, HN, and MCHN on numerically generated multivariate Gaussian variables and the Tiny ImageNet dataset [6]. It also compares it with the experimentally observed performance of humans in discriminating familiarity of natural images [12]. The first row displays the average accuracy as a function of the number of presented patterns or stimuli and the second row shows the number of patterns retained in memory (defined in [12] and given in Appendix A). In the leftmost column, pixels are uncorrelated and the three models have similar, decent accuracy. Remarkably, when features become weakly correlated in the middle column, only PCN keeps a similar level of performance. It is worth pointing out that real-world images tend to have correlated pixels, especially among pixels that are spatially proximate to each other. This is indeed what testing on the Tiny ImageNet reveals in the right column of Fig. 3: HN and MCHN perform poorly even for a small number of stored patterns, implying that the Hopfield Networks may perform poorly in real-world scenarios. By contrast PCN achieves accuracy even higher than human participants obtained while classifying if presented images had been seen by them before [12].

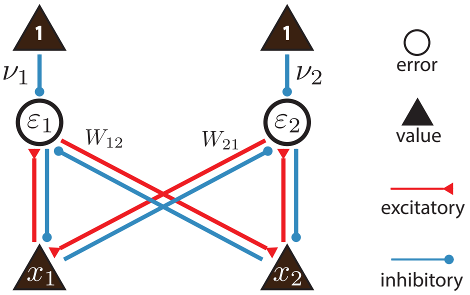

Figure 2: An illustration of Predictive Coding Network (PCN) in [14] for two-dimensional inputs, which requires only local information to perform its computations.

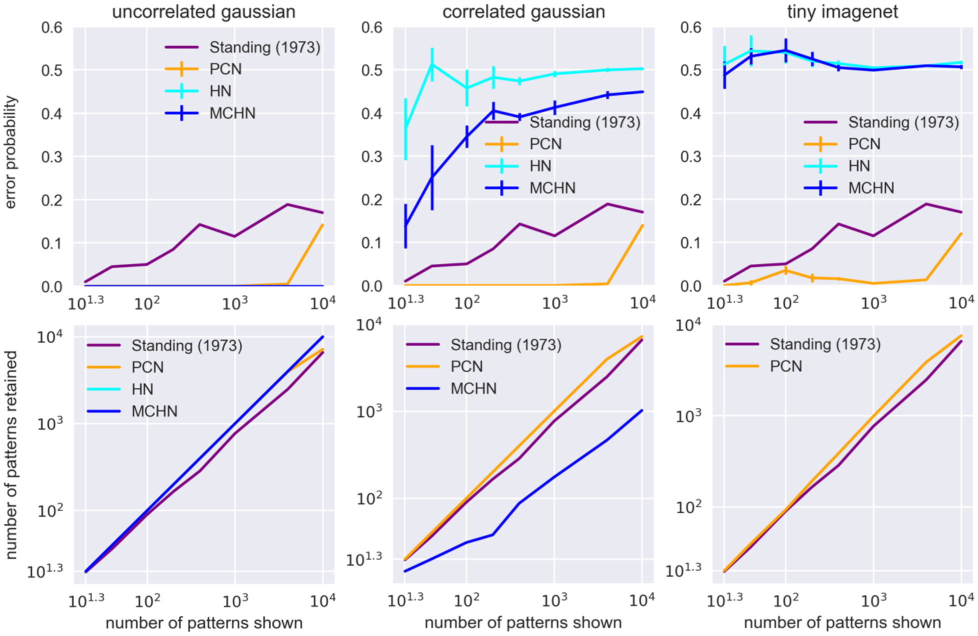

Figure 3: Top row: Error probability as a function of the number of training samples. An error probability of 0.5 corresponds to a baseline level equivalent to guessing. The first two columns are produced with $d = 500$. In the second column, there is a 0.4 covariance/correlation between any two distinct pixels. For the last column, d matches the dimensions of the dataset $d = 4096$ for (grayscale) Tiny ImageNet [6]. Bottom row: Number of patterns retained in memory as a function of the number of training samples. With both axes logarithmic, the plots reveal a power-law relationship observed for human participants [12]. HN are not shown in the middle plot and both HN and MCHN are not shown in the last plot due to their almost random performance. Classical results about the memory capacity of HN can be found in [1, 2].

3 Theoretical Analysis

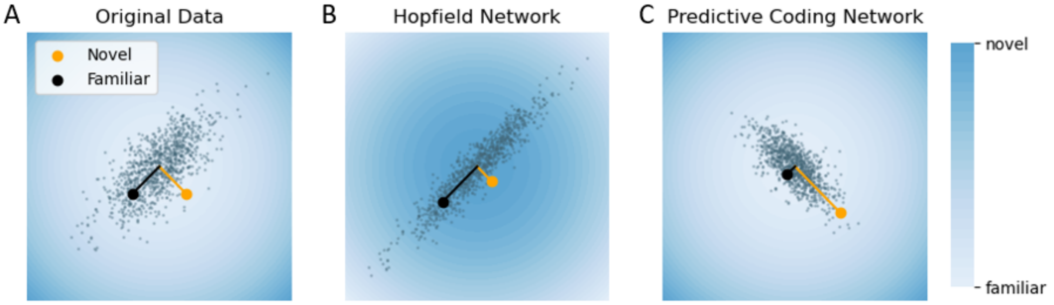

Figure 4: The effect of various transformations when $d = 2$ and underlying distribution $\mathbf{p}$ is correlated. The shade of blue indicates energy, which is inverted in B because of the negative sign with HNs.

Here we provide a unifying mathematical account for HN and PCN in RM, showing that they can both be considered as measuring the Euclidean distance between a query and the mean of the training data, but on different transformed planes. Formally, when the stored patterns $\mathbf{X}$ has mean $\bar{\mathbf{x}}$, there exists a parameterized class of distance of the form

$$d^2_{L,\mathbf{X}}(\mathbf{q}) := (\mathbf{q} - \bar{\mathbf{x}})^T L^T L(\mathbf{q} - \bar{\mathbf{x}}) = \|L(\mathbf{q} - \bar{\mathbf{x}})\|_2^2 \tag{3}$$

which can be seen as the squared Euclidean distance from the $\mathbf{q}$ to $\bar{\mathbf{x}}$ on the space transformed by a $d \times d$ matrix L. Using HN for RM can be seen as metric learning of this form since:

$$E_{HN}(\mathbf{q}, \mathbf{X}) = -\mathbf{q}^T\mathbf{X}\mathbf{X}^T\mathbf{q} \propto -\mathbf{q}^T\Sigma\mathbf{q} = -\mathbf{q}^T(\Sigma^{\frac{1}{2}})^T\Sigma^{\frac{1}{2}}\mathbf{q} = -\|\Sigma^{\frac{1}{2}}\mathbf{q}\|_2^2 \tag{4}$$

where we used the fact that $\Sigma = (\mathbf{X} - \bar{\mathbf{x}}\mathbf{1}^T)^T(\mathbf{X} - \bar{\mathbf{x}}\mathbf{1}^T)/N$ and patterns are assumed to be centered around $\mathbf{0}$. On the other hand, the energy function of PCN can be written as (derivation in Appendix):

$$E_{PCN}(\mathbf{q}, \mathbf{X}) = \|diagMat(\mathbf{1} \oslash diag(\Sigma^{-1}))\Sigma^{-1}\mathbf{q}\|_2^2 \tag{5}$$

where $\oslash$ is the Hadamard (element-wise) division, $diag$ extracts the diagonal elements of a matrix and converts them into a d-dimensional vector, and $diagMat$ converts a vector into a diagonal matrix. Note in this way, both Eq. 4 (HN) and Eq. 5 (PCN) can be seen as learning a metric from data through the covariance matrix Σ. Fig. 4 visualizes what each model does, when trained on two-dimensional patterns (grey dots), randomly generated from a normal distribution *with a positive correlation*. The correlation makes RM challenging since a typical familiar point (black) and a typical novel point (orange) may be equally away from the mean ($\mathbf{0}$ here, Fig. 4A). Fig. 4B illustrates that HN first transforms all points from Fig. 4A by $L = \Sigma^{\frac{1}{2}}$, and then measures the *negative* distance to the mean $\mathbf{0}$. Because of the negative sign, the further the distance (away from the mean), the more familiar is the query. Although this seems to address the particular problem of indistinguishable black and orange dots, it will actually classify the closest dots to the origin as the most novel. Fig. 4C shows that PCN also transforms all points from Fig. 4A, but by a different matrix. There, without the negative sign, the further the distance (away from the mean), the more novel is the query. Importantly, this has the effect of whitening the data cloud to an extent (i.e., pulling the orange dot from the origin), making it easy to measure familiarity by the distance between the query and the origin.

4 Discussion & Conclusion

Partly inspired by the relationship between two related tasks—AM and RM, we end up discovering a RM model in PCN that is attractive because **(1)** through an energy-based approach, it is an integrated part of a general memory network that can also perform AM; **(2)** it requires only local computations and plasticity [14] and thus has a degree of biological plausibility; **(3)** it performs more robustly in RM tasks, especially when the patterns have an explicit correlated structure such as real-world images. Moreover, we have shown analytically that this superior performance results from the inverse covariance encoded in the recurrent weights of PCN, stretching any query patterns and the mean of stored patterns accordingly before determining their novelty and familiarity. Overall, our work on RM combines recent advances in energy-based models for AM with experimental observation in neuroscience. This symbiosis eventually leads us to a plausible, effective, and general computational mechanism underlying the discrimination of novel and familiar stimuli in the brain or artificial agents.

5 Acknowledgments

This work has been supported by Medical Research Council UK grant MC_UU_00003/1 to RB, an E.P. Abraham Scholarship in the Chemical, Biological/Life and Medical Sciences to MT.

References

[1] Rafal Bogacz, Malcolm Brown, and Christophe Giraud-Carrier. Model of familiarity discrimination in the perirhinal cortex. *Journal of Computational Neuroscience*, 10, 11 2001.

[2] Rafal Bogacz and Malcolm W. Brown. Comparison of computational models of familiarity discrimination in the perirhinal cortex. *Hippocampus*, 13(4), 2003.

[3] Rafal Bogacz, M.W. Brown, and Christophe Giraud-Carrier. High capacity neural networks for familiarity discrimination. volume 2, pages 773 – 778 vol.2, 02 1999.

[4] Malcolm W Brown and John P Aggleton. Recognition memory: what are the roles of the perirhinal cortex and hippocampus? *Nature Reviews Neuroscience*, 2(1):51–61, 2001.

[5] Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168(2):288–299, may 2017.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] Karl Friston. A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360:815 – 836, 2005.

[8] J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

[9] Travis Meyer and Nicole C. Rust. Single-exposure visual memory judgments are reflected in inferotemporal cortex. *eLife*, 7, 2018.

[10] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Milena Pavlovic, Geir Kjetil Sandve, Victor Greiff, David P. Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. *CoRR*, abs/2008.02217, 2020.

[11] Edmund T. Rolls, Peter M. B. Cahusac, Janet Feigenbaum, and Yasushi Miyashita. Responses of single neurons in the hippocampus of the macaque related to recognition memory. *Experimental Brain Research*, 93:299–306, 2004.

[12] Lionel Standing. Learning 10,000 pictures. *The Quarterly journal of experimental psychology*, 25:207–22, 06 1973.

[13] Wendy A. Suzuki. The long and the short of it: Memory signals in the medial temporal lobe. *Neuron*, 24(2):295–298, October 1999.

[14] Mufeng Tang, Tommaso Salvatori, Beren Millidge, Yuhang Song, Thomas Lukasiewicz, and Rafal Bogacz. Recurrent predictive coding models for associative memory employing covariance learning. *PLOS Computational Biology*, 19(4):1–27, 04 2023.

[15] Indre Viskontas, Barbara Knowlton, Peter Steinmetz, and Itzhak Fried. Differences in mnemonic processing by neurons in the human hippocampus and parahippocampal regions. *Journal of cognitive neuroscience*, 18:1654–62, 10 2006.

[16] J.-Z. Xiang and M.W. Brown. Differential neuronal encoding of novelty, familiarity and recency in regions of the anterior temporal lobe. *Neuropharmacology*, 37(4):657–676, 1998.

A Details on the Experimental Procedure

To compare the model performances in Fig. 3, we

1. draw N i.i.d samples from $\mathbf{p}$ as stored patterns as the training set for the model.
2. draw N more i.i.d samples from $\mathbf{p}$ as novel patterns, each time making sure the samples are different from any of the N stored patterns through rejection sampling—rejecting until the sample drawn satisfy this requirement.
3. feed a pair of patterns—one seen, one unseen—into the model as queries (i.e., keep the weights, W, constant) and evaluate each model's energy value on these two patterns. A model's judgment on this pair is correct if its energy value for the novel pattern is higher, and vice versa.
4. Repeat this step for all N seen-unseen pairs and calculate the error rate of a model as the number of incorrect judgments divided by N.

In particular, we calculate the *number of patterns retained* $N_{retained}$ for the bottom row of fig. 3 as

$$N_{retained} = (1 - 2r)N \tag{6}$$

following [12], where $r \in [0, 1]$ is the error rate.

B Comparing variants of Hopfield Networks for RM

To analyze how HN or MCHN performs RM, let us first consider a simple algorithm of RM for comparison—calculating the squared Euclidean distance between the query $\mathbf{q}$ and the mean of all stored patterns.

$$\begin{aligned}
\|\mathbf{q} - \bar{\mathbf{x}}\|_2^2 &= \sum_{i=1}^{d} (q_i - \bar{x}_i)^2 \\
&= \sum_{i=1}^{d} q_i^2 + \bar{x}_i^2 - 2q_i\bar{x}_i \\
&= \|\mathbf{q}\|_2^2 + \|\bar{\mathbf{x}}\|_2^2 - 2\sum_{i=1}^{d} q_i\bar{x}_i \\
&= \|\mathbf{q}\|_2^2 + \|\bar{\mathbf{x}}\|_2^2 - \frac{2}{N}\sum_{\mu=1}^{N}\sum_{i=1}^{d} q_i x_i^{\mu}
\end{aligned}$$

where $\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, ..., \bar{x}_d)$ stands for the mean of all stored patterns in matrix $\mathbf{X}$.

In the classical setting [8] where patterns are binary—i.e., a neuron can either take value 1 or -1, it is always the case that $\|\mathbf{q}\|_2^2 = N$. Since both N and $\|\bar{\mathbf{x}}\|_2^2$ do not depend on the query $\mathbf{q}$, for RM, we can safely ignore them as well as the multiplicative constants $\frac{2}{N}$. This effectively gives us

$$E_{euc}(\mathbf{q}, \mathbf{X}) = -\sum_{\mu=1}^{N}\sum_{i=1}^{d} q_i x_i^{\mu} = -\sum_{\mu=1}^{N} \mathbf{q}^T\mathbf{x}^{\mu} \tag{7}$$

This shares a similar form with HN energy in eq. 1,

$$E_{HN}(\mathbf{q}, \mathbf{X}) = -\sum_{\mu=1}^{N} (\mathbf{q}^T\mathbf{x}^{\mu})^2 \tag{8}$$

And also with MCHN energy, ignoring $\frac{1}{2}\|\mathbf{q}\|_2^2$ term we have

$$E_{MCHN}(\mathbf{q}, \mathbf{X}) = -\log(\sum_{\mu=1}^{N} \exp(\mathbf{q}^T\mathbf{x}^{\mu})) \tag{9}$$

Notice that in [5], eq. 9 is obtained as the limit of $-\sum_{\mu=1}^{N}(\sum_{i=1}^{d} q_i x_i^{\mu})^n$ as $n \to \infty$. We can now see how these models perform RM: each stored pattern, say $\mathbf{x}^{\mu}$, contributes to the evaluation of familiarity through a similarity score determined by dot product with query $\mathbf{q}$. The more similar the two patterns are under the dot product, the lower the energy, the more familiar is the query. The difference between using squared Euclidean distance with the mean as opposed to HN is the weighting of these scores. Consider the case for MCHN where $F(x) = x^n$, as $n \to \infty$, the energy value becomes dominated by the contribution of the *most similar pattern* to the query, which is exactly what the task of RM asks for.

Even when we drop the binary assumptions deal with continuous patterns instead, although the $\|\mathbf{q}\|_2^2$ are no longer constant and depends on the query now, notice that they are on the same order of magnitude with $\mathbf{q}^T\mathbf{x}^{\mu}$, which is just one summand out of N. Thus, in most cases, the overall energy function will be dominated by the sum anyway.

C PCN Performs Metric Learning

The training phase of PCN can be seen as a constrained optimization problem from Eq. 2,

$$\min_{W} \frac{1}{2}\|\mathbf{X} - \mathbf{X}W\|_F^2 \text{ s.t. } diag(W) = \mathbf{0} \tag{10}$$

we can equivalently write the constraints into the Lagrangian

$$\mathcal{L}(W, \lambda) = \frac{1}{2}\|\mathbf{X} - \mathbf{X}W\|_F^2 + \lambda^T diag(W) \tag{11}$$

where $\lambda = (\lambda_1, ..., \lambda_d)$ is a vector of Lagrangian multipliers. Taking gradient w.r.t W yields that,

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial W} &= \frac{\partial}{\partial W}(\frac{1}{2}Tr(\mathbf{X}^T\mathbf{X} - \mathbf{X}^T\mathbf{X}W - W^T\mathbf{X}^T\mathbf{X} + W^T\mathbf{X}^T\mathbf{X}W) + \lambda^T diag(W)) \\
&= -\frac{\partial}{\partial W}Tr(W\mathbf{X}^T\mathbf{X}) + \frac{\partial}{\partial W}Tr(W^T\mathbf{X}^T\mathbf{X}W) + diagMat(\lambda) \\
&= -\mathbf{X}^T\mathbf{X} + \mathbf{X}^T\mathbf{X}W + diagMat(\lambda) \\
&= \mathbf{X}^T\mathbf{X}(W - I) + diagMat(\lambda)
\end{aligned}$$

setting it to $\mathbf{0}$ yields

$$\hat{\lambda} = \mathbf{1} \oslash diag(\Sigma^{-1}) \tag{12}$$

and

$$\hat{W} = I - \Sigma^{-1} diagMat(\hat{\lambda}) \tag{13}$$

where Eq. 12 follows from the constraints that $diag(\hat{W}) = \mathbf{0}$. It can also be verified that $(\hat{W}, \hat{\lambda})$ is indeed the global minimum by plugging it in Eq. 10.

Now, to obtain PCN as performing metric learning in the form Eq. 3, note that

$$\begin{aligned}
E_{PCN}(\mathbf{q}, W) &\propto \|\mathbf{q} - \hat{W}^T\mathbf{q}\|_2^2 \\
&= \|(I - \hat{W})^T\mathbf{q}\|_2^2 \\
&= \|diagMat(\mathbf{1} \oslash diag(\Sigma^{-1}))\Sigma^{-1}\mathbf{q}\|_2^2
\end{aligned}$$

as desired.