METRIC-EXTENDED SUPERVISED NEURAL GAS: WHEN DOES COOPERATION HELP METRIC LEARNING?

Nana Abeka Otoo
Authentic-Network
Chemnitz, Germany
nana.abekaotoo@authentic.network

Asirifi Boa
Independent Researcher
Colorado, USA
boaasirifi@gmail.com

ABSTRACT

Supervised Neural Gas (SNG) introduces the margin-optimized cost function of Generalized Learning Vector Quantization (GLVQ) equipped with neighborhood cooperation from Neural Gas (NG). This yields a highly interpretable learner that is less sensitive to the prototype initialization problem observed in the LVQ family of classification algorithms. Although Hammer, Strickert and Villmann presented the SNG framework as an extension applicable to any differentiable dissimilarity measure, the diagonal relevance weighting adaptation remains the only scheme to have been explored. In this paper, we investigate extensions of SNG with full matrix, local matrix, tangent and class-wise matrix distances by examining how rank-based cooperation interacts with metric parameterization learning. We provide gradient-level analysis of two failure modes: (I) gradient dilution and (II) subtraction amplification, which explain learning behavioral challenges in cooperation-based metric learning. We propose a Supervised Class-Wise Matrix Neural Gas variant (SCMNG) that identifies three structural conditions under which rank-based metric relevance learning can be achieved in light of these recorded failure modes. Ablation experiments against non-cooperative reference models indicate that the impact of rank-based cooperation depends on initialization quality, metric structure and class-wise conditions. When prototypes are placed in a representative manner, NG cooperation mainly results in gradient interference. In contrast, with poor initialization, NG cooperation facilitates the distribution of prototypes across class regions.

Keywords: learning vector quantization, neural gas, metric adaptation, prototype-based classification, gradient analysis

1 Introduction

Learning Vector Quantization (LVQ) [16, 17] represents each class by a set of labelled reference input space vector(s) with which new inputs are classified by nearest-prototype assignment. The resulting classifier is a learner with a decision boundary determined by the prototype positions with highly interpretable classification outcomes [32, 30]. Although LVQs are primed with high generalisation ability, the performance of this family of prototype-based classification algorithms has, among other things, observable limitations that have influenced the formulation of advanced variants. A persistent limitation of LVQ models is their sensitivity to prototype initialisation [23]. The cost function in LVQ, which is conceptually modelled as a multimodal objective with gradient-based training, can be prone to poor local optima [24]. Advances in these areas led to the introduction of a prototype ranked-based neighbourhood cooperation algorithm, Neural Gas (NG) [18], designed for addressing this problem. A remarkable related strategy in soft LVQ [26], which replaces hard winner-takes-all assignment with probabilistic memberships, is worth mentioning. The NG learning scheme updates each prototype for each training input sample by weighting its rank proximity to the input sample rather than distance alone. The consequence is that prototype vectors spread across the data manifold irrespective of their initial positions in the input space, eliminating dependence on initial conditions. Hammer, Strickert and Villmann [13] incorporated the ranked-based prototype cooperation mechanism into the margin-based cost function of Generalized LVQ (GLVQ) [23], yielding a new model, Supervised Relevance NG (SRNG) [9, 10]. Another limitation of LVQs backed by literature and empirical evidence relates to the dependence on a typical fixed distance, namely the squared Euclidean norm, which is inconsiderate of discriminative structure.

Adaptive distance learning [20, 2, 25] such as the Generalized Matrix LVQ (GMLVQ) [25], learns a full linear transformation of the input space and has proven to be a powerful classifier by limiting the impact of such restrictions. The local variant of Matrix GLVQ (LGMLVQ) is designed to provide separate transformations for each prototype and for relevance matrix assignments, whereas Limited Rank Matrix LVQ (LiRaM-LVQ) [4] imposes a fixed-rank constraint on the mapping. The Generalised Tangent distance variants [28, 29] complementarily learn invariances in the data subspaces rather than discriminative projections. The propositions from global through class-wise to per-prototype parameterisation to learning relevances are sufficiently represented in the LVQ literature [3, 4] within the wider metric learning research community. Notably, where the topic of Large Margin Nearest Neighbour (LMNN) [34] cum per-class extensions analogously point towards trade-offs. Deep metric learning methodologies [19] learn nonlinear embedding spaces that subsume linear projections, whilst prototype-based learners [27] combine learned embeddings with nearest-prototype classification, the outcome of which is an interaction between a specified metric structure and prototype organisation, leading to a question that goes beyond classical LVQ algorithms. The authors conclude that the SRNG framework already accommodates arbitrary differentiable dissimilarity measures [33]. In addition, there are demonstrative and pragmatic insights from learned metrics in the GMLVQ family of advanced learners. Therefore, combining the two approaches is a natural extension worth investigating. Topographic maps for metric learning have been extensively researched [15]. However, the interaction in transitioning rank-based cooperation from NG throughout various tiers of metric parameterisation remains an open question. In this regard, we set out to investigate what happens when SNG is equipped with a full matrix, a per-prototype matrix, or tangent distance.

We thus instantiate the SNG framework [13] with the aforementioned additional metric parameterizations and the diagonal relevance used by Hammer et al. The resulting models are equipped with full-matrix (SMNG), per-prototype-matrix (SLNG), tangent (STNG) and class-wise-matrix (SCMNG) distances. While these respective distance functions are well established in the baseline LVQ family of classification models, this work focuses on analysing how NG cooperation affects each level of metric adaptation. A gradient-level analysis uncovers two insightful failure modes: gradient dilution and subtraction amplification. These failure modes are responsible for the consistent degradation of SMNG and STNG learners relative to their non-cooperative baselines. From these failure modes are three observable structural conditions enabling prototype rank-based cooperation-compatible metrics (additivity, gradient alignment and locality preservation). We progressively note a baseline recovery property pointing in the direction that the NG framework is incapable of reducing expressiveness in principle. We separate the effect of cooperation from that of per-class metric parameterisation and perform an ablation study against Class-Wise GMLVQ (CW-GMLVQ), which is actually LGMLVQ tied to per-class parameters, or equivalently the non-cooperative limit of SCMNG obtained by fixing neighbourhood range $\gamma \to 0$. The performance evaluation gains of SCMNG over the baseline GMLVQ are attributable to the per-class metric structure utilised in its framework, not to NG cooperation. Under class-conditional representative vector initialisation, prototypes are already distributed within their respective class regions, thereby shifting the role of NG cooperation towards redundancy and gradient interference. Beyond classification generalisation improvements, the SCMNG model's per-class metric tensors Λ_c permit class-level interpretability, directly indicating which features are relevant for discriminating class c as opposed to the regular need to aggregate across prototypes. Section 2 reviews LVQs with adaptive metrics and SNG. Section 3 introduces the four metric extensions and their respective adaptation rules. Section 4 provides a theoretical evaluation that incorporates the observed behavioral learning dynamics. Section 5 presents the experimental evaluation conducted on four benchmark data sets. The final section offers the study's conclusions.

2 Learning Vector Quantization and Supervised Neural Gas

Consider a training set $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_N\} \subset \mathbb{R}^d$ with class labels $c_{\mathbf{v}} \in \mathcal{L}$, where $\mathcal{L}$ is finite. A prototype-based classifier maintains a codebook $\mathbf{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_K\}$ with labels $c_r \in \mathcal{L}$. We write $\mathbf{W}_c = \{\mathbf{w}_r \mid c_r = c\}$ for the subset belonging to class c and $P_c = |\mathbf{W}_c|$ for the number of prototypes per class. An input $\mathbf{v}$ is assigned to the class of its nearest prototype: $\mathbf{v} \mapsto c_r$ where $d(\mathbf{v}, \mathbf{w}_r)$ is minimal under some differentiable dissimilarity measure d. Generalized LVQ [23] casts this as an optimization problem by minimizing

$$E_{\mathrm{GLVQ}} = \sum_{\mathbf{v} \in V} \Phi(\mu(\mathbf{v})), \qquad \mu(\mathbf{v}) = \frac{d(\mathbf{v}, \mathbf{w}^+) - d(\mathbf{v}, \mathbf{w}^-)}{d(\mathbf{v}, \mathbf{w}^+) + d(\mathbf{v}, \mathbf{w}^-)}, \tag{1}$$

where $\mathbf{w}^+$ and $\mathbf{w}^-$ denote the closest correct-class and wrong-class prototypes,

$$\mathbf{w}^+ = \underset{\mathbf{w}_r:\, c_r = c_{\mathbf{v}}}{\arg\min}\; d(\mathbf{v}, \mathbf{w}_r), \qquad \mathbf{w}^- = \underset{\mathbf{w}_r:\, c_r \neq c_{\mathbf{v}}}{\arg\min}\; d(\mathbf{v}, \mathbf{w}_r).$$

The transfer function Φ is conditioned to be monotonically increasing and mostly exemplified without loss of generality as the logistic sigmoid $\Phi(x) = (1 + e^{-\beta x})^{-1}$ with steepness $\beta > 0$ for stochastic gradient descent learning

[6, 7]. The normalised margin $\mu(\mathbf{v})$ takes values in $[-1, 1]$ with negative values indicating the correct classification and vice versa. This formulation instantiates the minimum classification error principle, as seen in the work of Juang and Katagiri [14] and Crammer et al. [8] on theoretical generalization bounds, via its connection to hypothesis-margin optimization. A number of LVQ extensions replace the fixed Euclidean distance in GLVQ with learnable metrics. The Generalized Matrix LVQ (GMLVQ) [25] parameterises a global linear transformation $\Omega \in \mathbb{R}^{l\times d}$ which gives rise to the distance

$$d_\Omega(\mathbf{v}, \mathbf{w}) = \|\Omega(\mathbf{v} - \mathbf{w})\|^2 = (\mathbf{v} - \mathbf{w})^T \Lambda(\mathbf{v} - \mathbf{w}), \tag{2}$$

where $\Lambda = \Omega^T\Omega \succeq 0$ is the derived metric tensor, normalised so that $\sum_i \Lambda_{ii} = 1$. The local extension LGMLVQ equips each prototype with its own transformation Ω_k. Generalized Tangent LVQ (GTLVQ) [29] pursues a different learning objective per prototype invariance subspaces through

$$d_T(\mathbf{v}, \mathbf{w}_k) = \|\boldsymbol{\delta}_k\|^2 - \|\Omega_k^T \boldsymbol{\delta}_k\|^2, \qquad \boldsymbol{\delta}_k = \mathbf{v} - \mathbf{w}_k, \tag{3}$$

Directions in the column span of $\Omega_k \in \mathbb{R}^{d\times s}$ are treated as invariant near $\mathbf{w}_k$ and are projected out of the distance. A simpler but often effective alternative is Generalized Relevance LVQ (GRLVQ) [9], which restricts the metric to a diagonal weighting:

$$d_\lambda(\mathbf{v}, \mathbf{w}) = \sum_{j=1}^{d} \lambda_j \, (v_j - w_j)^2, \qquad \lambda_j \geq 0, \quad \sum_j \lambda_j = 1, \tag{4}$$

Each weight λ_j captures the relevance of feature j and the non-negativity constraint matters here because any increase in λ_j can only increase distances, a property whose significance will become apparent when relevance weighting is combined with NG cooperation. Though the metric adaptation formulation tackles the distance problem, it does not address the prototype initialization sensitivity problem. This sensitivity therefore requires a separate mechanism.

Neural Gas [18, 21] is an unsupervised vector quantization method that distributes prototypes across the data manifold by means of rank-based cooperation. Given an input $\mathbf{v}$, the prototypes are sorted based on their distance to $\mathbf{v}$ and updated as

$$\Delta\mathbf{w}_r = \varepsilon \cdot h_\gamma(k_r(\mathbf{v}, \mathbf{W})) \cdot (\mathbf{v} - \mathbf{w}_r), \tag{5}$$

Here $k_r(\mathbf{v}, \mathbf{W})$ denotes the rank of $\mathbf{w}_r$, which is defined as the number of prototypes closer to $\mathbf{v}$ than $\mathbf{w}_r$ and $h_\gamma(k) = \exp(-k/\gamma)$ is the neighbourhood function controlled by a range parameter $\gamma > 0$. Over the course of training γ is annealed toward zero, conditioned on the principle that cooperation gradually concentrates on nearby prototypes and eventually recovers winner-takes-all behaviour [5, 11]. Since neighbourhood cooperation is rank-based as opposed to distance-based, the resulting quantization preserves topological structure [31]. A practical consequence is robustness to initialisation occurs when prototypes that start far from the data manifold receive updates during the initial epochs when γ is large, preventing dead units. Hammer, Strickert and Villmann [12, 13] proceeded to introduce these together by embedding NG cooperation in the GLVQ cost function, resulting in the Supervised Neural Gas (SNG):

$$E_{\mathrm{SNG}} = \sum_{\mathbf{v}\in V} \sum_{\mathbf{w}_r \in \mathbf{W}_{c_\mathbf{v}}} \frac{h_\gamma(k_r(\mathbf{v}, \mathbf{W}_{c_\mathbf{v}}))}{C(\gamma, K_{c_\mathbf{v}})} \cdot \Phi(\mu_r(\mathbf{v})), \tag{6}$$

with $K_{c_\mathbf{v}} = |\mathbf{W}_{c_\mathbf{v}}|$ and normalisation constant $C(\gamma, K_{c_\mathbf{v}}) = \sum_{\mathbf{w}_r \in \mathbf{W}_{c_\mathbf{v}}} h_\gamma(k_r)$. The ranking $k_r(\mathbf{v}, \mathbf{W}_{c_\mathbf{v}})$ is computed among same-class prototypes only and the generalised margin

$$\mu_r(\mathbf{v}) = \frac{d(\mathbf{v}, \mathbf{w}_r) - d^-(\mathbf{v})}{d(\mathbf{v}, \mathbf{w}_r) + d^-(\mathbf{v})}, \qquad d^-(\mathbf{v}) = \min_{\mathbf{w}_j : c_j \neq c_\mathbf{v}} d(\mathbf{v}, \mathbf{w}_j) \tag{7}$$

extends the GLVQ margin from the single winner to every same-class prototype. The normalised weights $\bar{h}_r = h_\gamma(k_r)/C(\gamma, K_{c_\mathbf{v}})$ give all correct-class prototypes a stake in the cost, with influence that decays monotonically with rank. When γ is large, many prototypes contribute substantially by importing the topological coordination of NG into the supervised classification setting. The cost (6) is defined for any differentiable distance d^λ with metric parameters λ, so update rules for a particular metric are followed by straightforward differentiation. For a training sample $\mathbf{v}$ of class $c_\mathbf{v}$, we write $\boldsymbol{\delta}_r = \mathbf{v} - \mathbf{w}_r$ for the displacement to prototype $\mathbf{w}_r$ and $\boldsymbol{\delta}_{r-} = \mathbf{v} - \mathbf{w}_{r-}$ for the displacement to the closest wrong-class prototype. Defining

$$\xi_r^+ = \frac{2\, d_{r-}}{(d_r + d_{r-})^2}, \qquad \xi_r^- = \frac{2\, d_r}{(d_r + d_{r-})^2}, \tag{8}$$

where $d_r = d(\mathbf{v}, \mathbf{w}_r)$ and $d_{r-} = d(\mathbf{v}, \mathbf{w}_{r-})$, applying gradient descent on (6) produces the updates below [13]:

$$\Delta \mathbf{w}_r = -\varepsilon^+ \cdot \bar{h}_r \cdot \Phi' |_{\mu^r} \cdot \xi_r^+ \cdot \frac{\partial d_r}{\partial \mathbf{w}_r}, \qquad \mathbf{w}_r \in \mathbf{W}_{c_\mathbf{v}}, \tag{9}$$

$$\Delta \mathbf{w}_{r-} = \varepsilon^- \cdot \sum_{\mathbf{w}_r \in \mathbf{W}_{c_\mathbf{v}}} \bar{h}_r \cdot \Phi' |_{\mu^r} \cdot \xi_r^- \cdot \frac{\partial d_{r-}}{\partial \mathbf{w}_{r-}}, \tag{10}$$

$$\Delta \lambda = -\varepsilon_\lambda \cdot \sum_{\mathbf{w}_r \in \mathbf{W}_{c_\mathbf{v}}} \bar{h}_r \cdot \Phi' |_{\mu^r} \cdot \Big(\xi_r^+ \cdot \frac{\partial d_r}{\partial \lambda} - \xi_r^- \cdot \frac{\partial d_{r-}}{\partial \lambda} \Big), \tag{11}$$

We note here Φ' as the derivative of the logistic transfer function evaluated at $\mu^r(\mathbf{v})$. These three distinct update rules manifest significant roles: The equation (9) attracts each same-class prototype toward $\mathbf{v}$ whilst (10) repels the closest wrong-class prototype with the metric adaptation parameters handled by equation (11). What differs from one metric extension to another is the distance gradients $\partial d/\partial \mathbf{w}$ and $\partial d/\partial \lambda$ with the overall structure remaining the same. Following Hammer et al. [13], we utilise the same parameter setup for all NG models in this work with separate learning rates ε^+ and ε^- for correct-class and wrong-class prototype updates. In a modern algorithmic differentiation framework where a single optimiser controls the step size, this separation is realised by a stop-gradient scaling of d^- given by:

$$\tilde{d}^- = \mathrm{sg}(d^-) + \rho \cdot (d^- - \mathrm{sg}(d^-)), \qquad \rho = \varepsilon^- / \varepsilon^+, \tag{12}$$

where the operator $\mathrm{sg}(\cdot)$ stops gradient propagation. In the forward pass $\tilde{d}^- = d^-$ identically and in the backward pass $\partial L / \partial d^-$ is scaled by ρ. This scales the wrong-class contribution to all parameter gradients, inclusive of the metric parameters Ω and not only prototype positions. Consider the baseline model in the work of Hammer et al., where they apply separate learning rates ε^+ and ε^- to prototype updates only, along with a distinct rate ε_λ for the metric. The stop-gradient construction couples the prototype and metric learning rate ratios by design. Pragmatically observed within algorithmic differentiation frameworks in which a single optimizer controls all parameters, this coupling follows automatically. Setting $\rho \approx 0.5$ attenuates repulsive updates relative to attractive ones as recommended in the original work.

3 Metric Extensions of Supervised Neural Gas

We now introduce the extension of the SNG cost (6) to four additional distance functions, using SRNG as a baseline. In each case, the cost function retains the structure of (6), with changes only in the distance d and the learnable parameters. We note that prototype ranks within $\mathbf{W}_{c_\mathbf{v}}$ are always computed under the respective metric.

3.1 SRNG: Diagonal Relevance

The base instantiation due to Hammer et al. [13] is Supervised Relevance Neural Gas (SRNG), which utilizes the diagonal relevance distance (4) and formulated as:

$$d_\lambda(\mathbf{v}, \mathbf{w}_r) = \sum_{j=1}^{d} \lambda_j \, (v_j - w_{r,j})^2. \tag{13}$$

The resulting cost function is

$$E_{\mathrm{SRNG}}(\mathbf{W}, \boldsymbol{\lambda}) = \sum_{\mathbf{v} \in V} \sum_{\mathbf{w}_r \in \mathbf{W}_{c_\mathbf{v}}} \bar{h}_r \cdot \Phi(\mu_r^\lambda(\mathbf{v})), \tag{14}$$

with $\bar{h}_r = h_\gamma(k_r) / C(\gamma, K_{c_\mathbf{v}})$ and μ_r^λ evaluated under the diagonal relevance distance (13). The learnable parameters are the prototypes and the relevance vector $\boldsymbol{\lambda} \in \mathbb{R}^d$. In the limit $\gamma \to 0$, the SRNG model recovers the GRLVQ model. The distance gradients are computed by

$$\frac{\partial d_\lambda}{\partial w_{r,j}} = -2 \, \lambda_j \, (v_j - w_{r,j}), \qquad \frac{\partial d_\lambda}{\partial \lambda_j} = (v_j - w_{r,j})^2. \tag{15}$$

Since $\partial d_\lambda / \partial \lambda_j$ is non-negative for every prototype, cooperation gradients from distant prototypes produce relevance updates that are consistently aligned. We emphasize that this sign guarantee does not carry over to full matrix parameterizations.

3.2 SMNG: Global Matrix

SMNG moves beyond diagonal weighting by adopting the full GMLVQ matrix metric (2). A single transformation $\Omega \in \mathbb{R}^{l\times d}$ is shared among all prototypes:

$$d_\Omega(\mathbf{v}, \mathbf{w}_r) = \|\Omega(\mathbf{v} - \mathbf{w}_r)\|^2. \tag{16}$$

The cost function to optimise is given by

$$E_{\mathrm{SMNG}}(\mathbf{W}, \Omega) = \sum_{\mathbf{v}\in V} \sum_{\mathbf{w}_r \in \mathbf{W}_{c_\mathbf{v}}} \bar{h}_r \cdot \Phi(\mu_r^\Omega(\mathbf{v})), \tag{17}$$

where $\bar{h}_r = h_\gamma(k_r)/C(\gamma, K_{c_\mathbf{v}})$ and μ_r^Ω evaluated under the matrix distance (16). Similarly, the optimization goal regarding the parameters to learn is the prototypes and a single global transformation $\Omega \in \mathbb{R}^{l\times d}$. Setting $\gamma \to 0$ recovers GMLVQ. The distance gradients are

$$\frac{\partial d_\Omega}{\partial \mathbf{w}_r} = -2\,\Lambda\,\boldsymbol{\delta}_r, \qquad \frac{\partial d_\Omega}{\partial \Omega} = 2\,\Omega\,\boldsymbol{\delta}_r\boldsymbol{\delta}_r^T, \tag{18}$$

with $\Lambda = \Omega^T\Omega$ and $\boldsymbol{\delta}_r = \mathbf{v} - \mathbf{w}_r$. Substituting into the general update rules (9)–(11) gives the equations

$$\Delta\mathbf{w}_r = 2\varepsilon^+ \cdot \bar{h}_r \cdot \Phi'|_{\mu^r} \cdot \xi_r^+ \cdot \Lambda\,\boldsymbol{\delta}_r, \tag{19}$$

$$\Delta\mathbf{w}_{r-} = -2\varepsilon^- \cdot \Lambda\,\boldsymbol{\delta}_{r-} \cdot \sum_{\mathbf{w}_r \in \mathbf{W}_{c_\mathbf{v}}} \bar{h}_r \cdot \Phi'|_{\mu^r} \cdot \xi_r^-, \tag{20}$$

together with the matrix adaptations

$$\Delta\Omega = -2\varepsilon_\Omega \cdot \sum_{\mathbf{w}_r \in \mathbf{W}_{c_\mathbf{v}}} \bar{h}_r \cdot \Phi'|_{\mu^r} \cdot \Omega\big(\xi_r^+ \cdot \boldsymbol{\delta}_r\boldsymbol{\delta}_r^T - \xi_r^- \cdot \boldsymbol{\delta}_{r-}\boldsymbol{\delta}_{r-}^T\big), \tag{21}$$

followed by Ω matrix normalisation $\Omega \leftarrow \Omega/\sqrt{\mathrm{tr}(\Lambda)}$. The key point is that (21) folds rank-1 contributions from *all* cooperating same-class prototypes into a single shared matrix. It is this aggregation that gives rise to the gradient dilution phenomenon observed and analyzed in Section 4.

3.3 SLNG: Per-Prototype Matrix

A considerable direction to avoid sharing a single matrix that brings out tendencies of gradient dilution is to give each prototype its own transformation exactly as utilized in the local base learner LGMLVQ. The resulting formulation of Supervised Local NG does exactly this by ensuring that each prototype $\mathbf{w}_k$ carries a private $\Omega_k \in \mathbb{R}^{l\times d}$ where

$$d_k(\mathbf{v}, \mathbf{w}_k) = \|\Omega_k(\mathbf{v} - \mathbf{w}_k)\|^2. \tag{22}$$

for the cost function

$$E_{\mathrm{SLNG}}(\mathbf{W}, \{\Omega_k\}) = \sum_{\mathbf{v}\in V} \sum_{\mathbf{w}_r \in \mathbf{W}_{c_\mathbf{v}}} \bar{h}_r \cdot \Phi(\mu_r^{\Omega_k}(\mathbf{v})), \tag{23}$$

with $\bar{h}_r = h_\gamma(k_r)/C(\gamma, K_{c_\mathbf{v}})$ and $\mu_r^{\Omega_k}$ evaluated under the per-prototype distance (22). The optimization problem entails learning the prototypes and the K transformation matrices $\{\Omega_k\}_{k=1}^K$, as in the baseline local model. Observe that we obtain the same baseline LGMLVQ learner as $\gamma \to 0$. The distance gradients take the familiar form

$$\frac{\partial d_k}{\partial \mathbf{w}_k} = -2\,\Lambda_k\,\boldsymbol{\delta}_k, \qquad \frac{\partial d_k}{\partial \Omega_k} = 2\,\Omega_k\,\boldsymbol{\delta}_k\boldsymbol{\delta}_k^T, \tag{24}$$

with $\Lambda_k = \Omega_k^T\Omega_k$. The prototype updates are computed similarly by

$$\Delta\mathbf{w}_k = 2\varepsilon^+ \cdot \bar{h}_k \cdot \Phi'|_{\mu^k} \cdot \xi_k^+ \cdot \Lambda_k\,\boldsymbol{\delta}_k, \tag{25}$$

$$\Delta\mathbf{w}_{r-} = -2\varepsilon^- \cdot \Lambda_{r-}\,\boldsymbol{\delta}_{r-} \cdot \sum_{\mathbf{w}_r \in \mathbf{W}_{c_\mathbf{v}}} \bar{h}_r \cdot \Phi'|_{\mu^r} \cdot \xi_r^-. \tag{26}$$

Recall also that Ω_k belongs to $\mathbf{w}_k$ alone and hence only the cost term involving d_k contributes to its gradient:

$$\Delta\Omega_k = -2\varepsilon_\Omega \cdot \bar{h}_k \cdot \Phi'|_{\mu^k} \cdot \xi_k^+ \cdot \Omega_k\,\boldsymbol{\delta}_k\boldsymbol{\delta}_k^T, \tag{27}$$

The wrong-class prototype's matrix, in contrast, accumulates the full repulsive signal:

$$\Delta\Omega_{r-} = 2\varepsilon_\Omega \cdot \Omega_{r-}\, \boldsymbol{\delta}_{r-}\boldsymbol{\delta}_{r-}^T \cdot \sum_{\mathbf{w}_r \in \mathbf{W}_{c_\mathbf{v}}} \bar{h}_r \cdot \Phi'|_{\mu^r} \cdot \xi_r^-. \tag{28}$$

After each adaptation step, the Ω matrix normalization, $\Omega_k \leftarrow \Omega_k / \sqrt{\mathrm{tr}(\Lambda_k)}$ takes place to prevent the degeneration of the solution in accordance with the recommendations in [13]. The per-prototype ownership means that cooperation from distant same-class prototypes can provide extra measurement directions without interfering with other prototypes' metric adaptation. In Section 4, we theoretically examine whether this is beneficial overall or in contrast.

3.4 STNG: Tangent Distance

Supervised Tangent NG departs from the additive framework entirely by combining SNG with per-prototype tangent distance (3) given by

$$d_T(\mathbf{v}, \mathbf{w}_k) = \|\boldsymbol{\delta}_k\|^2 - \|\Omega_k^T \boldsymbol{\delta}_k\|^2, \tag{29}$$

and the optimisable learning cost as a function given by

$$E_{\mathrm{STNG}}(\mathbf{W}, \{\Omega_k\}) = \sum_{\mathbf{v}\in V} \sum_{\mathbf{w}_r \in \mathbf{W}_{c_\mathbf{v}}} \bar{h}_r \cdot \Phi(\mu_r^T(\mathbf{v})), \tag{30}$$

with $\bar{h}_r = h_\gamma(k_r)/C(\gamma, K_{c_\mathbf{v}})$ and μ_r^T evaluated under the tangent distance (29). The prototypes and K orthonormal bases $\{\Omega_k\}_{k=1}^K$ remain the parameters to learn. The $\Omega_k \in \mathbb{R}^{d\times s}$ herein used is an orthonormal basis spanning the invariance subspace of prototype $\mathbf{w}_k$. Setting $\gamma \to 0$ recovers the baseline GTLVQ learner. The distance gradients are

$$\frac{\partial d_T}{\partial \mathbf{w}_k} = -2\,(I - \Omega_k\Omega_k^T)\,\boldsymbol{\delta}_k, \qquad \frac{\partial d_T}{\partial \Omega_k} = -2\,\boldsymbol{\delta}_k\boldsymbol{\delta}_k^T\Omega_k. \tag{31}$$

with the prototype updates executed as

$$\Delta\mathbf{w}_k = 2\varepsilon^+ \cdot \bar{h}_k \cdot \Phi'|_{\mu^k} \cdot \xi_k^+ \cdot (I - \Omega_k\Omega_k^T)\,\boldsymbol{\delta}_k, \tag{32}$$

$$\Delta\mathbf{w}_{r-} = -2\varepsilon^- \cdot (I - \Omega_{r-}\Omega_{r-}^T)\,\boldsymbol{\delta}_{r-} \cdot \sum_{\mathbf{w}_r \in \mathbf{W}_{c_\mathbf{v}}} \bar{h}_r \cdot \Phi'|_{\mu^r} \cdot \xi_r^-. \tag{33}$$

The update for the correct-class subspace reads

$$\Delta\Omega_k = 2\varepsilon_\Omega \cdot \bar{h}_k \cdot \Phi'|_{\mu^k} \cdot \xi_k^+ \cdot \boldsymbol{\delta}_k\boldsymbol{\delta}_k^T\Omega_k, \tag{34}$$

The motivational effect is to rotate Ω_k toward $\boldsymbol{\delta}_k$ by declaring the displacement direction invariant and thereby reducing d_T. The wrong-class subspace receives the update

$$\Delta\Omega_{r-} = -2\varepsilon_\Omega \cdot \boldsymbol{\delta}_{r-}\boldsymbol{\delta}_{r-}^T\Omega_{r-} \cdot \sum_{\mathbf{w}_r \in \mathbf{W}_{c_\mathbf{v}}} \bar{h}_r \cdot \Phi'|_{\mu^r} \cdot \xi_r^-. \tag{35}$$

After each step, the Ω_k is re-orthonormalized by QR decomposition. The interaction with cooperation is problematic when a distant sample $\mathbf{v}$ generates a large $\boldsymbol{\delta}_k$ pointing along the inter-prototype axis, causing Ω_k to absorb that direction as invariant. We refer to this as the subtraction amplification mechanism and provide a detailed explanation accounting for it in Section 4.

3.5 SCMNG: Class-Wise Matrix

The models above occupy two extremes: SMNG shares a single Ω across the entire codebook (cheap in parameters but dilution-prone), whereas SLNG gives every prototype its own Ω_k (maximally expressive but parameter-heavy). SCMNG sits between these two learners as a middle-ground trade-off. It replaces the global Ω of SMNG with per-class transformation matrices, where all prototypes of class c share a common $\Omega_c \in \mathbb{R}^{l\times d}$, by formulating the dissimilarity measure as

$$d_c(\mathbf{v}, \mathbf{w}_r) = \|\Omega_{c_r}(\mathbf{v} - \mathbf{w}_r)\|^2, \qquad c_r = c(\mathbf{w}_r). \tag{36}$$

for the learnable task captured as a function of the form

$$E_{\mathrm{SCMNG}}(\mathbf{W}, \{\Omega_c\}) = \sum_{\mathbf{v}\in V} \sum_{\mathbf{w}_r \in \mathbf{W}_{c_\mathbf{v}}} \bar{h}_r \cdot \Phi(\mu_r^{\Omega_c}(\mathbf{v})), \tag{37}$$

Algorithm 1 SCMNG Training

Require: Training data V with labels, P_c prototypes per class, epochs T, learning rate ε, γ_{init}, γ_{final}, learning rate ratio ρ

Initialise prototypes $\mathbf{W}$ with P_c representatives from V
Initialise $\Omega_c \leftarrow \mathbf{I}_{l\times d}$ for each class c
$\gamma \leftarrow \gamma_{\text{init}}$
for $t = 1$ to T **do**
  for each $\mathbf{v} \in V$ **do**
    Compute $d_c(\mathbf{v}, \mathbf{w}_r)$ for all $\mathbf{w}_r$ using (36)
    Rank same-class prototypes: $k_r \leftarrow \text{rank}(\mathbf{w}_r, \mathbf{v}, \mathbf{W}_{c_\mathbf{v}})$
    Compute normalised weights: $\bar{h}_r \leftarrow e^{-k_r/\gamma} / \sum_{r'} e^{-k_{r'}/\gamma}$
    Find $d^-(\mathbf{v}) = \min_{j:c_j \neq c_\mathbf{v}} d_{c_j}(\mathbf{v}, \mathbf{w}_j)$
    Scale wrong-class gradient: $\tilde{d}^- \leftarrow \text{sg}(d^-) + \rho \cdot (d^- - \text{sg}(d^-))$
    For each $\mathbf{w}_r \in \mathbf{W}_{c_\mathbf{v}}$: compute μ_r using d_r and $\tilde{d}^-$
    Compute cost $E = \sum_r \bar{h}_r \cdot \Phi(\mu_r)$
    Update $\mathbf{W}$ and $\{\Omega_c\}$ via gradient of E
    Normalise: $\Omega_c \leftarrow \Omega_c / \sqrt{\text{tr}(\Omega_c^T \Omega_c)}$ for each c
  end for
  Decay: $\gamma \leftarrow \gamma_{\text{init}} \cdot (\gamma_{\text{final}}/\gamma_{\text{init}})^{t/T}$
end for
return Prototypes $\mathbf{W}$, class-wise metrics $\{\Omega_c\}_{c\in\mathcal{L}}$

with $\bar{h}_r = h_\gamma(k_r)/C(\gamma, K_{c_\mathbf{v}})$ and $\mu_r^{\Omega_c}$ evaluated under the class-wise dissimilarity measure in equation (36). We stipulate that the learnable parameters which are the prototypes and C transformation matrices $\{\Omega_c\}_{c\in\mathcal{L}}$ are jointly trained. Regarding the parameterization as a perspective, it is significant to recall that SMNG maintains ld metric parameters (one global Ω), SCMNG maintains $C \cdot ld$ (one per class) and SLNG maintains $K \cdot ld = C \cdot P_c \cdot ld$ (one per prototype). Since $P_c \geq 1$, SCMNG never exceeds SLNG in number of parameters, with equality only when $P_c = 1$. In the typical multi-prototype model setup, SCMNG is P_c times more parameter-efficient. Differentiating with respect to the model parameters gives the distance gradients

$$\frac{\partial d_c}{\partial \mathbf{w}_r} = -2\,\Lambda_{c_r}\,\boldsymbol{\delta}_r, \qquad \frac{\partial d_c}{\partial \Omega_{c_r}} = 2\,\Omega_{c_r}\,\boldsymbol{\delta}_r\boldsymbol{\delta}_r^T, \tag{38}$$

with $\Lambda_{c_r} = \Omega_{c_r}^T\Omega_{c_r}$. The prototype updates follow the usual pattern:

$$\Delta\mathbf{w}_r = 2\varepsilon^+ \cdot \bar{h}_r \cdot \Phi'|_{\mu^r} \cdot \xi_r^+ \cdot \Lambda_{c_\mathbf{v}}\,\boldsymbol{\delta}_r, \tag{39}$$

$$\Delta\mathbf{w}_{r-} = -2\varepsilon^- \cdot \Lambda_{c_{r-}}\,\boldsymbol{\delta}_{r-} \cdot \sum_{\mathbf{w}_r\in\mathbf{W}_{c_\mathbf{v}}} \bar{h}_r \cdot \Phi'|_{\mu^r} \cdot \xi_r^-. \tag{40}$$

The metric update is where SCMNG differs structurally from SMNG. The correct-class distance d_r depends on $\Omega_{c_\mathbf{v}}$ while the wrong-class distance d_{r-} depends on $\Omega_{c_{r-}}$, which are both *distinct* matrices with own adaptions during learning. This separation means the metric updates decompose into independent per-class contributions:

$$\Delta\Omega_{c_\mathbf{v}} = -2\varepsilon_\Omega \cdot \sum_{\mathbf{w}_r\in\mathbf{W}_{c_\mathbf{v}}} \bar{h}_r \cdot \Phi'|_{\mu^r} \cdot \xi_r^+ \cdot \Omega_{c_\mathbf{v}}\,\boldsymbol{\delta}_r\boldsymbol{\delta}_r^T, \tag{41}$$

$$\Delta\Omega_{c_{r-}} = 2\varepsilon_\Omega \cdot \Omega_{c_{r-}}\,\boldsymbol{\delta}_{r-}\boldsymbol{\delta}_{r-}^T \cdot \sum_{\mathbf{w}_r\in\mathbf{W}_{c_\mathbf{v}}} \bar{h}_r \cdot \Phi'|_{\mu^r} \cdot \xi_r^-, \tag{42}$$

each followed by normalization $\Omega_c \leftarrow \Omega_c/\sqrt{\text{tr}(\Lambda_c)}$, as in the SMNG and SLNG learners. The structural advantage is visible in (41) where the outer products $\boldsymbol{\delta}_r\boldsymbol{\delta}_r^T$ are summed only over prototypes of class $c_\mathbf{v}$ and these rank-1 terms satisfy the non-negative alignment property (58) derived in Section 4. The SCMNG learners framework is grounded on the fact that prototypes of different classes update different matrices, so cross-class gradient contamination is eliminated by construction.

4 Theoretical Analysis

This section reviews the interaction between NG cooperation and metric learning within the LVQ family of classifiers. The analysis starts by identifying two critical failure modes, followed by an examination of metrics designed to address them.

4.1 Gradient Dilution in Shared Metrics

The first notable failure mode concerns SMNG, which comes from the way cooperation gradients accumulate in the shared Ω matrix. The full gradient of the SMNG cost with respect to Ω receives contributions through both d_r (correct-class) and d_{r-} (wrong-class), as reflected in (21). The problematic aspect in the formulation is apparent in the correct-class component, which sums over all cooperating same-class prototypes

$$\left.\frac{\partial E}{\partial \Omega}\right|_{d_r} = \sum_{\mathbf{v}} \sum_{\mathbf{w}_r \in \mathbf{W}_{c_\mathbf{v}}} \bar{h}_r \cdot \frac{\partial \Phi(\mu_r)}{\partial \mu_r} \cdot \frac{\partial \mu_r}{\partial d_r} \cdot \frac{\partial d_r}{\partial \Omega}. \tag{43}$$

The wrong-class component involves only a single prototype $\mathbf{w}_{r-}$ per sample and is not subject to dilution. The term that matters in (43) is the distance gradient:

$$\frac{\partial d_\Omega(\mathbf{v}, \mathbf{w}_r)}{\partial \Omega} = 2\,\Omega\,\boldsymbol{\delta}_r \boldsymbol{\delta}_r^T, \qquad \boldsymbol{\delta}_r = \mathbf{v} - \mathbf{w}_r. \tag{44}$$

Each such term pushes Ω to strengthen its projection along $\boldsymbol{\delta}_r$. For the winning prototype ($k_r = 0$), $\boldsymbol{\delta}_r$ is minimal and consistent with the local class boundary geometry near a point $\mathbf{v}$ in the input space. For cooperating prototypes ($k_r > 0$), $\boldsymbol{\delta}_r$ is maximal and dominated by the displacement between prototype receptive fields, meaning it encodes inter-cluster geometry instead of a decision boundary structure at $\mathbf{v}$.

Definition 1 (Gradient Dilution)**.** *Let $\mathbf{g}_k = \bar{h}_k \cdot \alpha_k \cdot \Omega\,\boldsymbol{\delta}_k \boldsymbol{\delta}_k^T$ be the contribution of prototype $\mathbf{w}_k$ to $\partial E / \partial \Omega$ for sample $\mathbf{v}$, where α_k absorbs the transfer function derivative and margin scaling. Gradient dilution occurs when*

$$\left\| \sum_k \mathbf{g}_k \right\|_F \ll \sum_k \|\mathbf{g}_k\|_F, \tag{45}$$

that is, the cooperating gradient contributions partially cancel rather than reinforce (where $\sum_k \|\mathbf{g}_k\|_F > 0$).

To make the severity of dilution precise, we define the *dilution ratio*

$$\rho_{\text{dil}} = 1 - \frac{\left\|\sum_k \mathbf{g}_k\right\|_F}{\sum_k \|\mathbf{g}_k\|_F} \in [0, 1], \tag{46}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. At one extreme, $\rho_{\text{dil}} = 0$ means all gradients are perfectly aligned; at the other, $\rho_{\text{dil}} \to 1$ means they nearly cancel out. The triangle inequality ensures $\rho_{\text{dil}} \geq 0$ unconditionally. Each outer product $\boldsymbol{\delta}_k \boldsymbol{\delta}_k^T$ is a rank-1 matrix oriented along the direction $\mathbf{v} - \mathbf{w}_k$. When cooperating prototypes occupy different regions of input space, the pairwise inner products

$$\langle \boldsymbol{\delta}_i, \boldsymbol{\delta}_j \rangle = (\mathbf{v} - \mathbf{w}_i)^T (\mathbf{v} - \mathbf{w}_j) \tag{47}$$

have no systematic relationship to the direction that would be informative for discrimination, giving rise to partial cancellation in the gradient sum and an elevated ρ_{dil}. The situation worsens as the prototype count grows. With P_c prototypes per class, $P_c - 1$ non-informative cooperation signals progressively overflow out of the winner's local gradient. In a learner setup of large P_c and large γ, we can bound the aggregate gradient; since $\bar{h}_k \leq 1$ for all k and $\sum_k \bar{h}_k = 1$,

$$\left\| \frac{\partial E}{\partial \Omega} \right\| \leq \sum_k \bar{h}_k \cdot |\alpha_k| \cdot \|\Omega\| \cdot \|\boldsymbol{\delta}_k\|^2, \tag{48}$$

The sum is dominated by the cooperators' large $|\boldsymbol{\delta}_k|^2$ terms, which encode inter-prototype displacement rather than boundary structure. Compare this to the non-cooperative case in GMLVQ ($\gamma \to 0$) where only the winner contributes

$$\frac{\partial E_{\text{GMLVQ}}}{\partial \Omega} = \sum_{\mathbf{v}} \alpha_{\text{win}} \cdot \Omega\,\boldsymbol{\delta}_{\text{win}} \boldsymbol{\delta}_{\text{win}}^T. \tag{49}$$

Every gradient term involves a small, locally relevant $\boldsymbol{\delta}_{\text{win}}$ and there are no conflicting contributions. The metric adapts to discriminate classes at the decision boundary guided solely by the nearest prototype's perspective. Diagonal relevance escapes this failure mode because in SRNG, the relevance gradient for feature j is $\partial d / \partial \lambda_j = (v_j - w_{k,j})^2 \geq 0$. The squared form guarantees non-negativity, so cooperation from distant prototypes yields a weaker but *aligned* signal. No cancellation is possible because the gradient contributions cannot take opposite signs. This explains the asymmetry between SRNG (where relevance adaptation is intact) and SMNG (where matrix adaptation degrades).

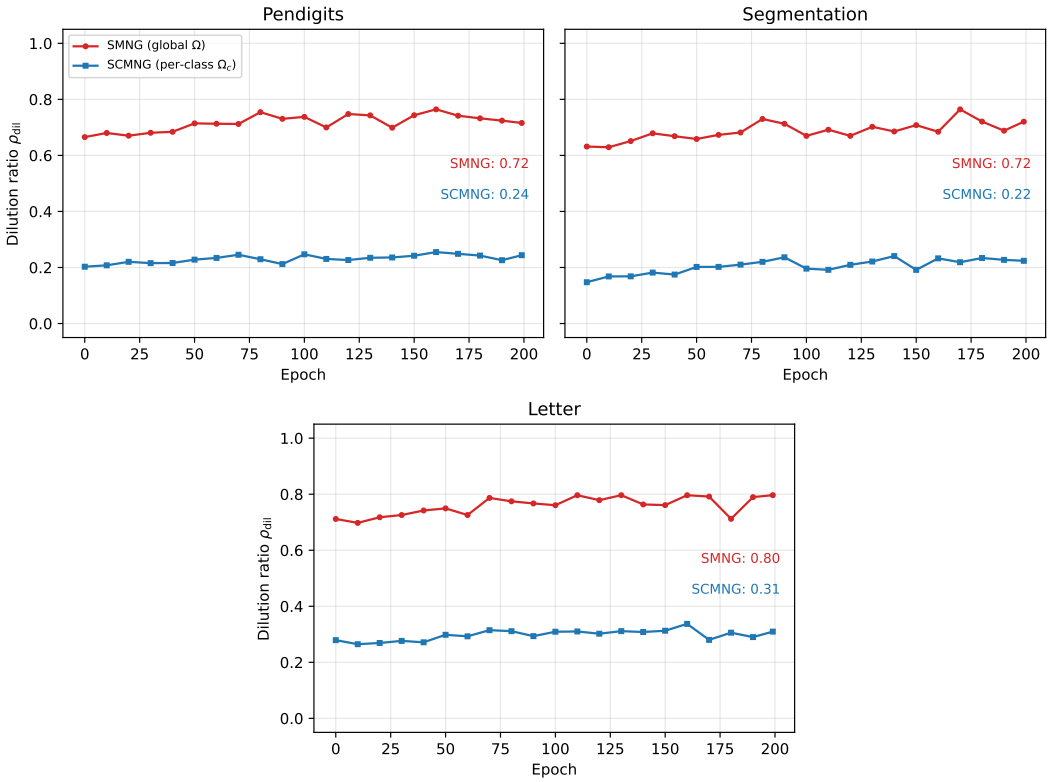

Figure 1: Empirical dilution ratio ρ_{dil} measured during training on all three datasets ($P_c = 5$, SGD, $\varepsilon = 0.01$). Across Pendigits, Segmentation and Letter, SMNG's global Ω yields $\rho_{\text{dil}} \approx 0.72$–$0.80$ (roughly 70–80% of the gradient signal cancels), whereas SCMNG's per-class Ω_c keeps dilution at $\rho_{\text{dil}} \approx 0.22$–$0.31$.

4.2 Subtraction Amplification in Tangent Distance

The second failure mode entails a subtraction amplification which arises in STNG from the subtractive structure of tangent distance. The Euclidean gradient of d_T with respect to Ω_k is $-2\,\boldsymbol{\delta}_k\boldsymbol{\delta}_k^T\Omega_k$ (cf. (31)). To preserve the orthonormality of Ω_k, we project onto the tangent space of the Stiefel manifold, obtaining the Riemannian gradient

$$\nabla_{\mathcal{S}} d_T = -2\,(I - \Omega_k\Omega_k^T)\,\boldsymbol{\delta}_k\boldsymbol{\delta}_k^T\Omega_k. \tag{50}$$

In both forms, the gradient pushes Ω_k to increase $\|\Omega_k^T\boldsymbol{\delta}_k\|^2$, which *declares the direction $\boldsymbol{\delta}_k$ invariant* near $\mathbf{w}_k$. The analysis that follows holds for either gradient form under NG cooperation. A sample $\mathbf{v}$ that lies close to the winner $\mathbf{w}_j$ but far from a cooperating prototype $\mathbf{w}_k$ generates a large displacement $\boldsymbol{\delta}_k = \mathbf{v} - \mathbf{w}_k$ pointing from $\mathbf{w}_k$ toward $\mathbf{w}_j$'s receptive field. The gradient drives Ω_k to absorb this inter-prototype direction into its invariance subspace. However, that direction may well be *discriminative* in the vicinity of $\mathbf{w}_k$ and declaring it invariant reduces the model's ability to separate classes locally.

Definition 2 (Subtraction Amplification)**.** *Let $\mathcal{S}_k = \text{span}(\Omega_k)$ denote prototype k's learned invariance subspace. Subtraction amplification occurs when cooperation gradients expand $\mathcal{S}_k$ to include directions that are discriminative near $\mathbf{w}_k$, causing $d_T(\mathbf{v}, \mathbf{w}_k)$ to decrease for wrong-class samples $\mathbf{v}$ in the Voronoi cell of $\mathbf{w}_k$ (samples with $c_{\mathbf{v}} \neq c_k$ for which maintaining large distance is necessary to preserve decision margins).*

The cause is an asymmetry between additive and subtractive distance formulations:

Proposition 1 (Additive vs. Subtractive Cooperation)**.** *Let $d_\theta(\mathbf{v}, \mathbf{w})$ be a parametric distance with learnable θ and let ϵ be the perturbation to θ from NG cooperation gradients.*

1. ***Additive case.*** *If d has the form $d(\mathbf{v}; \Omega) = \boldsymbol{\delta}^T\Omega^T\Omega\,\boldsymbol{\delta}$ and the perturbation $\Omega \to \Omega'$ satisfies $\text{span}(\Omega) \subseteq \text{span}(\Omega')$ with the existing rows of Ω preserved, then $d(\mathbf{v}; \Omega') \geq d(\mathbf{v}; \Omega)$ for all $\mathbf{v}$: the perturbation is regularising.*

2. ***Subtractive case.*** *If d has the form $d_T(\mathbf{v};\Omega_k) = \|\boldsymbol{\delta}\|^2 - \|\Omega_k^T\boldsymbol{\delta}\|^2$ and the perturbation expands the column space of Ω_k, then $d_T(\mathbf{v};\Omega_k') \leq d_T(\mathbf{v};\Omega_k)$: the perturbation is destructive.*

Proof. For matrix distance, the metric tensor $\Lambda = \Omega^T\Omega$ is positive semi-definite and any perturbation $\Omega \to \Omega'$ yields $\Lambda' = \Omega'^T\Omega' \succeq 0$. Writing $\Lambda' = \Lambda + \Delta\Lambda$, the distance change is

$$d_M(\mathbf{v};\Omega') - d_M(\mathbf{v};\Omega) = \boldsymbol{\delta}^T\Delta\Lambda\,\boldsymbol{\delta}. \tag{51}$$

If the cooperation perturbation increases the metric tensor ($\Delta\Lambda \succeq 0$, equivalently $\Lambda' \succeq \Lambda$), distances can only increase. A sufficient condition for $\Delta\Lambda \succeq 0$ is $\mathrm{span}(\Omega) \subseteq \mathrm{span}(\Omega')$ with the existing rows of Ω preserved, implying the perturbation adds measurement directions without removing or rescaling existing ones. (The converse does not hold: $\Lambda' \succeq \Lambda$ is possible even without span inclusion.) In practice, small perturbations approximately satisfy this condition. In this regard, cooperation noise in additive metrics is regularising, meaning it sharpens the margin. For tangent distance, let $P_k = \Omega_k\Omega_k^T$ denote the projection onto the invariance subspace. When the cooperation perturbation expands the column space ($\mathrm{span}(\Omega_k) \subseteq \mathrm{span}(\Omega_k')$), we can write $P_k' = P_k + P_k^{\mathrm{coop}}$ with $P_k^{\mathrm{coop}} \succeq 0$. Then

$$d_T(\mathbf{v};\Omega_k') = \|\boldsymbol{\delta}\|^2 - \boldsymbol{\delta}^T P_k'\boldsymbol{\delta} = d_T(\mathbf{v};\Omega_k) - \boldsymbol{\delta}^T P_k^{\mathrm{coop}}\boldsymbol{\delta} \leq d_T(\mathbf{v};\Omega_k), \tag{52}$$

since $P_k^{\mathrm{coop}} \succeq 0$ implies $\boldsymbol{\delta}^T P_k^{\mathrm{coop}}\boldsymbol{\delta} \geq 0$. This holds without qualification, as expanding the invariance subspace can only decrease tangent distances. Thus, cooperation noise in subtractive metrics is destructive since it erodes the margin. □

Note the asymmetry in strength between the two halves. The subtractive result holds unconditionally that any expansion of the invariance subspace decreases tangent distances. The additive result is conditional on the span-inclusion property, which holds when the perturbation adds measurement directions without removing existing ones. In practice, small perturbations approximately satisfy this condition, but sufficiently large cooperation updates could violate it. The matrix distance falls into the additive category, whilst the tangent distance falls into the subtractive one. This explains why cooperation noise regularises matrix metrics but erodes tangent metrics. A natural approach is to try the same procedure that resolved SMNG by sharing invariance subspaces at the class level, analogous to SCMNG. However, we note the tangent distance is inherently *local*. Invariance is defined relative to a specific point in input space and directions that are invariant near $\mathbf{w}_1$ may be discriminative near $\mathbf{w}_2$, even when both prototypes belong to the same class. Forcing all class-c prototypes to share one invariance subspace amounts to assuming that the class has a globally uniform invariance structure, which rarely holds in practice. Experimental evaluation confirms that class-wise tangent sharing performs worse than both STNG and GTLVQ.

A second possible remedy is to stop the gradient through Ω_k for all cooperating (non-winner) prototypes, so that only the winner's tangent subspace is updated while prototype positions still benefit from NG cooperation. This preserves the forward pass but restricts Ω-gradient flow to the winner. In practice, the winner's gradient is still determined by the cooperation-weighted loss and the normalized weight $\bar{h}_0 = 1/(1 + \sum_{r>0} e^{-r/\gamma})$ can be well below unity when γ is large. The net effect is to attenuate the winner's metric learning rate, causing the tangent subspace to lag behind prototype movement.

4.3 Metrics That Escape: SLNG and SCMNG

From the two established failure modes, we now proceed to the scope of metrics unaffected by them. In SLNG each prototype $\mathbf{w}_k$ owns its metric Ω_k and the gradient $\partial d_k/\partial\Omega_k = 2\,\Omega_k\,\boldsymbol{\delta}_k\boldsymbol{\delta}_k^T$ pushes Ω_k to *increase* its projection strength along $\boldsymbol{\delta}_k = \mathbf{v} - \mathbf{w}_k$. Under cooperation, distant samples therefore encourage Ω_k to measure more and not less. The matrix distance is additive, $\|\Omega_k\boldsymbol{\delta}\|^2 = \sum_i(\omega_i^T\boldsymbol{\delta})^2$ with ω_i the rows of Ω_k, therefore expanding the set of measurement directions can only increase distances. This is captured by the monotonicity property

$$\mathrm{span}(\Omega_k) \subseteq \mathrm{span}(\Omega_k') \implies d_{\Omega_k'}(\mathbf{v},\mathbf{w}_k) \geq d_{\Omega_k}(\mathbf{v},\mathbf{w}_k) \tag{53}$$

which holds for all $\mathbf{v} \in \mathbb{R}^d$. To see why, decompose Ω_k' into Ω_k and a complementary part $\Delta\Omega_k$ spanning the new directions. Since $\Delta\Omega_k$ lies in the orthogonal complement of $\mathrm{span}(\Omega_k)$, we have $\Omega_k^T\Delta\Omega_k = 0$ and therefore $(\Omega_k\boldsymbol{\delta}_k)^T(\Delta\Omega_k\boldsymbol{\delta}_k) = \boldsymbol{\delta}_k^T\Omega_k^T\Delta\Omega_k\boldsymbol{\delta}_k = 0$, so

$$d_{\Omega_k'}(\mathbf{v},\mathbf{w}_k) = \|\Omega_k\boldsymbol{\delta}_k\|^2 + \|\Delta\Omega_k\boldsymbol{\delta}_k\|^2 \geq d_{\Omega_k}(\mathbf{v},\mathbf{w}_k). \tag{54}$$

Of course, cooperation perturbations need not be exactly orthogonal to $\mathrm{span}(\Omega_k)$ in practice. However, the cooperation weights $\bar{h}_k \ll 1$ for $k > 0$ keep these perturbations small, so the net distance change stays close to the monotonic regime. Cooperation from distant samples thus introduces weak measurement axes that the winner's immediate data

distribution would not emphasize on its own. One can think of this as implicit regularization, quantifiable through the effective rank [22] of $\Lambda_k = \Omega_k^T \Omega_k$:

$$r_{\text{eff}}(\Lambda_k) = \frac{\left(\sum_i \sigma_i(\Lambda_k)\right)^2}{\sum_i \sigma_i(\Lambda_k)^2}, \tag{55}$$

where $\sigma_i(\Lambda_k)$ are the eigenvalues of Λ_k. Cooperation tends to raise r_{eff} by discouraging Ω_k from collapsing onto the narrow set of directions that dominate within its immediate Voronoi cell. The cooperation weights $\bar{h}_k$ guarantee that these secondary contributions stay attenuated relative to the winner's local signal, preserving the primacy of the principal discriminative directions. Whether this metric-level regularisation is able to outweigh the simultaneous perturbation of prototype positions is ultimately an empirical question, one whose answer depends on the quality of initialisation (Section 5). SLNG avoids dilution by making each Ω_k private, but at the cost of $K \cdot ld$ metric parameters. SCMNG occupies an intermediate position as it shares Ω_c within each class while keeping classes separate. The resulting gradient alignment can be checked directly. The gradient of E_{SCMNG} with respect to Ω_c sums over all prototypes of class c:

$$\frac{\partial E}{\partial \Omega_c} = \sum_{\mathbf{v}} \sum_{\mathbf{w}_r : c_r = c} \bar{h}_r \cdot \alpha_r \cdot 2\,\Omega_c\, \boldsymbol{\delta}_r \boldsymbol{\delta}_r^T. \tag{56}$$

Unlike SMNG, the sum is only carried out on prototypes of class c, all of which share the same objective by separating c from its competitors. Each gradient term takes the form $2\,\Omega_c\,\mathbf{G}_r$ with $\mathbf{G}_r = \boldsymbol{\delta}_r \boldsymbol{\delta}_r^T$, a rank-1 positive semi-definite matrix since for any $\mathbf{x} \in \mathbb{R}^d$:

$$\mathbf{x}^T(\boldsymbol{\delta}_r \boldsymbol{\delta}_r^T)\mathbf{x} = (\boldsymbol{\delta}_r^T \mathbf{x})^2 \geq 0. \tag{57}$$

Moreover, the pairwise alignment between any two within-class direction matrices is non-negative:

$$\langle \mathbf{G}_i, \mathbf{G}_j \rangle_F = \operatorname{tr}(\mathbf{G}_i \mathbf{G}_j) = (\boldsymbol{\delta}_i^T \boldsymbol{\delta}_j)^2 \geq 0, \tag{58}$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product. The shared left factor Ω_c acts as a common multiplier and the non-negative alignment of the direction matrices $\mathbf{G}_r$ guarantees that within-class gradient contributions reinforce one another rather than cancelling, the opposite of what happens in SMNG. The cross-class conflict is removed entirely because prototypes of different classes update *distinct* matrices (Ω_a and Ω_b). Formally, the gradient decomposition is orthogonal across classes:

$$\frac{\partial E}{\partial \Omega_a} \perp \frac{\partial E}{\partial \Omega_b} \quad \text{for } a \neq b, \tag{59}$$

since Ω_a and Ω_b live in disjoint parameter spaces. Within class c, prototypes stationed in different parts of the input space emphasize different boundary directions. The prototype $\mathbf{w}_i$ near the c/a boundary pushes Ω_c to separate c from a, while the prototype $\mathbf{w}_j$ near the c/b boundary pushes it to separate c from b. Far from conflicting, these signals are *complementary* and Ω_c converges to a metric that separates c from *all* its competitors, which is a more complete representation than any single prototype's viewpoint. Appendix B extends the pairwise antisymmetry argument to SMNG, SLNG and STNG.

4.4 Cooperation-Compatibility Conditions

The analyses of Sections 4.1–4.3 distil into three structural conditions for combining NG cooperation with metric adaptation. We note that these conditions were formulated on the basis of the gradient-level analysis and then checked against the experimental results; they are explanatory rather than independently predictive.

Definition 3 (Cooperation-Compatible Metric)**.** *A parametric distance $d_\theta(\mathbf{v}, \mathbf{w})$ is* cooperation-compatible *if for any non-local sample $\mathbf{v}$ ($\|\mathbf{v} - \mathbf{w}_k\| > \|\mathbf{v} - \mathbf{w}_j\|$ for some $j \neq k$ with $c_j = c_k$), the induced parameter perturbation $\Delta\theta_k \propto -\bar{h}_k \nabla_{\theta_k} E$ satisfies*

$$d_{\theta_k + \eta\Delta\theta_k}(\mathbf{v}', \mathbf{w}_k) \geq d_{\theta_k}(\mathbf{v}', \mathbf{w}_k) \tag{60}$$

for all $\mathbf{v}'$ in the Voronoi cell of $\mathbf{w}_k$ (computed under the current parameters θ_k) with $c_{\mathbf{v}'} = c_k$ and sufficiently small step size $\eta > 0$. That is, the perturbation does not decrease same-class distances in the local neighbourhood of $\mathbf{w}_k$.

The first condition, *additivity*, requires that the distance function be additive in its learnable parameters, which means expanding the parameter space can only increase distances. Additive metrics turn cooperation noise into regularization (adding measurement capacity) and never into destruction. Tangent distance is subtractive and therefore cooperation-incompatible. Secondly, *gradient alignment* requires that whenever a metric parameter is shared across prototypes, the sharing boundary must ensure that all contributing prototypes have aligned gradient objectives. Per-class sharing satisfies this condition because all class-c prototypes pursue the same goal. Global sharing violates it because prototypes of different classes need to be separated by different boundaries. Thirdly, the *locality preservation* condition requires

that if the metric's expressiveness rests on locality (invariance defined relative to a specific point in input space), the metric must not be shared across prototypes at any granularity. Tangent distance is intrinsically local and sharing it destroys the very property that gives it power. A critical question worthy of concern is whether NG cooperation can reduce the model's expressiveness relative to the baseline.

Remark 1 (Baseline Recovery)**.** *As the neighbourhood range $\gamma \to 0$, the neighbourhood function satisfies $h_\gamma(k) = e^{-k/\gamma} \to \mathbf{1}[k = 0]$ and the normalised weight concentrates entirely on the rank-0 prototype: $\bar{h}_0 \to 1$, $\bar{h}_{k>0} \to 0$. Concretely, for $k \geq 1$:*

$$\bar{h}_k = \frac{e^{-k/\gamma}}{\sum_{r=0}^{P_c-1} e^{-r/\gamma}} \leq e^{-k/\gamma} \xrightarrow{\gamma \to 0} 0. \tag{61}$$

Only the nearest same-class prototype contributes, reducing (6) *to the standard GLVQ cost with the respective distance function. Hence every NG variant recovers its non-cooperative baseline as a special case: SRNG → GLVQ with relevances, SMNG → GMLVQ, SLNG → LGMLVQ, STNG → GTLVQ, SCMNG → CW-GMLVQ. In particular, $\inf_{\gamma>0} E^*_{\mathrm{NG}}(\gamma) \leq E^*_{\mathrm{base}}$ where $E^*_{\mathrm{NG}}(\gamma)$ and E^*_{base} denote the respective optimal costs, so cooperation cannot reduce expressiveness in principle.*

Table 1: Cooperation compatibility of distance functions under the NG framework. The three conditions classify which metrics are structurally exposed to cooperation-induced failure modes.

Distance	Formula	NG effect on metric	Compatible?
Squared Euclidean	$\|\boldsymbol{\delta}\|^2$	No learnable metric	Yes (trivially)
Diagonal relevance	$\sum_j \lambda_j \delta_j^2$	"Weight this feature"	Yes ($\lambda_j \geq 0$)
Matrix (global)	$\|\Omega\boldsymbol{\delta}\|^2$	"Measure this direction"	No (cross-class conflict)
Matrix (per-class)	$\|\Omega_c\boldsymbol{\delta}\|^2$	"Measure this direction"	Yes (aligned)
Matrix (per-proto)	$\|\Omega_k\boldsymbol{\delta}\|^2$	"Measure this direction"	Yes (additive)
Tangent (per-proto)	$\|\boldsymbol{\delta}\|^2 - \|\Omega_k^T\boldsymbol{\delta}\|^2$	"Ignore this direction"	No (subtractive)

The practical question is whether $\gamma > 0$ actually helps in practice. The remark guarantees that cooperation cannot reduce expressiveness in principle, but compatibility (Table 1) speaks only to whether the metric parameters θ are adversely perturbed by cooperation gradients. The overall effect on classification correctness also depends on what cooperation does to the prototype positions themselves. Under class-conditional initialization, where each prototype is drawn from its class distribution and therefore starts in a reasonable part of the input space, the topological organising benefit of cooperation is largely redundant. What remains is a perturbation cost; rank-weighted updates from distant training samples displace prototypes from positions that the initialization had placed correctly. For incompatible metrics (SMNG, STNG), both the metric parameters and the prototypes suffer, resulting in severe degradation. For compatible metrics (SLNG, SCMNG), the metric adaptation mechanism stays intact but prototype positioning is still perturbed, yielding milder but consistently negative effects. The experiments in Section 5 exemplify this graduated prediction.

4.5 Interpretability

Beyond classification correctness, the per-class parameterization of SCMNG offers an interpretability advantage over both GMLVQ and LGMLVQ. After training, the model delivers C learned matrices $\Omega_c \in \mathbb{R}^{l \times d}$ from which the class-specific metric tensor

$$\Lambda_c = \Omega_c^T \Omega_c \in \mathbb{R}^{d \times d} \tag{62}$$

can be read off directly. The diagonal elements $(\Lambda_c)_{jj}$ quantify how important feature j is for classifying class c; off-diagonal elements $(\Lambda_c)_{ij}$ reveal class-specific feature correlations. Since the granularity is the class itself, no aggregation across prototypes is needed. Table 2 summarises the interpretive properties of the models considered in this paper. LGMLVQ provides finer-grained per-prototype relevance, but extracting a class-level summary requires aggregation across all prototypes of that class. GMLVQ's single global relevance matrix cannot express class-specific differences at all. SCMNG sits between the two, where each class-level metric tells the practitioner which features matter for each class, without post-hoc processing.

5 Experiments

We evaluate all models on three UCI benchmark datasets [1], chosen to cover a range of class cardinalities and feature dimensionalities (Table 3). The feature set is standardized to zero mean and unit variance using training-fold statistics

Table 2: Interpretability comparison across metric LVQ models.

Model	Learned metric	Interpretation
GLVQ	None (Euclidean)	No metric adaptation
GMLVQ	$\Lambda = \Omega^T\Omega$	"Feature j matters overall"
LGMLVQ	$\Lambda_k = \Omega_k^T\Omega_k$	"Feature j matters near $\mathbf{w}_k$"
SCMNG	$\Lambda_c = \Omega_c^T\Omega_c$	"Feature j matters for class c"

only. Twelve models are compared, entailing six non-cooperative baselines and their six NG counterparts. Each NG variant is paired with the baseline that shares its distance function. CW-GMLVQ deserves explicit mention. It is not a novel model but an ablation baseline where LGMLVQ with per-class parameter tying ($\Omega_k = \Omega_c$ for all k in class c) or equivalently, the non-cooperative limit of SCMNG obtained by fixing $\gamma_{\text{init}} = \gamma_{\text{final}} = 0.01$. At this value, the ratio $\bar{h}_1/\bar{h}_0 = e^{-1/0.01} \approx 10^{-43}$, so only the nearest same-class prototype receives a meaningful neighborhood weight at all times. Per-class metric parameterization occupies a known position on the global-to-local spectrum [4]. The

Table 3: Datasets used in the experiments.

Dataset	Samples	Features	Classes
Pendigits	10 992	16	10
Segmentation	2 310	19	7
Letter	20 000	16	26
Synthetic	1 600	16	8

primary comparison uses $P_c = 5$ prototypes per class. To study how prototype density interacts with cooperation, we additionally vary $P_c \in \{3, 5, 10, 15\}$ across the three pragmatic datasets (Pendigits, Segmentation and Letter) in Table 3. Learning is executed by a stochastic gradient descent (SGD) optimization with a learning rate $\varepsilon = 0.01$ for 200 epochs. Pendigits and Segmentation are trained with online updates (batch size 1). Letter uses mini-batches of 32 to keep runtime manageable given its larger sample count, though all models within each dataset use the same batch size to guarantee a fair comparison. For the NG models, γ is initialised to $P_c/2$ and annealed exponentially to 0.01, with the sigmoid steepness set to $\beta = 10$. For all matrix-based models used in this experiment, the projection matrix Ω is full-rank ($l = d$, where d is the feature dimensionality), initialised as the identity. STNG uses subspace dimension $s = 2$. Prototypes are initialised by class-conditional selection from the training data (random seed 42 for the stratified cross-validation (CV) splits). No regularization or weight decay is applied. Performance is measured by 10-fold stratified CV; we report mean classification correctness $\pm$ standard deviation. Following Hammer et al. [13], we also examine the effect of the learning rate ratio $\rho = \varepsilon^-/\varepsilon^+$, evaluating each NG variant at $\rho \in \{0.5, 1.0\}$ and reporting the better configuration. SMNG and SCMNG consistently favour $\rho = 1.0$, with particularly large effects on Letter ($+20.4$ pp for SMNG, $+6.2$ pp for SCMNG). SLNG is dataset-dependent, preferring $\rho = 1.0$ on Pendigits and Segmentation but $\rho = 0.5$ on Letter. For STNG, $\rho = 0.5$ offers marginal improvement. In the main results, we always report the best ρ for each model-dataset pair. This selection favours NG variants, which receive two configurations per dataset, while baselines have no analogous free parameter.

5.1 Prototype Density Ablation

The effect of prototype density is examined in Tables 9, 10 and 11, which report full accuracy results for $P_c \in \{3, 10, 15\}$ (the $P_c = 5$ results are in Table 4). Table 7 collects the accuracy differences Δ between each NG variant (best ρ) and its baseline across all four prototype densities. Table 8 zeroes in on the cooperation effect by comparing SCMNG against CW-GMLVQ, its non-cooperative limit. Every NG variant underperforms its non-cooperative baseline on all three UCI datasets (Table 5). This holds across all learning rate ratios, all prototype densities and all metric parameterizations tested. It is not an isolated failure of one model or one dataset; rather, it is an observed systematic pattern. SMNG suffers the worst, trailing GMLVQ by -8.3 pp on Pendigits, -4.2 pp on Segmentation and -10.9 pp on Letter (all at best $\rho = 1.0$). The gradient dilution analysis delivers a plausible mechanism where cooperating same-class prototypes feed conflicting rank-1 contributions into the shared Ω and the interference worsens as the prototype count increases. Direct measurement across all three datasets ($P_c = 5$) confirms this behavior. Figure 1 shows SMNG's dilution ratio at $\rho_{\text{dil}} \approx 0.72$ on Pendigits, ≈ 0.72 on Segmentation and ≈ 0.80 on Letter, meaning 70–80% of the gradient signal cancels throughout training. SCMNG's per-class parameterization keeps dilution at $\rho_{\text{dil}} \approx 0.22$–$0.31$ across all datasets, confirming that the mechanism is not dataset-specific.

Table 4: Test accuracy (%) with $P_c = 5$. Mean $\pm$ std over 10-fold stratified CV. Bold indicates the best result per metric group (baseline or NG variant). Gray shading marks the CW-GMLVQ ablation baseline throughout.

Model	Pendigits	Segmentation	Letter
GLVQ	94.82 ± 0.80	88.14 ± 2.81	68.62 ± 1.37
SNG ($\rho=0.5$)	82.32 ± 1.13	85.41 ± 2.42	59.82 ± 2.11
SNG ($\rho=1.0$)	81.93 ± 0.94	86.23 ± 2.01	61.74 ± 2.19
GRLVQ	92.94 ± 0.67	83.72 ± 3.29	59.54 ± 1.97
SRNG ($\rho=0.5$)	85.45 ± 1.20	80.17 ± 3.83	52.33 ± 1.86
SRNG ($\rho=1.0$)	83.52 ± 1.40	80.91 ± 3.35	56.56 ± 2.37
GMLVQ	95.52 ± 0.76	95.19 ± 1.55	80.39 ± 1.26
SMNG ($\rho=0.5$)	80.26 ± 2.19	80.91 ± 2.84	49.07 ± 7.11
SMNG ($\rho=1.0$)	87.23 ± 1.35	90.95 ± 1.75	69.45 ± 1.50
CW-GMLVQ	$\mathbf{98.89 \pm 0.34}$	$\mathbf{96.41 \pm 1.13}$	$\mathbf{91.05 \pm 0.71}$
SCMNG ($\rho=0.5$)	95.78 ± 0.92	93.94 ± 2.21	82.07 ± 1.30
SCMNG ($\rho=1.0$)	98.63 ± 0.37	95.63 ± 1.61	88.30 ± 0.79
LGMLVQ	$\mathbf{99.38 \pm 0.26}$	$\mathbf{96.49 \pm 1.54}$	$\mathbf{93.18 \pm 0.71}$
SLNG ($\rho=0.5$)	96.66 ± 0.58	93.94 ± 1.71	87.70 ± 0.99
SLNG ($\rho=1.0$)	98.58 ± 0.41	95.24 ± 1.63	85.23 ± 1.05
GTLVQ	$\mathbf{98.24 \pm 0.38}$	$\mathbf{92.68 \pm 1.83}$	$\mathbf{82.61 \pm 1.27}$
STNG ($\rho=0.5$)	94.69 ± 0.55	87.40 ± 2.02	75.23 ± 0.95
STNG ($\rho=1.0$)	93.54 ± 0.67	87.19 ± 2.13	74.65 ± 0.89

Table 5: Accuracy difference Δ in percentage points (pp) at $P_c = 5$, using the best ρ per model and dataset. Negative values indicate that NG cooperation degraded performance relative to the non-cooperative baseline.

Pair	Metric	Pendigits	Segm.	Letter
SNG $-$ GLVQ	Euclidean	-12.50	-1.90	-6.87
SRNG $-$ GRLVQ	Diagonal relevance	-7.49	-2.81	-2.98
SMNG $-$ GMLVQ	Global matrix	-8.30	-4.24	-10.93
SCMNG $-$ CW-GMLVQ	Class-wise matrix	-0.26	-0.78	-2.74
SLNG $-$ LGMLVQ	Per-proto matrix	-0.80	-1.26	-5.48
STNG $-$ GTLVQ	Per-proto tangent	-3.55	-5.28	-7.38
CW-GMLVQ $-$ GMLVQ	Per-class vs global	$+3.37$	$+1.21$	$+10.66$
SCMNG $-$ GMLVQ	Per-class + coop	$+3.10$	$+0.43$	$+7.92$

Indirect evidence from the P_c ablation (Table 7) is consistent. The SMNG learner deficit on Letter widens from -7.3 pp at $P_c = 3$ to -13.5 pp at $P_c = 15$, as more cooperating prototypes introduce more conflicting gradients. STNG falls below GTLVQ by -3.6 pp, -5.3 pp and -7.4 pp across the three datasets (at best $\rho = 0.5$). Although tangent distance is parameterized per-prototype, which eliminates sharing conflicts, its subtractive formulation exposes it to a different threat such that cooperation expands the invariance subspace and thereby erodes discriminative capacity (Proposition 1). The deficit again scales with P_c, reaching -10.5 pp on Letter at $P_c = 15$. The observed SCMNG results call for a critical analysis. The SCMNG model improves over GMLVQ by $+3.1$ pp on Pendigits, $+0.4$ pp on Segmentation and $+7.9$ pp on Letter at $P_c = 5$. However, this comparison conflates two known factors: the per-class metric parameterization and NG cooperation. The CW-GMLVQ ablation separates them cleanly. CW-GMLVQ shares SCMNG's per-class metric structure but eliminates cooperation ($\gamma \to 0$) and it exceeds SCMNG on every dataset and at every prototype density (Table 8). The gap widens with P_c; on Letter, it grows from -1.7 pp at $P_c = 3$ to -6.1 pp at $P_c = 15$, consistent with gradient interference scaling with the number of cooperating prototypes. The entire accuracy advantage over GMLVQ therefore comes from the per-class metric structure, not from cooperation. On Letter at $P_c = 3$, CW-GMLVQ gains $+13.1$ pp over GMLVQ while SCMNG gains $+11.4$ pp: the non-cooperative model outperforms the cooperative one. SLNG underperforms LGMLVQ on all datasets with -0.8 pp on Pendigits, -1.3 pp on Segmentation and -5.5 pp on Letter at $P_c = 5$. On theoretical grounds (Section 4.3) we had expected per-prototype additive metrics to benefit from cooperation functioning as implicit regularization. The monotonicity property (53) does ensure that cooperation noise cannot corrupt the metric parameters, yet the overall picture is negative because cooperation simultaneously perturbs the prototype positions. Under class-conditional initialization, prototypes already sit within their class distributions and the topological organizing benefit that cooperation is designed to provide is simply redundant. What remains is a pure perturbation effect, pushing prototypes away from positions the initialization

Table 6: Effect of the learning rate ratio ρ on NG model accuracy (pp). Difference = $\rho = 0.5$ minus $\rho = 1.0$; negative values indicate $\rho = 1.0$ is superior.

Model	Pendigits	Segm.	Letter
SNG	$+0.39$	-0.82	-1.92
SRNG	$+1.93$	-0.74	-4.23
SMNG	-6.97	-10.04	-20.38
SCMNG	-2.85	-1.69	-6.24
SLNG	-1.92	-1.30	$+2.47$
STNG	$+1.15$	$+0.22$	$+0.58$

Table 7: Δ = NG variant $-$ baseline (pp, best ρ) across prototype densities.

Pair / Dataset		$P_c = 3$	$P_c = 5$	$P_c = 10$	$P_c = 15$
SCMNG $-$ GMLVQ	Pendigits	$+4.33$	$+3.10$	$+1.91$	$+1.54$
	Segmentation	$+0.95$	$+0.43$	-0.52	-0.30
	Letter	$+11.39$	$+7.92$	$+2.54$	-0.24
SMNG $-$ GMLVQ	Pendigits	-7.48	-8.30	-9.30	-9.64
	Segmentation	-4.07	-4.24	-4.59	-3.55
	Letter	-7.31	-10.93	-13.78	-13.48
SLNG $-$ LGMLVQ	Pendigits	-0.70	-0.80	-0.96	-0.87
	Segmentation	-1.04	-1.26	-3.33	-3.20
	Letter	-2.65	-5.48	-9.11	-10.09
STNG $-$ GTLVQ	Pendigits	-2.91	-3.55	-3.45	-3.02
	Segmentation	-3.03	-5.28	-6.32	-7.32
	Letter	-5.21	-7.38	-9.69	-10.54
SNG $-$ GLVQ	Pendigits	-11.03	-12.50	-11.68	-11.39
	Segmentation	-0.35	-1.90	-5.41	-5.37
	Letter	-5.56	-6.87	-7.26	-6.29
SRNG $-$ GRLVQ	Pendigits	-7.31	-7.49	-6.69	-5.06
	Segmentation	-2.12	-2.81	-1.99	-1.17
	Letter	-1.54	-2.98	-2.58	-2.42

had placed well. The deficit grows markedly with P_c, reaching -10.1 pp on Letter at $P_c = 15$. The SNG results confirm that cooperation hurts even without any metric adaptation. SNG loses -12.5 pp, -1.9 pp and -6.9 pp against GLVQ on the three datasets. When prototypes are drawn from the training data, the topological organization that cooperation provides is superfluous. The rank-weighted updates from distant training samples simply displace prototypes from a configuration that was already serviceable. SRNG falls below GRLVQ by -7.5 pp on Pendigits, -2.8 pp on Segmentation and -3.0 pp on Letter. Diagonal relevance is classified as cooperation-compatible in Table 1: the non-negative gradient $\partial d/\partial \lambda_j = (v_j - w_{k,j})^2 \geq 0$ assures that metric updates continue aligned. Yet the overall model still degrades because cooperation perturbs prototype positions. Considering these results, we refine the design principles of Section 4.4. Cooperation-compatibility broadly tracks *how badly* cooperation degrades performance. Incompatible metrics (SMNG, STNG) suffer the largest losses on the UCI datasets (up to 13.8 pp), while compatible metrics (SCMNG, SLNG) sustain milder deficits. No metric tested benefits from cooperation under class-conditional initialization. Compatibility is necessary to avoid metric-level failure modes; it is not sufficient for cooperation to help overall.

NG cooperation was introduced into supervised learning by Hammer, Strickert and Villmann [13] to address the sensitivity of Generalized Learning Vector Quantization (GLVQ) to prototype initialization. Rank-based updates distribute prototypes across the data manifold irrespective of their starting positions, rescuing units that would otherwise remain stranded in uninformative regions. This topological organisation is the main contribution of cooperation and under random or arbitrary initialization, it is indispensable. Class-conditional initialization, which is a standard practice in modern LVQ implementations, already places each prototype within its class distribution. The topological benefit that cooperation would provide is therefore redundant. What remains is the perturbation cost, which implies rank-weighted updates from distant training samples displace prototypes from positions that the initialization had placed correctly. For cooperation to justify itself, its organisational benefit must exceed this perturbation cost. That condition is met when initialization is poor; it is not met when prototypes are drawn from the training data.

Table 8: Δ = SCMNG $-$ CW-GMLVQ (pp, best ρ): effect of cooperation alone, controlling for per-class metric parameterization. All values are negative, indicating that cooperation degrades performance at every prototype density tested.

Dataset	$P_c=3$	$P_c=5$	$P_c=10$	$P_c=15$
Pendigits	-0.21	-0.26	-0.41	-0.39
Segmentation	-0.04	-0.78	-0.95	-0.74
Letter	-1.70	-2.74	-4.83	-6.10

Table 9: Test accuracy (%) with $P_c = 3$. Mean $\pm$ std over 10-fold stratified CV.

Model	Pendigits	Segmentation	Letter
GLVQ	92.84 ± 0.72	86.54 ± 2.44	64.93 ± 2.23
SNG ($\rho=0.5$)	81.81 ± 1.24	85.45 ± 2.42	57.39 ± 2.33
SNG ($\rho=1.0$)	81.60 ± 1.01	86.19 ± 2.03	59.37 ± 2.85
GRLVQ	90.03 ± 0.84	81.30 ± 4.07	51.37 ± 1.57
SRNG ($\rho=0.5$)	82.72 ± 2.15	77.49 ± 4.19	45.57 ± 1.66
SRNG ($\rho=1.0$)	81.44 ± 1.23	79.18 ± 3.84	49.83 ± 1.64
GMLVQ	94.28 ± 0.88	94.98 ± 1.79	76.80 ± 1.47
SMNG ($\rho=0.5$)	79.87 ± 3.34	79.05 ± 3.00	44.58 ± 8.29
SMNG ($\rho=1.0$)	86.80 ± 1.47	90.91 ± 1.54	69.50 ± 1.08
CW-GMLVQ	98.82 ± 0.34	95.97 ± 1.01	89.89 ± 0.57
SCMNG ($\rho=0.5$)	95.62 ± 0.71	92.86 ± 2.49	81.25 ± 1.39
SCMNG ($\rho=1.0$)	98.61 ± 0.38	95.93 ± 1.40	88.19 ± 0.78
LGMLVQ	99.32 ± 0.33	96.62 ± 1.45	91.97 ± 0.55
SLNG ($\rho=0.5$)	96.32 ± 0.58	94.03 ± 1.86	89.32 ± 0.87
SLNG ($\rho=1.0$)	98.62 ± 0.44	95.58 ± 1.59	86.21 ± 1.11
GTLVQ	97.44 ± 0.48	91.52 ± 2.62	79.98 ± 1.08
STNG ($\rho=0.5$)	94.53 ± 0.79	88.48 ± 1.87	74.40 ± 1.35
STNG ($\rho=1.0$)	93.49 ± 0.82	87.27 ± 2.35	74.77 ± 1.51

Turning to the learning rate ratio, Table 6 reveals a pattern that tracks the metric structure. SMNG and SCMNG strongly prefer $\rho = 1.0$ (same learning rates), with the largest effects on Letter, where it can be observed $+20.4$ pp for SMNG and $+6.2$ pp for SCMNG. For models with shared matrix parameterization (global or class-wise), equal gradient flow through correct-class and wrong-class distances gives the metric a fuller picture of the decision boundary. Attenuating the wrong-class gradient ($\rho = 0.5$) starves the matrix of repulsive information that it needs for discriminative projections. SLNG shows a dataset-dependent preference for $\rho = 1.0$ on Pendigits (-1.9 pp) and Segmentation (-1.3 pp), but $\rho = 0.5$ on Letter ($+2.5$ pp), suggesting that the per-prototype metric's demand for balanced gradients interacts with class cardinality. For STNG, $\rho = 0.5$ is marginally better across all datasets. Its subtractive structure makes cooperation noise in the wrong-class gradient destructive (Proposition 1) and attenuating it via $\rho = 0.5$ partially lessens the damage. The fact that the optimal ρ varies with the metric reinforces the conclusion that the interaction between cooperation and metric learning depends on the metric's structural properties. The prototype density ablation (Table 7) indicates a pronounced scaling effect. For cooperation-incompatible metrics the degradation *intensifies* with increasing P_c. SMNG's deficit on Letter grows from -7.3 pp ($P_c = 3$) to -13.5 pp ($P_c = 15$) and STNG's grows from -5.2 pp to -10.5 pp. More cooperating prototypes mean increased conflicting gradient contributions and hence severe metric corruption. The CW-GMLVQ ablation (Table 8) shows the same scaling in a controlled setting: on Letter, SCMNG trails CW-GMLVQ by -1.7 pp at $P_c = 3$ but by -6.1 pp at $P_c = 15$. Even for a cooperation-compatible metric, the net effect of cooperation becomes more negative as the prototype count rises. Per-class metric parameterization as embodied by the CW-GMLVQ learner captures the accuracy enhancements previously attributed to SCMNG without incurring the cost of NG cooperation. CW-GMLVQ reaches 91.1% on Letter at $P_c = 5$, compared to 80.4% for GMLVQ ($+10.7$ pp) and 88.3% for SCMNG. With 26 classes, each Ω_c specializes for its class boundary, whereas GMLVQ's single global Ω must compromise across all classes. It is this per-class structure that drives the improvement and not neighborhood cooperation as one would have expected. Consider a comparison between CW-GMLVQ and LGMLVQ in this test scenario; LGMLVQ learns a fully independent Ω_k for each prototype. LGMLVQ achieves the highest accuracy on all three datasets (99.38% Pendigits, 93.18% Letter at $P_c = 5$), exceeding CW-GMLVQ (98.89%, 91.05%). This is expected as per-prototype parameterization is strictly more expressive than per-class tying and the additional parameters enable local adaptation that class-level matrices cannot capture. CW-GMLVQ trades this accuracy for two pragmatic advantages: the number of parameters that scales with the number of classes rather than prototypes

Table 10: Test accuracy (%) with $P_c = 10$. Mean $\pm$ std over 10-fold stratified CV.

Model	Pendigits	Segmentation	Letter
GLVQ	96.32 ± 0.60	91.08 ± 1.88	73.95 ± 0.63
SNG ($\rho=0.5$)	84.63 ± 0.96	85.41 ± 2.52	65.15 ± 1.41
SNG ($\rho=1.0$)	82.93 ± 0.81	85.67 ± 2.40	66.69 ± 1.64
GRLVQ	95.35 ± 0.87	87.71 ± 2.26	67.24 ± 1.14
SRNG ($\rho=0.5$)	88.66 ± 1.40	85.71 ± 3.04	60.86 ± 1.18
SRNG ($\rho=1.0$)	87.22 ± 1.28	83.90 ± 2.82	64.67 ± 1.49
GMLVQ	96.64 ± 0.78	96.15 ± 1.65	85.18 ± 1.11
SMNG ($\rho=0.5$)	79.86 ± 2.56	80.69 ± 3.92	55.70 ± 6.78
SMNG ($\rho=1.0$)	87.35 ± 1.17	91.56 ± 2.02	71.39 ± 0.92
CW-GMLVQ	98.96 ± 0.33	96.58 ± 1.39	92.55 ± 0.59
SCMNG ($\rho=0.5$)	95.85 ± 0.90	93.77 ± 2.04	82.75 ± 0.91
SCMNG ($\rho=1.0$)	98.55 ± 0.45	95.63 ± 1.60	87.72 ± 0.92
LGMLVQ	99.53 ± 0.23	97.71 ± 1.03	94.16 ± 0.53
SLNG ($\rho=0.5$)	97.12 ± 0.46	94.37 ± 1.47	85.05 ± 1.11
SLNG ($\rho=1.0$)	98.57 ± 0.47	94.37 ± 2.05	83.79 ± 1.22
GTLVQ	98.67 ± 0.30	94.63 ± 2.06	85.29 ± 0.63
STNG ($\rho=0.5$)	95.22 ± 0.78	88.31 ± 2.48	75.60 ± 0.93
STNG ($\rho=1.0$)	94.25 ± 0.84	87.92 ± 1.62	75.09 ± 1.19

Table 11: Test accuracy (%) with $P_c = 15$. Mean $\pm$ std over 10-fold stratified CV.

Model	Pendigits	Segmentation	Letter
GLVQ	96.82 ± 0.60	91.95 ± 1.76	77.94 ± 1.19
SNG ($\rho=0.5$)	85.43 ± 0.95	86.10 ± 2.08	70.62 ± 0.95
SNG ($\rho=1.0$)	85.13 ± 0.98	86.58 ± 2.23	71.65 ± 1.10
GRLVQ	95.79 ± 0.63	88.05 ± 2.10	73.18 ± 1.01
SRNG ($\rho=0.5$)	90.73 ± 1.68	86.88 ± 3.77	67.29 ± 1.00
SRNG ($\rho=1.0$)	88.69 ± 0.96	86.75 ± 2.35	70.76 ± 1.20
GMLVQ	97.13 ± 0.67	95.93 ± 1.89	87.71 ± 1.07
SMNG ($\rho=0.5$)	79.63 ± 2.46	83.03 ± 2.86	58.72 ± 8.13
SMNG ($\rho=1.0$)	87.49 ± 1.21	92.38 ± 1.99	74.23 ± 0.97
CW-GMLVQ	99.06 ± 0.33	96.36 ± 1.41	93.57 ± 0.49
SCMNG ($\rho=0.5$)	96.21 ± 0.66	93.90 ± 2.14	82.76 ± 1.07
SCMNG ($\rho=1.0$)	98.67 ± 0.26	95.63 ± 1.57	87.47 ± 0.85
LGMLVQ	99.51 ± 0.22	97.58 ± 1.16	94.14 ± 0.50
SLNG ($\rho=0.5$)	97.48 ± 0.42	94.37 ± 1.51	84.04 ± 1.04
SLNG ($\rho=1.0$)	98.64 ± 0.37	94.07 ± 1.91	83.70 ± 1.21
GTLVQ	98.76 ± 0.34	95.93 ± 1.61	86.84 ± 1.03
STNG ($\rho=0.5$)	95.74 ± 0.82	88.61 ± 2.16	76.13 ± 1.17
STNG ($\rho=1.0$)	94.79 ± 0.76	87.92 ± 1.90	76.30 ± 1.34

($C \cdot ld$ vs $K \cdot ld$) and direct class-level interpretability, since each Λ_c summarises the discriminative structure for class c without requiring aggregation across prototype-specific matrices. The per-class level thus occupies a Pareto point on the expressiveness-interpretability curve rather than claiming to be the most accurate parameterization. NG models also incur increased training costs compared to their baselines because every sample updates all same-class prototypes, not just the winner prototype. CW-GMLVQ avoids this overhead entirely as it retains the per-class metric of SCMNG but trains at baseline speed. For SCMNG with C classes, each sample triggers C independent matrix computations, compared to one for GMLVQ and $K = C \cdot P_c$ for LGMLVQ. The per-epoch cost is $\mathcal{O}(N \cdot K \cdot ld)$ for all NG variants, which is asymptotically the same as the baselines. However, the constant factor is larger because cooperation weights $\bar{h}_r$ for $r > 0$ produce nonzero gradient contributions that baselines avoid. Given that CW-GMLVQ matches or exceeds SCMNG in accuracy while training at baseline speed, the computational comparison reinforces the case for the non-cooperative model.

Table 12: Random vs. class-conditional initialization on Pendigits ($P_c = 5$). Accuracy (%) over 10-fold CV. Δ is the change from class-conditional to random init (best ρ per model).

Model	Class-cond	Random	Δ
GLVQ	94.82 ± 0.80	89.33 ± 0.83	-5.5
GMLVQ	95.52 ± 0.76	92.97 ± 0.49	-2.6
CW-GMLVQ	98.89 ± 0.34	98.86 ± 0.24	-0.0
SNG ($\rho=1.0$)	81.93 ± 0.94	82.26 ± 0.90	$+0.3$
SMNG ($\rho=1.0$)	87.23 ± 1.35	84.55 ± 0.83	-2.7
SCMNG ($\rho=1.0$)	98.63 ± 0.37	98.81 ± 0.40	$+0.2$

Table 13: Random vs. class-conditional initialization on the synthetic multimodal dataset (8 classes, 16D, 3 imbalanced modes per class, $P_c = 3$). Best ρ per model.

Model	Class-cond	Random	Δ
GLVQ	68.94 ± 3.00	56.63 ± 2.85	-12.3
GMLVQ	99.69 ± 0.64	99.56 ± 0.49	-0.1
CW-GMLVQ	99.31 ± 0.65	99.44 ± 0.59	$+0.1$
SNG ($\rho=0.5$)	78.12 ± 3.54	70.94 ± 3.38	-7.2
SMNG ($\rho=1.0$)	99.69 ± 0.42	99.75 ± 0.41	$+0.1$
SCMNG ($\rho=1.0$)	99.56 ± 0.63	99.50 ± 0.67	-0.1

Random Prototype Initialization: We noted, however, that all the preceding results were based on class-conditional initialization. One might expect cooperation to recover its value when prototypes are poorly placed, since the topological organizing property of NG is designed for that setting [13] in practice. We test this hypothesis on two datasets: Pendigits ($P_c = 5$), where classes are well separated and a synthetic multimodal dataset designed to stress prototype placement. In both cases, prototypes are drawn uniformly from the data range, ignoring class labels entirely. The artificial dataset has 8 classes in 16 dimensions, each comprising 3 isotropic Gaussian modes ($\sigma = 1.0$) with imbalanced sizes (150, 30 and 20 samples per mode; 1600 total). Class centers are placed at equal angles on a circle of radius 8 in dimensions 1–2; mode offsets within each class are $(0, 0)$, $(5, 0)$ and $(2.5, 4.3)$ in dimensions 3–4; dimensions 5–16 carry only the isotropic noise from the Gaussian sampling. The data are generated with NumPy random seed 42. We use $P_c = 3$ to ensure that prototype placement across modes is non-trivial. The summary Tables 12 and 13 report the results. The two datasets exhibit complementary observations. On Pendigits, cooperation confers initialization robustness with the SNG learner nearly unaffected ($\Delta = +0.3$ pp) while GLVQ drops 5.5 pp. However, this apparent robustness does not translate into a competitive advantage as cooperating models still trail their baselines in absolute accuracy. On the artificial dataset, cooperation provides a clear benefit for Euclidean distance as SNG (70.94%) outperforms GLVQ (56.63%) by 14.3 pp under random initialization, confirming that the topological organizing property helps prototypes discover uncovered modes. For metric learning, the picture is dataset-dependent. On Pendigits, SMNG (84.55%) lags GMLVQ (92.97%) by 8.4 pp even under random initialization, indicating that gradient dilution dominates when classes are well separated. On the synthetic multimodal data, SMNG (99.75%) matches GMLVQ (99.56%), suggesting that cooperation's placement benefit offsets gradient dilution when prototype coverage of multiple modes is critical. In both datasets, CW-GMLVQ and GMLVQ are virtually unaffected by initialization quality.

6 Conclusion

Metric-adaptive extensions of Supervised Neural Gas beyond diagonal relevance weighting, namely SMNG, SLNG, STNG and SCMNG, have been presented in this paper. This work establishes a systematic investigation of how rank-based neighbourhood cooperation from Neural Gas interacts with increasingly expressive distance functions. The mathematical formulations of the cost functions, distance gradients and update rules for all variants have been derived within the SNG framework. Two failure modes have been identified through gradient-level analysis: (i) gradient dilution and (ii) subtraction amplification, explaining the consistent degradation observed in cooperation-based metric learners. A class-wise matrix variant SCMNG has been introduced based on three structural conditions under which rank-based cooperation remains compatible with metric parameterisation learning. The numerical evaluation of experimental results on four benchmark datasets indicates that per-class metric localisation, not NG cooperation, is the source of classification improvement. Prototype initialisation strategies that are representative of the data class distribution cause the cooperation mechanism to contribute gradient interference rather than improved metric adaptation.

Cooperation therefore aids classification only when prototypes require topological spreading, a condition that arises under poor initialization. Future studies will investigate adaptive γ schedules that separate early cooperative spreading from later metric refinement and comparison with deep metric learning on larger-scale problems.

References

[1] D. Dua and C. Graff. *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences, 2017. `https://archive.ics.uci.edu/ml`.

[2] T. Bojer, B. Hammer, D. Schunk and K. Tluk von Toschanowitz. Relevance determination in learning vector quantization. In *Proc. European Symposium on Artificial Neural Networks (ESANN'01)*, pages 271–276, Bruges, Belgium, 2001.

[3] M. Biehl, B. Hammer and T. Villmann. Prototype-based models in machine learning. *WIREs Cognitive Science*, 7(2):92–111, 2016.

[4] K. Bunte, P. Schneider, B. Hammer, F.-M. Schleif, T. Villmann and M. Biehl. Limited rank matrix learning, discriminative dimension reduction and visualization. *Neural Networks*, 26:159–173, 2012.

[5] E. Erwin, K. Obermayer and K. Schulten. Self-organizing maps: ordering, convergence properties, and energy functions. *Biological Cybernetics*, 67(1):47–55, 1992.

[6] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[7] S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*. Springer Science & Business Media, 2000.

[8] K. Crammer, R. Gilad-Bachrach, A. Navot and N. Tishby. Margin analysis of the LVQ algorithm. In *Advances in Neural Information Processing Systems*, 2002.

[9] B. Hammer and T. Villmann. Generalized relevance learning vector quantization. *Neural Networks*, 15:1059–1068, 2002.

[10] B. Hammer and T. Villmann. Batch-RLVQ. In M. Verleysen, editor, *European Symposium on Artificial Neural Networks (ESANN'02)*, pages 295–300. D-side publications, 2002.

[11] T. Heskes. On self-organizing maps, vector quantization, and mixture modeling. *IEEE Transactions on Neural Networks*, 12(6):1299–1305, 2001.

[12] B. Hammer, M. Strickert and T. Villmann. Learning vector quantization for multimodal data. In J. R. Dorronsoro, editor, *Artificial Neural Networks — ICANN 2002*, pages 370–375. Springer, 2002.

[13] B. Hammer, M. Strickert and T. Villmann. Supervised neural gas with general similarity measure. *Neural Processing Letters*, 21(1):21–44, 2005.

[14] B. H. Juang and S. Katagiri. Discriminative learning for minimum error classification. *IEEE Transactions on Signal Processing*, 40(12):3043–3054, 1992.

[15] S. Kaski and J. Sinkkonen. A topography-preserving latent variable model with learning metrics. In N. Allinson, H. Yin, L. Allinson and J. Slack, editors, *Advances in Self-Organizing Maps*, pages 224–229. Springer, 2001.

[16] T. Kohonen. Learning vector quantization. In M. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 537–540. MIT Press, 1995.

[17] T. Kohonen. *Self-Organizing Maps*. Springer, Berlin, 1997.

[18] T. Martinetz, S. Berkovich and K. Schulten. "Neural-gas" network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks*, 4(4):558–569, 1993.

[19] K. Musgrave, S. Belongie and S.-N. Lim. A metric learning reality check. In *Proc. European Conference on Computer Vision (ECCV)*, pages 681–699, 2020.

[20] M. Pregenzer, G. Pfurtscheller and D. Flotzinger. Automated feature selection with distinction sensitive learning vector quantization. *Neurocomputing*, 11:19–29, 1996.

[21] H. Ritter, T. Martinetz and K. Schulten. *Neural Computation and Self-Organizing Maps: An Introduction*. Addison-Wesley, 1992.

[22] O. Roy and M. Vetterli. The effective rank: A measure of effective dimensionality. In *15th European Signal Processing Conference (EUSIPCO)*, pages 606–610, 2007.

[23] A. S. Sato and K. Yamada. Generalized learning vector quantization. In *Advances in Neural Information Processing Systems*, volume 8, pages 423–429. MIT Press, 1995.

[24] A. S. Sato and K. Yamada. An analysis of convergence in generalized LVQ. In *ICANN'98*, pages 172–176. Springer, 1998.

[25] P. Schneider, M. Biehl and B. Hammer. Adaptive relevance matrices in learning vector quantization. *Neural Computation*, 21(12):3532–3561, 2009.

[26] S. Seo and K. Obermayer. Soft learning vector quantization. *Neural Computation*, 15(7):1589–1604, 2003.

[27] J. Snell, K. Swersky and R. S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems 30*, pages 4077–4087, 2017.

[28] P. Simard, Y. LeCun and J. Denker. Efficient pattern recognition using a new transformation distance. In *Advances in Neural Information Processing Systems 5*, pages 50–58, 1993.

[29] S. Saralajew and T. Villmann. Adaptive tangent distances in generalized learning vector quantization for transformation and distortion invariant classification. In *Proc. Int. Joint Conf. Neural Networks (IJCNN)*, pages 2672–2679, 2016.

[30] M. Strickert, T. Bojer and B. Hammer. Generalized relevance LVQ for time series. In *Artificial Neural Networks — ICANN 2001*, pages 677–683. Springer, 2001.

[31] T. Villmann, R. Der, M. Herrmann and T. M. Martinetz. Topology preservation in self-organizing feature maps: exact definition and precise measurement. *IEEE Transactions on Neural Networks*, 8(2):256–266, 1997.

[32] T. Villmann, E. Merényi and B. Hammer. Neural maps in remote sensing image analysis. *Neural Networks*, 16(3–4):389–403, 2003.

[33] T. Villmann and S. Haase. Divergence-based vector quantization. *Neural Computation*, 23(5):1343–1392, 2011.

[34] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009.

A Derivation of the SCMNG Update Rules

This appendix verifies that the SCMNG update rules (39)–(42) are indeed obtained by gradient descent on the SCMNG cost (37) with class-wise distance (36). The technical complication is that the cost involves the rank ordering of same-class prototypes and the identity of the nearest wrong-class prototype, both of which are discontinuous in the model parameters. We demonstrate that the gradient contributions generated by these discontinuities cancel identically.

Notation. Let H denote the Heaviside step function, $H(x) = 1$ for $x > 0$ and $H(x) = 0$ otherwise, with distributional derivative $H'(x) = \delta(x)$ (the Dirac delta). The rank of prototype $\mathbf{w}_p$ among same-class prototypes can be expressed as

$$k_p\big(\mathbf{v}, \mathbf{W}_{c_\mathbf{v}}\big) = \sum_{\mathbf{w}_q \in \mathbf{W}_{c_\mathbf{v}}} H\big(d_c(\mathbf{v}, \mathbf{w}_p) - d_c(\mathbf{v}, \mathbf{w}_q)\big), \tag{63}$$

where $d_c(\mathbf{v}, \mathbf{w}_p) = \|\Omega_{c_p}(\mathbf{v} - \mathbf{w}_p)\|^2$ is the class-wise distance (36). Let $\mathbf{W}^C_{c_\mathbf{v}} = \mathbf{W} \setminus \mathbf{W}_{c_\mathbf{v}}$ collect the wrong-class prototypes. The nearest wrong-class indicator is $n_{p'}(\mathbf{v}) = H\big(-k_{p'}(\mathbf{v}, \mathbf{W}^C_{c_\mathbf{v}})\big)$, where $k_{p'}$ denotes the rank of $\mathbf{w}_{p'}$ among $\mathbf{W}^C_{c_\mathbf{v}}$, so that $n_{p'} = 1$ if and only if $\mathbf{w}_{p'}$ is the closest wrong-class prototype to $\mathbf{v}$.

Writing out both the rank and the nearest wrong-class indicator explicitly, the cost (37) becomes

$$E_{\mathrm{SCMNG}} = \sum_{\mathbf{v} \in V} \sum_{\substack{\mathbf{w}_p \in \mathbf{W}_{c_\mathbf{v}} \\ \mathbf{w}_{p'} \in \mathbf{W}^C_{c_\mathbf{v}}}} \frac{h_\gamma(k_p) \cdot n_{p'}}{C(\gamma, K_{c_\mathbf{v}})}\ \Phi(\mu^{p,p'}), \tag{64}$$

where $\mu^{p,p'} = (d_p - d_{p'})/(d_p + d_{p'})$. We abbreviate $d_p = d_c(\mathbf{v}, \mathbf{w}_p)$ and $d_{p'} = d_c(\mathbf{v}, \mathbf{w}_{p'})$.

Key lemma (pairwise antisymmetry). Let $\{a_p\}$ be quantities indexed by same-class prototypes that depend on distances only through the value d_p and consider a double sum of the form

$$S = \sum_{\mathbf{w}_p,\, \mathbf{w}_q \in \mathbf{W}_{c_\mathbf{v}}} a_p\ \delta(d_p - d_q)\ \big(g_p - g_q\big), \tag{65}$$

with g_p arbitrary. Exchanging the summation indices $p \leftrightarrow q$ gives $S = \sum a_q\, \delta(d_q - d_p)\, (g_q - g_p)$. The Dirac delta is symmetric and $a_p = a_q$ whenever $d_p = d_q$ (equal distances force equal ranks $k_p = k_q$ and hence equal neighborhood weights and cost values), so $S = -S$ and therefore $S = 0$. We invoke this cancellation repeatedly below; an entirely analogous argument handles sums over wrong-class prototypes.

Derivative with respect to $\mathbf{w}_r \in \mathbf{W}_{c_\mathbf{v}}$. Differentiating (64) with respect to a same-class prototype $\mathbf{w}_r$ yields two types of terms.

Direct terms. The distance $d_r = d_c(\mathbf{v}, \mathbf{w}_r)$ appears in the summand $p = r$ through $\Phi(\mu^{r,p'})$. Using the SCMNG distance gradient (38), $\partial d_c/\partial \mathbf{w}_r = -2\,\Lambda_{c_\mathbf{v}}\boldsymbol{\delta}_r$ and noting that $n_{p'} = 1$ for exactly one wrong-class prototype (the nearest, $\mathbf{w}_{r-}$), one obtains

$$\left.\frac{\partial E_{\mathrm{SCMNG}}}{\partial \mathbf{w}_r}\right|_{\text{direct}} = \bar{h}_r\ \Phi'|_{\mu^r}\ \ \xi_r^+\ (-2\,\Lambda_{c_\mathbf{v}}\,\boldsymbol{\delta}_r). \tag{66}$$

Multiplication by $-\varepsilon^+$ yields (39): $\Delta\mathbf{w}_r = 2\varepsilon^+ \cdot \bar{h}_r \cdot \Phi'|_{\mu^r} \cdot \xi_r^+ \cdot \Lambda_{c_\mathbf{v}}\,\boldsymbol{\delta}_r$.

Rank-change terms. Changing $\mathbf{w}_r$ modifies d_r and thereby potentially changes the ranks of all same-class prototypes via (63). These contributions involve

$$\sum_{\mathbf{w}_p,\,\mathbf{w}_q \in \mathbf{W}_{c_\mathbf{v}}} A_p\ \frac{\partial h_\gamma}{\partial k_p}\ \delta(d_p - d_q)\ \left(\frac{\partial d_p}{\partial \mathbf{w}_r} - \frac{\partial d_q}{\partial \mathbf{w}_r}\right),$$

where A_p collects the remaining factors. For $p \neq r$ and $q \neq r$, neither d_p nor d_q depends on $\mathbf{w}_r$, so both partial derivatives vanish. The surviving terms pair $(p, q) = (r, j)$ with $(p, q) = (j, r)$ for each $j \in \mathbf{W}_{c_\mathbf{v}} \setminus \{\mathbf{w}_r\}$. The Dirac delta $\delta(d_r - d_j) = \delta(d_j - d_r)$ enforces $d_r = d_j$, at which point $k_r = k_j$ and hence $A_r = A_j$ and $h'_\gamma(k_r) = h'_\gamma(k_j)$. Yet the difference factors have opposite signs:

$$(r, j)\colon\ +\frac{\partial d_r}{\partial \mathbf{w}_r}, \qquad\qquad (j, r)\colon\ -\frac{\partial d_r}{\partial \mathbf{w}_r}.$$

These cancel pairwise. The terms involving $\partial C/\partial \mathbf{w}_r$ vanish by the same argument, since $C = \sum_p h_\gamma(k_p)$ and each $\partial k_p/\partial \mathbf{w}_r$ produces Dirac deltas with the same antisymmetric structure.

Derivative with respect to $\mathbf{w}_{r-} \in \mathbf{W}^C_{c_\mathbf{v}}$. The nearest wrong-class prototype $\mathbf{w}_{r-}$ affects (64) through the distance $d_{r-} = d_c(\mathbf{v}, \mathbf{w}_{r-})$ and through the indicator n_{r-}.

Direct terms. When $n_{r-} = 1$, differentiating with respect to $\mathbf{w}_{r-}$ and using $\partial d_c/\partial \mathbf{w}_{r-} = -2\,\Lambda_{c_{r-}}\,\boldsymbol{\delta}_{r-}$ gives

$$\left.\frac{\partial E_{\mathrm{SCMNG}}}{\partial \mathbf{w}_{r-}}\right|_{\text{direct}} = \sum_{\mathbf{w}_r \in \mathbf{W}_{c_\mathbf{v}}} \bar{h}_r\ \Phi'|_{\mu^r}\ \ (-\xi_r^-)\ (-2\,\Lambda_{c_{r-}}\,\boldsymbol{\delta}_{r-}). \tag{67}$$

The summation over all same-class prototypes $\mathbf{W}_{c_\mathbf{v}}$ arises because $\mathbf{w}_{r-}$ appears in the cost term for *every* cooperating correct-class prototype: each $\mathbf{w}_r$ pairs with the same nearest wrong-class prototype and contributes a repulsive signal weighted by its own $\bar{h}_r$. Multiplication by $-\varepsilon^-$ yields (40): $\Delta\mathbf{w}_{r-} = -2\varepsilon^- \cdot \Lambda_{c_{r-}}\,\boldsymbol{\delta}_{r-} \cdot \sum_{\mathbf{w}_r \in \mathbf{W}_{c_\mathbf{v}}} \bar{h}_r \cdot \Phi'|_{\mu^r} \cdot \xi_r^-$.

Winner-change terms. These arise from differentiating $n_{p'}$ with respect to $\mathbf{w}_{r-}$. Through (63) applied to $\mathbf{W}^C_{c_\mathbf{v}}$, each $\partial n_{p'}/\partial \mathbf{w}_{r-}$ produces Dirac deltas $\delta(d_{p'} - d_{q'})$ with factors $(\partial d_{p'}/\partial \mathbf{w}_{r-} - \partial d_{q'}/\partial \mathbf{w}_{r-})$. By pairwise antisymmetry over the wrong-class summation indices p' and q', these vanish.

Derivative with respect to $\Omega_{c_\mathbf{v}}$ (correct-class matrix). Since all same-class prototypes $\mathbf{w}_r \in \mathbf{W}_{c_\mathbf{v}}$ use the shared matrix $\Omega_{c_\mathbf{v}}$, the distances $d_r = d_c(\mathbf{v}, \mathbf{w}_r) = \|\Omega_{c_\mathbf{v}}(\mathbf{v} - \mathbf{w}_r)\|^2$ all depend on $\Omega_{c_\mathbf{v}}$. The wrong-class distance d_{r-}, however, depends on $\Omega_{c_{r-}}$ with $c_{r-} \neq c_\mathbf{v}$, and is therefore independent of $\Omega_{c_\mathbf{v}}$.

Direct terms. Using $\partial d_c/\partial \Omega_{c_\mathbf{v}} = 2\,\Omega_{c_\mathbf{v}}\,\boldsymbol{\delta}_r\boldsymbol{\delta}_r^T$ from (38) and noting that the ξ_r^- term vanishes (since $\partial d_{r-}/\partial \Omega_{c_\mathbf{v}} = 0$):

$$\left.\frac{\partial E_{\mathrm{SCMNG}}}{\partial \Omega_{c_\mathbf{v}}}\right|_{\text{direct}} = \sum_{\mathbf{w}_r \in \mathbf{W}_{c_\mathbf{v}}} \bar{h}_r\ \Phi'|_{\mu^r}\ \ \xi_r^+\ 2\,\Omega_{c_\mathbf{v}}\,\boldsymbol{\delta}_r\boldsymbol{\delta}_r^T. \tag{68}$$

Multiplication by $-\varepsilon_\Omega$ yields (41): $\Delta\Omega_{c_\mathbf{v}} = -2\varepsilon_\Omega \cdot \sum_{\mathbf{w}_r \in \mathbf{W}_{c_\mathbf{v}}} \bar{h}_r \cdot \Phi'|_{\mu^r} \cdot \xi_r^+ \cdot \Omega_{c_\mathbf{v}}\,\boldsymbol{\delta}_r\boldsymbol{\delta}_r^T$.

Rank-change terms. These have the form of (65) with $g_p = \partial d_p/\partial \Omega_{c_\mathbf{v}} = 2\,\Omega_{c_\mathbf{v}}\,\boldsymbol{\delta}_p\boldsymbol{\delta}_p^T$. At $d_p = d_q$, the prefactors (neighborhood weights and cost values) are symmetric in (p, q) while the difference $g_p - g_q$ is antisymmetric under index exchange. Hence these vanish by pairwise antisymmetry.

Derivative with respect to $\Omega_{c_{r-}}$ (wrong-class matrix). The wrong-class distance $d_{r-} = \|\Omega_{c_{r-}}(\mathbf{v} - \mathbf{w}_{r-})\|^2$ depends on $\Omega_{c_{r-}}$, while the correct-class distances d_r are independent of $\Omega_{c_{r-}}$ (since $c_{r-} \neq c_{\mathbf{v}}$).

Direct terms. Using $\partial d_c/\partial\Omega_{c_{r-}} = 2\,\Omega_{c_{r-}}\,\boldsymbol{\delta}_{r-}\boldsymbol{\delta}_{r-}^T$ and noting that the ξ_r^+ term vanishes (since $\partial d_r/\partial\Omega_{c_{r-}} = 0$):

$$\left.\frac{\partial E_{\mathrm{SCMNG}}}{\partial\Omega_{c_{r-}}}\right|_{\text{direct}} = \sum_{\mathbf{w}_r\in\mathbf{W}_{c_{\mathbf{v}}}} \bar{h}_r\;\Phi'|_{\mu^r}\;\;(-\xi_r^-)\;2\,\Omega_{c_{r-}}\,\boldsymbol{\delta}_{r-}\boldsymbol{\delta}_{r-}^T. \tag{69}$$

Multiplication by $-\varepsilon_\Omega$ yields (42): $\Delta\Omega_{c_{r-}} = 2\varepsilon_\Omega \cdot \Omega_{c_{r-}}\,\boldsymbol{\delta}_{r-}\boldsymbol{\delta}_{r-}^T \cdot \sum_{\mathbf{w}_r\in\mathbf{W}_{c_{\mathbf{v}}}} \bar{h}_r \cdot \Phi'|_{\mu^r} \cdot\xi_r^-$.

Winner-change terms. By the analogous pairwise antisymmetry argument applied to $\mathbf{W}_{c_{\mathbf{v}}}^C$, these vanish.

Hence only the direct gradient terms (66)–(69) survive, confirming the SCMNG update rules (39)–(42). The class-wise separation of (41) and (42) is a structural consequence of each class maintaining an independent Ω_c: correct-class and wrong-class gradients flow into *different* matrices, eliminating the cross-class gradient contamination identified in Section 3.2.

B Extension to SMNG, SLNG and STNG

The pairwise antisymmetry argument used in Appendix A rests on three properties of the cost: (i) ranks are defined via Heaviside indicators on pairwise distance differences (63), (ii) equal distances entail equal ranks, neighborhood weights and cost values and (iii) the distance gradient with respect to a parameter θ_k vanishes unless θ_k enters the distance computation. All three properties hold for every SNG variant of the form (6) with a differentiable distance. We verify the remaining models by substituting the appropriate distance gradients.

SMNG. The interesting case is the derivative with respect to Ω. Since SMNG shares a single Ω across all prototypes, every distance in the cost depends on it. The distance gradients $\partial d_\Omega/\partial\mathbf{w}_r = -2\,\Lambda\,\boldsymbol{\delta}_r$ and $\partial d_\Omega/\partial\Omega = 2\,\Omega\,\boldsymbol{\delta}_r\boldsymbol{\delta}_r^T$ (cf. (18)) yield the SMNG updates (19)–(21) by direct substitution. The rank-change argument for the prototype gradient proceeds exactly as in the SCMNG case: perturbing $\mathbf{w}_r$ affects only d_r, producing the same $(r,j)/(j,r)$ Dirac pairing.

For the Ω derivative, rank-change terms involve all same-class pairs (p,q) since every $d_p = \|\Omega\boldsymbol{\delta}_p\|^2$ depends on Ω. At $d_p = d_q$, the prefactors are symmetric while $g_p - g_q = 2\,\Omega(\boldsymbol{\delta}_p\boldsymbol{\delta}_p^T - \boldsymbol{\delta}_q\boldsymbol{\delta}_q^T)$ changes sign under $p \leftrightarrow q$, so these vanish by (65). A complication absent from the SCMNG case: because wrong-class distances also depend on Ω, differentiating the indicator $n_{p'}$ generates winner-change terms $\delta(d_{p'} - d_{q'})$ over $\mathbf{W}_{c_{\mathbf{v}}}^C$. These vanish by the same antisymmetric pairing over wrong-class indices.

What the proof does *not* remove is the consequence of sharing itself: the verified update (21) funnels gradient contributions from training samples of every class into a single matrix, which is precisely the source of gradient dilution (Section 4.1).

SLNG. Per-prototype ownership simplifies the argument. Since Ω_k belongs to prototype k alone, $\partial d_p/\partial\Omega_k = 0$ for all $p \neq k$. Only the pairs (k,j) and (j,k) survive the rank-change expansion and they cancel by antisymmetry. No winner-change terms appear because wrong-class distances use a different matrix Ω_{r-}. Substituting the gradients (24) directly recovers (25)–(28).

STNG. The subtractive form of tangent distance ($d_T = \|\boldsymbol{\delta}_k\|^2 - \|\Omega_k^T\boldsymbol{\delta}_k\|^2$) might seem to threaten the cancellation, but it does not. The antisymmetry argument depends on the structure of the rank function (pairwise distance comparisons via Heaviside indicators), not on the sign of the distance gradient $\partial d_T/\partial\Omega_k = -2\,\boldsymbol{\delta}_k\boldsymbol{\delta}_k^T\Omega_k$ (cf. (31)). Since Ω_k is private to its prototype, the same simplification as for SLNG applies: only $(k,j)/(j,k)$ pairs survive and they cancel by antisymmetry. The post-update QR re-orthonormalization is a projection onto the Stiefel manifold applied *after* the gradient step; it does not enter the cost and creates no additional terms. Substitution recovers (32)–(35).