AN INDUCTIVE BIAS FOR DISTANCES: NEURAL NETS THAT RESPECT THE TRIANGLE INEQUALITY

Parviz Karimov

MIPT, 2024

April 8, 2024

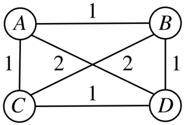
1. Motivation & Background
2. Deep Norm
3. Wide Norm
4. Application: Modelling Graph Distances
5. Computational considerations

Motivation

When defining distances, the triangle inequality has proven to be a useful constraint, both theoretically — to prove convergence and optimality guarantees — and empirically — as an inductive bias.

Deep metric learning architectures that respect the triangle inequality rely, almost exclusively, on Euclidean distance in the latent space. Though effective, this fails to model two broad classes of subadditive distances, common in graphs and reinforcement learning: asymmetric metrics, and metrics that cannot be embedded into Euclidean space.

Example



Norm	MSE
Euclidean, $\mathbb{R}^n, \forall n$	0.057
Deep Norm, $\mathbb{R}^2$	0.000
Wide Norm, $\mathbb{R}^2$	0.000

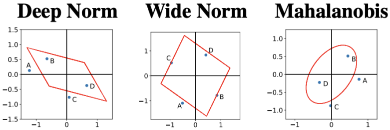


Figure: The nodes in the graph (left) cannot be embedded into any $\mathbb{R}^n$ so that edge distances are represented by the Euclidean metric: points $\phi(A)$ and $\phi(D)$ must lie at the midpoint of the segment from $\phi(B)$ to $\phi(C)$—but then $\phi(A)$ and $\phi(D)$ coincide, which is incorrect.

Background

Metric

A **metric** is a function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ satisfying $\forall x, y, z \in \mathcal{X}$

M1 $d(x, y) \geq 0$

M2 $d(x, y) = 0 \iff x = y$

M3 $d(x, z) \leq d(x, y) + d(y, z)$

M4 $d(x, y) = d(y, x)$

Norm

A **norm** is a function $||\cdot|| : \mathcal{X} \to \mathbb{R}^+$ satisfying $\forall x, y \in \mathcal{X}, \alpha \in \mathbb{R}^+$

N1 $||x|| > 0$ unless $x = 0$

N2 $\alpha||x|| = ||\alpha x||$

N3 $||x + y|| \leq ||x|| + ||y||$

N4 $||x|| = ||-x||$

Convex function

Function $f : \mathcal{X} \to \mathbb{R}$ is called **convex** if

C1 $\forall x, y \in \mathcal{X}, \alpha \in [0, 1] : f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$

Background

Auxilliary

A **quasi-metric** is **M1** and **M3**.
An **assymetric norm** is **N1-N3**.
An **(assymetric) semi-norm** is nonnegative, **N2** and **N3** (and **N4**).

Prop. 0.1

Any assymetric semi-norm induces a quasi-metric. Any induces quasi-metric is translation-inveriant and positive homogeneous.

Prop. 0.2

Any N2 and N3 function is convex — thus, all asymmetric semi-norms are convex.

Deep Norm

Proposition 1

All positive homogeneous convex functions are subadditive; i.e., **C1** $\wedge$ **N2** $\implies$ **N3**.

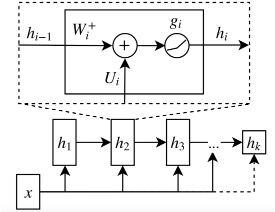


Figure: Deep norm architecture

$$||x|| = h_k$$

$$h_i = g_i(W_i^+ h_{i-1} + U_i x)$$

$$h_0 = 0, W_1^+ = 0$$

g_i preserves **C1**, **N2**

g_k is non-negative

W_i^+ is a non-negative matrix

Deep Norm

Proposition 2

If $||\cdot|$ is an assymetric semi-norm, then $||x|| = ||x| + || - x|$ is a semi-norm

Proposition 3

if $||\cdot||_a$ is an assymetric semi-norm, $||\cdot||_b$ is a norm, $\lambda > 0$, then $||x||_{a+\lambda b} = ||x||_a + \lambda ||x||_b$ is an assymetric norm.

Def. (MaxReLU)

$$\text{maxrelu}(x, y) = [\max(x, y), \alpha \text{relu}(x) + \beta \text{relu}(y)], \quad \alpha, \beta \geq 0$$

Wide Norm

Def. (MaxMean)

$$\text{maxmean}(x_1, ..., x_n) = \alpha \max(x_1, ..., x_n) + (1-\alpha)\text{mean}(x_1, ..., x_n)$$

Def. (Wide Norm)

A **Wide Norm** (or k-component mixture of Mahalanobis norms) is defined as

$$||x|| = \text{maxmean}_i(||W_i x||_2), \text{ where } W_i \in \mathbb{R}^{m_i \times n}, m_i \leq n$$

Wide Norm

Monotonic Norm (in the positive orthant)

N5 $||\cdot||$ is **monotonic in the positive orthant** if
$0 \leq x \leq y \implies ||x|| \leq ||y||$

Proposition 4

if $||\cdot||$ is an **N5** (semi-)norm on $\mathbf{R}^{2n}$, then $||x|| = ||\text{relu}(x :: -x)||$ is an assymetric (semi-)norm on $\mathbb{R}^n$.

Proposition 5

Mahalanobis norm with $W = DU$, with D diagonal and U non-negative, is **N5**.

Universal Approximation Theorem

Theorem

The families $\mathcal{D}$ of Deep Norms (using MaxReLU) and $\mathcal{W}$ of Wide Norms (using MaxMean) are dense in the family $\mathcal{N}$ of asymmetric semi-norms.

	N1 (M1-2)	**N2 (Homo.)**	**N3 (M3)**	**N4 (M4)**	**UA**	**Notes**
Euclidean	✓	✓	✓	✓	✗	
MLP	✗	✗	✗	✗	✓	
Deep Norm	✻	✓	✓	✻	✓	
Wide Norm	✻	✓	✓	✻	✓	works for large minibatches (§§3.5)
Neural Metric	✻	✻	✓	✻	✓	based on Deep Norm or Wide Norm

Figure: Norm (metric) properties of different architectures. As compared to Euclidean architectures, ours are universal asymmetric semi-norm approximators (UA) and can use propositions to optionally satisfy (*) **N1** and **N4**. Neural metrics relax the unnecessary homogeneity constraint on metrics.

Application: Modelling Graph Distances

The task is of modeling shortest path lengths in a weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. So long as edge weights are positive and the graph is connected, shortest path lengths are discrete quasi-metrics ($n = |\mathcal{V}|$), and provide an ideal domain for a comparison to the standard Euclidean approach.

	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\max(d)$	σ_d	Sym?
`to`	278K	611K	145.7	24.5	↔
`3d`	125K	375K	86.7	13.2	↔
`taxi`	391K	752K	111.2	13.4	↔
`push`	390K	1498K	113.1	14.3	→
`3dr`	123K	368K	86.5	13.1	→
`3dd`	125K	375K	97.8	13.4	→

(a) Graph statistics

	Eucl.	WN	DN_I	DN_N	MLP
`to`	12.5	**6.6**	6.7	6.7	12.3
`3d`	31.2	17.3	15.4	**12.9**	20.6
`taxi`	14.4	**10.6**	11.8	11.4	**5.8**
`push`	22.2	14.0	14.7	**13.5**	**11.3**
`3dr`	22.0	**17.5**	21.8	18.3	25.5
`3dd`	211.8	177.1	199.5	**157.7**	252.7

(b) Final test MSE @ $|D| = 50000$

Figure: Graph experiments. (a) Statistics for different graphs. (b) Test MSE after 1000 epochs at training size $|D| = 50000$ (3 seeds). The best metric (and overall result if different) is bolded.

Computational considerations

	32	128	512	2048
Euclidean	0.18	0.27	0.45	1.06
WN 3x600	1.59	1.57	1.75	2.36
WN 64x64	15.7	13.4	17.7	26.3
DN 2x400	0.97	5.73	76.9	293
DN 3x600	1.50	11.4	174	OOM

Figure: Mean computation time (ms) for different mini-batch sizes (250 trials).

Literature

1. **Main article** AN INDUCTIVE BIAS FOR DISTANCES: NEURAL NETS THAT RESPECT THE TRIANGLE INEQUALITY.