Lazy Quotient Metrics:

approximate symmetries for ML models

BIRS-CMO
Mathematics of Deep Learning

Josué Tonelli-Cueto
Johns Hopkins University

joint ONGOING work with...

Dustin MIXON
(Ohio State)

Soledad VILLAR
(Johns Hopkins)

Brantley VOSE
(Ohio State)

WARNING

Talk on ONGOING work,

we have still...

unanswered questions

untraversed paths

unthought directions

But... feel free to give us

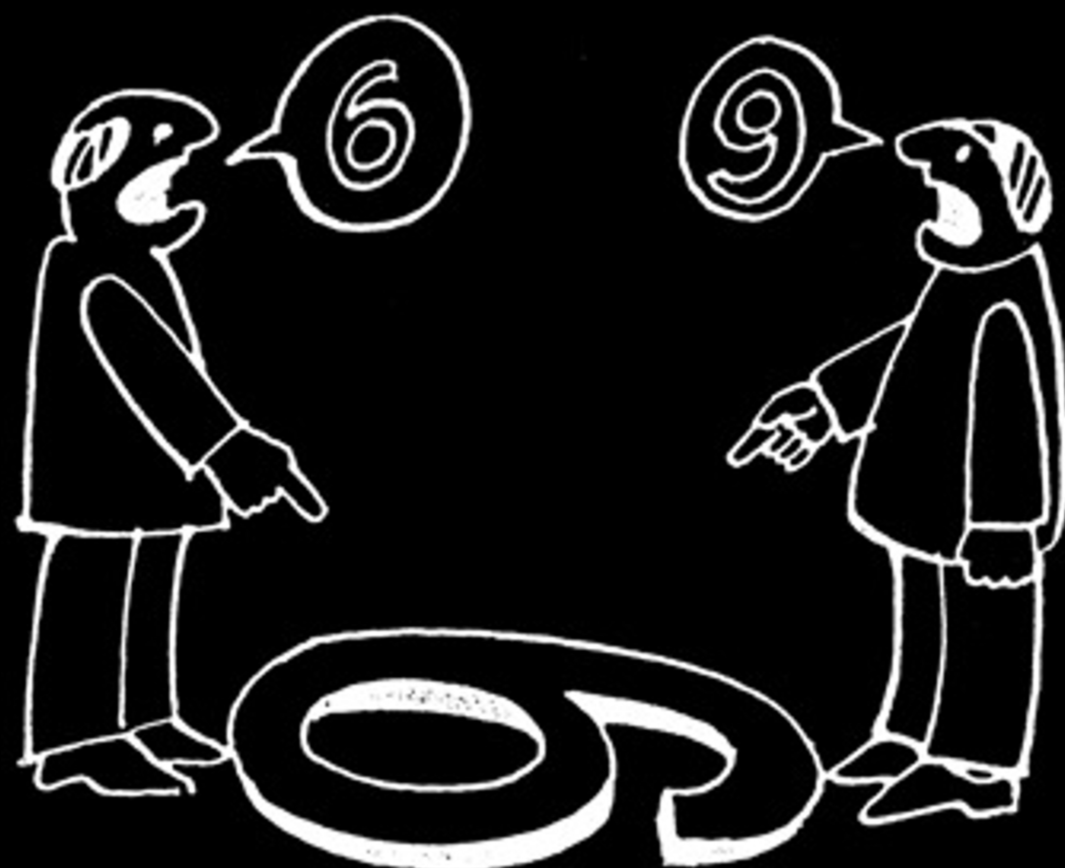
questions, comments, impressions...

Your input is welcome!

Motivating Example I

In shrot, his wtis bneig qtuie gnoe, he hit upon the saetrngst ntoion taht eevr mdaamn in tihs wrlod hit upon, and that was taht he fncaied it was rgiht and reuiiqste, as well for the sppruot of his own hnoour as for the serivce of his conutry, that he souhld mkae a kinght-earrnt of hiemslf, roaimng the wrlod over in full aourmr and on hroseabck in qsuet of aeduventrs, and puittng in prcaitce hiemslf all taht he had read of as bneig the uuasl pairectcs of kngihts-earrnt; rgiihtng eevry knid of worng, and eoixpsng hmiself to peirl and daengr form wihch, in the isuse, he was to reap etrenal rnweon and fame.

Motivating Example II



1st Idea

Small transformations do not change the object, but too many might make it unrecognizable

Setting:

G group

$(\mathbb{X}, d)$ metric space

$G \curvearrowright (\mathbb{X}, d)$ isometries

Quotient 'Metric'

$$d_0(x,y) := \inf_{g \in G} d(gx, y)$$

Lazy Quotient Metric

$$d_\lambda(x,y) := \inf_{g \in G} \sqrt{d(gx,y)^2 + \lambda^2 \nu(g)^2}$$

penalizes large transformations

where $\nu : G \to \mathbb{R}$ group-norm

$\nu(1) = 0, \quad \nu(g) = \nu(g^{-1}),$

$\nu(gh) \leq \nu(g) + \nu(h)$

$$\begin{tikzcd} d_0(x,y) \\ d_r(x,y) \arrow[u, "r \to 0"'] \arrow[d, "r \to \infty"] \\ d(x,y) \end{tikzcd}$$

$d_0(x,y)$ quotient metric

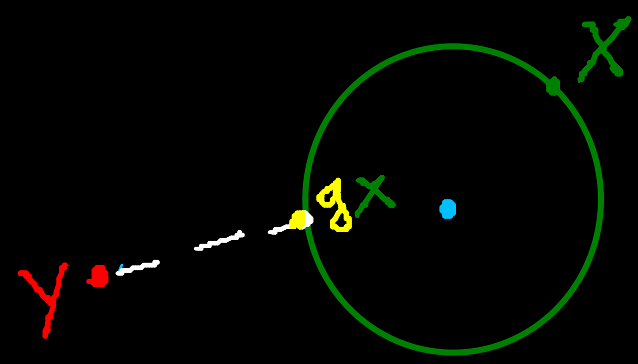
$d(x,y)$ original metric

Theoretical Example

$SO(n) \curvearrowright (\mathbb{R}^n, d)$

$$\mathbb{R}^n / SO(n) \cong [0, \infty)$$

$\hookrightarrow d_0(x, y) = \big| \|x\| - \|y\| \big|$



Theoretical Example

$SO(n) \curvearrowright (\mathbb{R}^n, d)$ $\qquad \nu = \mathrm{dist}_{geod}(\mathbb{I}, \cdot)$

Prop. d_λ is Riemannian with Riemannian metric:

$$g_x(v,w) = \langle v, w \rangle - \frac{\|x\|^2}{2\lambda^2 + \|x\|^2} \langle P_x v, P_x w \rangle$$

where $P_x = \mathbb{I} - \frac{xx^T}{\|x\|^2}$ is the orthoprojection onto $x^\perp$

$$\frac{\langle v, x\rangle \langle x, w\rangle}{\|x\|^2} \xleftarrow{\lambda \to 0} g_x(v,w) \xrightarrow{\lambda \to \infty} \langle v, w \rangle$$

A practical case:

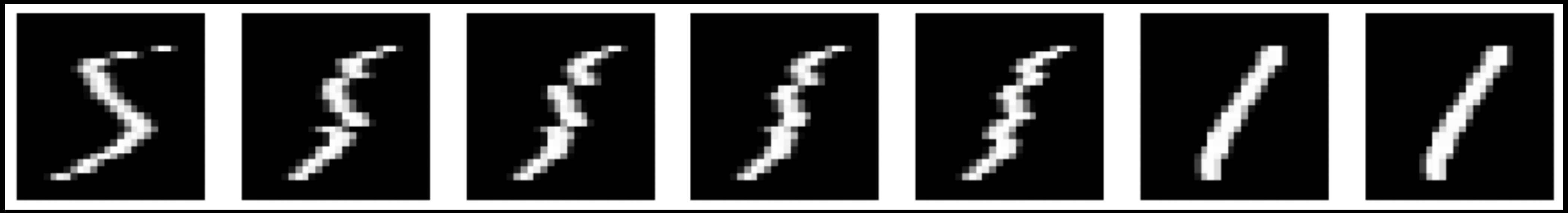
MNIST

Handwritten Digit

with horizontal shifts

Lazy quotient: Practical case

We can cyclically shift the rows of each image



We perform 10-nearest neighborhood classifier for several values of n

Lazy quotient: Practical case

$$X = \mathbb{R}^{28 \times 28}$$

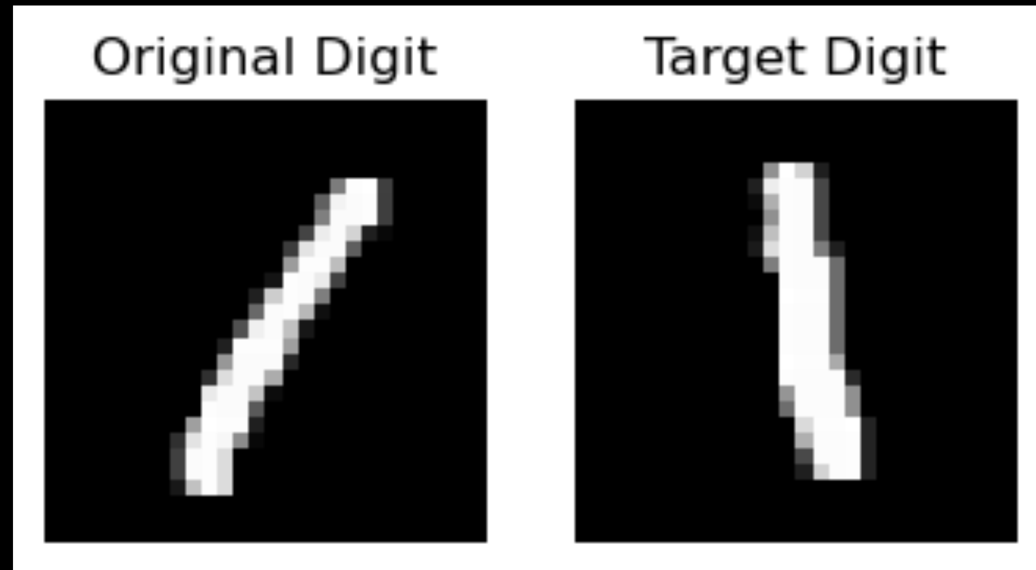
$$g \cdot x := (x_{i,\, j+g_i})_{i,j}$$

$$G = (\mathbb{Z}/28)^{28}$$

$$\nu(g) = \sqrt{\sum_{i=1}^{28} \delta(g_i)^2}$$

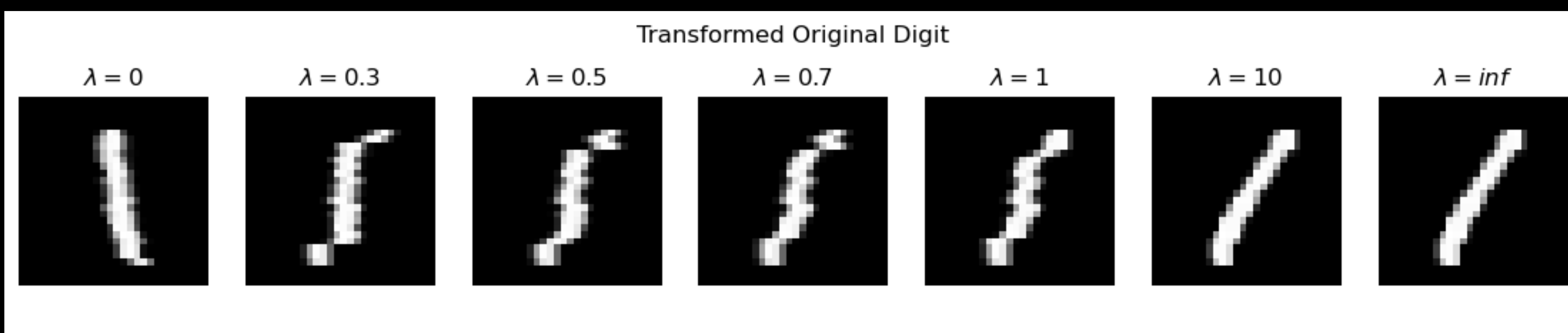
with $\nu(g_i) := \min\{|n| \mid n \equiv g_i \pmod{28}\}$

Lazy quotient: Practical case



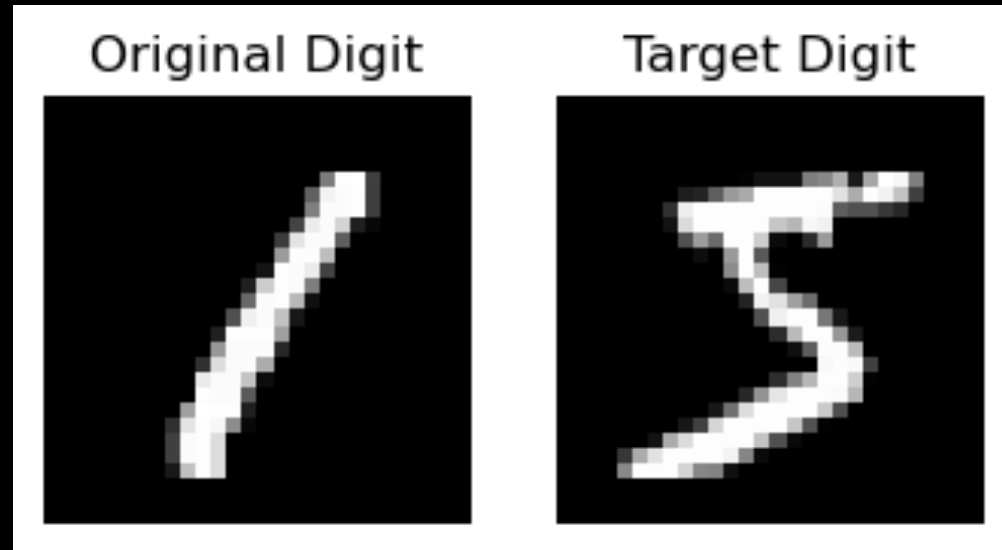
x y

$$d_\lambda(x,y) = \inf_{g\in G}\sqrt{d(gx,y)^2+\lambda^2\nu(g)^2} = \sqrt{d(g_\lambda x,y)^2+\lambda^2\nu(g_\lambda)^2}$$



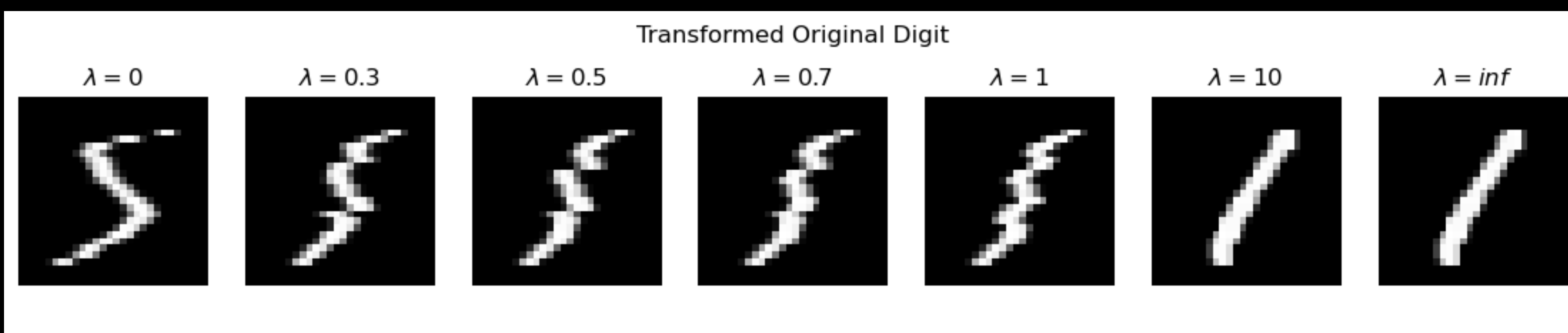
$g_\lambda x$

Lazy quotient: Practical case



x y

$$d_\lambda(x,y) = \inf_{g\in G}\sqrt{d(gx,y)^2+\lambda^2\nu(g)^2} = \sqrt{d(g_\lambda x, y)^2+\lambda^2\nu(g_\lambda)^2}$$



$g_\lambda x$

Lazy quotient: Practical case

λ	Classifier Accuracy
0 (Quotient metric)	0.6630
0.3	0.7819
0.5	0.7827
0.7	**0.7840**
1	0.7837
10	0.7665
∞	0.7665

2nd Idea

(Bietti, Venturi, Bruna; '21)

Average lazyly

the output

Usual Averaging

$$R_G f(x) := \mathbb{E}_{g \in G} f(gx)$$

Lazy Averaging

$$S_G^{\eta} f(x) := \mathbb{E}_{g \sim \eta} f(gx)$$

where $\eta \in \mathcal{P}(G)$ prob. dist.

Variation of (Bietti, Venturi, Bruna; '21)

Problem:

Can we choose

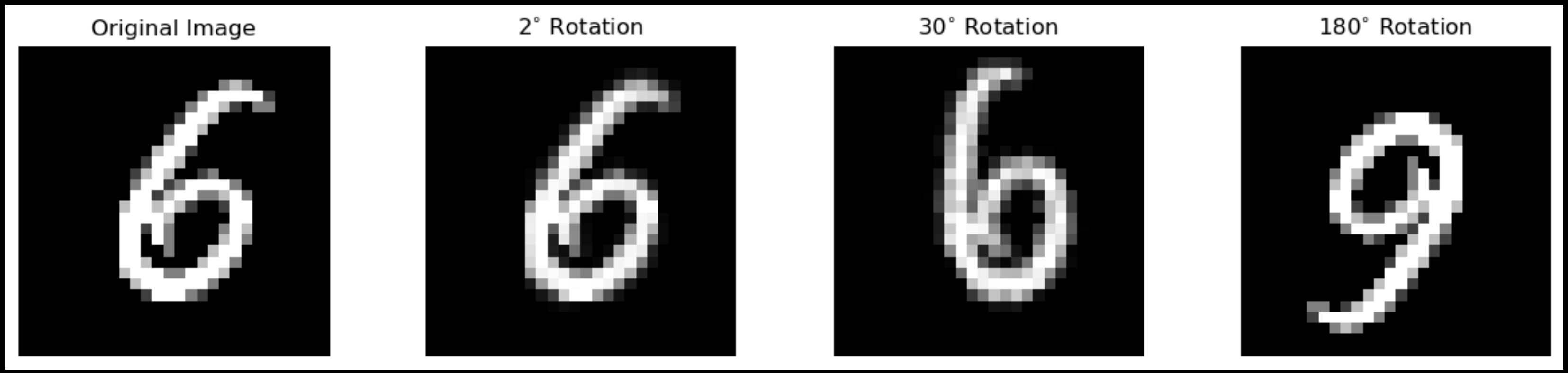
$$\eta \in \mathcal{P}(G)$$

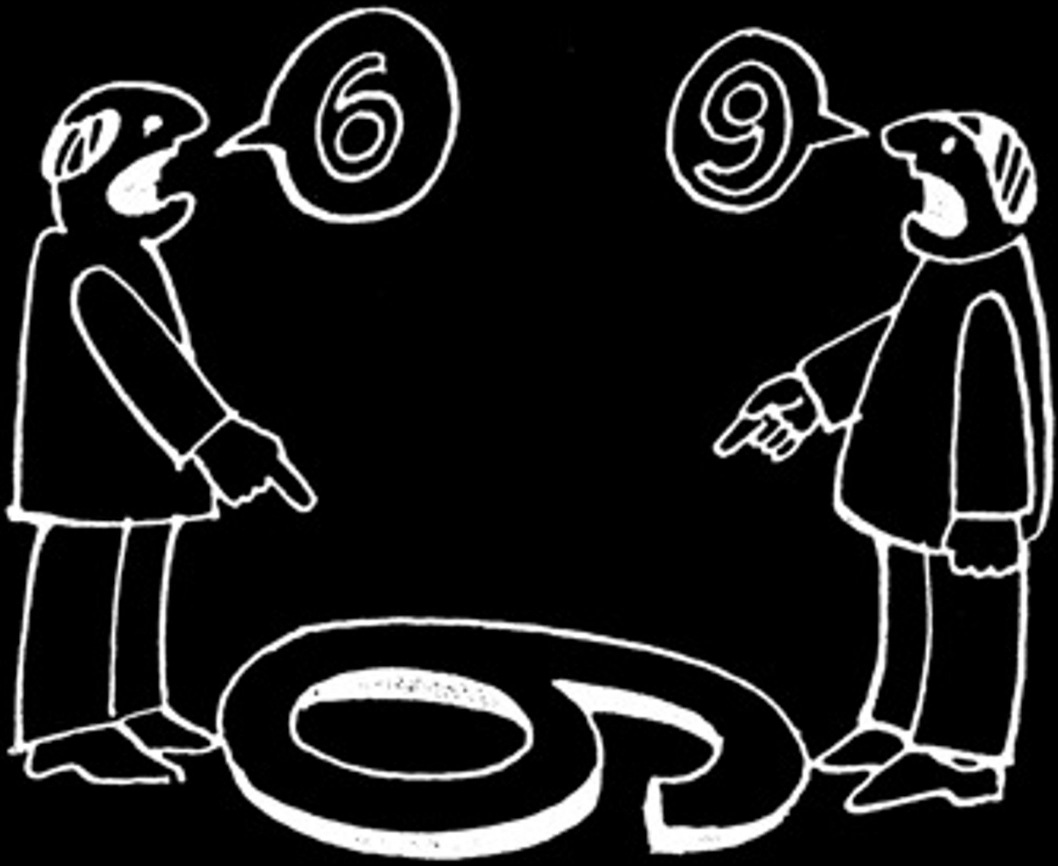
so that S^η_G turns arbitrary $f:(\mathbb{X}, d) \to \mathbb{R}$ into 'quasi-invariant'?

A Practical Case:

MNIST Handwritten Digit with small rotations

Lazy Averaging: Practical case

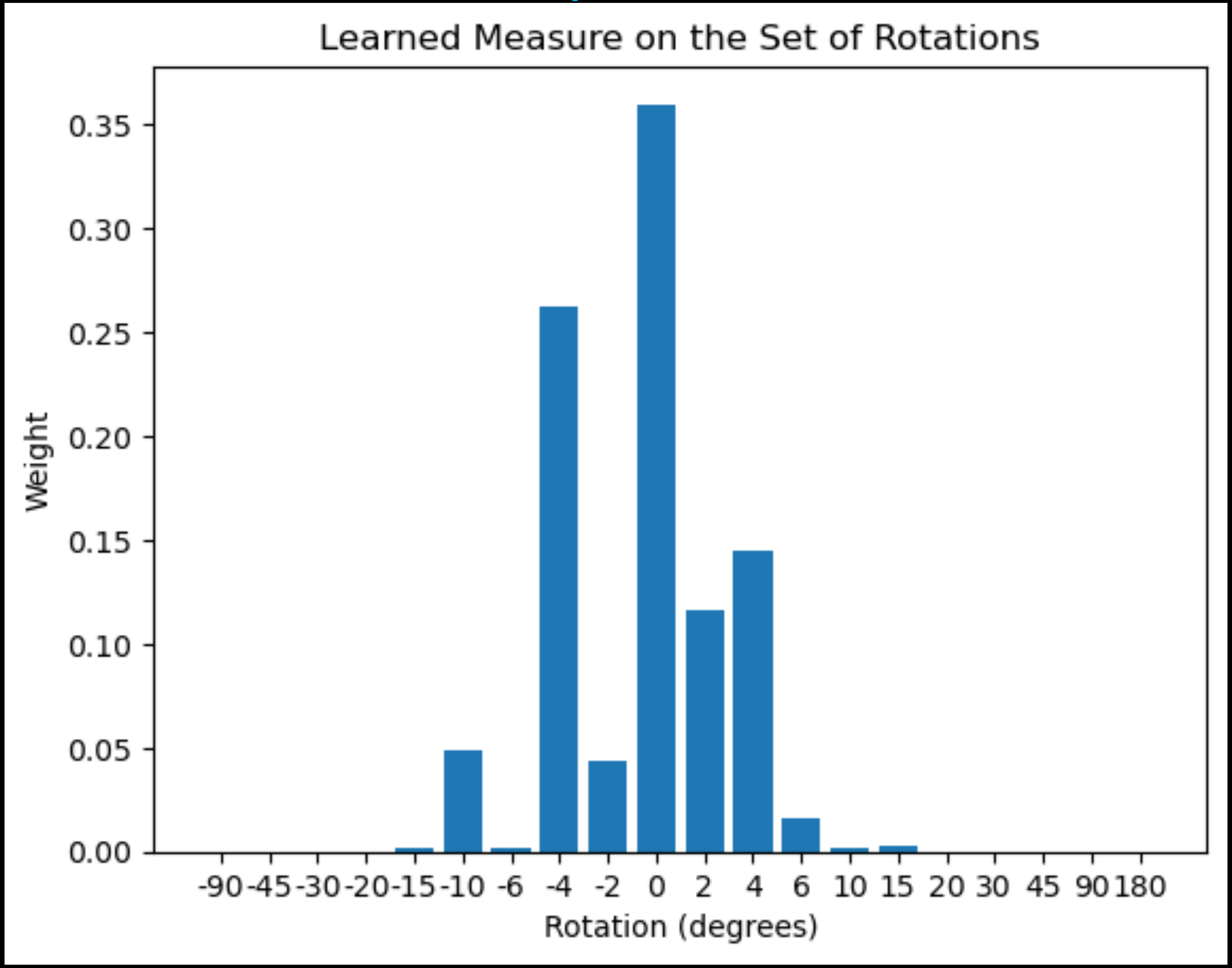




Lazy Averaging: Practical case

$$\pi \in \mathcal{P}(\{0, \pm 2^\circ, \pm 4^\circ, \pm 6^\circ, \pm 10^\circ, \pm 15^\circ, \pm 20^\circ, \dots\})$$

We learn π by gradient descent on the log-loss



Lazy Averaging: Practical case

Model	f	$S_G^\eta f$
Log-loss	0.428	**0.187**
Accuracy	0.9516	**0.9532**

A theoretical result

$f : (X, d) \to \mathbb{R}$ L-Lipschitz

$$\left| S_G^\eta f(gx) - S_G^\eta f(x) \right|$$

$$\leq$$

$$L \underbrace{\inf_{\gamma \in \Pi(\eta, g\eta)} \mathbb{E}_{(h_1, h_2) \sim \gamma} d(h_1 x, h_2 x)}_{\mathrm{Was}(\eta, g\eta)}$$

where $g\eta(A) := \eta(A g^{-1})$.

Conclusions

- $d_\mathcal{R}$ & $S^\mathcal{R}_G$ are formalization attemps at 'approximate' symmetries
- Initial examples and results show promise
- Can we find new settings where to test $d_\mathcal{R}$ & $S^\mathcal{R}_G$?

theoretical & practical

¡Muchas Gracias

por su Atención!