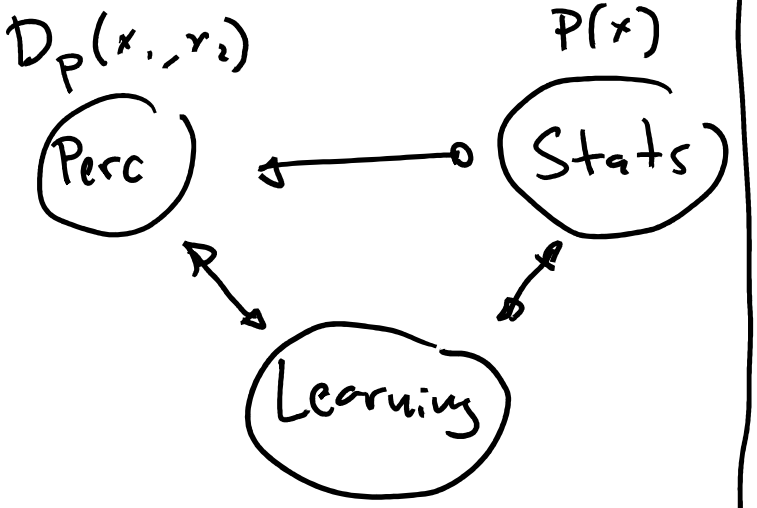
SOME OBSERVATIONS ON THE RELATION BETWEEN PERCEPTUAL DISTANCE AND STATISTICAL LEARNING IN AUTOENCODERS

A. Hepburn, V. Laparra, R. Santos, J. Ballé & J. Malo

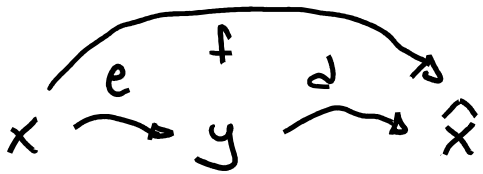
https://arxiv.org/abs/2106.04427

7th july 2021

ON THE RELATION BETWEEN STAT. LEARNING AND PERCEPTUAL DISTANCE



Autoencoders



$f = d \circ e$

* PERCEPTION: $x \xrightarrow{S} y$

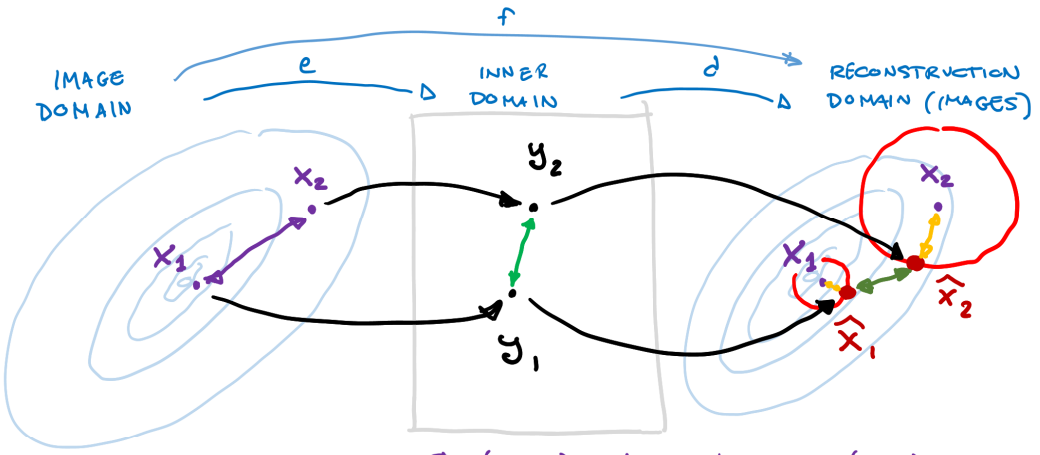
$$D_p(x_1, x_2) = |y_1 - y_2|_2 = |S(x_1) - S(x_2)|_2$$

* LEARNING

- Optimization $f_\theta(x)$

$$\mathcal{L}(\theta) = \mathbb{E}_x\left[D(x, f_\theta(x))\right] = \int p(x) D(x, f(x))\, dx$$

- Induced distances ?



EUCLIDEAN DISTANCE $D_e(x_1, x_2) = |x_1 - x_2|_2 = \text{RMSE}(x_1, x_2)$

INNER DISTANCE $D_i(x_1, x_2) = |y_1 - y_2|_2 = |e(x_1) - e(x_2)|_2$

RECONSTRUCT. DISTANCE $D_r(x_1, x_2) = |\hat{x}_1 - \hat{x}_2|_2 = |f(x_1) - f(x_2)|_2$

SELF RECONSTRUCT. DISTANCE $D_s(x, \hat{x}) = |x - \hat{x}|_2 = |x - f(x)|_2$

Proof of conjecture 1 (sensitivity of D_p) assuming S equalizes p(x)

Obs. 1 Sensitivity $D_p \sim p(x)$

$D_p = (\Delta s^T \Delta s)^{\frac{1}{2}}$ ↓ Δs

(A) $D_p(x_1, x_2) = |s(x_1) - s(x_2)|_2 = |s(x_1) - s(x_1 + \Delta x)|_2 = |s(x_1) - s(x_1) - \frac{\partial s}{\partial x}\Delta x|_2$

$D_p(x_1, x_2) = \left|\frac{\partial s}{\partial x}\Delta x\right|_2 = \left(\Delta x^T \left(\frac{\partial s}{\partial x}\right)^T \left(\frac{\partial s}{\partial x}\right) \Delta x\right)^{\frac{1}{2}} = \frac{\partial s}{\partial x}\Delta x \Rightarrow \boxed{\frac{\partial s}{\partial x} = \frac{D_p}{\Delta x}}$ *

($\Delta s^T \cdot \Delta s = D_p^2$; 1D; RMSE)

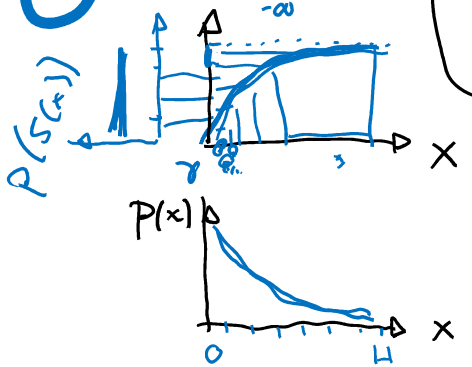
(B) Sensitivity $\frac{\partial D_p}{\partial x}$ ($s(x_2) - s(x_1)$)

$\frac{\partial D_p}{\partial x} = \frac{\partial}{\partial x}(\Delta s^T \Delta s)^{\frac{1}{2}} = \frac{2(\Delta s^T \Delta s)^{\frac{1}{2}}}{2(\Delta s^T \Delta s)} \cdot \frac{2 \cdot (\Delta s^T \Delta s)}{2 \Delta s} \cdot \frac{2(\Delta s)}{2x}$

$= \frac{1}{2}(\Delta s^T \cdot \Delta s)^{-\frac{1}{2}} \cdot 2 \Delta s^T \frac{\partial s}{\partial x} = \frac{1}{D_p} \cdot \Delta s^T \cdot \frac{\partial s}{\partial x}$

1D: $\frac{\partial D_p}{\partial x} = \frac{1}{D_p} \Delta s \frac{\partial s}{\partial x} \Rightarrow \boxed{\frac{\partial D_p}{\partial x} = \frac{\partial s}{\partial x}}$ **

(C) $s(x) = \int_{-\infty}^{x} p(x)^{\gamma} dx$



$s(x) = \int_{-\infty}^{x} p(x)^{\gamma} dx \Rightarrow \boxed{\frac{\partial s}{\partial x} = p(x)^{\gamma}}$ ***

Laughlin 81, Mcleod 01, Malo 06, Laparra 12, Laparra 15

$\frac{\partial D_p}{\partial x} : \frac{D_p}{RMSE} = p(x)^{\gamma}$

SENSITIVITY OF D_p

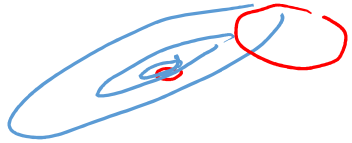
Obs. 2 Self reconst. distance $D_s \propto \frac{1}{p(x)}$

$$\mathcal{L}(\theta) = \int p(x)\, D(x, f_\theta(x))\, dx = \int p(x) \|x - f(x)\|_2 dx$$

$D \equiv$ RMSE / MSE

$D_s = \|x - \hat{x}\|_2$

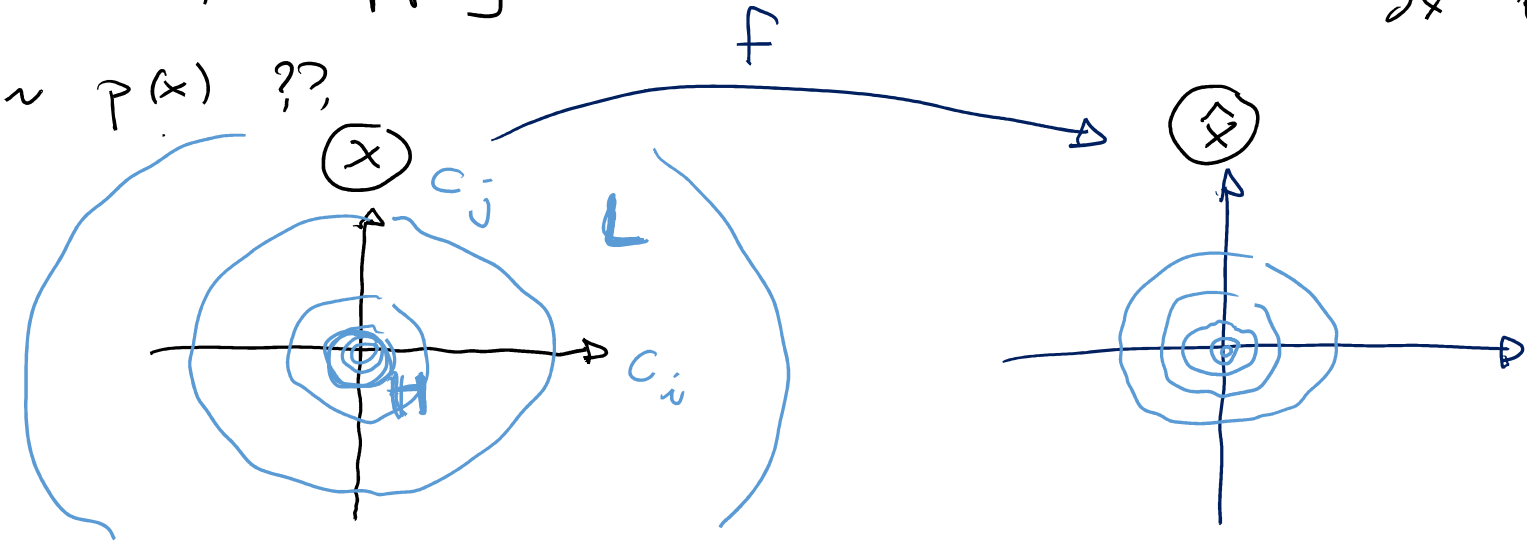
$$D_s \propto \frac{1}{p(x)}$$

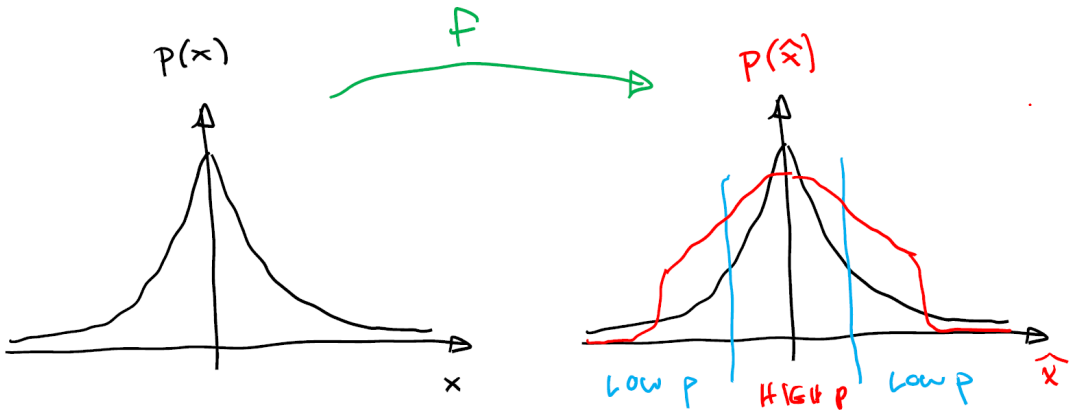


Obs. 3 sensit. of $D_r \propto p(x)$

why? ● A, B apply to f (as well as SJ) $\Rightarrow \frac{\partial D_r}{\partial x} = \frac{D_r}{RMSE} = \frac{\partial \epsilon}{\partial x}$

● $\frac{\partial f}{\partial x} \sim p(x)$??





f is "kind of" equalizing...

Not totally, but a little bit!!

$$\Rightarrow \frac{\partial f}{\partial x} \sim p(x)^{\gamma} \Rightarrow \frac{\partial D_r}{\partial x} \propto p(x)$$

Observation 3

Sensitivity of D_r

$$\frac{\partial D_r}{\partial x} = \frac{|f(x_1) - f(x_2)|_2}{|x_1 - x_2|_2} \propto p(x)^{\gamma}$$

Observation 4: sensitivity of D_i

$$\frac{\partial D_i}{\partial x} = \frac{|e(x_1) - e(x_2)|_2}{|x_1 - x_2|_2} \propto p(x)^{\gamma}$$

Conjectures / observations

(1) Sensitivity of D_p ——— $\frac{\partial D_p}{\partial x} \equiv \frac{|S(x_1) - S(x_2)|_2}{|x_1 - x_2|_2} = p(x)^r$

(2) Self reconstruct. distance D_s ——— $D_s(x, \hat{x}) \propto \frac{1}{p(x)}$

(3) Sensitivity of D_r ——— $\frac{\partial D_r}{\partial x} \equiv \frac{|f(x_1) - f(x_2)|_2}{|x_1 - x_2|_2} \propto p(x)$

(4) Sensitivity of D_i ——— $\frac{\partial D_i}{\partial x} \equiv \frac{|e(x_1) - e(x_2)|_1}{|x_1 - x_2|_2} \propto p(x)$

Conjectures / observations

(1) Sensitivity of D_p ——— $\frac{\partial D_p}{\partial x} \equiv \frac{|S(x_1) - S(x_2)|_2}{|x_1 - x_2|_2} = p(x)^r$

Equalization Hypothesis

Laughlin 81, Mcleod 01, Malo 06, Laparra 12, Laparra 15

Malo 20 $I(x, S(x)) = \sum_i h(S(x)_i) - T(S(x)) - h(u)$

Equalization is good. Redundancy & Noise are bad!

(2) Self reconstruct. distance D_s ——— $D_s(x, \hat{x}) \propto \frac{1}{p(x)}$

(3) Sensitivity of D_r ——— $\frac{\partial D_r}{\partial x} \equiv \frac{|f(x_1) - f(x_2)|_2}{|x_1 - x_2|_2} \propto p(x)$

(4) Sensitivity of D_i ——— $\frac{\partial D_i}{\partial x} \equiv \frac{|e(x_1) - e(x_2)|_1}{|x_1 - x_2|_2} \propto p(x)$

Conjectures / observations

(1) Sensitivity of D_p ——— $\frac{\partial D_p}{\partial x} \equiv \frac{|S(x_1) - S(x_2)|_2}{|x_1 - x_2|_2} = p(x)^r$

Equalization Hypothesis

Laughlin 81, Mcleod 01, Malo 06, Laparra 12, Laparra 15

Malo 20 $I(x, S(x)) = \sum_i h(S(x)_i) - T(S(x)) - h(u)$

Equalization is good. Redundancy & Noise are bad!

(2) Self reconstruct. distance D_s ——— $D_s(x, \hat{x}) \propto \frac{1}{p(x)}$

$\mathcal{L}(\theta) = \int p(x)\, D(x, f_\theta(x))\, dx$

(3) Sensitivity of D_r ——— $\frac{\partial D_r}{\partial x} \equiv \frac{|f(x_1) - f(x_2)|_2}{|x_1 - x_2|_2} \propto p(x)$

(4) Sensitivity of D_i ——— $\frac{\partial D_i}{\partial x} \equiv \frac{|e(x_1) - e(x_2)|_1}{|x_1 - x_2|_2} \propto p(x)$

Conjectures / observations

(1) Sensitivity of D_p ——— $\frac{\partial D_p}{\partial x} \equiv \frac{|S(x_1) - S(x_2)|_2}{|x_1 - x_2|_2} = p(x)^\gamma$

Equalization Hypothesis

Laughlin 81, Mclead 01, Malo 06, Laparra 12, Laparra 15

[Malo 20 $I(x, S(x)) = \sum_i h(S(x)_i) - T(S(x)) - h(u)$

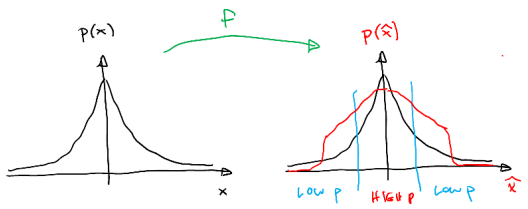
Equalization is good. Redundancy & Noise are bad!]

(2) Self reconstruct. distance D_s ——— $D_s(x, \hat{x}) \propto \frac{1}{p(x)}$

$\mathcal{L}(\theta) = \int p(x)\, D(x, f_\theta(x))\, dx$

(3) Sensitivity of D_r ——— $\frac{\partial D_r}{\partial x} \equiv \frac{|f(x_1) - f(x_2)|_2}{|x_1 - x_2|_2} \propto p(x)$

f kind of equalizes



(4) Sensitivity of D_i ——— $\frac{\partial D_i}{\partial x} \equiv \frac{|e(x_1) - e(x_2)|_1}{|x_1 - x_2|_2} \propto p(x)$

Conjectures / observations

(1) Sensitivity of D_p ——— $\frac{\partial D_p}{\partial x} \equiv \frac{|S(x_1) - S(x_2)|_2}{|x_1 - x_2|_2} = p(x)^\gamma$

Equalization Hypothesis

Laughlin 81, Mcleod 01, Malo 06, Laparra 12, Laparra 15 [Malo 20 $I(x, S(x)) = \sum_i h(S(x)_i) - T(S(x)) - h(u)$

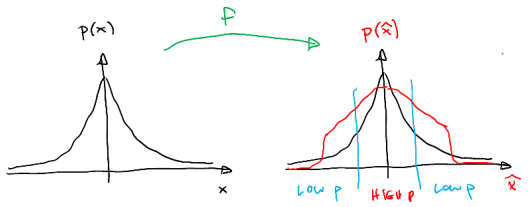
Equalization is good. Redundancy & Noise are bad!]

(2) Self reconstruct. distance D_s ——— $D_s(x, \hat{x}) \propto \frac{1}{p(x)}$

$\mathcal{L}(\theta) = \int p(x)\, D(x, f_\theta(x))\, dx$

(3) Sensitivity of D_r ——— $\frac{\partial D_r}{\partial x} \equiv \frac{|f(x_1) - f(x_2)|_2}{|x_1 - x_2|_2} \propto p(x)$

f kind of equalizes



(4) Sensitivity of D_i ——— $\frac{\partial D_i}{\partial x} \equiv \frac{|e(x_1) - e(x_2)|_1}{|x_1 - x_2|_2} \propto p(x)$

the inner domain maybe is equalizing

Conjectures / observations

$D_p \equiv$ Div. Norm. (NLPD) / MS-SSIM

$p(x) \equiv$ PixelCNN+ https://arxiv.org/abs/1701.05517

(1) Sensitivity of D_p ——— $\frac{\partial D_p}{\partial x} \equiv \frac{|S(x_1) - S(x_2)|_2}{|x_1 - x_2|_2} = p(x)^\gamma$

Equalization Hypothesis

Laughlin 81, Mcleod 01, Malo 06, Laparra 12, Laparra 15

Malo 20 $I(x, S(x)) = \sum_i h(S(x)_i) - T(S(x)) - h(u)$

Equalization is good, Redundancy & Noise are bad!

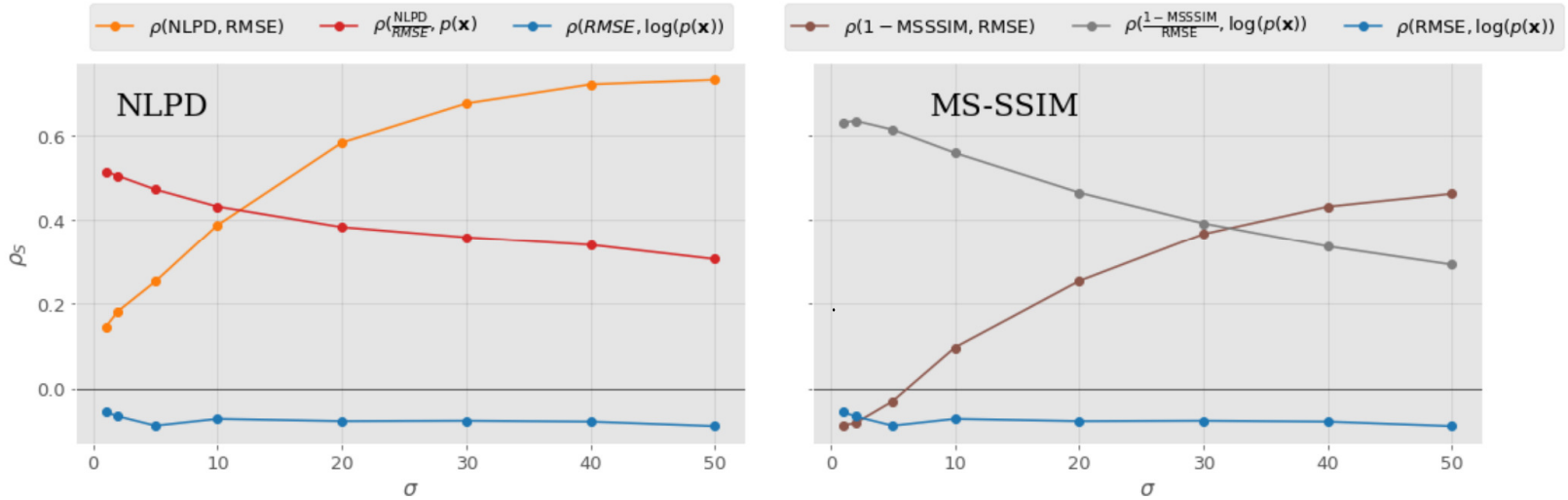


Figure 2: **Illustration of Oberservation 1.** Spearman correlations ρ_S between the sensitivity of the perceptual distances NLPD and MS-SSIM and $\log(p(\mathbf{x}))$ (in red/gray). Distances are computed between $\mathbf{x}$, and a distorted version with additive Gaussian noise, $\mathbf{x}+\Delta\mathbf{x}$, with deviation σ. Correlation of RMSE with perceptual distortions (in orange/brown) and of RMSE with $\log(p(\mathbf{x}))$ (in blue) are included for comparison. MS-SSIM is a similarity index, so 1-(MS-SSIM) is a distortion measure.

Conjectures / observations $\mathcal{L} = E_x\left[H(G) + \lambda D_e(x, \hat{x})\right]$ https://arxiv.org/pdf/2007.03034.pdf

(2) Self reconstruct. distance D_S —— $D_s(x, \hat{x}) \propto \frac{1}{p(x)}$ Empirical Risk Minimiz

(3) Sensitivity of D_r —— $\frac{\partial D_r}{\partial x} \equiv \frac{|f(x_1) - f(x_2)|_2}{|x_1 - x_2|_2} \propto p(x)$ Equalizat. ?

(4) Sensitivity of D_i —— $\frac{\partial D_i}{\partial x} \equiv \frac{|e(x_1) - e(x_2)|_1}{|x_1 - x_2|_2} \propto p(x)$ Equalizat ?

https://github.com/tensorflow/compression/blob/master/models/toy_sources/toy_sources.ipynb

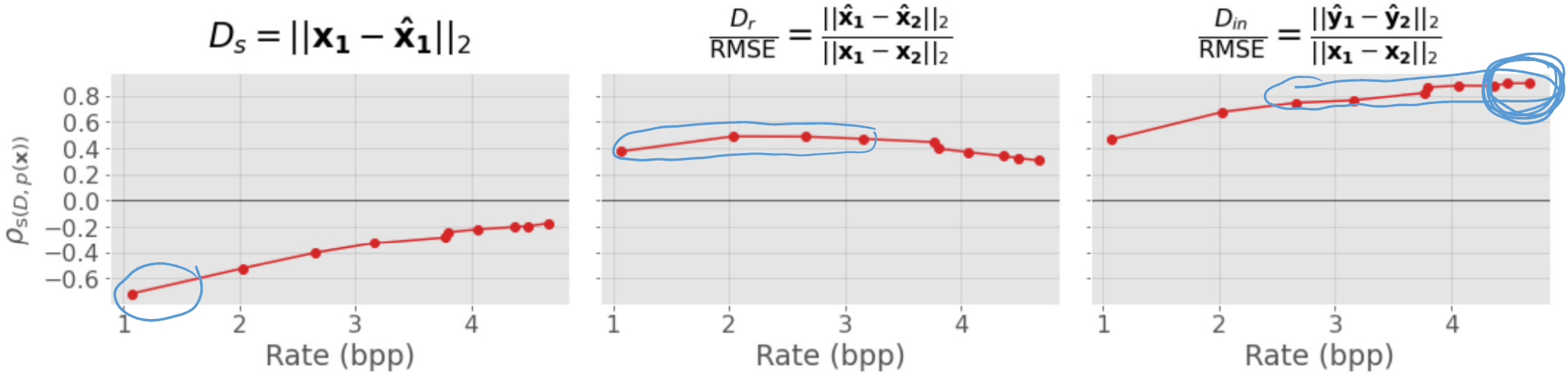


Figure 3: **Illustrating Observations 2, 3, and 4.** Spearman correlations $\rho_S(D, p(\mathbf{x_1}))$ between different distances and the probability of point $\mathbf{x_1}$ are shown. Each point corresponds to the correlation for one autoencoder trained for a particular Rate regime.

Conjectures / Observations

(5) The induced distances D_r and D_i are related to D_p

(1) Sensitivity of D_p ——— $\frac{\partial D_p}{\partial x} \equiv \frac{|S(x_1) - S(x_2)|_2}{|x_1 - x_2|_2} = p(x)^\gamma$

(3) Sensitivity of D_r ——— $\frac{\partial D_r}{\partial x} \equiv \frac{|f(x_1) - f(x_2)|_2}{|x_1 - x_2|_2} \propto p(x)$

(4) Sensitivity of D_i ——— $\frac{\partial D_i}{\partial x} \equiv \frac{|e(x_1) - e(x_2)|_1}{|x_1 - x_2|_2} \propto p(x)$

$\Rightarrow$ $D_i \propto D_p$ (MOS)

$D_r \propto D_p$ (MOS)

(6) As D_p is related to $p(x)$, $\Rightarrow$ using $\mathcal{L}(\theta) = \int p(x)\, D_p(x, f(x))\, dx$ implies considering $p(x)$ twice

(6.A) GOOD When your data is poor using D_p regularizes

(6.B) BAD When training with D_p performance may be poor in $p(x)$

Conjectures / Observations

https://arxiv.org/pdf/2007.03034.pdf

(5) The induced distances D_r and D_i are related to D_p

(2) Sensitivity of D_p —— $\frac{\partial D_p}{\partial x} = \frac{|S(x_1) - S(x_2)|_2}{|x_1 - x_2|_2} = p(x)^\gamma$

(3) Sensitivity of D_r —— $\frac{\partial D_r}{\partial x} = \frac{|f(x_1) - f(x_2)|_2}{|x_1 - x_2|_2} \propto p(x)$

(4) Sensitivity of D_i —— $\frac{\partial D_i}{\partial x} = \frac{|e(x_1) - e(x_2)|_1}{|x_1 - x_2|_2} \propto p(x)$

$\Rightarrow$ $D_i \propto D_p$ (MOS)

$D_r \propto D_p$ (MOS)

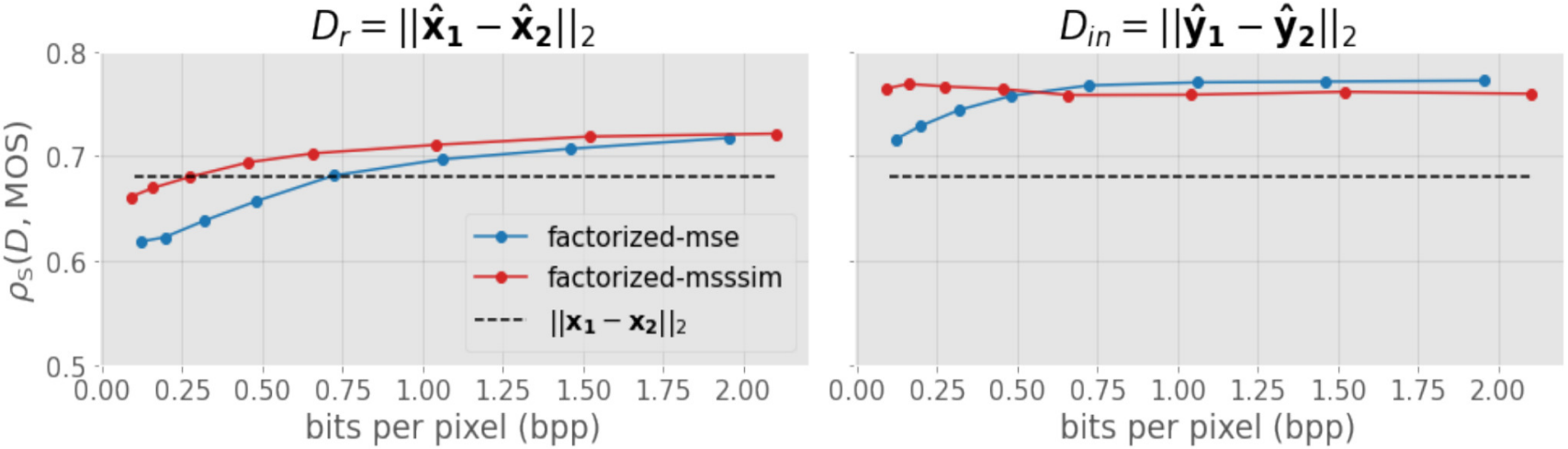


Figure 4: **Checking Observation 5**. Spearman correlations ρ_S between induced distances (D_r or D_{in}) and mean opinion score (MOS) for images from TID 2013 dataset [33]. Pretrained compressive autoencoders at different bitrates were used. *factorized-mse* denotes networks trained using MSE $\mathcal{D} = ||\mathbf{x}_1 - \mathbf{x}_2||_2^2$ in Eq. 3, and *factorized-msssim* networks use $\mathcal{D} = 1 - (\text{MS-SSIM})$ in Eq. 3.

Conjectures / Observations

(6) As D_p is related to $p(x)$, $\Rightarrow$ using $\mathcal{L}(\theta) = \int p(x)\, D_p(x, f(x))\, dx$ implies considering $p(x)$ twice

(6.A) GOOD When your data is poor using D_p regularizes

(a) $\min_{u(\mathbf{x})} ||\mathbf{x} - \hat{\mathbf{x}}||_2$ (b) $\min_{u(\mathbf{x})} \text{NLPD}(\mathbf{x}, \hat{\mathbf{x}})$ (c) $\min_{u(\mathbf{x})} 1 - \text{MS-SSIM}(\mathbf{x}, \hat{\mathbf{x}})$

Figure 6: **Checking observation 8.** Decoded image (compressed at 0.25bpp) encoded with networks trained using data from a uniform distribution and Euclidean vs perceptual losses.

Conjectures / Observations

(6) As D_p is related to $p(x)$, $\Rightarrow$ using $\mathcal{L}(\theta) = \int p(x)\, D_p(x, f(x))\, dx$ implies considering $p(x)$ twice

(6.A) GOOD When your data is poor using D_p regularizes

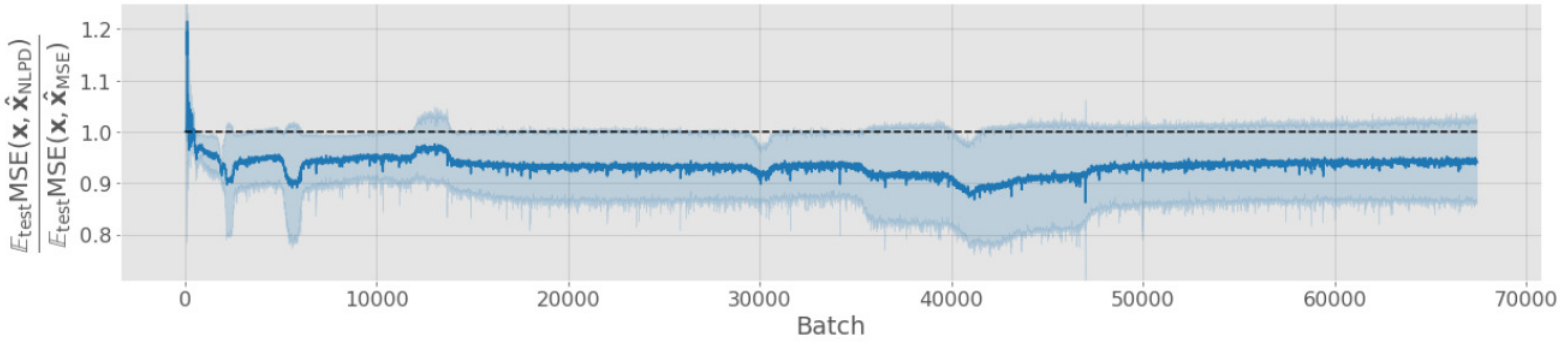


Figure 16: Gain in using NLPD over MSE as a loss function evaluated in terms of MSE loss on test set (Kodak dataset) using batch size of 1 and a small learning rate, fixing random seeds. $\hat{\mathbf{x}}_{\text{NLPD}}$ denotes the reconstruction of $\mathbf{x}$ with a network optimized for NLPD, and $\hat{\mathbf{x}}_{\text{MSE}}$ for a network optimized for MSE. The mean (solid line) and standard deviation (solid fill) was taken over 5 runs with different random seeds, i.e. different network initialization and training image ordering. The dashed line represents if the two networks had the same expected MSE on the test set.

Conjectures / Observations

(6) As D_p is related to $p(x)$, $\Rightarrow$ using $\mathcal{L}(\theta) = \int p(x)\, D_p(x, f(x))\, dx$ implies considering $p(x)$ twice

(6.B) BAD When training with D_p performance may be poor in $p(x)$

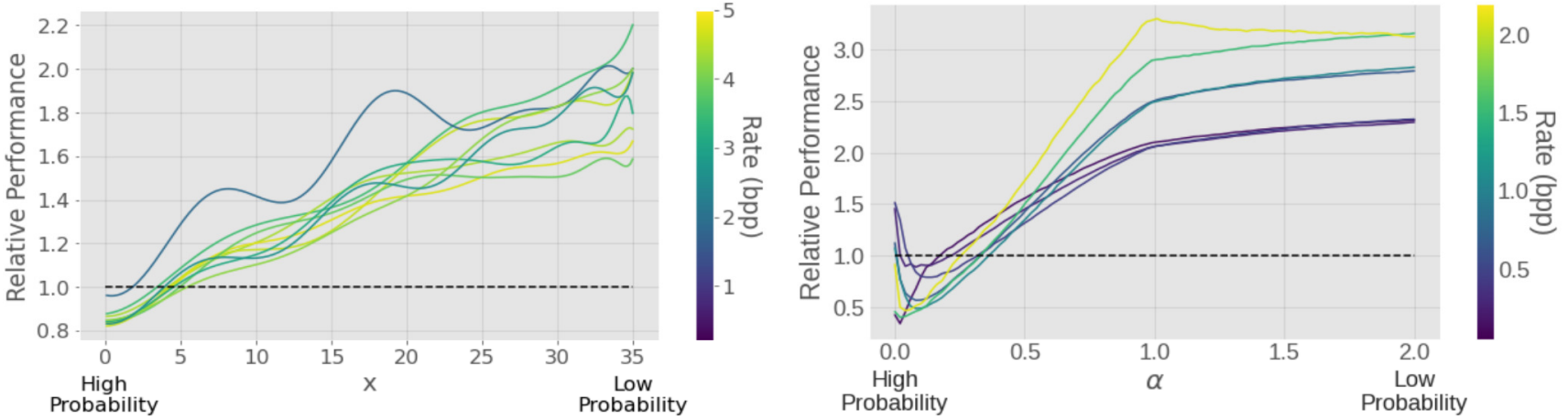


Figure 5: **Checking observation 7.** Relative performance of networks for samples along a line through the support of the respective distributions. Left: networks trained with $\mathcal{D} = p(\mathbf{x}) \cdot ||\mathbf{x}_1 - \mathbf{x}_2||_2^2$ in Eq. 3 divided by performance of networks trained with $\mathcal{D} = ||\mathbf{x}_1 - \mathbf{x}_2||_2^2$ on the 2D Student-t and evaluated using samples along x-axis (see Appendix B). Right: networks trained with $\mathcal{D} = 1 - (\text{MS-SSIM})$ in Eq. 3 divided by performance of networks trained with $\mathcal{D} = ||\mathbf{x}_1 - \mathbf{x}_2||_2^2$ for image coding using the Tensorflow Compression package.

CONCLUSIONS

(0) We considered the relations between NATURAL PERCEPTION / ARTIFICIAL AUTOENCODERS

- D_p Perceptual distance
- D_r Reconstruction distance
- D_i Inner domain distance
- D_s Self-reconstr. distance

and PROBABILITY $p(x)$

(1) Assuming the equalization property of the perceptual response, S, we proved $\frac{\partial D_p}{\partial x} = \frac{|S(x_1) - S(x_2)|_2}{|x_1 - x_2|_2} = p(x)^{\gamma}$

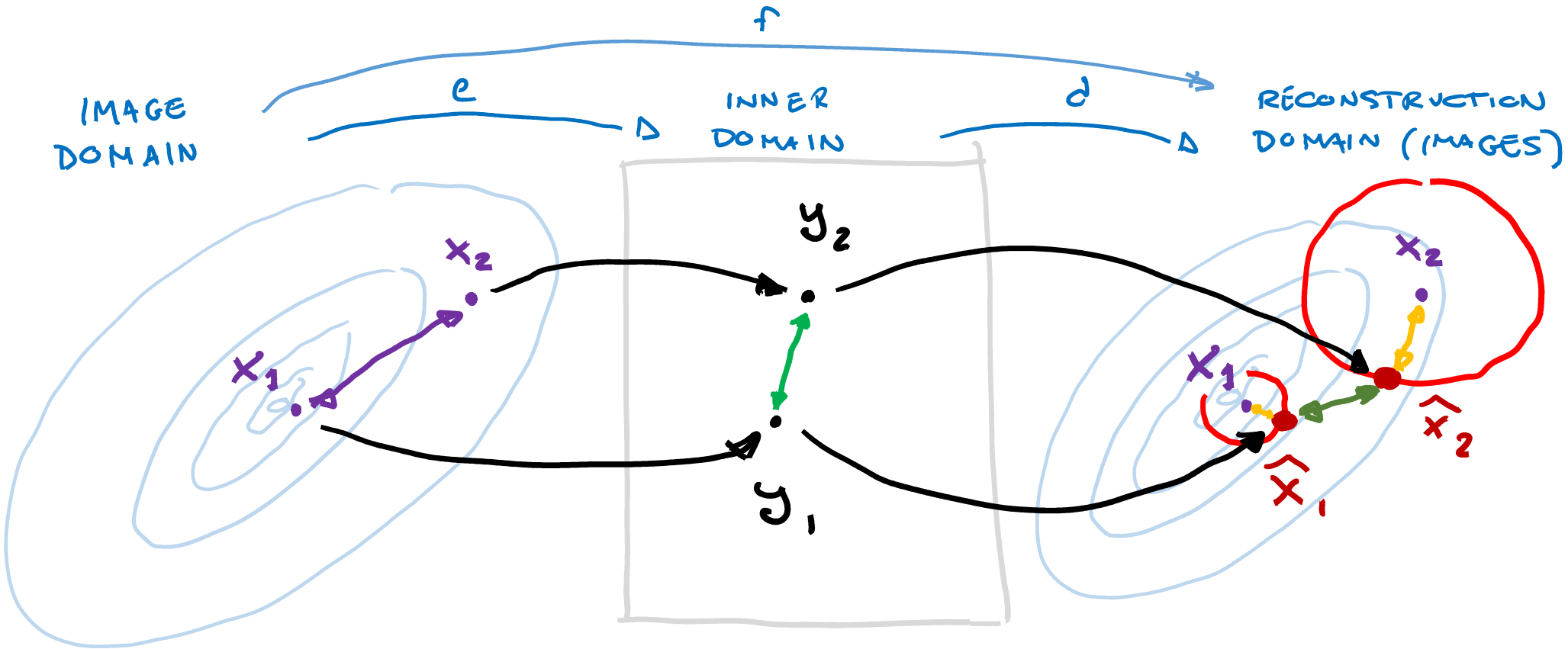
(2) Considering $\mathcal{L}(\theta) = \int p(x)\, D(x, f_\theta(x))\, dx$ we conjecture $D_s \sim p(x)^{-1}$

(3) Considering that the autoencoder response, $f(x)$, flattens $p(\hat{x})$ we conjecture $\frac{\partial D_r}{\partial x} = \frac{|f(x_1) - f(x_2)|_2}{|x_1 - x_2|_2} \sim p(x)$

(4) Considering equalization in the inner domain we conjecture $\frac{\partial D_i}{\partial x} = \frac{|e(x_1) - e(x_2)|_2}{|x_1 - x_2|_2} \sim p(x)$

(5) Assuming (1), (3) and (4) we conjecture D_r and D_i are correlated with D_p

(6) Using D_p may imply taking into account $p(x)$ twice
- 6.a Positive regularization effect of D_p
- 6.b Negative effect in some regions



EUCLIDEAN DISTANCE $D_e(x_1, x_2) = |x_1 - x_2|_2 = \text{RMSE}(x_1, x_2)$

INNER DISTANCE $D_i(x_1, x_2) = |y_1 - y_2|_2 = |e(x_1) - e(x_2)|_2$

RECONSTRUCT. DISTANCE $D_r(x_1, x_2) = |\hat{x}_1 - \hat{x}_2|_2 = |f(x_1) - f(x_2)|_2$

SELF RECONSTRUCT. DISTANCE $D_s(x, \hat{x}) = |x - \hat{x}|_2 = |x - f(x)|_2$