A geometric analysis of task-specific natural image statistics

Daniel Herrera-Esposito

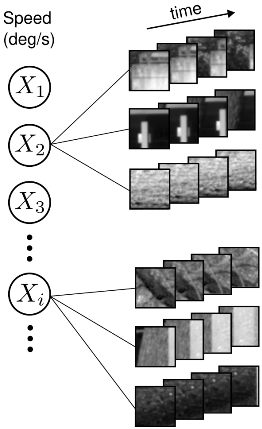
Johannes Burge Lab
Department of Psychology
University of Pennsylvania

Presentation Outline

- **Introduction: Task-specific natural image statistics (NIS)**
  - Conditioning image statistics on task variables
  - Useful for solving visual tasks
  - Draw a curve in SPD manifold
- **Part 1: Describing NIS curve geometry**
  - Choosing the right metric
  - Fit locally with geodesics
- **Part 2: Learning using NIS geometry**
  - Using distances in manifold as loss
  - Choosing the right metric
- **Part 3: Geometry across tasks**
  - Shape of curve across tasks, filters, and metrics

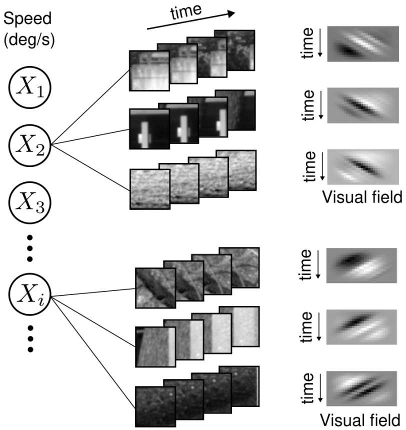
Task-specific natural image statistics

- Visual task: Estimating latent variable (X) from image
- Many natural scene patches for each X value



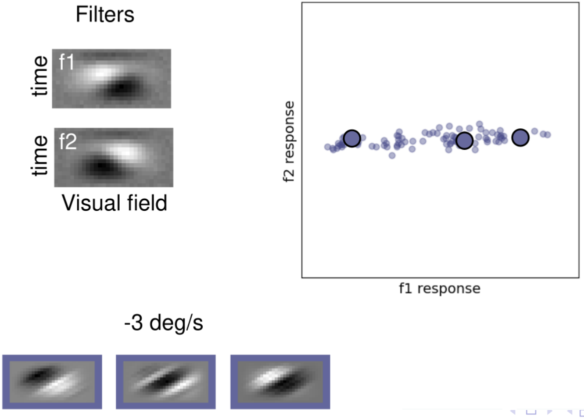
Task-specific natural image statistics

- Visual task: Estimating latent variable (X) from image
- Many natural scene patches for each X value



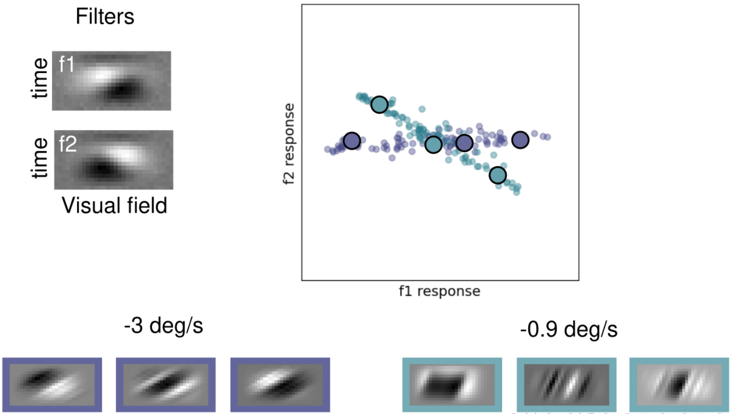
Task-specific natural image statistics

- Natural image variability for fixed X values
-



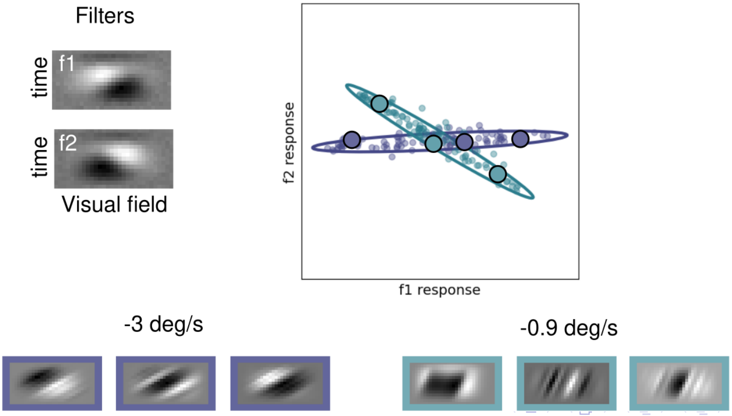
Task-specific natural image statistics

- Natural image variability for fixed X values
- Image feature statistics depend on X value



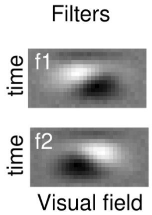
Task-specific natural image statistics

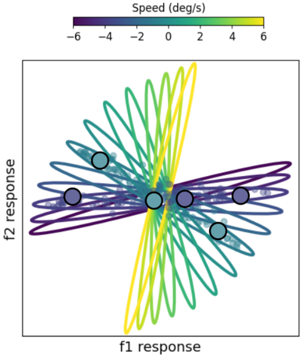
- Natural image variability for fixed X values
- Image feature statistics depend on X value



Task-specific natural image statistics

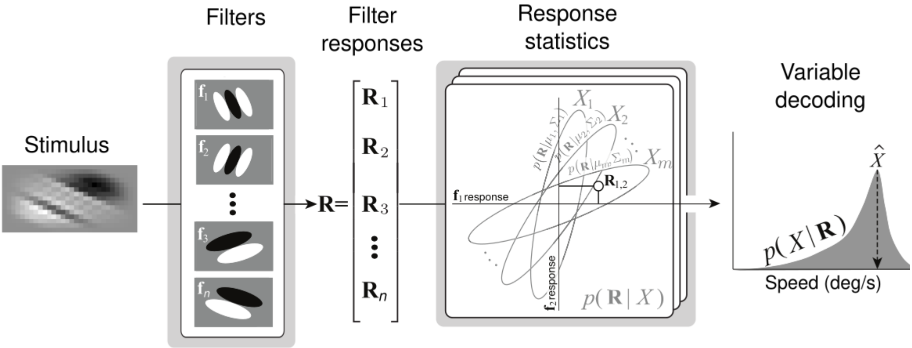
- Natural image variability for fixed X values
- Image feature statistics depend on X value





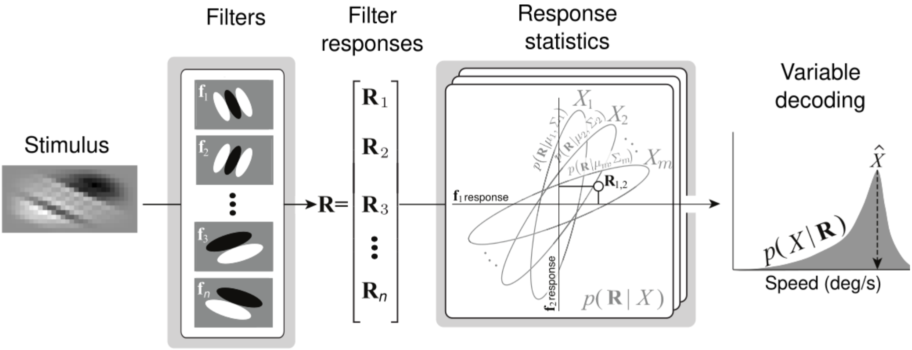
Task-specific natural image statistics

- Task-specific NIS for estimating X
-
-



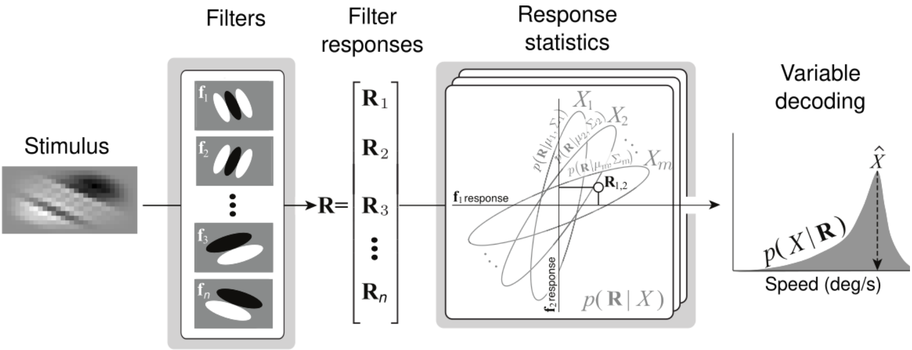
Task-specific natural image statistics

- Task-specific NIS for estimating X
- Ideal observer models use probabilistic decoding
-



Task-specific natural image statistics

- Task-specific NIS for estimating X
- Ideal observer models use probabilistic decoding
- Accuracy Maximization Analysis: Learn optimal linear filters for task



Task-specific natural image statistics

Accuracy Maximization Analysis has 3 steps:

1. **Preprocess stimuli** (<u>fixed</u>):
   Convert image to contrast: $\boldsymbol{s} = \frac{I - \bar{I}}{\bar{I}}$
   Add noise (γ) and normalize: $\boldsymbol{c} = \frac{\boldsymbol{s}+\gamma}{\|\boldsymbol{s}+\gamma\|}$, $\gamma \sim \mathcal{N}(0, \boldsymbol{I}\sigma_p^2)$

2. **Linear encoding** (<u>learnable</u>):

$$\boldsymbol{R} = \boldsymbol{f}^T \boldsymbol{c} + \boldsymbol{\lambda}$$

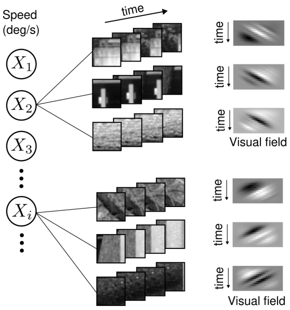
$\boldsymbol{c} \in \mathbb{R}^k$, $\boldsymbol{f} \in \mathbb{R}^{k \times n}$, $\boldsymbol{R} \in \mathbb{R}^n$, and $\boldsymbol{\lambda} \sim \mathcal{N}(0, \mathrm{I}\sigma_r^2)$

3. **Probabilistic decoding** (<u>determined by NIS</u>):

$$\hat{X} = \arg\max_{X_i} p(X_i | \boldsymbol{R})$$

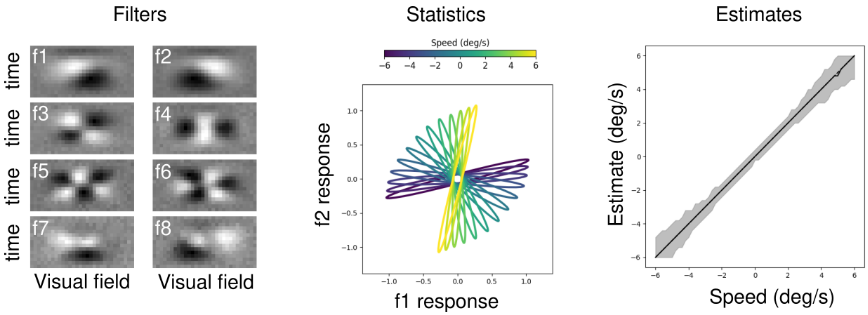
Task-specific natural image statistics

- Dataset composed of pairs $(\boldsymbol{s}_{ij}, X_i)$
- Finite number of X values: $\{X_1, \ldots, X_m\}$
- Filters are learned with loss $\mathcal{L}(\boldsymbol{R}_{ij}) = -\log p(X_i|\boldsymbol{R}_{ij})$
- We assume $p(\boldsymbol{R}|X_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ (empirically verified)



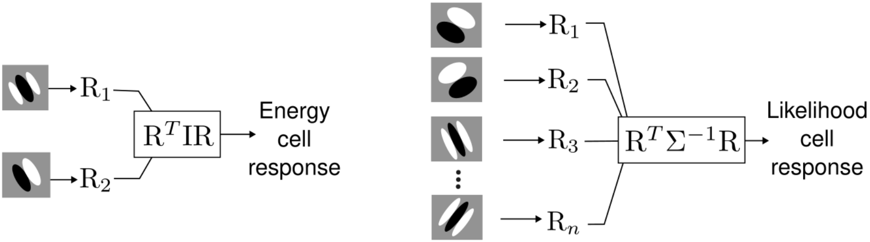
Task-specific natural image statistics

- Learning results:



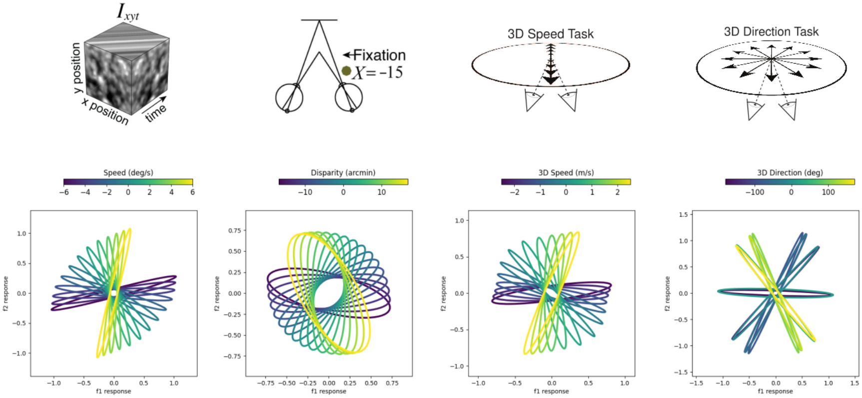
Task-specific natural image statistics

- Side note: Gaussian distribution implies quadratic combination of responses for decoding
- Biologically plausible



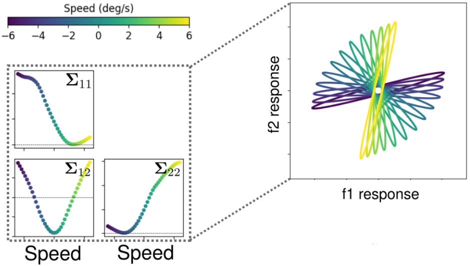
Task-specific natural image statistics

- Multiple tasks well approximated by zero-mean Gaussians



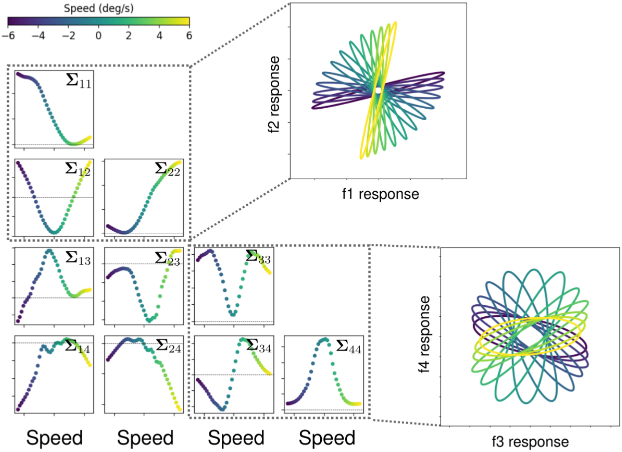
Geometric description of statistics

- $\Sigma(X)$: high-dimensional curve parametrized by X
- Constrained by NIS



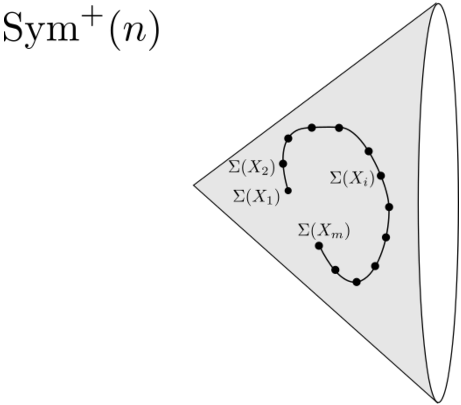
Geometric description of statistics

- $\Sigma(X)$: high-dimensional curve parametrized by X
- Constrained by NIS



Geometric description of statistics

- $\boldsymbol{\Sigma}(X)$ is a curve in SPDM manifold $\mathrm{Sym}^+(n)$
- What can we learn from this geometric perspective?



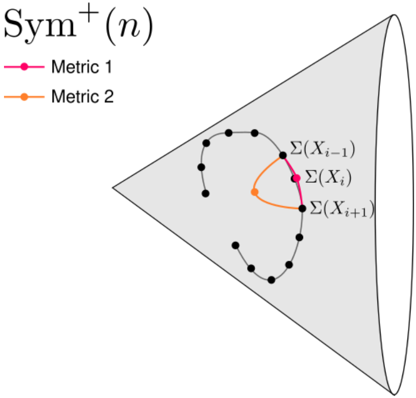
Geometric description of statistics

- First we need to specify a metric. Which one best fits the curve?

Metric	$d(\boldsymbol{A}, \boldsymbol{B})$
Euclidean	$\|\boldsymbol{A} - \boldsymbol{B}\|_F$
Affine-invariant	$\|\log(\boldsymbol{A}^{-\frac{1}{2}}\boldsymbol{B}\boldsymbol{A}^{-\frac{1}{2}})\|_F$
Bures-Wasserstein	$\left(\operatorname{tr}[\boldsymbol{A}] + \operatorname{tr}[\boldsymbol{B}] - 2\operatorname{tr}\left[\sqrt{\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{B}\boldsymbol{A}^{\frac{1}{2}}}\right]\right)^{\frac{1}{2}}$
Log-Euclidean	$\|\log(\boldsymbol{A}) - \log(\boldsymbol{B})\|_F$
Log-Cholesky	$\sqrt{\|\lfloor\boldsymbol{K}\rfloor - \lfloor\boldsymbol{L}\rfloor\|_F^2 + \|\log\mathbb{D}(\boldsymbol{K}) - \log\mathbb{D}(\boldsymbol{L})\|_F^2}$

Geometric description of statistics

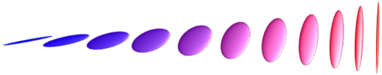
- Which geodesics best approximate the curve?
- For each $\boldsymbol{\Sigma}(X_i)$ compute mid-point between $\boldsymbol{\Sigma}(X_{i-1})$ and $\boldsymbol{\Sigma}(X_{i+1})$, compare to ground-truth



Metrics: Euclidean

Euclidean metric:

Distance	$d(\boldsymbol{A}, \boldsymbol{B}) = \|\boldsymbol{A} - \boldsymbol{B}\|_F$
Interpolation	$W(\boldsymbol{A}, \boldsymbol{B}, t) = (1 - t)\boldsymbol{A} + t\boldsymbol{B}$

- Invariant to orthogonal transformations
- Swelling in interpolation: 

Metrics: Affine-Invariant

Affine-invariant metric:

Distance	$d(\boldsymbol{A},\boldsymbol{B})^2 = \|\log\left(\boldsymbol{A}^{-\frac{1}{2}}\boldsymbol{B}\boldsymbol{A}^{-\frac{1}{2}}\right)\|_F = \sum_{i=1}^{n}(\log\lambda_i)^2$
Interpolation	$W(\boldsymbol{A},\boldsymbol{B},t) = \boldsymbol{A}^{\frac{1}{2}}\exp\{t\log\left(\boldsymbol{A}^{-\frac{1}{2}}\boldsymbol{B}\boldsymbol{A}^{-\frac{1}{2}}\right)\}\boldsymbol{A}^{\frac{1}{2}}$

λ_i generalized eigenvalues of (A, B): $\boldsymbol{A}\boldsymbol{v}_i = \lambda_i\boldsymbol{B}\boldsymbol{v}_i$

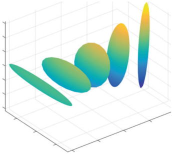
- Invariant to affine transformations
- Equals **Fisher information** metric for zero-mean Gaussians
- Flattening in interpolation: 

Metrics: Bures-Wasserstein

Bures-Wasserstein metric:

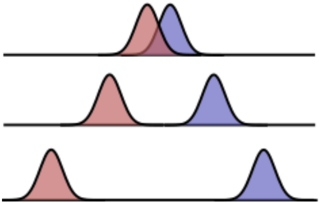
Distance	$d(\boldsymbol{A},\boldsymbol{B}) = \left(\operatorname{tr}[\boldsymbol{A}] + \operatorname{tr}[\boldsymbol{B}] - 2\operatorname{tr}\left[\left(\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{B}\boldsymbol{A}^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]\right)^{\frac{1}{2}}$
Interpolation	$W(\boldsymbol{A},\boldsymbol{B},t) = [(1-t)\boldsymbol{I} + t\boldsymbol{T}]\,\boldsymbol{A}\,[(1-t)\boldsymbol{I} + t\boldsymbol{T}]$ with $\boldsymbol{T} = \boldsymbol{B}^{\frac{1}{2}}\left[\boldsymbol{B}^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{B}^{\frac{1}{2}}\right]^{-\frac{1}{2}}\boldsymbol{B}^{\frac{1}{2}}$

- Invariant to orthogonal transformations
- Equals **optimal transport** distance between zero-mean Gaussians
- Geodesics are optimal transport plans
- Some swelling and flattening in interpolation:



Metrics: Intuition

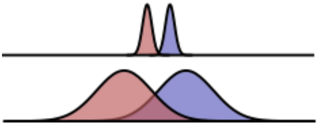
- Intuition of distributions distances



	Fisher-Rao (Affine-invariant)	Optimal transport (Bures-Wasserstein)
	Close	Close
	Far	Medium
	Far	Far

Metrics: Intuition

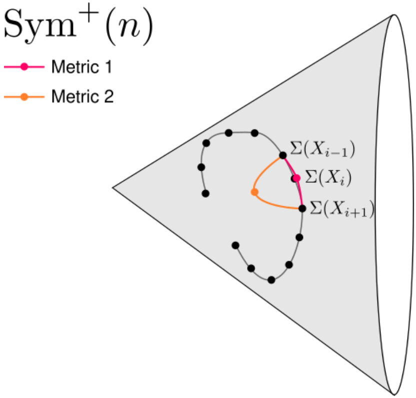
- Intuition of distributions distances



	Fisher-Rao (Affine-invariant)	Optimal transport (Bures-Wasserstein)
(top pair)	Far	Close
(bottom pair)	Close	Far

Geometric description of statistics

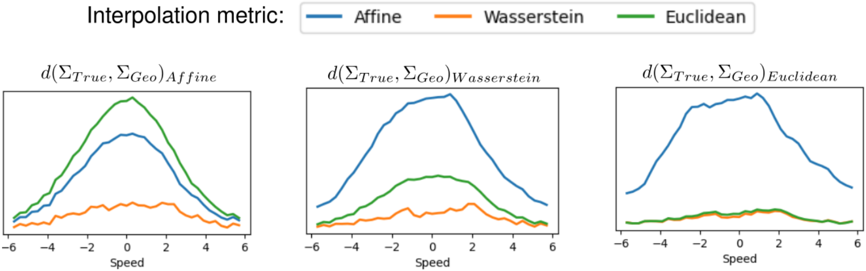
- Which geodesics best approximate the curve?
- For each $\boldsymbol{\Sigma}(X_i)$ compute mid-point between $\boldsymbol{\Sigma}(X_{i-1})$ and $\boldsymbol{\Sigma}(X_{i+1})$, compare to ground-truth



Geometric description of statistics

- Bures-Wasserstein (OT) geodesics best approximate the curve

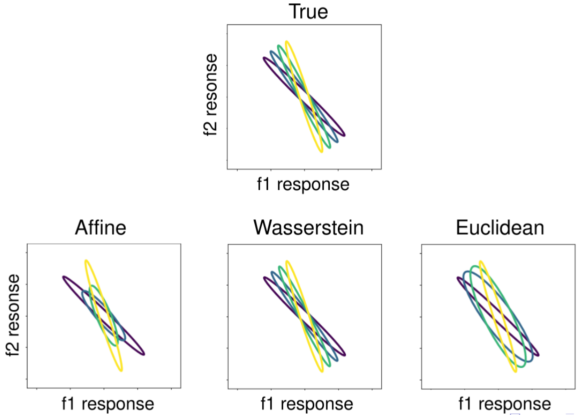
Interpolation errors:



Geometric description of statistics

- Bures-Wasserstein (OT) geodesics best approximate the curve

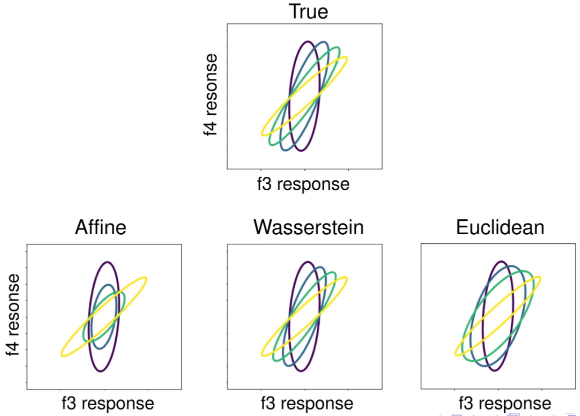
Interpolations examples:



Geometric description of statistics

- Bures-Wasserstein (OT) geodesics best approximate the curve

Interpolations examples:

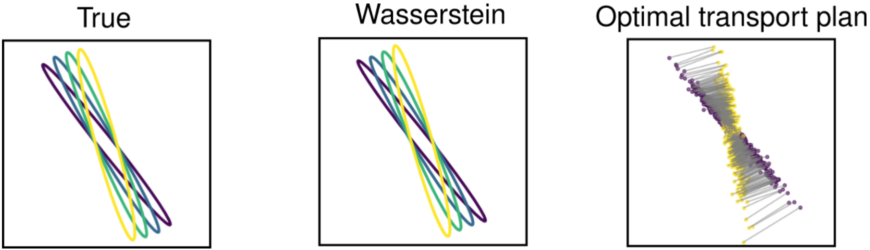


Geometric description of statistics

- Why Bures-Wasserstein geodesics fit best?
-

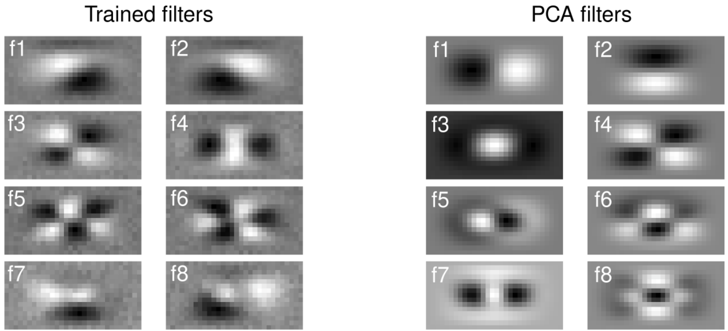
Geometric description of statistics

- Why Bures-Wasserstein geodesics fit best?
- Intuition: Optimal transport gets closest to ellipses rotation



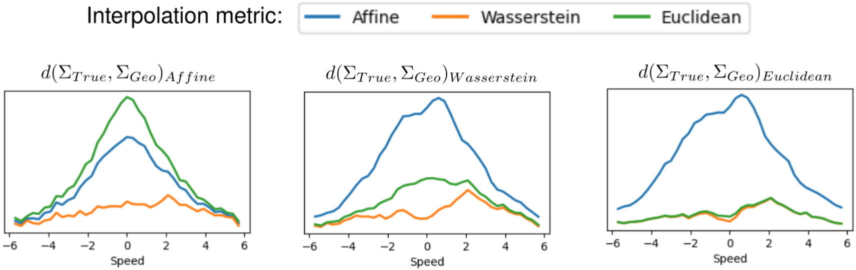
Geometric description of statistics

- Is this geometrical property (BW-like) a product of optimal filters?
- Do PCA filter statistics look different?



- BW best approximates PCA filter statistics curve

PCA interpolation errors:



Conclusions

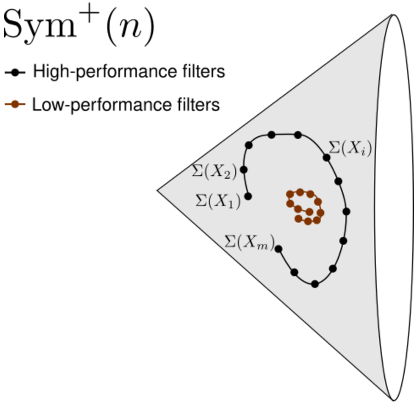
- Metric is important for covariance interpolation
- Geometry of NIS curve is best approximated by Bures-Wasserstein geodesics
- This geometry is maintained across filters, tasks (not shown) and levels of latent variable

Geometry as a training goal

- What insights can geometry provide?
- How does NIS geometry relate to visual tasks?
-

Geometry as a training goal

- What insights can geometry provide?
- How does NIS geometry relate to visual tasks?
- Intuition: More distant classes are more discriminable



Geometry as a training goal

Test this intuition:

- Use the pairwise distances as a loss to learn filters

$$\mathcal{L} = -\sum_{i=1}^{m-1} \sum_{j=i}^{m} d(\boldsymbol{\Sigma}(X_i), \boldsymbol{\Sigma}(X_j))$$

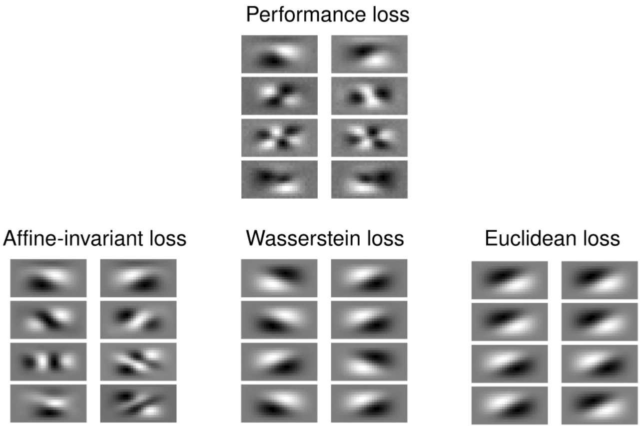
- Only requires stimulus statistics:

$$\Sigma(X_i) = \boldsymbol{f}^T \Psi(X_i) \boldsymbol{f}$$

$\Psi(X_i)$ is the covariance of $X = X_i$ stimuli

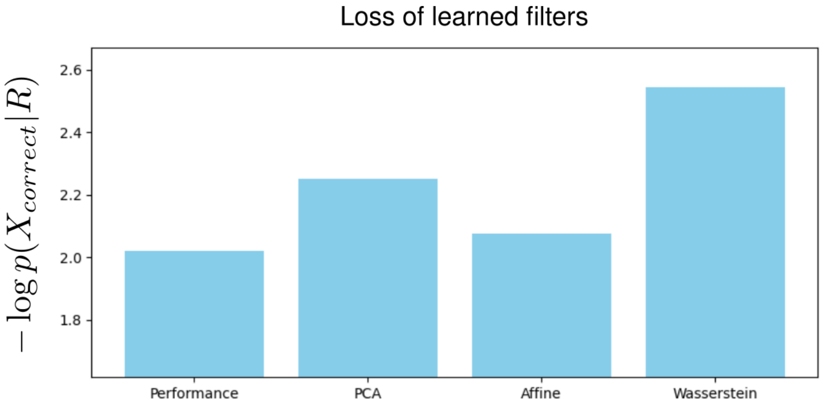
Geometry as a training goal

- Geometric learning is metric-dependent:
  - Affine-invariant loss learns good filters
  - Wasserstein and Euclidean losses do not



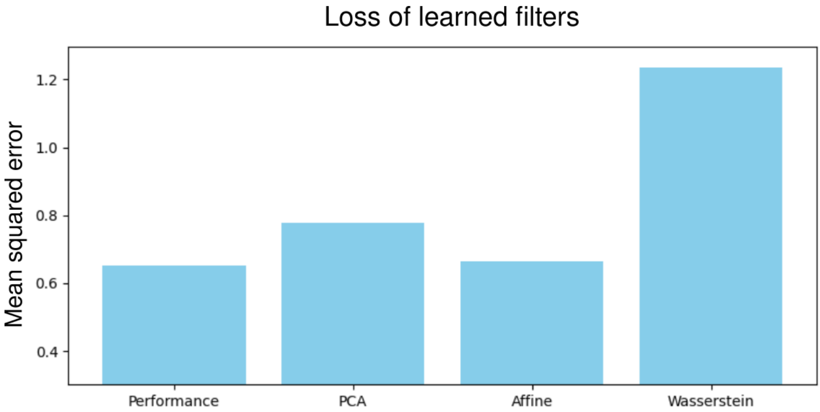
Geometry as a training goal

- Geometric learning is metric-dependent:
  - Affine-invariant loss learns good filters
  - Wasserstein and Euclidean losses do not



Geometry as a training goal

- Geometric learning is metric-dependent:
  - Affine-invariant loss learns good filters
  - Wasserstein and Euclidean losses do not



Geometry as a training goal

Why are some metrics better for training?

- Affine-Invariant metric measures local discriminability
- Affine-Invariant distance also relates to discriminability:

$$\boldsymbol{A} v_k = \lambda_k \boldsymbol{B} v_k$$

$$d(\boldsymbol{\Sigma}(X_i), \boldsymbol{\Sigma}(X_j)) = \sum_{k=1}^{n} (\log \lambda_k)^2$$

$$\frac{\mathbb{E}\left[(v_k^T R)^2 | X = X_i\right]}{\mathbb{E}\left[(v_k^T R)^2 | X = X_j\right]} = \frac{v_k^T \boldsymbol{\Sigma}(X_i) v_k}{v_k^T \boldsymbol{\Sigma}(X_j) v_k} = \lambda_k$$

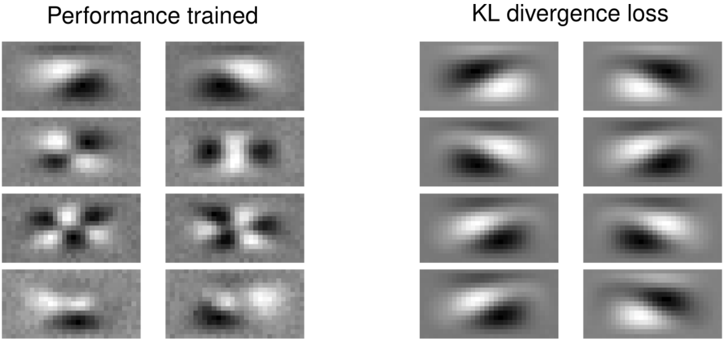
- Bures-Wasserstein is not invariant to scale

Geometry as a training goal

- KL divergence is related to Fisher-Rao metric
- It also relates to discriminability. Is it a good loss?
-

Geometry as a training goal

- KL divergence is related to Fisher-Rao metric
- It also relates to discriminability. Is it a good loss?
- KL divergence is not a good loss for training



Geometry as a training goal

Conclusions:

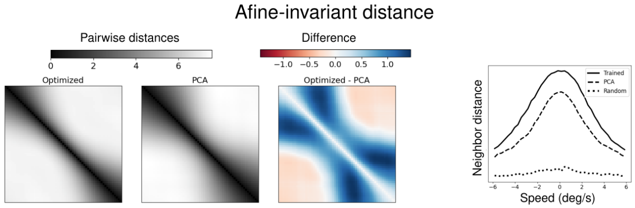
- Geometrical intuition can be used for training
- Choosing the right metric is important
- The best metric for training is not the same as for interpolation
- What makes a good metric for training?

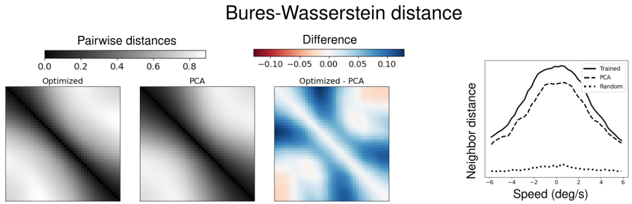
Curve shape

- Metric choice affects interpolation and learning
- Filters affect performance
- How do these affect curve shape?

Curve shape

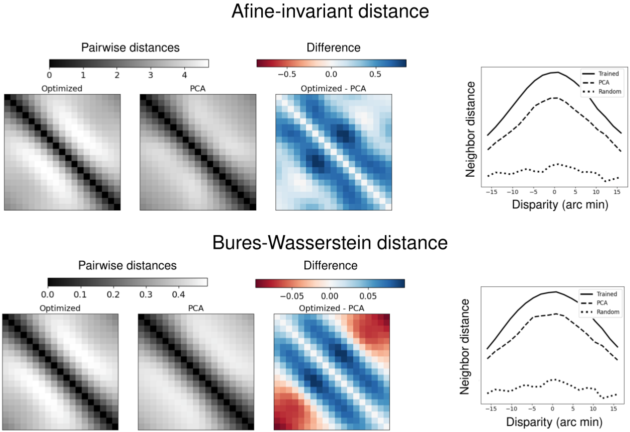
- Optimal filters generally (not always) farther than PCA filters
- Shape is similar across filters and metrics
- Shape changes with task





Curve shape

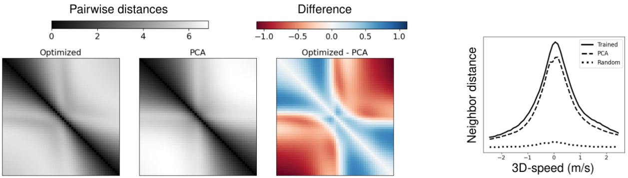
- Optimal filters generally (not always) farther than PCA filters
- Shape is similar across filters and metrics
- Shape changes with task



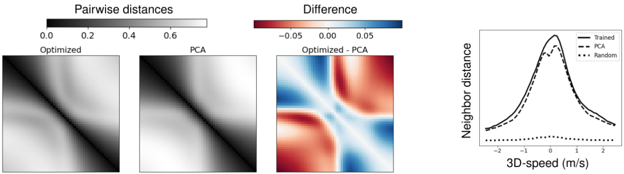
Curve shape

- Optimal filters generally (not always) farther than PCA filters
- Shape is similar across filters and metrics
- Shape changes with task

Afine-invariant distance

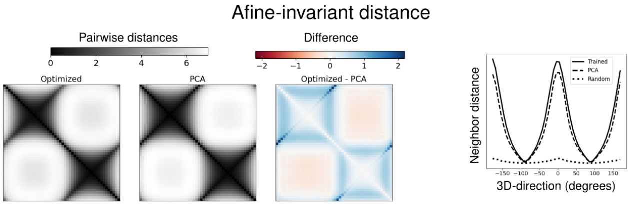


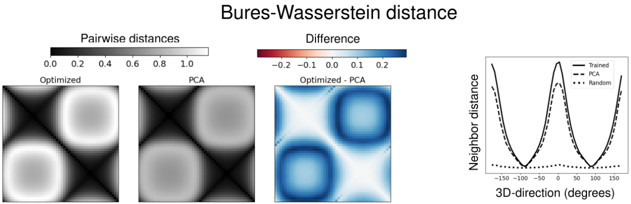
Bures-Wasserstein distance



Curve shape

- Optimal filters generally (not always) farther than PCA filters
- Shape is similar across filters and metrics
- Shape changes with task





Overview

- Task-specific NIS are a good system to explore geometric perspective on representations and learning
  - Zero-mean Gaussians have rich, well developed geometry
- Used SPDM manifold to interpolate and train
  - Chosing the right metric is important!
  - Bures-Wasserstein (OT) best for interpolation
  - Affine-Invariant (FR) best for training
- Geometry relates to performance and learning (given the right metric)
- Same results across tasks

Questions

- How generalizable are results for zero-mean Gaussian to other distributions?
- Why NIS covariances have this geometry?
- What makes a good metric for training?
- How does this relate to neural activity geometry? (e.g. is activity geometry something we can compare to real neurons?)
- Other geometric features as training objectives? (e.g. smoothness)

Thanks!

More information:

- Accuracy Maximization Analysis in Pytorch: `https://github.com/dherrera1911/accuracy_maximization_analysis`
- P. Jaini and J. Burge (2017). "**Linking normative models of natural tasks to descriptive models of neural response**". *Journal of Vision*
- J. Burge and P. Jaini (2017). "**Accuracy Maximization Analysis for Sensory-Perceptual Tasks: Computational Improvements, Filter Robustness, and Coding Advantages for Scaled Additive Noise**". *PLOS Computational Biology*
- D. Herrera-Esposito; J. Burge (2023). "**Optimal motion-in-depth estimation with natural stimuli**". *bioRxiv*