Ventral Visual Stream and Deep Networks

Matilde Marcolli and Doris Tsao

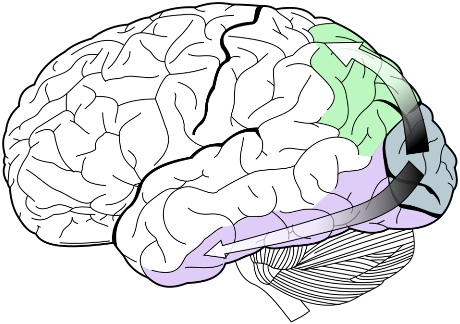
Ma191b Winter 2017
Geometry of Neuroscience

References for this lecture:

- Tomaso A. Poggio and Fabio Anselmi, *Visual Cortex and Deep Networks*, MIT Press, 2016
- F. Cucker, S. Smale, *On the mathematical foundations of learning*, Bulletin of the American Math. Society 39 (2001) N.1, 1–49.

Modeling Ventral Visual Stream via Deep Neural Networks

- Ventral Visual Stream considered responsible for object recognition abilities



dorsal (green) and ventral (purple) visual streams

- responsible for first $\sim$ 100msc time of processing visual information from initial visual stimulus to activation of inferior temporal cortex neurons

- mathematical model describing learning of *invariant representations* in the Ventral Visual Stream
- working hypothesis: main computational goal of the Ventral Visual Stream is compute neural representations of images that are invariant with respect to certain groups of transformations (mostly affine transformations: translations, rotations, scaling)
- model based on unsupervised learning
- far fewer examples are needed to train a classifier for recognition if using an *invariant representation*
- Gabor functions and frames optimal templates for simultaneously maximizing invariance with respect to translations and scaling
- architecture: hierarchy of Hubel–Wiesel modules

• a significant difference between (supervised) learning algorithms and functioning of the brain is that learning in the brain seems to require a very small number of labelled examples

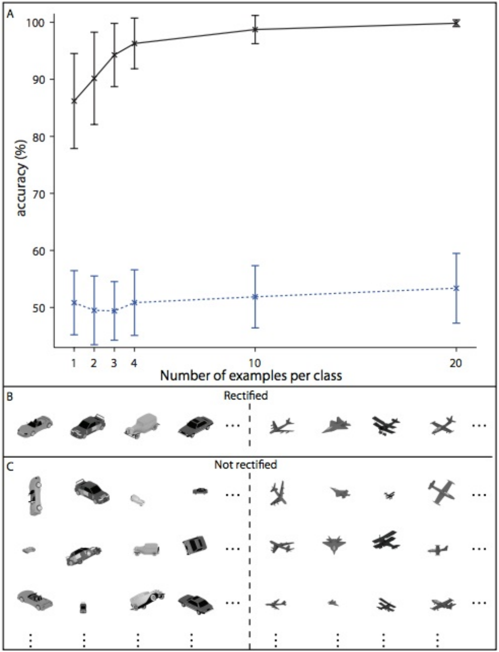
• conjecture: key to reducing sample complexity of object recognition is invariance under transformations

• two aspects: recognition and categorization

• for recognition it is clear that complexity is greatly increased by transformations (same objects seen from different perspectives, in different light conditions, etc.)

• for categorizations also (distinguishing between different classes of objects: cats/dogs, etc.) transformations can hide intrinsic characteristics of an object

• empirical evidence: accuracy of a classifier per number of examples greatly improved in the presence of an oracle that factors out transformations (solid curve, rectified; dashed, non-rectified)


A
100
90
80
70
60
50
accuracy (%)
1
2
3
4
10
20
Number of examples per class
B
Rectified
C
Not rectified

• order of magnitude for number of different object categorizations (e.g. distinguishable different types of dogs) smaller than magnitude for different viewpoints generated by group actions

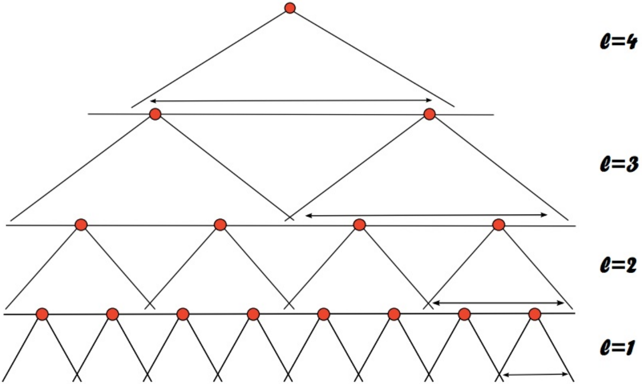
• reducing the variability by transformations makes greatly reduces the learning task complexity

• refer to sample complexity as number of examples needed for estimating a target function within an assigned error rate

• transform problem of distinguishing images into problem of distinguishing orbits under a given group action

Feedforward architecture in the ventral stream

- two main stages
  1. retinotopic areas computing a representation that is invariant under affine transformations
  2. approximate invariance to other object-specific dependencies, not described by group actions (parallel pathways)
- first stage realized through Gabor frames analysis
- overall model relies on a mathematical model of learning (Cucker-Smale)



- architecture layers: red circle = vector computed by one of the modules, double arrow = its receptive field; image at level zero (bottom), vector computed at top layer consists of invariant features (fed as input to a supervised learning classifier) biologically plausible algorithm (Hubel–Wiesel modules)

- two types of neurons roles:
  - *simple cells*: perform an operation of inner product with a template $t \in \mathcal{H}$ Hilbert space; a further non-linear operation (a threshold) is also applied
  - *complex cells*: aggregate the outputs of several simple cells

- steps: (assume G finite subgroup of affine transformations)
  1. unsupervised learning of group G by storing memory of orbit $G \cdot t = \{gt \,:\, g \in G\}$ of a set of templates $t \in \mathcal{H}$
  2. computation of invariant representation: new image $\mathcal{I} \in \mathcal{H}$ compute $\langle gt^k, \mathcal{I} \rangle$ for $g \in G$ and t^k, $k = 1, \dots, K$ templates and
$$\mu_h^k(\mathcal{I}) = \frac{1}{\#G} \sum_{g \in G} \sigma_h(\langle gt^k, \mathcal{I} \rangle)$$
     σ_h a set of nonlinear functions (e.g. threshold functions)

- computed $\mu_h^k(\mathcal{I})$ called signature of $\mathcal{I}$
- signature $\mu_h^k(\mathcal{I})$ clearly G-invariant
- Selectivity Question: how well does $\mu_h^k(\mathcal{I})$ distinguish different objects? different meaning $G \cdot \mathcal{I} \neq G \cdot \mathcal{I}'$

- Main Selectivity Result (Poggio-Anselmi)
  - want to be able to distinguish images within a given set of N images $\mathcal{I}$, with an error of at most a given $\epsilon > 0$
  - the signatures $\mu_h^k(\mathcal{I})$ can ϵ-approximate the distance between pairs among the N images with probability $1 - \delta$
  - provided that the number of templates used is at least

$$K > \frac{c}{\epsilon^2} \log \frac{N}{\delta}$$

- more detailed discussion of this statement below; main point: need of the order of $\log(N)$ templates to distinguish N images

- General problem: when two sets of random variables x, y are probabilistically related
  - relation described by probability distribution $P(x, y)$
  - some square loss problem (minimization problem)

$$E(f) = \int (y - f(x))^2 \, P(x, y) \, dx \, dy$$

  - distribution itself unknown, but minimize empirical error

$$E_N(f) = \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i))^2$$

    over a set of random sampled data points $\{(x_i, y_i)\}_{i=1,\dots,N}$
  - if f_N minimizes empirical error, want that the probability

$$\mathbb{P}(\| E(f_N) - E_N(f_N) \| > \epsilon)$$

    is sufficiently small
  - Problem depends on the function space where f_N lives

General setting

• F. Cucker, S. Smale, *On the mathematical foundations of learning*, Bulletin of the American Math. Society 39 (2001) N.1, 1–49.

- X compact manifold, $Y = \mathbb{R}^k$ (for simplicity $k = 1$), $Z = X \times Y$ with Borel measure ρ
- ξ random variable (real valued) on probability space (Z, ρ)
- expectation value and variance

$$\mathbb{E}(\xi) = \int_Z \xi \, d\rho, \qquad \sigma^2(\xi) = \mathbb{E}((\xi - \mathbb{E}(\xi))^2) = \mathbb{E}(\xi^2) - \mathbb{E}(\xi)^2$$

- function $f : X \to Y$, *least squares error* of f

$$\mathcal{E}(f) = \int_Z (f(x) - y)^2 \, d\rho$$

  measures average error incurred in using $f(x)$ as a model of the dependence between y and x
- Problem: how to minimize the error?

- conditional probability $\rho(y|x)$ (probability measure on Y)
- marginal probability $\rho_X(S) = \rho(\pi^{-1}(S))$ on X, with projection $\pi : Z = X \times Y \to X$
- relation between these measures

$$\int_Z \phi(x,y)\, d\rho = \int_X \left(\int_Y \phi(x,y)\, d\rho(y|x) \right) d\rho_X$$

- breaking of $\rho(x,y)$ into $\rho(y|x)$ and $\rho_X(S)$ is breaking of Z into input X and output Y

- regression function $f_\rho : X \to Y$

$$f_\rho(x) = \int_Y y \, d\rho(y|x)$$

- *assumption*: f_ρ is bounded
- for fixed $x \in X$ map Y to $\mathbb{R}$ via

$$y \mapsto y - f_\rho(x)$$

- expectation value is zero so variance

$$\sigma^2(x) = \int_Y (y - f_\rho(x))^2 \, d\rho(y|x)$$

- averaged variance

$$\sigma^2_\rho = \int_X \sigma^2(x) \, d\rho_X = \mathcal{E}(f_\rho)$$

  measures how "well conditioned" ρ is
- Note: in general ρ and f_ρ not known but ρ_X known

- error, regression, and variance:

$$\mathcal{E}(f) = \int_X ((f(x) - f_\rho(x))^2 + \sigma_\rho^2)\, d\rho_X$$

- What this says: σ_ρ^2 is a lower bound for the error $\mathcal{E}(f)$ for all f, and $f = f_\rho$ has the smallest possible error (which depends only on ρ)
- why identity holds:

$$\begin{aligned}
\mathcal{E}(f) &= \int_Z (f(x) - f_\rho(x) + f_\rho(x) - y)^2 \\
&= \int_X (f(x) - f_\rho(x))^2 + \int_X \int_Y (f_\rho(x) - y)^2 \\
&\qquad\qquad +2 \int_X \int_Y (f(x) - f_\rho(x))(f_\rho(x) - y) \\
&= \int_X (f(x) - f_\rho(x))^2 + \sigma_\rho^2.
\end{aligned}$$

Goal: "learn" (= find a good approximation for) f_ρ given random samples of Z

- $Z^N \ni z = ((x_1, y_1), \ldots, (x_N, y_N))$ sample set of points (x_i, y_i) independently drawn with probability ρ
- empirical error

$$\mathcal{E}_z(f) = \frac{1}{N} \sum_{i=1}^{N} (f(x_i) - y_i)^2$$

- for random variable ξ *empirical mean*

$$\mathbb{E}_z(\xi) = \frac{1}{N} \sum_{i=1}^{N} \xi(z_i, y_i)$$

- given $f : X \to Y$ take $f_Y : Z \to Y$ to be $f_Y : (x, y) \mapsto f(x) - y$

$$\mathcal{E}(f) = \mathbb{E}(f_Y^2), \qquad \mathcal{E}_z(f) = \mathbb{E}_z(f_Y^2)$$

Facts of Probability Theory

(quantitative versions of law of large numbers)

- ξ random variable on probability space Z with mean $\mathbb{E}(\xi) = \mu$ and variance $\sigma^2(\xi) - \sigma^2$
- Chebyshev: for all $\epsilon > 0$

$$\mathbb{P}\left\{ z \in Z^m \, : \, \left| \frac{1}{m} \sum_{i=1}^{m} \xi(z_i) - \mu \right| \geq \epsilon \right\} \leq \frac{\sigma^2}{m\epsilon^2}$$

- Bernstein: if $|\xi(z) - \mathbb{E}(\xi)| \leq M$ for almost all $z \in Z$ then $\forall \epsilon > 0$

$$\mathbb{P}\left\{ z \in Z^m \, : \, \left| \frac{1}{m} \sum_{i=1}^{m} \xi(z_i) - \mu \right| \geq \epsilon \right\} \leq 2 \exp\left(- \frac{m\epsilon^2}{2(\sigma^2 + \frac{1}{3} M\epsilon)} \right)$$

- Hoeffding:

$$\mathbb{P}\left\{ z \in Z^m \, : \, \left| \frac{1}{m} \sum_{i=1}^{m} \xi(z_i) - \mu \right| \geq \epsilon \right\} \leq 2 \exp\left(- \frac{m\epsilon^2}{2M^2} \right)$$

Defect Function of $f : X \to Y$

$$L_z(f) := \mathcal{E}(f) - \mathcal{E}_z(f)$$

discrepancy between error and empirical error (only $\mathcal{E}_z(f)$ measured directly)

- estimate of defect if $|f(x) - y| \leq M$ almost everywhere, then $\forall \epsilon > 0$, with σ^2 variance of f_Y^2

$$\mathbb{P}\{z \in Z^m \,:\, |L_z(f)| \leq \epsilon\} \geq 1 - 2\epsilon \exp\left(- \frac{m\epsilon^2}{2(\sigma^2 + \frac{1}{3}M^2\epsilon)} \right)$$

- from previous Bernstein estimate taking $\xi = f_Y^2$
- when is $|f(x) - y| \leq M$ a.e. satisfied? e.g. for $M = M_\rho + P$

$$M_\rho = \inf\{\bar{M} \,:\, \{(x,y) \in Z \,:\, |y - f_\rho(x)| \geq \bar{M}\} \text{ measure zero } \}$$

$$P \geq \sup_{x \in X} |f(x) - f_\rho(x)|$$

Hypothesis Space

- a learning process requires a datum of a class of functions (hypothesis space) within which the best approximation for f_ρ
- $C(X)$ algebra of continuous functions on topological space X
- $\mathcal{H} \subset C(X)$ compact subset (not necessarily subalgebra)
- look for minimizer (not necessarily unique)

$$f_{\mathcal{H}} = \mathrm{argmin}_{f\in\mathcal{H}} \int_Z (f(x)-y)^2$$

because $\mathcal{E}(f) = \int_X (f-f_\rho)^2 + \sigma_\rho^2$ also minimizer

$$f_{\mathcal{H}} = \mathrm{argmin}_{f\in\mathcal{H}} \int_X (f-f_\rho)^2$$

- continuity: if for $f \in \mathcal{H}$ have $|f(x)-y| \leq M$ a.e., bounds

$$|\mathcal{E}(f_1) - \mathcal{E}(f_2)| \leq 2M \|f_1 - f_2\|_\infty$$

and for $\mathcal{E}_z$ also, so $\mathcal{E}$ and $\mathcal{E}_z$ continuous

- compactness of $\mathcal{H}$ ensures existence of minimizer but not uniqueness (a uniqueness result when $\mathcal{H}$ convex)

Empirical target function $f_{\mathcal{H},z}$

- minimizer (non unique in general)

$$f_{\mathcal{H},z} = \mathrm{argmin}_{f\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m}(f(x_i)-y_i)^2$$

Normalized Error

$$\mathcal{E}_{\mathcal{H}}(f) = \mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}})$$

$\mathcal{E}_{\mathcal{H}}(f) \geq 0$ vanishing at $f_{\mathcal{H}}$

Sample Error $\mathcal{E}_{\mathcal{H}}(f_{\mathcal{H},z})$

$$\mathcal{E}(f_{\mathcal{H},z}) = \mathcal{E}_{\mathcal{H}}(f_{\mathcal{H},z}) + \mathcal{E}(f_{\mathcal{H}}) = \int_X (f_{\mathcal{H},z} - f_\rho)^2 + \sigma_\rho^2$$

estimating $\mathcal{E}(f_{\mathcal{H},z})$ by estimating sample and approximation errors, $\mathcal{E}_{\mathcal{H}}(f_{\mathcal{H},z})$ and $\mathcal{E}(f_{\mathcal{H}})$ one on $\mathcal{H}$ the other independent of sample z

bias-variance trade-off

- bias = approximation error; variance = sample error
  - fix $\mathcal{H}$: sample error $\mathcal{E}_{\mathcal{H}}(f_{\mathcal{H},z})$ decreases by increasing number m of samples
  - fix m: approximation error $\mathcal{E}(f_{\mathcal{H}})$ decreases when enlarging $\mathcal{H}$

- procedure:
  1. estimate how close $f_{\mathcal{H},z}$ and $f_{\mathcal{H}}$ depending on m
  2. how to choose $\dim \mathcal{H}$ when m is fixed

- first problem: how many examples need to draw to say with confidence $\geq 1-\delta$ that $\int_X (f_{\mathcal{H},z} - f_{\mathcal{H}})^2 \leq \epsilon$?

Uniformity Estimate (Vapnik's Statistical Learning Theory)

• covering number: S metric space, $s > 0$, number $\mathcal{N}(S, s)$ minimal $\ell \in \mathbb{N}$ so that $\exists$ disks in S radii s covering S; for S compact $\mathcal{N}(S, s)$ finite

• uniform estimate: $\mathcal{H} \subset C(X)$ compact, if for all $f \in \mathcal{H}$ have $|f(x) - y| \leq M$ a.e., then $\forall \epsilon > 0$

$$\mathbb{P}\{z \in Z^m \,:\, \sup_{f \in \mathcal{H}} |L_z(f)| \leq \epsilon\} \geq 1 - \mathcal{N}(\mathcal{H}, \frac{\epsilon}{8M}) 2 \exp\left(- \frac{m\epsilon^2}{4(2\sigma^2 + \frac{1}{3}M^2\epsilon)} \right)$$

with $\sigma^2 = \sup_{f \in \mathcal{H}} \sigma^2(f_Y^2)$

• main idea: like previous "estimate of defect" but passing from a single function to a family of functions, using a uniformity based on "covering number"

Estimate of Sample Error

• $\mathcal{H} \subset C(X)$ compact, with $|f(x) - y| \leq M$ a.e. for all $f \in \mathcal{H}$, and $\sigma^2 = \sup_{f \in \mathcal{H}} \sigma^2(f_Y^2)$, then $\forall \epsilon > 0$

$$\mathbb{P}\{z \in Z^m \,:\, \mathcal{E}_{\mathcal{H}}(f_z) \leq \epsilon\} \geq 1 - \mathcal{N}(\mathcal{H}, \frac{\epsilon}{16M}) 2 \exp\left(- \frac{m\epsilon^2}{8(4\sigma^4 + \frac{1}{3}M^2\epsilon)} \right)$$

• obtained from previous estimate using $L_z(f) = \mathcal{E}(f) - \mathcal{E}_z(f)$

• so answer to first question: to ensure probability above $\geq 1 - \delta$ need to take at least

$$m \geq \frac{8(4\sigma^4 + \frac{1}{3}M^2\epsilon)}{\epsilon^2} \left(\log(2\mathcal{N}(\mathcal{H}, \frac{\epsilon}{16M})) + \log(\frac{1}{\delta}) \right)$$

obtained by setting

$$\delta = \mathcal{N}(\mathcal{H}, \frac{\epsilon}{16M}) 2 \exp\left(- \frac{m\epsilon^2}{8(4\sigma^4 + \frac{1}{3}M^2\epsilon)} \right)$$

• need various techniques for estimating covering numbers $\mathcal{N}(\mathcal{H}, s)$ depending on the choice of the compact set $\mathcal{H}$

Second Question: Estimating the Approximation Error

$$\mathcal{E}(f_{\mathcal{H},z}) = \mathcal{E}_{\mathcal{H}}(f_{\mathcal{H},z}) + \mathcal{E}(f_{\mathcal{H}})$$

focus on $\mathcal{E}(f_{\mathcal{H}})$, which depends on $\mathcal{H}$ and ρ

$$\int_X (f_{\mathcal{H}} - f_\rho)^2 + \sigma_\rho^2$$

second term independent of $\mathcal{H}$ so focus on first; f_ρ bounded, but not in $\mathcal{H}$ nor necessarily in $C(X)$

- Main idea: use finite dimensional hypothesis space $\mathcal{H}$; estimate in terms of growth of eigenvalues of an operator
- Main technique: Fourier analysis; Hilbert spaces

Fourier Series: start with case of $X = T^n = (S^1)^n$ torus

• Hilbert space $L^2(X)$ Lebesgue measure with complete orthonormal system

$$\phi_\alpha(x) = (2\pi)^{-n/2} \exp(i\alpha \cdot x), \quad \alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{Z}^n$$

Fourier series expansion

$$f = \sum_{\alpha \in \mathbb{Z}^n} c_\alpha \, \phi_\alpha$$

• finite dimensional subspaces $\mathcal{H}_N \subset L^2(X)$ spanned by ϕ_α with $\|\alpha\| \leq B$, dimension $N(B)$ number of lattice points in ball radius B in $\mathbb{R}^n$

$$N(B) \leq (2B)^{n/2}$$

• $\mathcal{H}$ hypothesis space: ball $\mathcal{H}_{N,R}$ of radius R in $\| \cdot \|_\infty$ norm in $\mathcal{H}_N$

Laplacian

- on torus $X = T^n$ Laplacian $\Delta : \mathcal{C}^\infty(X) \to \mathcal{C}^\infty(X)$

$$\Delta(f) = \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}$$

Fourier series basis ϕ_α are eigenfunctions of $-\Delta$ with eigenvalue $\|\alpha\|^2$

- more general X: bounded domain $X \subset \mathbb{R}^n$ with smooth boundary ∂X and a complete orthonormal system ϕ_k of $L^2(X)$ (Lebesgue measure) of eigenfunctions of Laplacian with

$$-\Delta(\phi_k) = \zeta_k \, \phi_k, \quad \phi_k|_{\partial X} \equiv 0, \quad \forall k \geq 1$$

$$0 < \zeta_1 \leq \zeta_2 \leq \cdots \leq \zeta_k \leq \cdots$$

- subspace $\mathcal{H}_N$ of $L^2(X)$ generated by $\{\phi_1, \ldots, \phi_N\}$
- hypothesis space $\mathcal{H} = \mathcal{H}_{N,R}$ ball of radius R for $\| \cdot \|_\infty$ in $\mathcal{H}_N$

Construction of $f_{\mathcal{H}}$

- Lebesgue measure μ on X and measure ρ (marginal probability ρ_X induced by ρ on $Z = X \times Y$)
- consider regression function

$$f_\rho(x) = \int_Y y \, d\rho(y|x)$$

- assumption f_ρ bounded on X so in $L^2_\rho(X)$ and in $L^2_\mu(X)$
- choice of R: assume also that $R \geq \|f_\rho\|_\infty$, which implies $R \geq \|f_\rho\|_\rho$
- then $f_{\mathcal{H}}$ is orthogonal projection of f_ρ onto $\mathcal{H}_N$ using inner product in $L^2_\rho(X)$
- goal: estimate approximation error $\mathcal{E}(f_{\mathcal{H}})$ for this $f_{\mathcal{H}}$

Distorsion factor:

- identity function on bounded functions extends to

$$J : L^2_\mu(X) \to L^2_\rho(X)$$

- distorsion of ρ with respect to μ

$$D_{\rho\mu} = \|J\|$$

operator norm: how much ρ distorts the ambient measure μ

- reasonable assumption: distorsion is finite
- in general ρ not known, but ρ_X is known, so $D_{\rho\mu}$ can be computed

Weyl Law

• Weyl law on rate of growth of eigenvalues of the Laplacian (acting on functions vanishing on boundary of domain $X \subset \mathbb{R}^n$)

$$\lim_{\lambda \to \infty} \frac{N(\lambda)}{\lambda^{n/2}} = (2\pi)^{-n} B_n Vol(X)$$

B_n volume of unit ball in $\mathbb{R}^n$; $N(\lambda)$ number of eigenvalues (with multiplicity) up to λ

• Weyl law: Li–Yau version

$$\zeta_k \geq \frac{n}{n+2} 4\pi^2 \left(\frac{k}{B_n \, Vol(X)} \right)^{2/n}$$

P. Li and S.-T. Yau, *On the parabolic kernel of the Schrödinger operator*, Acta Math. 156 (1986), 153–201

• from this get a weaker estimate, using explicit volume B_n

$$\zeta_k \geq \left(\frac{k}{Vol(X)} \right)^{2/n}$$

Approximation Error and Weyl Law

- norm $\|\cdot\|_K$: for $f = \sum_{k=1}^\infty c_k \, \phi_k$ with ϕ_k eigenfunctions of $-\Delta$

$$\|f\|_K := \left(\sum_{k=1}^\infty c_k^2 \zeta_k \right)^{1/2}$$

like L^2-norm but weighted by eigenvalues of Laplacian in ℓ^2 measure of $c = (c_k)$

- Approximation Error Estimate: for $\mathcal{H}$ and $f_\mathcal{H}$ as above

$$\mathcal{E}(f_\mathcal{H}) \leq D_{\rho\mu}^2 \left(\frac{k}{Vol(X)} \right)^{2/n} \|f_\rho\|_K^2 + \sigma_\rho^2$$

- proved using Weyl law and estimates

$$\|f_\rho - f_\mathcal{H}\|_\rho = d_\rho(f_\rho, \mathcal{H}_N) \leq \|J\| \, d_\mu(f_\rho, \mathcal{H}_N)$$

$$d_\mu(f_\rho, \mathcal{H}_N)^2 = \| \sum_{k=N+1}^\infty c_k \phi_k \|_\mu^2 = \sum_{k=N+1}^\infty c_k^2 = \sum_{k=N+1}^\infty c_k^2 \zeta_k \frac{1}{\zeta_k} \leq \frac{1}{\zeta_{N+1}} \|f_\rho\|_K^2$$

where $f_\rho = \sum_k c_k \phi_k$

Solution of the bias-variance problem

• mimimize $\mathcal{E}(f_{\mathcal{H},z})$ by minimizing both sample error and approximation error

• minimization as a function of $N \in \mathbb{N}$ (for the choice of hypothesis space $\mathcal{H} = \mathcal{H}_{N,R}$)

• select integer $N \in \mathbb{N}$ that minimizes $\mathcal{A}(N) + \epsilon(N)$ where $\epsilon = \epsilon(N)$ as in previous estimate of sample error and

$$\mathcal{A}(N) = D^2_{\rho\mu} \left(\frac{k}{Vol(X)} \right)^{2/n} \|f_\rho\|^2_K + \sigma^2_\rho$$

• from previous relation between m, $R = \|f_\rho\|_\infty$, δ and ϵ obtain

$$\epsilon - \frac{288 M^2}{m} \left(N \log(\frac{96 RM}{\epsilon}) + 1 + \log(\frac{1}{\delta}) \right) \geq 0$$

find N that minimizes ϵ with this constraint

• no explicit closed form solution for N minimizing $\mathcal{A}(N) + \epsilon(N)$ but can be estimated numerically in specific cases

back to the visual cortex modeling (Poggio-Anselmi)

- stored templates t^k, $k = 1, \ldots, K$ and new images $\mathcal{I}$ in some finite dimensional approximation $\mathcal{H}_N$ to a Hilbert space
- simple cells perform inner products $\langle gt^k, \mathcal{I} \rangle$ in $\mathcal{H}_N$
- estimate in terms of 1D-projections: $\mathcal{I} \in \mathbb{R}^d$ some in general large d; projections $\langle t^k, \mathcal{I} \rangle$ for a set of normalized vectors $t^k \in S^{d-1}$ (unit sphere)

$$Z : S^{d-1} \to \mathbb{R}_+, \quad Z(t) = \left| \mu^t(\mathcal{I}) - \mu^t(\mathcal{I}') \right|$$

- distance between images $d(\mathcal{I}, \mathcal{I}')$ think of as a distance between two probability distributions $P_{\mathcal{I}}$, $P_{\mathcal{I}'}$ on $\mathbb{R}^d$
- measure distance in terms of

$$d(P_{\mathcal{I}}, P_{\mathcal{I}'}) \sim \int_{S^{d-1}} Z(t)\, dvol(t)$$

• model this in terms of

$$\hat{d}(P_{\mathcal{I}}, P_{\mathcal{I}'}) := \frac{1}{K} \sum_{k=1}^{K} Z(t^k)$$

want to evaluate the error incurred in using $\hat{d}(P_{\mathcal{I}}, P_{cI'})$ (1D projections and templates) to estimate $d(P_{\mathcal{I}}, P_{\mathcal{I}'})$

• as in the Cucker-Smale setting, evaluate the error and the probability of error in terms of the Hoeffding estimate

$$\left| d(P_{\mathcal{I}}, P_{\mathcal{I}'}) - \hat{d}(P_{\mathcal{I}}, P_{cI'}) \right| = \left| \frac{1}{K} \sum_{k=1}^{K} Z(t^k) - \mathbb{E}(Z) \right|$$

• probability of error

$$\mathbb{P}\left(\left| \frac{1}{K} \sum_{k=1}^{K} Z(t^k) - \mathbb{E}(Z) \right| > \epsilon \right) \leq 2e^{-\frac{K\epsilon^2}{2M^2}}$$

if a.e. bound $|Z(t) - \mathbb{E}(Z)| \leq M$

• want this estimate to hold uniformly over a set of N images: want same bound to hold over each pair so error probability is at most

$$N(N-1)\exp\left(-\frac{K\epsilon^2}{2M_{\min}^2}\right) \sim N^2 \exp\left(-\frac{K\epsilon^2}{2M_{\min}^2}\right) \leq \delta^2$$

with $M_{\min}$ the smallest M over all pairs

• This is at most a given δ^2 whenever

$$K \geq \frac{4M_{\min}^2}{\epsilon^2} \log \frac{N}{\delta}$$

Group Actions and Orbits

- $\{t^k\}_{k=1,\dots,K}$ given templates
- G finite subgroup of the affine group (translations, rotations, scaling)
- G acts on set of images $\mathcal{I}$: orbit $G\mathcal{I}$
- projection $P : \mathbb{R}^d \to \mathbb{R}^K$ of images $\mathcal{I}$ onto span of templates t^k
- Johnson–Lindenstrauss lemma: low distortion embeddings of sets of points from a high-dimensional to a low-dimensional Euclidean space (special case with map an orthogonal projection)
  - given $0 < \epsilon < 1$; given finite set X of n points in $\mathbb{R}^d$
  - take $K > 8\log(n)/\epsilon^2$
  - then there is a linear map f given by a multiple of an orthogonal projection onto a (random) subspace of dimension K such that, for all $u, v \in X$

$$(1-\epsilon)\|u-v\|^2_{\mathbb{R}^d} \le \|f(u)-f(v)\|^2_{\mathbb{R}^K} \le (1+\epsilon)\|u-v\|^2_{\mathbb{R}^d}$$

- result depends on concentration of measure phenomenon

• up to a scaling, for a good choice of subspace spanned by templates, can take P to satisfy Johnson-Lindenstrauss lemma

• starting from finite set $X = \{u\}$ of images, can generate another set by including all group translates $X_G = \{g \cdot u \,:\, g \in G,\, u \in X\}$

• then for Johnson-Lindenstrauss lemma required accuracy for X_G

$$K > 8 \frac{\log(n \cdot \#G)}{\epsilon^2}$$

• so can estimate sufficiently well the distance between images in $\mathbb{R}^d$ using the distance between projections $\langle t^k, g\mathcal{I}\rangle$ of their group orbits onto the space of templates

• by $\langle t^k, g\mathcal{I}\rangle = \langle g^{-1} t^k, \mathcal{I}\rangle$ for unitary representations it would seem one needs to increase by $K \mapsto \#G \cdot K$ the number of templates to distinguish orbits, but in fact by argument above need an increase $K \mapsto K + 8\log(\#G)/\epsilon^2$

• given $\langle t^k, g\mathcal{I}\rangle = \langle g^{-1}t^k, \mathcal{I}\rangle$ computed by the simple cells, pooling by complex cells by computing

$$\mu_h^k(\mathcal{I}) = \frac{1}{\#G}\sum_{g\in G}\sigma_h(\langle gt^k, \mathcal{I}\rangle)$$

σ_h a set of nonlinear functions: examples

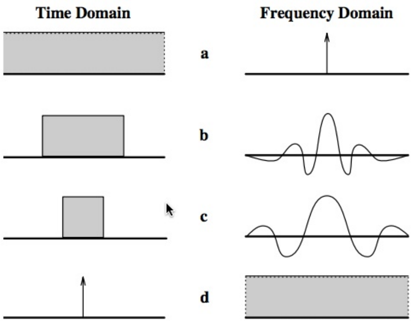
- $\mu^k_{\rm average}(\mathcal{I}) = \frac{1}{\#G}\sum_{g\in G}\left|\langle gt^k, \mathcal{I}\rangle\right|$
- $\mu^k_{\rm energy}(\mathcal{I}) = \frac{1}{\#G}\sum_{g\in G}\langle gt^k, \mathcal{I}\rangle^2$
- $\mu^k_{\rm max}(\mathcal{I}) = \max_{g\in G}\left|\langle gt^k, \mathcal{I}\rangle\right|$
- other nonlinear functions: especially useful case, when $\sigma_h : \mathbb{R}\to\mathbb{R}^+$ is *injective*

• Note: stored knowledge of gt^k for $g\in G$ allows the system to be automatically invariant wrt G action on images $\mathcal{I}$

Localization and uncertainty principle

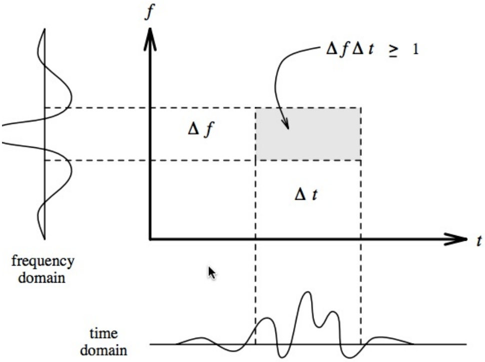
- would like templates $t(x)$ to be localized in x: small outside of some interval Δx
- would also like $\hat{t}$ to be localized in frequency: small outside an interval $\Delta\omega$
- but... uncertainty principle: localized in x / delocalized in ω

$$\Delta x \cdot \Delta\omega \geq 1$$



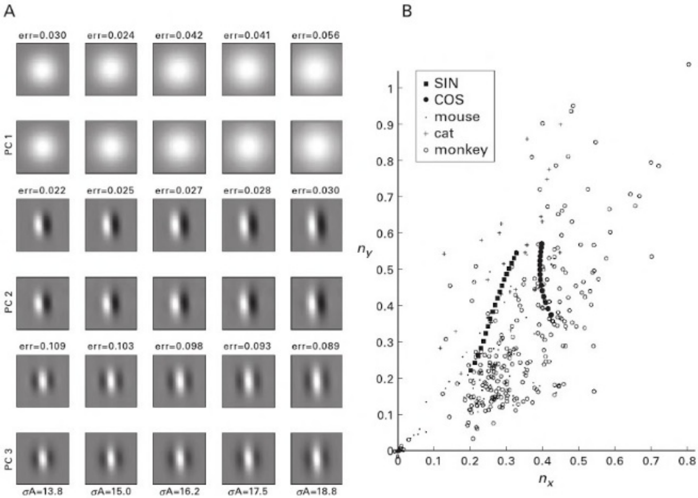
Optimal localization

- optimal possible localization when $\Delta x \cdot \Delta\omega = 1$



- realized by the Gabor functions

$$t(x) = e^{i\omega_0 x} e^{-\frac{x^2}{2\sigma^2}}$$



each cell applies a Gabor filter; plotted n_y/n_x anisotropy ratios