Stat 212b: Topics in Deep Learning
Lecture 3

Joan Bruna
UC Berkeley

Berkeley
UNIVERSITY OF CALIFORNIA

Marvin Minsky 1927-2016

Talking about his book *Perceptrons*:
"We really spent one year too much on it. We finished off all the easy conjectures, and so no beginner could do anything. We didn't leave anything for students to do. We got too greedy. As a result, ten years went by without another significant paper on the subject. It's a fact about the sociology of science that the people who should work in a field like this are the students and the graduate students. If we had given some of these problems to students, they would have got as good at it as we were, since there was nothing special about what we did except that we worked together for several years. Furthermore, I now believe that the book was overkill in another way. What we showed came down to the fact that a Perceptron can't put things together that are visually nonlocal."

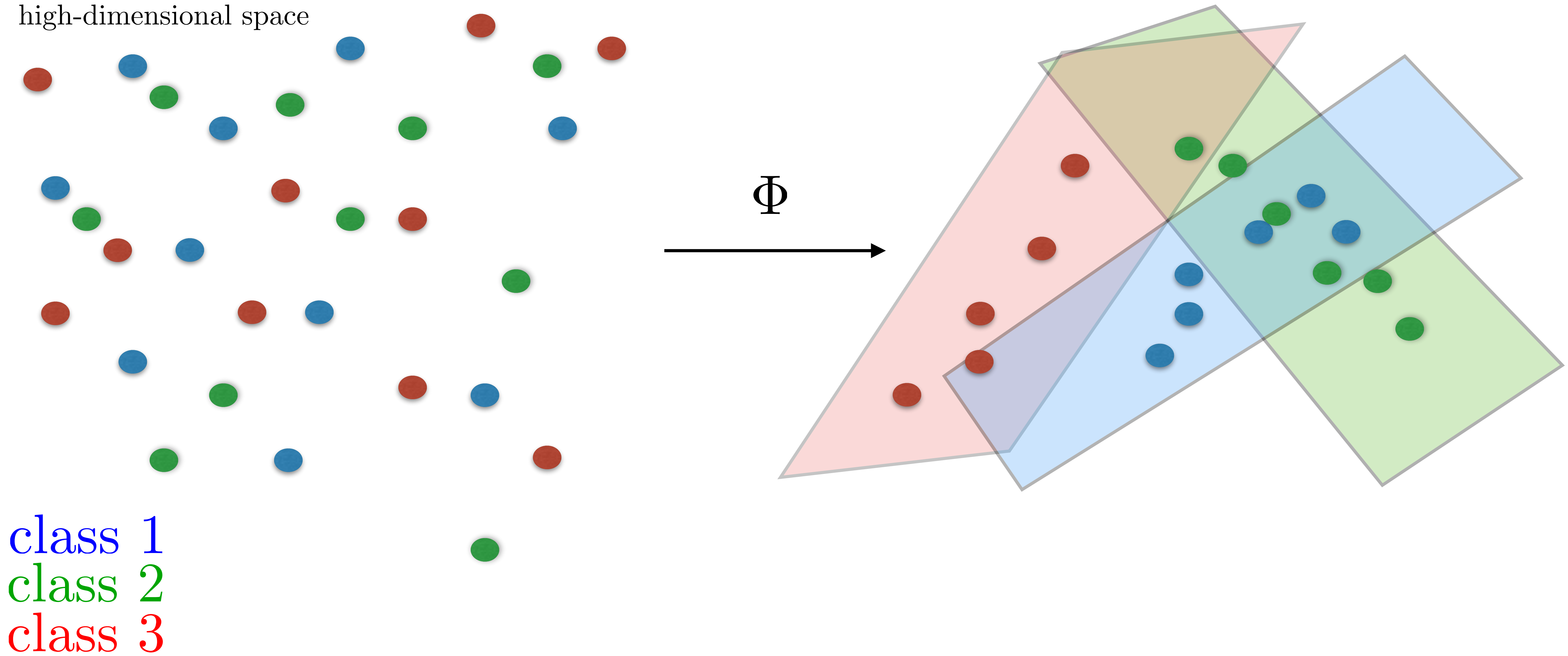
The New Yorker, 1981

Last Lecture Review

- Representations for recognition
  - curse of dimensionality
  - invariance/covariance
  - discriminability

- Variability models
  - transformation groups and symmetries
  - deformations
  - stationarity
  - clutter and class-specific

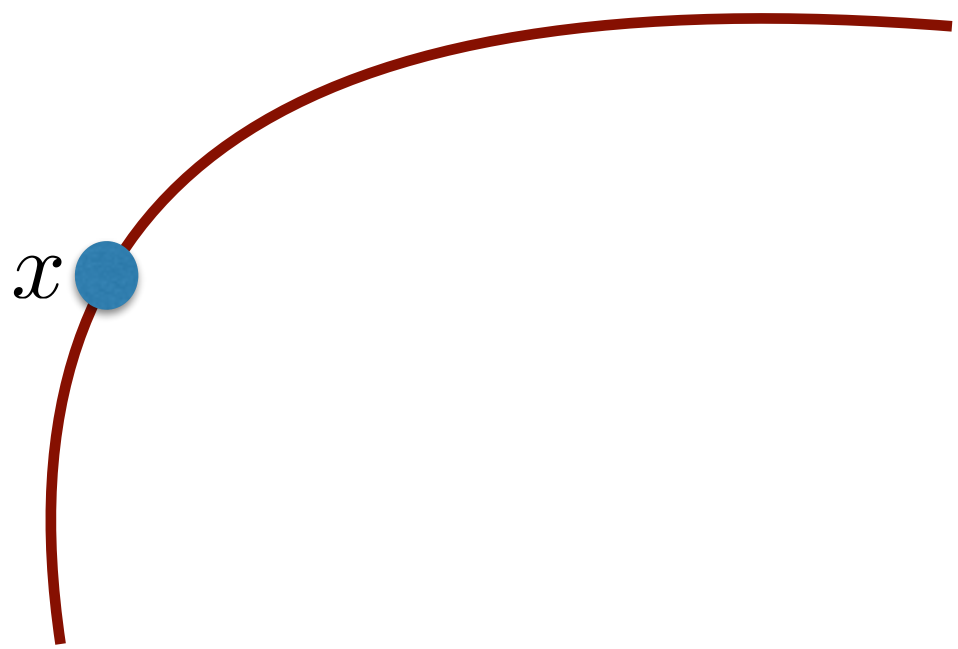
Review: Linearization

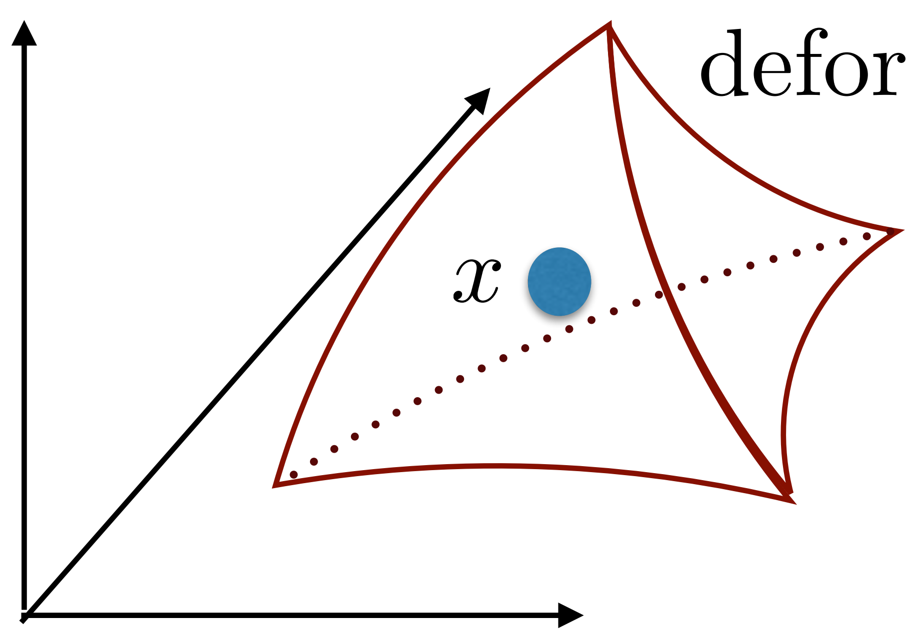


In order to beat the curse of dimensionality, we need features that ***linearize intra-class variability*** *and* ***preserve inter-class variability.***

Review: Filling the space with deformations

symmetry group: low dimension





deformations fill the space

Review: From Invariance to Stability




- Informally, if $\|\tau\|$ measures the amount of deformation, many recognition tasks satisfy

$$\forall\ x, \tau,\ \ |f(x) - f(x_\tau)| \lesssim \|\tau\|$$

- If our representation is stable, then

$$\forall\ x, \tau,\ \ \|\Phi(x) - \Phi(x_\tau)\| \leq C\|\tau\| \Longrightarrow |\hat{f}(x) - \hat{f}(x_\tau)| \leq \tilde{C}\|\tau\|$$

Objectives

1. Groups, invariants and filters.

2. Review of Wavelet Decompositions.

3. Examples

Transformation Groups

- We discussed about "universal" transformation groups acting on images, audio and video:

  - Translations: $\{\varphi_v\,;\, v \in \mathbb{R}^2\}$, with $\varphi_v(x)(u) = x(u-v)$.

  - Dilations: $\{\varphi_s\,;\, s \in \mathbb{R}_+\}$, with $\varphi_s(x)(u) = s^{-1}x(s^{-1}u)$.

  - Rotations: $\{\varphi_\theta\,;\, \theta \in [0, 2\pi)\}$, with $\varphi_\theta(x)(u) = x(R_\theta u)$.

Transformation Groups

- We discussed about "universal" transformation groups acting on images, audio and video:

  - Translations: $\{\varphi_v\,;v \in \mathbb{R}^2\}$, with $\varphi_v(x)(u) = x(u-v)$.

  - Dilations: $\{\varphi_s\,;s \in \mathbb{R}_+\}$, with $\varphi_s(x)(u) = s^{-1}x(s^{-1}u)$.

  - Rotations: $\{\varphi_\theta\,;\theta \in [0,2\pi)\}$, with $\varphi_\theta(x)(u) = x(R_\theta u)$.

- Systematic approach to obtain representations invariant to these groups?

One-parameter Unitary Groups

- A particularly simple example is given by *continuous one-parameter unitary transformations*:

Definition: A one-parameter unitary group $\{\varphi_t \in Aut(\Omega)\}_{t\in\mathbb{R}}$ satisfies

1. $\forall\ t, s,\ \ \varphi_{s+t} = \varphi_t \varphi_s\ \ ,$
2. $\lim_{s\to t} \|\varphi_s - \varphi_t\| = 0,$
3. $\forall\ t \in \mathbb{R}\,,\ \ x \in \Omega\ ,\ \ \|\varphi_t x\| = \|x\|\ .$

One-parameter Unitary Groups

- A particularly simple example is given by *continuous one-parameter unitary transformations*:

Definition: A one-parameter unitary group $\{\varphi_t \in Aut(\Omega)\}_{t\in\mathbb{R}}$ satisfies

1. $\forall\ t, s,\ \ \varphi_{s+t} = \varphi_t \varphi_s\ ,$
2. $\lim_{s\to t} \|\varphi_s - \varphi_t\| = 0,$
3. $\forall\ t \in \mathbb{R}\,,\ \ x \in \Omega\ ,\ \ \|\varphi_t x\| = \|x\|\ .$

- In particular, these are Abelian groups.
  - Rotations and Translations are 1-parameter unitary groups
  - Dilations can be made unitary: $\varphi_s x(u) = s^{1/2} x(su)\ .$

Stone's theorem

Theorem: Suppose Ω is a Hilbert space. There is a one-to-one correspondence between self-adjoint operators on Ω and one-parameter unitary groups of $Aut(\Omega)$.
Given $\{\varphi_t\}_{t\in\mathbb{R}}$, there exists A self-adjoint such that $\forall\, t\ ,\ \varphi_t = e^{itA}$. Conversely, if A is self-adjoint, the family $\{e^{itA}\}_t$ is a one-parameter unitary group.

Stone's theorem

Theorem: Suppose Ω is a Hilbert space. There is a one-to-one correspondence between self-adjoint operators on Ω and one-parameter unitary groups of $Aut(\Omega)$.
Given $\{\varphi_t\}_{t\in\mathbb{R}}$, there exists A self-adjoint such that $\forall\ t$, $\varphi_t = e^{itA}$. Conversely, if A is self-adjoint, the family $\{e^{itA}\}_t$ is a one-parameter unitary group.

Remark: In finite dimensions, we define the matrix exponential e^A, $A \in \mathbb{C}^{n\times n}$, as $e^A := \sum_{k\geq 0} \frac{A^k}{k!}$.

Proof: [class notes, or see http://www2.maths.lth.se/media/thesis/2010/MATX01.pdf]

Fourier transform Defrost

Definition The Fourier transform of a function $x \in L^2(\mathbb{R})$ is defined as

$$\hat{x}(\omega) = \int x(u) e^{-i\omega u} du \ .$$

Fourier transform Defrost

Definition The Fourier transform of a function $x \in L^2(\mathbb{R})$ is defined as

$$\hat{x}(\omega) = \int x(u) e^{-i\omega u} du \ .$$

[Main Properties]:

- Linear: $z = \alpha x + \beta y \Longrightarrow \hat{z} = \alpha \hat{x} + \beta \hat{y}$.
- Parseval identity: $\|\hat{x}\| = \|x\|$, $\langle x, y \rangle = \langle \hat{x}, \hat{y} \rangle$.
- Inverse Fourier transform: $x(u) = \int \hat{x}(\omega) e^{i\omega u} d\omega$.
- Translation: $y(u) = x(u - u_0) \Longrightarrow \hat{y}(\omega) = e^{i\omega u_0} \hat{x}(\omega)$.
- Dilation: $y(u) = x(su)$ for $s > 0 \Longrightarrow \hat{y}(\omega) = s^{-1} \hat{x}(s^{-1}\omega)$.

Stone theorem, Fourier and Global Invariants

- Translations are simultaneously diagonalized by Fourier atoms.

Stone theorem, Fourier and Global Invariants

- Translations are simultaneously diagonalized by Fourier atoms.
- The Stone theorem formalizes the fact that a collection of "nice" commuting operators simultaneously diagonalizes (in a complex basis):

$$A = V^* \mathrm{diag}(\lambda_1, \dots, \lambda_n) V$$

  - Unitary condition implies that eigenvalues are unitary complex numbers.

Stone theorem, Fourier and Global Invariants

- Translations are simultaneously diagonalized by Fourier atoms.
- The Stone theorem formalizes the fact that a collection of "nice" commuting operators simultaneously diagonalizes (in a complex basis):

$$\boxed{A = V^* \mathrm{diag}(\lambda_1, \dots, \lambda_n) V}$$

  - Unitary condition implies that eigenvalues are unitary complex numbers.
- What happens on larger Abelian (commuting) groups?
  - Factorization of Abelian groups into one-parameter groups (eg translations in R2)

$$G = G_1 \times G_2 \times \dots G_l$$

Stone theorem, Fourier and Global Invariants

- Translations are simultaneously diagonalized by Fourier atoms.
- The Stone theorem formalizes the fact that a collection of "nice" commuting operators simultaneously diagonalizes (in a complex basis):

$$A = V^* \mathrm{diag}(\lambda_1, \ldots, \lambda_n) V$$

  - Unitary condition implies that eigenvalues are unitary complex numbers.
- What happens on larger Abelian (commuting) groups?
  - Factorization of Abelian groups into one-parameter groups (eg translations in R2)

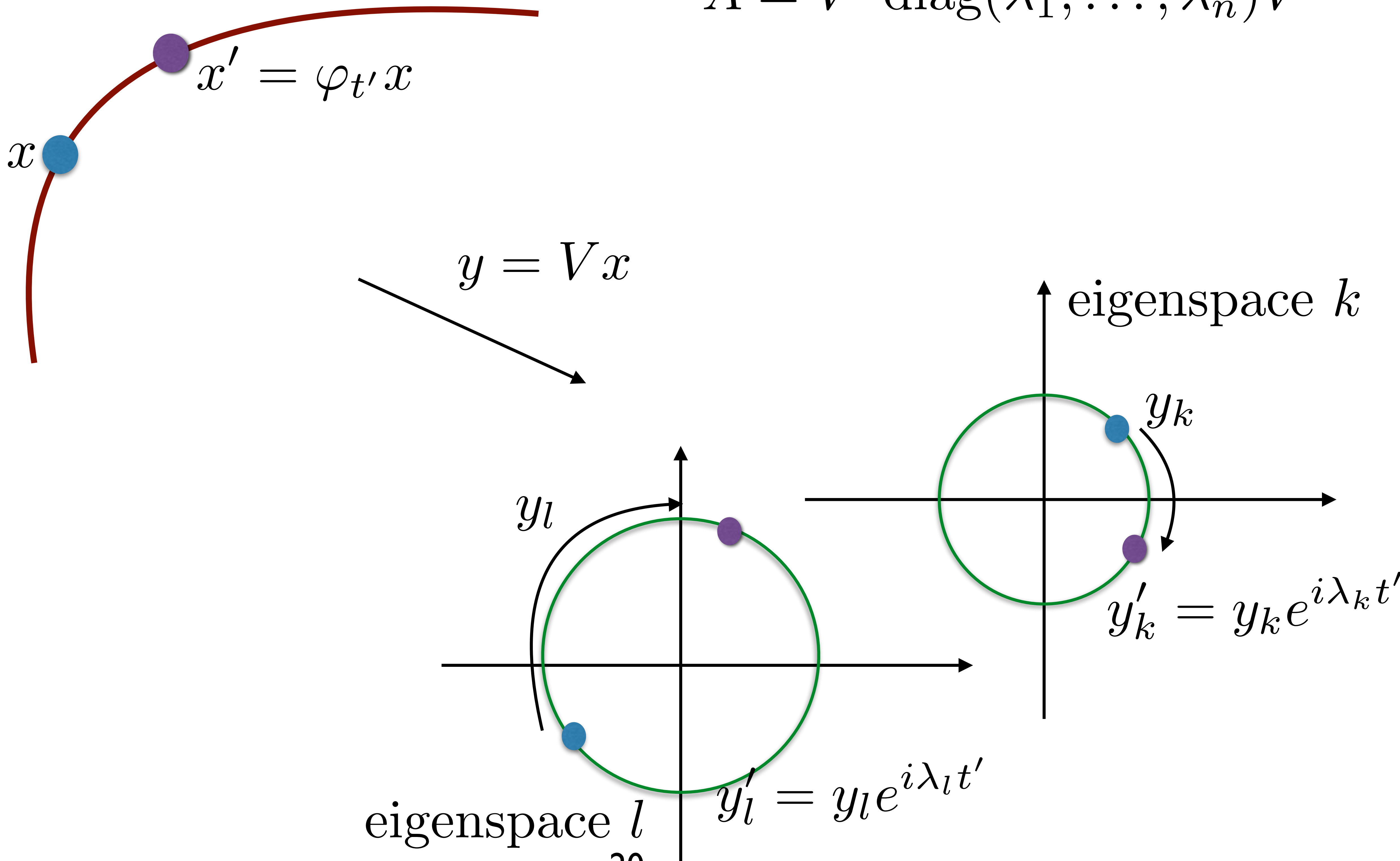
$$G = G_1 \times G_2 \times \ldots G_l$$

- Q: How to obtain global invariants in that case?

Stone theorem, Fourier and Global Invariants

$\{\varphi_t x\}_{t\in\mathbb{R}}$ one-parameter group

$$A = V^*\mathrm{diag}(\lambda_1, \ldots, \lambda_n)V$$



- Thus $\Phi(x) = |Vx|$ satisfies

$$\forall\ x, t\ ,\ \Phi(\varphi_t(x)) = \Phi(x)\ .$$

- Thus $\Phi(x) = |Vx|$ satisfies

$$\forall\ x, t\ ,\ \ \Phi(\varphi_t(x)) = \Phi(x)\ .$$

- Indeed,

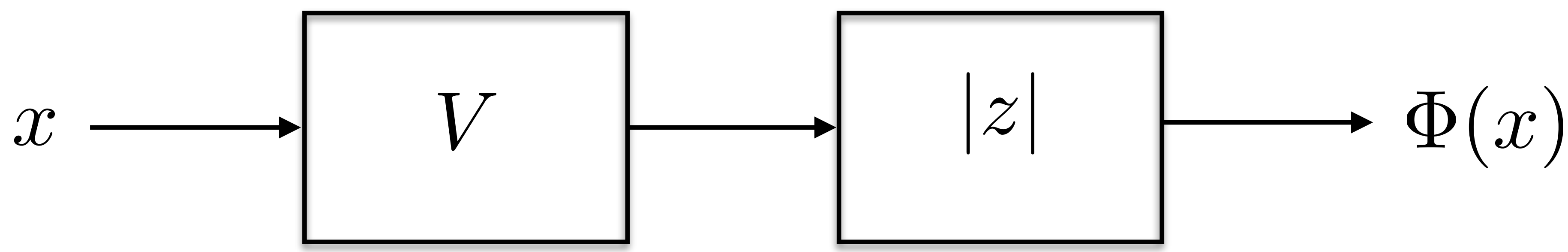
$$A = V^* \mathrm{diag}(\lambda_1, \ldots, \lambda_n) V \implies e^{itA} = V^* \mathrm{diag}(e^{it\lambda_1}, \ldots, e^{it\lambda_n}) V\ .$$

$$\begin{aligned} V \varphi_t x = V e^{itA} x &= V V^* \mathrm{diag}(e^{it\lambda_1}, \ldots, e^{it\lambda_n}) V x \\ &= \mathrm{diag}(e^{it\lambda_1}, \ldots, e^{it\lambda_n}) V x \end{aligned}$$

$$\text{thus}\ \Phi(\varphi_t x) = |V \varphi_t x| = |Vx|\ .$$

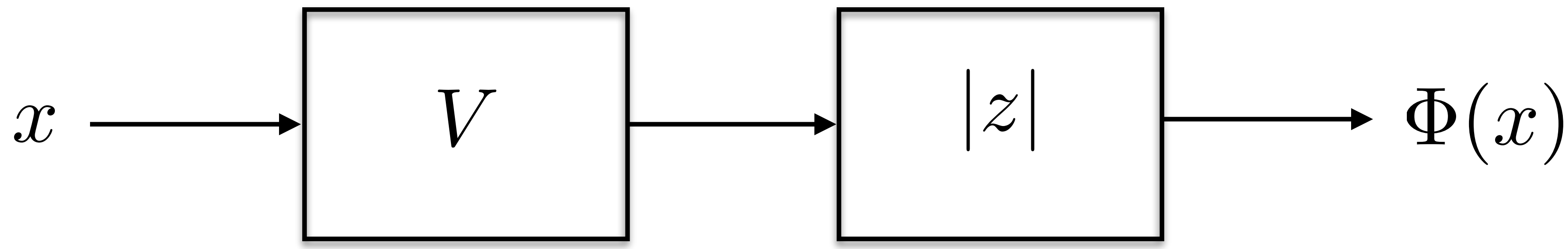
Limits of Group Diagonalisation

- A shallow (1 layer) network is thus sufficient to achieve invariance to commutative group transformations:



Limits of Group Diagonalisation

- A shallow (1 layer) network is thus sufficient to achieve invariance to commutative group transformations:



- However, this architecture has a number of shortcomings.

Limits of Group Diagonalisation

- Non-commutative Groups:

Proposition: If $G = \{\varphi_t\}_t$ is non-commutative, then there is no basis V that diagonalises simultaneously all φ_t.

Limits of Group Diagonalisation

- Non-commutative Groups:

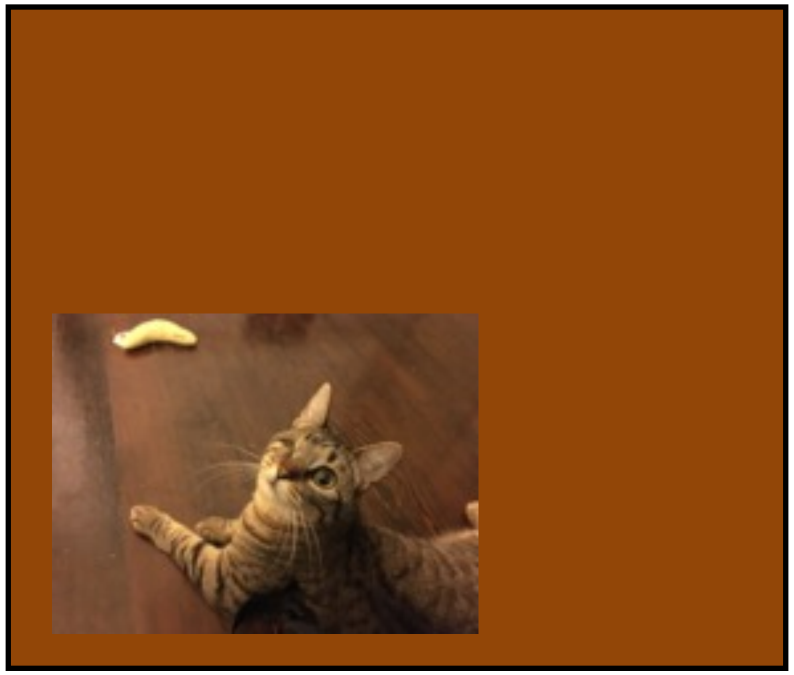
Proposition: If $G = \{\varphi_t\}_t$ is non-commutative, then there is no basis V that diagonalises simultaneously all φ_t.

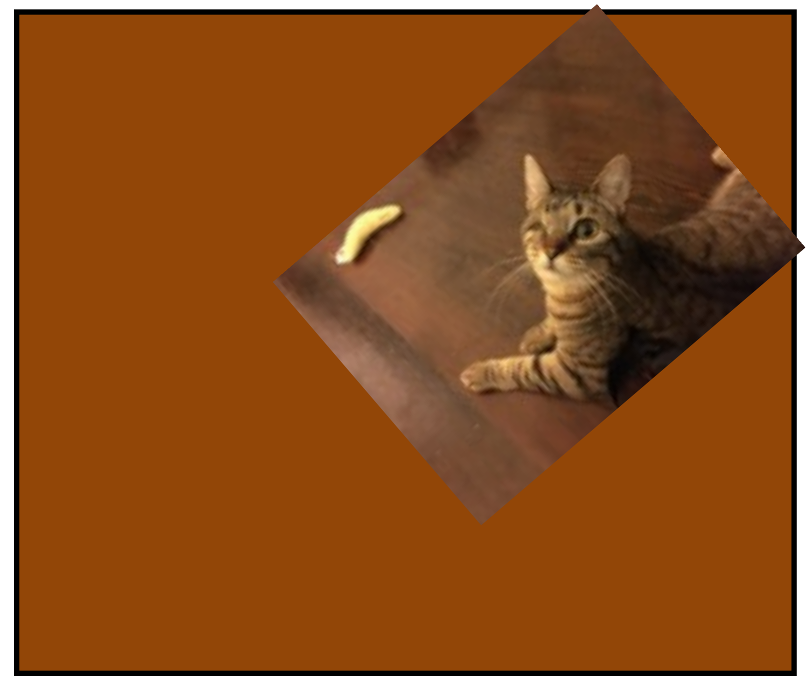
$$\text{Square matrices } A \text{ and } B \text{ commute}$$

$$\Updownarrow$$

$$A \text{ and } B \text{ share the same eigenvectors.}$$

Example: the Roto-Translation Group

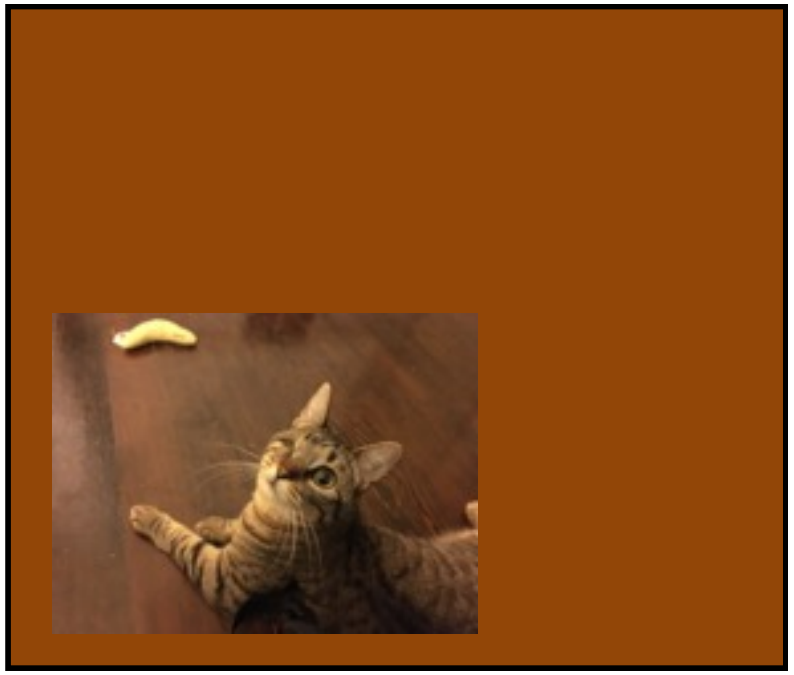
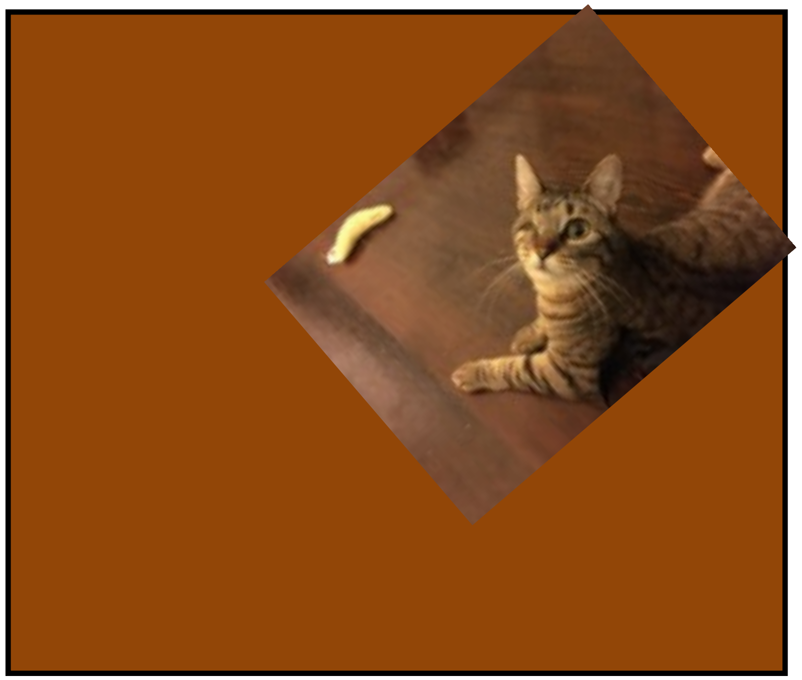




Roto-translation group: $\{\varphi_{v,\theta}\,; v \in \mathbb{R}^2,\ \theta \in [0, 2\pi)\}$.

$$\varphi_{v,\theta} \ : \ u \mapsto R_\theta(u - v) \ .$$

Example: the Roto-Translation Group

Roto-translation group: $\{\varphi_{v,\theta}\,;v\in\mathbb{R}^2,\ \theta\in[0,2\pi)\}$.

$$\varphi_{v,\theta}\ :\ u\mapsto R_\theta(u-v)\ .$$

$$\begin{aligned}\varphi_{v',\theta'}\cdot\varphi_{v,\theta}u &= R_{\theta'}(\varphi_{v,\theta}u-v') = R_{\theta'}(R_\theta u - R_\theta v - v')\\ &= R_{\theta'}R_\theta u - (R_{\theta'}R_\theta v + R_{\theta'}v')\\ &= R_{\theta+\theta'}\left(u-(v+R_{-\theta}v')\right)\end{aligned}$$

Thus $(v',\theta')\cdot(v,\theta)=(v+R_{-\theta}v',\theta+\theta')$

- We will see later how to deal with such groups.

Limits of Group Diagonalisation

- How discriminative is $\Phi(x) = |Vx|$?

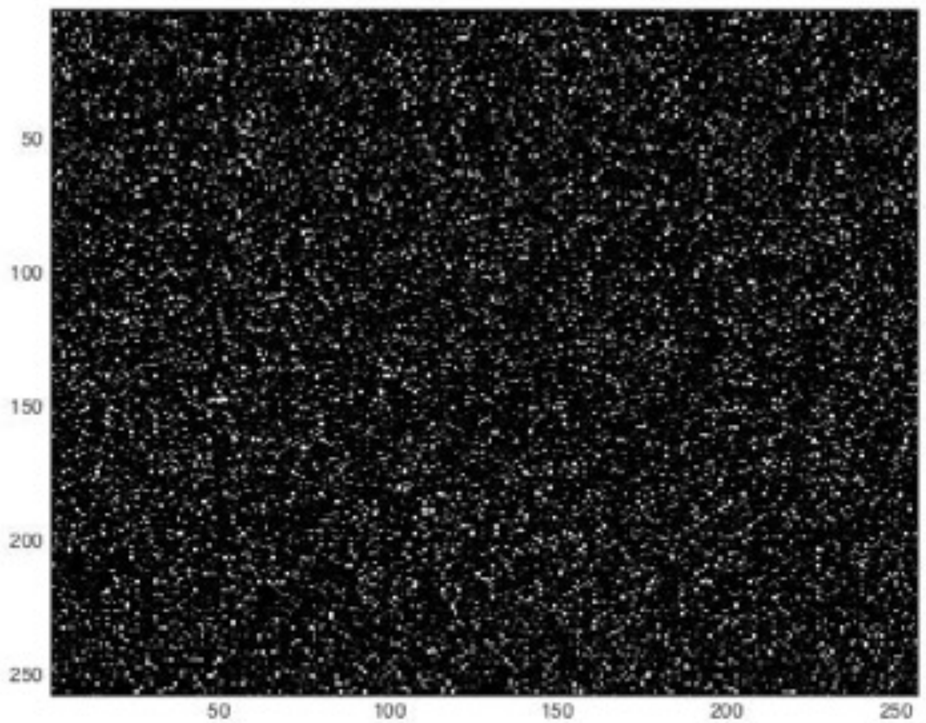
Limits of Group Diagonalisation

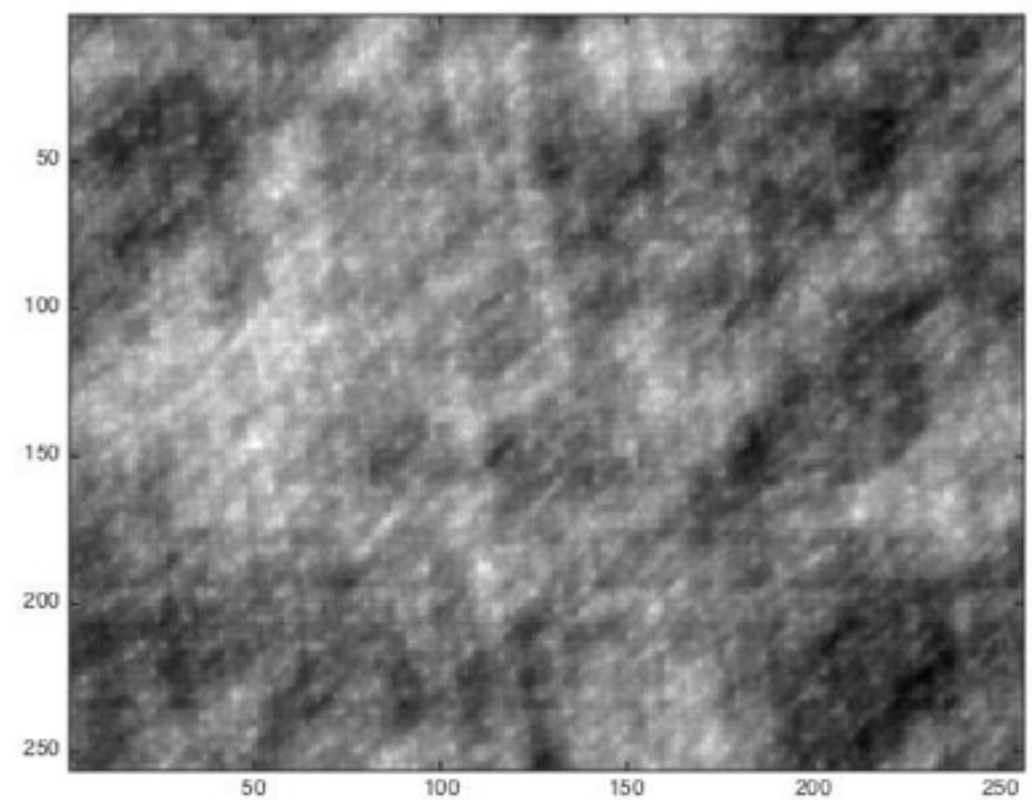
- How discriminative is $\Phi(x) = |Vx|$?
  - Because of Hermitic symmetry, $\Phi : \mathbb{R}^n \to \mathbb{R}^{\lceil n/2 \rceil}$
  - We "pay" n/2 degrees of freedom to remove group variability, *independently of the group dimensionality.*

Limits of Group Diagonalisation

- How discriminative is $\Phi(x) = |Vx|$?
  - Because of Hermitic symmetry, $\Phi : \mathbb{R}^n \to \mathbb{R}^{\lceil n/2 \rceil}$
  - We "pay" n/2 degrees of freedom to remove group variability, *independently of the group dimensionality.*
- If the group has dimension *p*, a G-invariant representation could have up to *n-p* d.f. : we are losing discriminability when *p* is small.

Limits of Group Diagonalisation





Limits of Group Diagonalisation

- How discriminative is $\Phi(x) = |Vx|$?
  - Because of Hermitic symmetry, $\Phi \ : \ \mathbb{R}^n \to \mathbb{R}^{\lceil n/2 \rceil}$
  - We "pay" n/2 degrees of freedom to remove group variability, *independently of the group dimensionality.*
  - If the group has dimension p, a G-invariant representation has at most n-p d.f. : we are losing discriminability when p is small.
- Fourier Phases encode most of the relevant signal information.

Limits of Group Diagonalisation

- Stable to deformations?

Limits of Group Diagonalisation

- Stable to deformations?

- The diagonalisation ensures that $\Phi(\varphi_t x) = \Phi(x) \ \forall \, t, x$, but we have no control outside the group $\{\varphi_t\}_t$ in general.

Limits of Group Diagonalisation

- Stable to deformations?

- The diagonalisation ensures that $\Phi(\varphi_t x) = \Phi(x) \; \forall \, t, x$, but we have no control outside the group $\{\varphi_t\}_t$ in general.

- To evaluate stability, we first need to quantify the amount of deformation.

- Also, we need the notion of **scale:** in many applications, we are interested in *local* invariance rather than *global* group invariance.

Deformation Metric

- Assume $\tau : \mathbb{R}^d \to \mathbb{R}^d$ differentiable, and denote
  $\varphi_\tau x(u) := x(u - \tau(u))$.
- $\|\nabla \tau(u)\|$: operator norm of Jacobian of τ at u.
- If $\|\nabla \tau\|_\infty = \sup_u \|\nabla \tau(u)\| < 1$,
  then φ_τ is invertible, and it defines a diffeomorphism.
- We consider the following deformation cost:

$$\|\tau\| := 2^{-J}\|\tau\|_\infty + \|\nabla \tau\|_\infty \ .$$

Deformation Metric

- We consider the following deformation cost:

$$\|\tau\| := 2^{-J}\|\tau\|_\infty + \|\nabla\tau\|_\infty \ .$$

- Scale J controls how much we pay for absolute displacements

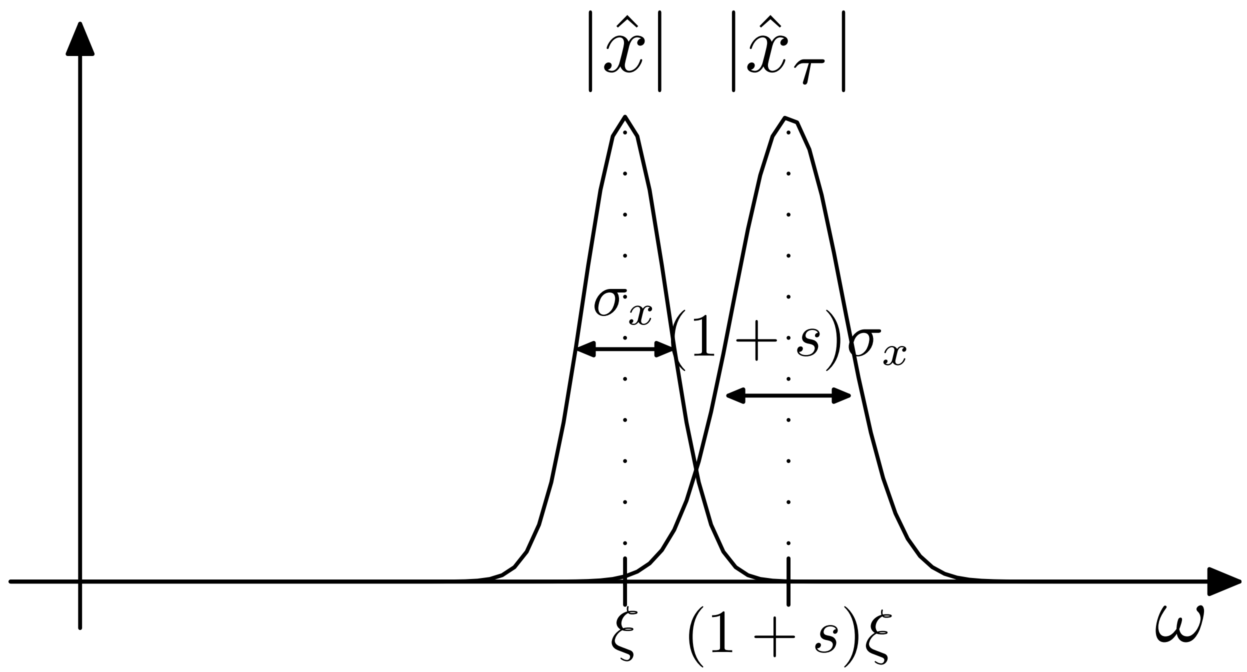
- Stability criterion: $\forall \ \|x\| = 1, \tau, \ \ \|\Phi(x) - \Phi(x_\tau)\| \leq C\|\tau\|.$

- We can define similar metrics for diffeomorphisms associated with other transformation groups (e.g. rotation).

Shallow invariants are unstable

- Consider a lowpass window $h(u)$ of bandwidth σ_h and $x(u) = h(u)e^{i\xi u}$. (bandwidth: $\sigma_h^2 = \int |\hat{h}(\omega)|^2 |\omega|^2 d\omega$.)

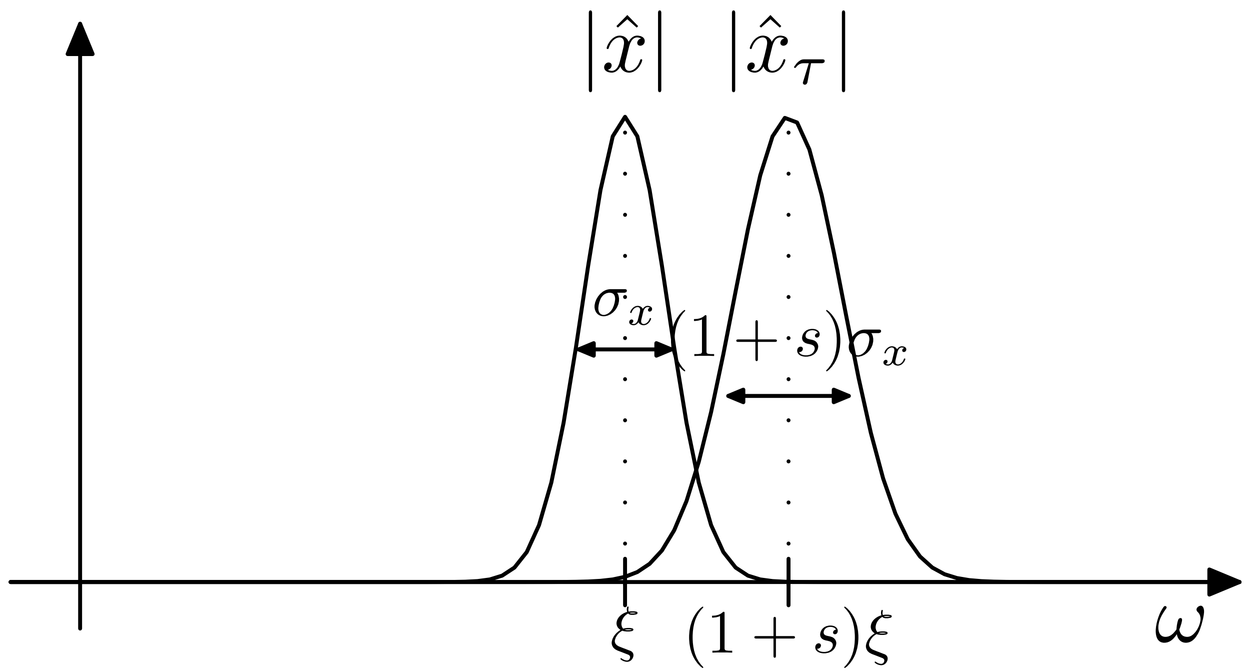
- Consider a deformation of the form $\varphi_\tau x(u) = x((1+s)u)$ with $s \ll 1$.



Shallow invariants are unstable

- Consider a lowpass window $h(u)$ of bandwidth σ_h and $x(u) = h(u)e^{i\xi u}$. (bandwidth: $\sigma_h^2 = \int |\hat{h}(\omega)|^2 |\omega|^2 d\omega$.)

- Consider a deformation of the form $\varphi_\tau x(u) = x((1+s)u)$ with $s \ll 1$.



If $(1+s)\xi - \xi = s\xi \gg \sigma_h(2+s)$
(central frequency separation $\gg$ bandwidth)

$$\implies \| |\hat{x}| - |\widehat{\varphi_\tau x}| \| \sim \|x\|$$

Shallow invariants are unstable

- Fourier Modulus is therefore unstable: high-frequency information spans a large linear subspace as soon as there is non-rigid deformation.

Shallow invariants are unstable

- Fourier Modulus is therefore unstable: high-frequency information spans a large linear subspace as soon as there is non-rigid deformation.

- Similarly, we can obtain a translation-invariant representation with the signal auto-correlation:

$$R_x(v) = \int x(u)x^*(u+v)du$$

  - This suffers from the same problem as Fourier.

$$\left(\|R_x - R_y\| = \|\hat{R}_x - \hat{R}_y\| = \| |\hat{x}|^2 - |\hat{y}|^2 \| \right)$$

Shallow invariants are unstable

- Fourier Modulus is therefore unstable: high-frequency information spans a large linear subspace as soon as there is non-rigid deformation.

- Similarly, we can obtain a translation-invariant representation with the signal auto-correlation:

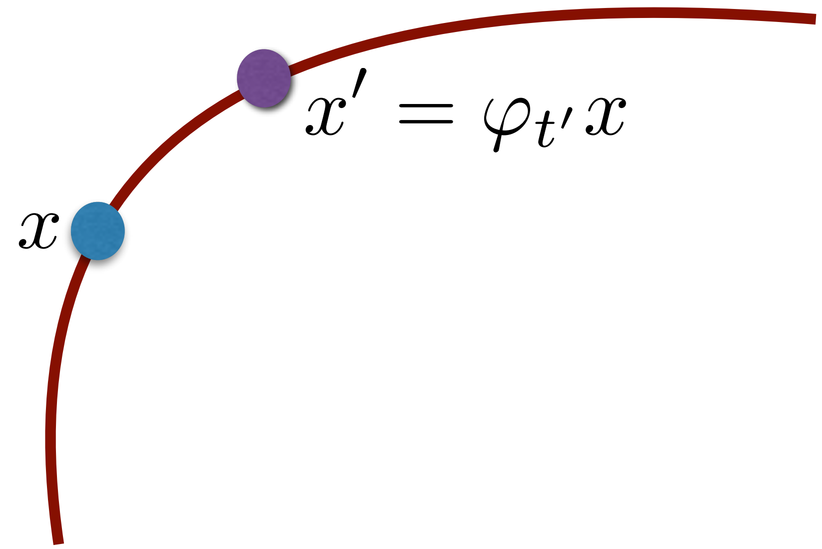
$$R_x(v) = \int x(u) x^*(u+v) du$$

  - This suffers from the same problem as Fourier.

$$\left(\|R_x - R_y\| = \|\hat{R}_x - \hat{R}_y\| = \| |\hat{x}|^2 - |\hat{y}|^2 \| \right)$$

- How to fix it?

Local invariants and convolution

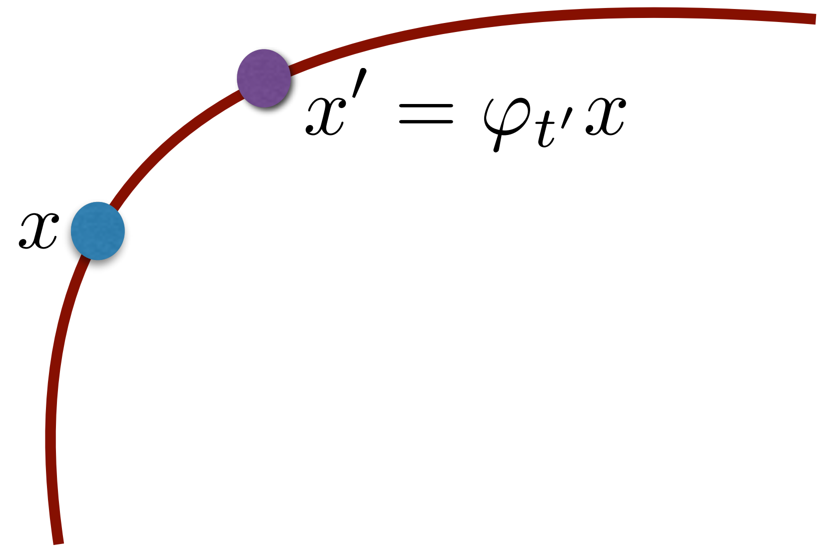


- Local translation invariance:

$$\|\Phi(x) - \Phi(\varphi_v x)\| \leq C 2^{-J} \|v\| \text{ , or}$$

$$\forall\ v, \|x\| = 1\ ,\ \frac{\|\Phi(x) - \Phi(\varphi_v x)\|}{\|v\|} \leq C 2^{-J}\ .$$

Local invariants and convolution



- Local translation invariance:

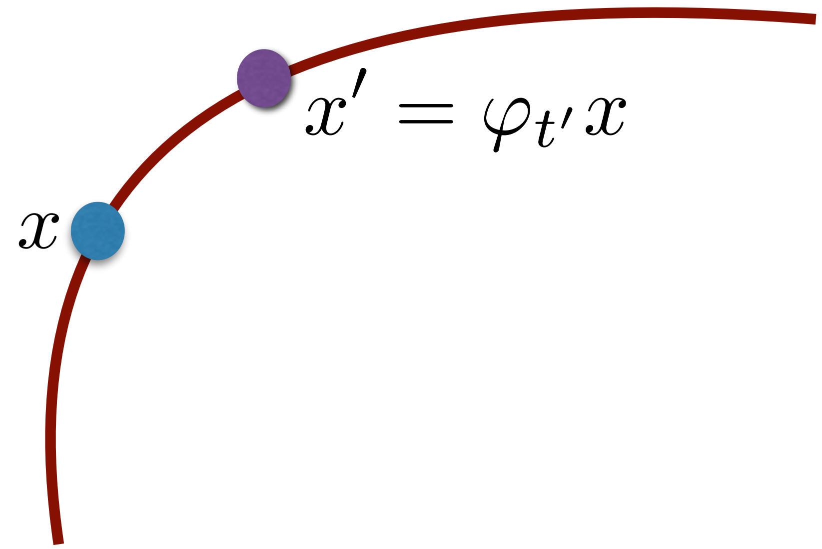
$$\|\Phi(x) - \Phi(\varphi_v x)\| \leq C 2^{-J} \|v\| \text{ , or}$$

$$\forall\ v, \|x\| = 1\ ,\ \frac{\|\Phi(x) - \Phi(\varphi_v x)\|}{\|v\|} \leq C 2^{-J}\ .$$

- So, we want to *smooth* along the orbits.
- Local averaging within the translation orbit:

$$\Phi(x) = 2^{-dJ} \int_v \phi(2^{-J} v) \varphi_v x dv\ ,\ \left(\int \phi(v) dv = 1, \phi \geq 0 \right)\ .$$

Local invariants and convolution



- Local averaging within the translation orbit:

$$\Phi(x) = 2^{-dJ} \int_v \phi(2^{-J}v)\varphi_v x dv \ , \quad \left(\int \phi(v) dv = 1, \phi \geq 0 \right) \ .$$

- In coordinates, it becomes

$$\Phi(x)(u) = \int \phi_J(v) x(u-v) dv = x * \phi_J(u) \ , \text{with}$$

$$\phi_J(v) = 2^{-Jd} \phi(2^{-J}v)$$

Local average and stability

Proposition: The local averaging $\Phi(x) = x * \phi_J$ satisfies
$\forall \ \|x\| = 1 \in L^2 \ , \tau \ , \ \|\Phi(x) - \Phi(\varphi_\tau x)\| \leq C\|\tau\| \ .$

Local average and stability

Proposition: The local averaging $\Phi(x) = x * \phi_J$ satisfies $\forall \ \|x\| = 1 \in L^2 \ , \tau \ , \ \|\Phi(x) - \Phi(\varphi_\tau x)\| \leq C\|\tau\| \ .$

- Not surprising, since this operator removes the problematic high-frequencies.
- Are there other linear operators with the same property?

Average and uniqueness

- The only linear, translation-invariant operator is the average:

$$\forall\ v\ ,\ \ \Phi(x) = \Phi(\varphi_v x) \Longrightarrow \Phi(x) = \frac{1}{|G|}\int \Phi(\varphi_v x) dv$$

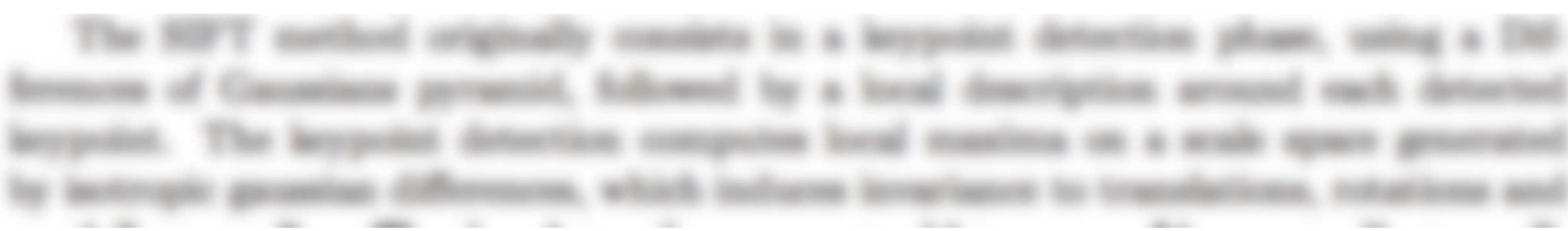
$$\Longrightarrow \Phi(x) = \Phi\left(\frac{1}{|G|}\int \varphi_v x dv\right) = \Phi\left(\frac{1}{|G|}\int x(u) du\right)\ .$$

- And a similar argument can be used locally.

From averages to Wavelets

- Low-pass information is insufficient:

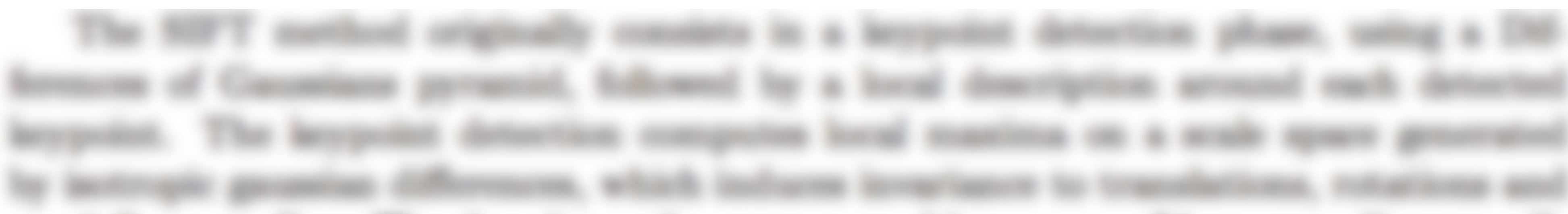
The SIFT method originally consists in a keypoint detection phase, using a Differences of Gaussians pyramid, followed by a local description around each detected keypoint. The keypoint detection computes local maxima on a scale space generated by isotropic gaussian differences, which induces invariance to translations, rotations and



From averages to Wavelets

- Low-pass information is insufficient:

The SIFT method originally consists in a keypoint detection phase, using a Differences of Gaussians pyramid, followed by a local description around each detected keypoint. The keypoint detection computes local maxima on a scale space generated by isotropic gaussian differences, which induces invariance to translations, rotations and

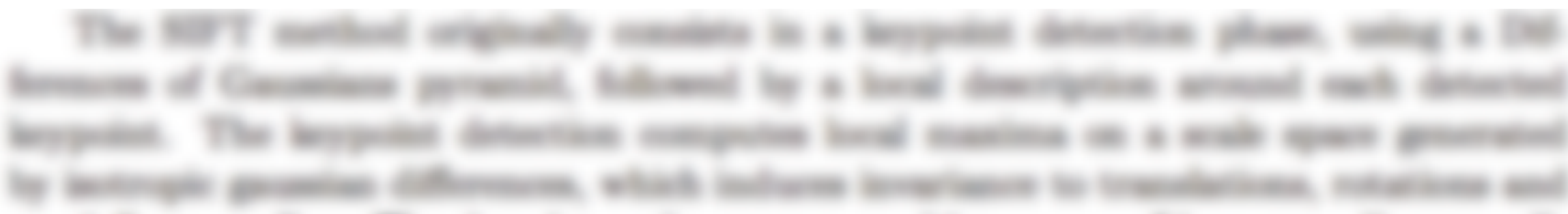


- Thus, we must capture high-frequency.
- These new measurements must involve a non-linearity.

From averages to Wavelets

- Low-pass information is insufficient:

The SIFT method originally consists in a keypoint detection phase, using a Differences of Gaussians pyramid, followed by a local description around each detected keypoint. The keypoint detection computes local maxima on a scale space generated by isotropic gaussian differences, which induces invariance to translations, rotations and

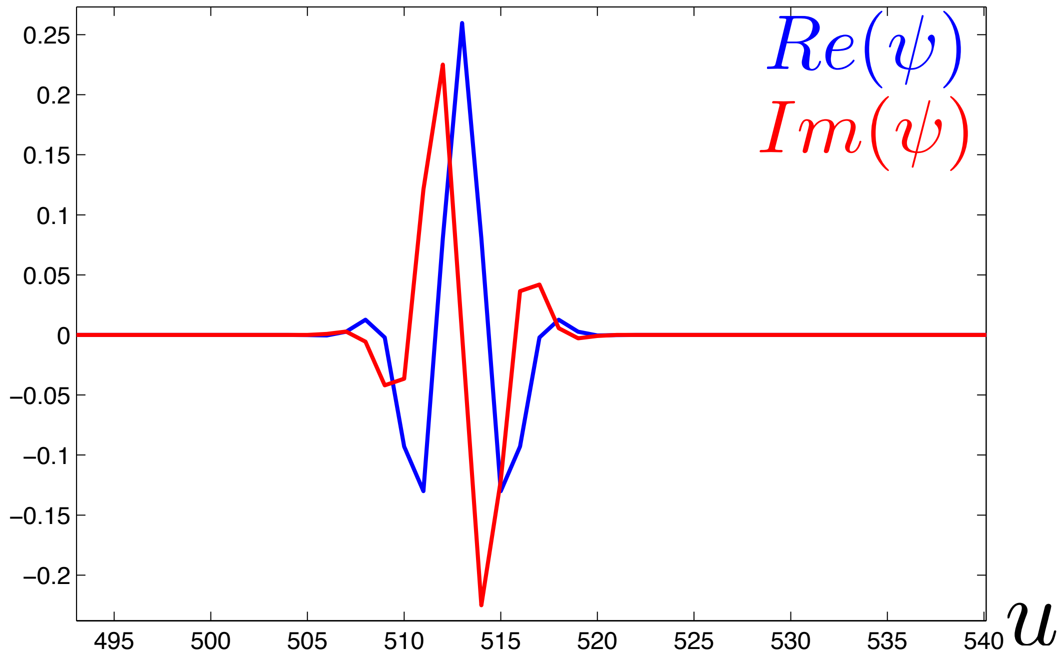


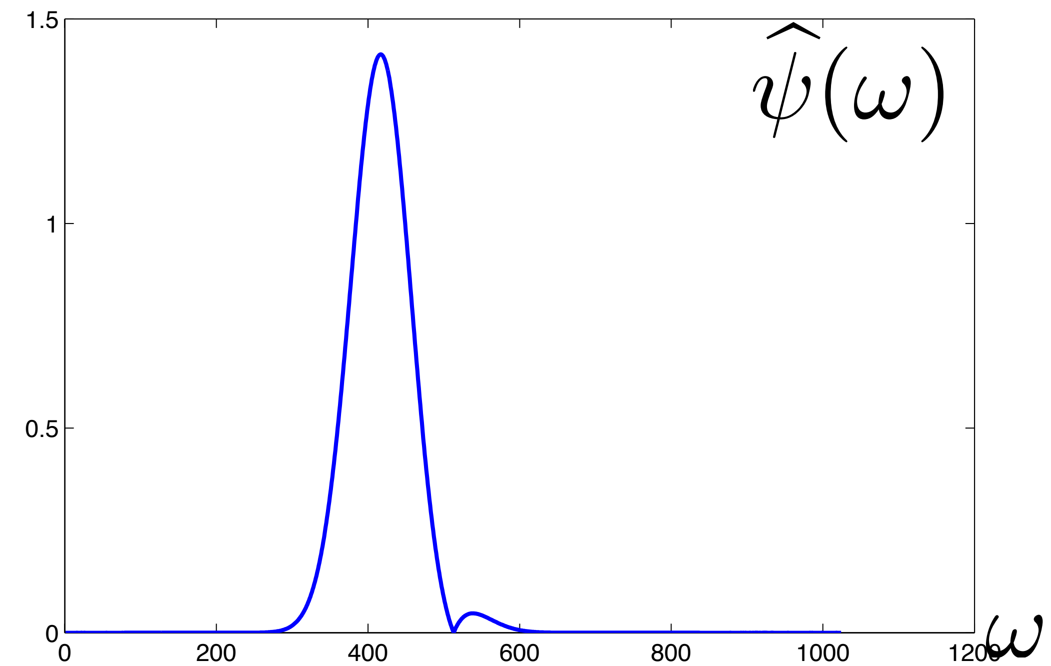
- Thus, we must capture high-frequency.
- These new measurements must involve a non-linearity.
- We want them to preserve stability to deformations.
- And we want them to preserve inter-class variability.

Wavelets

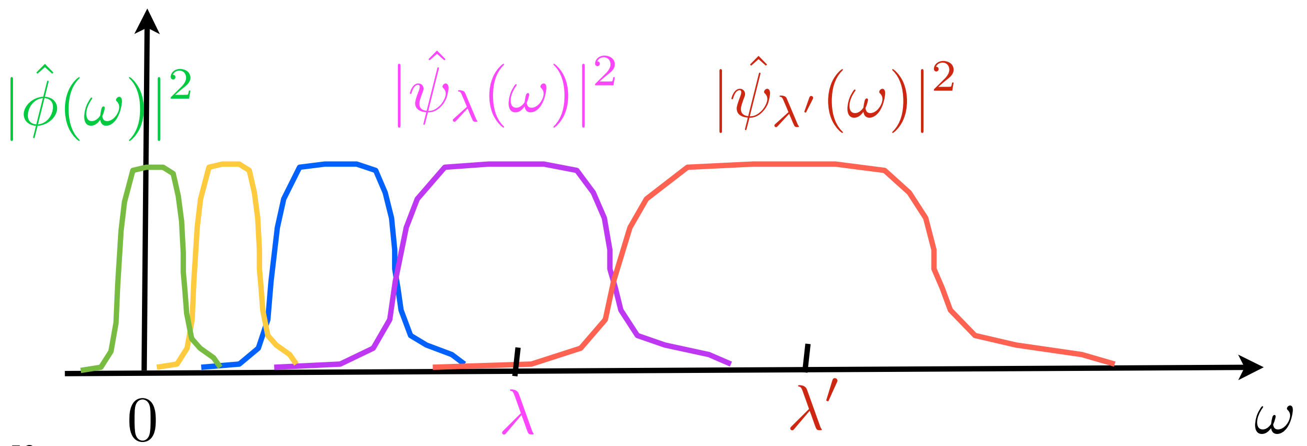
- ψ well localized in space and frequency.
- At least one vanishing moment: $$\int \psi(u)du = 0$$

Ex: Morlet wavelet





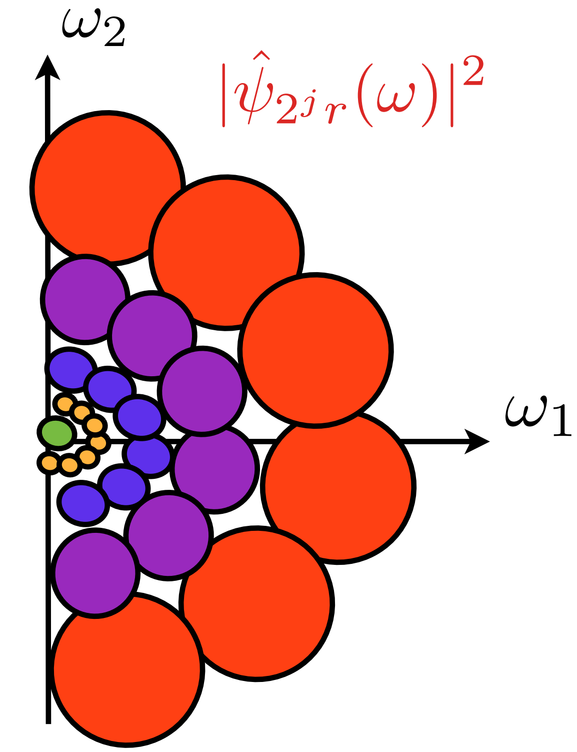
Dilated wavelets: $\psi_j(u) = 2^{-j}\psi(2^{-j}u)\ ,\ j \in \mathbb{Z}$



Littlewood-Paley Wavelet Filter Banks

- For images, dilated and rotated wavelets:

$$\psi_\lambda(u) = 2^{-j/2}\psi(2^{-j}ru)\ ,\ \text{with } \lambda = 2^j r$$



- Wavelet transform convolutional filter bank:

$$Wx = \{x \star \phi(u)\ ,\ x \star \psi_\lambda(u)\}_{\lambda \in \Lambda}$$

$$x \star \psi(u) = \int x(v)\psi(u-v)dv\ .$$

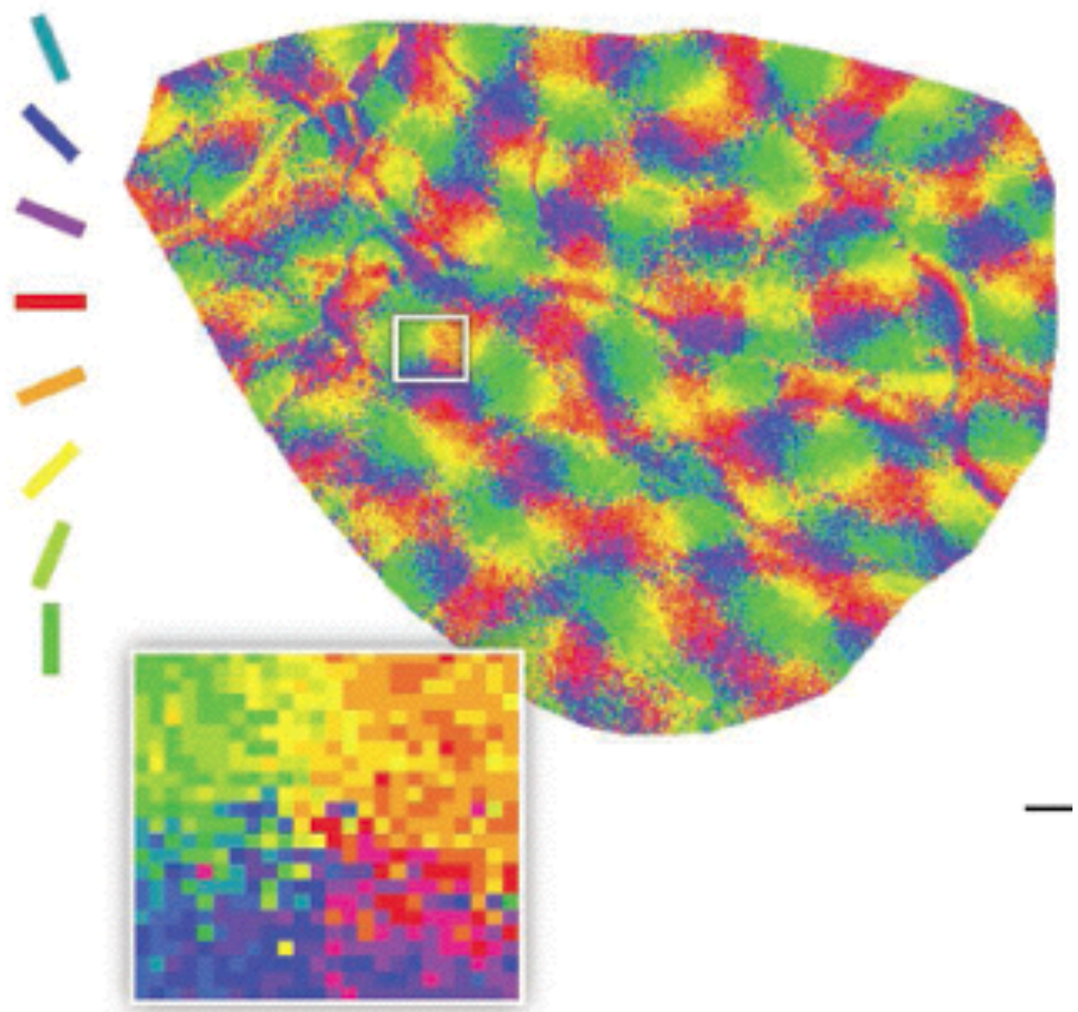
Theorem (Littlewood-Paley): If there exists $\delta > 0$ such that

$$\forall \omega > 0\ ,\ 1 - \delta \leq |\hat{\phi}(\omega)|^2 + \frac{1}{2}\sum_\lambda |\hat{\psi}(\lambda^{-1}\omega)|^2 \leq 1\ ,$$

then $\forall\, x \in L^2\ ,\ (1-\delta)\|x\|^2 \leq \|Wx\|^2 \leq \|x\|^2\ .$

Wavelets in Vision

- V1 Model of Simple and Complex cells: First layer of processing is selective in orientation, scale and position.



  - cells are organized in *pinwheels*. (more on that later).

Wavelets and learning

- Why are wavelets a good idea?
  - We will see that they provide stability to deformations because they *commute* nicely with diffeomorphisms:

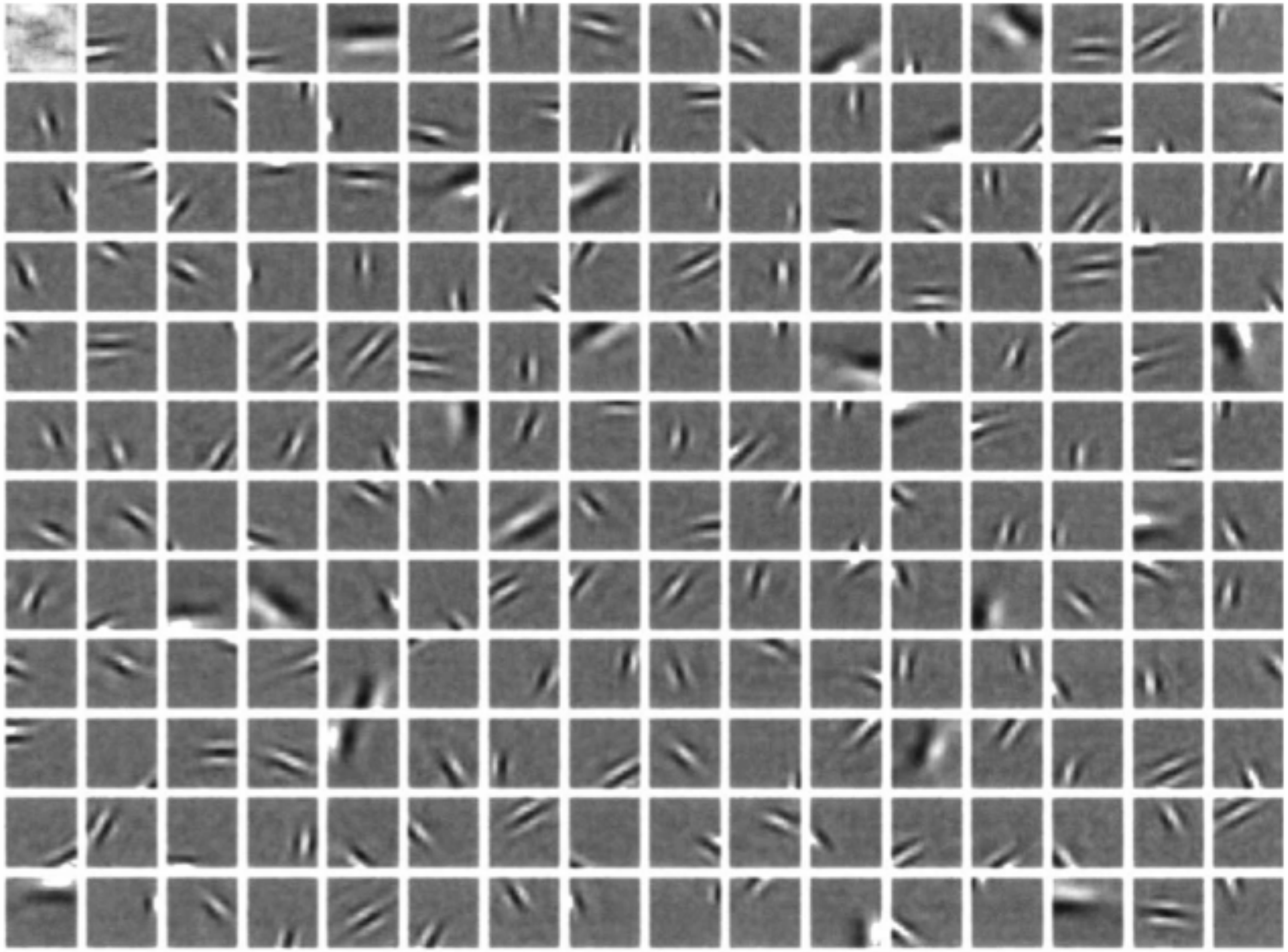
$$\|W\varphi_\tau x - \varphi_\tau W x\| \lesssim \|\tau\| \ .$$

  - We will also see that the discriminability of $\Phi(x) = \rho(Wx)$ is controlled by the *sparsity* produced by W:

$$\{x * \psi_\lambda(u)\}_{\lambda,u} \text{ has few non-zero coefficients.}$$

Examples

- Olshausen and Field Sparse coding model trained on natural images:



[Olshausen and Field,'96]

Examples

- Top performing shallow network unsupervised learning:

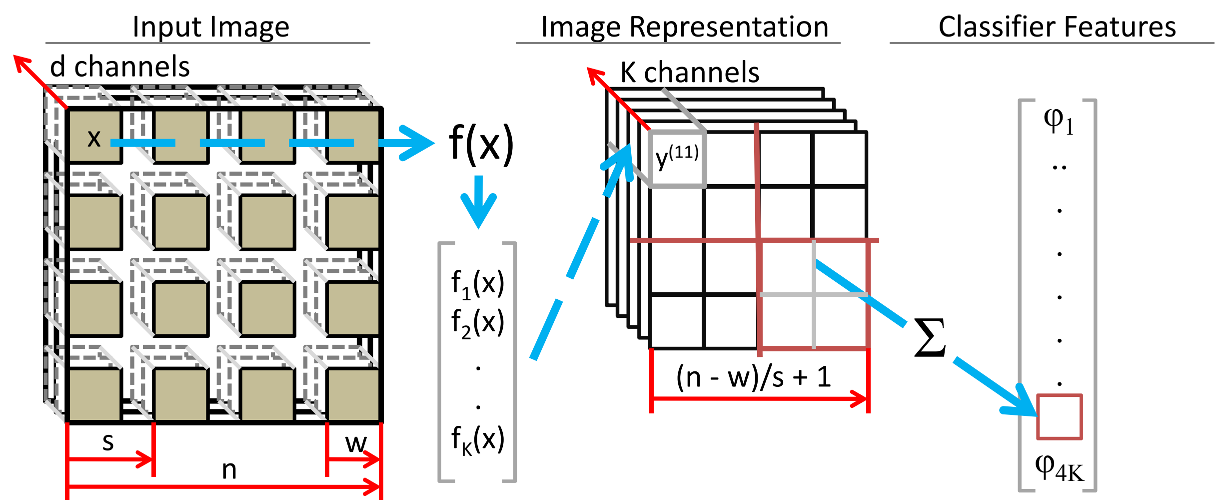
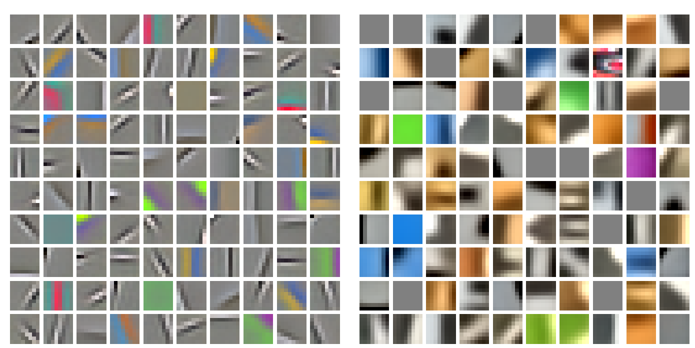
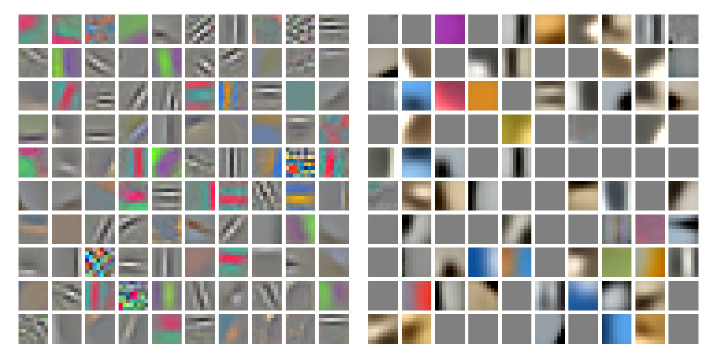


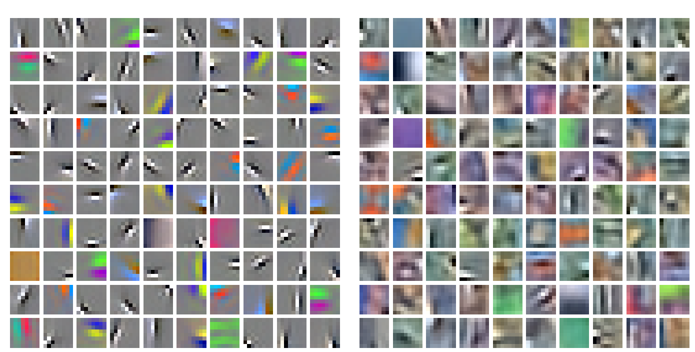
Figure 1: Illustration showing feature extraction using a w-by-w receptive field and stride s. We first extract w-by-w patches separated by s pixels each, then map them to K-dimensional feature vectors to form a new image representation. These vectors are then pooled over 4 quadrants of the image to form a feature vector for classification. (For clarity we have drawn the leftmost figure with a stride greater than w, but in practice the stride is almost always smaller than w.



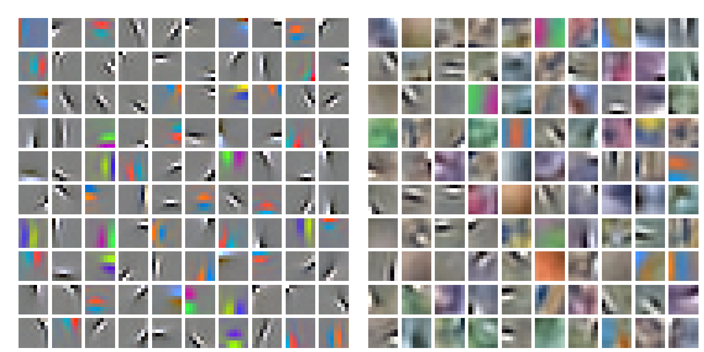
(a) K-means (with and without whitening)



(b) GMM (with and without whitening)



(c) Sparse Autoencoder (with and without whitening)



(d) Sparse RBM (with and without whitening)

[Coates, Le & Ng, '10]