Bayesian inference

image generation

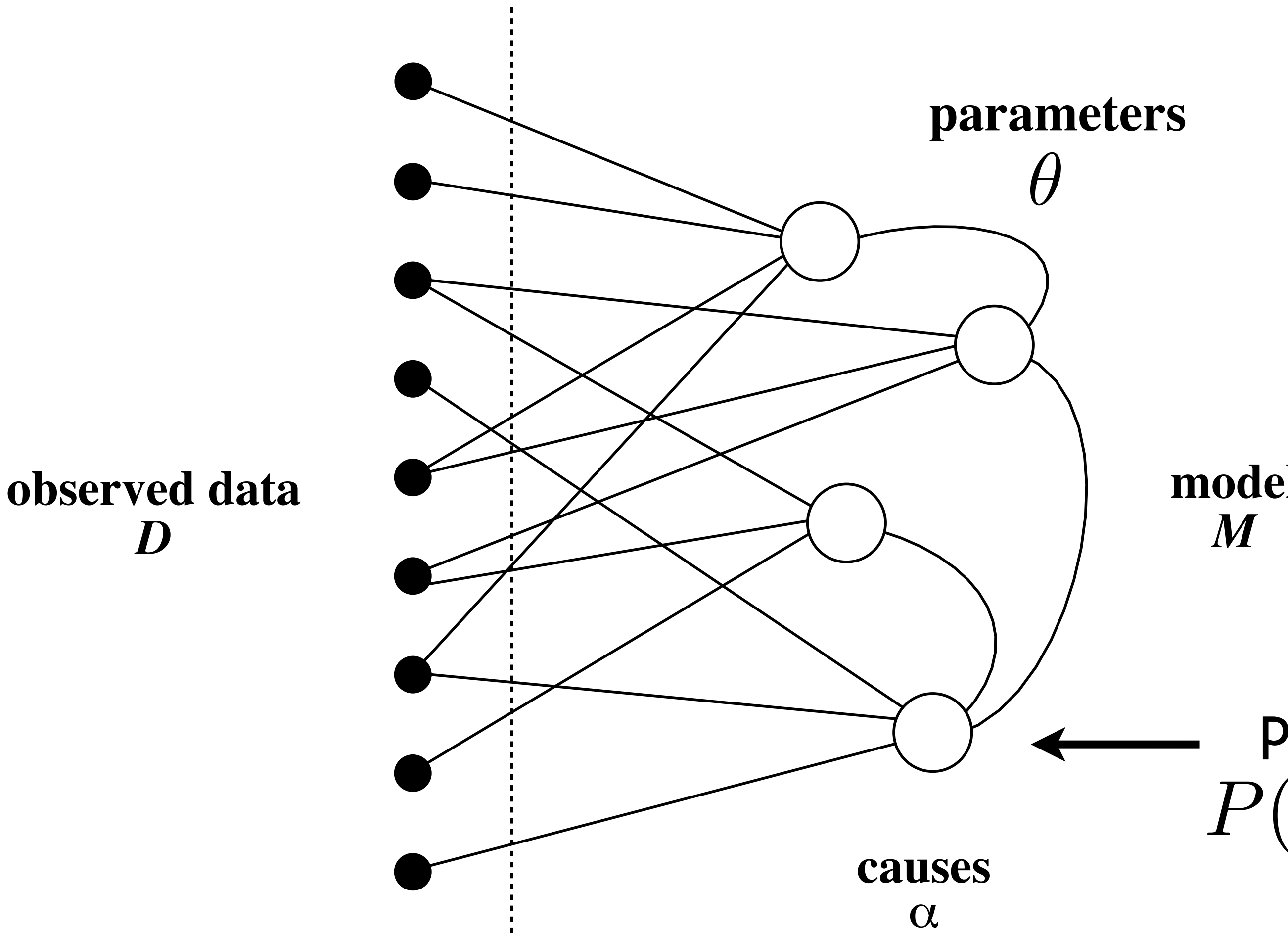
prior knowledge

$$P(H|D) = \frac{P(D|H)\,P(H)}{P(D)}$$

?

Generative models

inference $\longrightarrow$ $P(\alpha|D;\theta)$



prior $\longleftarrow$
$P(\alpha;\theta)$

$P(D|\alpha;\theta)$ $\longleftarrow$

explanation or prediction

Inference:

$$P(\alpha|D;\theta) = \frac{P(D|\alpha;\theta)\,P(\alpha;\theta)}{P(D|\theta)}$$

"Posterior"

Explanation or prediction:

$$P(D|\hat{\alpha};\theta) \quad \textbf{with} \quad \hat{\alpha} = \arg\max_{\alpha} P(\alpha|D;\theta)$$

Objective for learning:

$$\hat{\theta} = \arg\max_{\theta} \langle \log P(D|\theta) \rangle$$

"Log likelihood"

$$P(D|\theta) = \sum_{\alpha} P(D|\alpha;\theta)\,P(\alpha;\theta)$$

We can keep on going…

$$P(\theta|D) = \frac{\overbrace{P(D|\theta)}^{\text{likelihood}}\,\overbrace{P(\theta)}^{\text{prior}}}{\underbrace{P(D)}_{\text{evidence}}}$$

$$P(D) = \int P(D|\theta)\,P(\theta)\,d\theta$$

David MacKay Ph.D. thesis (1991)

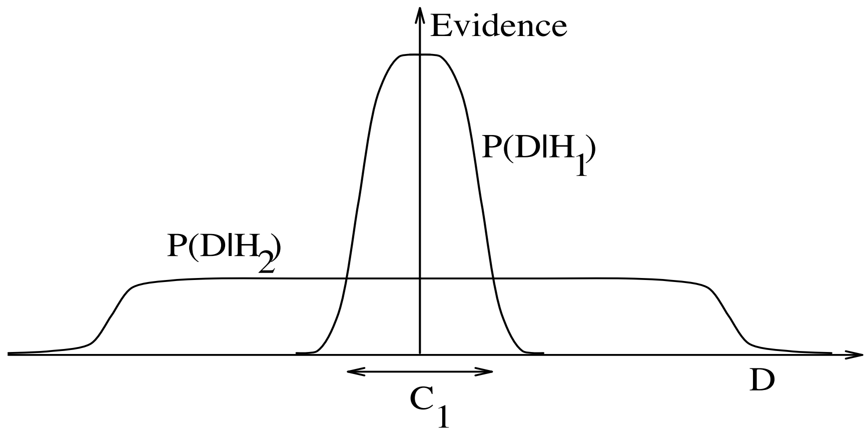


Figure 2.2: **Why Bayes embodies Occam's razor**

This figure gives the basic intuition for why complex models are penalised. The horizontal axis represents the space of possible data sets D. Bayes' rule rewards models in proportion to how much they *predicted* the data that occurred. These predictions are quantified by a normalised probability distribution on D. In this paper, this probability of the data given model $\mathcal{H}_i$, $P(D|\mathcal{H}_i)$, is called the evidence for $\mathcal{H}_i$.

A simple model $\mathcal{H}_1$ makes only a limited range of predictions, shown by $P(D|\mathcal{H}_1)$; a more powerful model $\mathcal{H}_2$, that has, for example, more free parameters than $\mathcal{H}_1$, is able to predict a greater variety of data sets. This means however that $\mathcal{H}_2$ does not predict the data sets in region $\mathcal{C}_1$ as strongly as $\mathcal{H}_1$. Assume that equal prior probabilities have been assigned to the two models. Then if the data set falls in region $\mathcal{C}_1$, the *less powerful* model $\mathcal{H}_1$ will be the *more probable* model.

David MacKay Ph.D. thesis (1991)

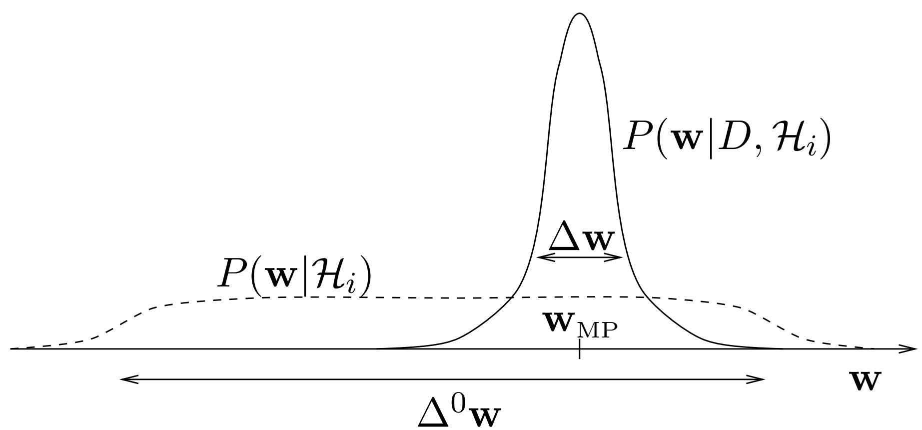
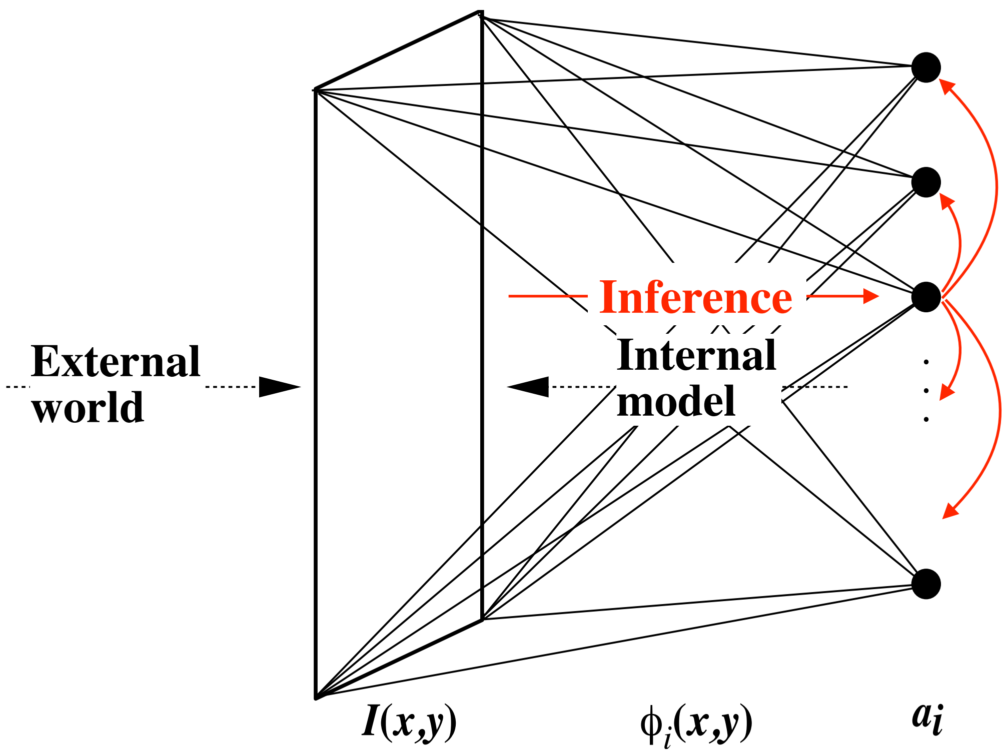


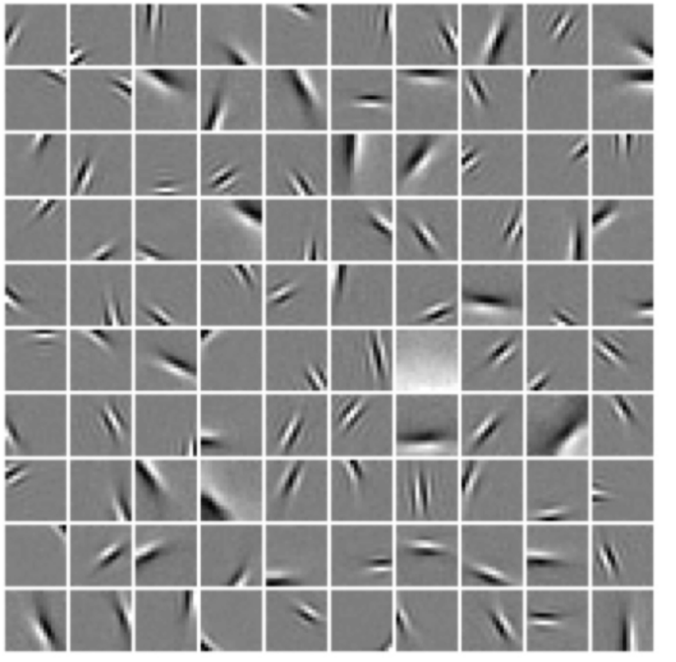
Figure 2.3: **The Occam factor**
This figure shows the quantities that determine the Occam factor for a hypothesis $\mathcal{H}_i$ having a single parameter $\mathbf{w}$. The prior distribution (dotted line) for the parameter has width $\Delta^0\mathbf{w}$. The posterior distribution (solid line) has a single peak at $\mathbf{w}_{\mathrm{MP}}$ with characteristic width $\Delta\mathbf{w}$. The Occam factor is $\frac{\Delta\mathbf{w}}{\Delta^0\mathbf{w}}$.

$$\underset{\text{Evidence}}{P(D\,|\mathcal{H}_i)} \simeq \underbrace{P(D\,|\mathbf{w}_{\mathrm{MP}},\mathcal{H}_i)}_{\text{Best fit likelihood}}\ \underbrace{P(\mathbf{w}_{\mathrm{MP}}|\mathcal{H}_i)\,\Delta\mathbf{w}}_{\text{Occam factor}}. \tag{2.5}$$

$$\boxed{\text{Occam factor} = \frac{\Delta\mathbf{w}}{\Delta^0\mathbf{w}}}$$

Sparse coding model





$\phi_i(x,y)$

Inference: $$P(\mathbf{a}|\mathbf{I};\mathbf{\Phi}) \propto P(\mathbf{I}|\mathbf{a};\mathbf{\Phi})\, P(\mathbf{a})$$

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} |\mathbf{I} - \Phi\,\mathbf{a}|^2 + \lambda \sum_i C(a_i)$$

Learning: $$\Delta\mathbf{\Phi} \propto \frac{d}{d\mathbf{\Phi}} \log \int P(\mathbf{I}|\mathbf{a};\mathbf{\Phi})\, P(\mathbf{a})\, d\mathbf{a}$$

NOISE REMOVAL VIA BAYESIAN WAVELET CORING

Eero P. Simoncelli

Computer and Information Science Dept.
University of Pennsylvania
Philadelphia, PA 19104

Edward H. Adelson

Brain and Cognitive Science Dept.
Massachusetts Institute of Technology
Cambridge, MA 02139

The classical solution to the noise removal problem is the Wiener filter, which utilizes the second-order statistics of the Fourier decomposition. Subband decompositions of natural images have significantly non-Gaussian higher-order point statistics; these statistics capture image properties that elude Fourier-based techniques. We develop a Bayesian estimator that is a natural extension of the Wiener solution, and that exploits these higher-order statistics. The resulting nonlinear estimator performs a "coring" operation. We provide a simple model for the subband statistics, and use it to develop a semi-blind noise-removal algorithm based on a steerable wavelet pyramid.

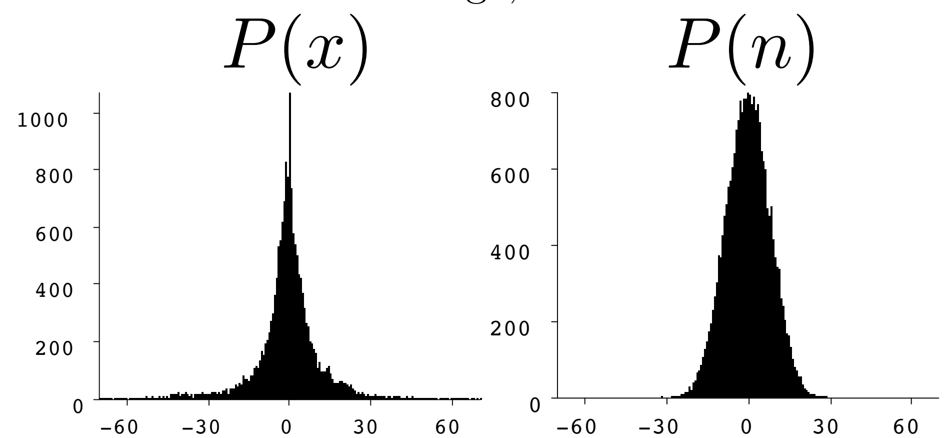


Figure 1: Histograms of a mid-frequency subband in an octave-bandwidth wavelet decomposition for two different images. Left: The "Einstein" image. Right: A white noise image with uniform pdf.

$$y = x + n$$

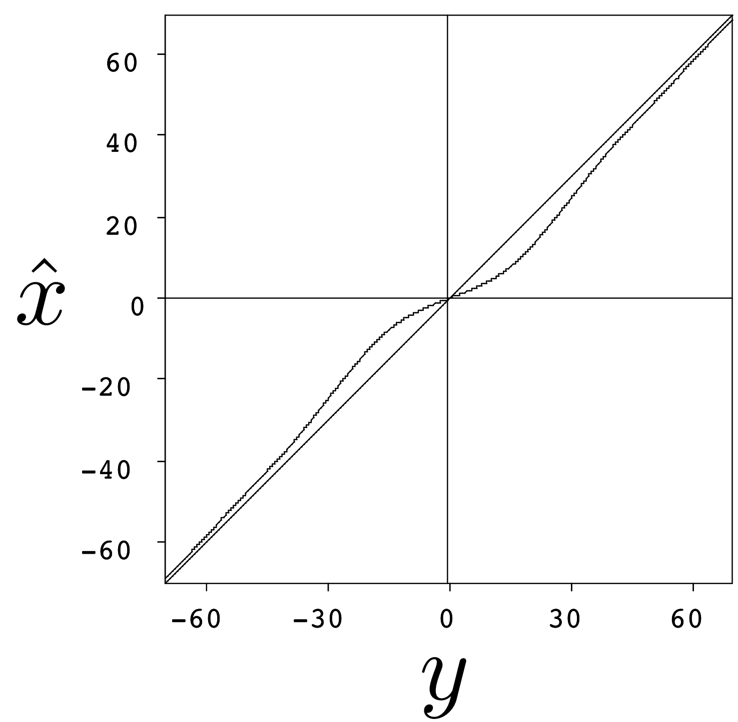
$$y = x + n$$

$$P(x) = \frac{1}{Z_s} e^{-|\frac{x}{s}|^p}$$

$$P(x|y) \propto P(y|x)\, P(x)$$

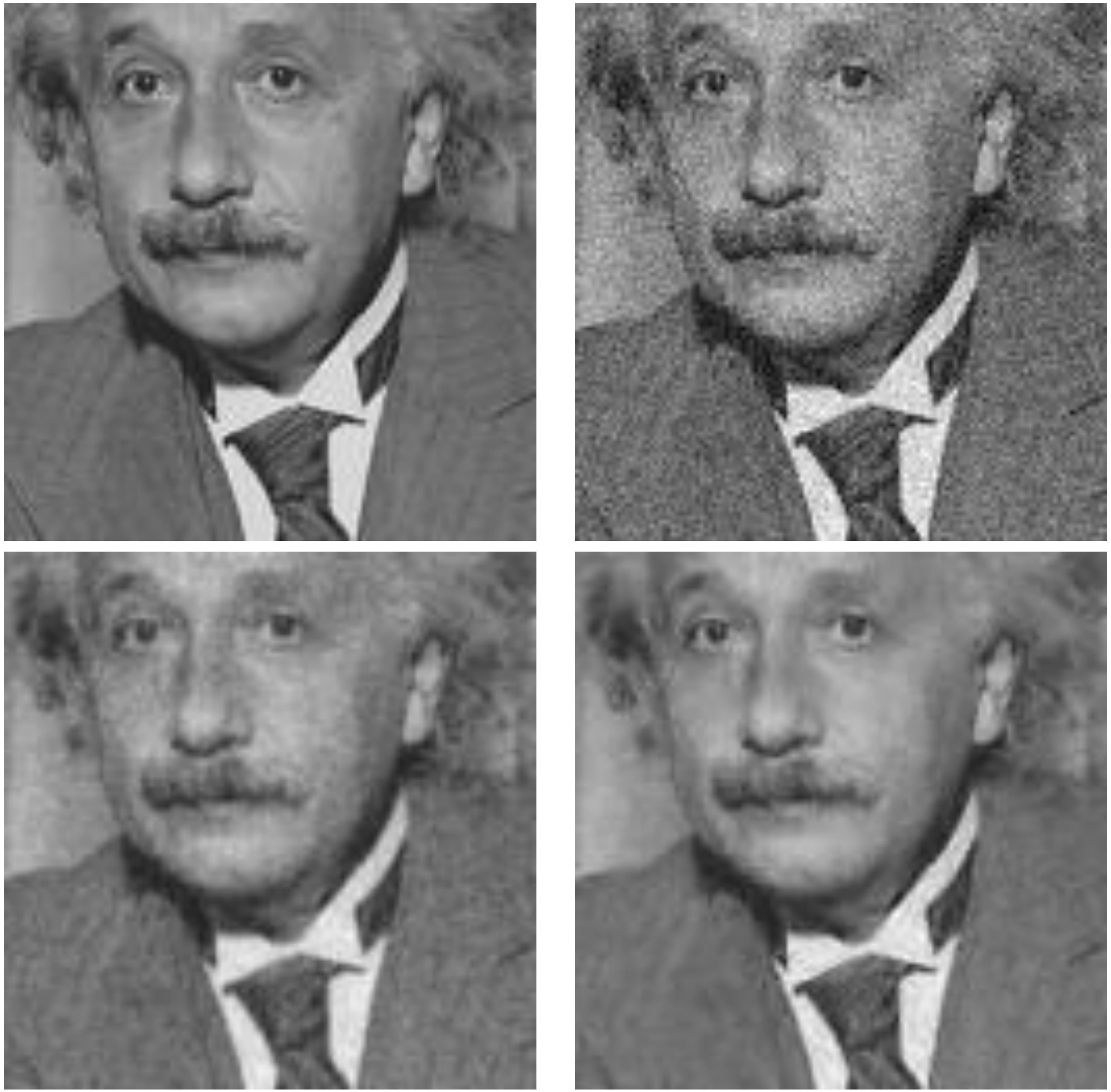
MAP estimate:

$$\hat{x} = \arg\min_x \left[\frac{|y - x|^2}{2\sigma_n^2} + |\frac{x}{s}|^p \right]$$



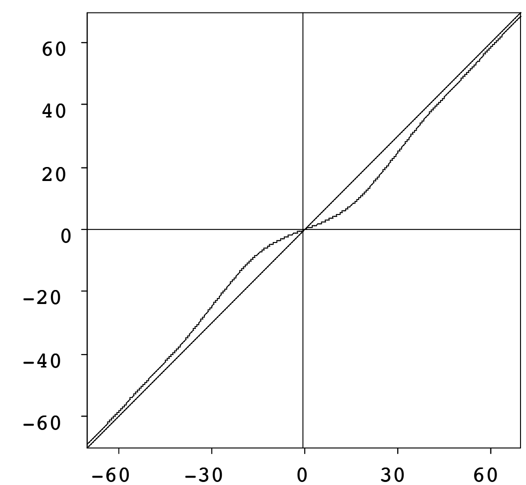
original image

image + noise

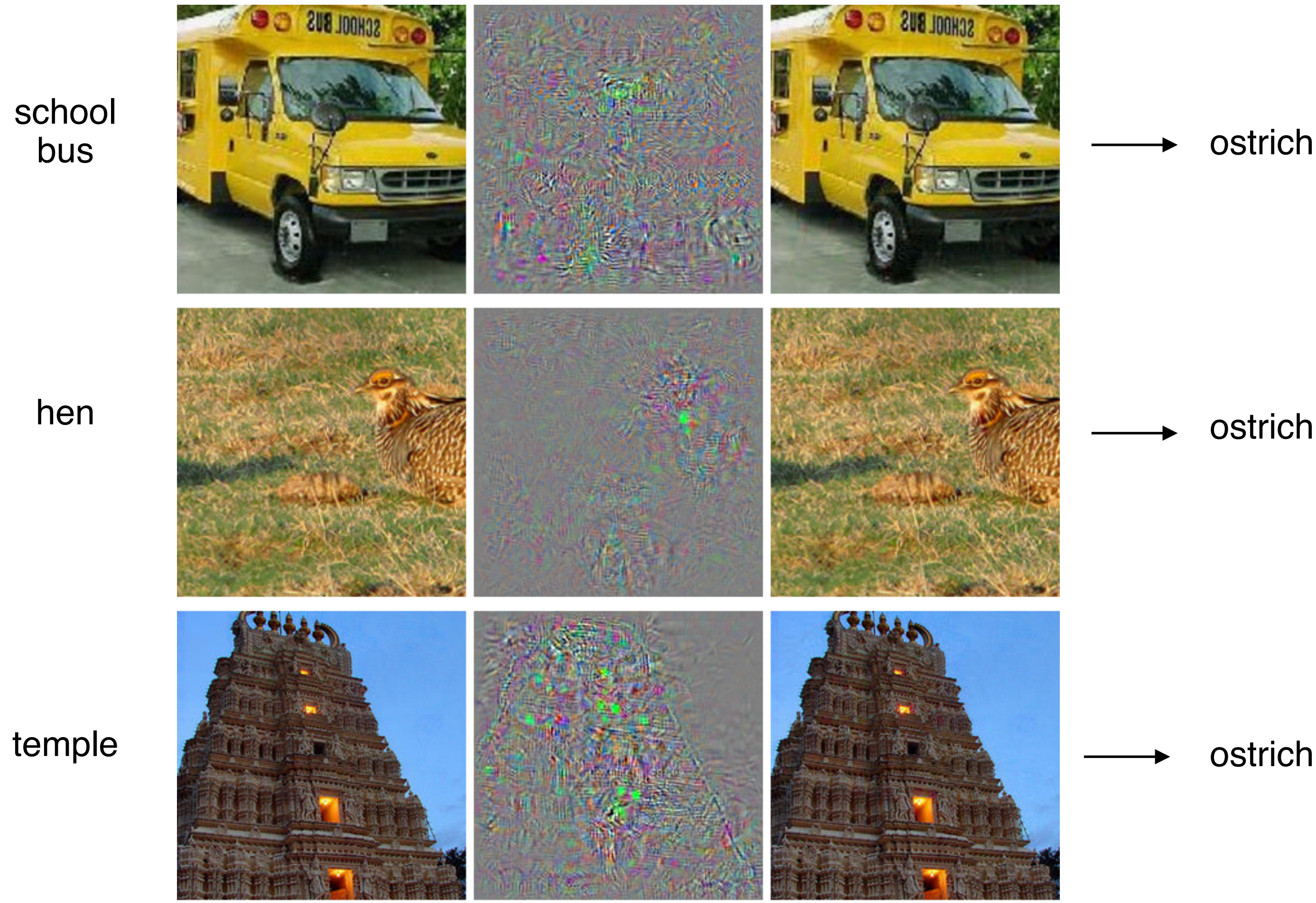


Wiener filter

wavelet coring



Deep convnets are easily fooled by imperceptible perturbations (adversarial examples)



Szegedy et al. (2013)

Sparse inference protects against adversarial attack

(Paiton, Frye, Lundquist, Bowen, Zarcone & Olshausen 2020)