CS 103: Representation Learning, Information Theory and Control

Lecture 8, Mar 1, 2019

Recap

Group nuisances

- Group convolutions
- Canonical reference frames
- SIFT descriptors

General nuisances

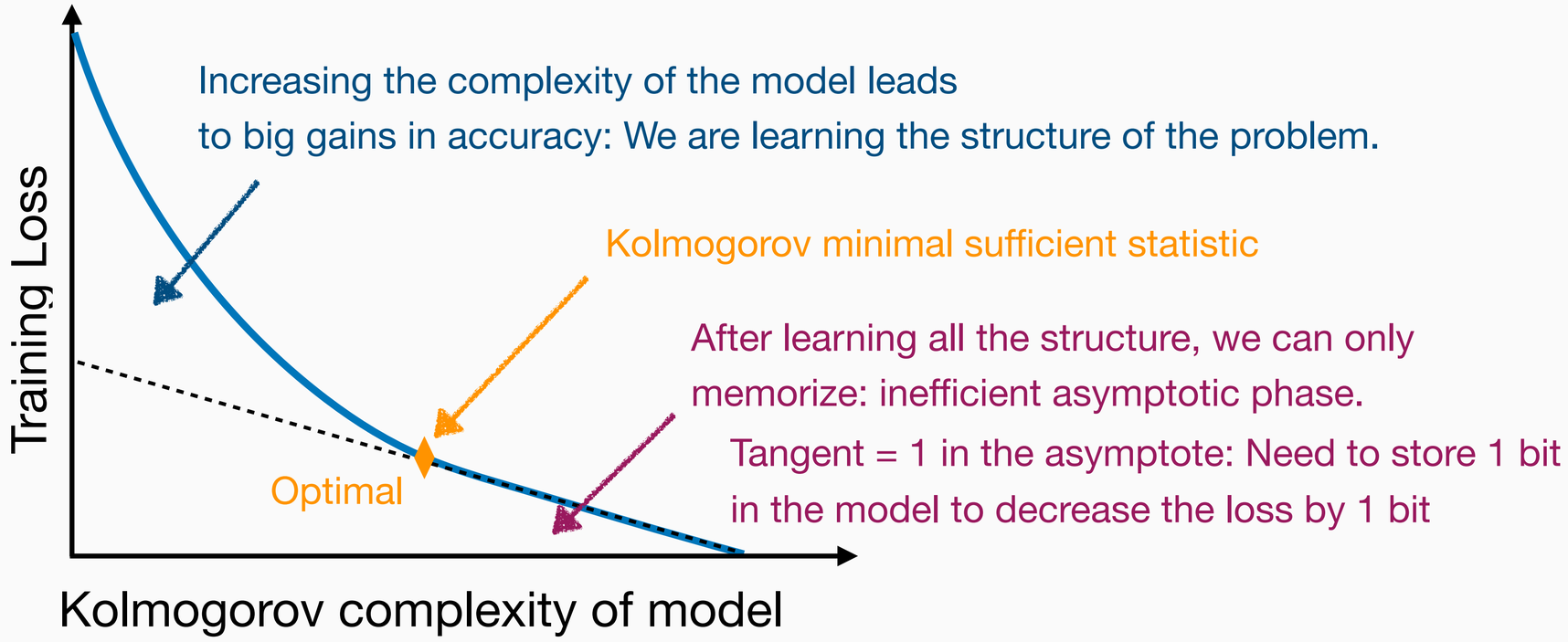
- Minimal information in activation ⇒ Invariance to nuisances
- Information Bottleneck
- IB loss can upper-bounded by introducing an auxiliary variable
- Aside: Variational Auto-Encoder can be seen as a particular case
- Aside: Disentanglement in VAE

How does this relate to standard deep learning?

The Kolmogorov Structure of a Task

How can we define the structure of a task? Define the Kolmogorov Structure Function:

$$S_{\mathscr{D}}(t) = \min_{K(M) \leq t} L(\mathscr{D}; M)$$



Kolmogorov's Structure Functions and Model Selection, Vereshchagin and Vitanyi, *2002*
Information Complexity of Tasks, their Structure and their Distance, Achille et al., *2018*

Optimizing using Deep Neural Networks

How do we find the optimal solution?

$$S_{\mathscr{D}}(t) = \min_{K(M)<t} L(\mathscr{D}; M)$$

↓ Corresponding Lagrangian

$$\mathscr{L}(M) = L(\mathscr{D}; M) + \lambda K(M)$$

↓ Let w be the parameters of the model

↓ Use the bound $K(M) \leq \mathrm{KL}(q(w \mid \mathscr{D}) \| p(w))$

$$\mathscr{L}(M) = L(\mathscr{D}; M) + \lambda\, \mathrm{KL}(q(w \mid \mathscr{D}) \| p(w))$$

This loss can be implemented using a DNN and the local reparametrization trick.*

* *Variational Dropout and the Local Reparameterization Trick*, Kingma et al., *2015*

Information Complexity of Tasks, their Structure and their Distance, Achille et al., *2018*

Let’s rewrite it using Information Theory

We used an upperbound, what is the best we value it can assume?

$$\mathscr{L}(M) = \mathbb{E}_{w\sim q(w|\mathscr{D})}[H_{p,q}(\mathscr{D}\,|\,w)] + \lambda\,\mathrm{KL}(q(w\,|\,\mathscr{D})\|p(w))\,.$$

Recall that:

$$I(w;\mathscr{D}) \leq \mathbb{E}_{\mathscr{D}}[\mathrm{KL}(q(w\,|\,\mathscr{D})\|p(w))],$$

which is obtained when *p(w) = q(w|D)*. Hence, on expectation over the datasets, the best function loss function to use to recover the task structure is:

$$\mathscr{L}(M) = \mathbb{E}_{\mathscr{D}}[H(\mathscr{D}\,|\,w)] + \lambda I(w;\mathscr{D})\,.$$

IB Lagrangian for the weights

A new Information Bottleneck

Weights IB
Overfitting

D → w → $p(y|x)$

dataset — weights — real distribution

$$\min_{w} \mathcal{L} = H_{p,q_w}(y|z) + \beta I(\mathcal{D}; w)$$

Activations IB
Invariance

x → z → y

data — activations — label

$$\min_{q(z|x)} \mathcal{L} = H_{p,q}(y|z) + \beta I(z; x)$$

The PAC-Bayes generalization bound

Catoni, 2007; McAllester 2013

PAC-Bayes bound (Catoni, 2007; McAllester 2013).

$$L_{\text{test}}(q(w|\mathcal{D})) \leq \frac{1}{N(1-1/2\beta)} \underbrace{\left[H_{p,q}(y|x,w) + \beta\ \mathrm{KL}(q(w|\mathcal{D})\|p(w))\right]}_{\text{IB Lagrangian for the weights}}$$

Corollary. Minimizing the IB Lagrangian for the weights minimizes an upper bound on the test error.

This gives non-vacuous generalization bounds! (Dziugaite and Roy, 2017)

Can we really minimize the IBL for the weights?

We are making an approximation by assuming that both q(w|D) and p(w) are Gaussian.

Let *w** be a local minimum. The optimal amount of gaussian noise is to add is:

$$\Sigma = \left(I + \frac{2\lambda^2}{\beta} F(w_0)\right)^{-1},$$

where *F(w*)* is the Fisher Information Matrix (equiv. Hessian) computed in *w**.

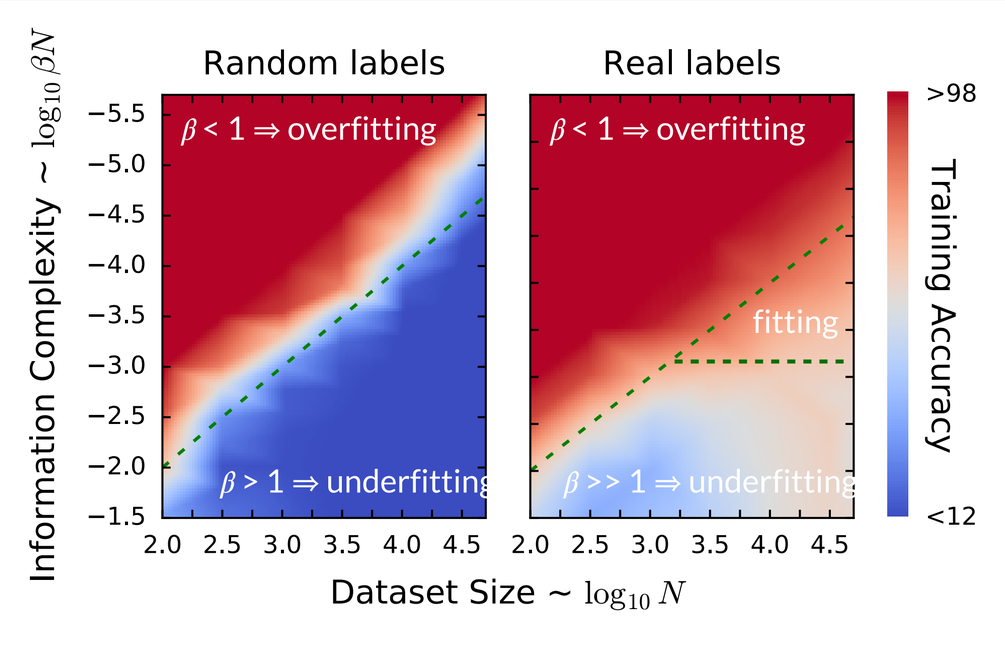
$$I(w; \mathcal{D}) \leq \frac{\|w\|^2}{\lambda^2} + \log |2\lambda^2 N F(w^*) + I|$$

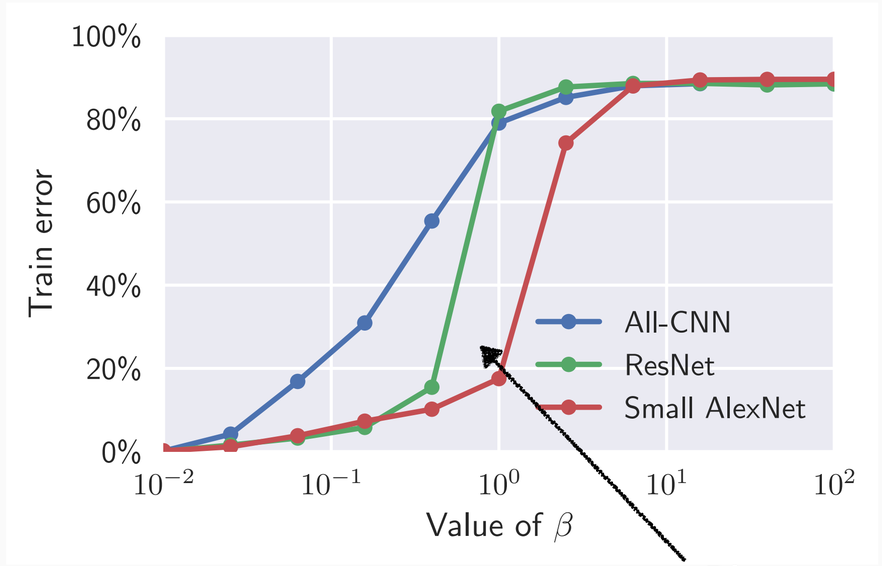
Weight Information is bounded by the geometry of the loss landscape*

Flat minima have low information in the weights.

Is this approximation good? Phase transition

Consider a dataset with random labels. There is no structure, so for $\lambda < 1$we should not fit anything, and for $\lambda > 1$ we should memorize everything (see the structure function).



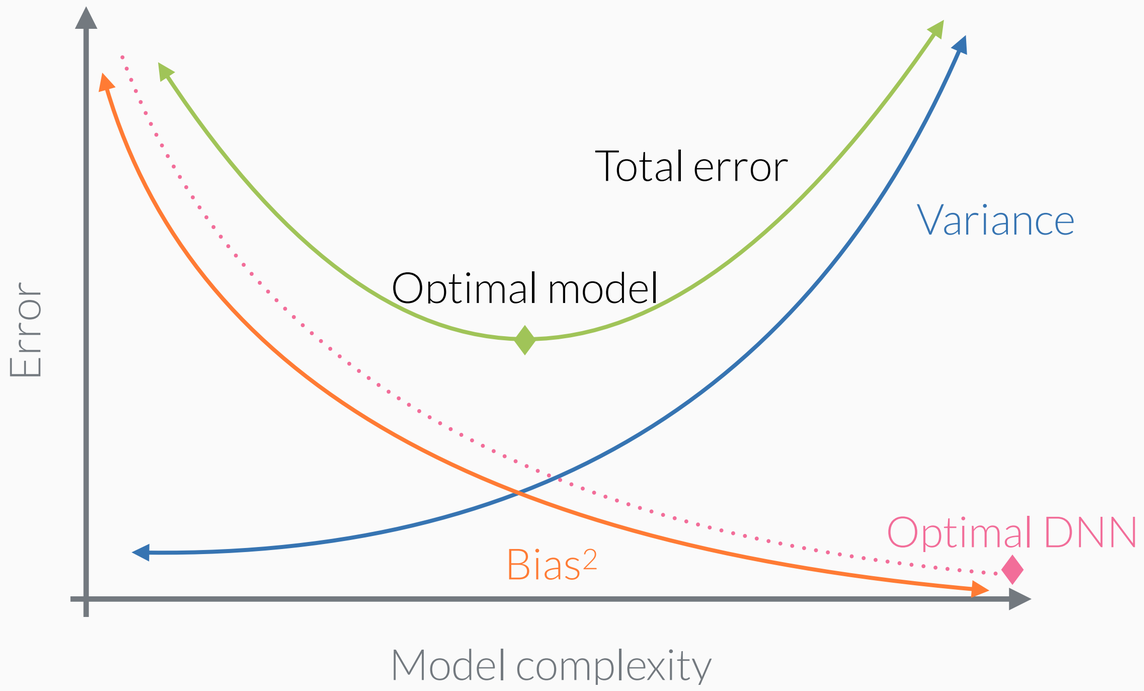


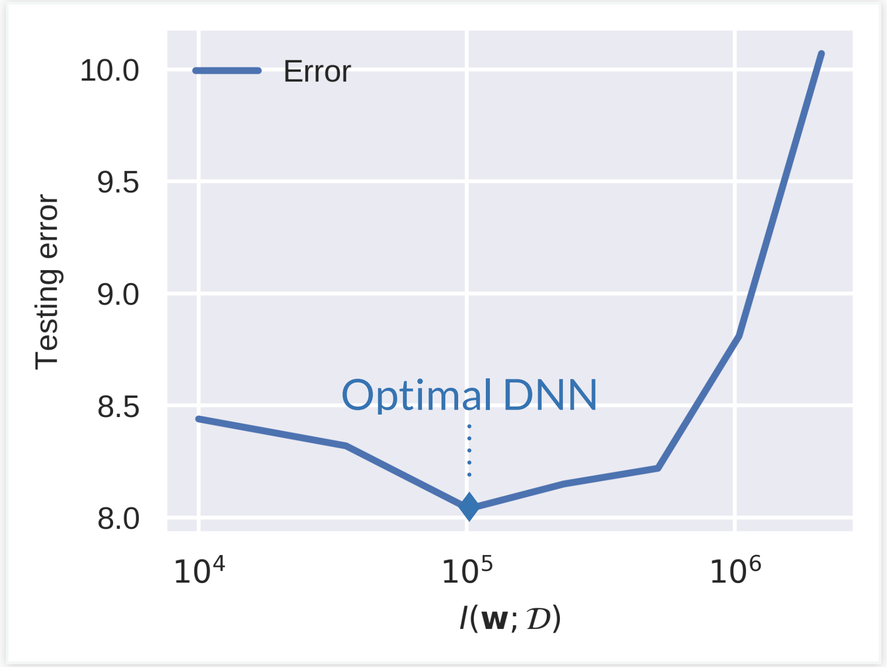
Even with local approximation, we can observe this behavior in real deep networks.
For real label, we have a “Goldilocks Zone” where we fit without overfitting for $\lambda > 1$.

Achille and Soatto, *Emergence of Invariance and Disentanglement in Deep Representations*, JMLR 2018

Bias-variance tradeoff

Information is a better measure of complexity than number of parameters





Information complexity

Parametrizing the complexity with information in the weights, we recover bias-variance trade-off trend.

Achille and Soatto, *Emergence of Invariance and Disentanglement in Deep Representations*, JMLR 2018

Arora et al., *Stronger generalization bounds for deep nets via a compression approach*, ICML 2018

A new Information Bottleneck

Weights IB
Overfitting

$D \longrightarrow w \longrightarrow p(y|x)$

dataset — weights — real distribution

$$\min_{w} \mathcal{L} = H_{p,q_w}(y|z) + \beta I(\mathcal{D}; w)$$

Activations IB
Invariance

$x \longrightarrow z \longrightarrow y$

data — activations — label

$$\min_{q(z|x)} \mathcal{L} = H_{p,q}(y|z) + \beta I(z; x)$$

The Emergence Bound

Info in activations

$$I(z;x) + TC(z)$$

Minimality of the weights (representation of the training set) induces minimality (hence **invariance**) and **disentanglement** of the activations.

Tight for one layer; for more layers we have:

$$I(z_L;x) \leq \min_{k<L}\{\dim(\mathrm{z_k})\left[\mathrm{g}(\frac{\mathrm{I}(\mathrm{W^k};\mathcal{D})}{\dim(\mathrm{W^k})}) + 1\right]\}$$

Achille and Soatto, *Emergence of Invariance and Disentanglement in Deep Representations*, JMLR 2018

How do minimize the information in the weights?

1. Explicitly minimize the IBL for the weights using local reparametrization trick.
2. Let SGD do it for you:
   - Empirically we know that SGD tend to find flatter minima.
   - We know from the local information bound that flat minima have less information in the weights.
   - Hence, SGD implicitly minimizes the information in the weights.
3. Modify SGD to reduce information more aggressively.

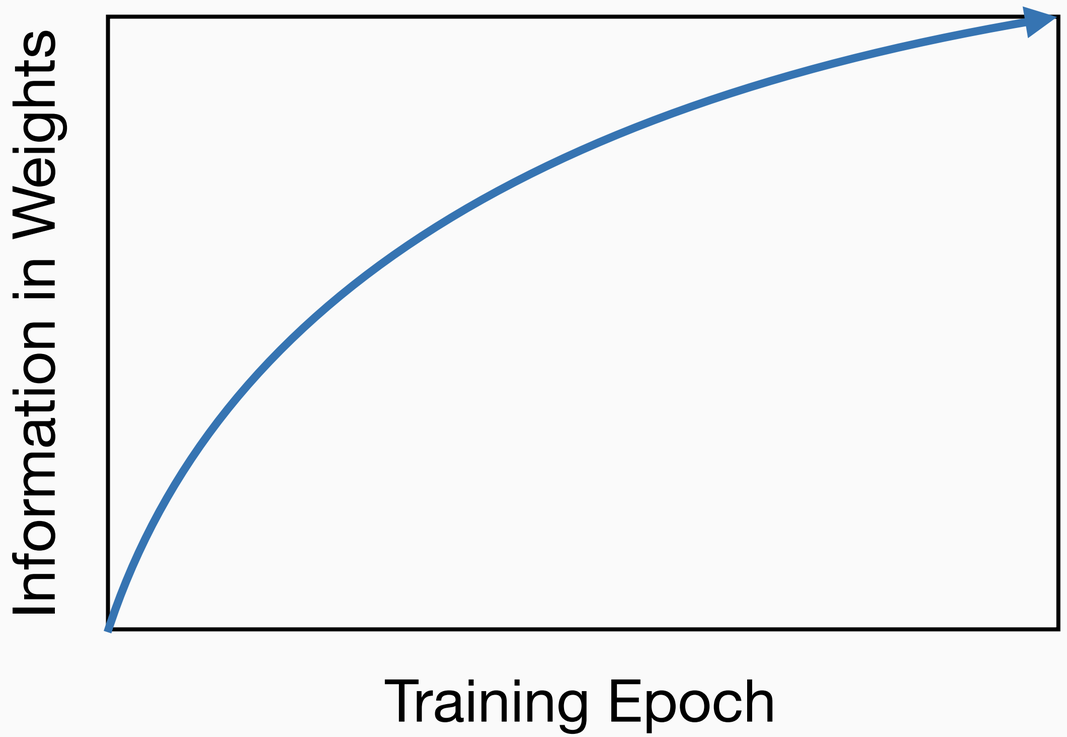
Counterpoint: Are flat minima the whole story?

Warning: These are still open questions and the claims are not proved.

Information in Weights during training

What should we expect from the information in the weights during training?

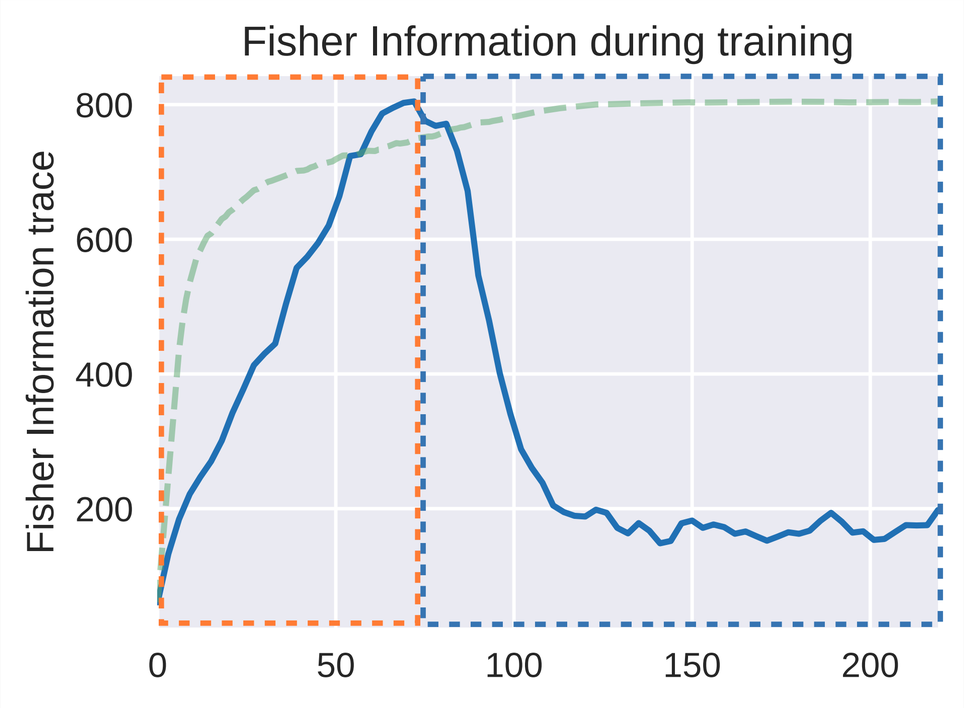
Maybe something like this?



Information during training

Information extraction

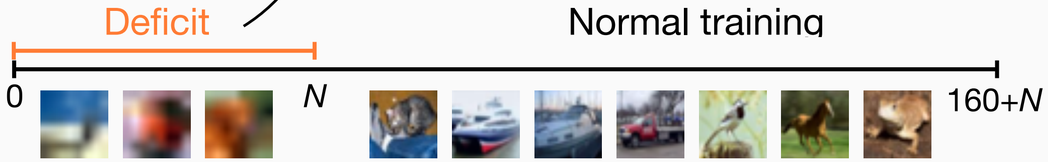
Information consolidation



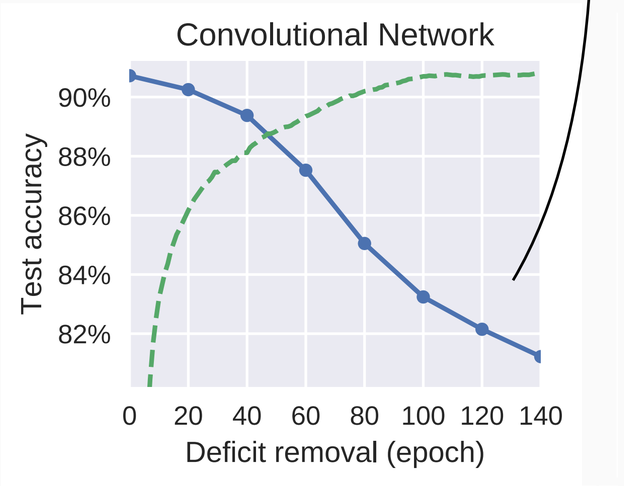
Achille, Rovere, Soatto, *Critical Learning Periods in Deep Networks, 2018*

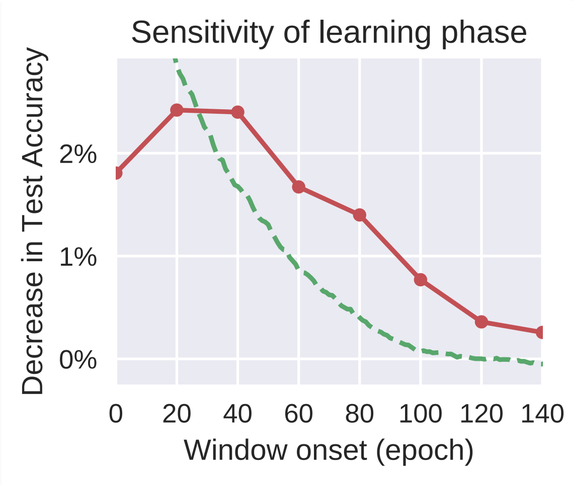
A snag: Critical periods in Deep Networks

Show network blurred images to simulate cataract



The network does not classify correctly if the deficit is removed to late





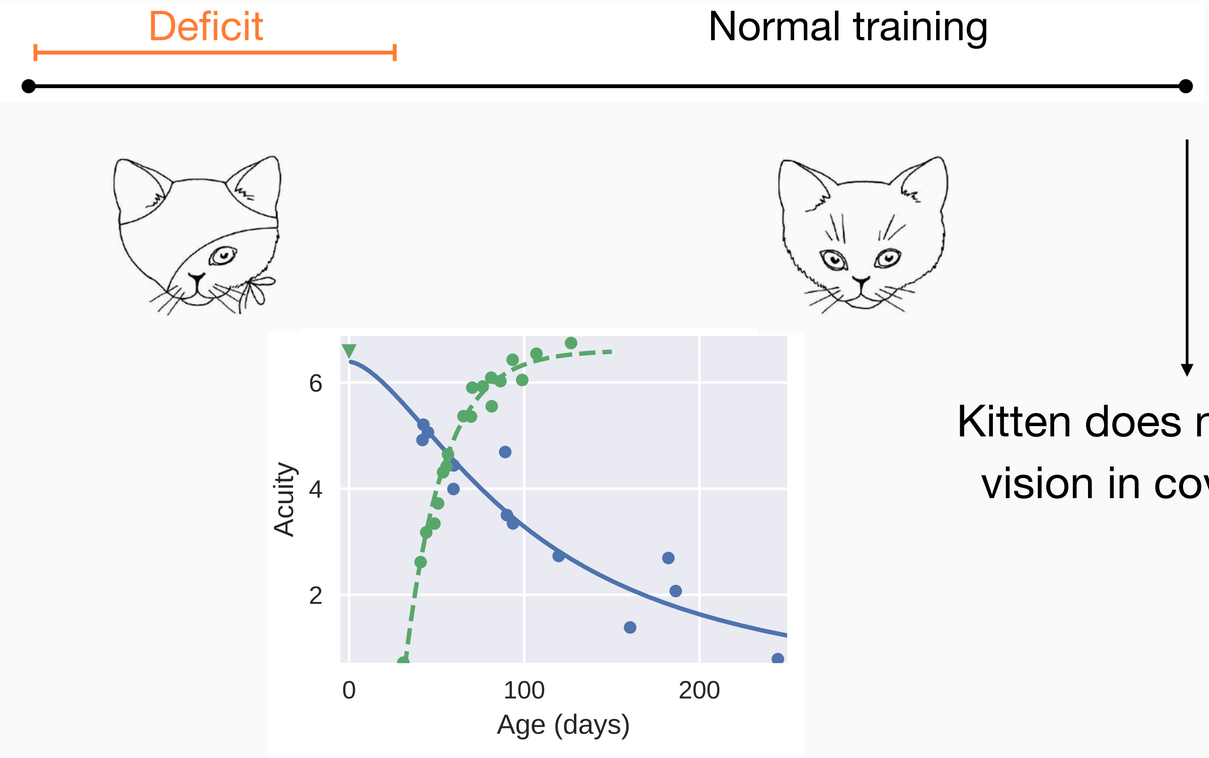
A short deficit at epoch ~40 is enough to permanently damage the network.

Achille, Rovere, Soatto, *Critical Learning Periods in Deep Networks, 2018*

Critical periods

Critical periods. A time-period in early development where sensory deficits can permanently impair the acquisition of a skill

Examples: *monocular deprivation, cataracts, imprinting, language acquisition, …*

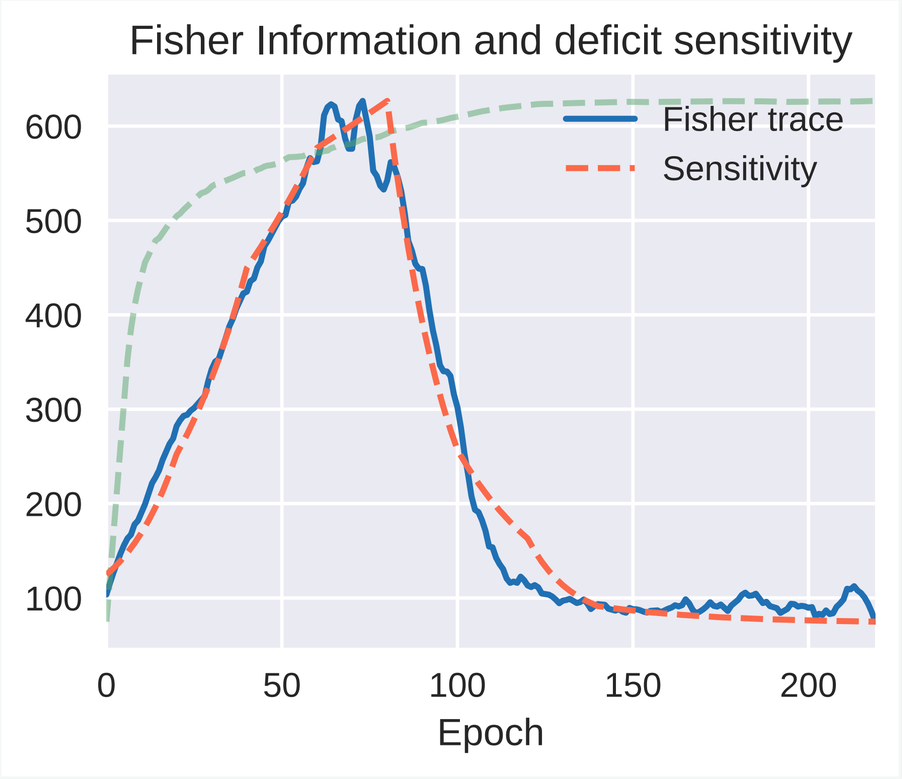


Hubel and Wiesel

Image from *Cnops et al.*, 2008

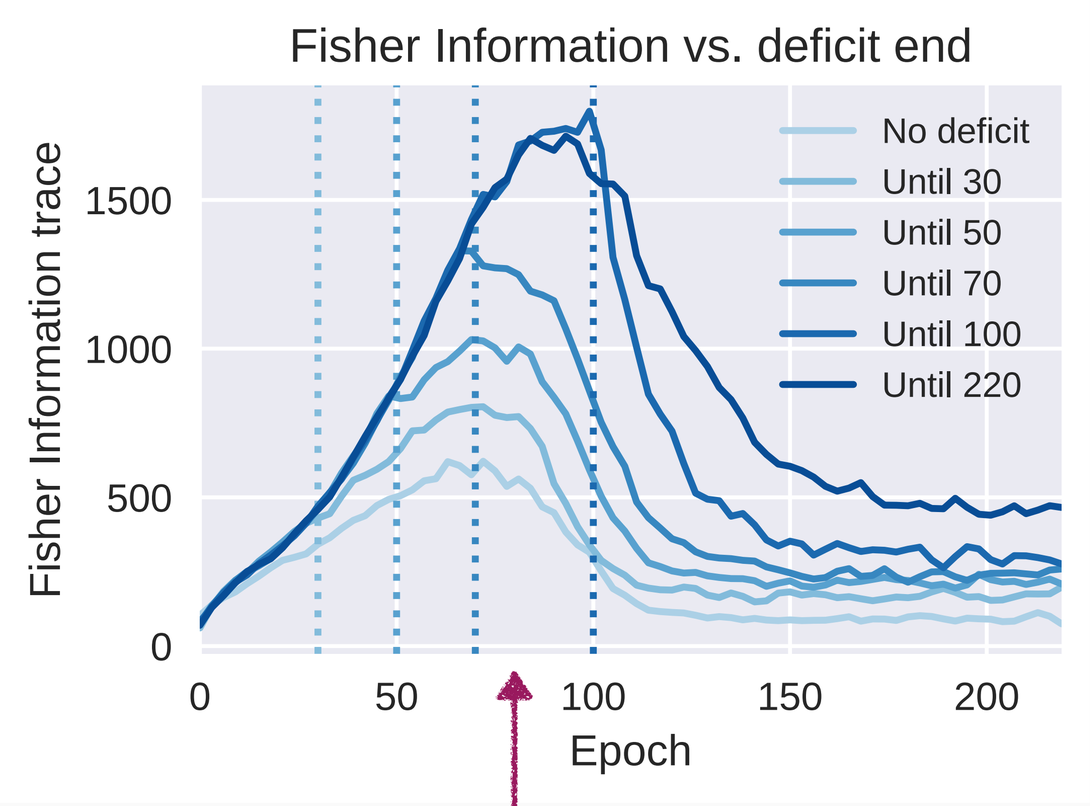
Critical learning periods and Information in Weights

Sensitivity to deficits peaks when network is absorbing information.
Is minimal when the network is consolidating information.



Achille, Rovere, Soatto, *Critical Learning Periods in Deep Networks, 2018*

Are flat minima an epiphenomenon?



Final sharpness correlates with generalization…

…but generalization quality is decided here, far from convergence to minima