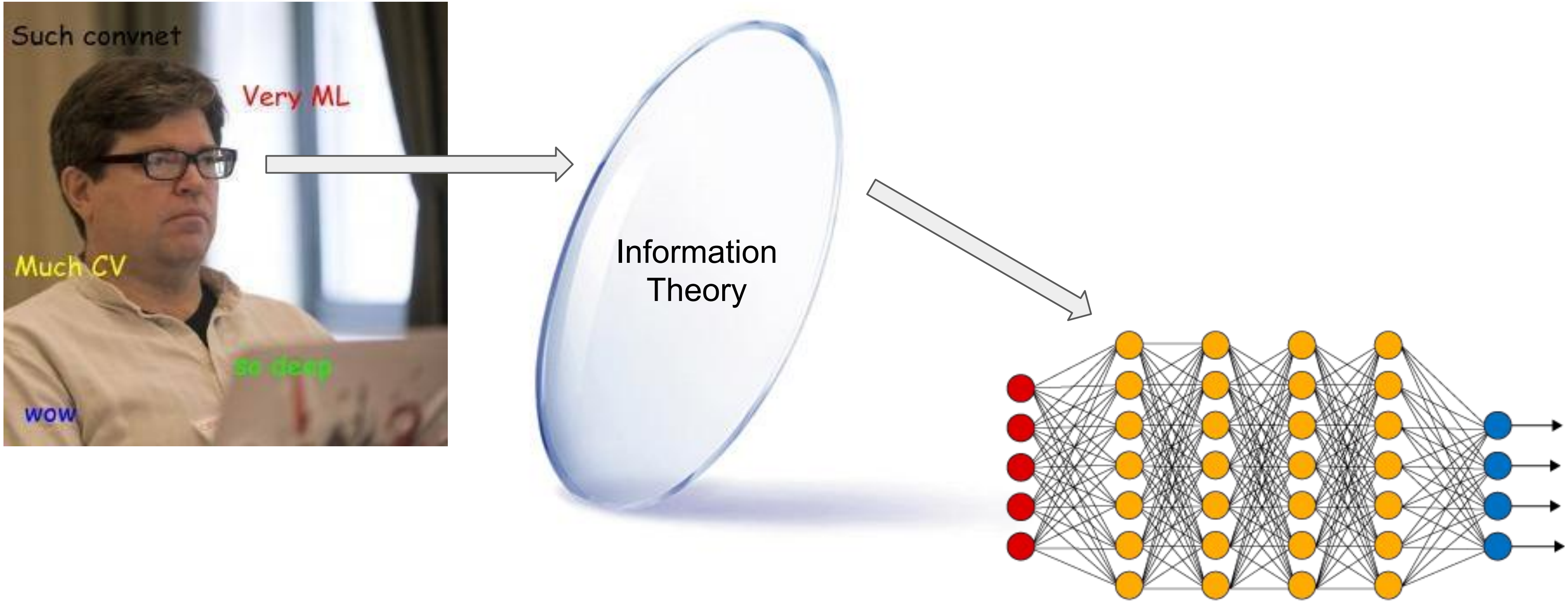
Emergence of Invariance and Disentangling in Deep Representations

A.Achille and S.Soatto
arXiv:1706.01350

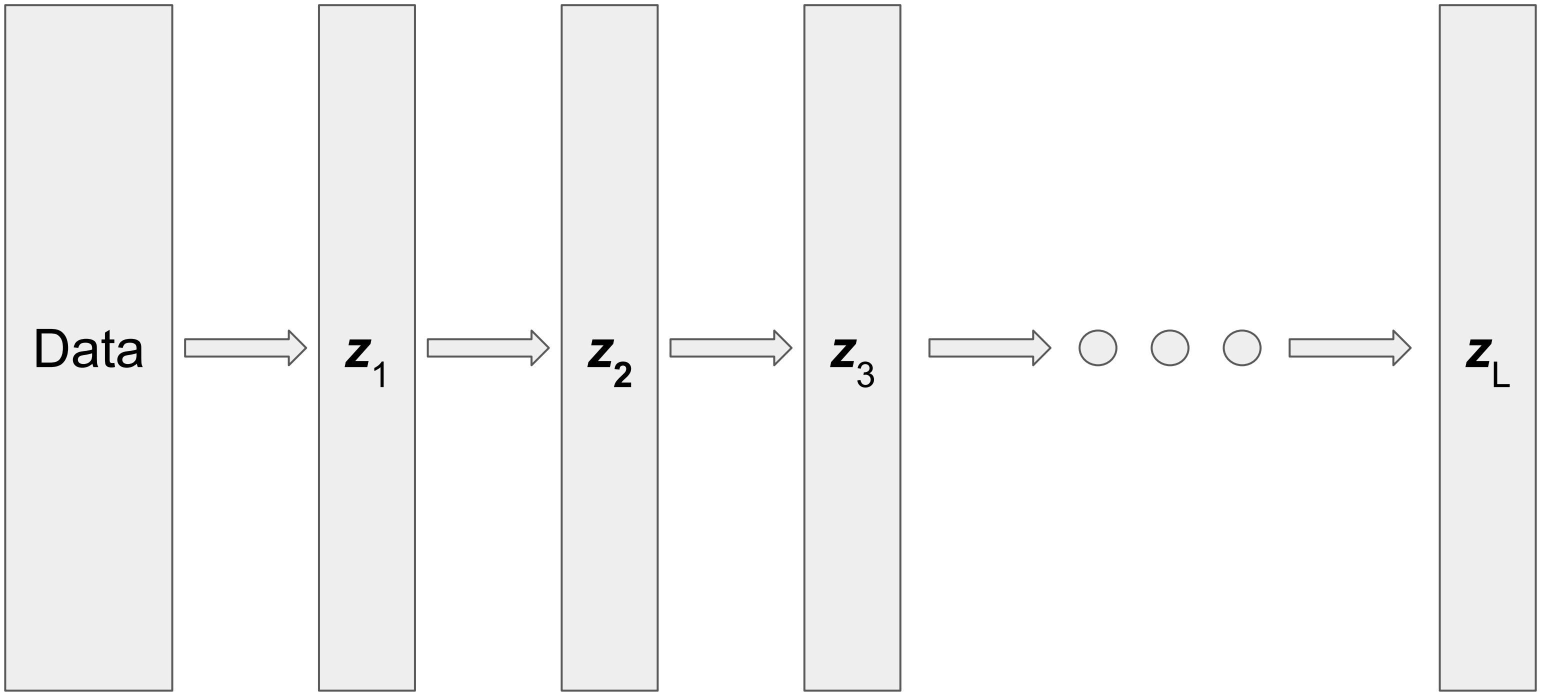
Presented by:
Aristide Baratin, Brady Neal, Nithin Vasisth

Deep learning through the lens of information theory



Representation learning

Representation: some function of the data that is useful for a given task



What makes a **good** representation?

- a function of future data
- constructed from past data
- that is useful for a task
- independent to nuisance factors
- and is easier to use than the data itself

sufficient
invariant
minimal &
disentangled

Information theory

Setting:

- **Task:** predict output y given input data x
- **Representation** $z \sim p(z \mid x)$ is a stochastic function of the data x

Entropy $H(x)$: amount of information in a random variable x

Conditional Entropy $H(y \mid x)$: amount of information in y when x is known

Mutual information $I(x; y)$: amount of information shared by x and y

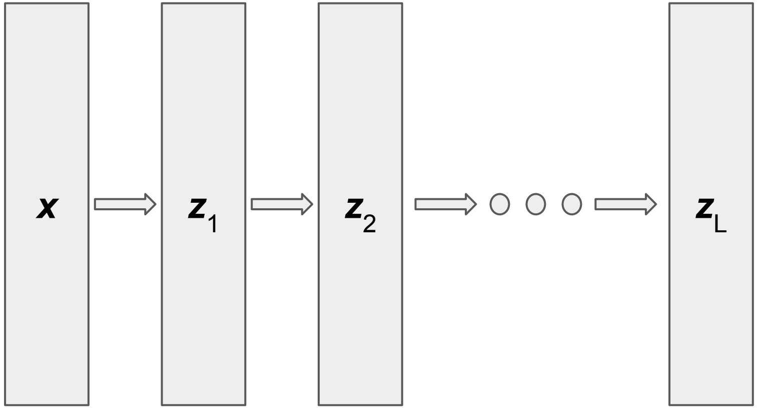
$$I(x; y) = H(y) - H(y \mid x)$$

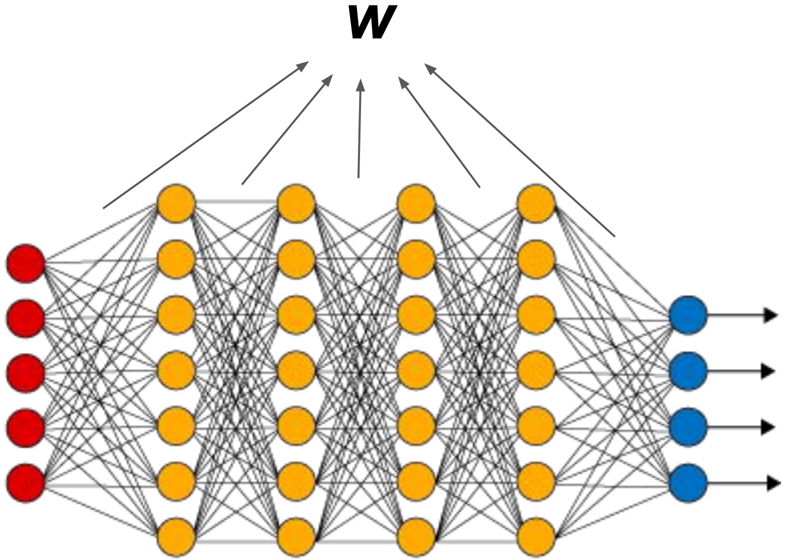
What makes a **good** representation? (formal)

- **Sufficient:** $I(z; y) = I(x; y)$
- **Minimal:** $I(z; x)$ is minimal among sufficient z
- **Invariant** to any *nuisance* n: $I(z; n) = 0$ for all n with $I(n; y) = 0$
- Maximally **disentangled:** minimize $TC(z) = KL(p(z) \,||\, \prod_i p(z_i))$

Representation perspective vs weights perspective

$\boldsymbol{z}_i$ vs $\boldsymbol{w}$





Outline

Introduction

Part 1: Learning minimal **representations**

Result: minimality implies invariance

Part 2: Learning minimal **weights**

Result: information in the weights is good measure of complexity

Part 3: Duality of **representation** and **weights**

Result: minimal weights → invariant & disentangled representation

Outline

Introduction

Part 1: Learning minimal **representations**

Part 2: Learning minimal **weights**

Part 3: Duality of **representation** and **weights**

IB Lagrangian

Recall:

Sufficiency: $y \perp\!\!\!\perp x \mid z$, or equivalently if *I*(z;y) = *I*(x;y)

Minimal: *I*(z;x) is smallest among all the sufficient representations

IB Lagrangian (Tischby et al. 1999):

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

Data Processing Inequality

For a Markov chain,

$$x \to z \to y$$

DPI ensures that:

$$I(x; z) \geq I(x; y)$$

Basically, we keep losing information as we propagate through the layers

Nuisance

Nuisance: Any random variable that affects the data *x*; but is irrelevant to the task *y*

$$y \perp\!\!\!\perp n, \text{ or equivalently } I(y;n) = 0.$$

A representation is *invariant* to a nuisance **n**, if:

$$z \perp\!\!\!\perp n, \text{ or } I(z;n) = 0.$$

A representation is *maximally insensitive* to a nuisance **n**, if:

It minimizes *I*(z;n) among all sufficient representations

Minimality implies Invariance

Proposition:

constant

$$I(n;z) \leq I(z;x) - I(x;y)$$

Invariance

minimality

Consequence: Minimality promotes Invariance

Invariance emerges from elimination of irrelevant information!

Ways to impose invariance

Explicit regularisation : IB Lagrangian

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

Implicit regularization:

- Stacking layers (due to DPI)
- Bottlenecks (Eg: max pooling)
- Noise (Eg: gradient variance, dropout)

Outline

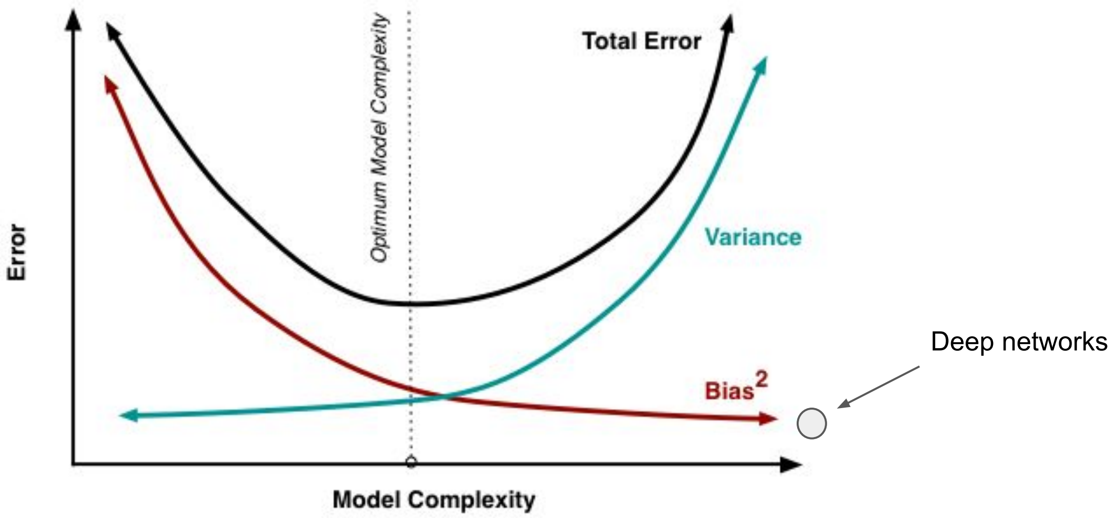
Introduction

Part 1: Learning minimal **representations**

Part 2: Learning minimal **weights**

Part 3: Duality of **representation** and **weights**

Generalization: the puzzle



E.g Zhang et al (2016): deep networks fit random labels (high Rademacher complexity)

One million dollar question: Is there a better notion of complexity for deep networks?

Overfitting: a view from information theory

Bayesian setting:

$$\theta \sim p(\theta), \qquad \mathcal{D} = (\mathbf{x}, \mathbf{y}) \sim p_\theta(x, y)$$

data distribution and dataset

$$q_w(x, y), \qquad w \sim q(w|\mathbf{x}, \mathbf{y})$$

learned distribution

$$p(\mathbf{x}, \mathbf{y}, \theta, \omega) \;=\; p(\theta)\, p(\mathbf{x}, \mathbf{y}|\theta)\, q(w|\mathbf{x}, \mathbf{y})$$

joint distribution

Cross-entropy loss:

$$\mathcal{L}_{p,q} \;=\; \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p(\mathbf{x},\mathbf{y})}\, \mathbb{E}_{w\sim q(w|\mathbf{x},\mathbf{y})}\left[-\log q_w(\mathbf{x}, \mathbf{y})\right]$$

Overfitting: a view from information theory

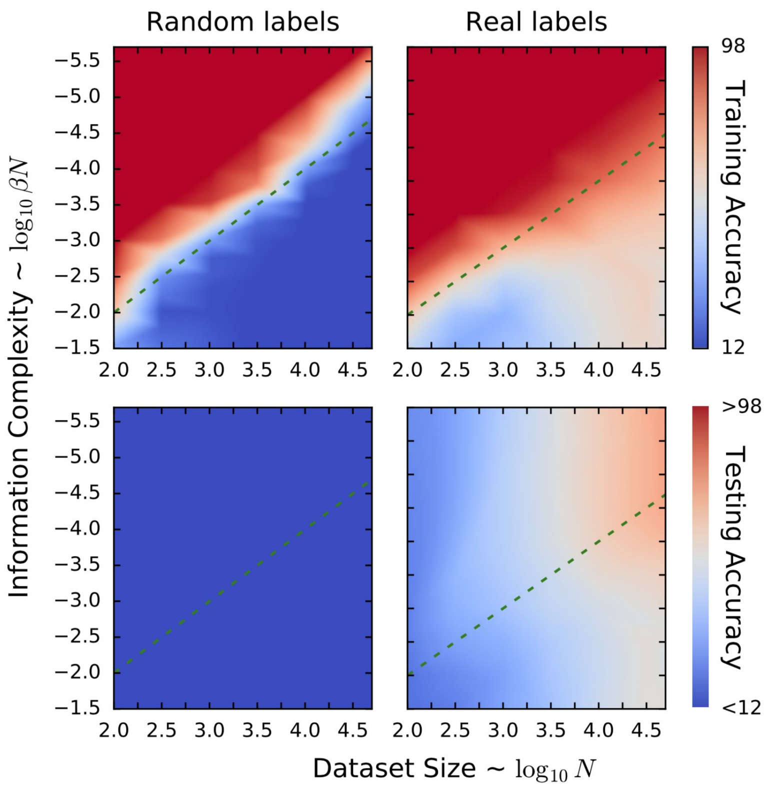
Information decomposition

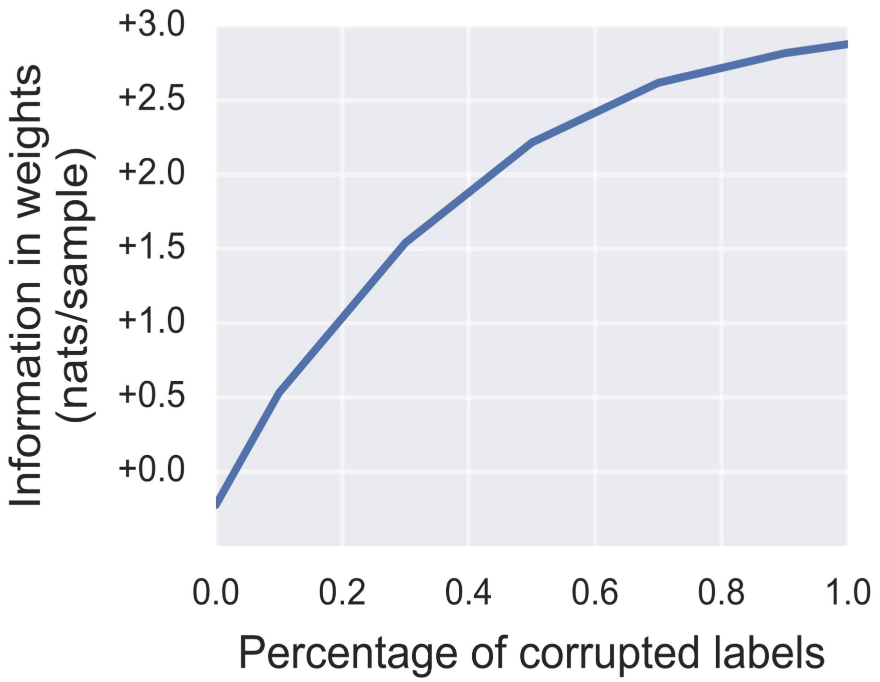
$$\mathcal{L}_{p,q} = \underbrace{H(\mathcal{D} \mid \theta)}_{\text{intrinsic error}} + \underbrace{I(\theta; \mathcal{D} \mid w)}_{\text{sufficiency}} + \underbrace{\mathrm{KL}(q \,||\, p)}_{\text{model efficiency}} - \underbrace{I(\mathcal{D}; w \mid \theta)}_{\text{overfitting}}$$

Suggests regularization: $$\mathcal{L}(q(w \mid \mathcal{D})) = \mathcal{L}_{p,q} + \beta I(w; \mathcal{D})$$

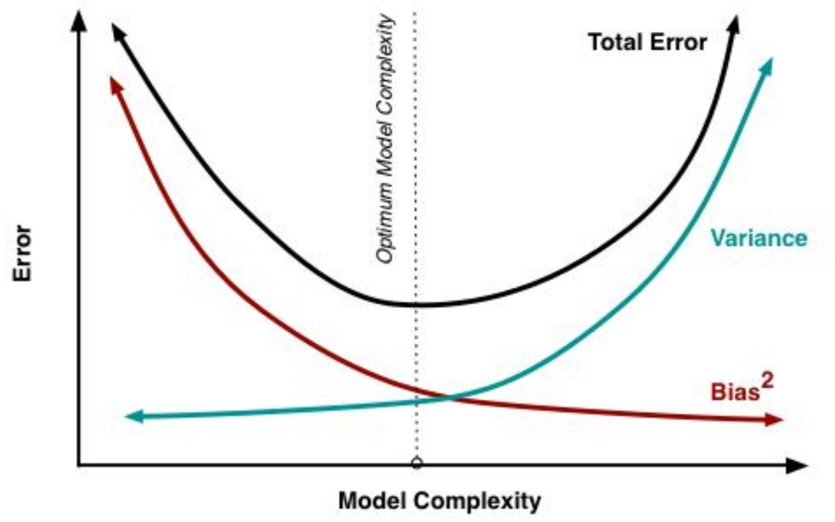
- Minimizing I(w,D) is an old idea — Hinton & Van Camp (1993)
- Reduces to variational lower-bound when $\beta = 1$
- Related to variational dropout — Kingma et al. (2015)

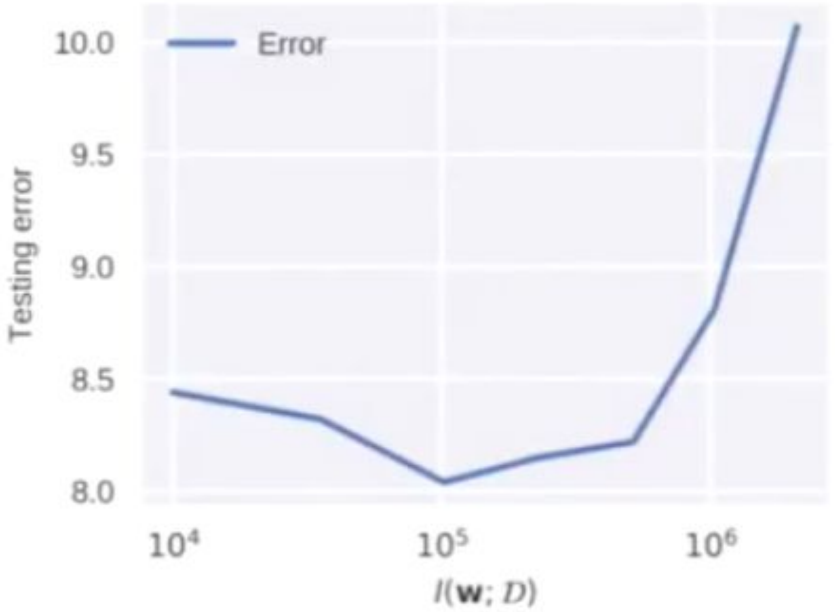
Experiments: random labels





Bias-variance trade off





Information in the weights is a good measure of complexity

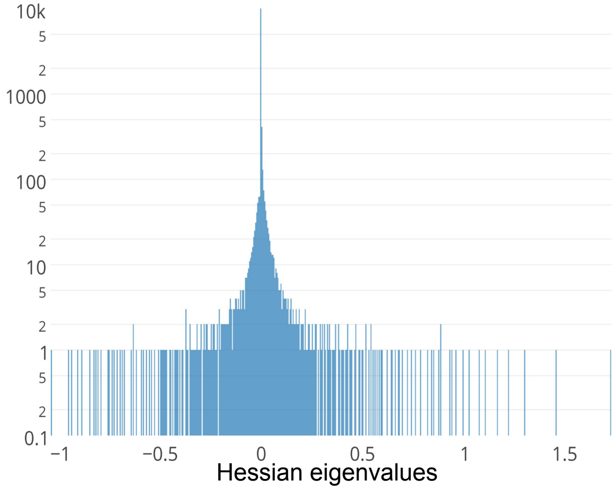
Bonus: SGD finds low information minima

Implicit regularization

SGD finds flat minima…

…and flat minima have low information !

Hochreiter & Schmidhuber (1997)



$$I(w;\mathcal{D}) \leq \frac{1}{2}K[\log\|\hat{w}\|_2^2 + \log\|\mathcal{H}\|_* - K\log(K^2\beta/2)]$$

$$K = \dim(w)$$

Outline

Introduction

Part 1: Learning minimal **representations**

Part 2: Learning minimal **weights**

Part 3: Duality of **representation** and **weights**

Disentanglement

Let’s say we find a representation that is:

- Sufficient
- Minimal
- Invariant (or maximally invariant) to nuisances

Such a representation is not unique… (no bijective mapping)

… And that is good!

Disentanglement

So, we can also try to make the representation *maximally disentangled*;
i.e minimize Total Correlation *TC(z)*;

$$\mathrm{TC}(z) = \mathrm{KL}(\, p(z) \,\|\, \textstyle\prod_i p(z_i) \,),$$

A bound on minimality

$$g(\alpha) \leq \frac{I(x;z) + TC(z)}{\dim(z)} \leq g(\alpha) + c,$$

where $c = O(1/\dim(x)) \leq 1$, $g(\alpha) = \log\left(1 - e^{-\alpha}\right)/2$ *and* α *is related to* $\tilde{I}(w;\mathcal{D})$ *by* $\alpha = \exp\{-I(W;\mathcal{D})/\dim(W)\}$. *In particular,* $I(x;z) + TC(z)$ *is tightly bounded by* $\tilde{I}(W;\mathcal{D})$ *and increases strictly with it.*

what it tells you is this:

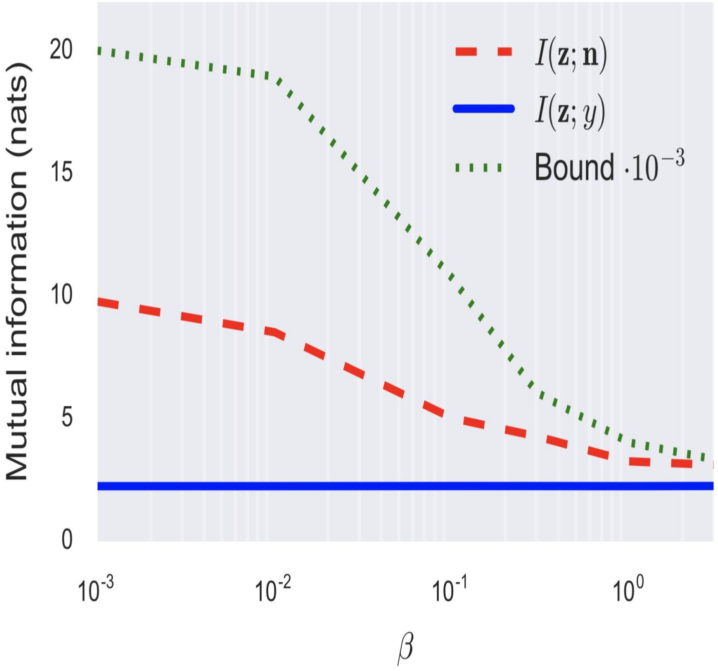
$I(x;z) + TC(z)$ is tightly bounded (on both sides) by an increasing function of $I(W;\mathcal{D})$

Recall:

- *TC(z)* 0; implies disentanglement
- Minimizing *I(x;z)* increases invariance

minimal & disentangled representations ⇔ minimal weights!

Experiment: nuisance invariance



IG Lagrangian (weights perspective):

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

- Sensitivity to nuisance n measured by I(z,n)
- I(z,n) decreases with beta: regularizer promotes invariance!

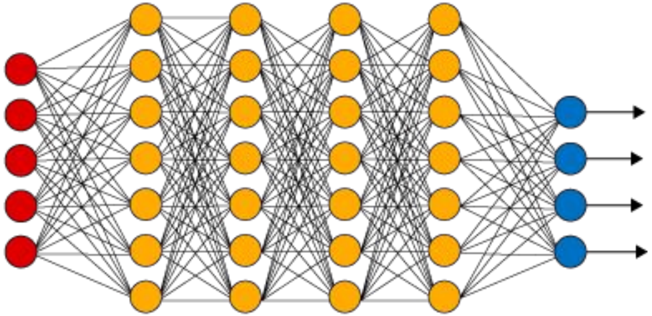
Takeaways

- Minimal (sufficient) representation are invariant
  explicit regularization (IB) or implicit architecture bias (depth) promote invariance

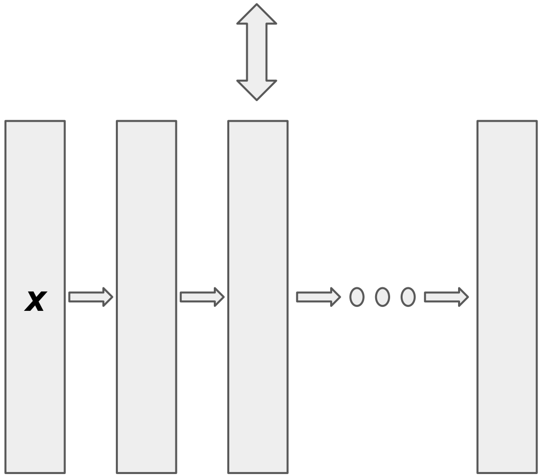
- Information in the weights as a measure of complexity of the network
  low information prevents overfitting

- Information in the weights is closely related to minimality and disentanglement

- SGD finds low information minima



$I(\boldsymbol{w}, \boldsymbol{D})$



Thank you