Deep Learning & Applied AI

Regularization

Emanuele Rodolà
rodola@di.uniroma1.it

SAPIENZA
UNIVERSITÀ DI ROMA

2nd semester a.y. 2019/2020 · April 15, 2020

Regularization

Overfitting often happens with limited training data:

$$\#\ \text{parameters} \gg \#\ \text{training examples}$$

Regularization

Overfitting often happens with limited training data:

$$\#\ \text{parameters} \gg \#\ \text{training examples}$$

Regularization is a general mechanism to reduce overfitting and thus improve generalization.

Regularization

Overfitting often happens with limited training data:

$$\#\text{ parameters} \gg \#\text{ training examples}$$

Regularization is a general mechanism to reduce overfitting and thus improve generalization.

General idea: reduce the number of free parameters.

Regularization

Overfitting often happens with limited training data:

$$\#\ \text{parameters} \gg \#\ \text{training examples}$$

Regularization is a general mechanism to reduce overfitting and thus improve generalization.

General idea: reduce the number of free parameters.

- Eliminate network weights.

  e.g. estimate network sensitivity w.r.t. each weight.

LeCun et al, "Handwritten digit recognition with a backpropagation network", 1990

Regularization

Overfitting often happens with limited training data:

$$\# \text{ parameters} \gg \# \text{ training examples}$$

Regularization is a general mechanism to reduce overfitting and thus improve generalization.

General idea: reduce the number of free parameters.

- Eliminate network weights.

  e.g. estimate network sensitivity w.r.t. each weight.

- Weight sharing (i.e. $\#$ weights $<$ $\#$ connections).

LeCun et al, "Handwritten digit recognition with a backpropagation network", 1990

Regularization

Overfitting often happens with limited training data:

$$\# \text{ parameters} \gg \# \text{ training examples}$$

Regularization is a general mechanism to reduce overfitting and thus improve generalization.

General idea: reduce the number of free parameters.

- Eliminate network weights.

  e.g. estimate network sensitivity w.r.t. each weight.

- Weight sharing (i.e. $\#$ weights $<$ $\#$ connections).

- Explicit penalties.

LeCun et al, "Handwritten digit recognition with a backpropagation network", 1990

Regularization

Overfitting often happens with limited training data:

$$\#\ \text{parameters} \gg \#\ \text{training examples}$$

Regularization is a general mechanism to reduce overfitting and thus improve generalization.

General idea: reduce the number of free parameters.

- Eliminate network weights.

  e.g. estimate network sensitivity w.r.t. each weight.

- Weight sharing (i.e. $\#$ weights $<$ $\#$ connections).

- Explicit penalties.

- Implicit regularization.

LeCun et al, "Handwritten digit recognition with a backpropagation network", 1990

Weight penalties

$$\underbrace{\ell(\boldsymbol{\Theta})}_{\text{loss}} + \lambda \underbrace{\rho(\boldsymbol{\Theta})}_{\text{regularizer}}$$

Weight penalties

$$\underbrace{\ell(\mathbf{\Theta})}_{\text{loss}} + \lambda \underbrace{\rho(\mathbf{\Theta})}_{\text{regularizer}}$$

The regularizer induces a trade-off:

data fidelity vs. model complexity

Even with a high learning rate, weights will not blow up.

Weight penalties

$$\underbrace{\ell(\boldsymbol{\Theta})}_{\text{loss}} + \lambda \underbrace{\rho(\boldsymbol{\Theta})}_{\text{regularizer}}$$

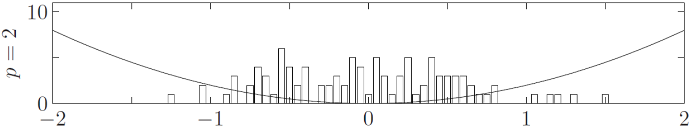
The regularizer induces a trade-off:

data fidelity vs. model complexity

Even with a high learning rate, weights will not blow up.

Typical penalties:

- Tikhonov (L_2) regularization



Weight penalties

$$\underbrace{\ell(\mathbf{\Theta})}_{\text{loss}} + \lambda \underbrace{\rho(\mathbf{\Theta})}_{\text{regularizer}}$$

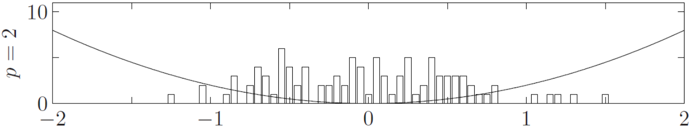
The regularizer induces a trade-off:

data fidelity vs. model complexity

Even with a high learning rate, weights will not blow up.

Typical penalties:

- Tikhonov (L_2) regularization $\Rightarrow$ promotes shrinkage



Weight penalties

$$\underbrace{\ell(\mathbf{\Theta})}_{\text{loss}} + \lambda \underbrace{\rho(\mathbf{\Theta})}_{\text{regularizer}}$$

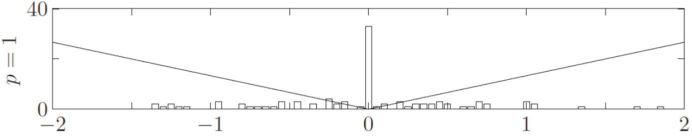
The regularizer induces a trade-off:

data fidelity vs. model complexity

Even with a high learning rate, weights will not blow up.

Typical penalties:

- Tikhonov (L_2) regularization $\Rightarrow$ promotes shrinkage
- Lasso (L_1) regularization



Weight penalties

$$\underbrace{\ell(\boldsymbol{\Theta})}_{\text{loss}} + \lambda \underbrace{\rho(\boldsymbol{\Theta})}_{\text{regularizer}}$$

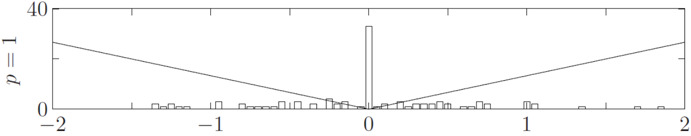
The regularizer induces a trade-off:

data fidelity vs. model complexity

Even with a high learning rate, weights will not blow up.

Typical penalties:

- Tikhonov (L_2) regularization $\Rightarrow$ promotes shrinkage
- Lasso (L_1) regularization $\Rightarrow$ promotes sparsity or weight selection



Weight penalties

$$\underbrace{\ell(\boldsymbol{\Theta})}_{\text{loss}} + \lambda \underbrace{\rho(\boldsymbol{\Theta})}_{\text{regularizer}}$$

The regularizer induces a trade-off:

data fidelity vs. model complexity

Even with a high learning rate, weights will not blow up.

Typical penalties:

- Tikhonov (L_2) regularization $\Rightarrow$ promotes shrinkage
- Lasso (L_1) regularization $\Rightarrow$ promotes sparsity or weight selection
- Bounded L_2 norm at each layer $\|\mathbf{W}^{(\ell)}\|_F \leq c^{(\ell)}$

Weight penalties

$$\underbrace{\ell(\boldsymbol{\Theta})}_{\text{loss}} + \lambda \underbrace{\rho(\boldsymbol{\Theta})}_{\text{regularizer}}$$

The regularizer induces a trade-off:

data fidelity vs. model complexity

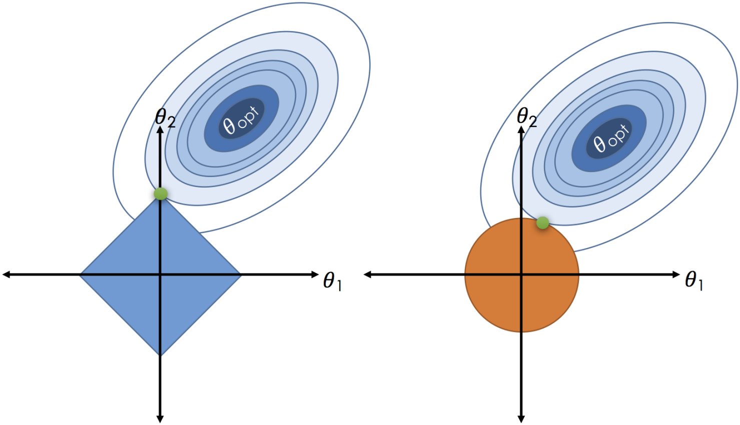
Even with a high learning rate, weights will not blow up.

Typical penalties:

- Tikhonov (L_2) regularization $\Rightarrow$ promotes shrinkage
- Lasso (L_1) regularization $\Rightarrow$ promotes sparsity or weight selection
- Bounded L_2 norm at each layer $\|\mathbf{W}^{(\ell)}\|_F \leq c^{(\ell)}$

After training, the L_p magnitude of each weight reflects its importance.

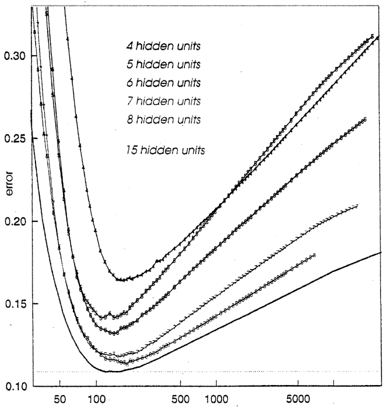
Sparsity



L_1 penalty $\qquad\qquad$ L_2 penalty

Detecting overfitting

Overfitting can be recognized by looking at the validation error:

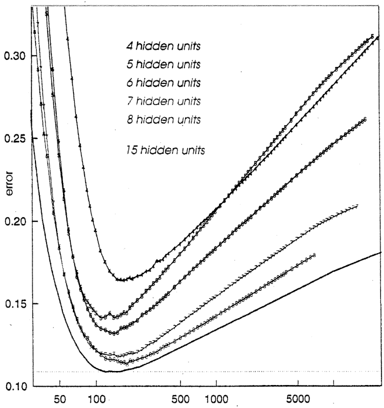


Weigend, "On overfitting and the effective number of hidden units", 1993

Detecting overfitting

Overfitting can be recognized by looking at the validation error:

- Small networks can also overfit.

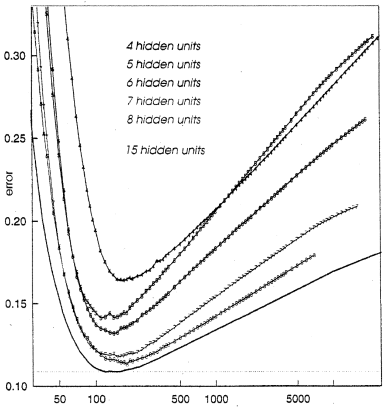


Weigend, "On overfitting and the effective number of hidden units", 1993

Detecting overfitting

Overfitting can be recognized by looking at the validation error:

- Small networks can also overfit.
- Large networks have best performance if they stop early.



Weigend, "On overfitting and the effective number of hidden units", 1993

Early stopping: Overview

Large network $\approx$ superposition of small networks.

In the first epochs of training, large nets tend to learn models similar to those learned by smaller nets.

Weigend, "On overfitting and the effective number of hidden units", 1993

Early stopping: Overview

Large network $\approx$ superposition of small networks.

In the first epochs of training, large nets tend to learn models similar to those learned by smaller nets.

However:

- This ability depends on the optimization algorithm.
- Not always true that more parameters lead to overfitting.

Weigend, "On overfitting and the effective number of hidden units", 1993

Early stopping: Overview

Large network $\approx$ superposition of small networks.

In the first epochs of training, large nets tend to learn models similar to those learned by smaller nets.

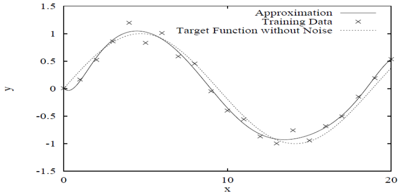
However:

- This ability depends on the optimization algorithm.
- Not always true that more parameters lead to overfitting.

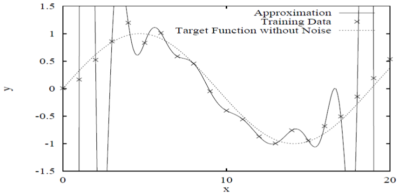
Early stopping: Stop training as soon as performance on a validation set decreases. This is where the network starts to overfit the training set.

Weigend, "On overfitting and the effective number of hidden units", 1993

Many parameters $\neq$ overfitting



Order 10

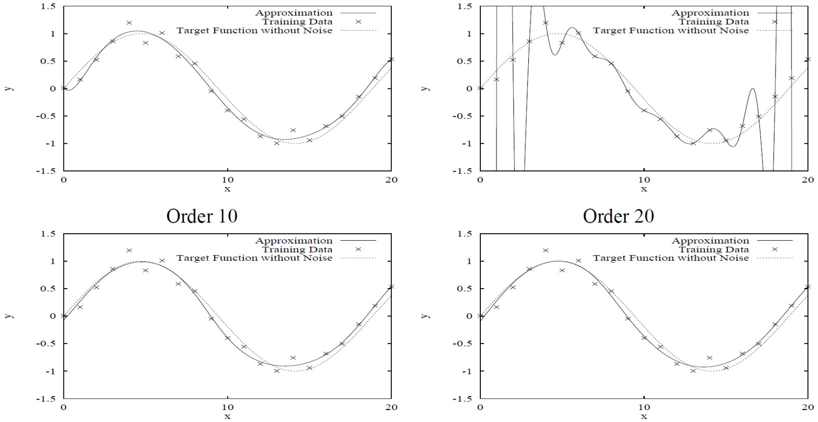


Order 20

Caruana et al, "Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping", NIPS 2001

Many parameters ≠ overfitting

More MLP parameters not leading to overfitting:



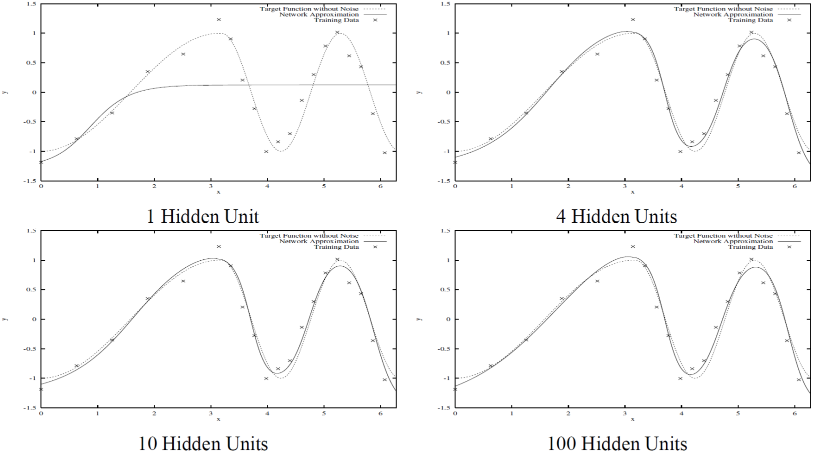
Order 10 | Order 20

10 Hidden Nodes | 50 Hidden Nodes

Caruana et al, "Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping", NIPS 2001

Many parameters $\neq$ overfitting

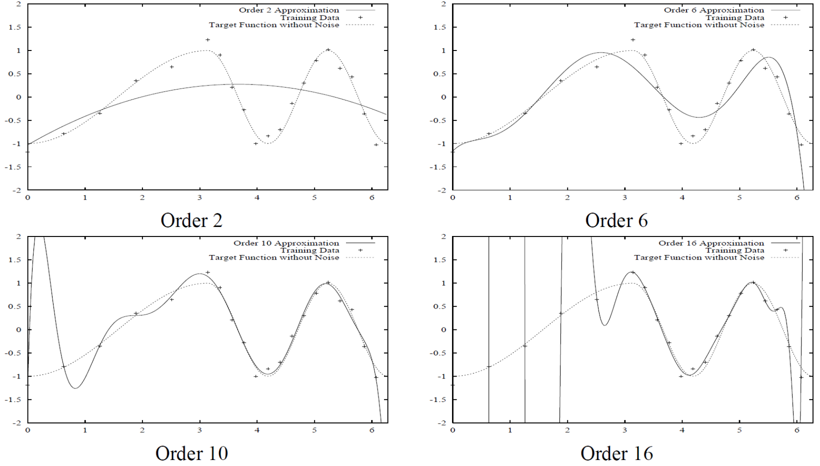
Good fit over all the different data regions:



Caruana et al, "Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping", NIPS 2001

Overfitting as a local phenomenon

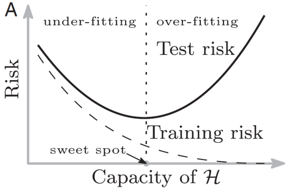
Overfitting is local and can vary significantly in different regions:



Caruana et al, “Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping”, NIPS 2001

Double descent

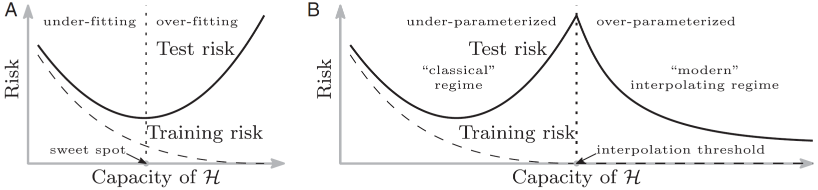
U-shaped curve as a function of # network parameters:



Belkin et al, "Reconciling modern machine-learning practice and the classical bias–variance trade-off", PNAS 2019

Double descent

U-shaped curve as a function of # network parameters:

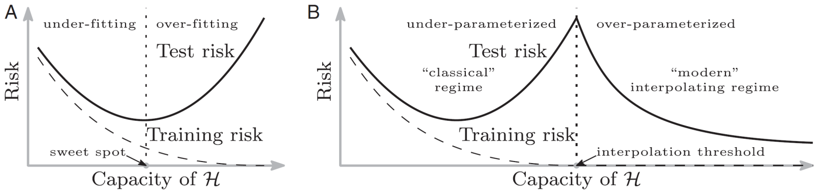


Interpolation: perfect fit on the training data.

Belkin et al, "Reconciling modern machine-learning practice and the classical bias–variance trade-off", PNAS 2019

Double descent

U-shaped curve as a function of # network parameters:



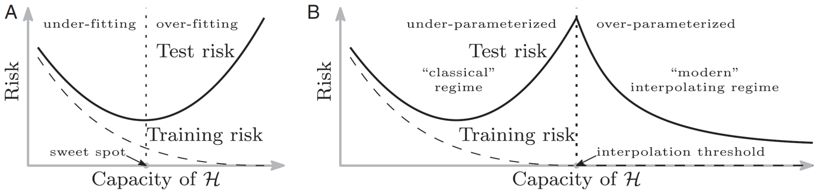
Interpolation: perfect fit on the training data.

"By considering larger function classes, which contain more candidate predictors compatible with the data, we are able to find interpolating functions that have smaller norm and are thus "simpler." Thus, increasing function class capacity improves performance of classifiers."

Belkin et al, "Reconciling modern machine-learning practice and the classical bias–variance trade-off", PNAS 2019

Double descent

U-shaped curve as a function of # network parameters:



Interpolation: perfect fit on the training data.

"By considering larger function classes, which contain more candidate predictors compatible with the data, we are able to find interpolating functions that have smaller norm and are thus "simpler." Thus, increasing function class capacity improves performance of classifiers."

The surprising fact is that SGD is able to find such good models.

Belkin et al, "Reconciling modern machine-learning practice and the classical bias–variance trade-off", PNAS 2019

Early stopping

Our requirements:

- Avoid underfitting regions of high nonlinearity.
- Avoid overfitting regions of low nonlinearity.

Better optimization methods (e.g. conjugate gradient) yield worse generalization because they tend to overfit regions of low non-linearity.

Caruana et al, "Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping", NIPS 2001

Early stopping

Our requirements:

- Avoid underfitting regions of high nonlinearity.
- Avoid overfitting regions of low nonlinearity.

Better optimization methods (e.g. conjugate gradient) yield worse generalization because they tend to overfit regions of low non-linearity.

Early stopping is based on the “smoothness” heuristic:

Representational power grows with weight magnitude

Caruana et al, “Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping”, NIPS 2001

Early stopping

Our requirements:

- Avoid underfitting regions of high nonlinearity.
- Avoid overfitting regions of low nonlinearity.

Better optimization methods (e.g. conjugate gradient) yield worse generalization because they tend to overfit regions of low non-linearity.

Early stopping is based on the "smoothness" heuristic:

Representational power grows with weight magnitude

- Initialize with small weights.

Caruana et al, "Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping", NIPS 2001

Early stopping

Our requirements:

- Avoid underfitting regions of high nonlinearity.
- Avoid overfitting regions of low nonlinearity.

Better optimization methods (e.g. conjugate gradient) yield worse generalization because they tend to overfit regions of low non-linearity.

Early stopping is based on the "smoothness" heuristic:

Representational power grows with weight magnitude

- Initialize with small weights.
- Simple hypotheses are considered before complex hypotheses.

Caruana et al, "Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping", NIPS 2001

Early stopping

Our requirements:

- Avoid underfitting regions of high nonlinearity.
- Avoid overfitting regions of low nonlinearity.

Better optimization methods (e.g. conjugate gradient) yield worse generalization because they tend to overfit regions of low non-linearity.

Early stopping is based on the "smoothness" heuristic:

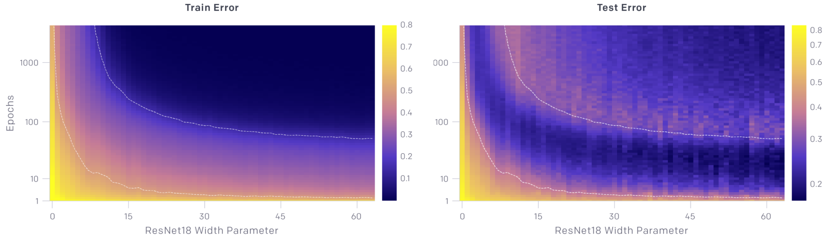
Representational power grows with weight magnitude

- Initialize with small weights.
- Simple hypotheses are considered before complex hypotheses.
- Training first explores models similar to what a smaller net of optimal size would have learned.

Caruana et al, "Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping", NIPS 2001

Epoch-wise double descent

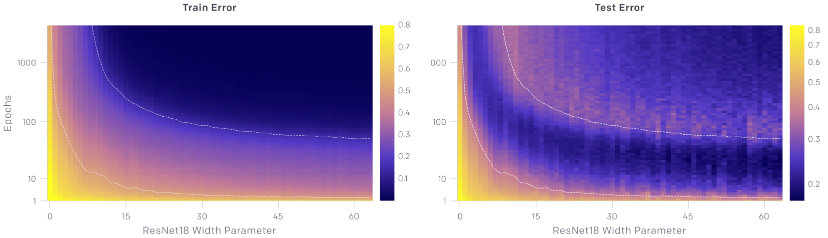
There is a regime where training longer reverses overfitting.



https://openai.com/blog/deep-double-descent/

Epoch-wise double descent

There is a regime where training longer reverses overfitting.

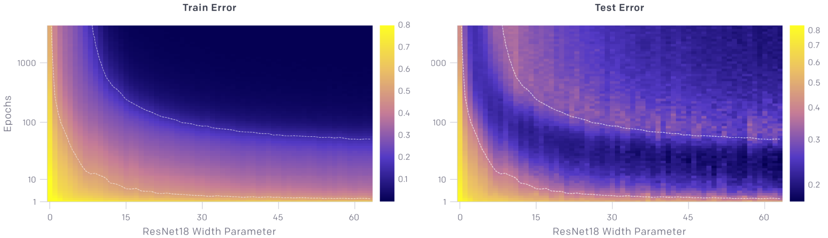


For a fixed number of epochs, the “usual” double descent.

https://openai.com/blog/deep-double-descent/

Epoch-wise double descent

There is a regime where training longer reverses overfitting.



For a fixed number of epochs, the “usual” double descent.

For a fixed number of parameters, we observe double descent as a function of training time.

https://openai.com/blog/deep-double-descent/

Batch normalization

$$\mathbf{x}^{(k)} = \sigma\left(\mathbf{W}^{(k)}\mathbf{x}^{(k-1)}\right)$$

Batch normalization

$$\mathbf{x}^{(k)} = \sigma\left(\mathbf{W}^{(k)}\mathbf{x}^{(k-1)}\right)$$

Internal covariate shift: The input distribution changes at each layer, and the layers need to continuously adapt to the new distribution.

Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function", 2000

Batch normalization

$$\mathbf{x}^{(k)} = \sigma\left(\mathbf{W}^{(k)}\mathbf{x}^{(k-1)}\right)$$

Internal covariate shift: The input distribution changes at each layer, and the layers need to continuously adapt to the new distribution.

Normalize the features by the statistics computed within the training set:

$$\hat{\mathbf{x}} = \mathrm{normalize}(\mathbf{x}, \mathcal{X})$$

where both $\mathbf{x}$ and $\mathcal{X}$ depend on $\mathbf{W}$.

Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function", 2000; Ioffe and Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", ICML 2015

Batch normalization

$$\mathbf{x}^{(k)} = \sigma\left(\mathbf{W}^{(k)}\mathbf{x}^{(k-1)}\right)$$

Internal covariate shift: The input distribution changes at each layer, and the layers need to continuously adapt to the new distribution.

Normalize the features by the statistics computed within the training set:

$$\hat{\mathbf{x}} = \mathrm{normalize}(\mathbf{x}, \mathcal{X})$$

where both $\mathbf{x}$ and $\mathcal{X}$ depend on $\mathbf{W}$.

In particular, backprop will need the partial derivatives:

$$\frac{\partial}{\partial \mathbf{x}}\mathrm{normalize}(\mathbf{x}, \mathcal{X}), \qquad \frac{\partial}{\partial \mathcal{X}}\mathrm{normalize}(\mathbf{x}, \mathcal{X})$$

Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function", 2000; Ioffe and Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", ICML 2015

Batch normalization: Transformation

For each dimension of $\mathbf{x}$, transform:

$$x_i \mapsto \frac{x_i - \mathrm{E}[x_i]}{\sqrt{\mathrm{var}(x_i)}}$$

where mean and variance are computed over the training set.

After the transformation, we get mean $= 0$ and var $= 1$.

Ioffe and Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", ICML 2015

Batch normalization: Transformation

For each dimension of $\mathbf{x}$, transform:

$$x_i \mapsto \frac{x_i - \mathrm{E}[x_i]}{\sqrt{\mathrm{var}(x_i)}}$$

where mean and variance are computed over the training set.

After the transformation, we get mean $= 0$ and var $= 1$.

Furthermore, introduce trainable weights:

$$x_i \mapsto \gamma_i \frac{x_i - \mathrm{E}[x_i]}{\sqrt{\mathrm{var}(x_i)}} + \beta_i$$

These allow to represent the identity $x_i \mapsto x_i$, if that was the optimal thing to do in the original network.

Ioffe and Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", ICML 2015

Batch normalization: Using mini-batches

Avoid analyzing the entire training set at each parameter update.

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1\ldots m}\}$;
Parameters to be learned: γ, β

Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma \widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

Ioffe and Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", ICML 2015

Batch normalization: Using mini-batches

Avoid analyzing the entire training set at each parameter update.

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β

Output: $\{y_i = \text{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_\mathcal{B} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma_\mathcal{B}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_\mathcal{B})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_\mathcal{B}}{\sqrt{\sigma_\mathcal{B}^2 + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \text{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

The batchnorm transformation makes each training example interact with the other examples in each mini-batch.

Ioffe and Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", ICML 2015

Batch normalization: Properties

Typically, batchnorm is applied right before the nonlinearity:

$$\sigma(\mathbf{W}\mathbf{x} + \mathbf{b}) \quad \text{becomes} \quad \sigma \circ \mathrm{BN}_{\gamma,\beta}(\mathbf{W}\mathbf{x})$$

The bias can be removed, since it is ruled out by the mean subtraction.

Ioffe and Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", ICML 2015

Batch normalization: Properties

Typically, batchnorm is applied right before the nonlinearity:

$$\sigma(\mathbf{W}\mathbf{x} + \mathbf{b}) \quad \text{becomes} \quad \sigma \circ \mathrm{BN}_{\gamma,\beta}(\mathbf{W}\mathbf{x})$$

The bias can be removed, since it is ruled out by the mean subtraction.

At test time, mini-batches are not legitimate.

Ioffe and Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", ICML 2015

Batch normalization: Properties

Typically, batchnorm is applied right before the nonlinearity:

$$\sigma(\mathbf{Wx} + \mathbf{b}) \quad \text{becomes} \quad \sigma \circ \mathrm{BN}_{\gamma,\beta}(\mathbf{Wx})$$

The bias can be removed, since it is ruled out by the mean subtraction.

At test time, mini-batches are not legitimate. Mean and variance are those estimated during training, and used for inference.

Ioffe and Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", ICML 2015

Batch normalization: Properties

Typically, batchnorm is applied right before the nonlinearity:

$$\sigma(\mathbf{W}\mathbf{x} + \mathbf{b}) \quad \text{becomes} \quad \sigma \circ \mathrm{BN}_{\gamma,\beta}(\mathbf{W}\mathbf{x})$$

The bias can be removed, since it is ruled out by the mean subtraction.

At test time, mini-batches are not legitimate. Mean and variance are those estimated during training, and used for inference.

Beneficial properties include:

- The stochastic uncertainty of the batch statistics acts as a regularizer that can benefit generalization.

Ioffe and Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", ICML 2015; Santurkar et al, "How does batch normalization help optimization?", NIPS 2018

Batch normalization: Properties

Typically, batchnorm is applied right before the nonlinearity:

$$\sigma(\mathbf{W}\mathbf{x} + \mathbf{b}) \quad \text{becomes} \quad \sigma \circ \mathrm{BN}_{\gamma,\beta}(\mathbf{W}\mathbf{x})$$

The bias can be removed, since it is ruled out by the mean subtraction.

At test time, mini-batches are not legitimate. Mean and variance are those estimated during training, and used for inference.

Beneficial properties include:

- The stochastic uncertainty of the batch statistics acts as a regularizer that can benefit generalization.
- Batchnorm leads to more stable gradients, thus faster training can be achieved with higher learning rates.

Ioffe and Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", ICML 2015; Santurkar et al, "How does batch normalization help optimization?", NIPS 2018

Normalization variants

Normalizing along the batch dimension can lead to inconsistency:

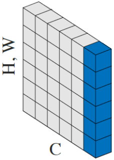
- Bad transfer across different data distributions.
- The stats must be reliable; reducing the mini-batch size increases the model error dramatically.

Normalization variants

Normalizing along the batch dimension can lead to inconsistency:

- Bad transfer across different data distributions.
- The stats must be reliable; reducing the mini-batch size increases the model error dramatically.

Several variants have been proposed:

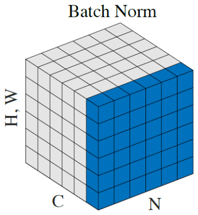


Normalization variants

Normalizing along the batch dimension can lead to inconsistency:

- Bad transfer across different data distributions.
- The stats must be reliable; reducing the mini-batch size increases the model error dramatically.

Several variants have been proposed:

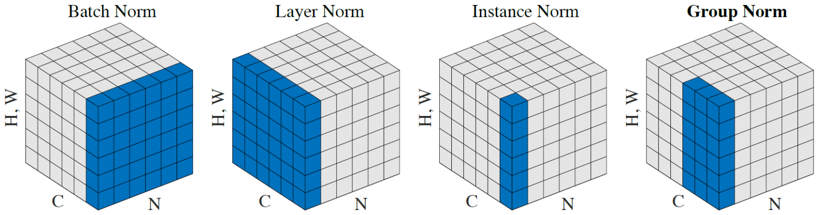


Normalization variants

Normalizing along the batch dimension can lead to inconsistency:

- Bad transfer across different data distributions.
- The stats must be reliable; reducing the mini-batch size increases the model error dramatically.

Several variants have been proposed:



Wu and He, "Group normalization", ECCV 2018

Ensemble deep learning?

Assume you have unlimited computational power.

Train an ensemble of deep nets and average their predictions.

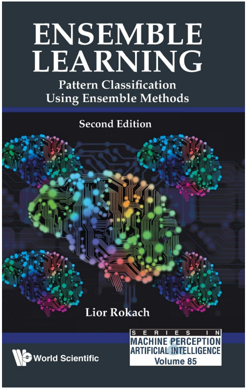
Ensemble deep learning?

Assume you have unlimited computational power.

Train an ensemble of deep nets and average their predictions.

Ensemble predictions (e.g. bayesian networks, random forests) are known to generalize better than the individual models.

Most successful methods in **Kaggle** are ensemble methods.



Neal, "Bayesian Learning for Neural Networks", 1996

Ensemble deep learning?

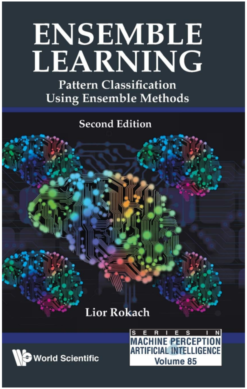
Assume you have unlimited computational power.

Train an ensemble of deep nets and average their predictions.

Ensemble predictions (e.g. bayesian networks, random forests) are known to generalize better than the individual models.

Most successful methods in **Kaggle** are ensemble methods.

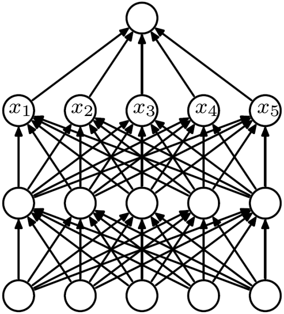
However, for deep nets this would come at a high computational cost.



Neal, "Bayesian Learning for Neural Networks", 1996

Dropout

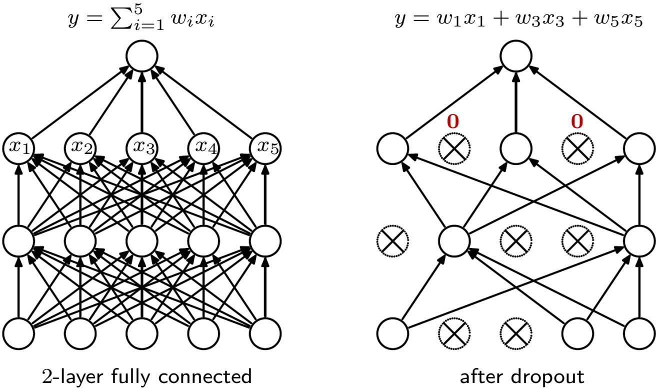
$y = \sum_{i=1}^{5} w_i x_i$



2-layer fully connected

Dropout

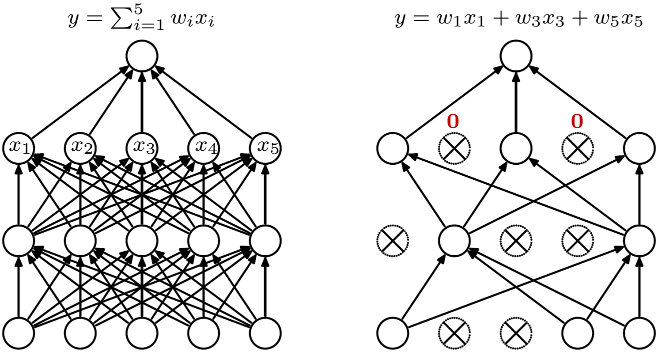
Main idea: Parametrize each model in the ensemble by dropping random units (i.e. nodes with their input/output connections):



Srivastava et al, "Dropout: A Simple Way to Prevent Neural Networks from Over tting", JMLR 2014

Dropout

Main idea: Parametrize each model in the ensemble by dropping random units (i.e. nodes with their input/output connections):



2-layer fully connected

after dropout

Crucially, all networks share the same parameters.

Srivastava et al, "Dropout: A Simple Way to Prevent Neural Networks from Over tting", JMLR 2014

Dropout

This be seen as sampling a network w.r.t to a probability distribution.

n nodes $\Rightarrow$ 2^n possible ways to sample them

Srivastava et al, "Dropout: A Simple Way to Prevent Neural Networks from Over tting", JMLR 2014

Dropout

This be seen as sampling a network w.r.t to a probability distribution.

$$n \text{ nodes} \Rightarrow 2^n \text{ possible ways to sample them}$$

This is way too costly.

- Training: All the networks must be trained.
- Test: All the predictions must be averaged.

Srivastava et al, "Dropout: A Simple Way to Prevent Neural Networks from Over tting", JMLR 2014

Dropout

This be seen as sampling a network w.r.t to a probability distribution.

$$n \text{ nodes} \Rightarrow 2^n \text{ possible ways to sample them}$$

Make it feasible by keeping one single network:

- Training: Generate a new sampling each time new training data is presented (e.g. at each mini-batch in SGD).

Srivastava et al, "Dropout: A Simple Way to Prevent Neural Networks from Over tting", JMLR 2014

Dropout

This be seen as sampling a network w.r.t to a probability distribution.

$$n \text{ nodes} \Rightarrow 2^n \text{ possible ways to sample them}$$

Make it feasible by keeping one single network:

- Training: Generate a new sampling each time new training data is presented (e.g. at each mini-batch in SGD).

  The ensemble is trained to convergence (e.g. with early stopping).

  The individual models are not trained to convergence.

Srivastava et al, "Dropout: A Simple Way to Prevent Neural Networks from Over tting", JMLR 2014

Dropout

This be seen as sampling a network w.r.t to a probability distribution.

$$n \text{ nodes} \Rightarrow 2^n \text{ possible ways to sample them}$$

Make it feasible by keeping one single network:

- Test: The trained weights from each model in the ensemble must be averaged somehow.

Srivastava et al, "Dropout: A Simple Way to Prevent Neural Networks from Over tting", JMLR 2014

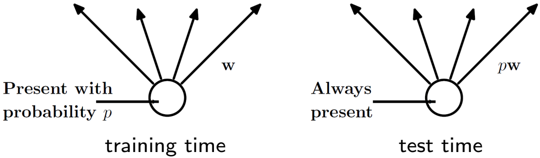
Dropout

This be seen as sampling a network w.r.t to a probability distribution.

$$n \text{ nodes} \Rightarrow 2^n \text{ possible ways to sample them}$$

Make it feasible by keeping one single network:

- Test: The trained weights from each model in the ensemble must be averaged somehow.



If a unit is retained with probability p during training (chosen by hand, even per layer), its outgoing weights are multiplied by p.

Srivastava et al, "Dropout: A Simple Way to Prevent Neural Networks from Over tting", JMLR 2014

Dropout as an ensemble method

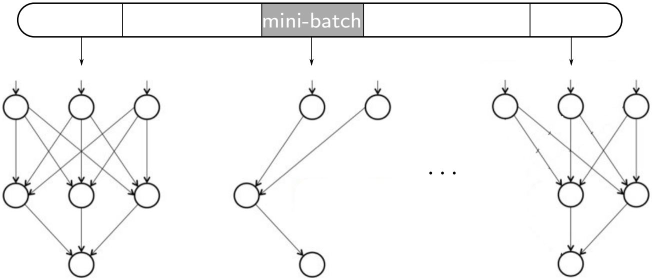
Dropout has two key features:

- It does bagging, i.e. each model is trained on random data.
- It does weight sharing, which is atypical in ensemble methods.

Dropout as an ensemble method

Dropout has two key features:

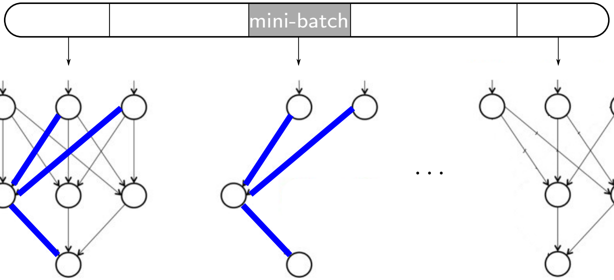
- It does bagging, i.e. each model is trained on random data.
- It does weight sharing, which is atypical in ensemble methods.



Dropout as an ensemble method

Dropout has two key features:

- It does bagging, i.e. each model is trained on random data.
- It does weight sharing, which is atypical in ensemble methods.



At each training step, the weight update is applied to all members of the ensemble simultaneously.

Dropout: Properties

In a standard neural network, weights are optimized jointly.

Co-adaptation: Small errors in a unit are absorbed by another unit.

Srivastava et al, "Dropout: A Simple Way to Prevent Neural Networks from Over
tting", JMLR 2014

Dropout: Properties

In a standard neural network, weights are optimized jointly.

Co-adaptation: Small errors in a unit are absorbed by another unit.

Some properties of dropout as a regularizer:

- Reduces co-adaptation by making units unreliable. This improves generalization to unseen data, and reduces overfitting.

Srivastava et al, "Dropout: A Simple Way to Prevent Neural Networks from Over tting", JMLR 2014

Dropout: Properties

In a standard neural network, weights are optimized jointly.

Co-adaptation: Small errors in a unit are absorbed by another unit.

Some properties of dropout as a regularizer:

- Reduces co-adaptation by making units unreliable. This improves generalization to unseen data, and reduces overfitting.
- Side-effect: sparse representations are learned.

Srivastava et al, "Dropout: A Simple Way to Prevent Neural Networks from Over tting", JMLR 2014

Dropout: Properties

In a standard neural network, weights are optimized jointly.

Co-adaptation: Small errors in a unit are absorbed by another unit.

Some properties of dropout as a regularizer:

- Reduces co-adaptation by making units unreliable. This improves generalization to unseen data, and reduces overfitting.
- Side-effect: sparse representations are learned.
- Performs closely to exact model averaging over all 2^n models.

Srivastava et al, "Dropout: A Simple Way to Prevent Neural Networks from Over tting", JMLR 2014; Warde-Farley et al, "An empirical analysis of dropout in piecewise linear networks", 2014

Dropout: Properties

In a standard neural network, weights are optimized jointly.

Co-adaptation: Small errors in a unit are absorbed by another unit.

Some properties of dropout as a regularizer:

- Reduces co-adaptation by making units unreliable. This improves generalization to unseen data, and reduces overfitting.
- Side-effect: sparse representations are learned.
- Performs closely to exact model averaging over all 2^n models.
- ...and much better if no weight sharing is done in the exact model.

Srivastava et al, "Dropout: A Simple Way to Prevent Neural Networks from Over tting", JMLR 2014; Warde-Farley et al, "An empirical analysis of dropout in piecewise linear networks", 2014

Dropout: Properties

In a standard neural network, weights are optimized jointly.

Co-adaptation: Small errors in a unit are absorbed by another unit.

Some properties of dropout as a regularizer:

- Reduces co-adaptation by making units unreliable. This improves generalization to unseen data, and reduces overfitting.
- Side-effect: sparse representations are learned.
- Performs closely to exact model averaging over all 2^n models.
- ...and much better if no weight sharing is done in the exact model.
- Longer training times, since parameter updates are now noisier.

Srivastava et al, "Dropout: A Simple Way to Prevent Neural Networks from Over tting", JMLR 2014; Warde-Farley et al, "An empirical analysis of dropout in piecewise linear networks", 2014

Dropout: Properties

In a standard neural network, weights are optimized jointly.

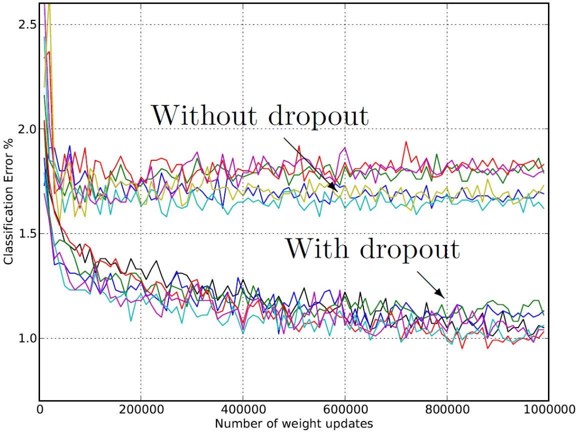
Co-adaptation: Small errors in a unit are absorbed by another unit.

Some properties of dropout as a regularizer:

- Reduces co-adaptation by making units unreliable. This improves generalization to unseen data, and reduces overfitting.
- Side-effect: sparse representations are learned.
- Performs closely to exact model averaging over all 2^n models.
- ...and much better if no weight sharing is done in the exact model.
- Longer training times, since parameter updates are now noisier.
- Typical choices: 20% of the input units and 50% of the hidden units.

Srivastava et al, "Dropout: A Simple Way to Prevent Neural Networks from Over tting", JMLR 2014; Warde-Farley et al, "An empirical analysis of dropout in piecewise linear networks", 2014

Dropout: Properties



Srivastava et al, "Dropout: A Simple Way to Prevent Neural Networks from Over tting", JMLR 2014

Dropout: Implicit bias

In a 2-layer network (input $\mathbf{U}$ + hidden $\mathbf{V}$) dropout is equivalent to:

$$\ell(\mathbf{U}, \mathbf{V}) + \frac{1-p}{p} \sum_{i=1}^{r} \|\mathbf{u}_i\|^2 \|\mathbf{v}_i\|^2$$

where the vectors are the columns of $\mathbf{U}$ and $\mathbf{V}$.

Mianjy et al, "On the implicit bias of dropout", 2018; Neyshabur et al, "Norm-based capacity control in neural networks", 2015

Dropout: Implicit bias

In a 2-layer network (input $\mathbf{U}$ + hidden $\mathbf{V}$) dropout is equivalent to:

$$\ell(\mathbf{U}, \mathbf{V}) + \frac{1-p}{p} \sum_{i=1}^{r} \|\mathbf{u}_i\|^2 \|\mathbf{v}_i\|^2$$

where the vectors are the columns of $\mathbf{U}$ and $\mathbf{V}$.

This is a form of path regularizer: Sum of products of squared weights along each possible path from input to output.

It equalizes the norms $\|\mathbf{u}_i\|^2, \|\mathbf{v}_i\|^2$ for all i, which is conjectured to reduce co-adaptation.

Mianjy et al, "On the implicit bias of dropout", 2018; Neyshabur et al, "Norm-based capacity control in neural networks", 2015

Dropout: Implicit bias

In a 2-layer network (input $\mathbf{U}$ + hidden $\mathbf{V}$) dropout is equivalent to:

$$\ell(\mathbf{U}, \mathbf{V}) + \frac{1-p}{p} \sum_{i=1}^{r} \|\mathbf{u}_i\|^2 \|\mathbf{v}_i\|^2$$

where the vectors are the columns of $\mathbf{U}$ and $\mathbf{V}$.

This is a form of path regularizer: Sum of products of squared weights along each possible path from input to output.

It equalizes the norms $\|\mathbf{u}_i\|^2, \|\mathbf{v}_i\|^2$ for all i, which is conjectured to reduce co-adaptation.

For small enough dropout rate, all minima are global.

Mianjy et al, "On the implicit bias of dropout", 2018; Neyshabur et al, "Norm-based capacity control in neural networks", 2015

Suggested reading

- All the references given throughout the slides.
- Interesting thread on the history of double descent:
  `https://twitter.com/hippopedoid/status/1243229021921579010`
- *Section 4.2.1* is a practical guide for batchnorm by the original authors:
  `https://arxiv.org/pdf/1502.03167`
- *Appendix A* is a practical guide for dropout by the original authors:
  `http://jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf`
- Talk by René Vidal on the implicit bias of dropout:
  `https://ipam.wistia.com/medias/30h1ay475i`