

Generalization

Understanding Machine Learning

October 2022

Felix Wick

Generalization

core of ML:

empirical risk minimization (training error) as proxy for minimizing unknown population risk (test error, aka generalization error or out-of-sample error)

generalization gap: difference between test and training error

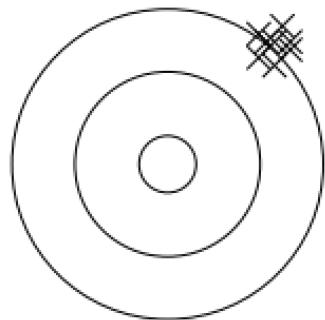
- **interpolation**: to unencountered samples from training environment
- **extrapolation**: to testing conditions differing from training environment

curse of dimensionality: “*learning in high dimensions always amounts to extrapolation*”

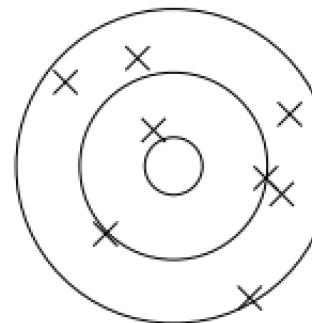
Bias-Variance Tradeoff

Bias, Variance, Irreducible Error

think of fitting ML algorithms as repeatable processes with different (i.i.d.) data sets



bias:
due to too simplistic model
(same for all training data sets)
“underfitting”



variance:
due to sensitivity to specifics (noise)
of different training data sets
“overfitting”

irreducible error (aka Bayes error):

inherent randomness (target generated from random variable following probability distribution)

→ limiting accuracy of ideal model

different potential reasons for inherent randomness (noise): complexity, missing information, ...

Mean Squared Error

fit function $\hat{f}(\mathbf{x})$ to training data set D to approximate assumed true $f(\mathbf{x})$

$$y_i = f(x_{i1}, x_{i2}, \dots, x_{ip}) + \varepsilon_i$$

- with $\mathbf{x} = (x_1, x_2, \dots, x_p)$
- $D = \{(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)\}$
- noise ε : Gaussian with zero mean and variance σ^2

measure mean squared error MSE

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(\mathbf{x}_i; D))^2$$

for training data set D (in-sample)

$$\text{MSE} = \frac{1}{q} \sum_{i=n+1}^{n+q} (y_i - \hat{f}(\mathbf{x}_i; D))^2$$

and test data set T (out-of-sample)

$$T = \{(y_{n+1}, \mathbf{x}_{n+1}), \dots, (y_{n+q}, \mathbf{x}_{n+q})\}$$

Bias-Variance Decomposition

expected squared error on individual sample in test data set T (average over T for MSE):

$$\begin{aligned} & E_{D,\varepsilon} \left[(y_i - \hat{f}(\mathbf{x}_i; D))^2 \right] \\ &= \underbrace{\left(E_D[\hat{f}(\mathbf{x}_i; D)] - f(\mathbf{x}_i) \right)^2}_{\text{bias}} + \underbrace{E_D \left[\left(E_D[\hat{f}(\mathbf{x}_i; D)] - \hat{f}(\mathbf{x}_i; D) \right)^2 \right]}_{\text{variance}} + \sigma^2 \end{aligned}$$

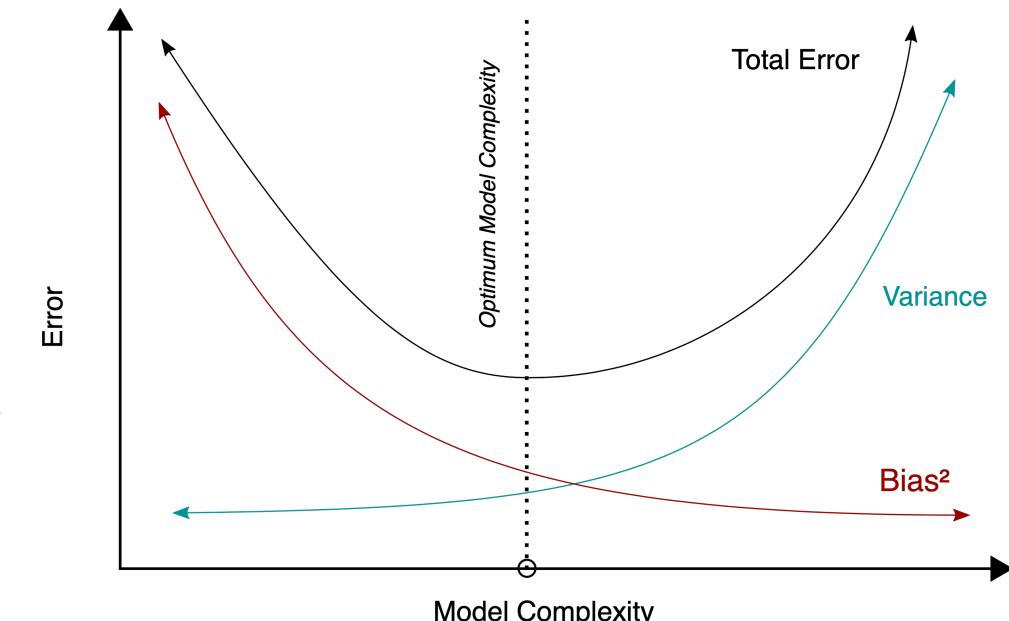
irreducible error

meaning of E_D :

“averaging” predictions from models trained on different choices of D
(independently sampled from same $p(y, \mathbf{x})$)

Bias-Variance Tradeoff

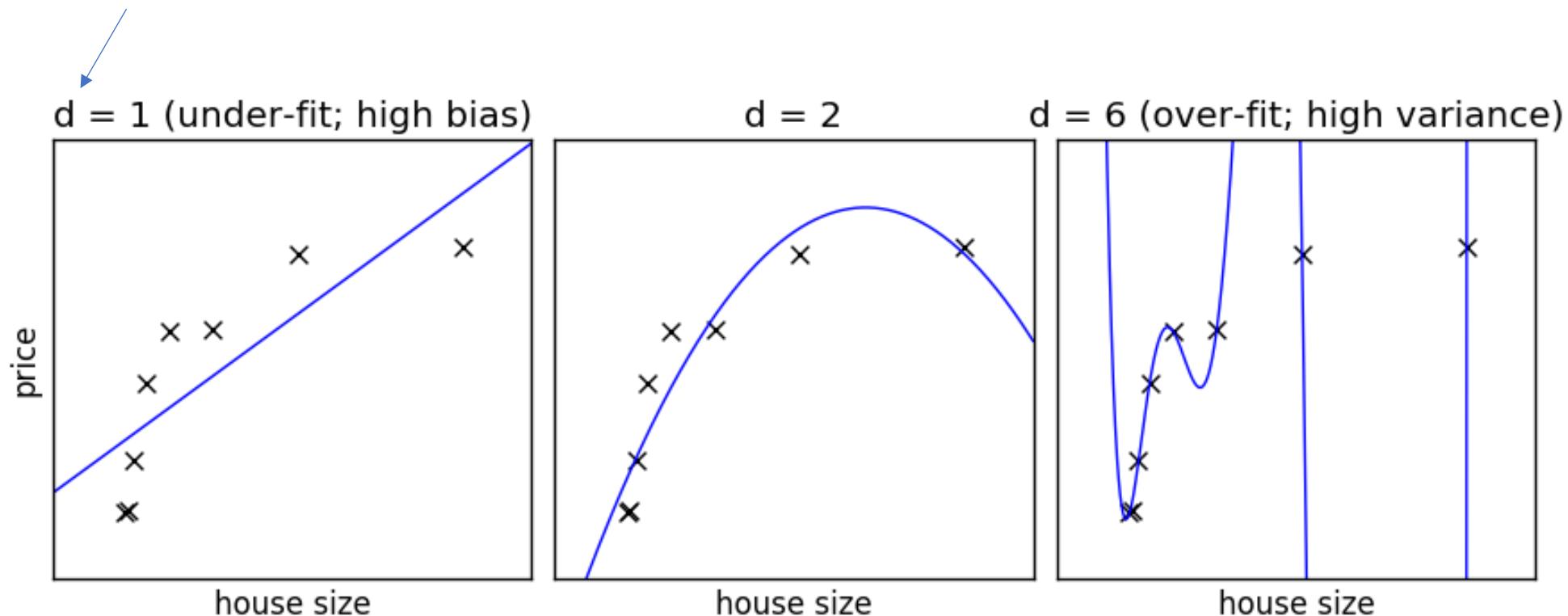
- fundamental concept in classical statistical learning theory
- models of higher complexity have lower bias but higher variance (given the same number of training examples)
- generalization error follows U-shaped curve:
overfitting once model complexity (number of parameters) passes certain threshold
- overfitting: variance term dominating test error
→ increasing model complexity increases test error



from wikipedia

Example: Non-Linear Function Approximation

degree of fitted polynomial

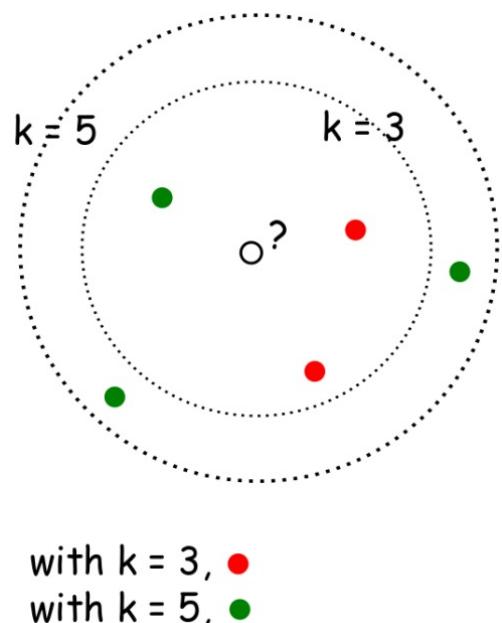


from scikit-learn documentation

Example: k-Nearest Neighbors

- local method, instance-based learning
- non-parametric
- distance defined by metric on x (e.g., Euclidean)

$$\hat{f}(\mathbf{x}_0) = \frac{1}{k} \sum_{j=1}^k y_j \quad \text{with } j \text{ running over } k \text{ nearest neighbors of } \mathbf{x}_0$$



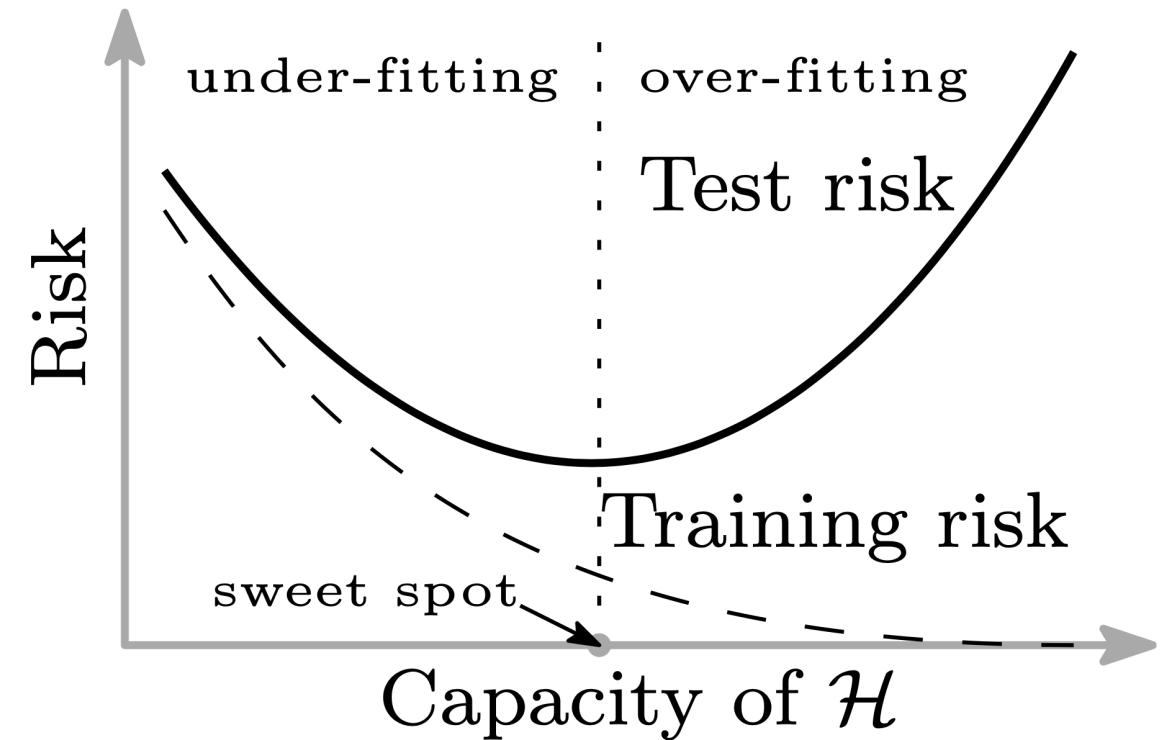
- low k : low bias but high variance
- high k : low variance but high bias

$$bias = f(\mathbf{x}) - \frac{1}{k} \sum_{j=1}^k y_j$$

$$var = \frac{\sigma^2}{k}$$

Problem of Finding Complexity Sweet Spot

- training/ in-sample error keeps decreasing with more complex model (less bias)
- test/out-of-sample error not easily accessible during training
→ e.g., cross-validation, early stopping



from <https://arxiv.org/abs/1812.11118>

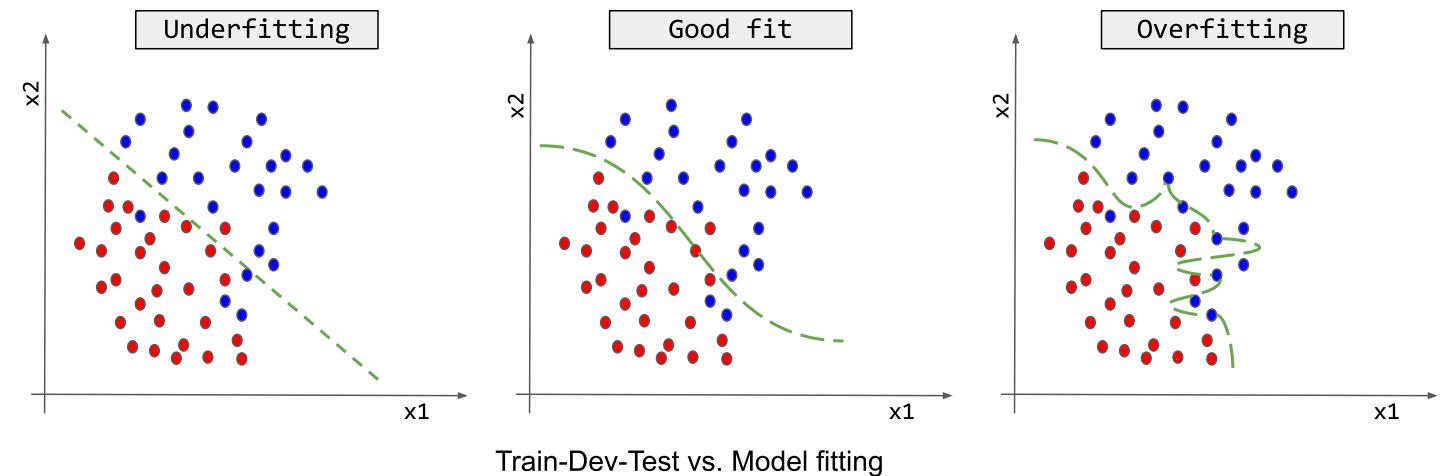
Hyperparameters

model complexity often controlled via hyperparameters (parameters not fitted but set in advance)

hyperparameter tuning

examples:

- degree of fitted polynomial d
- number of considered nearest neighbors k
- maximum depth of decision tree
- number of trees in random forest
- ...



Methods for ...

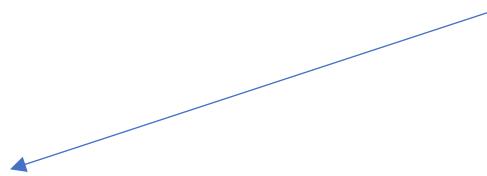
regularization

reducing bias:

- kNN: consider less neighbors
- decision tree: greater depth
- ensemble: boosting
- neural network: more hidden nodes
- more model features
- larger training data set

reducing variance:

- kNN: consider more neighbors
- decision tree: shallower depth
- ensemble: bagging (kind of regularization)
- neural network: less hidden nodes
- dimensionality reduction, feature selection
- larger training data set



Regularization

Supplement to General Recipe of Statistical Learning

adding one more piece to the general recipe of combining models, costs, and optimization methods: regularization

reduce test error, possibly at expense of increased training error

many ways: explicit constraints, adding penalties, priors, parameter sharing, data set augmentation, early stopping, dropout, bagging, ...

Idea of Regularization

3 possible scenarios: trained model family ...

1. excluded the true data-generating process (underfitting, high bias)
2. matched the true data-generating process
3. included the data-generating process, but also many others
(overfitting, high variance)

goal of regularization: take a model from 3 into 2

L^2 Regularization (Ridge Regression)

adding **penalty term** on parameter L^2 norm to cost function

$$\tilde{J}(\hat{\theta}) = J(\hat{\theta}) + \lambda \cdot \hat{\theta}^T \hat{\theta}$$

hyperparameter controlling
strength of regularization

aka weight decay (neural networks)

corresponds to imposing constraint on parameters to lie in L^2 region (size controlled by λ)

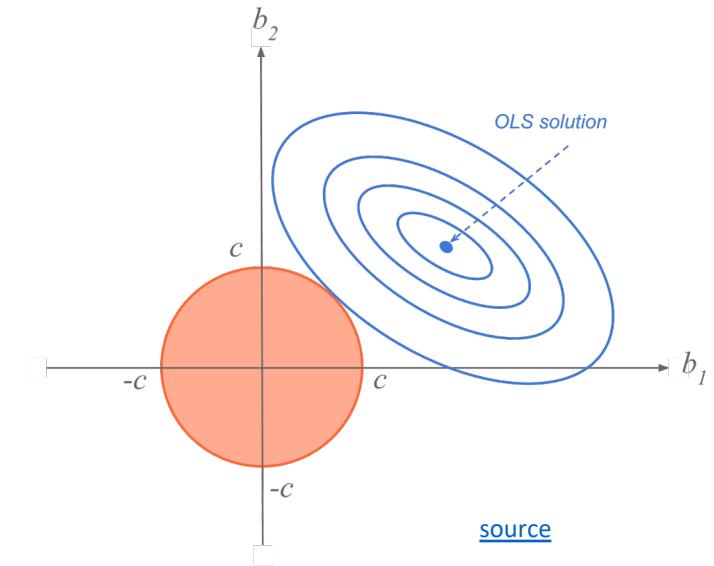
ridge regression for linear regression (MSE costs)

$$\tilde{J}(\hat{\beta}) = (y - X \hat{\beta})^T (y - X \hat{\beta}) + \lambda \cdot \hat{\beta}^T \hat{\beta}$$

analytical solution:

$$\hat{\beta}_{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T y$$

$p \times p$ identity matrix



[source](#)

L^1 Regularization (LASSO)

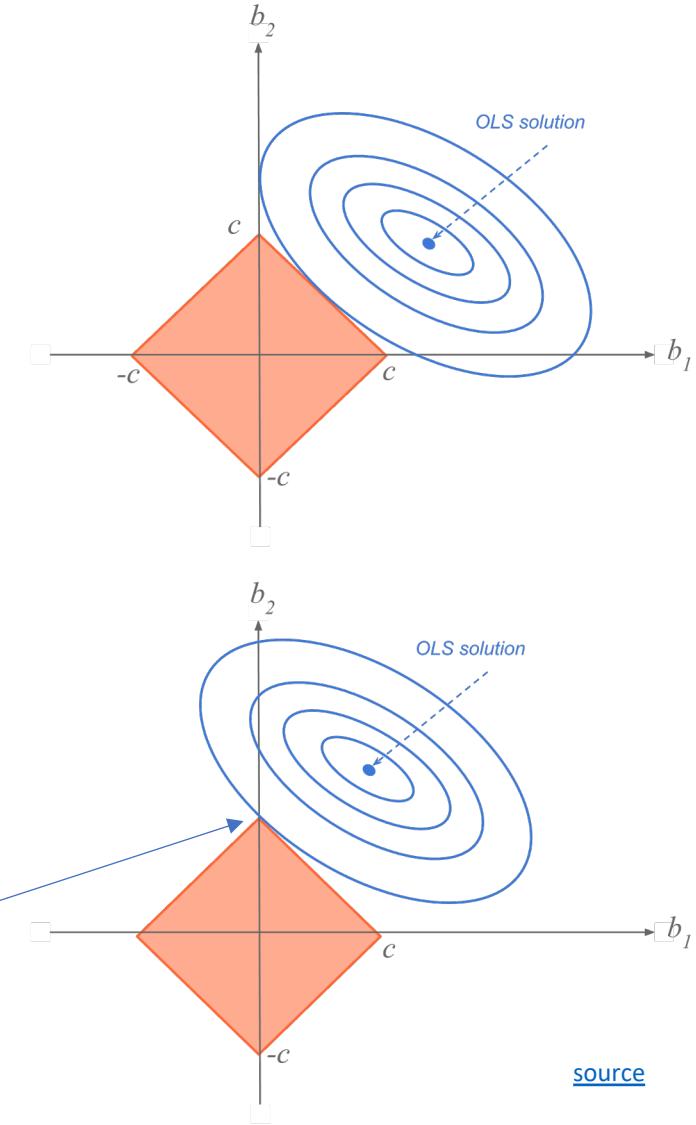
adding **penalty term** on parameter L^1 norm to cost function

$$\tilde{J}(\hat{\theta}) = J(\hat{\theta}) + \lambda \cdot \sum_j |\hat{\theta}_j|$$

corresponds to imposing constraint on parameters to lie in L^1 region
(size controlled by λ)

least absolute shrinkage and selection operator:

also performs feature selection



[source](#)

Maximum a Posteriori Estimation

frequentist maximum likelihood estimation:

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \operatorname{argmax}_{\hat{\boldsymbol{\theta}}} \mathcal{L}(\hat{\boldsymbol{\theta}}; \mathbf{y}|\mathbf{x}) = \operatorname{argmax}_{\hat{\boldsymbol{\theta}}} p(\mathbf{y}|\mathbf{x}; \hat{\boldsymbol{\theta}})$$

maximum a posteriori estimation as alternative applying Bayes theorem:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\text{MAP}} &= \operatorname{argmax}_{\hat{\boldsymbol{\theta}}} p(\hat{\boldsymbol{\theta}}|\mathbf{y}, \mathbf{x}) = \operatorname{argmax}_{\hat{\boldsymbol{\theta}}} \frac{p(\mathbf{y}|\mathbf{x}; \hat{\boldsymbol{\theta}}) \cdot p(\hat{\boldsymbol{\theta}})}{\int p(\mathbf{y}|\mathbf{x}; \boldsymbol{\vartheta}) \cdot p(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}} \\ &= \operatorname{argmax}_{\hat{\boldsymbol{\theta}}} p(\mathbf{y}|\mathbf{x}; \hat{\boldsymbol{\theta}}) \cdot p(\hat{\boldsymbol{\theta}})\end{aligned}$$

posterior distribution

likelihood

prior distribution

- approximation to Bayesian inference (using mode of posterior distribution)
- maximum likelihood special case of maximum a posteriori estimation with uniform prior

Priors as Regularizers

prior distribution: leveraging information that cannot be found in training data

many regularized estimation strategies can be interpreted as MAP

- L^2 regularization corresponds to MAP with Gaussian prior on parameters
- L^1 regularization corresponds to MAP with isotropic Laplace prior on parameters

usage of conjugate priors allows analytical solution of MAP (as mode of posterior distribution can be given in closed form)

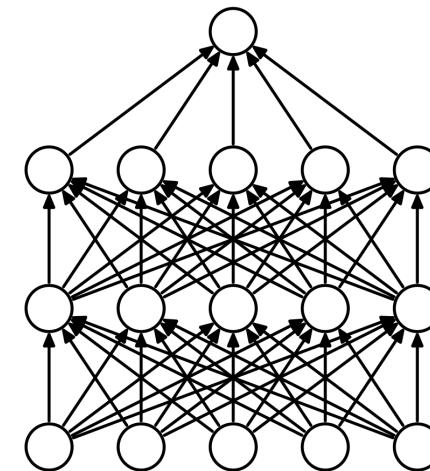
- conjugate distributions: prior $p(\boldsymbol{\theta})$ and posterior $p(\boldsymbol{\theta}|y, \mathbf{x})$ in same probability distribution family, e.g., exponential family)
- $p(\boldsymbol{\theta})$ conjugate prior for likelihood $p(y|\mathbf{x}; \boldsymbol{\theta})$

Dropout in Neural Networks

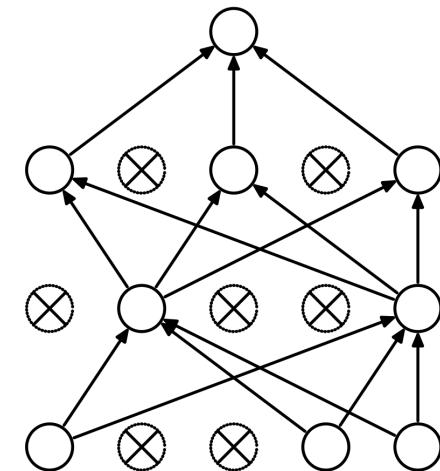
goal: prevent overfitting of large neural networks

idea: randomly drop non-output nodes (along with their connections) during training (not prediction)

- implicit ensemble method (kind of sampling from exponentially many differently thinned networks): for each mini-batch, randomly sample independent binary masks for the nodes
- inexpensive approximation to bagging (without training many different neural networks) with parameter sharing



(a) Standard Neural Net



(b) After applying dropout.

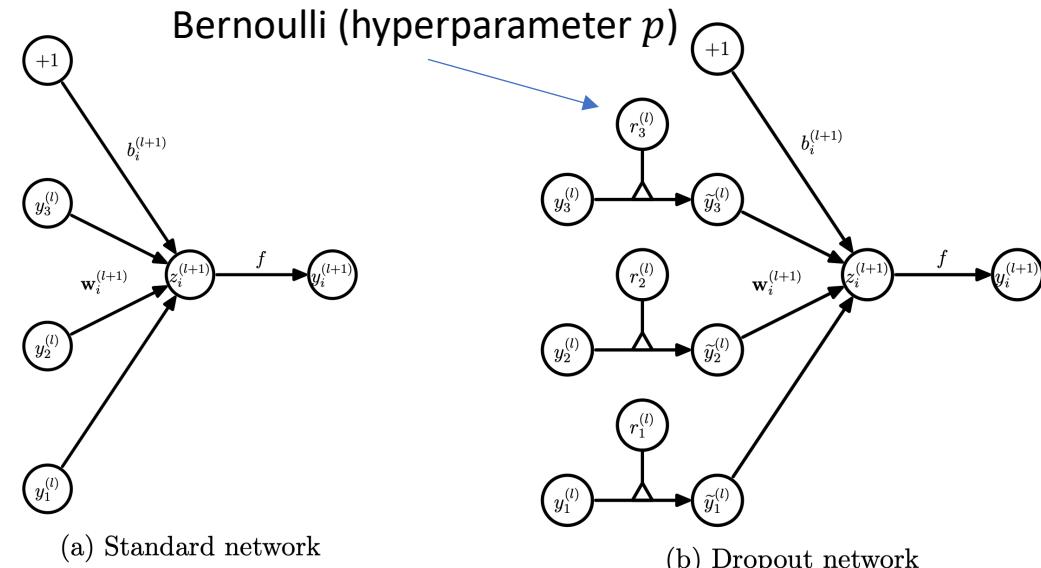
[source](#)

Dropout in Neural Networks

key advantage compared to averaging ensemble of independent models (typical bagging procedure):

better generalization by means of regularizing each hidden node to perform well regardless of which other hidden nodes are in the model (adaptability)

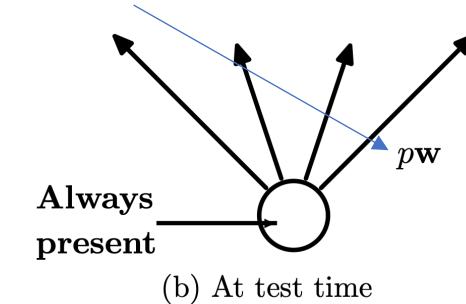
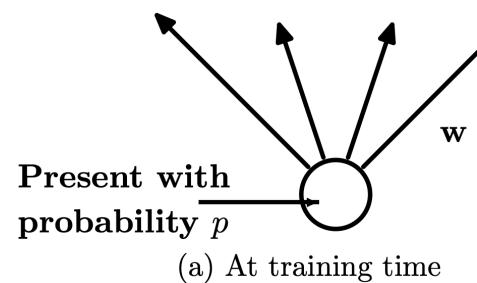
clever noise injection method: destroying extracted features rather than input values



(a) Standard network

(b) Dropout network

weight scaling inference rule
(approximating ensemble predictions)



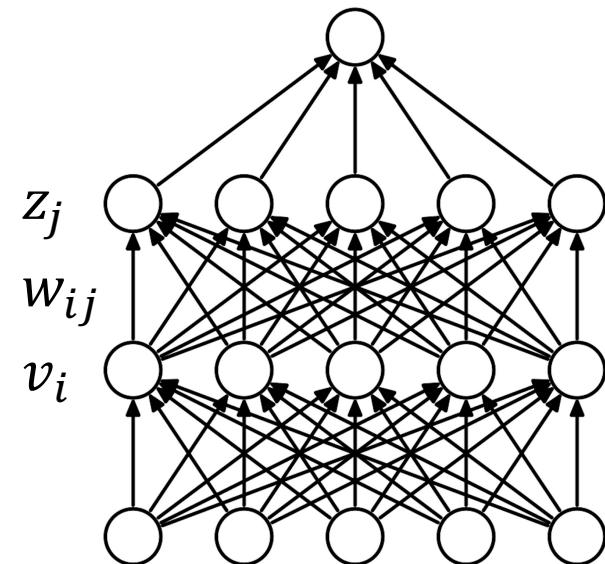
Convolutional Neural Networks (CNN)

Recap: Feed-Forward Neural Networks

computation in usual feed-forward network:

scalar input values z_j to activation function of nodes j in hidden or output layer as matrix multiplication of scalar output values v_i from activation function of nodes i from previous layer with connecting weights $w_{i,j}$

$$z_j = \sum_i v_i w_{i,j}$$



dropping dimension of different training observations in this view → loading full batch or mini-batches

Grid-Like Data

scalar value (like in usual feed-forward network)

data with grid-like topology (spatial structures), e.g.:

- time-series data: 1-D grid of data taken at regular time intervals (can also be done with recurrent neural networks or transformers)
- image data: 2-D grid of pixels (→ computer vision)

convolutional networks: neural networks using convolution in place of general matrix multiplication in at least one of their layers

→ highly regularized feed-forward networks



0	2	15	0	0	11	10	0	0	0	9	9	0	0	0	0
0	0	0	4	60	157	236	255	255	177	95	61	32	0	0	29
0	10	16	119	238	255	244	245	243	250	249	255	222	103	10	0
0	14	170	255	255	244	254	255	253	245	255	249	253	251	124	1
2	98	255	228	255	251	254	211	141	116	122	215	251	238	255	49
13	217	243	255	155	33	226	52	2	0	10	13	232	255	255	36
16	229	252	254	49	12	0	0	7	7	0	70	237	252	235	62
6	141	245	255	212	25	11	9	3	0	115	236	243	255	137	0
0	87	252	250	248	215	60	0	1	121	252	255	248	144	6	0
0	13	113	255	255	245	255	182	181	248	252	242	208	36	0	19
1	0	5	117	251	255	241	255	247	255	241	162	17	0	7	0
0	0	0	4	58	251	255	246	254	253	255	120	11	0	1	0
0	0	4	97	255	255	255	248	252	255	244	255	182	10	0	4
0	22	206	252	246	251	241	100	24	113	255	245	255	194	9	0
0	111	255	242	255	158	24	0	0	6	39	255	232	230	56	0
0	218	251	250	137	7	11	0	0	0	2	62	255	250	128	3
0	173	255	255	101	9	20	0	13	3	13	182	251	245	61	0
0	107	251	241	255	230	98	55	19	118	217	248	253	255	52	4
0	18	146	250	255	247	255	255	255	249	255	240	255	129	0	5
0	0	23	113	215	255	250	248	255	255	248	248	118	14	12	0
0	0	6	1	0	52	153	233	255	252	147	37	0	0	4	1
0	0	5	5	0	0	0	0	0	14	1	0	6	6	0	0

[source](#)

Convolution Operation

to be exact, usually rather cross-correlation instead of convolution operation (what would have – here):

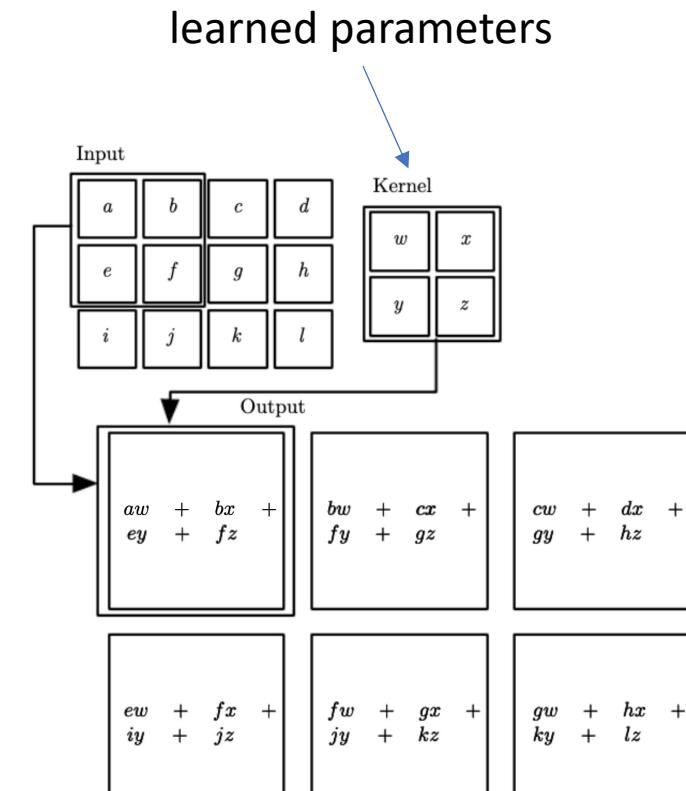
$$Z_{i,j} = (K * V)_{i,j} = \sum_{m,n} V_{i+m,j+n} K_{m,n}$$

feature map input (matrix) kernel (matrix)

again, dropping dimension of different training observations

matrices → tensors: several input channels c (e.g., RGB) and several output channels f (different feature maps, e.g., vertical edge, nose, ear, ...)

$$Z_{f,i,j} = \sum_{c,m,n} V_{c,i+m,j+n} K_{f,c,m,n}$$



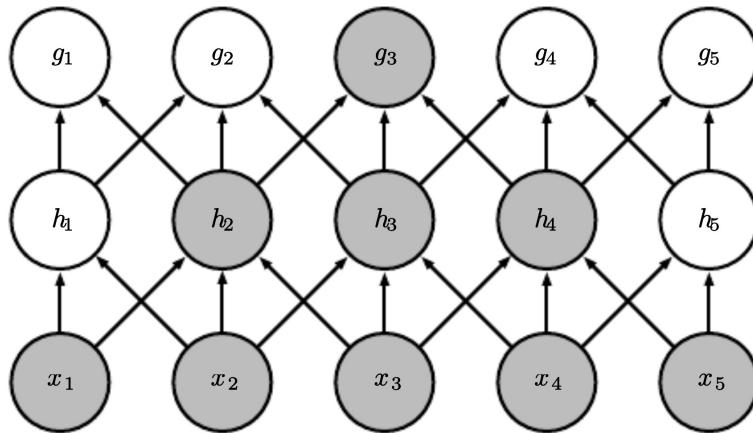
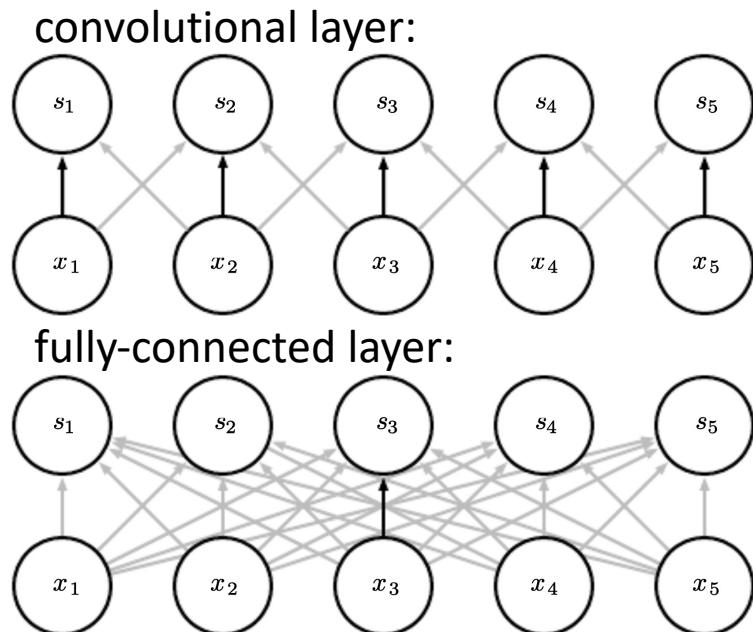
source

Regularization Effects

- sparse interactions: much less weights
- parameter sharing: use same weights for different connections
- equivariance to translation (not to rotation or scale): e.g., same edge detection across entire image

effect of receptive field:

- consider only locally restricted number of input values from previous layer
- grows for earlier layers (indirect interactions)
→ hierarchical patterns from simple building blocks



[source](#)

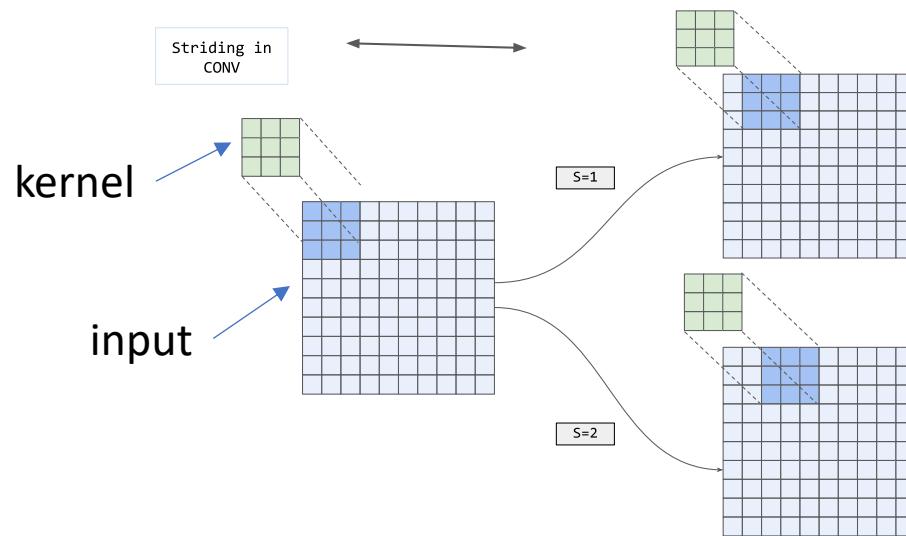
Important Details: Striding and Padding

need to define how to stride over image

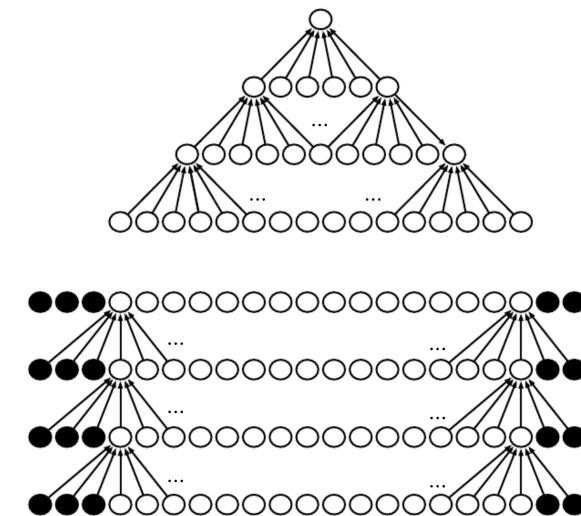
$$Z_{f,i,j} = \sum_{c,m,n} V_{c,i \times s + m, j \times s + n} K_{f,c,m,n}$$

$s > 1$ corresponds to down-sampling

→ fewer nodes after convolutional layer



zero-padding of input to make it wider:
otherwise shrinking of representation
with each layer (depending on kernel
size) → allowing large kernels and slow
shrinkage



[source](#)

Another Ingredient: Pooling

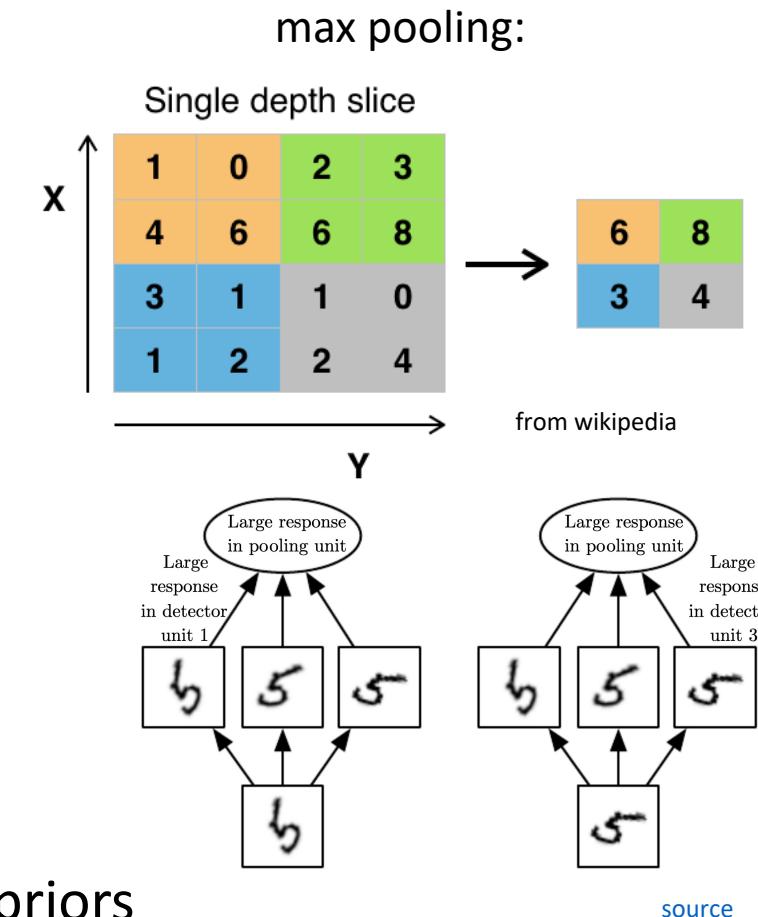
replacing outputs of neighboring nodes with summary statistic
(e.g., maximum or average value of nodes)

→ non-linear down-sampling (regularization)

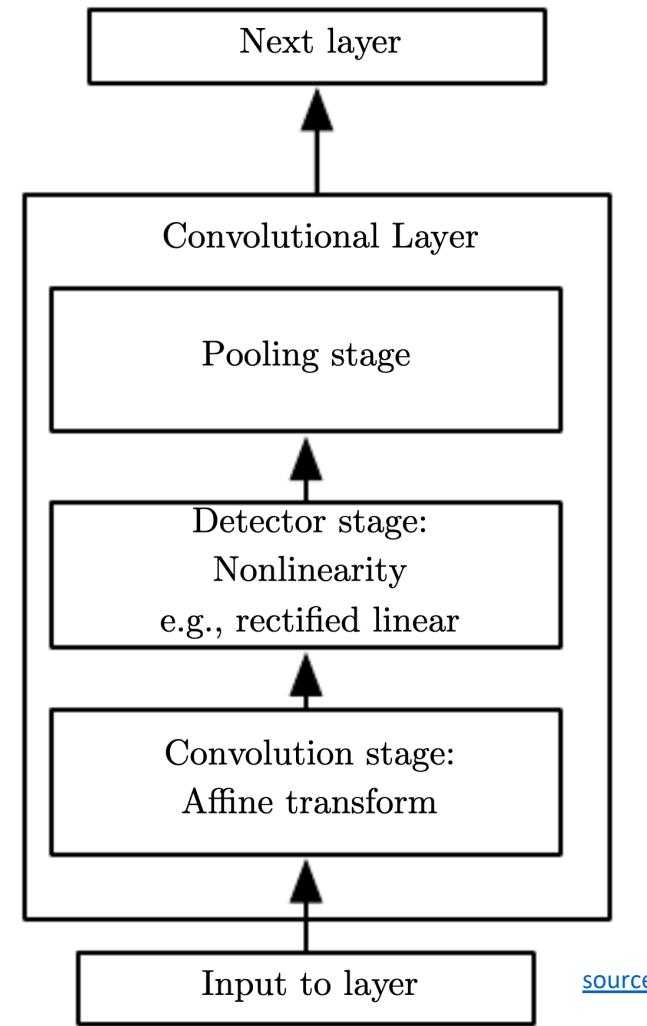
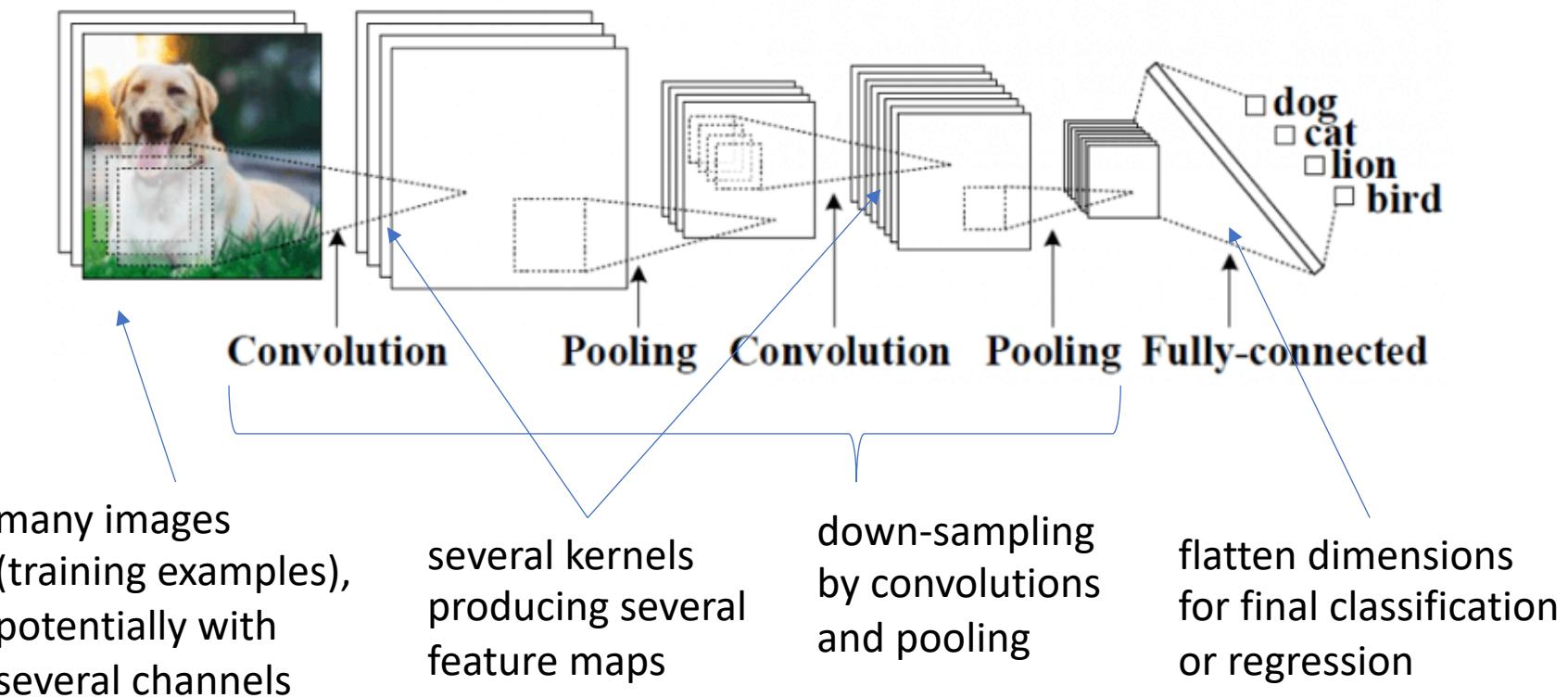
local translation invariance: no interest in exact position
pooling over features learned by separate kernels (multi-channel) can learn other transformation invariances

convolution and pooling can be interpreted as infinitely strong priors

- convolution: only local interactions, equivariant to translation
- pooling: invariant to small translations



Putting It All Together

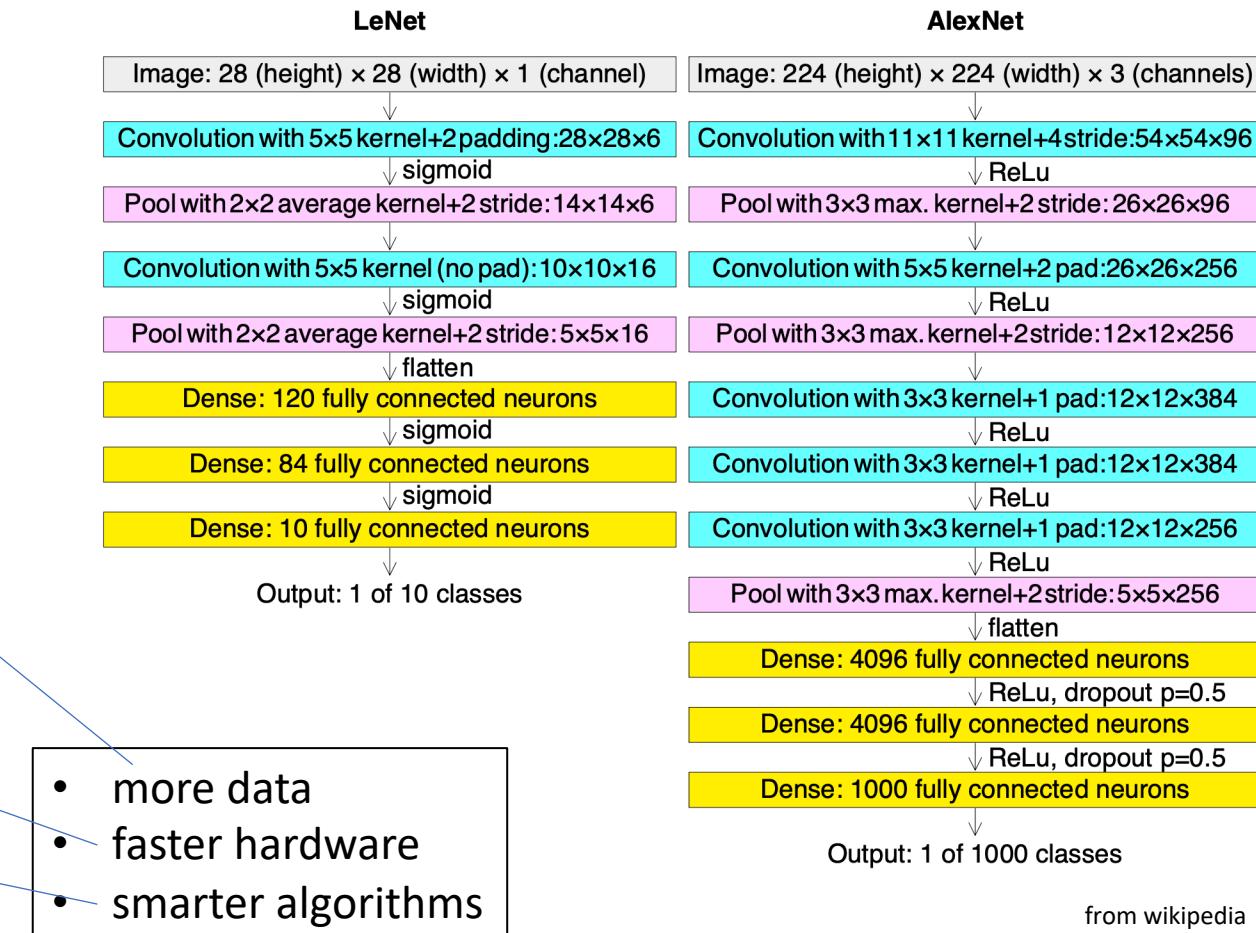


The Pivotal Moment: AlexNet

AlexNet finally started deep learning hype (won ImageNet challenge 2012)

some major improvements:

- GPU implementation (enabling more layers → better hierarchical representation)
- using ReLU activation functions
- using dropout



- more data
- faster hardware
- smarter algorithms

Inductive Bias

Inductive Bias (aka Learning Bias)

set of assumptions that a learning algorithm uses to predict outputs of given inputs that it has not encountered
(e.g., maximum margin in SVM, translation invariance in CNN)

- crucial piece of generalization
- data in disguise (replacement for missing information on specific situations in limited training data sets)

No Free Lunch Theorem

All optimization/ML algorithms (both sophisticated and simple ones) perform equally well when their performance is averaged across all possible problems.

(But deep learning is trying to solve many problems with very general-purpose forms of regularization.)

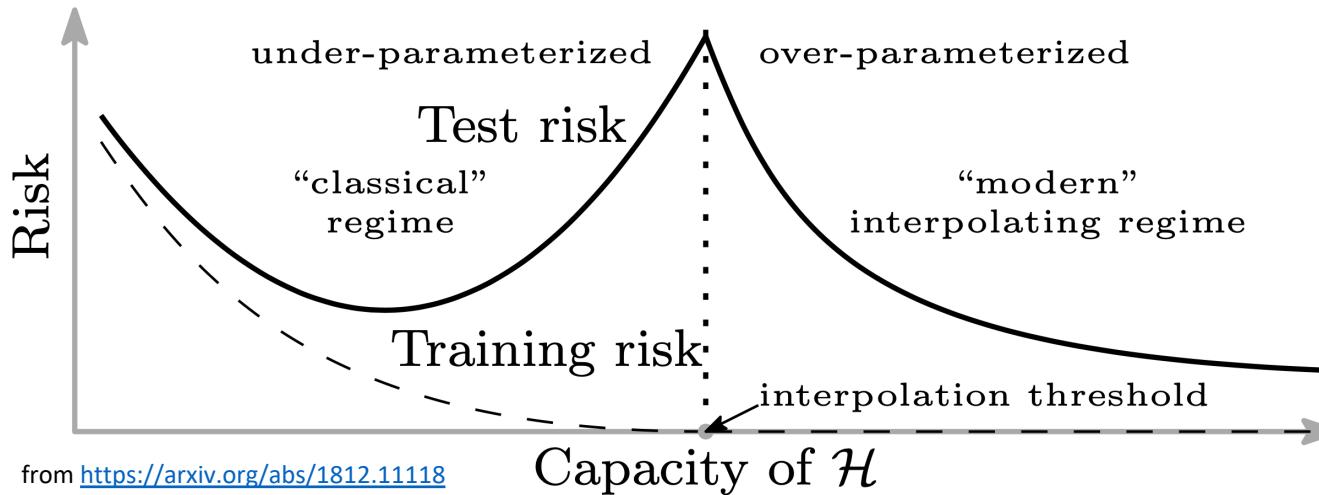
Particularly, model complexity doesn't reflect if inductive bias is appropriate for problem at hand.

→ choose right ML method for learning task at hand

Double Descent

Double Descent in Deep Learning

deep neural networks show additional structure for many parameters:



- generalization error first decreases
- then peaks (sharp increase in variance) at the interpolation threshold (where training error vanishes)
- then decreases monotonically again

- peak can be suppressed by regularization
- but potentially also suppression of second descent in case of sub-optimal inductive bias

not only Deep Learning: sufficiently complex, i.e., over-parametrized, ML models

Interpretation of Double Descent

all models right of interpolation threshold fit training data perfectly

find smoothest function that perfectly fits observed data (best inductive bias for task at hand)

→ a form of Occam's razor (the simplest explanation compatible with the observations should be preferred)

larger function classes contain more interpolating functions for training data at hand, including simpler/smooth ones (smaller norm)

→ increasing model complexity can improve generalization (but need more data)

again: over-parametrize and regularize

Memorization First Step Toward Generalization?

many interpolating models (perfectly fitting training data) in over-parameterization regime

empirical evidence that explicit regularization is not necessary for generalization (but helps to choose smooth solution)

→ implicit regularization (tendency of an algorithm to seek out solutions that generalize well on its own)

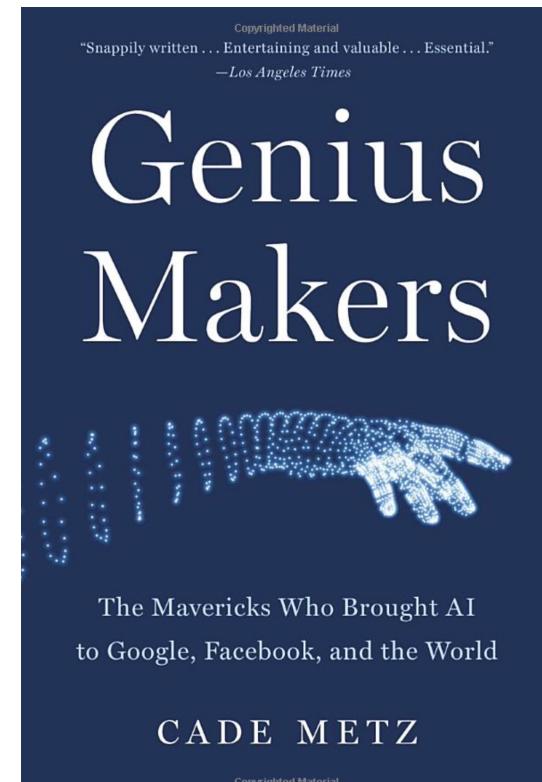
running **Stochastic Gradient Descent** to convergence: pick minimum norm solution of different interpolating models → implicitly regularizing

Literature

papers:

- CNN: neural networks work (at least for LeCun ☺)
- AlexNet: deep learning takes over
- dropout

nice historical overview:



Key Challenge for AI: Generalization

common sense of humans or animals enabled by world model
(generalized representation of the world, allows to be inattentive to vast amount of irrelevant details)

not yet reached this level of generalization in AI systems
without common sense: need for tons of data and hard-wired engineering of corner cases (example: self-driving cars)