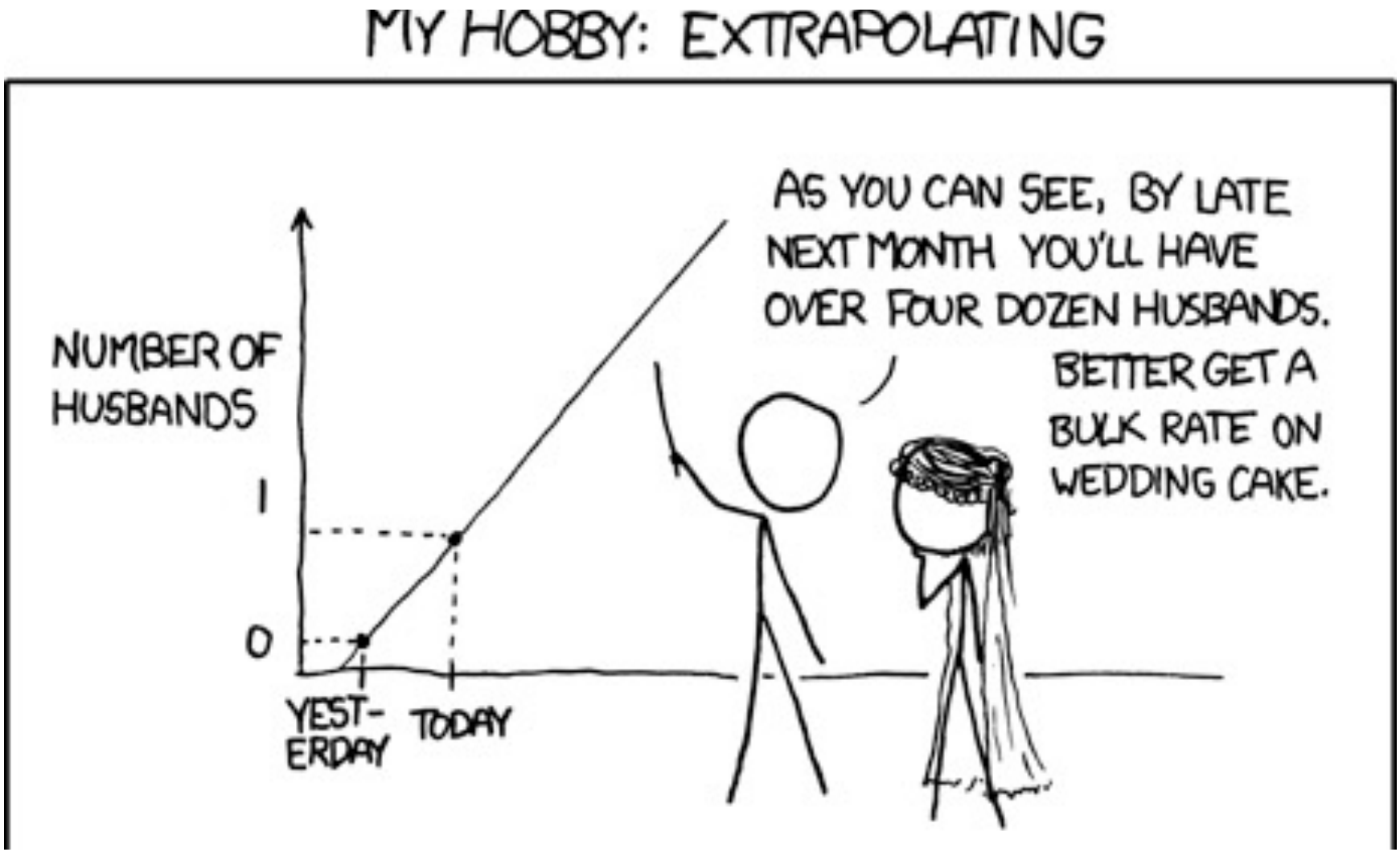
Stochastic Gradient Descent

On-line learning and intro to (Feedforward) Neural Networks

Beyond linear Regression

- almost linear methods
  - additive models, piecewise linear (e.g. CART); locally linear regression, ….
- nonlinear models
  - linear in fixed transform ("phi") space (Fourier, Gabor, polynomial…)
    - Hence linear in the transformed space.
  - general nonlinear forms (basis functions are adaptive as well)
    - e.g. feedforward neural networks (MLP…)
    - MLPs are universal approximators, just like polynomials.
  - Universal approximator: Given any continuous function, there is some member of that family that can exactly match it

Knowing Scope of Models



How to Learn the Weights?

- **Linear Regression** with Cost E(w) = SSE or MSE
  - **direct** solution ("pseudo-inverse solution") gives optimal weights, w*
  - **One Step:** Given initial guess, $w^{(0)}$, can find w* in one update step.
    - Realize that E(w)) is quadratic in **w:** $E(\mathbf{w}) = \text{SSE}(\mathbf{w}) = \text{SSE}(\mathbf{w}^*) + (\mathbf{w}-\mathbf{w}^*)^T\mathbf{Q}(\mathbf{w}-\mathbf{w}^*)$
    - **w*** is the optimum solution, and **Q** a positive definite matrix.
    - So only one update of Newton's "root finding" step is needed.
  - **Iterative using Gradient Descent:** Given initial guess, $w^{(0)}$, iteratively update w by "going down the gradient" till you reach w*.

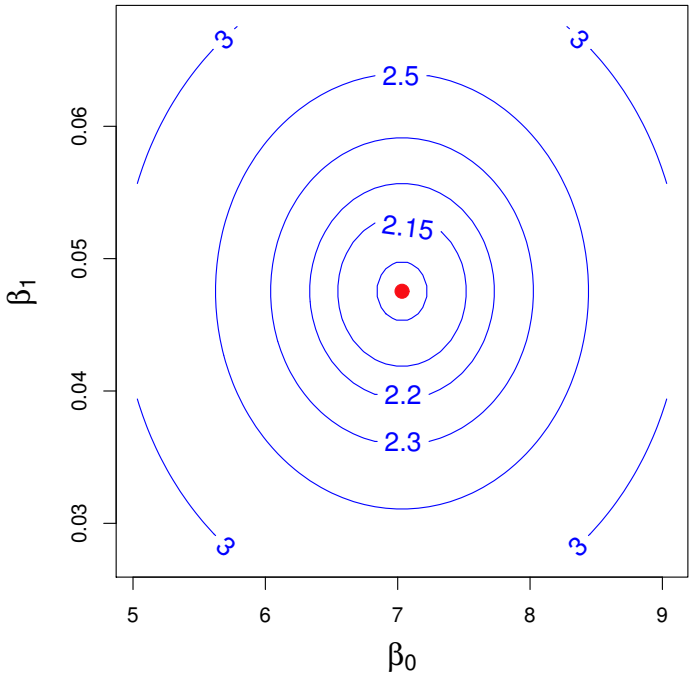
- **Nonlinear models:** E(w) is not quadratic, and in general non-convex.
  - weights are typically updated in an iterative manner (could be "on-line" or batch iterative)
  - SGD a popular choice, is on-line.

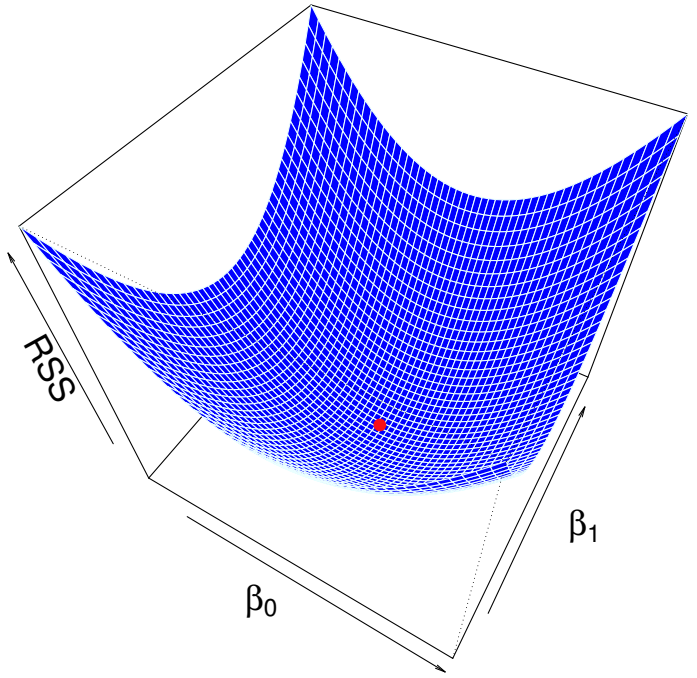
Gradient Descent for MLR

- **See (very introductory) Coursera Lectures by Andrew Ng, and Bishop Ch 5.1, 5.2**
- The cost function $E(\mathbf{w})$ is quadratic in $\mathbf{w}$
  - weights can be obtained by doing gradient descent **(incrementally moving down the cost surface in weight space)**

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E$$

Learning rate





Learning Rate

- Learning rate η is crucial
  - Too low: slow convergence
  - Somewhat high: convergence with oscillations
  - Too high: unstable/divergence

What happens when gradients differ greatly across the weight dimensions? (Sketch below).

Gradient Descent for **Non-Linear Models**

E(w) is not quadratic, and in general non-convex, with *multiple local minima*.

Visualize GD as well as some second-order methods

http://www.benfrederickson.com/numerical-optimization/

(look at Gradient Descent example first, then Nelder-Mead)

- motivates adaptive learning rates, and second order methods that look at curvatures (2^{nd} derivative) in addition to gradients.

- Note: True gradient descent is a batch algorithm, involving all of the training data. (why?)

Stochastic gradient descent (SGD)

- "on-line" version (**stochastic gradient descent or SGD**): replace true gradient by "instantaneous" gradient, that reduces error only on the new instance (pun intended!)
- e.g. for linear models, with τ as (inner loop) iteration number, n denoting the datapoint being considered, we get:

$$\begin{aligned} \mathbf{w}^{(\tau+1)} &= \mathbf{w}^{(\tau)} - \eta \nabla E_n \\ &= \mathbf{w}^{(\tau)} + \eta(t_n - \mathbf{w}^{(\tau)\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n))\boldsymbol{\phi}(\mathbf{x}_n). \end{aligned}$$

This is known as the least-mean-squares (LMS) algorithm or the Widrow-Hoff rule.

Note: Data items considered one at a time (a.k.a. online learning)

Stokhos = "to aim"

Toy illustration for SGD

y = cost;

x is scalar parameter that can be adjusted

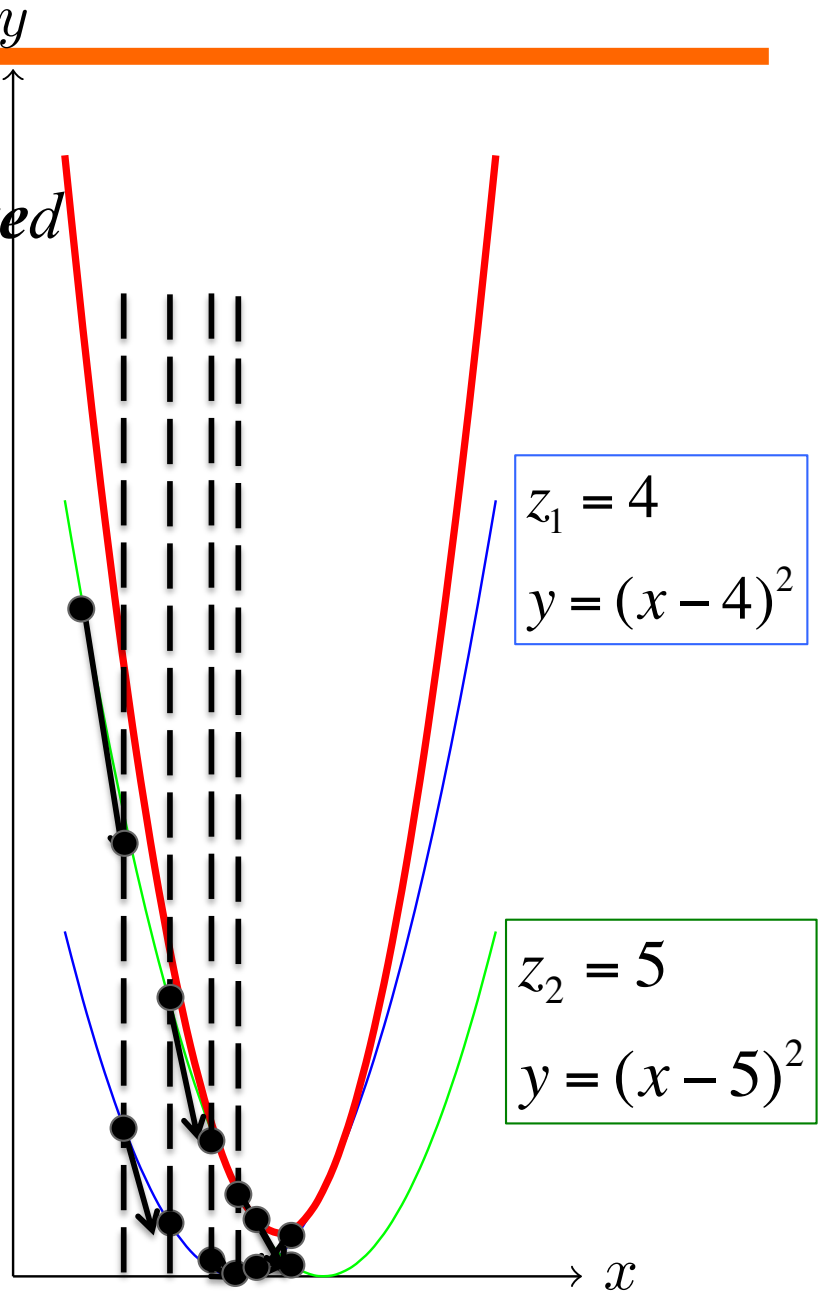
$$y = (x-4)^2 + (x-5)^2$$

$$\min_x \sum_i (x - z_i)^2 = \min_x \sum_i \ell(x, z_i)$$

$$\ell(x, z) = (x - z)^2$$

$$x^{t+1} = x^t - \eta \frac{\partial \ell(x^t, 4)}{\partial x}$$

$$x^{t+2} = x^{t+1} - \eta \frac{\partial \ell(x^{t+1}, 5)}{\partial x}$$



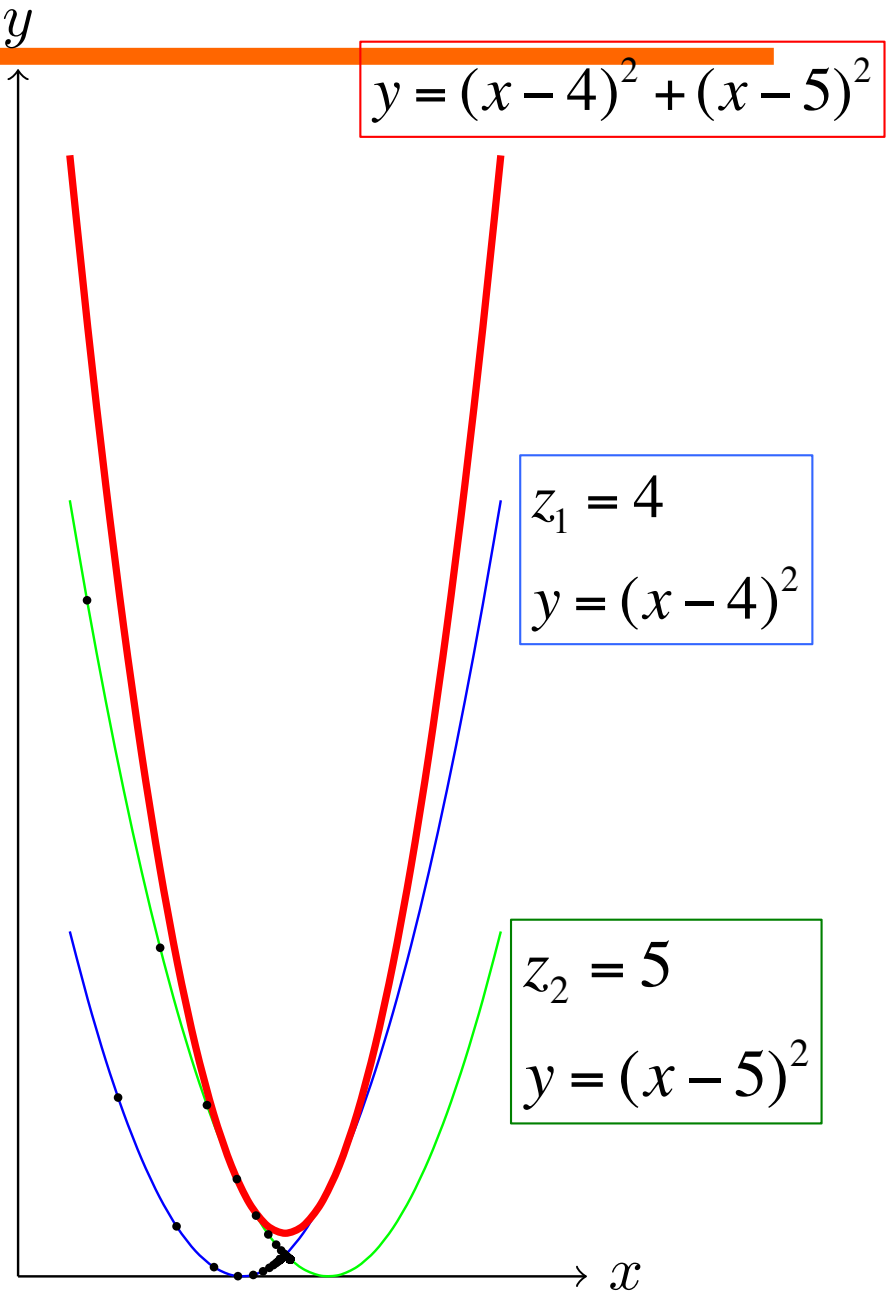
Toy illustration for SGD

$$\min_x \sum_i (x - z_i)^2 = \min_x \sum_i \ell(x, z_i)$$

$$\ell(x, z) = (x - z)^2$$

$$x^{t+1} = x^t - \eta \frac{\partial \ell(x^t, 4)}{\partial x}$$

$$x^{t+2} = x^{t+1} - \eta \frac{\partial \ell(x^{t+1}, 5)}{\partial x}$$



Sequential Learning with SGD

- Learning rate: too small → too slow
  - Too large → oscillatory, or may even diverge
- Should η be fixed or adaptive (second order methods)?

- Is convergence needed or not?
  - Streaming data?
    - Non-stationary? May not want to converge!
    - If convergence is desired, then η should decrease with time.
      - (e.g. choose $\eta = a/t$)
  - Batch training data?
    - Two loops:
      - outer: # of epochs
      - inner (one per epoch) = one pass over the training dataset.
    - Typically stop when validation error "flattens" vs. # of epochs.

Why SGD?

Better for large data sets

+ Often faster than batch gradient descent

Can do "mini-batches" as a practical trade-off

+ Less prone to local minima, so often applied to complex models (and correspondingly complex error surfaces)

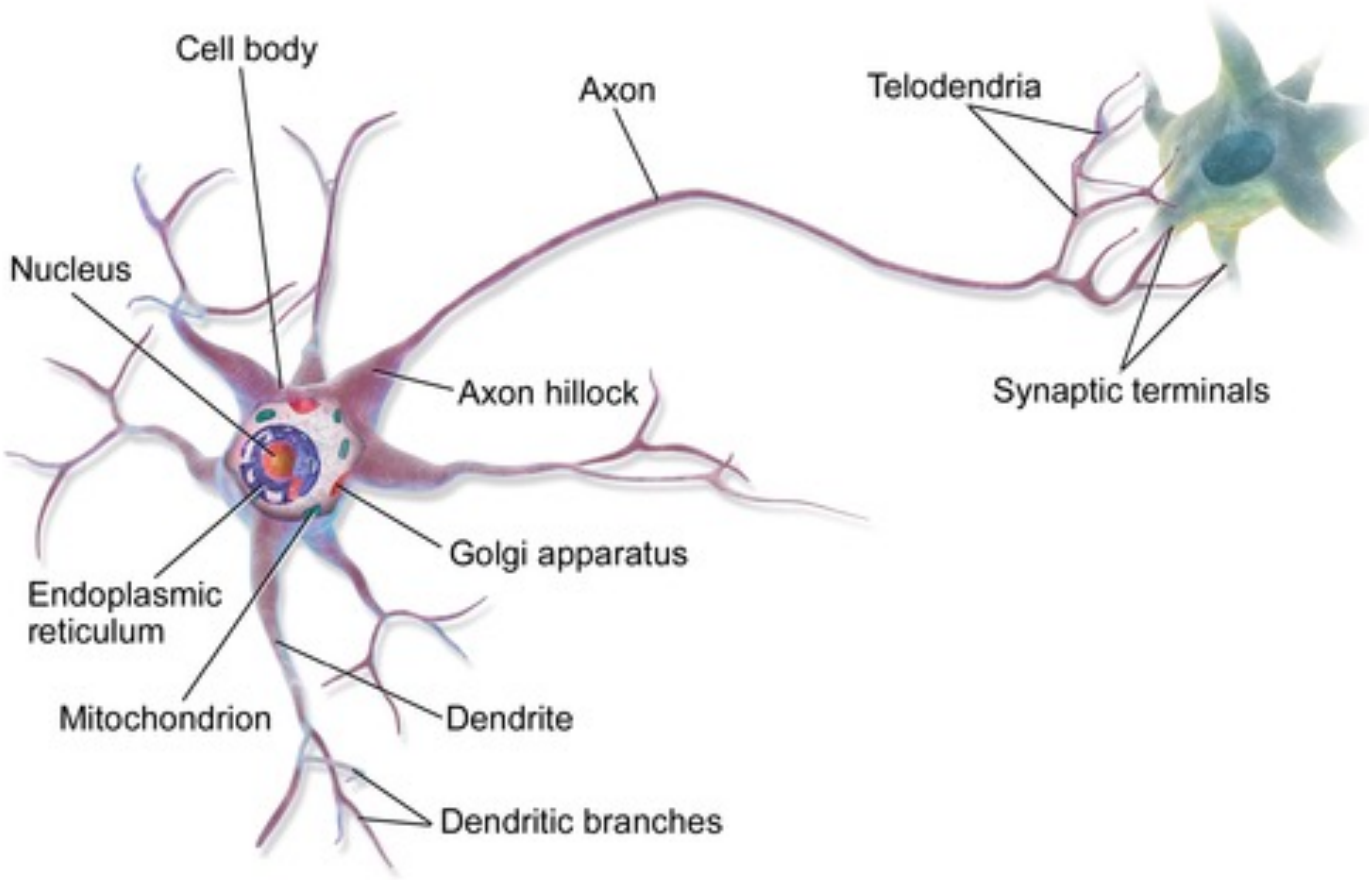
+ useful to scale inputs to help find learning rate (usually fixed)

+ works for non-stationary environments as well as online settings

Also used for more complex error surfaces, though many second-order methods, that also consider the Hessian (matrix of curvatures) in addition to the Jacobian, or vector of slopes, exist. (see "conjugate gradient" in the animation link on previous slide).

A Neuron

From https://en.wikipedia.org/wiki/Neuron



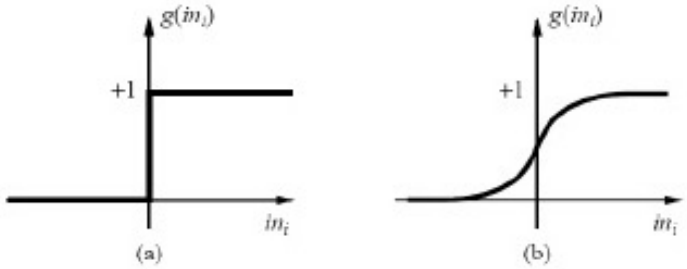
1-layer (Artificial) Neural Networks

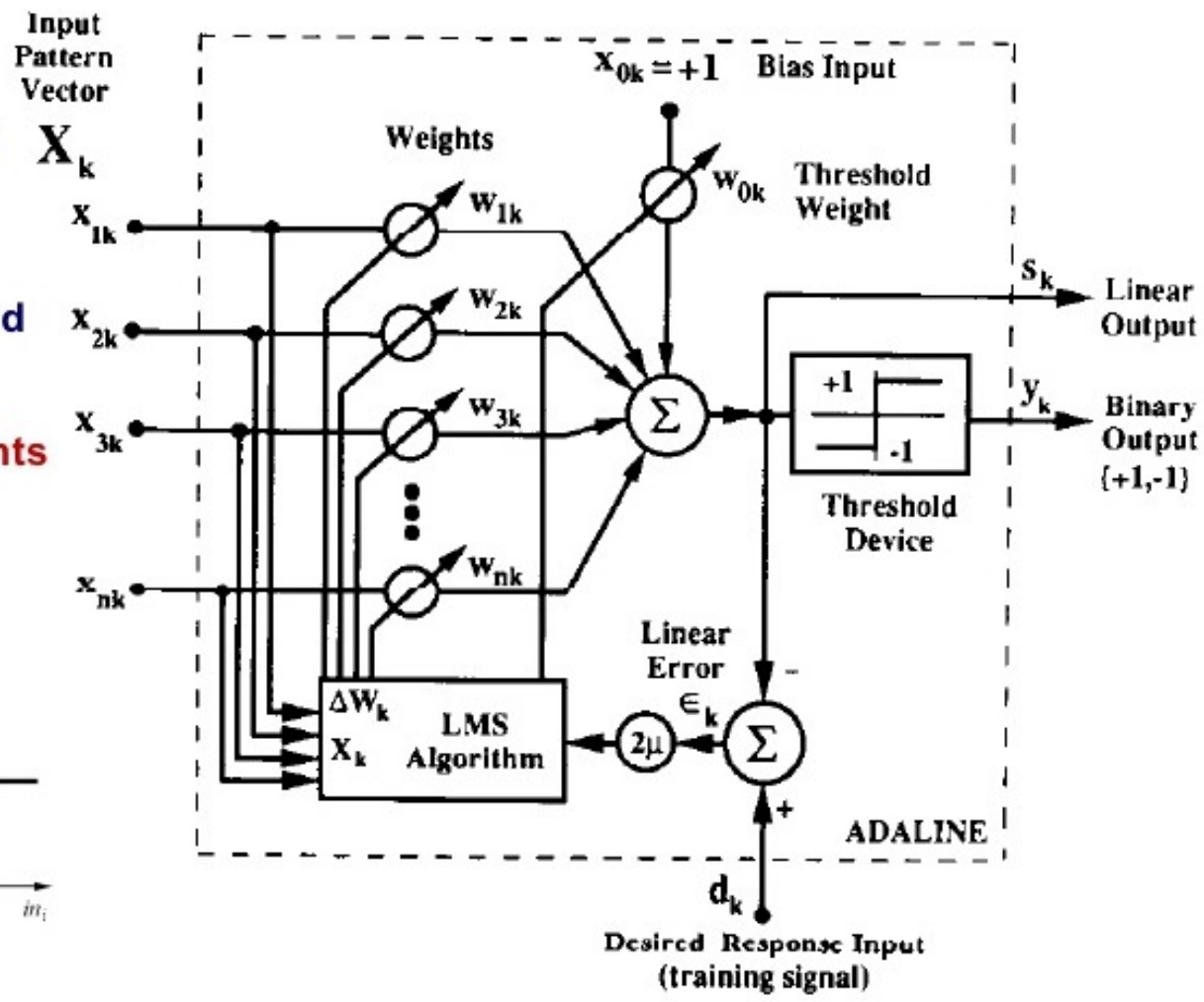
- Viewing (generalized) linear model as a “neural network” (draw below)

- Learning through SGD
- (déjà vu) LMS Algorithm, Widrow-Hoff rule from ADALINE (1960)

Adaline

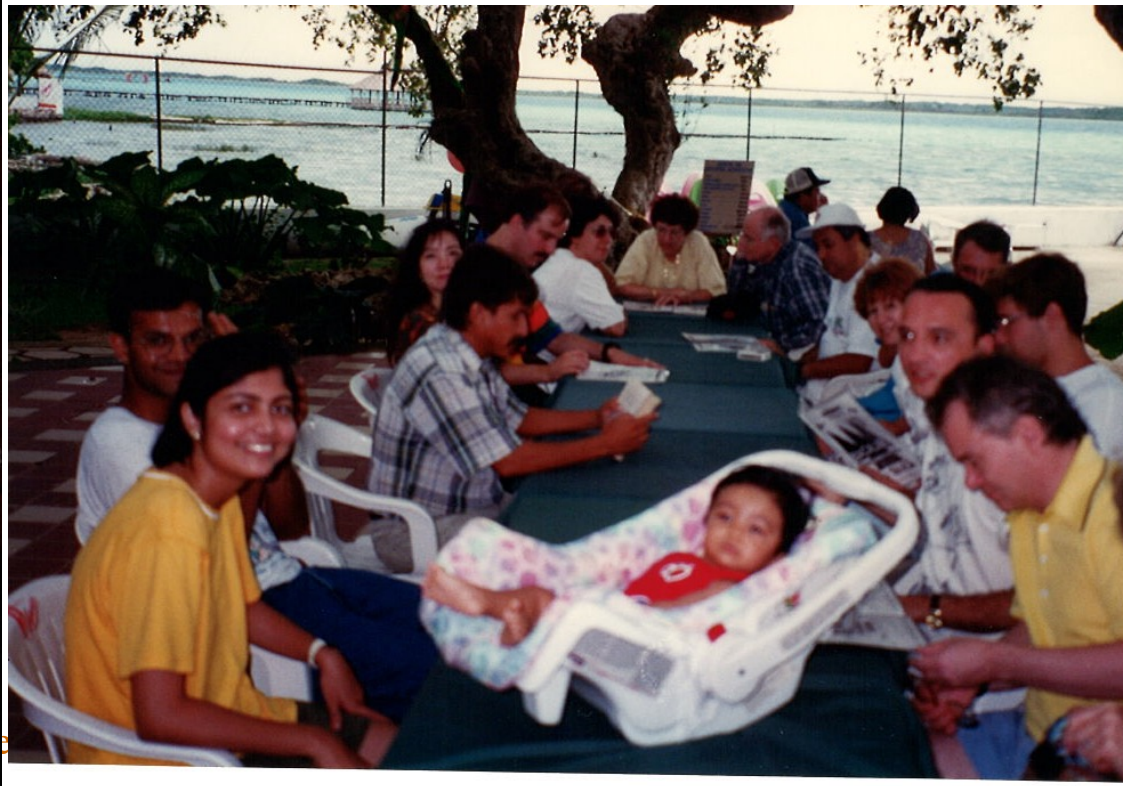
- Adaptive Linear Element
- Adaptive linear combiner cascaded with a hard-limiting quantizer
- Linear output transformed to binary by means of a threshold device
- Training = adjusting the weights
- Activation functions





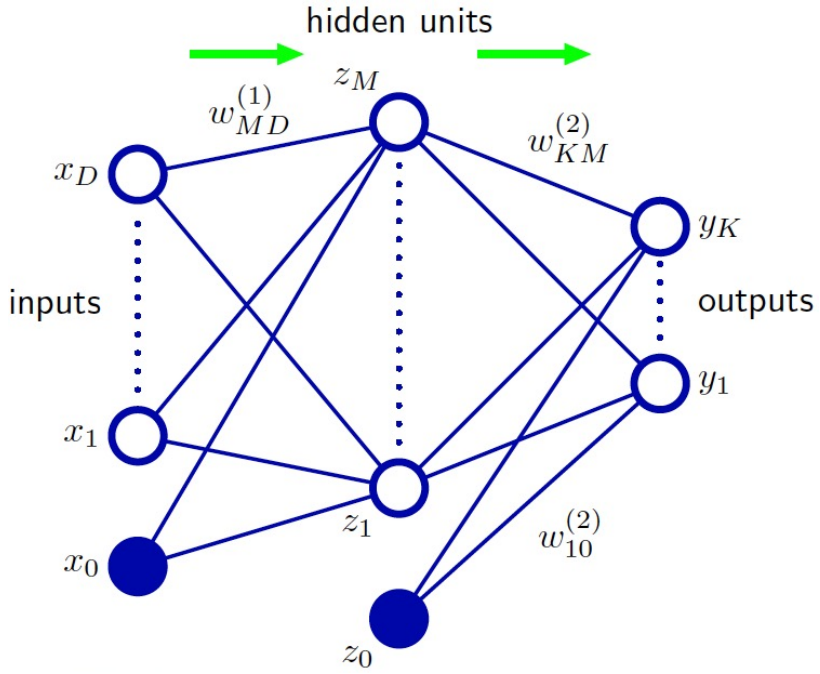
ADALINE

- https://www.youtube.com/watch?v=hc2Zj55j1zU



Multi-Layered Perceptrons (MLP) (most popular neural network before deep learning)

- Bishop Fig 5.1 (right). Typical 2-layered MLP for Regression (with one hidden layer of **M units** and 2 layers of adaptive weights):
  Choice of activation functions:
  - Hidden layer: tanh/sigmoid
  - Output layer: linear / sigmoid or softmax
- **Universal approximator** if output layer cells are linear and hiddent layer is nonlinear with h() = tanh/sigmoid/gaussian/…



- Hidden layer with activation function h():

$$a_j = \sum_{i=1}^{D} w^{(1)}_{ji} x_i + w^{(1)}_{j0} \qquad z_j = h(a_j)$$

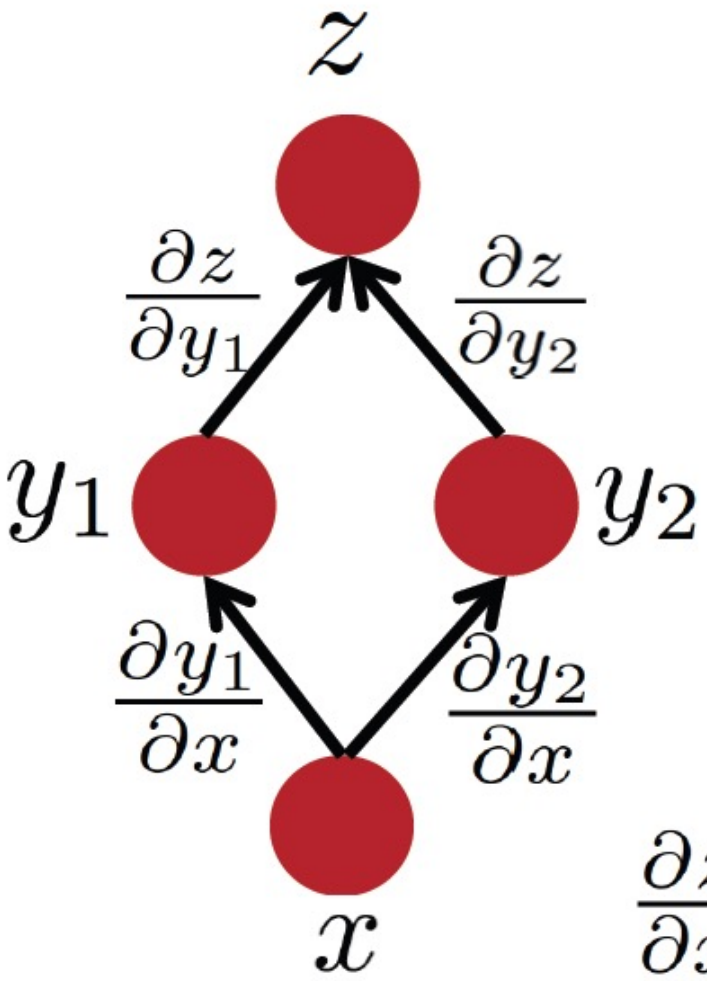
Output layer with activation functio[illegible]():

$$a_k = \sum_{j=1}^{M} w^{(2)}_{kj} z_j + w^{(2)}_{k0} \qquad y_k = \sigma(a_k)$$

Overall:

$$y_k(\mathbf{x}, \mathbf{w}) = \sigma\left(\sum_{j=1}^{M} w^{(2)}_{kj} h\left(\sum_{i=1}^{D} w^{(1)}_{ji} x_i + w^{(1)}_{j0} \right) + w^{(2)}_{k0} \right)$$

(Multi-path) Chain Rule from Calculus



$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y_1}\frac{\partial y_1}{\partial x} + \frac{\partial z}{\partial y_2}\frac{\partial y_2}{\partial x}$$

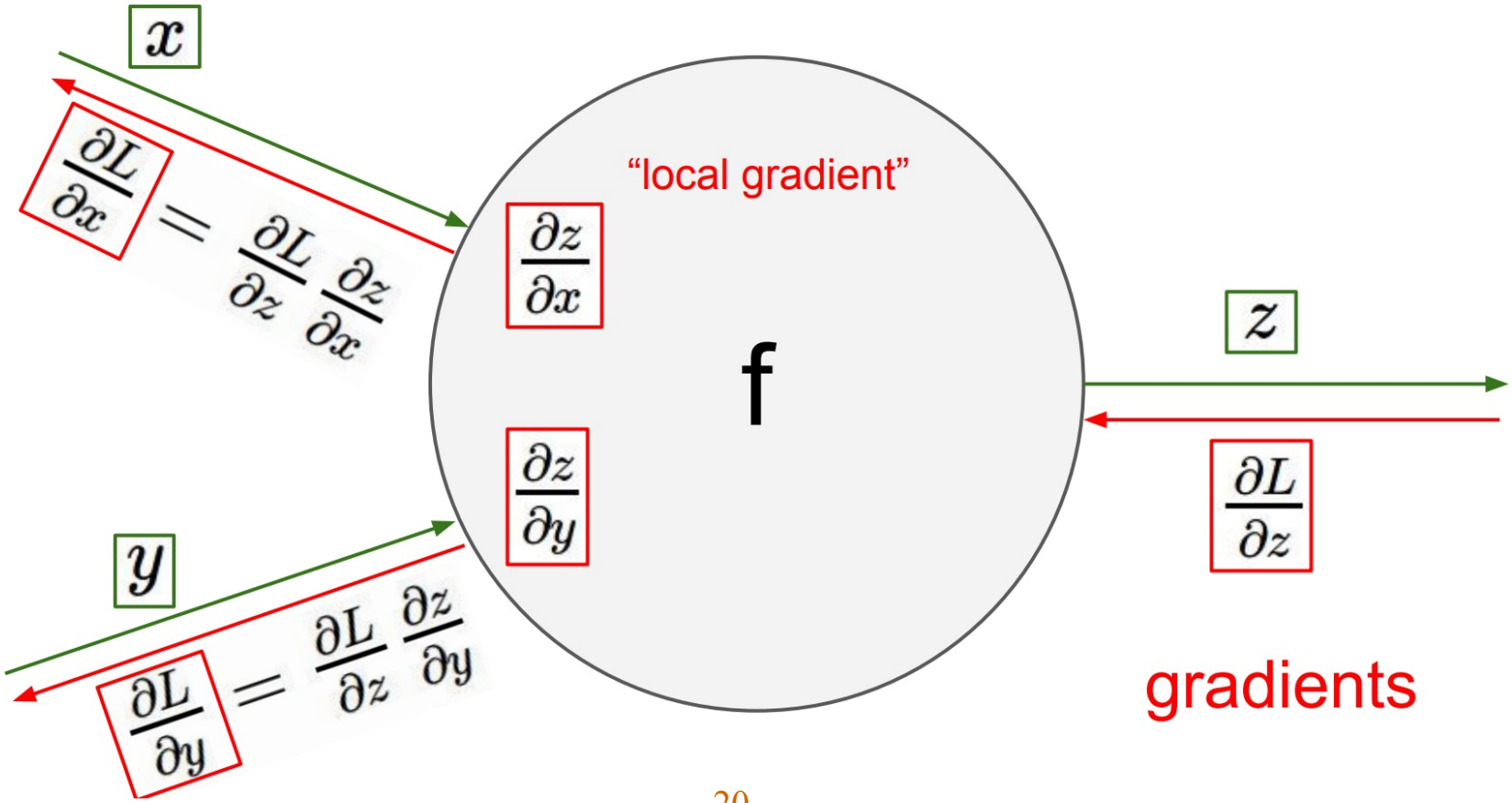
Learning via Error Backpropagation

- 1. forward pass (from input layer through to output layer): compute outputs of successive layers, given the inputs to that layer
- 2. Compute error (loss) at final output layer by comparing with desired output values.
- 3. backward pass (from output layer to input layer): compute gradients of weights w.r.t. loss, layer by layer; and update all the weights.

- Cycle through all the data:
  - For batch settings: ONE EPOCH = one pass through the training dataset.

Backpropagation: Intuition

See Bishop 5.3.1 for application to MLP
Z = cell output, feeding into next layer that is closer to the final output
f= cell activation function. (relates z to inputs x, y)
Gradient of Error w.r.t. x (or y) involve previously computed gradient w.r.t. z



Local gradient times upstream gradient

Design Parameters to be Aware about

- # of hidden units/nodes
  - model complexity
- # of epochs/iterations
  - how long do you train: best is use a cross-validation set to decide when to stop
- Activation function
  - typically tanh or logistic, a.k.a. sigmoid
- Learning rate (SGD is used to update weights)
  - Speed of training

- Momentum: (a second order gradient descent method) https://distill.pub/2017/momentum/

Visualizing the Workings of an MLP

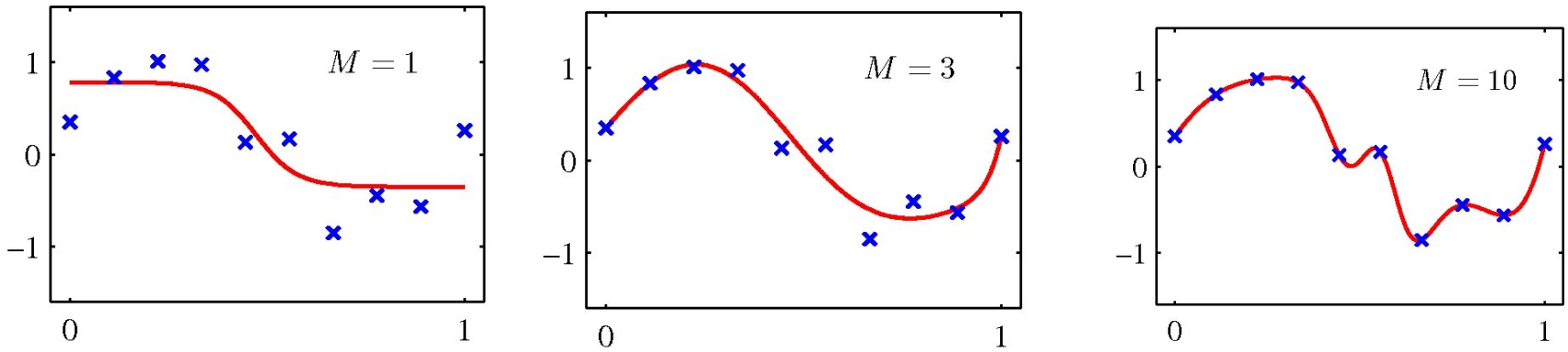
- http://playground.tensorflow.org/

- (both regression and classification examples with different degrees of difficulty).

- TRY IT OUT!

Model Complexity for MLP

- complexity related to # of hidden units AND amount of training (number of passes or epochs through the data)



MLPs with 1, 3 and 10 hidden units, trained till MSE_train flattened out

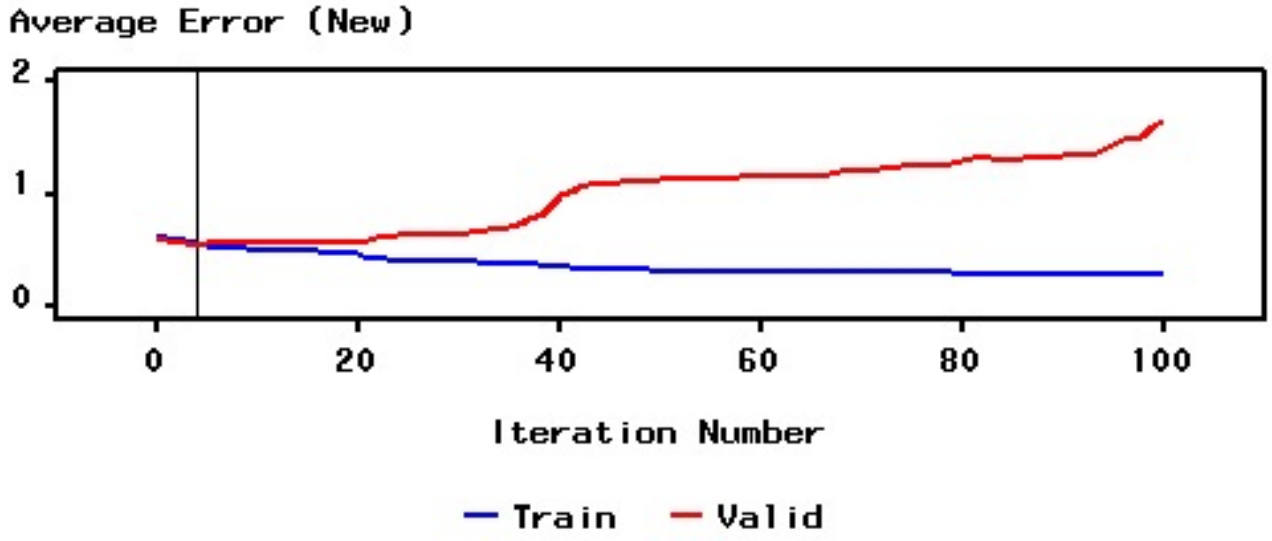
Fact: The "effective number of parameters" (or effective degrees of freedom) of an MLP increases with amount of training.

So, with less training M=10 solution looks like a network with M < 10 but with more training …

Adjusting Complexity

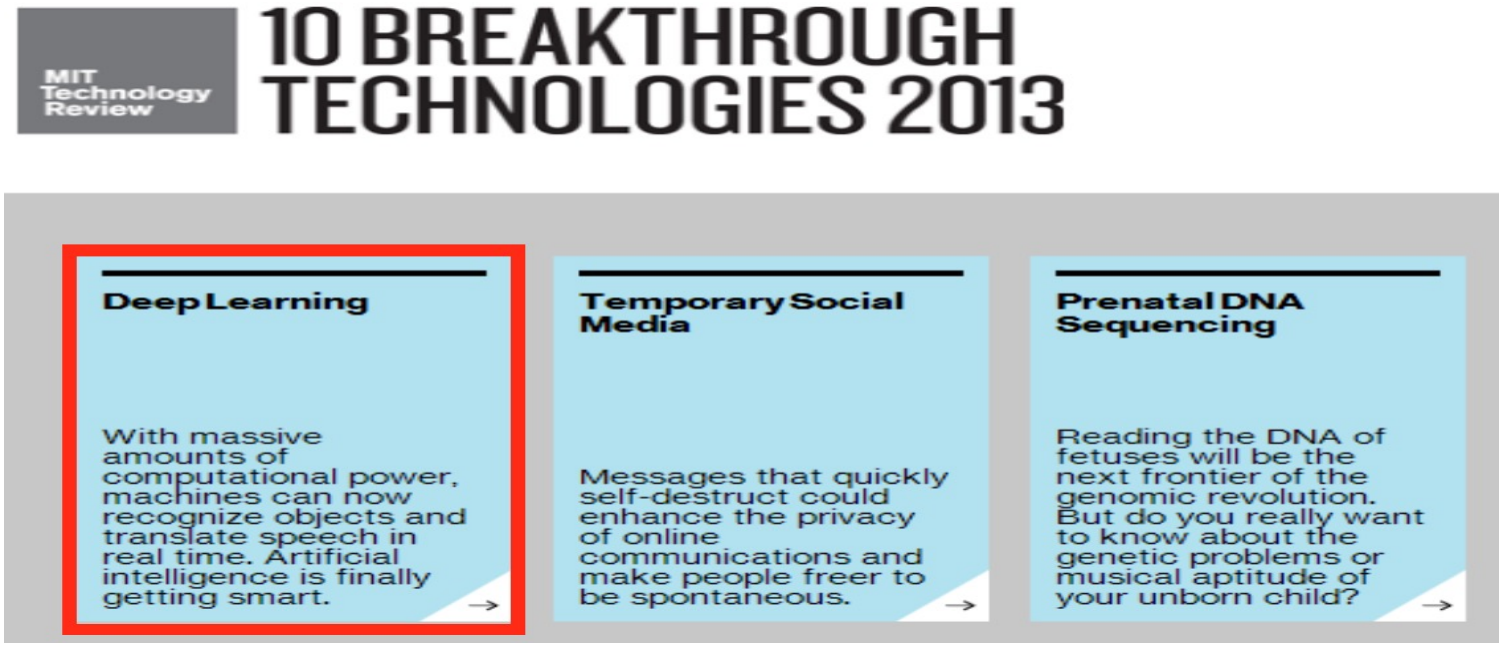
– "effective # of parameters" increases with # of epochs !!
  - Select adequately powerful model
  - while training, monitor performance using validation set
  - **stop training** when error on validation set reaches a minimum

  - SAS fig.



Deep Learning

- Amazing improvements in speech recognition, NLP, recognizing objects in images,…
- See http://deeplearning.net/
  - For hype/investment, see Frank Chen video https://www.youtube.com/watch?v=ht6fLrar91U starting 32:00



Going Deep

See tutorial at: http://www.iro.umontreal.ca/~bengioy/talks/mlss-austin.pdf

From Facebook's Deepface paper (2014)
https://www.cs.toronto.edu/~ranzato/publications/taigman_cvpr14.pdf
Our method reaches an accuracy of 97.35% on the Labeled Faces in the Wild (LFW) dataset, reducing the error of the current state of the art by more than 27%, closely approaching human-level performance.

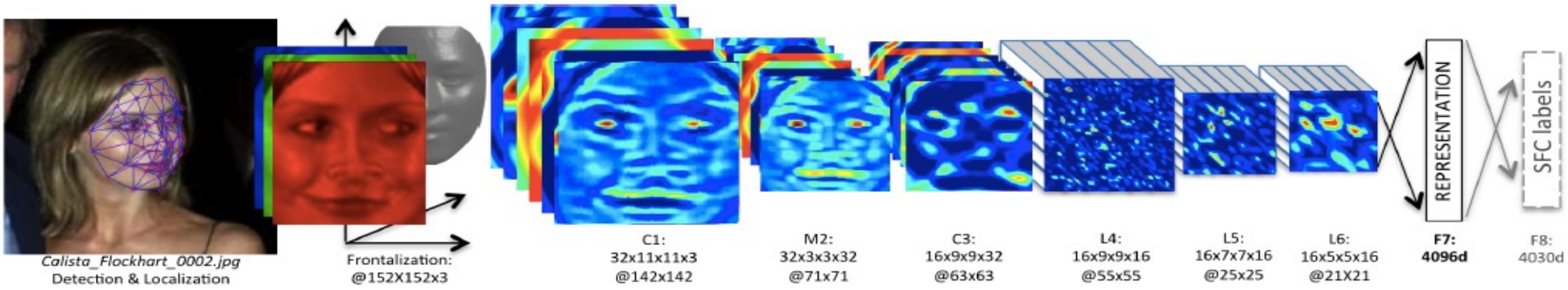
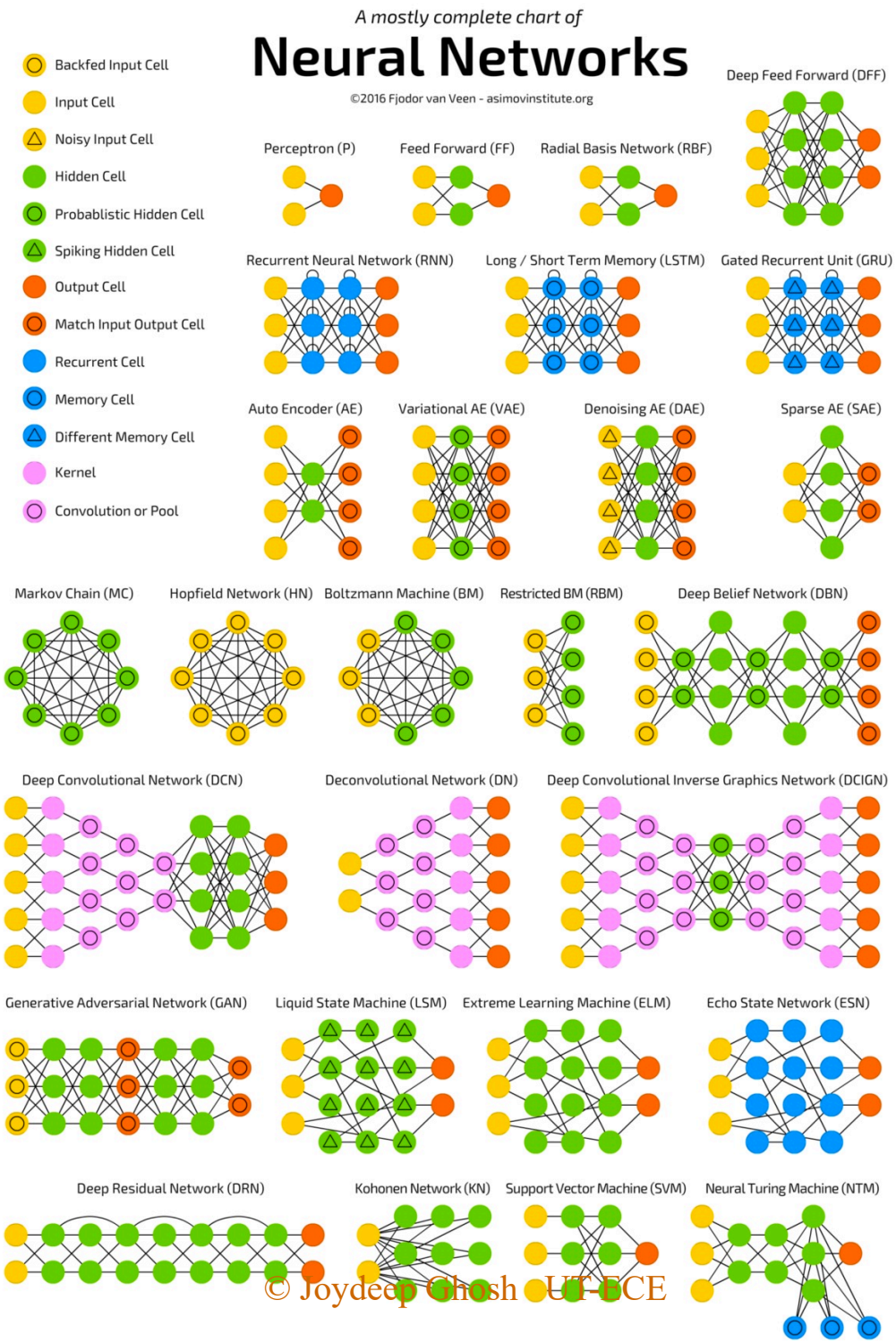


Figure 2. Outline of the *DeepFace* architecture. A front-end of a single convolution-pooling-convolution filtering on the rectified input, followed by three locally-connected layers and two fully-connected layers. Colors illustrate outputs for each layer. The net includes more than 120 million parameters, where more than 95% come from the local and fully connected layers.



So which method should I choose?

- Depends on type, complexity of problem; data size
  - (i) try linear regression first
    - Explore data; Study residues
    - do feature selection/transformation if needed
    - If robustness is needed, try "rlm" (R package), or use SVR.
  - (ii) Now try a set of powerful but not so interpretable models (MLP, deep learning, etc)
    - (estimate complexity of fit using a few trial runs)
  - How much is the gap between (i) and (ii)?
  - Consider Decision tree based regression if interpretation is important or a piecewise constant answer is more "actionable"

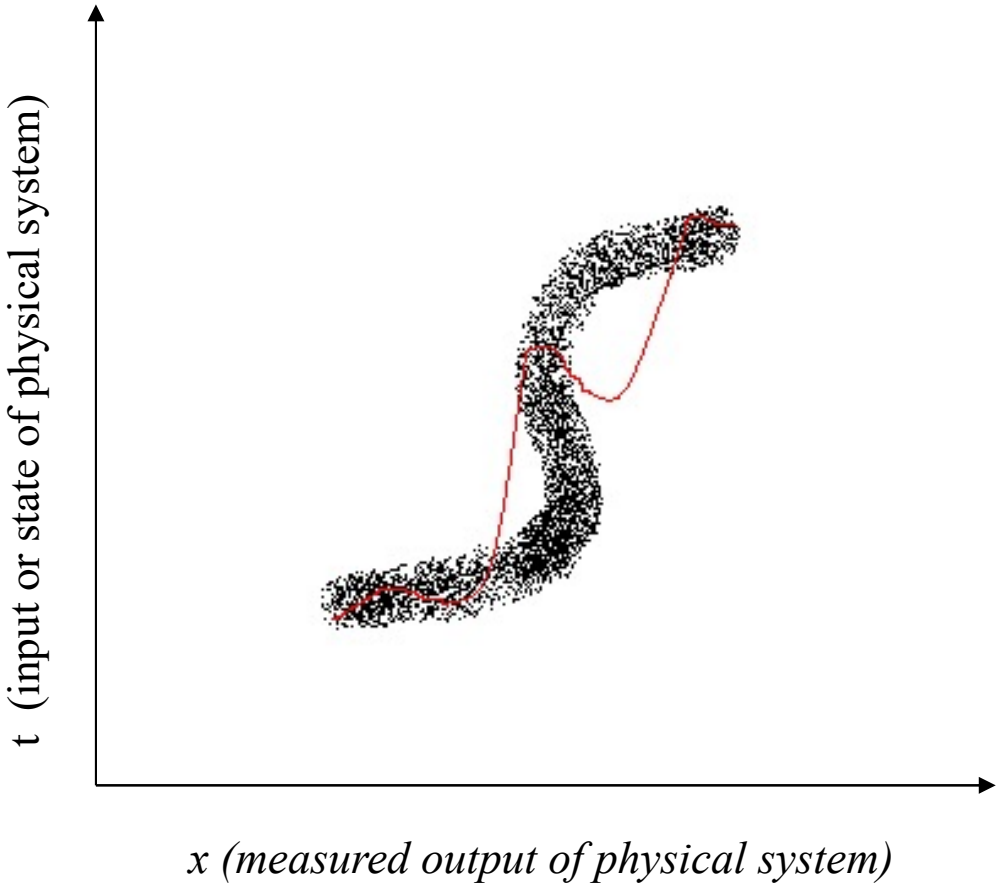
- Still lacking? try ensemble approaches specially GBDT/XGBoost, which also rank-orders the features.

Caution: Modeling **Inverse** Problems

- Forward problem reasonably characterized by a function, but not the reverse problem
- t is not h(**x**) + (zero-mean**, unimodal**) noise

So Least squares solution will bomb

- Solutions:
  - model joint pdf
  - build piecewise models
  - ….



Extras

(Recap) Linear Regression

- Studied extensively and have well-developed theory (variable selection methods, extensions for dealing with correlated data, evaluation of results,...)

  Model: $E(y \mid \mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ + i.i.d. Gaussian error term

- Evaluating the coefficients
  - consider standard errors
  - Consider coefficients if both x and y's were normalized
  - Doing significance test to see if each term should be dropped.
  - Many predictors? Better to do forward search…

Weight Update through Error Back-propagation*

Takeaway: weights are updated through (online) SGD, using chain rule to "propagate" the error back from output towards the input nodes.

- Derived from chain rule for partial derivatives, applied to SGD
- Three stages:

1. Evaluate an "error signal" at the output units with net input a_k

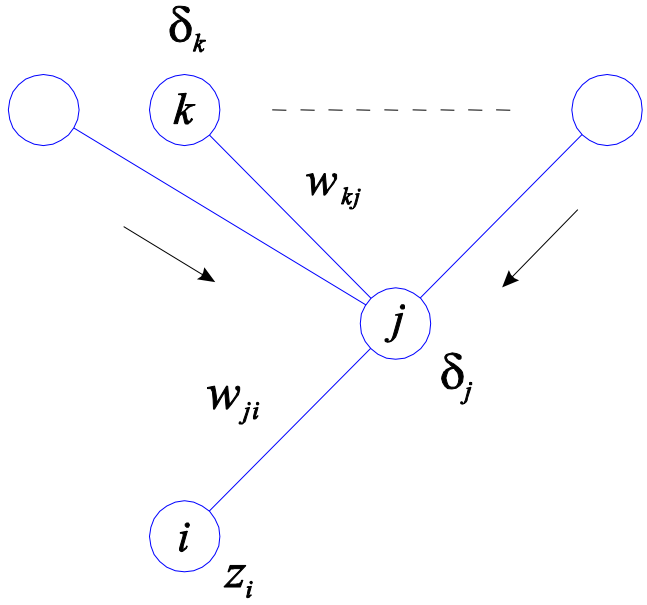
$$\delta_k = \frac{\partial E_n}{\partial a_k}$$

2. Propagate the signal backwards through the network

$$\delta_j = g'(a_j) \sum_k w_{kj} \delta_k$$

3. Evaluate derivatives

$$\frac{\partial E_n}{\partial w_{ji}} = \delta_j z_i$$



Sequential Learning with SGD

- Learning rate: too small → too slow
  - Too large → oscillatory, or may even diverge
- Should η be fixed or adaptive (second order methods)?
- Is convergence needed or not?
  - Non-stationary? May not want to converge!
  - If convergence is desired, then η should decrease with time.

  - * Robbins-Monroe Sequence is adequate (e.g. $\eta(t) = 1/t$)
    - Sum of absolute values (of sequence of η values) is unbounded
    - Sum of squared values is finite
    - Values in decreasing sequence