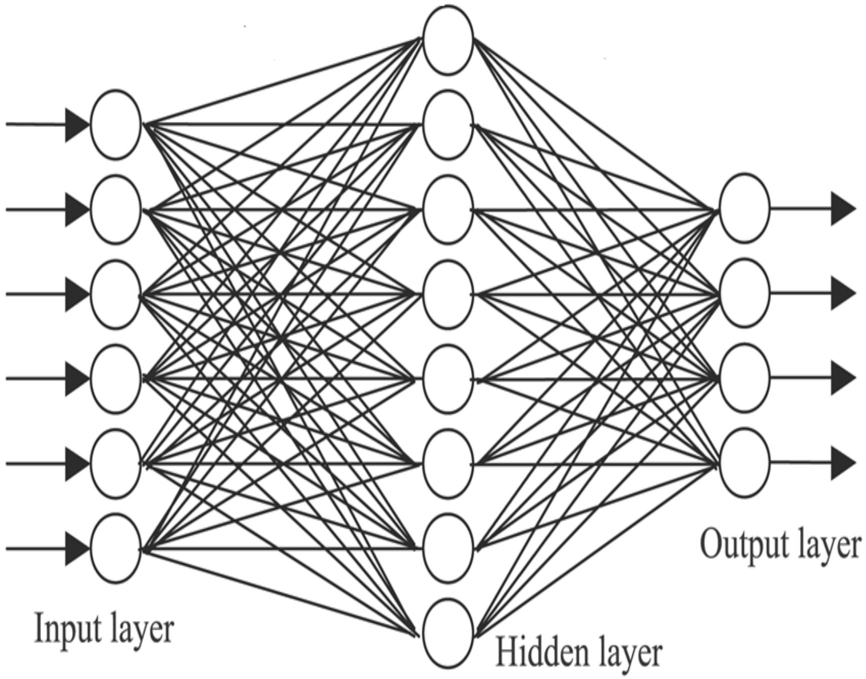
# Optimizing Deep Neural Networks

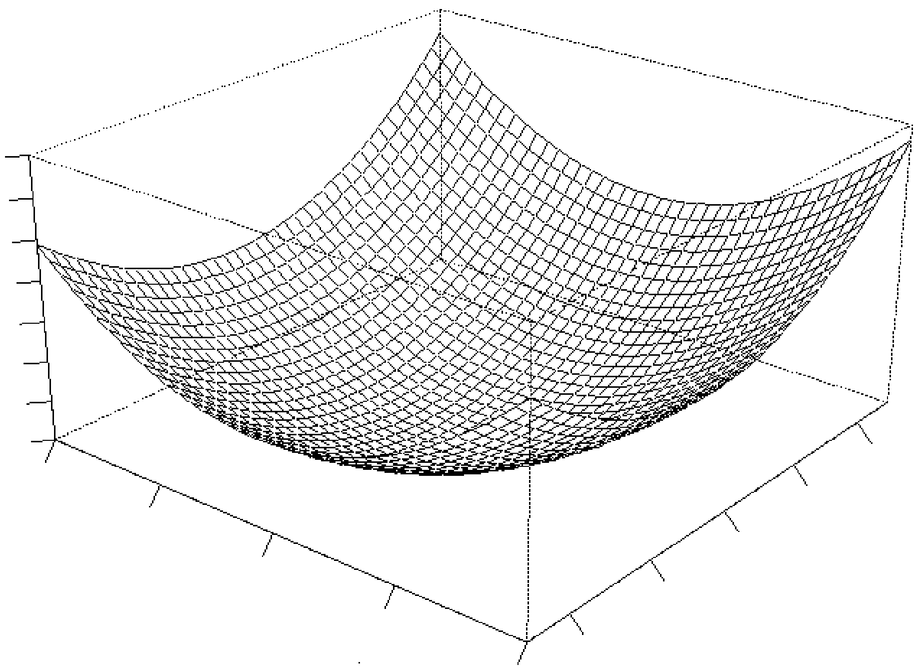
Leena Chennuru Vankadara

26-10-2015

# Table of Contents

# Neural Networks and Loss surfaces

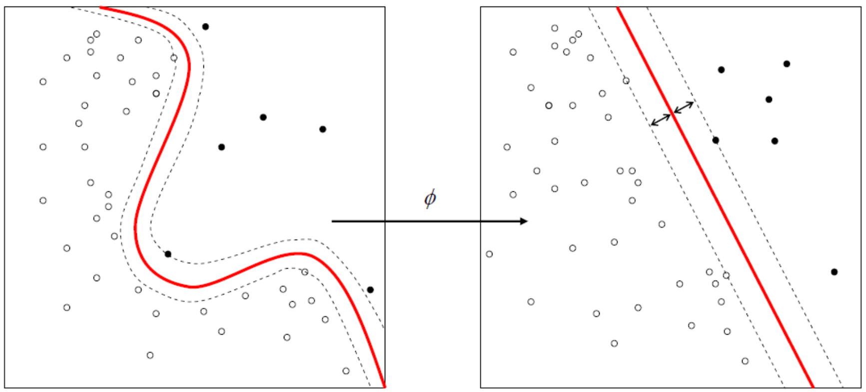


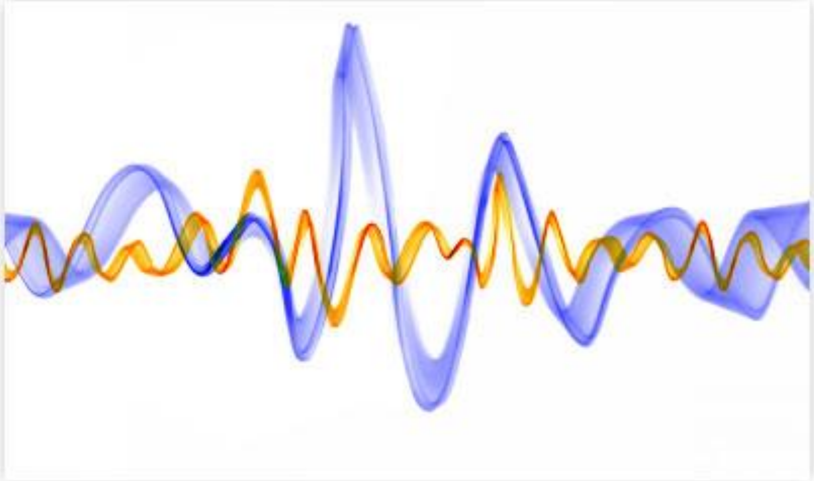
[1] Extremetech.com

[2] Willamette.edu

# Shallow architectures vs Deep architectures

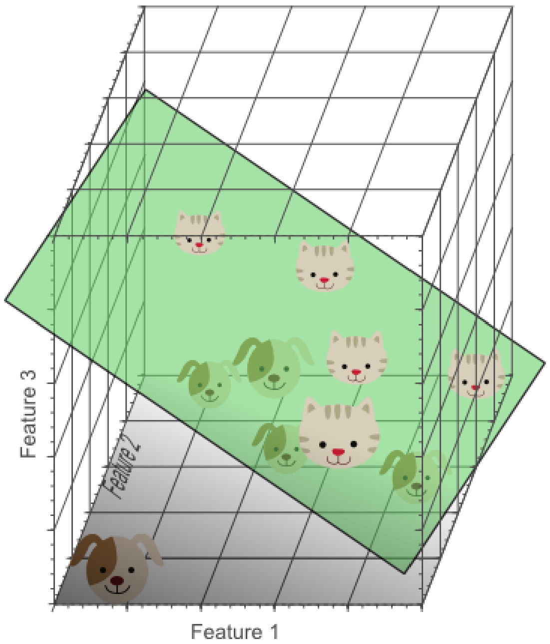


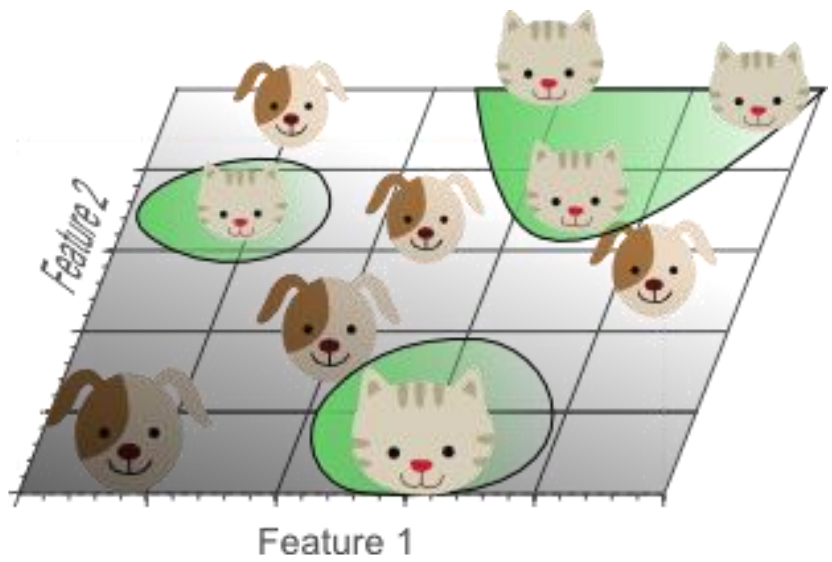

[3] Wikipedia.com



[4] Allaboutcircuits.com

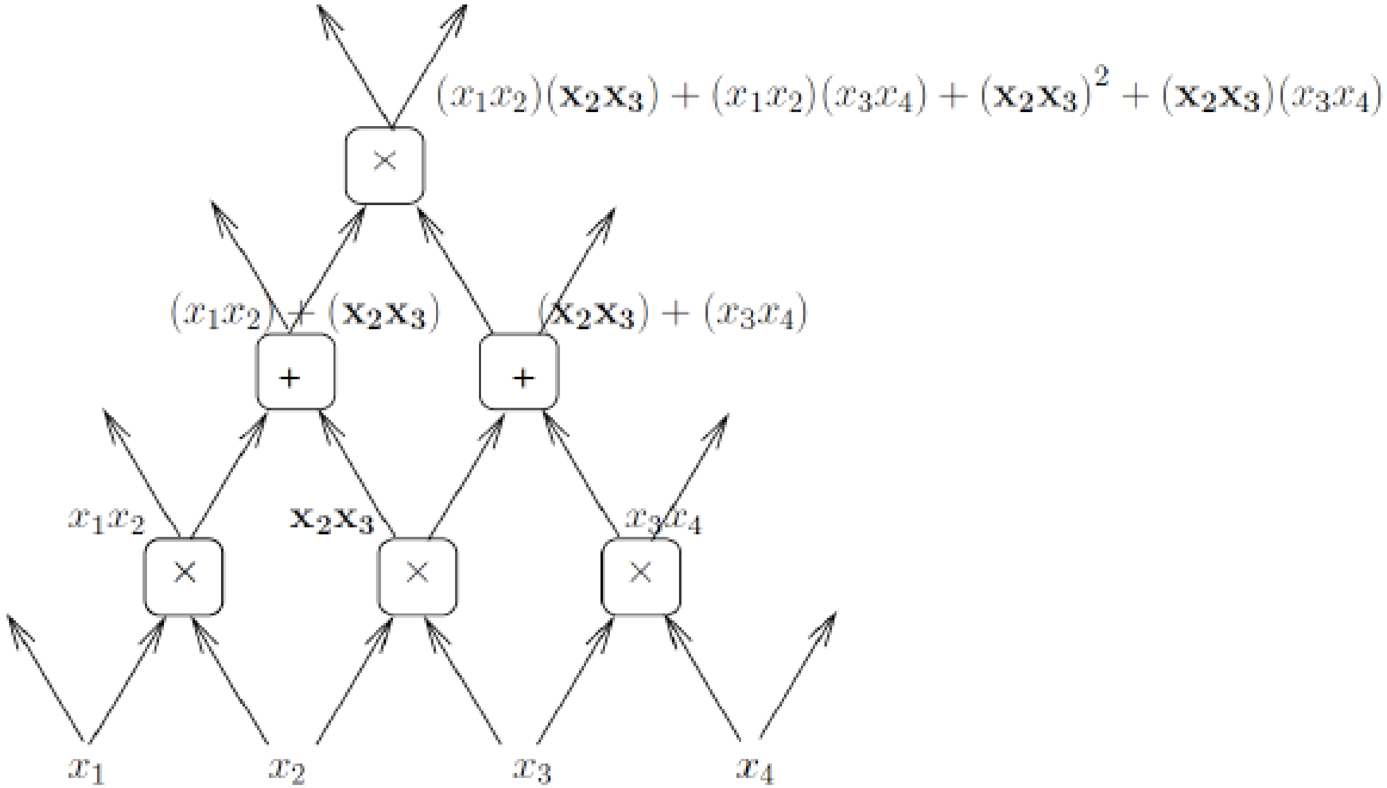
# Curse of dimensionality




[5] Visiondummy.com

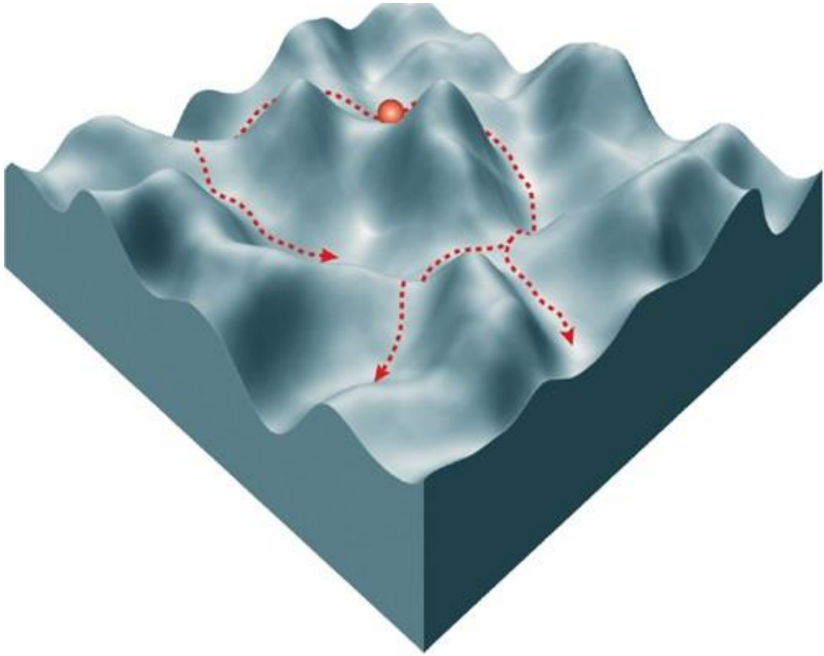
# Compositionality

- 


[6] Yoshua Bengio Deep Learning Summer School

# Problems of deep architectures

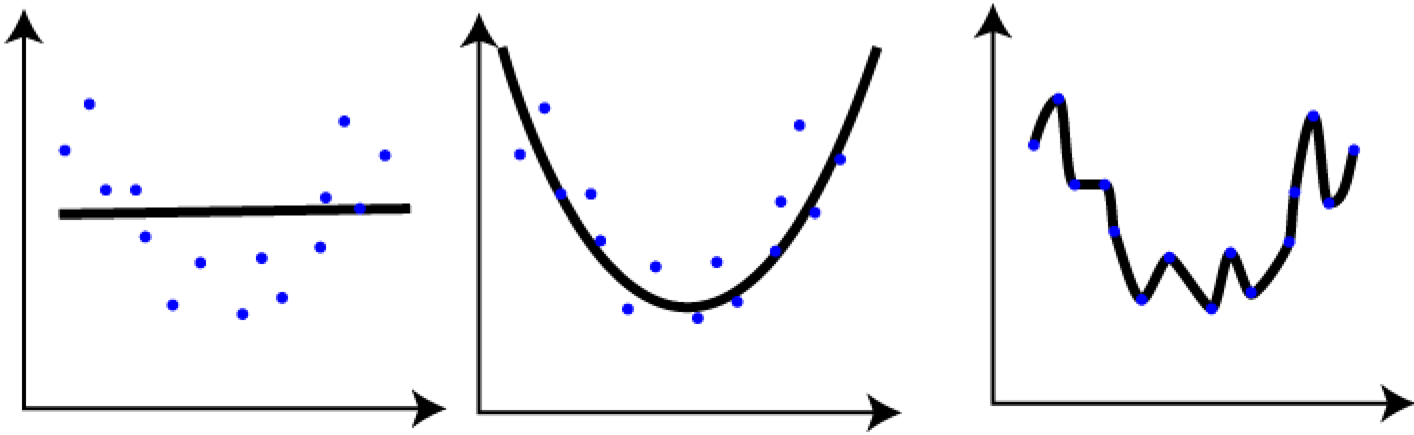
- ? Convergence to apparent local minima
- ? Saturating activation functions
- ? Overfitting
- ? Long training times
- ? Exploding gradients
- ? Vanishing gradients



[7] Nature.com

# Optimization in Neural networks(A broad perspective)

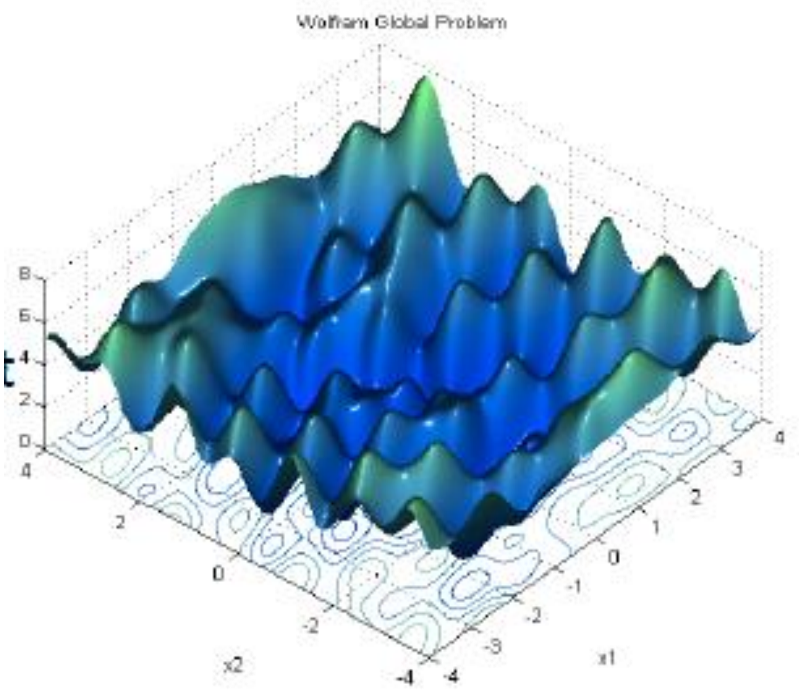
- Under fitting
- Training time
- Overfitting



[8] Shapeofdata.wordpress.com

# Proliferation of saddle points

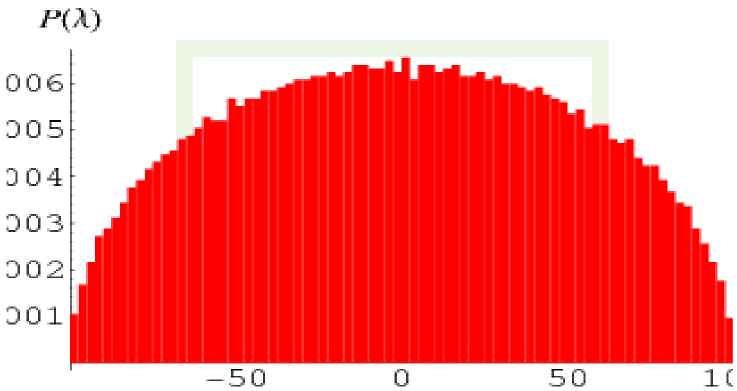
- Random Gaussian error functions.
- Analysis of critical points
- Unique global minima & maxima(Finite volume)
- Concentration of measure



# Proliferation of saddle points (Random Matrix Theory)

- Hessian at a critical point
  - Random Symmetric Matrix
- Eigenvalue distribution
  - A function of error/energy
- Proliferation of degenerate saddles
- Error(local minima) ≈ Error(global minima)

Wigner's Semicircular Distribution




[9] Mathworld.wolfram.com

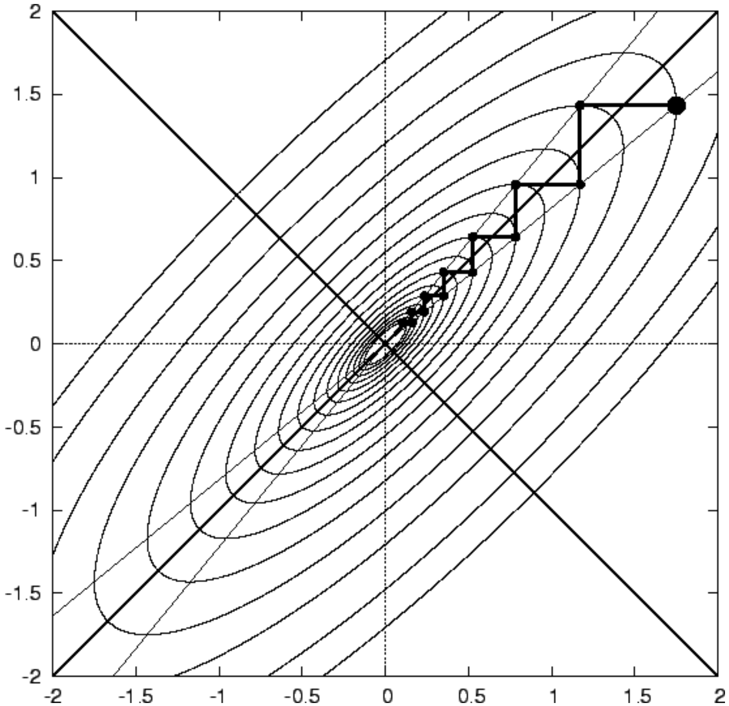
# Effect of dimensionality

- Single draw of a Gaussian process – unconstrained
  - Single valued Hessian
  - Saddle Point – Probability(0)
  - Maxima/Minima - Probability (1)
- Random function in N dimensions
  - Maxima/Minima – O(exp(-N))
  - Saddle points – O(exp(N))

# Analysis of Gradient Descent

- $$\Theta_{k+1} = \Theta_k - \alpha_k \nabla f_k$$
- Saddle points and pathological curvatures
- (Recall) High number of degenerate saddle points

+ Direction ? Step size
+ Solution1: Line search
  - Computational expense
+ Solution2: Momentum




[10] gist.github.com

# Analysis of momentum

- Idea: Add momentum in persistent directions
- Formally

$$\nu_{k+1} = \mu \nu_k - \varepsilon \nabla f(\Theta_k)$$

$$\Theta_{k+1} = \Theta_k + \nu_{k+1}$$

\+ Pathological curvatures.

? Choosing an appropriate momentum coefficient.

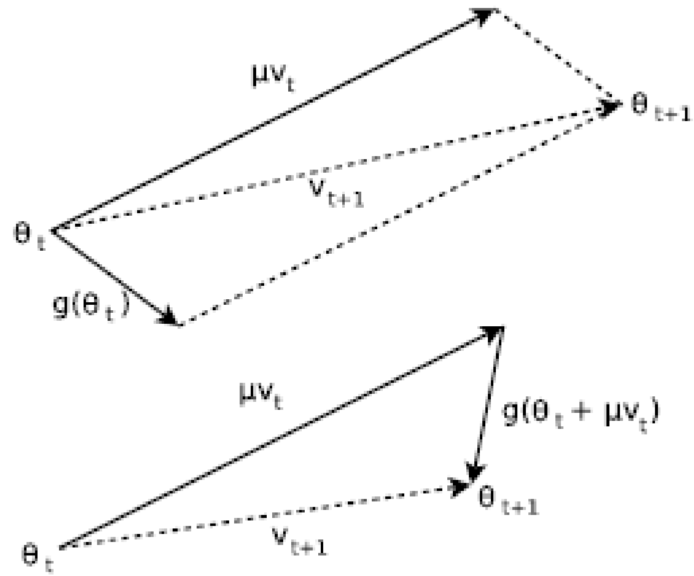
# Analysis of Nestrov's Accelerated Gradient(NAG)

- Formally

$$\nu_{k+1} = \mu\nu_k - \varepsilon\nabla f(\Theta_k + \mu\nu_k)$$

$$\Theta_{k+1} = \Theta_k + \nu_{k+1}$$

- Immediate correction of undesirable updates
- NAG vs Momentum
  - \+ Stability
  - \+ Convergence
  - = Qualitative behaviour around saddle points




[11] Sutskever, Martens, Dahl, Hinton On the importance of initialization and momentum in deep learning, [ICML 2013]

# Hessian based Optimization techniques

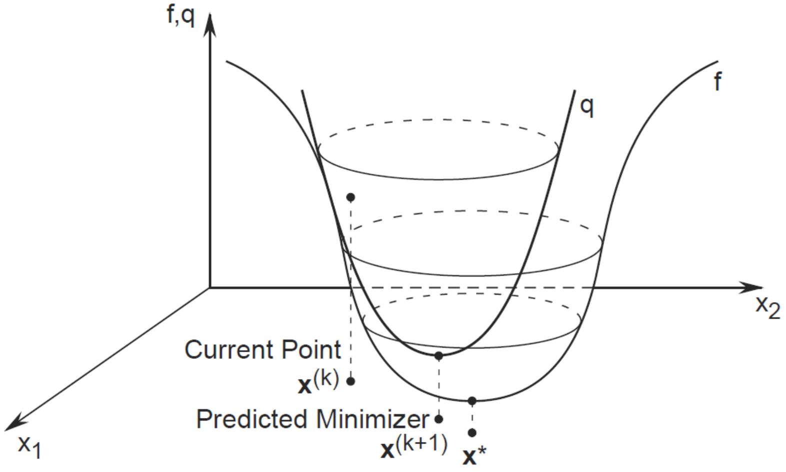
- Exploiting local curvature information
- Newton Method
- Trust Region methods
- Damping methods
- Fisher information criterion

# Analysis of Newton's method

- Local quadratic approximation
- Idea: Rescale the gradients by eigenvalues

+ Solves the slowness problem

- Problem: Negative curvatures
- Saddle points become attractors




[12] netlab.unist.ac.kr

# Analysis of Conjugate gradients

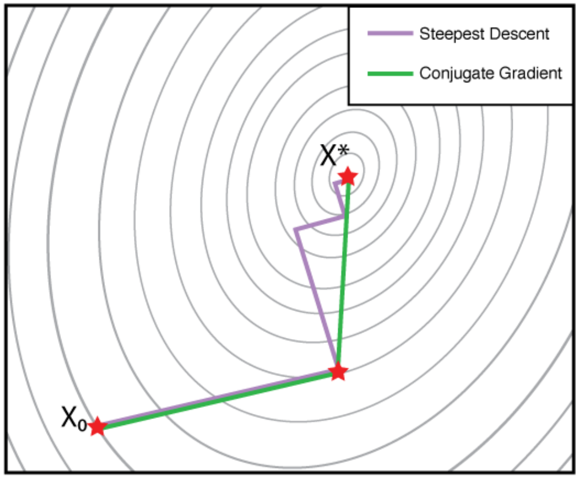
- Idea: Choose n ‘A’ – orthogonal search directions
  - Exact step size to reach the local minima
  - Step size rescaling by corresponding curvatures
  - Convergence in exactly n steps

+ Very effective with the slowness problem

? Problem: Computationally expensive

  - Saddle point structures

! Solution: Appropriate preconditioning




[13] Visiblegeology.com

# Analysis of Hessian Free Optimization

- Idea: Compute Hd through finite differences
  + Avoids computing the Hessian
- Utilizes the conjugate gradients method
- Uses Gauss Newton approximation(G) to Hessian
  + Gauss Newton method is P.S.D

+ Effective in dealing with saddle point structures

? Problem: Dampening to make the Hessian P.S.D

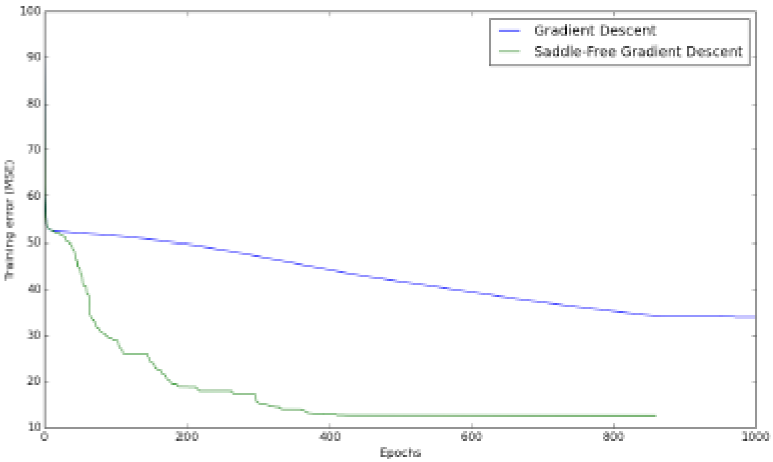
  - Anisotropic scaling $\longrightarrow$ slower convergence

# Saddle Free Optimization

- Idea: Rescale the gradients by the absolute value of eigenvalues

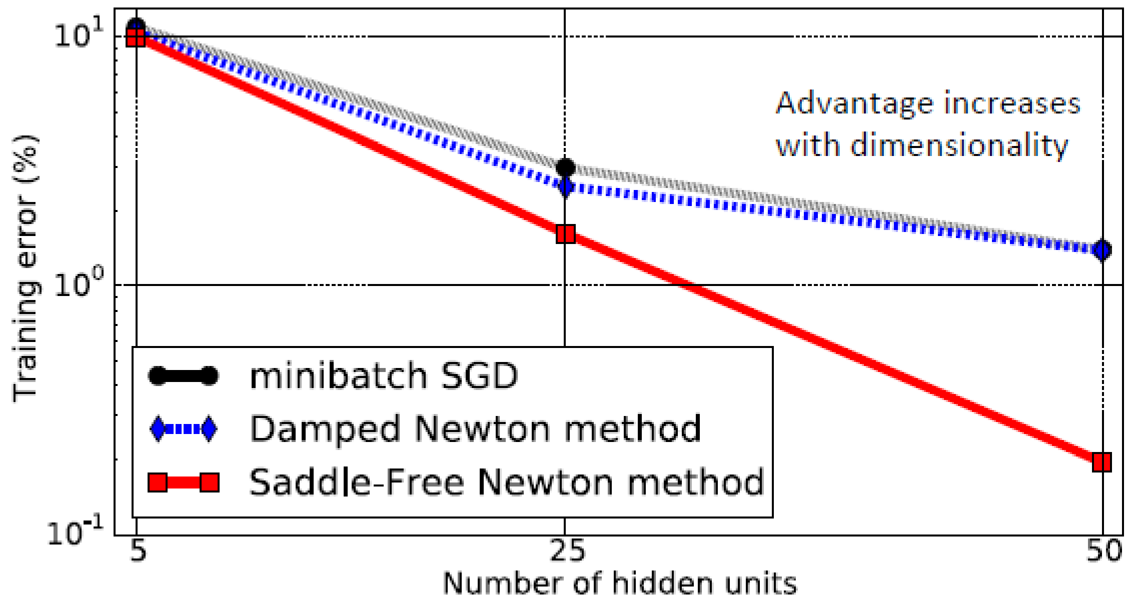
? Problem: Could change the objective!

! Solution: Justification by generalized trust region methods.




[14] Dauphin, Bengio Identifying and attacking the saddle point problem in high dimensional non-convex optimization arXiv 2014

# Advantage of saddle free method with dimensionality




[14] Dauphin, Bengio Identifying and attacking the saddle point problem in high dimensional non-convex optimization arXiv 2014

# Overfitting and Training time

- Dynamics of gradient descent
- Problem of inductive inference
- Importance of initialization
- Depth independent Learning times
- Dynamical isometry
- Unsupervised pre training

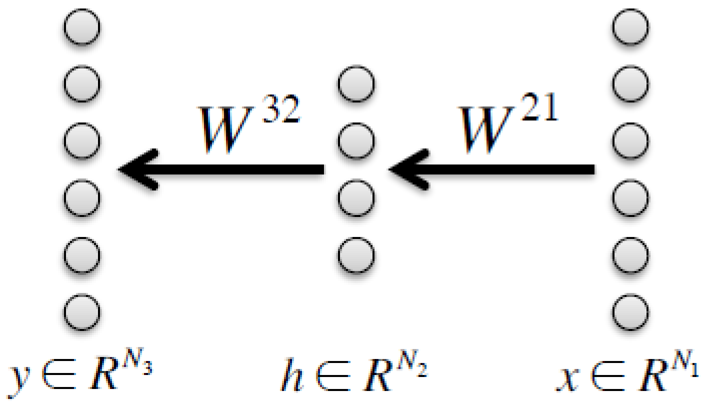
# Dynamics of Gradient Descent

- Squared loss – $\sum_{i=1}^{P} \|y^\mu - W^{32}W^{21}x^\mu\|^2$

- Gradient descent dynamics –

$$\nabla W^{21} = \lambda \sum_{i=1}^{P} W^{32^T}(y^\mu x^{\mu T} - W^{32}W^{21}x^\mu x^{\mu T})$$

$$\nabla W^{32} = \lambda \sum_{i=1}^{P} (y^\mu x^{\mu T} - W^{32}W^{21}x^\mu x^{\mu T})W^{21^T}$$



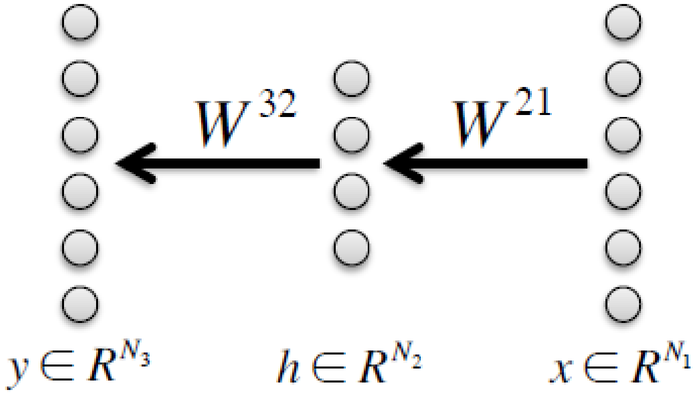

[15] Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. Andrew Saxe

# Learning Dynamics of Gradient Descent

- Input correlation to Identity matrix
- As t $\longrightarrow \infty$, weights approach the input output correlation.
- SVD of the input output map.

$$\Sigma^{31} = U^{33} S^{31} V^{11^T} = \sum_{\alpha=1}^{N_1} s_\alpha u_\alpha \nu_\alpha{}^T$$

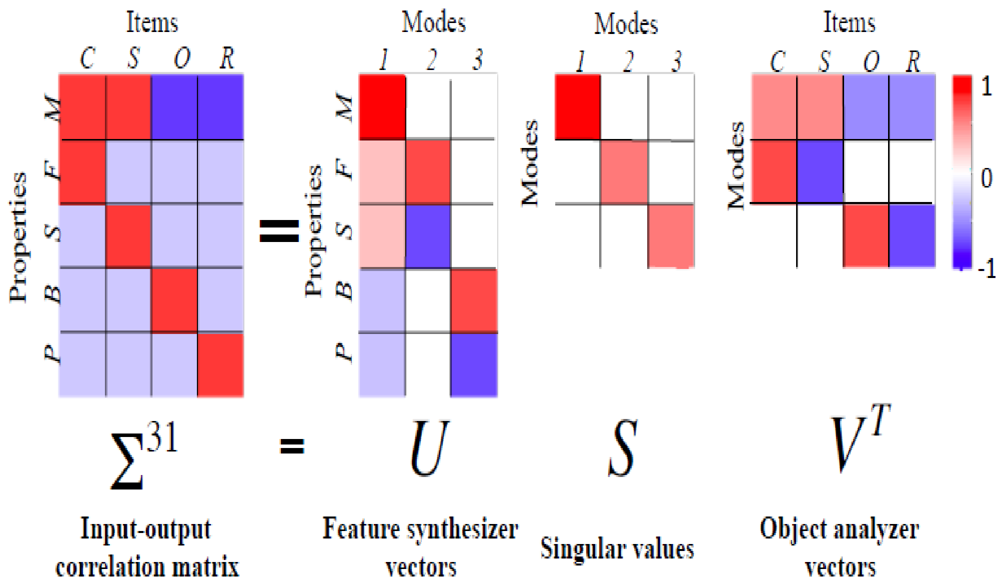
- What dynamics go along the way?




[15] Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. Andrew Saxe

# Understanding the SVD

- Canary, Salmon, Oak, Rose
- Three dimensions identified : plant -animal dimension, fish-birds, flowers-trees.
- S – Association strength
- U – Features of each dimension
- V – Item's place on each dimension.




[16] A.M. Saxe, J.L. McClelland, and S. Ganguli. Learning hierarchical category structure in deep neural networks. In Proceedings of the 35th Annual Conference of the Cognitive Science Society, 2013.

# Results

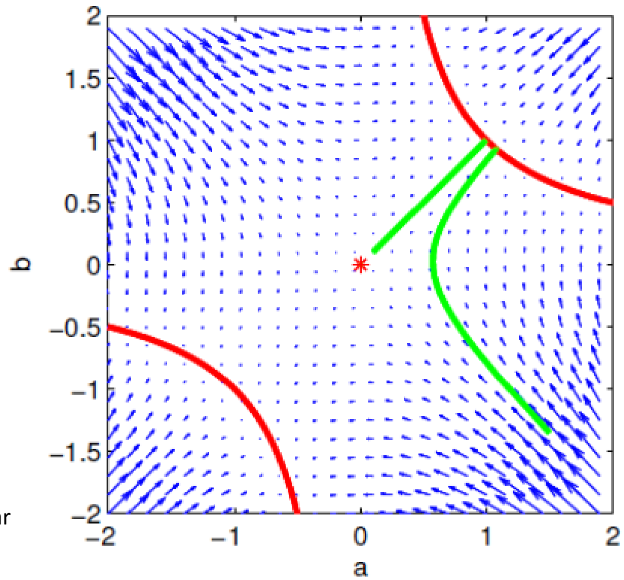
- Co-operative and competitive interactions across connectivity modes.
- Network driven to a decoupled regime
- Fixed points - saddle points
  - No non-global minima
- Orthogonal initialization of weights of each connectivity mode

$$W^{32} = U^{33} D_a R^T, W^{21} = R D_b {V^{11}}^T$$

  - R - an arbitrary orthogonal matrix
  - Eliminates the competition across modes

# Hyperbolic trajectories

- Symmetry under scaling transformations
- Noether's theorem → Conserved quantity
- Hyperbolic trajectories
- Convergence to a fixed point manifold
- Each mode learned in time O(t/s)
- Depth independent learning rates.
- Extension to non linear networks
- Just beyond the edge of orthogonal chaos

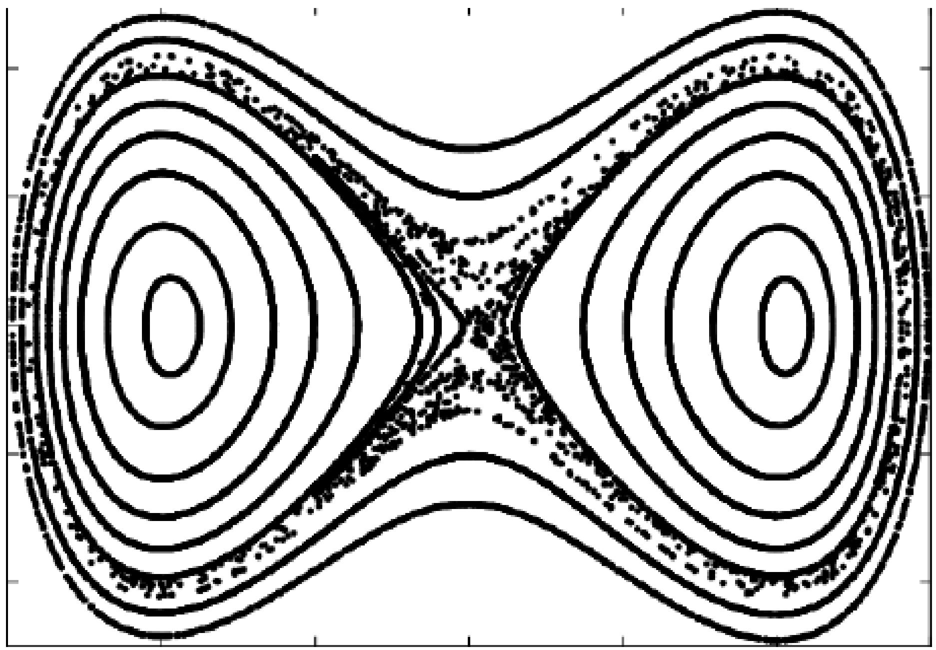



[15] Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. Andrew Saxe

# Importance of initialization

- Dynamics of deeper multi layer neural networks.
- Orthogonal initialization.
- Independence across modes.
- Existence of an invariant manifold in the weight space.
- Depth independent learning times.
- Normalized initialization
  - Can not achieve depth independent training times.
  - Anisometric projection onto different eigenvector directions
  - Slow convergence rates in some directions

# Importance of Initialization



[17] inspirehep.net

# Unsupervised pre-training

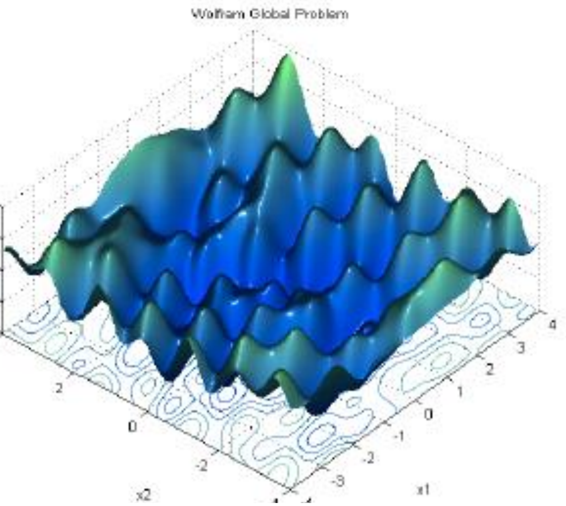
- No free lunch theorem
- Inductive bias
- Good basin of attraction
- Depth independent convergence rates.
- Initialization of weights in a near orthogonal regime
- Random orthogonal initializations
- Dynamical isometry with as many singular values of the Jacobian as possible at O(1)

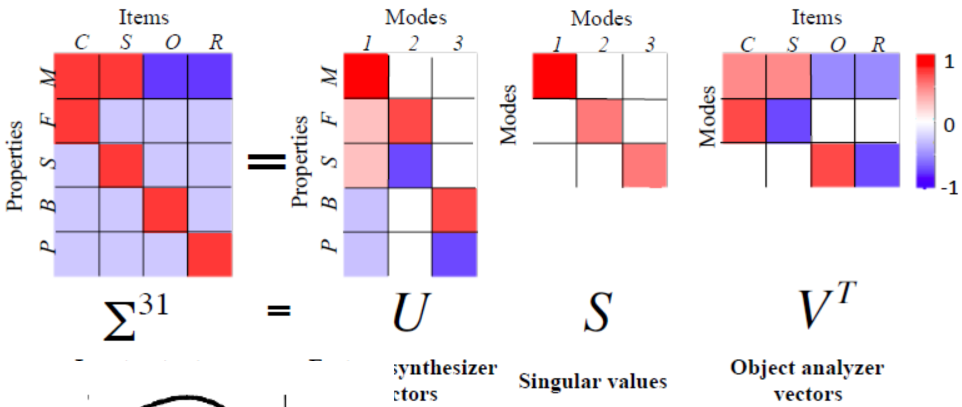
# Unsupervised learning as an inductive bias

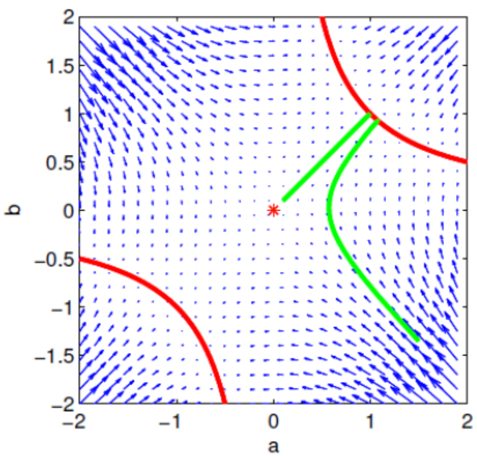
- Good regularizer to avoid overfitting
- Requirement:
  - Modes of variation in the input = Modes of variation in the input – output map.
- Saddle point symmetries in high dimensional spaces
- Symmetry breaking around saddle point structures
- Good basin of attraction of a good quality local minima.

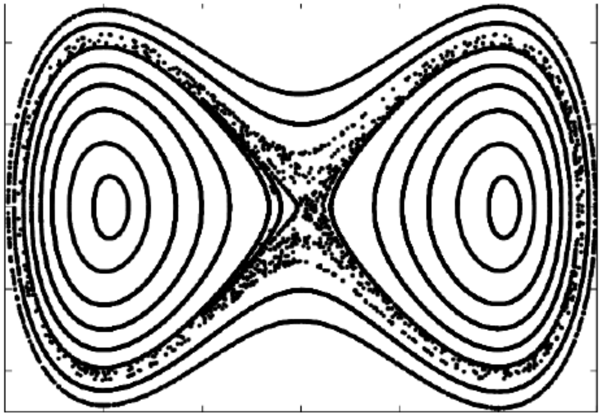
# Conclusion

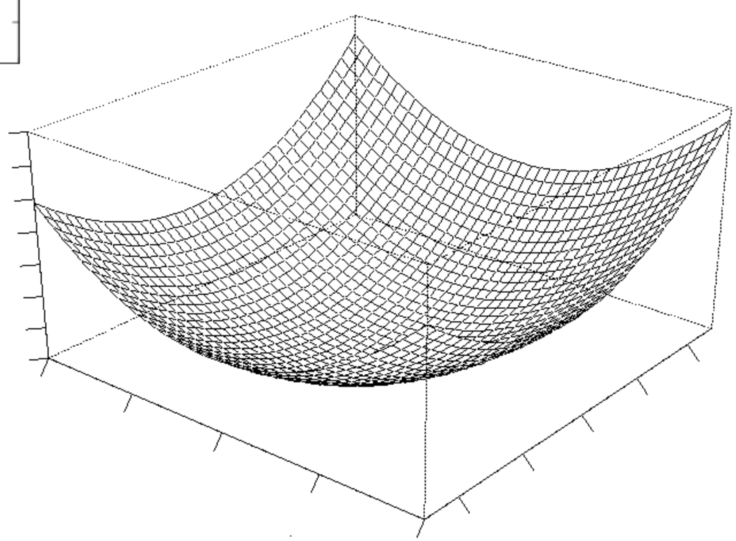
- Good momentum techniques such as Nestrov's accelerated gradient.
- Saddle Free optimization.
- Near orthogonal initialization of the weights of connectivity modes.
- Depth independent training times.
- Good initialization to find the good basin of attraction.
- Identify what good quality local minima are.


Wolfram Global Problem
x2
x1


Items
C S O R
Properties
M F S B P
Modes
1 2 3
1
0
-1
$\Sigma^{31}$ = $U$ $S$ $V^T$
Singular values
Object analyzer vectors

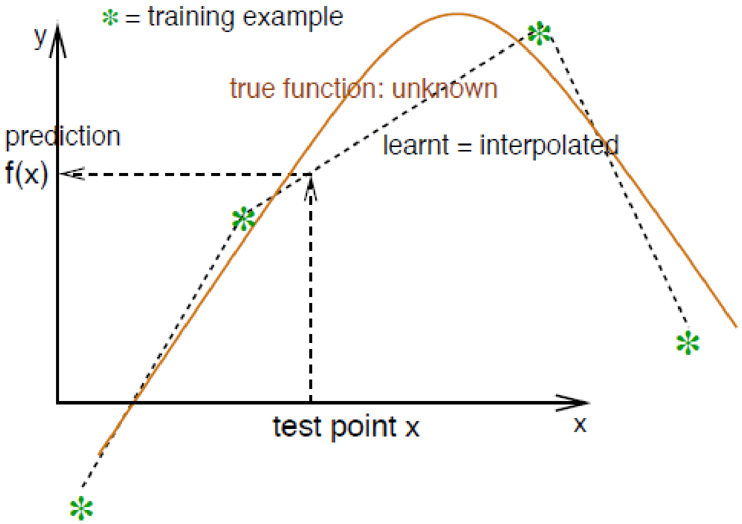

a
b

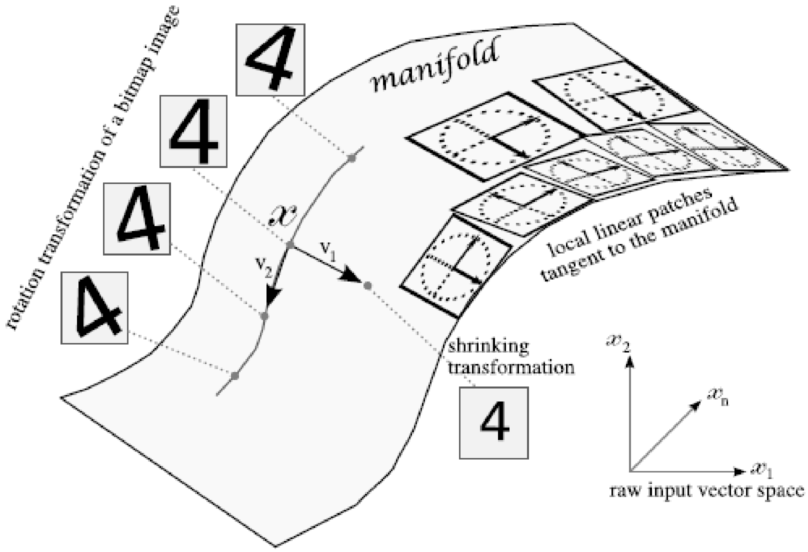




# Backup Slides

# Local Smoothness Prior vs curved submanifolds

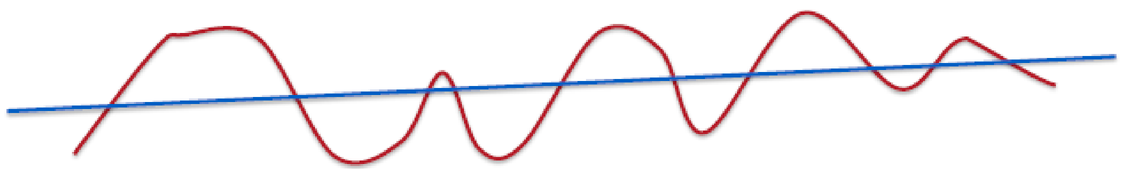



[18] Yoshua Bengio, Deep learning Summer school

# Number of variations vs dimensionality

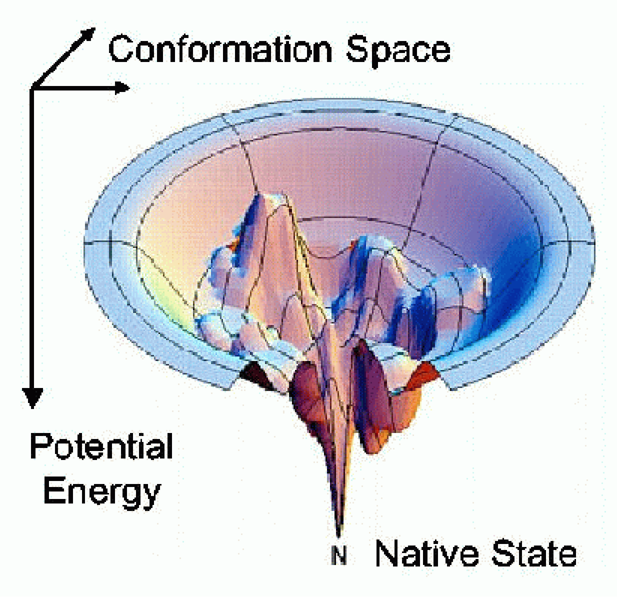
- Theorem: Gaussian kernel machines need at least k examples to learn a function that has 2k zero crossings along some line.*(Bengio, Dellalleau & Le Roux 2007)*
- Theorem: For a Gaussian kernel machine to learn some maximally varying functions over d inputs requires O(2^d) examples.



[18] Yoshua Bengio, Deep learning Summer school

# Theory of deep learning

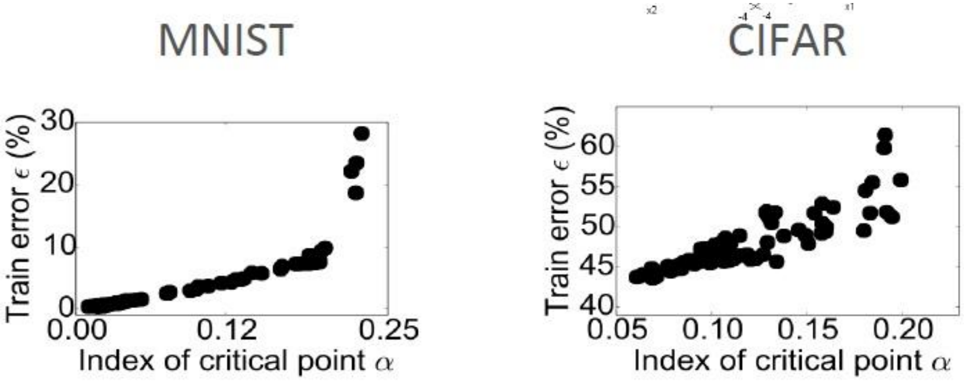
- Spin glass models
- String theory landscapes
- Protein folding
- Random Gaussian ensembles




[19] charlesmartin14.wordpress.com

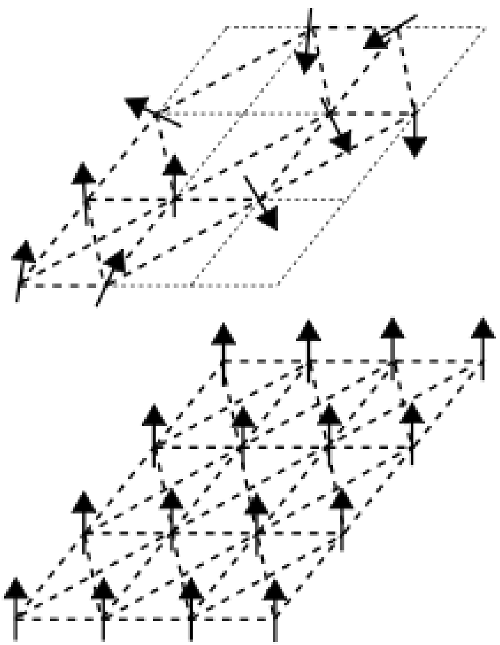
# Proliferation of saddle points(Cont'd…)

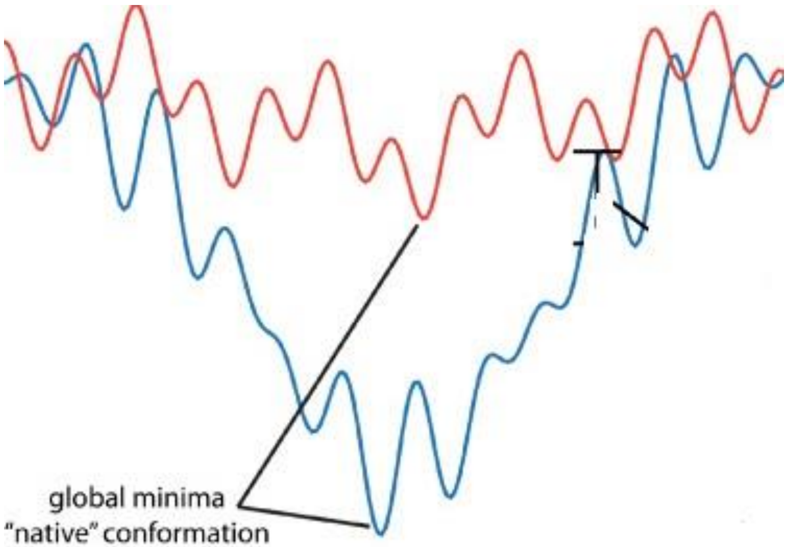
- Distribution of critical points as a function of index and energy.
  - Index – fraction/number of negative eigenvalues of the Hessian
- Error - Monotonically increasing function of index(0 to 1)
- Energy of local minima vs global minima
- Proliferation of saddle points




[20] Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, Yann Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, Yoshua Bengio

# Ising spin glass model and Neural networks



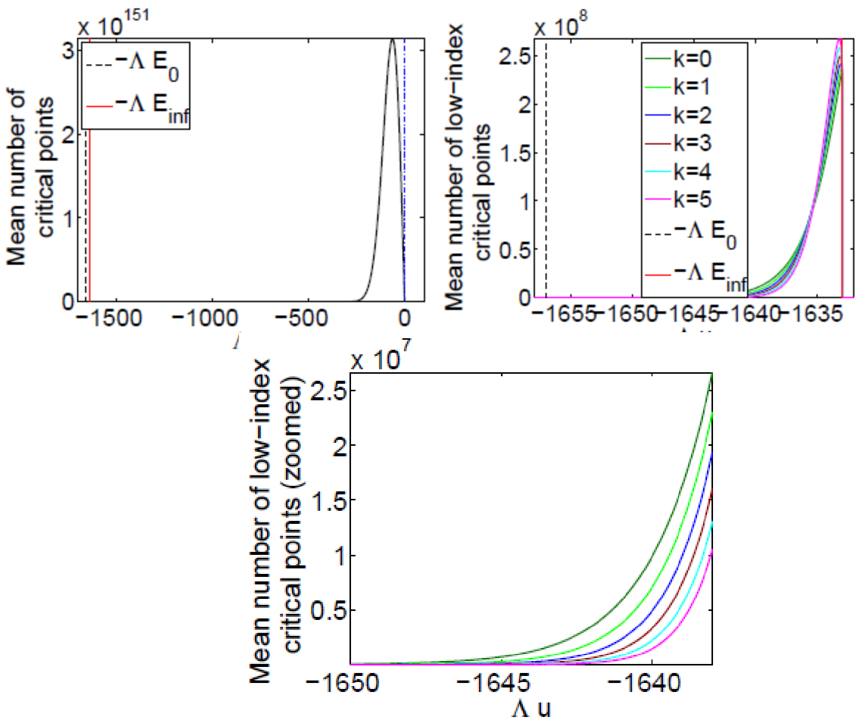


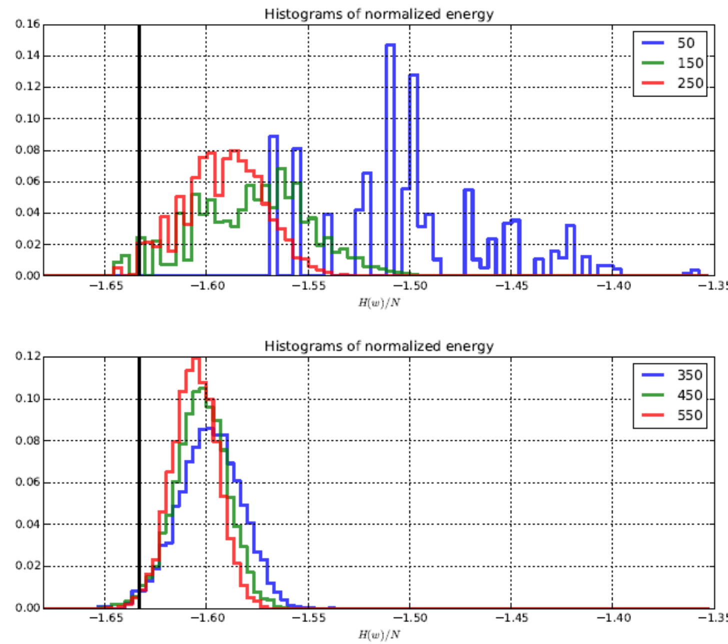

[19] charlesmartin14.wordpress.com

# Loss surfaces of multilayer neural networks(H layers)

- Equivalence to the Hamiltonian of the H-spin spherical spin glass model
  - Assumptions of Variable independence
  - Redundancy in network parametrization
  - Uniformity
- Existence of  a ground state
- Existence of an energy barrier (Floor)
- Layered structure of critical points in the energy band
- Exponential time to search for a global minima
- Experimental evidence for close energy values of ground state and Floor

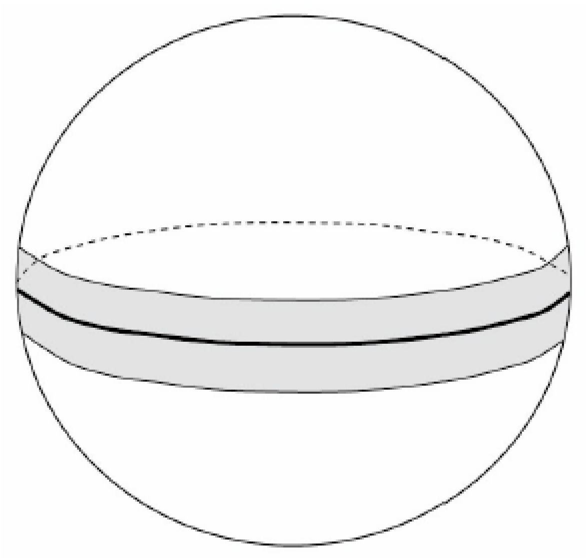
# Loss surfaces of multilayer neural networks




[20] Loss surfaces of Multilayer Neural Networks, Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, Yann LeCun

# Concentration of Measure



- Its very difficult for N independent random variables to work together and pull the sum or any function dependent on them very far away from its mean.
- Informally, A random variable that depends in a Lipschitz way on many independent random variables is essentially constant.

[21] High-dimensional distributions with convexity properties Bo’az Klartag Tel-Aviv University