Optimization of neural networks

Dmitry Yarotsky

Parametrized predictive models

True response function: $y = f(\mathbf{x})$, where $\mathbf{x}$ is the input vector

- $y \in \mathbb{R}$ for regression
- $y \in \{0, 1\}$ for binary classification

Predictive model: $y = \widetilde{f}(\mathbf{x}, \mathbf{W})$, and $\mathbf{W}$ are model parameters (e.g., network weights)

"Soft classification": $\widetilde{f}(\mathbf{x}, \mathbf{W}) \in [0, 1]$

Model training as a parametric optimization

Loss function: $L(\mathbf{W}) = \int l(f(\mathbf{x}), \widetilde{f}(\mathbf{x}, \mathbf{W}))d\mu(\mathbf{x})$

Sample average measure μ :

- $d\mu(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\delta(\mathbf{x} - \mathbf{x}_n)$ with Dirac's delta — "finite training set" scenario
- $d\mu(\mathbf{x}) = p(\mathbf{x})d\mathbf{x}$ with some (e.g. Gaussian) density $p(\mathbf{x})$ — "population average" scenario

Function $l(f(\mathbf{x}), \widetilde{f}(\mathbf{x}, \mathbf{W}))$ measures the discrepancy between f and $\widetilde{f}$, e.g.:

- Regression: $l(y, \widetilde{y}) = \frac{1}{2}(y - \widetilde{y})^2$
- Classification: $l(y, \widetilde{y}) = -y \log \widetilde{y} - (1 - y) \log(1 - \widetilde{y})$

Model training:

$$L(\mathbf{W}) \longrightarrow \min_{\mathbf{W}}$$

Gradient-based optimization

- $\mathbf{W}$ high-dimensional
- $L(\mathbf{W})$ non-smooth, non-nonconvex

Most popular approach: gradient-based optimization and its modifications

Basic gradient descent with learning rate $\alpha \in (0, 1)$:

$$\mathbf{W}^{(n+1)} = \mathbf{W}^{(n)} - \alpha \nabla_{\mathbf{W}} L(\mathbf{W}^{(n)})$$

Gradient descent with Nesterov momentum ($\alpha, \beta \in (0, 1)$):

$$\mathbf{W}^{(n+1)} = \mathbf{W}^{(n)} - \mathbf{V}^{(n)}$$
$$\mathbf{V}^{(n+1)} = \alpha \nabla_{\mathbf{W}} L(\mathbf{W}^{(n)}) + \beta \mathbf{V}^{(n)}$$

Exercise: how can we interprete the coefficients α and β?

Computation of $\nabla_{\mathbf{W}} L$: "Error backpropagation"

$$\nabla_{\mathbf{W}} L(\mathbf{W}) = \int \frac{\partial l}{\partial \widetilde{y}}(f(\mathbf{x}), \widetilde{f}(\mathbf{x}, \mathbf{W})) \cdot \nabla_{\mathbf{W}} \widetilde{f}(\mathbf{x}, \mathbf{W}) d\mu(\mathbf{x})$$

$$\nabla_{\mathbf{W}} \widetilde{f} = (\nabla_{\mathbf{w}_1} \widetilde{f}, \ldots, \nabla_{\mathbf{w}_K} \widetilde{f})$$

$\frac{\partial l}{\partial \widetilde{y}}(f(\mathbf{x}), \widetilde{f}(\mathbf{x}, \mathbf{W}))$: directly computed from y and $\widetilde{y}$

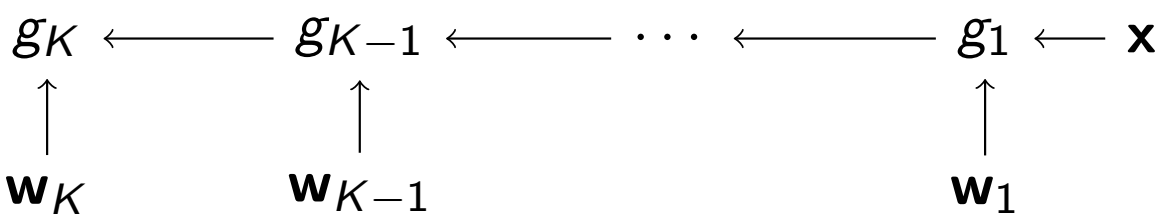
$\widetilde{y} = \widetilde{f}(\mathbf{x}, \mathbf{W})$: "forward propagation"

To find $\nabla_{\mathbf{W}} \widetilde{f}$: use layerwise representation

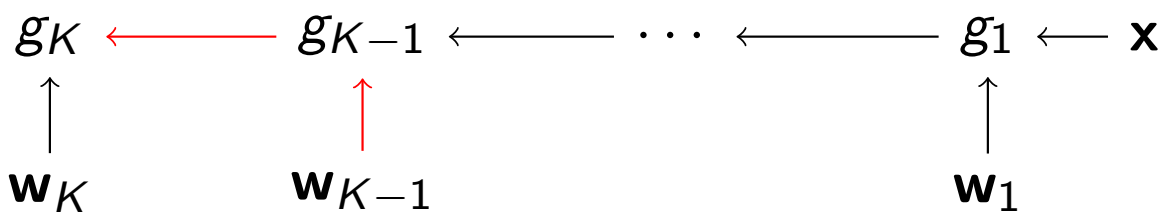
$$\widetilde{f}(\mathbf{x}, \mathbf{W}) = g_K(g_{K-1}(\ldots g_1(\mathbf{x}, \mathbf{w}_1), \ldots \mathbf{w}_{K-1}), \mathbf{w}_K)$$

$\mathbf{z}_k$: output of the k'th layer (known from "forward propagation")

$$\mathbf{z}_k = g_k(\mathbf{z}_{k-1}, \mathbf{w}_k)$$



“Error backpropagation”



$$\mathbf{z}_k = g_k(\mathbf{z}_{k-1}, \mathbf{w}_k)$$

$$\nabla_{\mathbf{w}_K} \widetilde{f}(\mathbf{x}, \mathbf{W}) = \frac{\partial g_K}{\partial \mathbf{w}_K}(\mathbf{z}_{K-1}, \mathbf{w}_K)$$

$$\begin{aligned}
\nabla_{\mathbf{w}_{K-1}} \widetilde{f}(\mathbf{x}, \mathbf{W}) &= \nabla_{\mathbf{w}_K} g_K(g_{K-1}(\mathbf{z}_{K-2}, \mathbf{w}_{K-1}), \mathbf{w}_K) \\
&= \frac{\partial g_K}{\partial \mathbf{z}_{K-1}}(\mathbf{z}_{K-1}, \mathbf{w}_K) \cdot \frac{\partial g_{K-1}}{\partial \mathbf{w}_{K-1}}(\mathbf{z}_{K-2}, \mathbf{w}_{K-1})
\end{aligned}$$

$$\cdots$$

$$\begin{aligned}
\nabla_{\mathbf{w}_k} \widetilde{f}(\mathbf{x}, \mathbf{W}) = &\frac{\partial g_K}{\partial \mathbf{z}_{K-1}}(\mathbf{z}_{K-1}, \mathbf{w}_K) \cdots \frac{\partial g_{k+1}}{\partial \mathbf{z}_k}(\mathbf{z}_k, \mathbf{w}_{k+1}) \\
&\cdot \frac{\partial g_k}{\partial \mathbf{w}_k}(\mathbf{z}_{k-1}, \mathbf{w}_k)
\end{aligned}$$

Basic theory: convergence of gradient descent

See e.g. Yu. Nesterov, Introductory Lectures on Convex Programming Volume I: Basic course.

- **Global minima**: $L(\mathbf{W}_*) = \min_{\mathbf{W} \in \mathbb{R}^W} L(\mathbf{W})$
- **Local minima**: $L(\mathbf{W}_*) = \min_{\mathbf{W} \in U} L(\mathbf{W})$, where U is an open neighborhood of $\mathbf{W}_*$
- **Stationary points**: $\nabla_{\mathbf{W}} L(\mathbf{W}_*) = 0$ (assuming $L(\mathbf{W})$ is smooth)

In general, gradient descent converges to a stationary point.

Proposition

Suppose function L is lower bounded and $L \in \mathcal{W}^{2,\infty}(\mathbb{R}^W)$, so that $|\nabla L(\mathbf{a}) - \nabla L(\mathbf{b})| \leq M|\mathbf{a} - \mathbf{b}|$ with some Lipschitz constant M. Let $\alpha < \frac{2}{M}$. Then $\nabla L(\mathbf{W}^{(n)}) \to 0$, and $\min_{n=1,\ldots,N} |\nabla L(\mathbf{W}^{(n)})| = O(N^{-1/2})$.

Proof

$$
\begin{aligned}
L(\mathbf{W}^{(n+1)}) &= L(\mathbf{W}^{(n)}) + \left\langle \mathbf{W}^{(n+1)} - \mathbf{W}^{(n)}, \int_0^1 \nabla L(\mathbf{W}^{(n)} + t(\mathbf{W}^{(n+1)} - \mathbf{W}^{(n)}))dt \right\rangle \\
&\leq L(\mathbf{W}^{(n)}) + \left\langle \mathbf{W}^{(n+1)} - \mathbf{W}^{(n)}, \nabla L(\mathbf{W}^{(n)}) \right\rangle \\
&\quad + |\mathbf{W}^{(n+1)} - \mathbf{W}^{(n)}| \int_0^1 Mt|\mathbf{W}^{(n+1)} - \mathbf{W}^{(n)}|dt \\
&\leq L(\mathbf{W}^{(n)}) + \left\langle \mathbf{W}^{(n+1)} - \mathbf{W}^{(n)}, \nabla L(\mathbf{W}^{(n)}) \right\rangle + \frac{M}{2}|\mathbf{W}^{(n+1)} - \mathbf{W}^{(n)}|^2 \\
&\leq L(\mathbf{W}^{(n)}) + (-\alpha + \frac{M}{2}\alpha^2)|\nabla L(\mathbf{W}^{(n)})|^2 \\
&< L(\mathbf{W}^{(n)})
\end{aligned}
$$

if $\alpha < \frac{2}{M}$. Let $c = \alpha(1 - \frac{M}{2}\alpha) > 0$, then

$$
\sum_{n=1}^{N} |\nabla L(\mathbf{W}^{(n)})|^2 \leq \frac{1}{c} \sum_{n=1}^{N} (L(\mathbf{W}^{(n)}) - L(\mathbf{W}^{(n+1)})) \leq \frac{1}{c}(L(\mathbf{W}^{(1)}) - \min_{\mathbf{W}} L(\mathbf{W})),
$$

$$
\min_{n=1,\ldots,N} |\nabla L(\mathbf{W}^{(n)})| \leq \frac{c^{-1/2}}{\sqrt{N}} \left(L(\mathbf{W}^{(1)}) - \min_{\mathbf{W}} L(\mathbf{W})\right)^{1/2}
$$

Formulation in terms of stopping condition

Assume the **stopping condition**: $|\nabla L(\mathbf{W}^{(n)})| < \epsilon$.

Then, optimization terminates in $O\Big(\frac{L(\mathbf{W}^{(1)}) - \min_{\mathbf{W}} L(\mathbf{W})}{\epsilon^2}\Big)$ steps.

Exercise: What is the optimal value of α, assuming M is known?

Exercise: Give an example of gradient descent converging to a stationary point which is not a (local or global) minimum.

Linearization and spectral analysis

Suppose that $L \in C^2(\mathbb{R})$ and $\mathbf{W}_*$ is a stationary point.
For $\mathbf{W}$ near $\mathbf{W}_*$:

$$\nabla L(\mathbf{W}) = D^2 L(\mathbf{W}_*) \cdot (\mathbf{W} - \mathbf{W}_*) + o(|\mathbf{W} - \mathbf{W}_*|),$$

where $D^2 L(\mathbf{W}_*)$ is the Hessian matrix. Optimization iterates:

$$\begin{aligned}
\mathbf{W}^{(n+1)} - \mathbf{W}_* &= \mathbf{W}^{(n)} - \mathbf{W}_* - \alpha \nabla L(\mathbf{W}^{(n)}) \\
&= \mathbf{W}^{(n)} - \mathbf{W}_* - \alpha D^2 L(\mathbf{W}_*) \cdot (\mathbf{W}^{(n)} - \mathbf{W}_*) + o(|\mathbf{W}^{(n)} - \mathbf{W}_*|) \\
&= (1 - \alpha D^2 L(\mathbf{W}_*)) \cdot (\mathbf{W}^{(n)} - \mathbf{W}_*) + o(|\mathbf{W}^{(n)} - \mathbf{W}_*|)
\end{aligned}$$

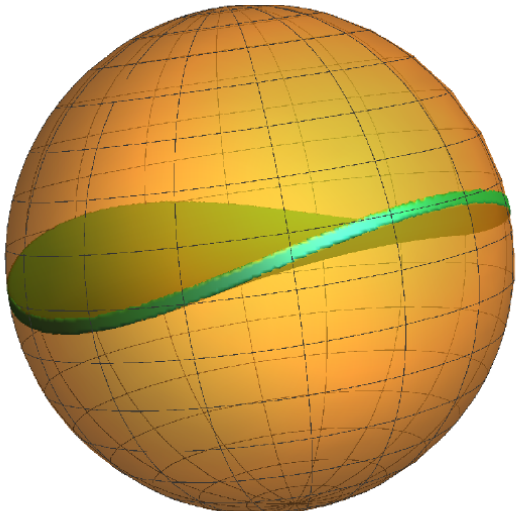
Convergence is determined by eigenvalues of $D^2 L(\mathbf{W}_*)$:

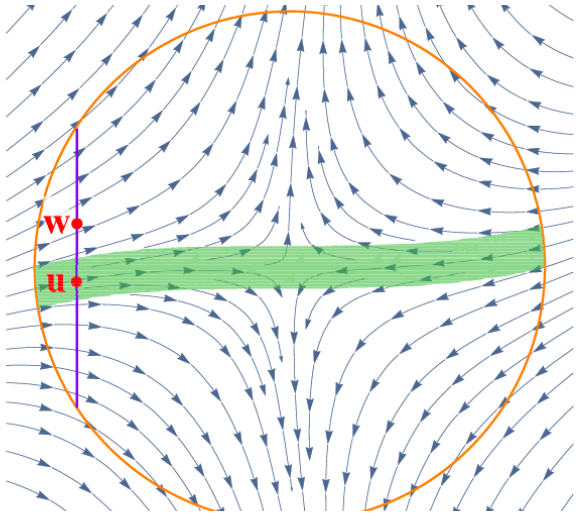
- positive: convergence
- negative: divergence

Evasion of saddle points

Saddle points: $D^2 L(\mathbf{W}_*)$ has both positive and negative eigenvalues

Typically, saddles are evaded by optimization, due to the presence of diverging components in $\mathbf{W}^{(n)} - \mathbf{W}_*$. The manifold of converging $\mathbf{W}^{(n)}$ has Lebesgue measure 0.[1]



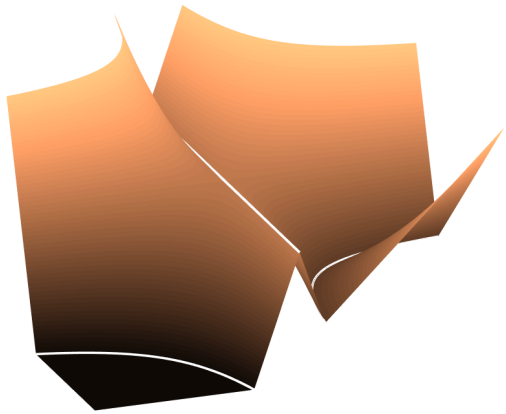


Chi Jin and M. Jordan, How to Escape Saddle Points Efficiently: Saddle points can slow down optimization; perturbing the GD can help.

[1] B. Recht, Saddles Again

Real-life ANNs

- No smoothness, in general (e.g. with ReLU): local minima of $L(\mathbf{W})$ are non-differentiable



Th. Laurent, J. von Brecht, The Multilinear Structure of ReLU Networks, arXiv:1712.10132

- Large size of the network and its structure are important

Empirical observations of real-life ANNs

From A. Choromanska et al., The Loss Surfaces of Multilayer Networks, arXiv:1412.0233:

- Large networks train well despite their size. Optimization can terminate at different local minima, but they seem to be equivalent and yield similar performance on a test set.
- The probability of finding a bad (high value) local minimum is non-zero for small-size networks and decreases quickly with network size.
- Struggling to find the global minimum on the training set (as opposed to one of the many good local ones) is not useful in practice and may lead to overfitting.

Conceptual pictures of the loss surface (conjectured)

From M. Baity-Jesi et al., Comparing Dynamics: Deep Neural Networks versus Glassy Systems: two alternatives

1. The loss landscape is very rough, has many isolated local minima, but GD tends to find good minima having low loss.
2. The loss function is highly nonlinear, but has few local minima, and the minima are connected. (Example: $L(w_1, w_2) = (w_2 - w_1^2)^2 + \epsilon w_1^2$.)

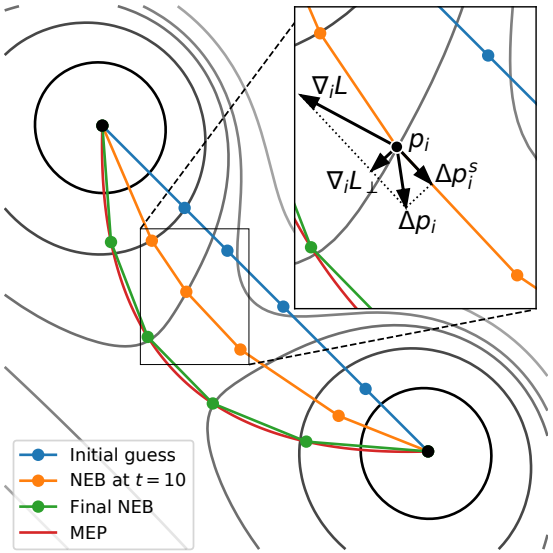
(Note: these conteptual pictures have only a limited value due to the "curse of dimensionality" in $\mathbb{R}^W$, lack of characterization of locality and depth of a local minimum, etc.)

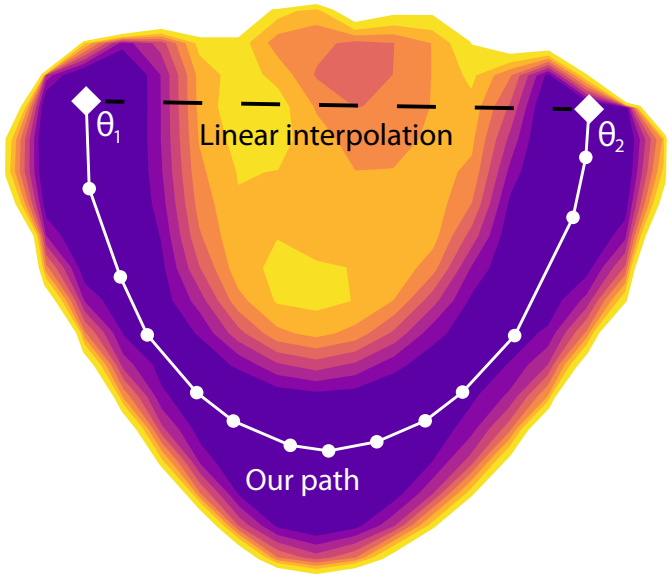
Some current research directions

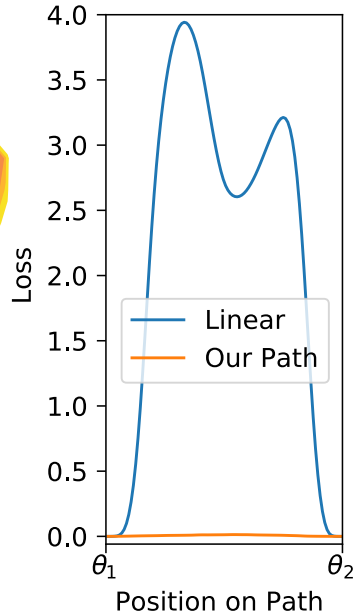
- Numerical studies of loss surface and gradient descent
- Direct analytic studies of simple (toy) scenarios:
  - Deep linear networks (no nonlinear activation)
  - Wide shallow networks with small training sets, pyramidal networks (no spurious local minima)
- Large-size limits:
  - Large-width limit: Gaussian approximation for signal propagation, connections to random matrices and spherical spin glasses
  - Phenomenological: Stochastic PDE and Langevin dynamics
- Specialized networks (e.g., convnets)

F. Draxler et al., Essentially No Barriers in Neural Network Energy Landscape, arXiv:1803.00885

- Two local minima are connected by a path, and then it is deformed to find a low loss trajectory
- ResNets and DenseNets on CIFAR10 and CIFAR100
- The optimized path: approximately constant loss







(Global description of the optimal set?)

I. Safran, O. Shamir, Spurious Local Minima are Common in Two-Layer ReLU Neural Networks, arXiv:1712.08968

Spurious local minimum W_0: $\min_{\mathbf{W}} L(\mathbf{W}) < L(\mathbf{W}_0) < L(\mathbf{W}')$ for $\mathbf{W}'$ in a small neighborhood of $\mathbf{W}_0$

Theorem

Consider the optimization problem

$$\min_{\mathbf{w}_1,\ldots,\mathbf{w}_n \in \mathbb{R}^k} \mathbb{E}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},I)} \Big(\sum_{i=1}^{n} (\mathbf{w}_i^\top \mathbf{x})_+ - \sum_{i=1}^{k} (\mathbf{v}_i^\top \mathbf{x})_+ \Big)^2,$$

where $\mathbf{v}_1, \ldots, \mathbf{v}_k$ *are orthogonal unit vectors in* $\mathbb{R}^k$*. Then for* $n = k \in \{6, 7, \ldots, 20\}$ *as well as* $(k, n) \in \{(8, 9), (10, 11), \ldots, (19, 20)\}$*, this objective function has spurious local minima.*

Proof: computer-assisted

Dependence on n, k

More spurious minima observed at larger k, but overparametrization (large n) appears to partly remove them.

Table: Spurious local minima found for $n = k$

k	n	% of runs converging to local minima	Average minimal eigenvalue	Average objective value
6	6	0.3%	0.0047	0.025
7	7	5.5%	0.014	0.023
8	8	12.6%	0.021	0.021
9	9	21.8%	0.027	0.02
10	10	34.6%	0.03	0.022
11	11	45.5%	0.034	0.022
12	12	58.5%	0.035	0.021
13	13	73%	0.037	0.022
14	14	73.6%	0.038	0.023
15	15	80.3%	0.038	0.024
16	16	85.1%	0.038	0.027
17	17	89.7%	0.039	0.027
18	18	90%	0.039	0.029
19	19	93.4%	0.038	0.031
20	20	94%	0.038	0.033

Table: Spurious local minima found for $n \neq k$

k	n	% of runs converging to local minima	Average minimal eigenvalue	Average objective value
8	9	0.1%	0.0059	0.021
10	11	0.1%	0.0057	0.018
11	12	0.1%	0.0056	0.017
12	13	0.3%	0.0054	0.016
13	14	1.5%	0.0015	0.038
14	15	5.5%	0.002	0.033
15	16	10.1%	0.004	0.032
16	17	18%	0.0055	0.031
17	18	20.9%	0.007	0.031
18	19	36.9%	0.0064	0.028
19	20	49.1%	0.0077	0.027

Linear networks

Linear networks: no nonlinear activation ($\sigma(x) = x$)

For simplicity also drop the constant terms in neurons

Then the model becomes:

$$\mathbf{y} = \widetilde{f}(\mathbf{x}) = W_K W_{K-1} \cdots W_1 \mathbf{x}, \qquad W_k \in \mathbb{R}^{d_k \times d_{k-1}}$$

- Parameters: $\mathbf{W} = (W_1, \ldots, W_k)$
- The k'th layer has d_k neurons; the input layer has d_0 neurons

$\widetilde{f}$ can model only linear maps $\mathbf{x} \mapsto \mathbf{y}$!

Linear networks with quadratic loss

Quadratic loss:

$$L(\mathbf{W}) = \frac{1}{2}\int \|f(\mathbf{x}) - W_K W_{K-1}\cdots W_1\mathbf{x}\|^2 d\mu(\mathbf{x})$$

Given two maps f, g, consider the scalar product

$$\langle f, g\rangle = \int \langle f(\mathbf{x}), g(\mathbf{x})\rangle d\mu(\mathbf{x})$$

Then $L(\mathbf{W}) = \|f - \widetilde{f}_{\mathbf{W}}\|^2$, where $\widetilde{f}_{\mathbf{W}}$ belongs to the $d_0 d_K$-dimensional space F_{lin} of linear maps. By orthogonally projecting f to F_{lin}, we can assume that f is also linear:

$$f(\mathbf{x}) = R\mathbf{x}$$

Linear networks with quadratic loss

Let $\Delta = \Delta(\mathbf{W}) = W_K \cdots W_1 - R$

$$L(\mathbf{W}) = \frac{1}{2}\int \|\Delta \mathbf{x}\|^2 d\mu(\mathbf{x}) = \frac{1}{2}\int \langle \mathbf{x}, \Delta^*\Delta \mathbf{x}\rangle d\mu(\mathbf{x})$$
$$= \frac{1}{2}\int \mathrm{tr}(\Delta^*\Delta |\mathbf{x}\rangle\langle \mathbf{x}|) d\mu(\mathbf{x}) = \frac{1}{2}\,\mathrm{tr}(\Delta^*\Delta\Sigma),$$

where

$$\Sigma = \Sigma^* = \int |\mathbf{x}\rangle\langle \mathbf{x}| d\mu(\mathbf{x})$$

("Bra"-"ket" notation: $|\mathbf{u}\rangle\langle \mathbf{v}| : \mathbf{z} \mapsto \langle \mathbf{v}, \mathbf{z}\rangle \mathbf{u}$)

Constant width networks ($d_k \equiv d$)

M. Hardt, T. Ma, Identity Matters in Deep Learning, arXiv:1611.04231

Expressiveness of stacked "nearly identical" linear layers:

Proposition

Any $A \in \mathbb{R}^{d\times d}$ *with* $\det A > 0$ *can be represented as* $(1+V_K)(1+V_{K-1})\cdots(1+V_1)$, *where* $\max_{k=1,\ldots,K} \|V_k\| = O(1/K)$

Exercise: Why $\det A > 0$?

Proof

Write $A = BO$, where $B = B^*$ is positive definite and O orthogonal:

$$B = (AA^*)^{1/2}, \quad O = B^{-1}A$$

Canonical forms[2]:

$$B = O_1 \operatorname{diag}(\lambda_1, \ldots, \lambda_d) O_1^*$$
$$O = O_2 \operatorname{diag}\left(\begin{pmatrix} \cos\phi_1 & \sin\phi_1 \\ -\sin\phi_1 & \cos\phi_1 \end{pmatrix}, \ldots, \begin{pmatrix} \cos\phi_s & \sin\phi_s \\ -\sin\phi_s & \cos\phi_s \end{pmatrix}, 1, \ldots, 1 \right) O_2^*$$

Then $B = (B^{1/K})^K, \quad O = (O^{1/K})^K$, where

$$B^{1/K} = O_1 \operatorname{diag}(\sqrt[K]{\lambda_1}, \ldots, \sqrt[K]{\lambda_d}) O_1^* = \mathbf{1} + O(\tfrac{1}{K})$$
$$O^{1/K} = O_2 \operatorname{diag}\left(\begin{pmatrix} \cos\frac{\phi_1}{K} & \sin\frac{\phi_1}{K} \\ -\sin\frac{\phi_1}{K} & \cos\frac{\phi_1}{K} \end{pmatrix}, \ldots, \begin{pmatrix} \cos\frac{\phi_s}{K} & \sin\frac{\phi_s}{K} \\ -\sin\frac{\phi_s}{K} & \cos\frac{\phi_s}{K} \end{pmatrix}, 1, \ldots, 1 \right) O_2^*$$
$$= \mathbf{1} + O(\tfrac{1}{K})$$

[2] **Exercise** (Euler's rotation theorem): any $O \in SO(3)$ is a rotation about some axis by some degree ϕ.

Computation of $\nabla_{\mathbf{W}} L(\mathbf{W})$

Think of ∇_{W_k} as a matrix $(\nabla_{(W_k)_{m,n}})_{m,n}$

$$
\begin{aligned}
\nabla_{W_k} L(\mathbf{W}) &= \nabla_{W_k} \frac{1}{2} \operatorname{tr}(\Delta(\mathbf{W})^* \Delta(\mathbf{W}) \Sigma) \\
&= \operatorname{tr}(\Delta(\mathbf{W})^* \nabla_{W_k}(\Delta(\mathbf{W})) \Sigma) \\
&= \operatorname{tr}(\nabla_{W_k}(\Delta(\mathbf{W})) \Sigma \Delta(\mathbf{W})^*) \\
&= \operatorname{tr}(\nabla_{W_k}(W_K \cdots W_{k+1} W_k W_{k-1} \cdots W_1) \Sigma \Delta(\mathbf{W})^*) \\
&= \left(\operatorname{tr}\left((W_K \cdots W_{k+1} |\mathbf{e}_m\rangle \langle \mathbf{e}_n| W_{k-1} \cdots W_1) \Sigma \Delta(\mathbf{W})^* \right) \right)_{m,n} \\
&= \left(\langle \mathbf{e}_n| W_{k-1} \cdots W_1 \Sigma \Delta(\mathbf{W})^* W_K \cdots W_{k+1} |\mathbf{e}_m\rangle \right)_{m,n} \\
&= W_{k+1}^* \cdots W_K^* \Delta(\mathbf{W}) \Sigma W_1^* \cdots W_{k-1}^*
\end{aligned}
$$

Local nondegenerate minima are global

Theorem (Hardt, Ma)

Suppose that Σ is strictly positive definite, $\mathbf{W} = (W_1, \ldots, W_K)$ is a local minimum of $L(\mathbf{W})$ and all matrices W_k are nondegenerate. Then $\mathbf{W}$ is a global minimum.

Proof. By nondegeneracy of Σ and W_k,

$$0 = \nabla_{W_k} L(\mathbf{W}) = W_{k+1}^* \cdots W_K^* \Delta \Sigma W_1^* \cdots W_{k-1}^*$$

implies $\Delta = 0$. □

Exercise: If $W_k = W_n = 0$ for some $k \neq n$, then $\mathbf{W}$ is a stationary point.

Dynamics in the subalgebra generated by R

Key idea: ensure that matrices produced by the algorithm are diagonalized by a common basis[3]

Assume $R = R^*, \Sigma = \mathbf{1}$, and the starting values $W_k^{(1)} = \mathbf{1}$ for all k

$$
\begin{aligned}
W_k^{(2)} &= W_k^{(1)} - \alpha \nabla_{W^k} L(\mathbf{W}^{(1)}) = \mathbf{1} - \alpha(\mathbf{1} - R) \\
W_k^{(3)} &= W_k^{(2)} - \alpha \nabla_{W^k} L(\mathbf{W}^{(2)}) = \mathbf{1} - \alpha(\mathbf{1} - R) \\
&\quad - \alpha(\mathbf{1} - \alpha(\mathbf{1} - R))^* \cdots (\mathbf{1} - \alpha(\mathbf{1} - R))^* \\
&\quad \times ((\mathbf{1} - \alpha(\mathbf{1} - R)) \cdots (\mathbf{1} - \alpha(\mathbf{1} - R)) - R) \\
&\quad \times (\mathbf{1} - \alpha(\mathbf{1} - R))^* \cdots (\mathbf{1} - \alpha(\mathbf{1} - R))^* \\
&= p_{k,3}(R) \\
&\dots \\
W_k^{(n)} &= p_{k,n}(R)
\end{aligned}
$$

with some polynomials $p_{k,n}$.

[3] P. Bartlett et al., Gradient descent with identity initialization efficiently learns positive definite linear transformations by deep residual networks

Diagonalization and 1D dynamics

If $R = O\,\mathrm{diag}(r_1, \ldots, r_d)O^*$, then

$$W_k^{(n)} = O\,\mathrm{diag}(p_{k,n}(r_1), \ldots, p_{k,n}(r_d))O^*$$

Optimization dynamics decouples into d independent 1D components

$d = 1$: $\mathbf{w} = (w_1, \ldots, w_K)$

$$\frac{dw_k}{dt} = w_{k+1} \cdots w_K (r - w_1 \cdots w_K) w_1 \cdots w_{k-1}; \qquad w_k(t = 0) = 1$$

$$\frac{dw_k^2}{dt} = 2q(r - q), \qquad q = w_1 \cdots w_K$$

Exercise:

- $r > 0$: $\mathbf{w}$ converges to a solution of $r = w_1 \cdots w_K$
- $r \leq 0$: ?

Conclusion and generalizations

Theorem (Bartlett et al.)

Suppose that $\Sigma = \mathbf{1}, R = R^*$ *and is positive definite. Then the gradient descent starting from* $W_k^{(1)} \equiv \mathbf{1}$ *converges to a global minimum of* L.

Exercise[4]**:** A generalization: assume that there are orthogonal operators $O_k \in SO(d_k)$ diagonalizing the initial weight matrices $W_k^{(1)}$ and the operators R and Σ (in the sense that $O_k W_k^{(1)} O_{k-1}^*$ is diagonal and similarly for R, Σ). Then GD decouples into independent 1D components.

An open question (?): Is there an asymptotic description of GD in linear networks for generic initial conditions?

[4] A. Saxe et al., Exact solutions to the nonlinear dynamics of learning in deep linear neural networks, arXiv:1312.6120

General layer widths and loss functions

Th. Laurent, J. von Brecht, Deep Linear Networks with Arbitrary Loss: All Local Minima Are Global, arXiv:1712.01473

Assumptions:

1. The loss function $\widetilde{\mathbf{y}} \mapsto l(\mathbf{y}, \widetilde{\mathbf{y}})$ is convex and differentiable.
2. The thinnest layer is either the input layer or the output layer (i.e., $\min(d_1, \ldots, d_{K-1}) \geq \min(d_0, d_K)$).

Theorem (1)

Under these assumptions, the linear network has no spurious (sub-optimal) local minima, i.e. any local minimum is global.

Theorem (2)

There exists a convex, Lipschitz, but non-differentiable loss function $\widetilde{\mathbf{y}} \mapsto l(\mathbf{y}, \widetilde{\mathbf{y}})$ *with which the network has sub-optimal local minima.*

Sketch of proof of Theorem 1

Two problems:

$$\text{(P1)} \quad \begin{cases} \text{Minimize} \quad g(A) \\ \text{over all } A \text{ in } \mathbb{R}^{d_L \times d_0} \end{cases}$$

$$\text{(P2)} \quad \begin{cases} \text{Minimize} \quad g(W_K W_{K-1} \cdots W_1) \\ \text{over all } K\text{-tuples } (W_1, \ldots, W_K) \\ \text{in } \mathbb{R}^{d_1 \times d_0} \times \cdots \times \mathbb{R}^{d_K \times d_{K-1}} \end{cases}$$

Lemma

For any differentiable g, if $(\widehat{W}_1, \ldots, \widehat{W}_K)$ is a local minimizer of (P2), then $\widehat{W}_K \cdots \widehat{W}_1$ is a stationary point of (P1).

Exercise: the lemma implies Theorem 1 by convexity of l.

Sketch of proof of the Lemma

Let $G(W_1, \ldots, W_K) = g(W_1 \cdots W_K)$. Stationarity of G at $\widehat{\mathbf{W}}$:

$$0 = \nabla_{W_k} G(\widehat{\mathbf{W}}) = \widehat{W}_{k+1}^* \cdots \widehat{W}_K^* \nabla g(\widehat{W}_K \cdots \widehat{W}_1) \widehat{W}_1^* \cdots \widehat{W}_{k-1}^*, \quad \forall k$$

Easy case: $\ker(\widehat{W}_{K-1} \cdots \widehat{W}_1) = \{0\}$. Then from stationarity of G with $k = K$, by transposing:

$$0 = \widehat{W}_{K-1} \cdots \widehat{W}_1 \big(\nabla g(\widehat{W}_K \cdots \widehat{W}_1)\big)^*$$

Then $\nabla g(\widehat{W}_K \cdots \widehat{W}_1) = 0$, Q.E.D.

Harder case: $\ker(\widehat{W}_{K-1} \cdots \widehat{W}_1) \neq \{0\}$. Main idea: construct a family of local perturbations $\widetilde{W}_1, \ldots, \widetilde{W}_K$ such that

$$\widetilde{W}_K \cdots \widetilde{W}_1 = \widehat{W}_K \cdots \widehat{W}_1 \quad \text{and} \quad \|\widetilde{W}_k - \widehat{W}_k\| < \epsilon \quad \forall k \tag{1}$$

Then $(\widetilde{W}_1, \ldots, \widetilde{W}_K)$ is also a local minimum of G.

Construction of local perturbations

$$\ker(\widehat{W}_1) \subset \ker(\widehat{W}_2\widehat{W}_1) \subset \ldots \subset \ker(\widehat{W}_{K-1}\cdots\widehat{W}_1) \neq \{0\}$$

Assume for simplicity that $\dim\ker(\widehat{W}_k\cdots\widehat{W}_1) > 0$ for all k

Suppose that $d_k \geq d_0$. Then

$$\text{co-dim}\,\text{Ran}(\widehat{W}_k\cdots\widehat{W}_1) \geq \dim\ker(\widehat{W}_k\cdots\widehat{W}_1) > 0,$$

so there is $0 \neq \mathbf{u}_k \in \mathbb{R}^{d_k} \ominus \text{Ran}(\widehat{W}_k\cdots\widehat{W}_1)$. Let

$$\widetilde{W}_k = \widehat{W}_k + \delta_k|\mathbf{w}_k\rangle\langle\mathbf{u}_{k-1}|$$

with any $\mathbf{w}_k$ and small δ_k. This fulfills conditions (1).

End of proof

From stationarity, with $k = 1$:

$$0 = \left(\nabla g(\widehat{W}_K \cdots \widehat{W}_1)\right)^* \widetilde{W}_K \cdots \widetilde{W}_2$$

Since we can add to $\widetilde{W}_2$ an arbitrary term $\delta_k |\mathbf{w}_2\rangle\langle \mathbf{u}_1|$:

$$0 = \left(\nabla g(\widehat{W}_K \cdots \widehat{W}_1)\right)^* \widetilde{W}_K \cdots \widetilde{W}_3 |\mathbf{w}_2\rangle\langle \mathbf{u}_1|$$

Since $\mathbf{w}_2$ was arbitrary:

$$0 = \left(\nabla g(\widehat{W}_K \cdots \widehat{W}_1)\right)^* \widetilde{W}_K \cdots \widetilde{W}_3$$

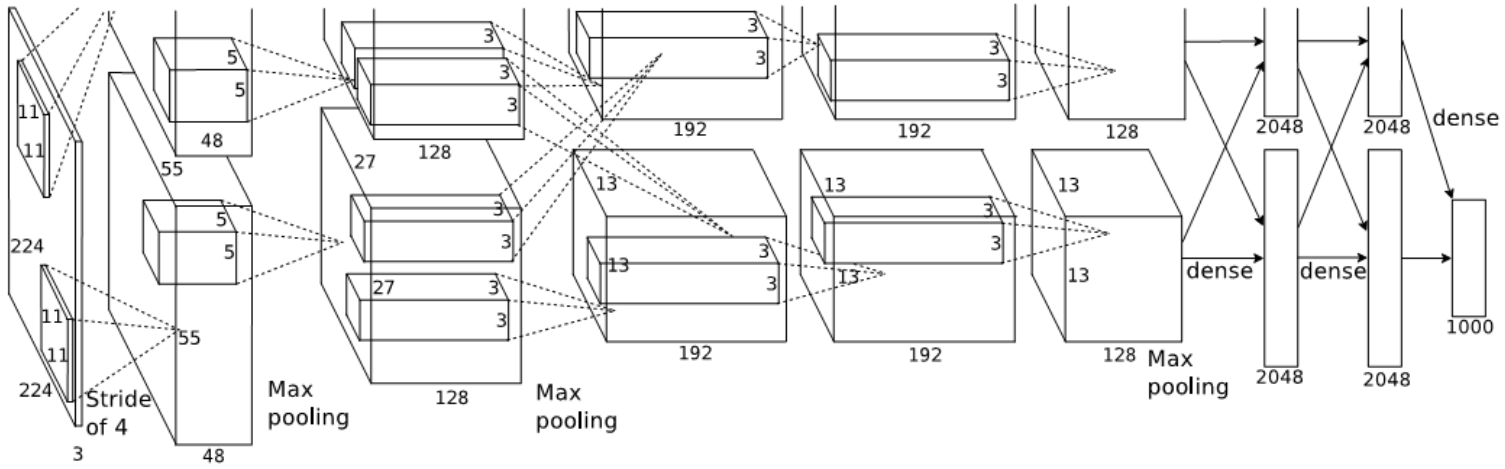
Removing in the same way $\widetilde{W}_3, \widetilde{W}_4, \ldots$:

$$0 = \left(\nabla g(\widehat{W}_K \cdots \widehat{W}_1)\right)^*,$$

hence $\nabla g(\widehat{W}_K \cdots \widehat{W}_1) = 0$. □

Standard vs. residual networks

An early standard multilayer architecture for image recognition:



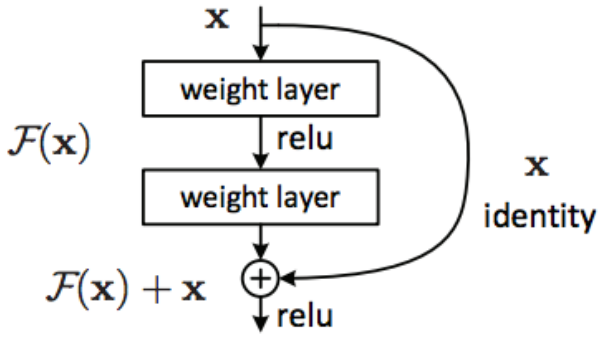
A. Krizhevsky et al., ImageNet Classification with Deep Convolutional Neural Networks, 2012

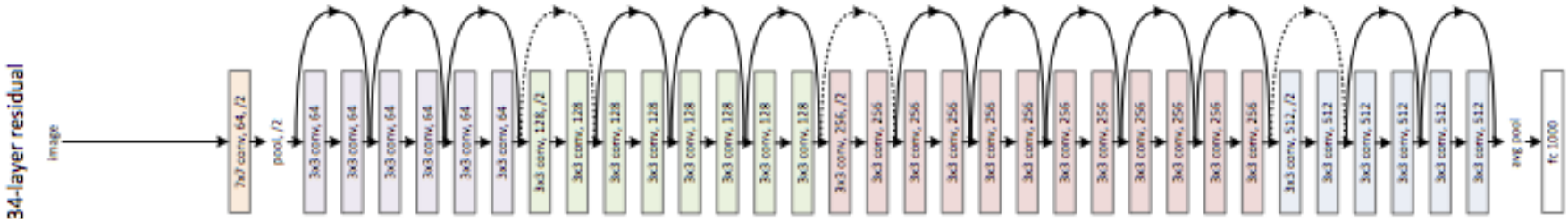
- Few layers
- "Rough" transition between layers

Residual networks: accuracy improved

K. He et al., Deep Residual Learning for Image Recognition, arXiv:1512.03385

- Shortcut connections: learn only residuals to Id maps
- Only 1x1 and 3x3 convolutions
- Up to 150 layers





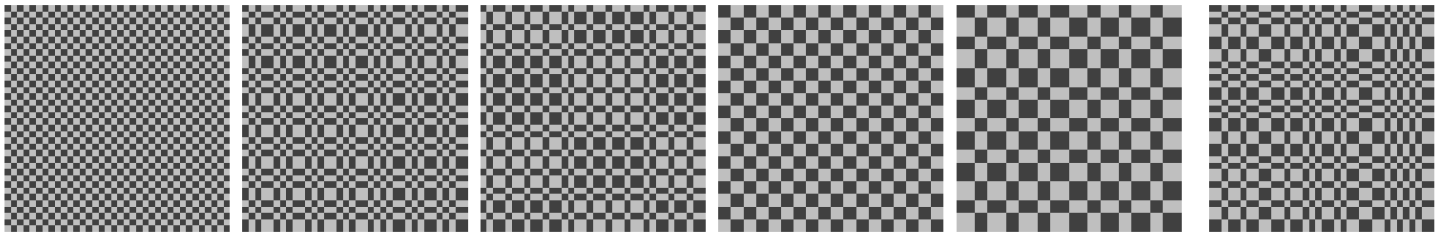
Benefits:

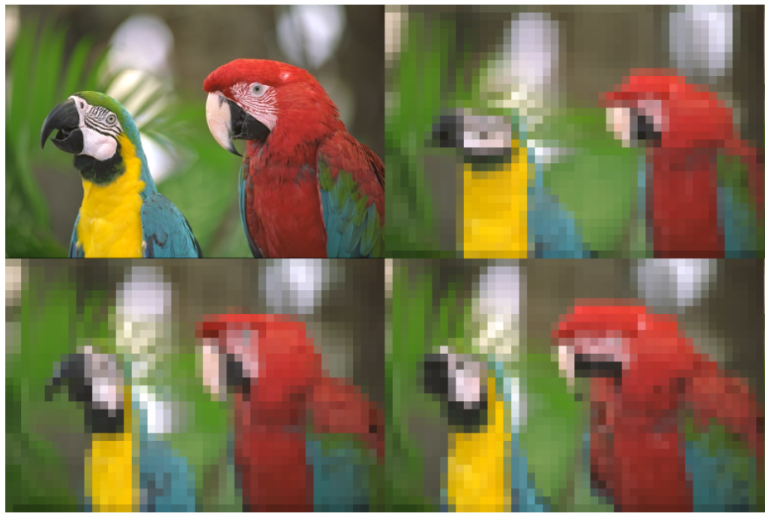
- Non-divergent initialization (a perturbation of Id)
- Smooth transition between length scales

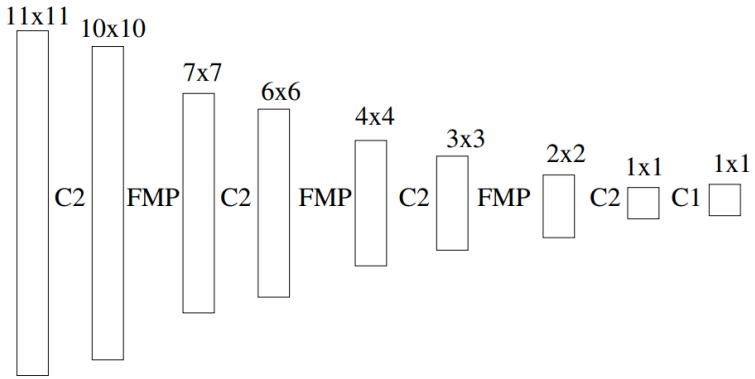
Fractional max-pooling

B. Graham, Fractional Max-Pooling, arXiv:1412.6071

Smoother transitions between length scales by a more complex pooling







Continuum limit of a deep constant-width linear network

Recall the constant-width linear network with small residuals:

$$\widetilde{f}(\mathbf{x}) = (1 + V_K)(1 + V_{K-1}) \cdots (1 + V_1)\mathbf{x},$$

where $\max_{k=1,\ldots,K} \|V_k\| = O(1/K)$

Let $V_k = \frac{1}{K} A_{k/K}$, for some $A_s, s \in [0, 1]$. Let $K \to \infty$ and

$$X_s = \lim_{K \to \infty} (1 + V_{\lfloor Ks \rfloor})(1 + V_{\lfloor Ks \rfloor - 1}) \cdots (1 + V_1)$$

Then

$$\frac{d}{ds} X_s = A_s X_s, \quad X_{s=0} = \mathbf{1}$$

Exercise: The solution can be written using time-ordered exponentials:

$$\begin{aligned} X_s &= \mathcal{T}\{e^{\int_0^s A_u du}\} \\ &\equiv \sum_{n=0}^{\infty} \int_0^s du_1 \int_0^{u_1} du_2 \cdots \int_0^{u_{n-1}} du_n \left[A_{u_1} A_{u_2} \cdots A_{u_n}\right] \end{aligned}$$

Bounds on X_s

How does X_s depend on the control variable A_s? Can X_s blow up or degenerate? Not if A_s is not too big:

Lemma (1)

$$e^{-\int_0^s \|A_u\| du}|\mathbf{x}| \leq |\mathcal{T}\{e^{\int_0^s A_u du}\}^* \mathbf{x}| \leq e^{\int_0^s \|A_u\| du}|\mathbf{x}|$$

(where $|\mathbf{x}| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$*)*

Exercise: If $|B^*\mathbf{x}| \geq c|\mathbf{x}|$ for all $\mathbf{x}$, then also $|B\mathbf{x}| \geq c|\mathbf{x}|$ for all $\mathbf{x}$.

Exercise: Prove the upper bound using the expansion of $\mathcal{T}\{e^{\int_0^s A_u du}\}$.

Proof of Lemma (1)

Consider $X_s X_s^*$:

$$\frac{d}{ds}(X_s X_s^*) = A_s X_s X_s^* + X_s X_s^* A_s^*$$

Let $\lambda_{\min}(s), \lambda_{\max}(s)$ be the minimum and maximum eigenvalues of $X_s X_s^*$:

$$\begin{aligned}\lambda_{\min \text{ (or } \max)}(s) &= \min \text{ (or } \max)_{\mathbf{e}:|\mathbf{e}|=1} \langle \mathbf{e}, X_s X_s^* \mathbf{e} \rangle \\ &= \min \text{ (or } \max)_{\mathbf{e}:|\mathbf{e}|=1} |X_s^* \mathbf{e}|^2\end{aligned}$$

Then

$$\lambda_{\min}^{1/2}(s)|\mathbf{x}| \leq |\mathcal{T}\{e^{\int_0^s A_u du}\}^* \mathbf{x}| \leq \lambda_{\max}^{1/2}(s)|\mathbf{x}|$$

Perturbation of eigenvalues

Lemma (2)

Let $B = B^*, C = C^*$, and B has a nondegenerate eigenvalue λ with the unit eigenvector $\mathbf{e}$. Then $B + \epsilon C$ has eigenvalue $\lambda_\epsilon = \lambda + \epsilon\langle \mathbf{e}, C\mathbf{e}\rangle + O(\epsilon^2)$.

Applying Lemma (2) with $B = X_s X_s^*, C = A_s X_s X_s^* + X_s X_s^* A_s^*$:

$$\frac{d}{ds}\lambda_{\text{min (or max)}}(s) = \langle \mathbf{e}_{\text{min (or max)}}, (A_s X_s X_s^* + X_s X_s^* A_s^*)\mathbf{e}_{\text{min (or max)}}\rangle$$

$$\frac{d}{ds}\lambda_{\max}(s) \le 2\|A_s\|\lambda_{\max}(s)$$

$$\frac{d}{ds}\ln\lambda_{\max}(s) \le 2\|A_s\|$$

$$\lambda_{\max}(s) \le e^{2\int_0^s \|A_s\| ds}$$

$$\frac{d}{ds}\lambda_{\min}(s) \ge -2\|A_s\|\lambda_{\min}(s)$$

$$\frac{d}{ds}\ln\lambda_{\min}(s) \ge -2\|A_s\|$$

$$\lambda_{\min}(s) \ge e^{-2\int_0^s \|A_s\| ds}$$

□

Proof of Lemma (2): spectral theory, the resolvent[5]

The **resolvent** of the operator B:

$$\mathcal{R}_B(z) = (B - z)^{-1}, \quad z \in \mathbb{C} \setminus \text{spec}(B)$$

Let $(\mathbf{e}_n)_{n=1}^d$ be the eigenbasis of B:

$$B = \sum_{n=1}^{d} \lambda_n |\mathbf{e}_n\rangle\langle\mathbf{e}_n|$$

Then

$$\mathcal{R}_B(z) = \sum_{n=1}^{d} \frac{1}{\lambda_n - z} |\mathbf{e}_n\rangle\langle\mathbf{e}_n|$$

Let $D \subset \mathbb{C}$ and f be analytic in D, then

$$\frac{1}{2\pi i} \int_{\partial D} f(z)\mathcal{R}_B(z)dz = -\sum_{n:\lambda_n \in D} f(\lambda_n)|\mathbf{e}_n\rangle\langle\mathbf{e}_n|$$

[5] T. Kato, Perturbation Theory for Linear Operators

The resolvent expansion

In particular, if D contains λ_n, but not the other points of the spectrum, then:

$$\lambda_n = -\operatorname{tr} \frac{1}{2\pi i} \int_{\partial D} z \mathcal{R}_B(z) dz$$

We can expand $\mathcal{R}_{B+\epsilon C}(z)$ for $z \in \partial D$ and small ϵ:

$$\begin{aligned}
\mathcal{R}_{B+\epsilon C}(z) &= (B - z + \epsilon C)^{-1} \\
&= (B - z)^{-1}(1 + \epsilon C (B - z)^{-1})^{-1} \\
&= \sum_{n=0}^{\infty} \epsilon^n \mathcal{R}_B(z)(-C\mathcal{R}_B(z))^n \\
&= \mathcal{R}_B(z) - \epsilon \mathcal{R}_B(z) C \mathcal{R}_B(z) + O(\epsilon^2)
\end{aligned}$$

Perturbation of the eigenvalue

Let $\lambda_n(B+\epsilon C)$ be the perturbed n'th eigenvalue of $B+\epsilon C$.

$$\begin{aligned}
\lambda_n(B+\epsilon C) &= \lambda_n(B) + \epsilon \operatorname{tr} \frac{1}{2\pi i} \int_{\partial D} z\mathcal{R}_B(z) C \mathcal{R}_B(z) dz + O(\epsilon^2) \\
&= \lambda_n(B) + \frac{\epsilon}{2\pi i} \int_{\partial D} \sum_{m=1}^{d} \langle \mathbf{e}_m, z\mathcal{R}_B(z) C \mathcal{R}_B(z) \mathbf{e}_m \rangle dz + O(\epsilon^2) \\
&= \lambda_n(B) + \frac{\epsilon}{2\pi i} \int_{\partial D} \sum_{m=1}^{d} \frac{C_{mm} z}{(\lambda_m - z)^2} dz + O(\epsilon^2) \\
&= \lambda_n(B) + \epsilon C_{nn} + O(\epsilon^2)
\end{aligned}$$

□

Exercise: Show that the second order correction to the eigenvalue λ_n is $\epsilon^2 \sum_{m:m\neq n} \frac{|C_{mn}|^2}{\lambda_m - \lambda_n}$

Gradient descent

Let $\mathbf{A} = \{A_s, s \in [0,1]\}$, $\Delta = X_{s=1} - R$ and $L = \frac{1}{2}\operatorname{tr}(\Delta^* \Delta \Sigma)$.

Exercise: $\nabla_{\mathbf{A}} L = \{\nabla_{A_s} L, s \in [0,1]\}$, where

$$\begin{aligned}\nabla_{A_s} L &= \mathcal{T}\{e^{\int_s^1 A_u du}\}^* (\mathcal{T}\{e^{\int_0^1 A_u du}\} - R)\Sigma \mathcal{T}\{e^{\int_0^s A_u du}\}^* \\ &= (X_1 X_s^{-1})^* \Delta \Sigma X_s^*\end{aligned}$$

Gradient descent:

$$\frac{d}{dt}\mathbf{A} = -\nabla_{\mathbf{A}} L$$

Exercise: Describe explicitely the gradient descent when dimension $d = 1$. Does the gradient descent always converge?

No spurious local minima (again)[6]

Scalar product of matrix-valued functions
$\mathbf{B} = \{B_s \in \mathbb{R}^{d\times d}\}, \mathbf{C} = \{C_s \in \mathbb{R}^{d\times d}\}, s \in [0,1]$:

$$\langle \mathbf{B}, \mathbf{C} \rangle = \int_0^1 \mathrm{tr}(B_s^* C_s) ds$$

The corresponding L^2 norm:

$$\|\mathbf{B}\|_{L^2}^2 = \int_0^1 \mathrm{tr}(B_s^* B_s) ds$$

Proposition

Suppose that Σ is strictly positive definite (sample non-degeneracy). Then

1. $\nabla_{\mathbf{A}} L(\mathbf{A}) = 0$ *iff* $L(\mathbf{A}) = 0$, *i.e. iff* $X_1 = R$
2. $\|\nabla_{\mathbf{A}} L(\mathbf{A})\|_{L^2} \geq \sigma e^{-\int_0^1 \|A_s\| ds} \|\Delta\|$, *where* $\sigma > 0$ *is the minimum eigenvalue of* Σ.

[6] A. Taghvaei et al., How regularization affects the critical points in linear networks

Proof of (2)

$$\|\nabla_{\mathbf{A}} L\|^2_{L^2[0,1]} = \int_0^1 \operatorname{tr}\left[(\nabla_{A_s} L)(\nabla_{A_s} L)^*\right] ds = \int_0^1 \operatorname{tr}[F_s^* \Delta G_s G_s^* \Delta^* F_s] ds,$$

where

$$F_s = \mathcal{T}\{e^{\int_s^1 A_u du}\}, \qquad G_s = \Sigma \mathcal{T}\{e^{\int_0^s A_u du}\}^*$$

By Lemma (1):

$$\langle \mathbf{x}, F_s F_s^* \mathbf{x} \rangle \geq e^{-2\int_s^1 \|A_u\| du} \langle \mathbf{x}, \mathbf{x} \rangle$$

$$\langle \mathbf{x}, G_s G_s^* \mathbf{x} \rangle \geq \sigma^2 e^{-2\int_0^s \|A_u\| du} \langle \mathbf{x}, \mathbf{x} \rangle$$

Then

$$\begin{aligned}
\operatorname{tr}[F_s^* \Delta G_s G_s^* \Delta^* F_s] &\geq \sigma^2 e^{-2\int_0^s \|A_u\| du} \operatorname{tr}[F_s^* \Delta \Delta^* F_s] \\
&= \sigma^2 e^{-2\int_0^s \|A_u\| du} \operatorname{tr}[\Delta^* F_s F_s^* \Delta] \\
&\geq \sigma^2 e^{-2\int_0^s \|A_u\| du} \cdot e^{-2\int_s^1 \|A_u\| du} \operatorname{tr}[\Delta^* \Delta] \\
&\geq \sigma^2 e^{-2\int_0^1 \|A_u\| du} \|\Delta\|^2
\end{aligned}$$

□

Wide nonlinear networks: finite sample expressiveness

Exercise[7]**:** There exists a two layer neural network with ReLU activations and $2n + d$ weights that can represent any function on a sample of size n in d dimensions.

[7] Ch. Zhang et al., Understanding deep learning requires rethinking generalization, arXiv:1611.03530

Manifestations of optimization failure

- Optimization can get stuck at suboptimal local minima
- Optimization iterates $\mathbf{W}^{(n)}$ may by unbounded and have no limit points
- Some regions of the loss surface may be "flat"

Exercise: Perform gradient descent numerically for a shallow network with 1D input. Observe whether GD converges to the global minimum and explain why if not.

Getting rid of bias terms

Standard layer representation: $\mathbf{z}_k = \sigma(W_k \mathbf{z}_{k-1} + \mathbf{h})$, with σ evaluated component-wise

We can get rid of bias terms by adding an extra neuron with output 1:

$$\widetilde{\mathbf{z}}_{k-1} = (\mathbf{z}_{k-1}, 1)$$

$$\mathbf{z}_k = \sigma(\widetilde{W}_k \widetilde{\mathbf{z}}_{k-1})$$

The "extended" network inputs:

$$\widetilde{\mathbf{x}} = (\mathbf{x}, 1)$$

In the wollowing assume W.L.O.G that the network layers are without bias terms:

$$\mathbf{z}_k = \sigma(W_k \mathbf{z}_{k-1})$$

and the network parameters are $\mathbf{W} = (W_1, \ldots, W_K)$

Gradient of nonlinear networks

- $\mathbf{y}_k := (y_{k,1}, \ldots, y_{k,d_k})$: pre-activation outputs in the k'th layer (d_k neurons)
- $F_k := (\sigma(y_{k,1}), \ldots, \sigma(y_{k,d_k}))$: post-activation outputs
- $F_k' := \mathrm{diag}(\sigma'(y_{k,1}), \ldots, \sigma'(y_{k,d_k}))$: diagonal matrix of respective derivatives of σ

Then, for the quadratic loss $L(\mathbf{W}) = \frac{1}{2}\int |f(\mathbf{x}) - \widetilde{f}(\mathbf{x}, \mathbf{W})|^2 d\mu(\mathbf{x})$:

$$\begin{aligned}\left(\nabla_{W_k} L(\mathbf{W})\right)_{ij} = \int d\mu(\mathbf{x}) \Big\langle \widetilde{f}(\mathbf{x}, \mathbf{W}) - f(\mathbf{x}), \\ W_K \cdot F_{K-1}'(\mathbf{x}) \cdot W_{K-1} \cdot F_{K-2}'(\mathbf{x}) \ldots F_k'(\mathbf{x}) \cdot |\mathbf{e}_i\rangle\langle \mathbf{e}_j| \cdot F_{k-1}(\mathbf{x}) \Big\rangle \\ = \int d\mu(\mathbf{x}) \Big[(F_{k-1}(\mathbf{x}))_j \times \\ \times \left(F_k'(\mathbf{x}) \cdot W_{k+1}^* \ldots F_{K-1}'(\mathbf{x}) \cdot W_K^* \cdot (\widetilde{f}(\mathbf{x}, \mathbf{W}) - f(\mathbf{x}))\right)_i \Big]\end{aligned}$$

Exercise: Check that this agrees with the formula for linear networks.

Pyramidal networks[8]

Pyramidal networks: layer widths do not increase (i.e., $d_{k+1} \leq d_k$)

Suppose that the activation function is differentiable, and $\sigma'(x) > 0$.

Consider training with the quadratic loss and on a *finite* training set $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$

Theorem

Let $\mathbf{W} = (W_1, \ldots, W_K)$ *be a stationary point of the gradient descent. Suppose that:*

1. *The extended input vectors* $\{\widetilde{\mathbf{x}}_n = (\mathbf{x}_n, 1)\}_{n=1}^N$ *in the training set are linearly independent (i.e.,* $\{\mathbf{x}_n\}_{n=1}^N$ *are affinely independent);*
2. *The weight matrices* W_k *have full ranks (i.e.,* $\text{rank}(W_k) = d_k$*).*

Then $\mathbf{W}$ *is a global minimum:* $L(\mathbf{W}) = 0$.

[8] M. Gori and A. Tesi, On the problem of local minima in backpropagation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 14:76-86, 1992

Proof

Applying the stationarity condition with $d\mu(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\delta(\widetilde{\mathbf{x}} - \widetilde{\mathbf{x}}^{(n)})$ and $k = 1$: for all i, j

$$
\begin{aligned}
0 &= \left(\nabla_{W_1} L(\mathbf{W})\right)_{ij} \\
&= \sum_{n=1}^{N} \widetilde{x}_j^{(n)} \times \left(F_1'(\widetilde{\mathbf{x}}^{(n)}) \cdot W_2^* \ldots F_{K-1}'(\widetilde{\mathbf{x}}^{(n)}) \cdot W_K^* \cdot (\widetilde{f}(\widetilde{\mathbf{x}}^{(n)}, \mathbf{W}) - f(\widetilde{\mathbf{x}}^{(n)}))\right)_i \\
&= \sum_{n=1}^{N} \widetilde{x}_j^{(n)} \times Z_i^{(n)}
\end{aligned}
$$

We want to prove that $f(\widetilde{\mathbf{x}}^{(n)}, \mathbf{W}) - f(\widetilde{\mathbf{x}}^{(n)}) = 0$ for all n.

Suppose that's not so. Since W_k have full ranges, $\ker W_k^* = 0$. Also, all matrices $F_k'(\mathbf{x})$ are nondegenerate, since $\sigma'(t) > 0$. Therefore $Z_i^{(n)} \not\equiv 0$. But then $\mathbf{x}^{(n)}$ are not linearly independent. Contradicton. $\square$

The Johnson – Lindenstrauss lemma

A set of size N in a high-dimensional space can be projected to $\mathbb{R}^{\sim \log N}$ while almost preserving its metric properties:

Theorem (Johnson – Lindenstrauss)

Given $\epsilon < 0.5$, a set X of N points in $\mathbb{R}^d$, and a number $n > \frac{20 \ln N}{\epsilon^2}$, there is a linear map $f : \mathbb{R}^d \to \mathbb{R}^n$ such that for all $\mathbf{u}, \mathbf{v} \in X$

$$(1-\epsilon)\|\mathbf{u}-\mathbf{v}\|^2 \leq \|f(\mathbf{u})-f(\mathbf{v})\|^2 \leq (1+\epsilon)\|\mathbf{u}-\mathbf{v}\|^2.$$

Key idea: with overwhelming probability, one can take a rescaled random projection from $\mathbb{R}^d$ to $\mathbb{R}^n$ as f.

Example of application[9]: a sample of size N can be fitted exactly by projecting to $\mathbb{R}^{\sim \log N}$ in the first layer and then using a residual-like ReLU network of width $\sim \log N$ and depth N.

[9] M. Hardt, T. Ma, Identity Matters in Deep Learning, arXiv:1611.04231

Idea of proof

Let $f : \mathbb{R}^d \to \mathbb{R}^n$ be a random linear operator with matrix $A = (\frac{1}{\sqrt{n}}\xi_{ks})$, where ξ_{ks} are i.i.d. standard normal $(k = 1, \dots, n; s = 1, \dots, d)$.

Let $\mathbf{u}, \mathbf{v}$ be a fixed pair of points. We want to find the distribution of $\|f(\mathbf{u}) - f(\mathbf{v})\|^2$. We may assume W.L.O.G that $\mathbf{u} - \mathbf{v} = \|\mathbf{u} - \mathbf{v}\|\mathbf{e}_1$, since the distribution of A is invariant under multiplication by $O(d)$-matrices.

Then

$$\|f(\mathbf{u}) - f(\mathbf{v})\|^2 = \|\mathbf{u} - \mathbf{v}\|^2 \cdot \frac{1}{n}\sum_{k=1}^{n} \xi_{k1}^2$$

Measure concentration: $\frac{1}{n}\sum_{k=1}^{n} \xi_{k1}^2 \approx 1$, and

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n} \xi_{k1}^2 - 1\right| > \epsilon\right) \lesssim e^{-I(\epsilon)n}$$

with some constant $I(\epsilon)$.

Sketch of proof

Since we have $\sim N^2$ pairs $(\mathbf{u}, \mathbf{v})$, we have nonzero probability that for all these pairs we have $\left| \frac{\|f(\mathbf{u}) - f(\mathbf{v})\|^2}{\|\mathbf{u} - \mathbf{v}\|^2} - 1 \right| < \epsilon$, if

$$e^{-I(\epsilon)n} \lesssim N^{-2}.$$

This condition is satisfied by choosing

$$n \gtrsim \frac{2 \ln N}{I(\epsilon)}$$

From large deviation theory, $I(\epsilon) = c\epsilon^2 + O(\epsilon^3)$

□

Measure concentration results

Let ξ_k be i.i.d. random variables with $\mathbb{E}\xi_k = 0$, and let $\eta_n = \sum_{k=1}^{n} \xi_k$

- Law of Large Numbers: $\frac{\eta_n}{n} \to 0$
- Central Limit Theorem: $\frac{\eta_n}{\sqrt{n}} \to \mathcal{N}(0, \mathbb{E}\xi^2)$ in distribution
- Large Deviation Theory[10]: Given $x > 0$, $\mathbb{P}(\frac{\eta_n}{n} > x) \simeq e^{-nI(x)}$ with a *rate function* $I(x)$ depending on the distribution of ξ

[10] A good introduction: J. T. Lewis, R. Russell, An Introduction to Large Deviations for Teletraffic Engineers

Large deviation theory: basic idea (Chernoff's bound)

For any t,

$$\mathbb{E}(e^{t\eta_n}) = \mathbb{E}(e^{t\sum_{k=1}^n \xi_k}) = \mathbb{E}\Big(\prod_{k=1}^n e^{t\xi_k}\Big) = \prod_{k=1}^n \mathbb{E}(e^{t\xi_k}) = (\mathbb{E}(e^{t\xi}))^n$$

$$= e^{n\ln \mathbb{E}(e^{t\xi})}$$

In particular, by Markov's inequality, fot $t \geq 0$

$$\mathbb{P}(\tfrac{\eta_n}{n} > x) \leq \mathbb{E}(e^{t(\eta_n - nx)}) = e^{-N(tx - \ln \mathbb{E}(e^{t\xi}))} = e^{-nI(x)},$$

where

$$I(x) = \sup_{t\geq 0}(tx - \ln \mathbb{E}(e^{t\xi}))$$

($I(x)$ is the Legendre transform of $f(t) = \ln \mathbb{E}(e^{t\xi})$ up to restriction $t \geq 0$)

Exercise: $f(t)$ is convex; $f(t) = \frac{t^2}{2}\mathbb{E}(\xi^2) + O(t^3)$ at small t; $I(x)$ is convex; $I(x) = \frac{x^2}{2\mathbb{E}(\xi^2)} + O(x^3)$.

Sketch of the lower bound

Suppose that $\mathbb{P}(\frac{\eta_n}{n} = x) \simeq e^{-n\widetilde{I}(x)}$ with some $\widetilde{I}(x)$; let us show that $\widetilde{I} = I$

Assume $\widetilde{I}(x)$ is monotone increasing, then

$$\mathbb{P}(\frac{\eta_n}{n} > x) \simeq \int_x^\infty e^{-n\widetilde{I}(z)} dz \simeq e^{-n\widetilde{I}(x)}$$

Also,

$$\mathbb{E}(e^{t\eta_n}) \simeq \int e^{-n\widetilde{I}(z)} e^{ntz} dz \simeq e^{n \sup_z (tz - \widetilde{I}(z))},$$

i.e.

$$\sup_z (tz - \widetilde{I}(z)) \simeq \ln \mathbb{E}(e^{t\xi})$$

By duality of Legendre transform, $\widetilde{I} = I$.

Exercises

Exercise: Compute the rate function $I(x)$ for:

- $\xi \sim \mathcal{N}(0,1)$ (check that $I(x) = \frac{x^2}{2}$)
- ξ Bernoulli ($P(\xi = \pm 1) = \frac{1}{2}$)

Exercise: Let $B_{p,n}$ be the ball $\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_p < r\}$ w.r.t. the l^p norm in $\mathbb{R}^n$, and the radius r is chosen so that $\mathrm{Vol}(B_{p,n}) = 1$. Using the large deviation theory, show that

$$\mathrm{Vol}(B_{1,n} \cap B_{\infty,n}) \simeq e^{-an} \quad (n \to \infty)$$

with some constant a, and find this a.

ANN in the large-width limit: independent weights[11]

Transformation from layer $k-1$ to k:

$$\mathbf{x}_k = \sigma(\mathbf{z}_k), \quad \mathbf{z}_k = W_k \mathbf{x}_{k-1} + \mathbf{h}_k, \qquad \mathbf{x}_k, \mathbf{z}_k \in \mathbb{R}^{d_k}$$

Assume $W_k, \mathbf{h}_k$ are random:

- $(W_k)_{ij}$ are i.i.d. with mean 0 and variance $\frac{\alpha_w^2}{d_{k-1}}$
- $(\mathbf{h}_k)_i$ are i.i.d. with mean 0 and variance α_h^2

By CLT: in the large-d_{k-1} limit, for a fixed $\mathbf{x}_{k-1}$, all components in $\mathbf{z}_k$ are i.i.d normal with mean 0 and variance $\alpha_w^2 \frac{|\mathbf{x}_{k-1}|^2}{d_{k-1}} + \alpha_h^2$

[11] B. Poole et al., Exponential expressivity in deep neural networks through transient chaos, arXiv:1606.05340

Reduction to 1D dynamics

For given input $\mathbf{x}_0$, define magnitude of the propagated signal in k'th layer:

$$q_k = \frac{1}{d_k} \sum_{i=1}^{d_k} z_{k,i}^2$$

Law of Large Numbers: q_k is approximately deterministic, $q_k \approx \mathbb{E} z_{k,i}^2$

Proposition (Poole et al.)

Evolution of q_k is given by

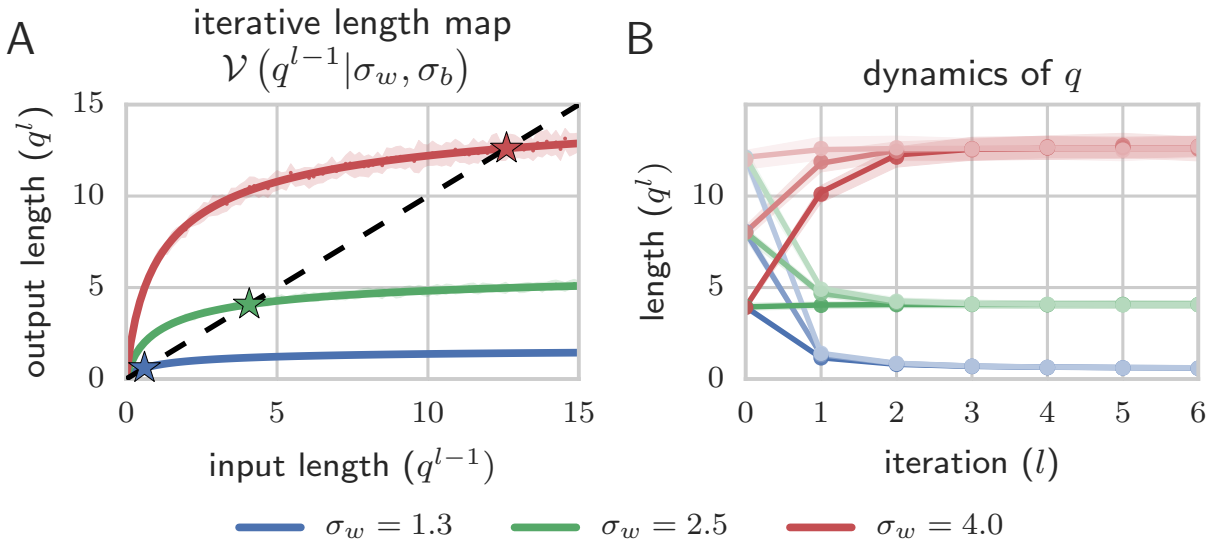
$$q_k = \nu(q_{k-1} | \alpha_w^2, \alpha_h^2) := \alpha_w^2 \int \sigma^2(\sqrt{q_{k-1}} s) \mathcal{D}s + \alpha_h^2, \quad \mathcal{D}s \equiv \frac{1}{\sqrt{2\pi}} e^{-s^2/2} ds$$

Proof: $q_k = \alpha_w^2 \frac{|\mathbf{x}_{k-1}|^2}{d_{k-1}} + \alpha_h^2 \approx \alpha_w^2 \mathbb{E}(x_{k-1,i}^2) + \alpha_h^2 = \alpha_w^2 \mathbb{E}(\sigma^2(z_{k-1,i})) + \alpha_h^2$, and $\mathbb{E}(\sigma^2(z_{k-1,i})) = \int \sigma^2(\sqrt{q_{k-1}} s) \mathcal{D}s$ since $z_{k-1,i} \sim \mathcal{N}(0, q_{k-1}^2)$ $\square$

Fixed points

Exercise: Find the map $q_k = \nu(q_{k-1})$ explicitely in the case of $\sigma = \mathrm{ReLU}$

In general, the map $q_k = \nu(q_{k-1})$ may have several stable or unstable fixed points:



Exercise:

- When is $q = 0$ a fixed point?
- When is a fixed point stable?
- Show that with $\sigma(z) = \tanh(z)$, the map ν has a stable fixed point $q_* \neq 0$.

Evolution of covariances

Let $\mathbf{x}_{0,1}$ and $\mathbf{x}_{0,2}$ be two input vectors. Define

$$q_{k,ab} = \frac{1}{d_k}\langle \mathbf{z}_{k,a}, \mathbf{z}_{k,b}\rangle = \frac{1}{d_k}\sum_{i=1}^{d_k}\langle z_{k,i,a}, z_{k,i,b}\rangle, \quad a, b \in \{1, 2\}$$

In particular, $q_{k,aa} = q_k$, and we already know its evolution. How does $q_{k,12}$ evolve?

We are looking for a map

$$q_{k,12} = C(c_{k-1,12}, q_{k-1,11}, q_{k-1,22}|\alpha_w, \alpha_h)$$

where $c_{k-1,12}$ is the *correlation coefficient*:

$$c_{k-1,12} = \frac{q_{k-1,12}}{\sqrt{q_{k-1,11}\, q_{k-1,22}}}$$

Evolution of covariances

$$
\begin{aligned}
q_{k,12} &= \frac{1}{d_k}\sum_{i=1}^{d_k}\langle \mathbf{z}_{k,1}, \mathbf{z}_{k,2}\rangle \approx \mathbb{E}(z_{k,1,i}z_{k,2,i}) \\
&= \mathbb{E}\left[\left(\sum_{j=1}^{d_{k-1}}(W_k)_{ij}x_{k-1,1,j} + h_{k,i}\right)\left(\sum_{m=1}^{d_{k-1}}(W_k)_{im}x_{k-1,2,m} + h_{k,i}\right)\right] \\
&= \sum_{j=1}^{d_{k-1}}\mathbb{E}((W_k)_{ij}^2)x_{k-1,1,j}x_{k-1,2,j} + \mathbb{E}(h_{k,i}^2) \\
&\approx \alpha_w^2\mathbb{E}(x_{k-1,1,j}x_{k-1,2,j}) + \alpha_h^2 \\
&= \alpha_w^2\mathbb{E}(\sigma(z_{k-1,1,j})\sigma(z_{k-1,2,j})) + \alpha_h^2 \\
&= \alpha_w^2\int \sigma(u_1)\sigma(u_2)\mathcal{D}\mu(u_1,u_2) + \alpha_h^2 \qquad (\mu \sim \mathcal{N}(0,(q_{k,ab}))) \\
&= \alpha_w^2\int \sigma(u_1)\sigma(u_2)\mathcal{D}s_1\mathcal{D}s_2 + \alpha_h^2 \qquad (\mathcal{D}s_a \sim \mathcal{N}(0,1)))
\end{aligned}
$$

where $u_1 = \sqrt{q_{k-1,11}}s_1$, $u_2 = \sqrt{q_{k-1,22}}(c_{k-1,12}s_1 + \sqrt{1-c_{k-1,12}^2}s_2)$

Evolution of correlations near a fixed point q^*

Assuming $|\mathbf{z}|^2$ is near the (stable) fixed point q_* of the map ν:

$$c_{k,12} = \widetilde{\mathcal{C}}_{q^*}(c_{k-1,12}) \equiv \frac{1}{q^*}\mathcal{C}(c_{k-1,12}, q^*, q^* | \sigma_w, \sigma_h)$$

Exercise: $c^* = 1$ is a fixed point of the map $\widetilde{\mathcal{C}}_{q^*}$

Is $c^* = 1$ stable or unstable?

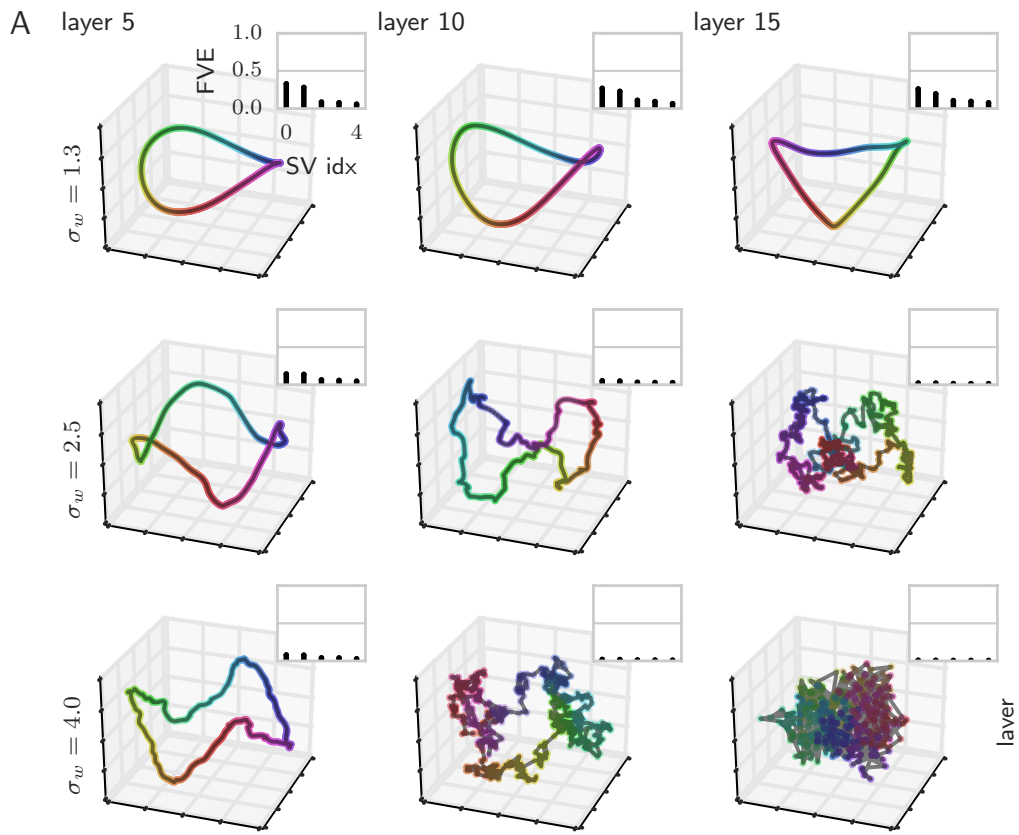
Exercise: Stability of $c^* = 1$ is determined by

$$\chi_1 = \left.\frac{\partial c_{k,12}}{\partial c_{k-1,12}}\right|_{c=1}$$

$$= \alpha_w^2 \int \left(\sigma'(\sqrt{q^*}s)\right)^2 \mathcal{D}s$$

(hint: $\int f(s) s \mathcal{D}s = \int f'(s) \mathcal{D}s$)

Stable vs. unstable evolution of a curve of input points:



Stable ("ordered") vs. unstable ("chaotic") phases

- Stable ($|\chi_1| < 1$):
  - different input vectors converge
  - initial details in the input get lost
  - deep network has low expressiveness
- Unstable ($|\chi_1| > 1$):
  - close input vectors tend to decorrelate
  - small changes in the input lead to major deviations
  - deep network has high expressiveness

In general, there are stable points $c^* \neq 1$; input vectors then tend to attain this fixed correlation.

Exercise: Find the explicit form of the correlation evolution map $\widetilde{\mathcal{C}}_{q^*}$ for the ReLU nonlinearity.

Depth of information propagation[12]

Exercise: Let $q_k = q^* + \delta q_k$. Then $\Delta q_{k+1} = a \Delta q_k + O(\Delta q_k^2)$, where

$$a = \chi_1 + \alpha_w^2 \int \sigma''(\sqrt{q^* s})\sigma(\sqrt{q^* s})\mathcal{D}s$$

Then $\Delta q_k \simeq e^{-k/\xi_q}$, where $\xi_q = -\frac{1}{\ln a}$ is the "characteristic depth scale of single input information propagation".

Exercise: Let c^* be a stable correlation fixed point and $c_{k,12} = c^* + \Delta c_k$. Then $\Delta c_{k+1} = b \Delta c_{k+1} + O(\Delta c_k^2)$, where

$$b = \alpha_w^2 \int \sigma'(u_1)\sigma'(u_2)\mathcal{D}s_1 \mathcal{D}s_2$$

with $u_1 = \sqrt{q_{k-1,11}} s_1$, $u_2 = \sqrt{q_{k-1,22}}(c_{k-1,12} s_1 + \sqrt{1 - c_{k-1,12}^2} s_2)$. If $c^* = 1$, then $b = \chi_1$.

Then $\Delta c_k \simeq e^{-k/\xi_c}$, where $\xi_c = -\frac{1}{\ln b}$ is the "characteristic depth scale of correlation information propagation".

[12] S. Schoenholz et al., Deep Information Propagation, arXiv:1611.01232

Error backpropagation

$$\begin{aligned}\left(\nabla_{W_k} L(\mathbf{W})\right)_{ij} = & \int d\mu(\mathbf{x}) \left[(F_{k-1}(\mathbf{x}))_j \times \right. \\ & \left. \times \left(F_k'(\mathbf{x}) \cdot W_{k+1}^* \dots F_{K-1}'(\mathbf{x}) \cdot W_K^* \cdot (\widetilde{f}(\mathbf{x}, \mathbf{W}) - f(\mathbf{x}))\right)_i\right] \\ = & \int d\mu(\mathbf{x}) \left[(F_{k-1}(\mathbf{x}))_j \times \right. \\ & \left. \times \left(T_k(\mathbf{x}) \cdots T_{K-1}(\mathbf{x}) \cdot (\widetilde{f}(\mathbf{x}, \mathbf{W}) - f(\mathbf{x}))\right)_i\right],\end{aligned}$$

where

$$T_k(\mathbf{x}) = F_k'(\mathbf{x}) \cdot W_{k+1}^*$$

In a deep network, the magnitude of $\nabla_{W_k} L(\mathbf{W})$ is mostly determined by $T_k(\mathbf{x}) \cdots T_{K-1}(\mathbf{x})$, since $\widetilde{f}(\mathbf{x}, \mathbf{W}) - f(\mathbf{x}) \asymp 1$ and $F_k'(\mathbf{x}) \asymp 1$

Error backpropagation

Let $\delta_k = (\delta_{k,i})_{i=1}^{d_k}$ be the backpropagated error:

$$\delta_{k,i} = (F_k'(\mathbf{x}) \cdot W_{k+1}^* \delta_{k+1})_i = \sigma'(z_{k,i}) \sum_{j=1}^{d_{k+1}} (W_{k+1})_{ji} \delta_{k+1,j}$$

Assume $z_{k,i}$ are independent of $\delta_{k+1,j}$, and recall that $(W_{k+1})_{ij}$ are i.i.d. with mean 0 and variance $\frac{\alpha_w^2}{d_k}$:

Assuming d_{k+1} is large:

$$\delta_{k,i}^2 \approx \mathbb{E}((\sigma'(z_{k,i}))^2) \sum_{j=1}^{d_{k+1}} \mathbb{E}((W_{k+1})_{ji}^2) \delta_{k+1,k}^2$$

$$\approx \int \left(\sigma'(\sqrt{q^*}s)\right)^2 \mathcal{D}s \cdot \frac{\alpha_w^2}{d_k} \sum_{j=1}^{d_{k+1}} \delta_{k+1,k}^2$$

$$= \chi_1 \frac{1}{d_k} \sum_{j=1}^{d_{k+1}} \delta_{k+1,k}^2$$

Vanishing and exploding gradients

Assume $d_k = d_{k+1}$, then

$$\|\delta_k\|^2 \approx \chi_1 \|\delta_{k+1}\|^2$$

- $\chi_1 < 1$: "vanishing gradient" ($\|\nabla_{W_k} L(\mathbf{W})\| \ll 1$ at small k)
- $\chi_1 > 1$: "exploding gradient" ($\|\nabla_{W_k} L(\mathbf{W})\| \gg 1$ at small k)

Optimal for learning: $\chi_1 \approx 1$ ("edge of chaos")

Hessian of ANN's loss: a random matrix theory approach[13]

The quadratic loss:

$$L(\mathbf{W}) = \frac{1}{2}\int \langle \widetilde{f}(\mathbf{x},\mathbf{W}) - f(\mathbf{x}), \widetilde{f}(\mathbf{x},\mathbf{W}) - f(\mathbf{x})\rangle d\mu(\mathbf{x})$$

The Hessian matrix:

$$H = D^2 L(\mathbf{W}))) = H_0 + H_1,$$

where

$$H_0 = \int \langle D_{\mathbf{W}}\widetilde{f}(\mathbf{x},\mathbf{W}), D_{\mathbf{W}}\widetilde{f}(\mathbf{x},\mathbf{W})\rangle d\mu(\mathbf{x}) \geq 0$$

$$H_1 = \int \langle \widetilde{f}(\mathbf{x},\mathbf{W}) - f(\mathbf{x}), D^2_{\mathbf{WW}}\widetilde{f}(\mathbf{x},\mathbf{W})\rangle d\mu(\mathbf{x})$$

[13] J. Pennington, Ya. Bahri, Geometry of Neural Network Loss Surfaces via Random Matrix Theory

Heuristics

- When loss is large, $\|\widetilde{f}(\mathbf{x}, \mathbf{W}) - f(\mathbf{x})\|$ is large, and H_1 makes a large (in magnitude) contribution to H. H_1 generally has both positive and negative eigenvalues, so critical points tend to be saddle points, and GD can easily decrease the loss.

- When loss is small, $\|\widetilde{f}(\mathbf{x}, \mathbf{W}) - f(\mathbf{x})\|$ is small and H_1 has a smaller contribution to H compared to H_0. H_0 does not have negative eigenvalues, so critical points have few directions of decreasing loss, and GD can get stuck.

Random matrix assumptions

Let $\mu(\mathbf{x}) = \sum_{s=1}^{m} \delta(\mathbf{x} - \mathbf{x}_s)$ (a finite training sample of size m)

Let $\mathbf{x} \in \mathbb{R}^{d_{\text{in}}}$, $f(\mathbf{x}) \in \mathbb{R}^{d_{\text{out}}}$, and $\mathbf{W} = (w_1, \ldots, w_n) \in \mathbb{R}^n$

$$(H_0)_{ij} = \frac{1}{m} \sum_{s=1}^{m} \sum_{r=1}^{d_{\text{out}}} \frac{\partial \widetilde{f}_r(\mathbf{x}_s, \mathbf{W})}{\partial w_i} \frac{\partial \widetilde{f}_r(\mathbf{x}_s, \mathbf{W})}{\partial w_j} = \frac{1}{m} (JJ^*)_{ij},$$

where $J \in \mathbb{R}^{n \times (m d_{out})}$ is the full Jacobian.

$$(H_1)_{ij} = \frac{1}{m} \sum_{s=1}^{m} \sum_{r=1}^{d_{\text{out}}} (\widetilde{f}_r(\mathbf{x}_s, \mathbf{W}) - f_r(\mathbf{x}_s)) \frac{\partial^2 \widetilde{f}_r(\mathbf{x}_s, \mathbf{W})}{\partial w_i \partial w_j}$$

Key assumptions:

- J is a random matrix with independent Gaussian entries $\sim \mathcal{N}(0, \alpha^2)$
- H_1 is a random matrix with independent Gaussian entries $\sim \mathcal{N}(0, 2\epsilon\alpha^2)$ (up to symmetricity constraint; ϵ reflects the magnitude of the errors $\widetilde{f}_r(\mathbf{x}_s, \mathbf{W}) - f_r(\mathbf{x}_s)$)

The strategy

Our goal: determine spectral properties of $H = H_0 + H_1$ for large m, n

- H_0 is described by **Wishart** ensemble; spec(H_0) is asymptotically given by Marchenko-Pastur distribution
- H_1 is described by **Wigner** ensemble; spec(H_1) is asymptotically given by Wigner semicircle law
- spec($H_0 + H_1$) can be derived under assumption of *free independence*
- spec($H_0 + H_1$) will depend on:
  - ϵ: determines the relative contribution of H_0 and H_1
  - $\phi = \frac{n}{md_{\text{out}}}$: ratio of the number of weights to the number of values to fit

Wigner ensemble[14]

- Can be real or complex
- Consists of symmetric $n \times n$ matrices $H = H^*$, so that $H_{ij} = \overline{H_{ji}}$
- Apart from the symmetricity constraint, all matrix entries are independent random variables
- The entries H_{ij} with $i \neq j$ are identically distributed for all i, j
- The entries H_{ii} are identically distributed for all i

Main example: Gaussian Orthogonal Ensemble (GOE or GOE(n))

$$H_{ij} \in \mathbb{R}, \quad H_{ij} = H_{ji} \sim \begin{cases} \mathcal{N}(0,1), & i \neq j \\ \mathcal{N}(0,2), & i = j \end{cases}$$

[14] Following T. Tao, Topics in random matrix theory

Properties of GOE

Exercise: Let A be a (non-symmetric) $\times$ matrix with independent entries $\sim \mathcal{N}(0, \frac{1}{2})$. Then $H = A + A^* \sim \mathrm{GOE}(n)$.

Exercise: The GOE distribution can be written as

$$Z_n^{-1} e^{-\frac{1}{4}\operatorname{tr}(HH^*)} \prod_{1 \le i \le j \le n} dH_{ij}$$

Exercise: GOE is invariant under conjugation with othogonal matrices:

$$H \sim \mathrm{GOE}(n) \Longrightarrow OHO^* \sim \mathrm{GOE}(n)$$

Wishart ensemble

- Can be real or complex
- Consists of symmetric $n \times n$ matrices $H = JJ^*$, where J is a (non-symmetric) $n \times p$ random matrix
- All matrix entries in J are i.i.d random variables

We will consider real matrices J with entries $\sim \mathcal{N}(0, 1)$

The Wigner semicircle law

$H_n \sim \mathrm{GOE}(n)$

Random spectral distribution:

$$\rho_{\frac{1}{\sqrt{n}}H_n} = \frac{1}{n}\sum_{k=1}^{n} \delta_{\frac{1}{\sqrt{n}}\lambda_k(H_n)},$$

where $\lambda_1(H_n) \leq \lambda_2(H_n) \leq \ldots \leq \lambda_n(H_n)$ are the eigenvalues of H_n

Theorem (Semicircle law)

With probability 1, $\lim_{n\to\infty} \rho_{\frac{1}{\sqrt{n}}H_n} = \rho_{sc}$, *where*

$$\rho_{sc}(x) = \frac{1}{2\pi}\sqrt{(4-x^2)_+}$$

Exercise: Why is $\frac{1}{\sqrt{n}}$ the right rescaling of eigenvalues?

Two different strategies of proof:

- Method of moments
- Stieltjes transform

Sketch of first proof: method of moments

It suffices to prove that for any k, with probability 1

$$\lim_{n\to\infty} \int x^k \rho_{\frac{1}{\sqrt{n}} H_n}(x)dx = \int x^k \rho_{sc}(x)dx$$

For slightly weaker convergence ("in probability") it suffices to prove that:

$$\lim_{n\to\infty} \mathbb{E}\Big(\int x^k \rho_{\frac{1}{\sqrt{n}} H_n}(x)dx \Big) = \int x^k \rho_{sc}(x)dx$$

$$\lim_{n\to\infty} \mathbb{V}\mathrm{ar}\Big(\int x^k \rho_{\frac{1}{\sqrt{n}} H_n}(x)dx \Big) = 0$$

By spectral decomposition:

$$\mathbb{E}\Big(\int x^k \rho_{\frac{1}{\sqrt{n}} H_n}(x)dx \Big) = \mathbb{E}\Big(\frac{1}{n} \mathrm{tr}\left[\Big(\frac{1}{\sqrt{n}} H_n \Big)^k \right] \Big)$$

Computation of moments

Denote $H_n = (\xi_{st})_{s,t=1}^n$

$k = 1$:

$$\mathbb{E}\Big(\frac{1}{n}\operatorname{tr}\Big[\Big(\frac{1}{\sqrt{n}}H_n\Big)\Big]\Big) = \frac{1}{n^{3/2}}\sum_{s=1}^{n}\mathbb{E}(\xi_{ss}) = 0$$

$k = 2$:

$$\begin{aligned}\mathbb{E}\Big(\frac{1}{n}\operatorname{tr}\Big[\Big(\frac{1}{\sqrt{n}}H_n\Big)^2\Big]\Big) &= \frac{1}{n^2}\sum_{s,t=1}^{n}\mathbb{E}(\xi_{st}\xi_{ts}) \\ &= \frac{1}{n^2}\sum_{s=1}^{n}\sum_{t:t\neq s}\mathbb{E}(\xi_{st}^2) + o(1) \\ &= 1 + o(1)\end{aligned}$$

$k = 3$:

$$\mathbb{E}\Big(\frac{1}{n}\operatorname{tr}\Big[\Big(\frac{1}{\sqrt{n}}H_n\Big)^3\Big]\Big) = \frac{1}{n^{5/2}}\sum_{s_1,s_2,s_3=1}^{n}\mathbb{E}(\xi_{s_1s_2}\xi_{s_2s_3}\xi_{s_3s_1}) = 0$$

Computation of moments

$k = 4$:

$$\begin{aligned}
\mathbb{E}\Big(\frac{1}{n}\operatorname{tr}\Big[\Big(\frac{1}{\sqrt{n}}H_n\Big)^4\Big]\Big) &= \frac{1}{n^3}\sum_{s_1,s_2,s_3,s_4=1}^{n}\mathbb{E}(\xi_{s_1s_2}\xi_{s_2s_3}\xi_{s_3s_4}\xi_{s_4s_1}) \\
&= \frac{1}{n^3}\sum_{s_1=1}^{n}\Bigg(\sum_{s_2:s_2\neq s_1}\ \sum_{s_3:s_3\neq s_1,s_2}\Big[\mathbb{E}(\xi_{s_1s_2}\xi_{s_2s_3}\xi_{s_3s_2}\xi_{s_2s_1}) \\
&\quad + \mathbb{E}(\xi_{s_1s_2}\xi_{s_2s_1}\xi_{s_1s_3}\xi_{s_3s_1})\Big]\Bigg) + o(1) \\
&= \frac{1}{n^3}\sum_{s_1=1}^{n}\Bigg(\sum_{s_2:s_2\neq s_1}\ \sum_{s_3:s_3\neq s_1,s_2}\Big[\mathbb{E}(\xi_{s_1s_2}^2)\mathbb{E}(\xi_{s_1s_3}^2) \\
&\quad + \mathbb{E}(\xi_{s_1s_2}^2)\mathbb{E}(\xi_{s_1s_3}^2)\Big]\Bigg) + o(1) \\
&= 2 + o(1)
\end{aligned}$$

Reduction to Dyck words

Exercise:

- Expectations vanish for odd k
- For even k, the leading term is equal to the number of closed paths of length k with each edge repeated exactly once, and traversing exactly $k/2+1$ vertices $s_1, s_2, \ldots, s_{k/2+1}$

Dyck word of length k: valid formulas made of $k/2$ symbols "(" and ")":

$$\begin{aligned} k=2: &\quad ()\\ k=4: &\quad (()),\ ()()\\ k=6: &\quad ((())),\ (()()),\ (())(),\ ()(()),\ ()()() \end{aligned}$$

Catalan numbers $C_{k/2}$: the number of Dyck words of length k

Exercise: For even k,

$$\mathbb{E}\Big(\frac{1}{n}\operatorname{tr}\Big[\Big(\frac{1}{\sqrt{n}}H_n\Big)^k\Big]\Big) = C_{k/2} + o(1)$$

Catalan numbers

Appear in many combinatoric problems:
https://en.wikipedia.org/wiki/Catalan_number

Exercise: $C_{t+1} = \sum_{r=0}^{t} C_r C_{t-r}$

Exercise: $C_t = \binom{2t}{t} - \binom{2t}{t+1} = \frac{1}{t+1}\binom{2t}{t} = \frac{(2t)!}{(t+1)!t!}$

Even moments of the semicircle distribution ρ:

$$\begin{aligned}
\int x^k \rho_{sc}(x) dx &= \frac{1}{2\pi} \int_{-2}^{2} x^k \sqrt{4 - x^2} dx \\
&= -\frac{1}{2\pi} \int_{0}^{\pi} (2\cos\phi)^k \sqrt{4 - 4\cos^2\phi} d(2\cos\phi) \\
&= \frac{1}{2\pi} \int_{0}^{\pi} (e^{i\phi} + e^{-i\phi})^k (e^{i\phi} - e^{-i\phi})^2 d\phi \\
&= \frac{1}{2}\left(\binom{k}{k/2+1} - 2\binom{k}{k/2} + \binom{k}{k/2-1} \right) = C_{k/2}
\end{aligned}$$

□

Sketch of second proof: Stieltjes transform

Stieltjes transform of the measure on $\mathbb{R}$ with density ρ:

$$s(z) := \int_{\mathbb{R}} \frac{1}{x-z}\rho(x)dx, \quad z \in \mathbb{C} \setminus \operatorname{supp}\rho$$

If $\rho = \rho_{\frac{1}{\sqrt{n}}H_n}$:

$$s_n(z) = \int_{\mathbb{R}} \frac{1}{x-z}\rho_{\frac{1}{\sqrt{n}}H_n}(x)dx = \frac{1}{n}\operatorname{tr}\left[\left(\frac{1}{\sqrt{n}}H_n - z\right)^{-1}\right]$$

Some properties:

- $s(z)$ is analytic in $\mathbb{C} \setminus \operatorname{supp}\rho$
- $|s(z)| \leq \frac{1}{|\operatorname{Im}(z)|}\int \rho(x)dx$
- ρ can be reconstructed from s:

$$s(a \pm ib) = \int_{\mathbb{R}} \frac{x-a \pm ib}{(x-a)^2+b^2}\rho(x)dx$$

$$\frac{s(a+ib)-s(a-ib)}{2\pi i} = \int_{\mathbb{R}} \frac{b}{(x-a)^2+b^2}\rho(x)dx \overset{b\to+0}{\longrightarrow} \rho(a)$$

"Reduction to predecessor"

Assuming concentration of measure, $s_n(z) = \mathbb{E}s_n(z) + o(1)$. Then

$$\begin{aligned} s_n(z) &= \mathbb{E}s_n(z) + o(1) \\ &= \frac{1}{n}\mathbb{E}\,\mathrm{tr}\left[\left(\frac{1}{\sqrt{n}}H_n - z\right)^{-1}\right] + o(1) \\ &= \mathbb{E}\left[\left(\frac{1}{\sqrt{n}}H_n - z\right)^{-1}\right]_{nn} + o(1) \\ &= \mathbb{E}\frac{1}{\frac{\xi_{nn}}{\sqrt{n}} - z - \frac{1}{n}X^*(\frac{1}{\sqrt{n}}H_{n-1} - z)^{-1}X} + o(1), \end{aligned}$$

where $X = (\xi_{1n}, \xi_{2n}, \ldots, \xi_{n-1,n})^t$

Exercise (Schur complement): Let A_n be a $n \times n$ matrix in the block form $A_n = \begin{pmatrix} A_{n-1} & B \\ C^* & a_{nn} \end{pmatrix}$. Then $[A_n^{-1}]_{nn} = \frac{1}{a_{nn} - C^* A_{n-1}^{-1} B}$.

Quadratic equation for $s(z)$

X and H_{n-1} are independent and normal. By concentration of measure:

$$\begin{aligned}\frac{1}{n}X^*\left(\frac{1}{\sqrt{n}}H_{n-1}-z\right)^{-1}X &= \mathbb{E}\frac{1}{n}X^*\left(\frac{1}{\sqrt{n}}H_{n-1}-z\right)^{-1}X+o(1)\\ &= \mathbb{E}\left[\left(\frac{1}{\sqrt{n-1}}H_{n-1}-z\right)^{-1}\right]_{nn}+o(1)\\ &= s_{n-1}(z)+o(1)\end{aligned}$$

Then

$$s_n(z)=\frac{1}{-z-s_{n-1}(z)}+o(1)$$

Assuming the limit $s(z)=\lim_{n\to\infty}s_n(z)$ exists,

$$s(z)=\frac{1}{-z-s(z)}$$

The limiting $s(z)$ and $\rho(x)$

$$s(z) = \frac{z \pm \sqrt{z^2 - 4}}{2}$$

$s(z)$ is analytic outside of the cut $[-2, 2]$

$$\rho(x) = \lim_{b \to +0} \frac{s(a + ib) - s(a - ib)}{2\pi i} = \frac{1}{2\pi} \sqrt{4 - x^2} = \rho_{sc}(x)$$

□