Fundamental Bounds on Learning Performance in Neural Circuits

Raman, Rotondo & O'Leary
Presented by Rylan Schaeffer

November 22, 2019

Motivation

- Three weeks ago, Mikail discussed Stable Memory with Unstable Synapses [10]

Motivation

- Three weeks ago, Mikail discussed Stable Memory with Unstable Synapses [10]
- Size and connectome of biological circuit change during learning

Motivation

- Three weeks ago, Mikail discussed Stable Memory with Unstable Synapses [10]
- Size and connectome of biological circuit change during learning
- Synapses in biological networks lack persistence, undergoing significant turnover [3, 9, 8], with magnitude rivaling Hebbian plasticity [5]

Motivation

- Three weeks ago, Mikail discussed Stable Memory with Unstable Synapses [10]
- Size and connectome of biological circuit change during learning
- Synapses in biological networks lack persistence, undergoing significant turnover [3, 9, 8], with magnitude rivaling Hebbian plasticity [5]
- Across species and regions, neurons frequently make multiple synaptic connections to same postsynaptic neuron [1, 2, 4, 6]

Motivation

- Three weeks ago, Mikail discussed Stable Memory with Unstable Synapses [10]
- Size and connectome of biological circuit change during learning
- Synapses in biological networks lack persistence, undergoing significant turnover [3, 9, 8], with magnitude rivaling Hebbian plasticity [5]
- Across species and regions, neurons frequently make multiple synaptic connections to same postsynaptic neuron [1, 2, 4, 6]
- What is the role of these processes? What (dis)advantages do these phenomena confer on biological circuits? [7]

Overview

- How does increasing neurons and/or adding redundant synapses affect learning?

Overview

- How does increasing neurons and/or adding redundant synapses affect learning?
- Consider gradient descent on error function, where weight change is comprised of three components: (a) task-specific gradient, (b) task-independent and (c) random noise

Overview

- How does increasing neurons and/or adding redundant synapses affect learning?
- Consider gradient descent on error function, where weight change is comprised of three components: (a) task-specific gradient, (b) task-independent and (c) random noise
- Rate of error reduction depends on interaction between gradient and Hessian

Overview

- How does increasing neurons and/or adding redundant synapses affect learning?
- Consider gradient descent on error function, where weight change is comprised of three components: (a) task-specific gradient, (b) task-independent and (c) random noise
- Rate of error reduction depends on interaction between gradient and Hessian
- For a given task, there is an optimal network size that maximizes rate of error reduction

Overview

- How does increasing neurons and/or adding redundant synapses affect learning?
- Consider gradient descent on error function, where weight change is comprised of three components: (a) task-specific gradient, (b) task-independent and (c) random noise
- Rate of error reduction depends on interaction between gradient and Hessian
- For a given task, there is an optimal network size that maximizes rate of error reduction
- Below optimal size, increasing network size causes network to learn faster by minimizing effect of curvature

Model

- Synaptic weights $w(t) \in \mathbb{R}^n$

Model

- Synaptic weights $w(t) \in \mathbb{R}^n$
- (Noisy) error function: $F[w(t)]$

Model

- Synaptic weights $w(t) \in \mathbb{R}^n$
- (Noisy) error function: $F[w(t)]$
- Consider network receives error feedback at $t = 0$, but no additional feedback till time $t = T$

Model

- Synaptic weights $w(t) \in \mathbb{R}^n$
- (Noisy) error function: $F[w(t)]$
- Consider network receives error feedback at $t = 0$, but no additional feedback till time $t = T$
- Define "learning rate" k over interval $t \in [0, T]$:

$$F[w(T)] = (1 - kT)F[w(0)]$$

Model

- Synaptic weights $w(t) \in \mathbb{R}^n$
- (Noisy) error function: $F[w(t)]$
- Consider network receives error feedback at $t = 0$, but no additional feedback till time $t = T$
- Define "learning rate" k over interval $t \in [0, T]$:

$$F[w(T)] = (1 - kT)F[w(0)]$$

- Goal: maximize k to learn!

Model

- Synaptic weights $w(t) \in \mathbb{R}^n$
- (Noisy) error function: $F[w(t)]$
- Consider network receives error feedback at $t = 0$, but no additional feedback till time $t = T$
- Define "learning rate" k over interval $t \in [0, T]$:

$$F[w(T)] = (1 - kT)F[w(0)]$$

- Goal: maximize k to learn!
- Notation: $c\hat{d}ot$ denotes normalized vector

Error Reduction Depends on Gradient, Hessian

How do gradient, Hessian affect the "learning rate" k?

$$F[w(T)] = [1 - kT]F[w(0)]$$

Error Reduction Depends on Gradient, Hessian

How do gradient, Hessian affect the "learning rate" k?

$$F[w(T)] = [1 - kT]F[w(0)]$$

$$-kTF[w(0)] = F[w(T)] - F[w(0)]$$

Error Reduction Depends on Gradient, Hessian

How do gradient, Hessian affect the "learning rate" k?

$$\begin{aligned} F[w(T)] &= [1 - kT]F[w(0)] \\ -kTF[w(0)] &= F[w(T)] - F[w(0)] \\ &= \int_{t=0}^{T} dt\, \frac{d}{dt} F[w(t)] \end{aligned}$$

Error Reduction Depends on Gradient, Hessian

How do gradient, Hessian affect the "learning rate" k?

$$
\begin{aligned}
F[w(T)] &= [1 - kT]F[w(0)] \\
-kTF[w(0)] &= F[w(T)] - F[w(0)] \\
&= \int_{t=0}^{T} dt \, \frac{d}{dt} F[w(t)] \\
&= T \left\langle \nabla_w F[w(t)]^T \frac{dw}{dt} \right\rangle_{t \sim \mathrm{Unif}(0,T)}
\end{aligned}
$$

Error Reduction Depends on Gradient, Hessian

How do gradient, Hessian affect the "learning rate" k?

$$
\begin{aligned}
F[w(T)] &= [1 - kT]F[w(0)] \\
-kTF[w(0)] &= F[w(T)] - F[w(0)] \\
&= \int_{t=0}^{T} dt \, \frac{d}{dt} F[w(t)] \\
&= T \left\langle \nabla_w F[w(t)]^T \frac{dw}{dt} \right\rangle_{t \sim \mathrm{Unif}(0,T)}
\end{aligned}
$$

Define $\dot{w}_T = \frac{w(T) - w(0)}{T}$ and Taylor-series expand around $F[w(0)]$

Error Reduction Depends on Gradient, Hessian

How do gradient, Hessian affect the "learning rate" k?

$$\begin{aligned} F[w(T)] &= [1 - kT]F[w(0)] \\ -kTF[w(0)] &= F[w(T)] - F[w(0)] \\ &= \int_{t=0}^{T} dt \, \frac{d}{dt} F[w(t)] \\ &= T \left\langle \nabla_w F[w(t)]^T \frac{dw}{dt} \right\rangle_{t \sim \mathrm{Unif}(0,T)} \end{aligned}$$

Define $\dot{w}_T = \frac{w(T) - w(0)}{T}$ and Taylor-series expand around $F[w(0)]$

$$-kTF[w(0)] = T\nabla_w F[w(t)]^T \dot{w}_T + \frac{1}{2} T^2 \dot{w}_T^T \nabla_w^2 F[w(0)] \dot{w}_T + O(T^3)$$

Error Reduction Depends on Gradient, Hessian

How do gradient, Hessian affect the "learning rate" k?

$$\begin{aligned}
F[w(T)] &= [1 - kT]F[w(0)] \\
-kTF[w(0)] &= F[w(T)] - F[w(0)] \\
&= \int_{t=0}^{T} dt\, \frac{d}{dt} F[w(t)] \\
&= T \left\langle \nabla_w F[w(t)]^T \frac{dw}{dt} \right\rangle_{t \sim \mathrm{Unif}(0,T)}
\end{aligned}$$

Define $\dot{w}_T = \frac{w(T) - w(0)}{T}$ and Taylor-series expand around $F[w(0)]$

$$-kTF[w(0)] = T \nabla_w F[w(t)]^T \dot{w}_T + \frac{1}{2} T^2 \dot{w}_T^T \nabla_w^2 F[w(0)] \dot{w}_T + O(T^3)$$

$$k \approx -\frac{||\nabla F[w(0)]||_2}{F[w(0)]} \left[\dot{w}_T^T \nabla \hat{F}[w(0)] + \frac{T ||\dot{w}_T||_2^2}{2||\nabla F[w(0)]||_2} \hat{\dot{w}}_T^T \nabla^2 F[w(0)] \hat{\dot{w}}_T \right]$$

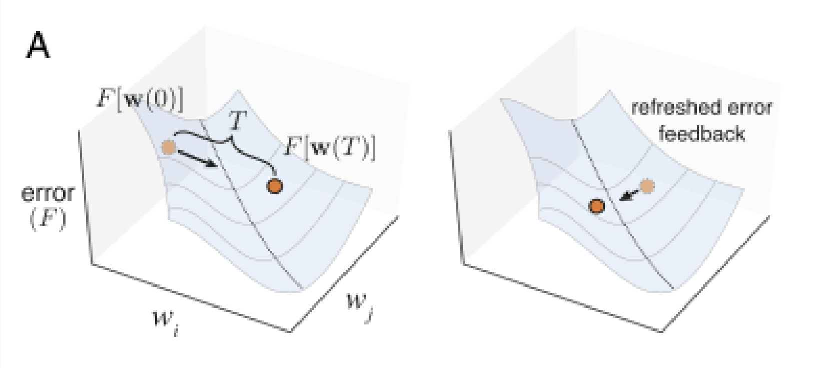
Error Reduction Depends on Gradient, Hessian

$$k \approx -\frac{||\nabla F[w(0)]||_2}{F[w(0)]}\left[\dot{w}_T^T \nabla \hat{F}[w(0)] + \frac{T||\dot{w}_T||_2^2}{2||\nabla F[w(0)]||_2}\dot{\hat{w}}_T^T \nabla^2 F[w(0)]\dot{\hat{w}}_T\right]$$

Error Reduction Depends on Gradient, Hessian

$$k \approx -\frac{||\nabla F[w(0)]||_2}{F[w(0)]}\left[\dot{w}_T^T \nabla \hat{F}[w(0)] + \frac{T||\dot{w}_T||_2^2}{2||\nabla F[w(0)]||_2}\dot{\hat{w}}_T^T \nabla^2 F[w(0)]\dot{\hat{w}}_T\right]$$

Curvature competes with gradient to accelerate, slow or reverse learning. Fig 3A:



Model (Continued)

- Suppose weight change comprised of 3 components:

Model (Continued)

- Suppose weight change comprised of 3 components:
- "Task relevant plasticity" (direction of error gradient): $\nabla_w \hat{F}[w(0)]$

Model (Continued)

- Suppose weight change comprised of 3 components:
- "Task relevant plasticity" (direction of error gradient): $\nabla_w \hat{F}[w(0)]$
- "Task irrelevant plasticity" e.g. homeostatic plasticity, learning on other tasks: $\hat{n}_2$

Model (Continued)

- Suppose weight change comprised of 3 components:
- "Task relevant plasticity" (direction of error gradient): $\nabla_w \hat{F}[w(0)]$
- "Task irrelevant plasticity" e.g. homeostatic plasticity, learning on other tasks: $\hat{n}_2$
- "Synaptic Noise" i.e. Additive, iid white noise at each synapse: $\hat{n}_3$

Model (Continued)

- Suppose weight change comprised of 3 components:
- "Task relevant plasticity" (direction of error gradient): $\nabla_w \hat{F}[w(0)]$
- "Task irrelevant plasticity" e.g. homeostatic plasticity, learning on other tasks: $\hat{n}_2$
- "Synaptic Noise" i.e. Additive, iid white noise at each synapse: $\hat{n}_3$
- Assume network has no second-order information!

Model (Continued)

- Suppose weight change comprised of 3 components:
- "Task relevant plasticity" (direction of error gradient): $\nabla_w \hat{F}[w(0)]$
- "Task irrelevant plasticity" e.g. homeostatic plasticity, learning on other tasks: $\hat{n}_2$
- "Synaptic Noise" i.e. Additive, iid white noise at each synapse: $\hat{n}_3$
- Assume network has no second-order information!
- Writing the weight change:

$$\dot{w}_T = -\gamma_1 \nabla \hat{F}[w(0)] + \gamma_2 \hat{n}_2 + \gamma_3 \sqrt{\frac{N}{T}} \hat{n}_3$$

Model (Continued)

$$\dot{w}_T = -\gamma_1 \nabla \hat{F}[w(0)] + \gamma_2 \hat{n}_2 + \gamma_3 \sqrt{\frac{N}{T}} \hat{n}_3$$

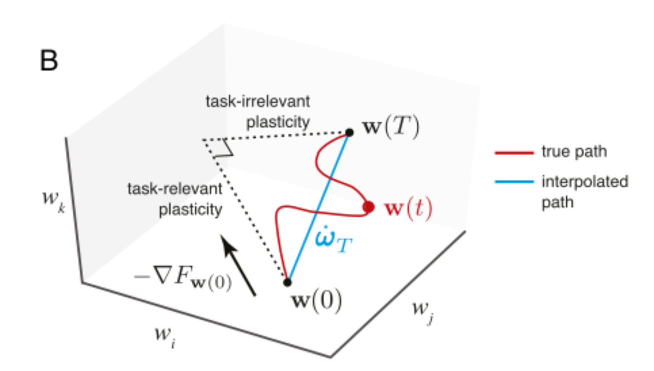


Fig 3. Synaptic noise not pictured!

Model (Continued)

$$\dot{w}_T = -\gamma_1 \nabla \hat{F}[w(0)] + \gamma_2 \hat{n}_2 + \gamma_3 \sqrt{\frac{N}{T}} \hat{n}_3$$

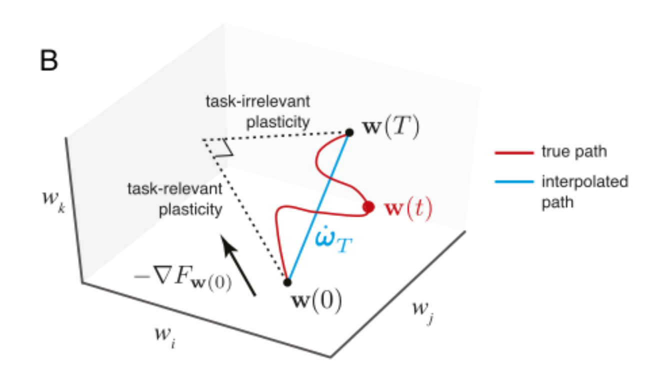


Fig 3. Synaptic noise not pictured! How does each factor affect k?

Weight Change Effect on Learning Rate

$$k \approx -\frac{||\nabla F||_2}{F}\left[\dot{w}_T^T \nabla \hat{F} + T\frac{||\dot{w}_T||_2^2}{2||\nabla F||_2}\dot{\hat{w}}_T^T \nabla^2 F \dot{\hat{w}}_T\right]$$

Weight Change Effect on Learning Rate

$$k \approx -\frac{||\nabla F||_2}{F}\left[\dot{w}_T^T \nabla \hat{F} + T\frac{||\dot{w}_T||_2^2}{2||\nabla F||_2}\dot{\hat{w}}_T^T \nabla^2 F \dot{\hat{w}}_T\right]$$

Assume (1) $n_2, n_3, \nabla F$ uncorrelated; (2) n_2, n_3 independent from $\nabla^2 F[w]$ i.e. $\langle n_i^T \nabla^2 F n_i \rangle_{\hat{n}_2,\hat{n}_3} = \frac{\mathrm{Tr}(\nabla^2 F)}{N}$.

Weight Change Effect on Learning Rate

$$k \approx -\frac{||\nabla F||_2}{F}\left[\dot{w}_T^T \nabla \hat{F} + T\frac{||\dot{w}_T||_2^2}{2||\nabla F||_2}\dot{\hat{w}}_T^T \nabla^2 F \dot{\hat{w}}_T\right]$$

Assume (1) $n_2, n_3, \nabla F$ uncorrelated; (2) n_2, n_3 independent from $\nabla^2 F[w]$ i.e. $\langle n_i^T \nabla^2 F n_i \rangle_{\hat{n}_2,\hat{n}_3} = \frac{\mathrm{Tr}(\nabla^2 F)}{N}$.

$\langle \bullet \rangle_{\hat{n}_2,\hat{n}_3}$

Weight Change Effect on Learning Rate

$$k \approx -\frac{||\nabla F||_2}{F}\left[\dot{w}_T^T \nabla \hat{F} + T\frac{||\dot{w}_T||_2^2}{2||\nabla F||_2}\dot{\hat{w}}_T^T \nabla^2 F \dot{\hat{w}}_T\right]$$

Assume (1) $n_2, n_3, \nabla F$ uncorrelated; (2) n_2, n_3 independent from $\nabla^2 F[w]$ i.e. $\langle n_i^T \nabla^2 F n_i \rangle_{\hat{n}_2,\hat{n}_3} = \frac{\mathrm{Tr}(\nabla^2 F)}{N}$.

$$\langle \bullet \rangle_{\hat{n}_2,\hat{n}_3} = (-\gamma_1 \nabla \hat{F} + \gamma_2 \hat{n}_2 + \gamma_3 \sqrt{\frac{N}{T}} \hat{n}_3)^T \nabla \hat{F}$$

Weight Change Effect on Learning Rate

$$k \approx -\frac{||\nabla F||_2}{F}\left[\dot{w}_T^T \nabla \hat{F} + T\frac{||\dot{w}_T||_2^2}{2||\nabla F||_2}\dot{\hat{w}}_T^T \nabla^2 F \dot{\hat{w}}_T\right]$$

Assume (1) $n_2, n_3, \nabla F$ uncorrelated; (2) n_2, n_3 independent from $\nabla^2 F[w]$ i.e. $\langle n_i^T \nabla^2 F n_i \rangle_{\hat{n}_2,\hat{n}_3} = \frac{\mathrm{Tr}(\nabla^2 F)}{N}$.

$$\begin{aligned}\langle \bullet \rangle_{\hat{n}_2,\hat{n}_3} &= (-\gamma_1 \nabla \hat{F} + \gamma_2 \hat{n}_2 + \gamma_3 \sqrt{\frac{N}{T}}\hat{n}_3)^T \nabla \hat{F} \\ &= -\gamma_1 + 0 + 0\end{aligned}$$

Weight Change Effect on Learning Rate

$$k \approx -\frac{||\nabla F||_2}{F}\left[\dot{w}_T^T \nabla \hat{F} + T\frac{||\dot{w}_T||_2^2}{2||\nabla F||_2}\hat{\dot{w}}_T^T \nabla^2 F \hat{\dot{w}}_T\right]$$

Assume (1) $n_2, n_3, \nabla F$ uncorrelated; (2) n_2, n_3 independent from $\nabla^2 F[w]$ i.e. $\langle n_i^T \nabla^2 F n_i \rangle_{\hat{n}_2, \hat{n}_3} = \frac{\mathrm{Tr}(\nabla^2 F)}{N}$.

$$\begin{aligned} \langle \bullet \rangle_{\hat{n}_2,\hat{n}_3} &= (-\gamma_1 \nabla \hat{F} + \gamma_2 \hat{n}_2 + \gamma_3 \sqrt{\frac{N}{T}}\hat{n}_3)^T \nabla \hat{F} \\ &= -\gamma_1 + 0 + 0 \end{aligned}$$

$$\langle \bullet \rangle_{\hat{n}_2,\hat{n}_3}$$

Weight Change Effect on Learning Rate

$$k \approx -\frac{||\nabla F||_2}{F}\left[\dot{w}_T^T \nabla \hat{F} + T\frac{||\dot{w}_T||_2^2}{2||\nabla F||_2}\hat{\dot{w}}_T^T \nabla^2 F \hat{\dot{w}}_T\right]$$

Assume (1) $n_2, n_3, \nabla F$ uncorrelated; (2) n_2, n_3 independent from $\nabla^2 F[w]$ i.e. $\langle n_i^T \nabla^2 F n_i \rangle_{\hat{n}_2,\hat{n}_3} = \frac{\mathrm{Tr}(\nabla^2 F)}{N}$.

$$\begin{aligned}
\langle \bullet \rangle_{\hat{n}_2,\hat{n}_3} &= (-\gamma_1 \nabla \hat{F} + \gamma_2 \hat{n}_2 + \gamma_3 \sqrt{\frac{N}{T}}\hat{n}_3)^T \nabla \hat{F} \\
&= -\gamma_1 + 0 + 0 \\
\langle \bullet \rangle_{\hat{n}_2,\hat{n}_3} &= \gamma_1^2 \nabla \hat{F}^T \nabla^2 F \nabla \hat{F} + \frac{\gamma_2^2 \hat{n}_2 \nabla^2 F \hat{n}_2 + \gamma_3^2 \hat{n}_3 \nabla^2 F \hat{n}_3}{2||\nabla F||_2}
\end{aligned}$$

Weight Change Effect on Learning Rate

$$k \approx -\frac{||\nabla F||_2}{F}\Big[\dot{w}_T^T \nabla \hat{F} + T\frac{||\dot{w}_T||_2^2}{2||\nabla F||_2}\hat{\dot{w}}_T^T \nabla^2 F \hat{\dot{w}}_T\Big]$$

Assume (1) $n_2, n_3, \nabla F$ uncorrelated; (2) n_2, n_3 independent from $\nabla^2 F[w]$ i.e. $\langle n_i^T \nabla^2 F n_i \rangle_{\hat{n}_2,\hat{n}_3} = \frac{\mathrm{Tr}(\nabla^2 F)}{N}$.

$$\begin{aligned}
\langle \bullet \rangle_{\hat{n}_2,\hat{n}_3} &= (-\gamma_1 \nabla \hat{F} + \gamma_2 \hat{n}_2 + \gamma_3 \sqrt{\frac{N}{T}}\hat{n}_3)^T \nabla \hat{F} \\
&= -\gamma_1 + 0 + 0 \\
\langle \bullet \rangle_{\hat{n}_2,\hat{n}_3} &= \gamma_1^2 \nabla \hat{F}^T \nabla^2 F \nabla \hat{F} + \frac{\gamma_2^2 \hat{n}_2 \nabla^2 F \hat{n}_2 + \gamma_3^2 \hat{n}_3 \nabla^2 F \hat{n}_3}{2||\nabla F||_2} \\
&= \gamma_1^2 \nabla \hat{F}^T \nabla^2 F \nabla \hat{F} + \frac{\mathrm{Tr}(\nabla^2 F)}{2||\nabla F||_2}\Big[\frac{\gamma_2^2}{N} + \frac{\gamma_3^2}{T}\Big]
\end{aligned}$$

Weight Change Effect on Learning Rate

$$\langle k \rangle_{\hat{n}_2,\hat{n}_3} \approx -\frac{-||\nabla F||_2}{F}\left[-\gamma_1 + T\gamma_1^2 \nabla \hat{F}^T \nabla^2 F \nabla \hat{F} + T\frac{\mathrm{Tr}(\nabla^2 F)}{2||\nabla F||_2^2}\left[\frac{\gamma_2^2}{N} + \frac{\gamma_3^2}{T}\right]\right]$$

- Authors call $G_F[\dot{\hat{w}}_T] = \frac{\bullet}{||\dot{w}_T||_2^2}$ the "local task difficulty"

Weight Change Effect on Learning Rate

$$\langle k \rangle_{\hat{n}_2,\hat{n}_3} \approx -\frac{-||\nabla F||_2}{F}\left[-\gamma_1 + T\gamma_1^2 \nabla \hat{F}^T \nabla^2 F \nabla \hat{F} + T\frac{\mathrm{Tr}(\nabla^2 F)}{2||\nabla F||_2^2}\left[\frac{\gamma_2^2}{N} + \frac{\gamma_3^2}{T}\right]\right]$$

- Authors call $G_F[\dot{\hat{w}}_T] = \frac{\bullet}{||\dot{w}_T||_2^2}$ the "local task difficulty"
- Authors argue that $\mathrm{sign}(G_F[\dot{\hat{w}}_T]) = \mathrm{sign}(\mathrm{Tr}(\nabla^2 F)) > 0$. Why?

Weight Change Effect on Learning Rate

$$\langle k \rangle_{\hat{n}_2,\hat{n}_3} \approx -\frac{-||\nabla F||_2}{F}\left[-\gamma_1 + T\gamma_1^2 \nabla \hat{F}^T \nabla^2 F \nabla \hat{F} + T\frac{\mathrm{Tr}(\nabla^2 F)}{2||\nabla F||_2^2}\left[\frac{\gamma_2^2}{N} + \frac{\gamma_3^2}{T}\right]\right]$$

- Authors call $G_F[\dot{\hat{w}}_T] = \frac{\bullet}{||\dot{w}_T||_2^2}$ the "local task difficulty"
- Authors argue that $\mathrm{sign}(G_F[\dot{\hat{w}}_T]) = \mathrm{sign}(\mathrm{Tr}(\nabla^2 F)) > 0$. Why?
- Learning occurs when:

$$\langle k \rangle_{\hat{n}_2,\hat{n}_3} > 0 \Rightarrow G_F[\dot{\hat{w}}_T] < \frac{\gamma_1}{T(\gamma_1^2 + \gamma_2^2 + \gamma_3^2 \frac{N}{T})}$$

Weight Change Effect on Learning Rate

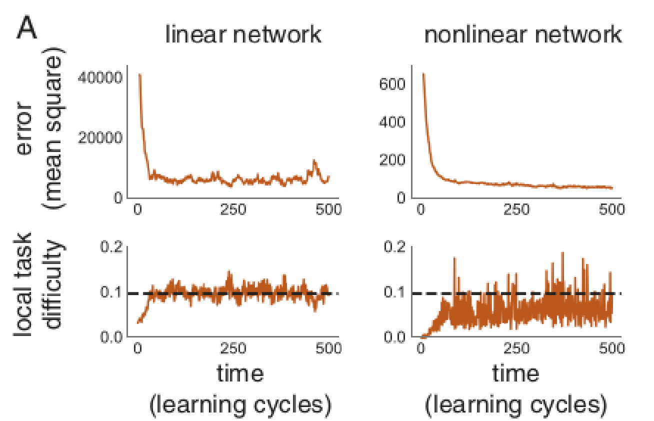
$$\langle k \rangle_{\hat{n}_2,\hat{n}_3} \approx -\frac{-||\nabla F||_2}{F}\left[-\gamma_1 + T\gamma_1^2 \nabla \hat{F}^T \nabla^2 F \nabla \hat{F} + T\frac{\mathrm{Tr}(\nabla^2 F)}{2||\nabla F||_2^2}\left[\frac{\gamma_2^2}{N} + \frac{\gamma_3^2}{T}\right]\right]$$

- Authors call $G_F[\dot{\hat{w}}_T] = \frac{\bullet}{||\dot{w}_T||_2^2}$ the "local task difficulty"
- Authors argue that $\mathrm{sign}(G_F[\dot{\hat{w}}_T]) = \mathrm{sign}(\mathrm{Tr}(\nabla^2 F)) > 0$. Why?
- Learning occurs when:

$$\langle k \rangle_{\hat{n}_2,\hat{n}_3} > 0 \Rightarrow G_F[\dot{\hat{w}}_T] < \frac{\gamma_1}{T(\gamma_1^2 + \gamma_2^2 + \gamma_3^2 \frac{N}{T})}$$

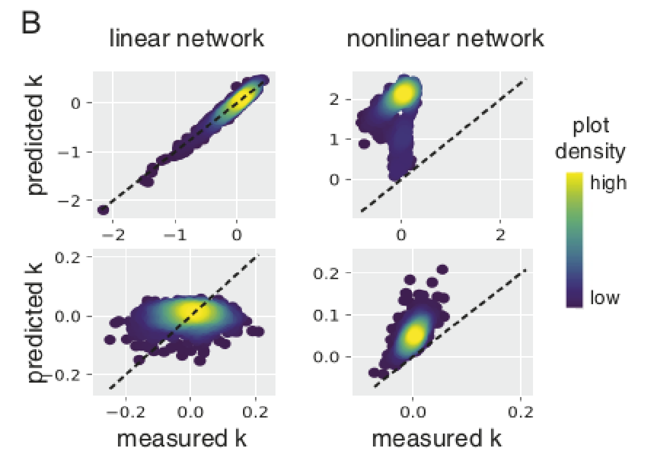
- There exists optimal size N^* that maximizes $\langle k \rangle_{\hat{n}_2,\hat{n}_3}$

Local Task Difficulty



Local Task Difficulty

Top: Low intrinsic noise ($\gamma_3 = 0.05$). Bottom: High intrinsic noise ($\gamma_3 = 0.1$).



Optimal Linear Network Size

- Student-teacher framework with $W \in \mathbb{R}^{o \times i}$:

$$y^* = W^* x \qquad y = Wx \qquad F(W) = \frac{1}{2}||y^* - y||_2^2$$

Optimal Linear Network Size

- Student-teacher framework with $W \in \mathbb{R}^{o \times i}$:

$$y^* = W^* x \qquad y = Wx \qquad F(W) = \frac{1}{2}||y^* - y||_2^2$$

- Choose

Optimal Linear Network Size

- Student-teacher framework with $W \in \mathbb{R}^{o \times i}$:

$$y^* = W^* x \qquad y = Wx \qquad F(W) = \frac{1}{2}||y^* - y||_2^2$$

- Choose
  - $c_1, c_2 > 1$

Optimal Linear Network Size

- Student-teacher framework with $W \in \mathbb{R}^{o \times i}$:

$$y^* = W^* x \qquad y = Wx \qquad F(W) = \frac{1}{2}||y^* - y||_2^2$$

- Choose
  - $c_1, c_2 > 1$
  - Two semi-orthogonal matrices $B \in \mathbb{R}^{c_1 i \times i}, D \in \mathbb{R}^{c_2 o \times o}$

Optimal Linear Network Size

- Student-teacher framework with $W \in \mathbb{R}^{o \times i}$:

$$y^* = W^* x \qquad y = Wx \qquad F(W) = \frac{1}{2}||y^* - y||_2^2$$

- Choose
  - $c_1, c_2 > 1$
  - Two semi-orthogonal matrices $B \in \mathbb{R}^{c_1 i \times i}, D \in \mathbb{R}^{c_2 o \times o}$
  - Random $W' \in \mathbb{R}^{c_2 o \times c_1 i}$ such that $W = D^T W' B$

Optimal Linear Network Size

- Student-teacher framework with $W \in \mathbb{R}^{o \times i}$:

$$y^* = W^* x \qquad y = Wx \qquad F(W) = \frac{1}{2} ||y^* - y||_2^2$$

- Choose
  - $c_1, c_2 > 1$
  - Two semi-orthogonal matrices $B \in \mathbb{R}^{c_1 i \times i}, D \in \mathbb{R}^{c_2 o \times o}$
  - Random $W' \in \mathbb{R}^{c_2 o \times c_1 i}$ such that $W = D^T W' B$
- Interpretation: add neurons or redundant synapses

Optimal Linear Network Size

- Student-teacher framework with $W \in \mathbb{R}^{o \times i}$:

$$y^* = W^* x \qquad y = Wx \qquad F(W) = \frac{1}{2}||y^* - y||_2^2$$

- Choose
  - $c_1, c_2 > 1$
  - Two semi-orthogonal matrices $B \in \mathbb{R}^{c_1 i \times i}, D \in \mathbb{R}^{c_2 o \times o}$
  - Random $W' \in \mathbb{R}^{c_2 o \times c_1 i}$ such that $W = D^T W' B$
- Interpretation: add neurons or redundant synapses
- Replace W with $D^T W' B$

$$y = D^T W' B x$$

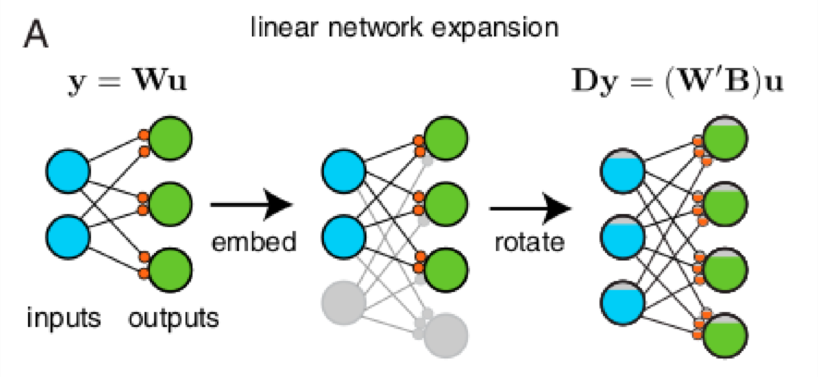
$$F[W'] = F[W]$$

$$||F[W']||_F^2 = ||F[W]||_F^2$$

$$\mathrm{Tr}(\nabla^2 F[W']) = c_2 \,\mathrm{Tr}(\nabla^2 F[W])$$

Optimal Linear Network Size

$$y = D^T W' Bx \iff Dy = W' Bx$$



Optimal Linear Network Size

- Define $N = io, \tilde{N} = c_1 c_2 io$

Optimal Linear Network Size

- Define $N = io, \tilde{N} = c_1 c_2 io$
- Compare learning rate $k(N)$ vs $k(\tilde{N})$:

$$\langle k(N) \rangle \approx \frac{-||\nabla F||_2}{F} \left[-\gamma_1 + T\gamma_1^2 \nabla \hat{F}^T \nabla^2 F \nabla \hat{F} + T \frac{\mathrm{Tr}(\nabla^2 F)}{2||\nabla F||_2^2} \left[\frac{\gamma_2^2}{N} + \frac{\gamma_3^2}{T} \right] \right]$$

Optimal Linear Network Size

- Define $N = io, \tilde{N} = c_1 c_2 io$
- Compare learning rate $k(N)$ vs $k(\tilde{N})$:

$$\langle k(N)\rangle \approx \frac{-||\nabla F||_2}{F}\left[-\gamma_1 + T\gamma_1^2 \nabla \hat{F}^T \nabla^2 F \nabla \hat{F} + T\frac{\mathrm{Tr}(\nabla^2 F)}{2||\nabla F||_2^2}\left[\frac{\gamma_2^2}{N} + \frac{\gamma_3^2}{T}\right]\right]$$

- If $\nabla F[W']$ projects equally onto Hessian eigenvectors:

$$\nabla \hat{F}[W']^T \nabla^2 F[W'] \nabla \hat{F}[W'] \approx c_2 \nabla \hat{F}[W]^T \nabla^2 F[W] \nabla \hat{F}[W]$$

Optimal Linear Network Size

- Define $N = io, \tilde{N} = c_1 c_2 io$
- Compare learning rate $k(N)$ vs $k(\tilde{N})$:

$$\langle k(N) \rangle \approx \frac{-||\nabla F||_2}{F} \left[-\gamma_1 + T \gamma_1^2 \nabla \hat{F}^T \nabla^2 F \nabla \hat{F} + T \frac{\mathrm{Tr}(\nabla^2 F)}{2||\nabla F||_2^2} \left[\frac{\gamma_2^2}{N} + \frac{\gamma_3^2}{T} \right] \right]$$

- If $\nabla F[W']$ projects equally onto Hessian eigenvectors:

$$\nabla \hat{F}[W']^T \nabla^2 F[W'] \nabla \hat{F}[W'] \approx c_2 \nabla \hat{F}[W]^T \nabla^2 F[W] \nabla \hat{F}[W]$$

- Previously:

$$\frac{\mathrm{Tr}(\nabla^2 F[W'])}{2||\nabla F[W']||_2} = \frac{c_2 \, \mathrm{Tr}(\nabla^2 F[W])}{2||\nabla F[W]||_2}$$

Optimal Linear Network Size

- Define $N = io, \tilde{N} = c_1 c_2 io$
- Compare learning rate $k(N)$ vs $k(\tilde{N})$:

$$\langle k(N) \rangle \approx \frac{-||\nabla F||_2}{F}\left[-\gamma_1 + T\gamma_1^2 \nabla \hat{F}^T \nabla^2 F \nabla \hat{F} + T\frac{\mathrm{Tr}(\nabla^2 F)}{2||\nabla F||_2^2}\left[\frac{\gamma_2^2}{N} + \frac{\gamma_3^2}{T}\right]\right]$$

- If $\nabla F[W']$ projects equally onto Hessian eigenvectors:

$$\nabla \hat{F}[W']^T \nabla^2 F[W'] \nabla \hat{F}[W'] \approx c_2 \nabla \hat{F}[W]^T \nabla^2 F[W] \nabla \hat{F}[W]$$

- Previously:

$$\frac{\mathrm{Tr}(\nabla^2 F[W'])}{2||\nabla F[W']||_2} = \frac{c_2\, \mathrm{Tr}(\nabla^2 F[W])}{2||\nabla F[W]||_2}$$

- Thus:

$$\langle k(\tilde{N}) \rangle \approx \frac{-||\nabla F||_2}{F}\left[-\gamma_1 + Tc_2\gamma_1^2 \nabla \hat{F}^T \nabla^2 F \nabla \hat{F} + Tc_2\frac{\mathrm{Tr}(\nabla^2 F)}{2||\nabla F||_2^2}\left[\frac{\gamma_2^2}{\tilde{N}} + \frac{\gamma_3^2}{T}\right]\right]$$

Optimal Linear Network Size

Find N^* that maximizes $k(\tilde{N})$:

$$N^* \approx \frac{T\gamma_2^2}{\gamma_3^2}(1 - \frac{\gamma_1^2}{\gamma_2^2})$$

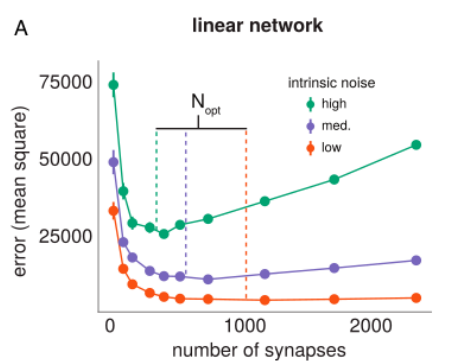
If no task-irrelevant plasticity, $\gamma_2 = 0 \Rightarrow N^* \approx 0 - \frac{T\gamma_1^2}{\gamma_3^2} < 0 \Rightarrow$ optimal network size is negative?

Optimal Linear Network Size

Find N^* that maximizes $k(\tilde{N})$:

$$N^* \approx \frac{T\gamma_2^2}{\gamma_3^2}(1 - \frac{\gamma_1^2}{\gamma_2^2})$$

If no task-irrelevant plasticity, $\gamma_2 = 0 \Rightarrow N^* \approx 0 - \frac{T\gamma_1^2}{\gamma_3^2} < 0 \Rightarrow$ optimal network size is negative?



Optimal Non-Linear Network Size

- Student-Teacher framework with logistic sigmoid activation functions :

$$h^{(k)} = \sigma(W^{(k)} h^{k-1})$$

Optimal Non-Linear Network Size

- Student-Teacher framework with logistic sigmoid activation functions :
$$h^{(k)} = \sigma(W^{(k)} h^{k-1})$$
- Replace W with larger W', with new synaptic weights initialized to zero

Optimal Non-Linear Network Size

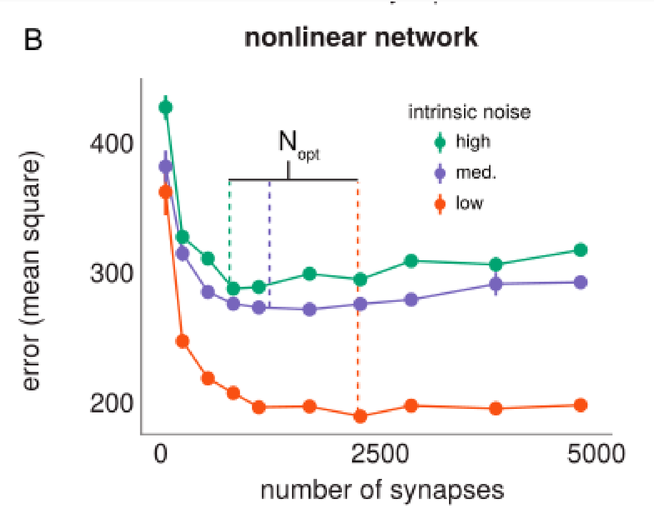
- Student-Teacher framework with logistic sigmoid activation functions :

$$h^{(k)} = \sigma(W^{(k)} h^{k-1})$$

- Replace W with larger W', with new synaptic weights initialized to zero
- Through some derivation I didn't have time to read:

$$N^* = \frac{T\gamma_2^2}{\gamma_3^2}\left[\frac{\gamma_1^2 N}{\gamma_2^2 N^*}\right]$$

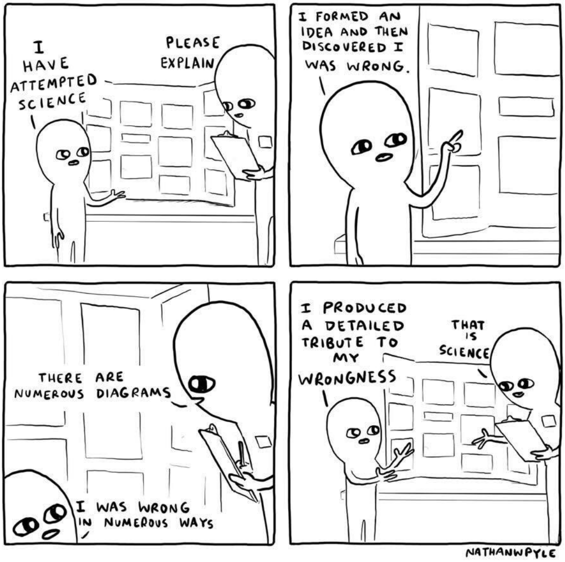
Optimal Non-Linear Network Size



Takeaways

- Larger networks learn better, but
- Intrinsically noisy synapses eventually negate benefits of larger network size
- Experimental Prediction: Circuit size should be inversely proportional to per-synaptic rate of change
- Experimental Prediction: suppression of synaptic noise allows for larger circuit formation

Questions?



I HAVE ATTEMPTED SCIENCE
PLEASE EXPLAIN
I FORMED AN IDEA AND THEN DISCOVERED I WAS WRONG.
THERE ARE NUMEROUS DIAGRAMS
I WAS WRONG IN NUMEROUS WAYS
I PRODUCED A DETAILED TRIBUTE TO MY WRONGNESS
THAT IS SCIENCE
NATHANWPYLE

References i

T. M. Bartol Jr, C. Bromer, J. Kinney, M. A. Chirillo, J. N. Bourne, K. M. Harris, and T. J. Sejnowski.
Nanoconnectomic upper bound on the variability of synaptic plasticity.
Elife, 4:e10778, 2015.

E. B. Bloss, M. S. Cembrowski, B. Karsh, J. Colonell, R. D. Fetter, and N. Spruston.
Single excitatory axons form clustered synapses onto ca1 pyramidal cell dendrites.
Nature neuroscience, 21(3):353, 2018.

References ii

C. Clopath, T. Bonhoeffer, M. Hübener, and T. Rose.
Variance and invariance of neuronal long-term representations.
Philosophical Transactions of the Royal Society B: Biological Sciences, 372(1715):20160161, 2017.

S. Druckmann, L. Feng, B. Lee, C. Yook, T. Zhao, J. C. Magee, and J. Kim.
Structured synaptic connectivity between hippocampal regions.
Neuron, 81(3):629–640, 2014.

References iii

R. Dvorkin and N. E. Ziv.
Relative contributions of specific activity histories and spontaneous processes to size remodeling of glutamatergic synapses.
PLoS biology, 14(10):e1002572, 2016.

K. Eichler, F. Li, A. Litwin-Kumar, Y. Park, I. Andrade, C. M. Schneider-Mizell, T. Saumweber, A. Huser, C. Eschbach, B. Gerber, et al.
The complete connectome of a learning and memory centre in an insect brain.
Nature, 548(7666):175, 2017.

References iv

D. Kappel, R. Legenstein, S. Habenschuss, M. Hsieh, and W. Maass.
A dynamic connectome supports the emergence of stable computational function of neural circuits through reward-based learning.
Eneuro, 5(2), 2018.

Y. Loewenstein, U. Yanover, and S. Rumpel.
Predicting the dynamics of network connectivity in the neocortex.
Journal of Neuroscience, 35(36):12535–12544, 2015.

References v

G. Mongillo, S. Rumpel, and Y. Loewenstein.
Intrinsic volatility of synaptic connections—a challenge to the synaptic trace theory of memory.
Current opinion in neurobiology, 46:7–13, 2017.

L. Susman, N. Brenner, and O. Barak.
Stable memory with unstable synapses.
Nature communications, 10(1):1–9, 2019.