Generalization Dynamics

Outline

- Introduction
- Problem setup
- Exact solutions

}**Arian**

- Technical dive
- Empirical results

}**Reyhane**

- Take home message
- Discussion

}**Mohammad**

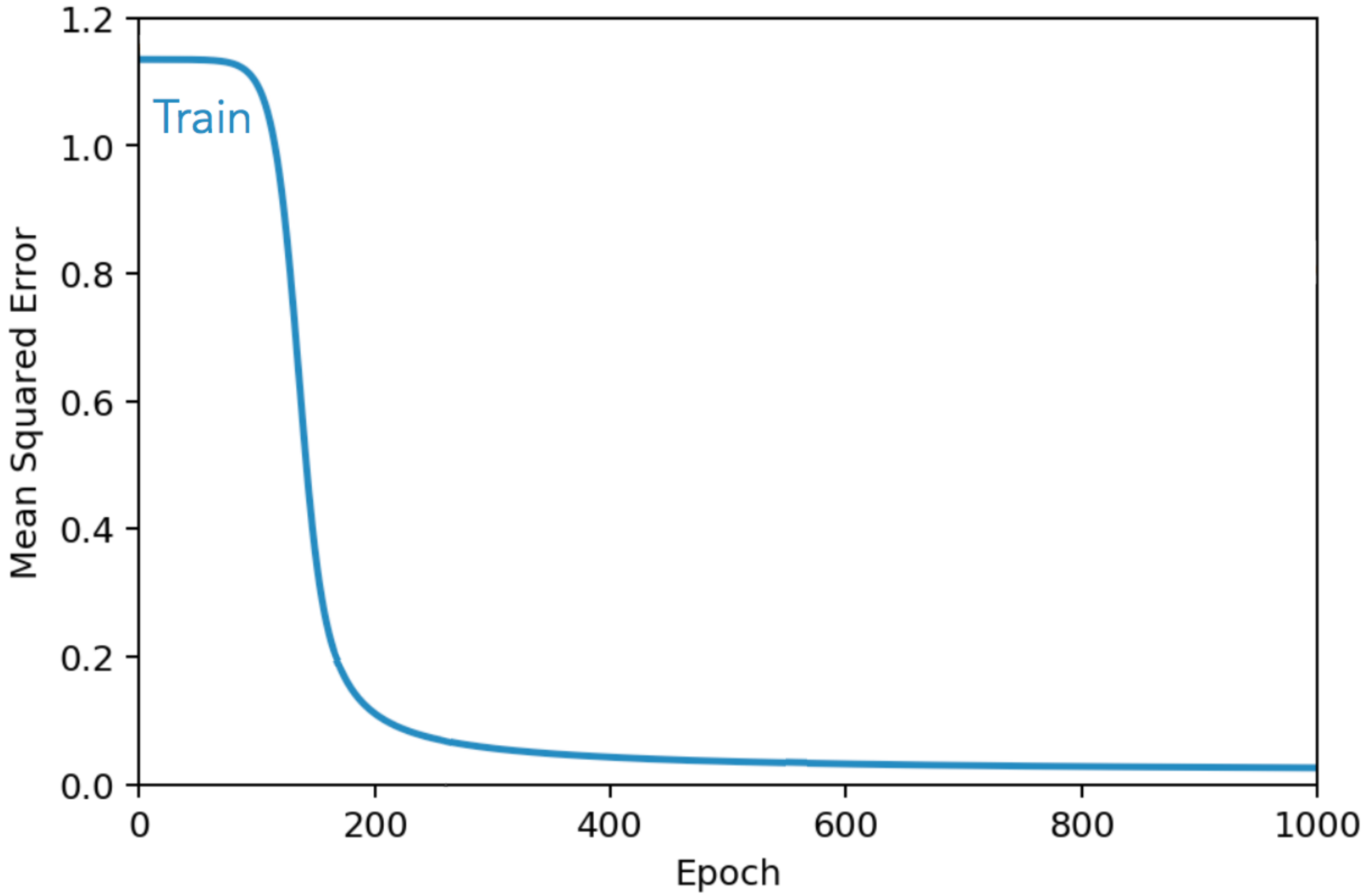
Paper: Advani, Madhu S., and Andrew M. Saxe. "**High-dimensional dynamics of generalization error in neural networks.**" arXiv preprint arXiv:1710.03667 (2017).

Generalization Dynamics

Introduction

MILA

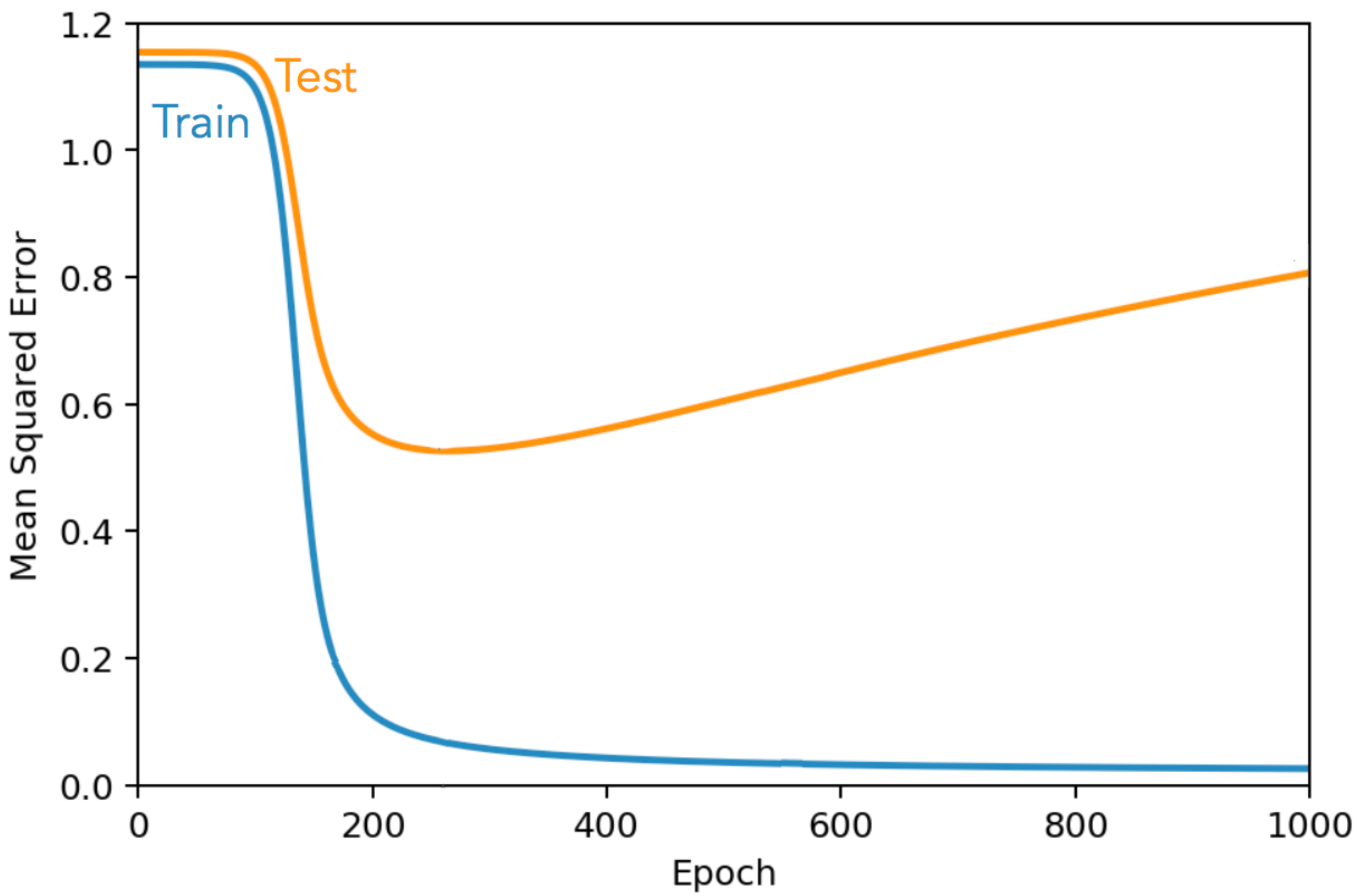
- How does training error E_t evolve?



Generalization Dynamics

Introduction

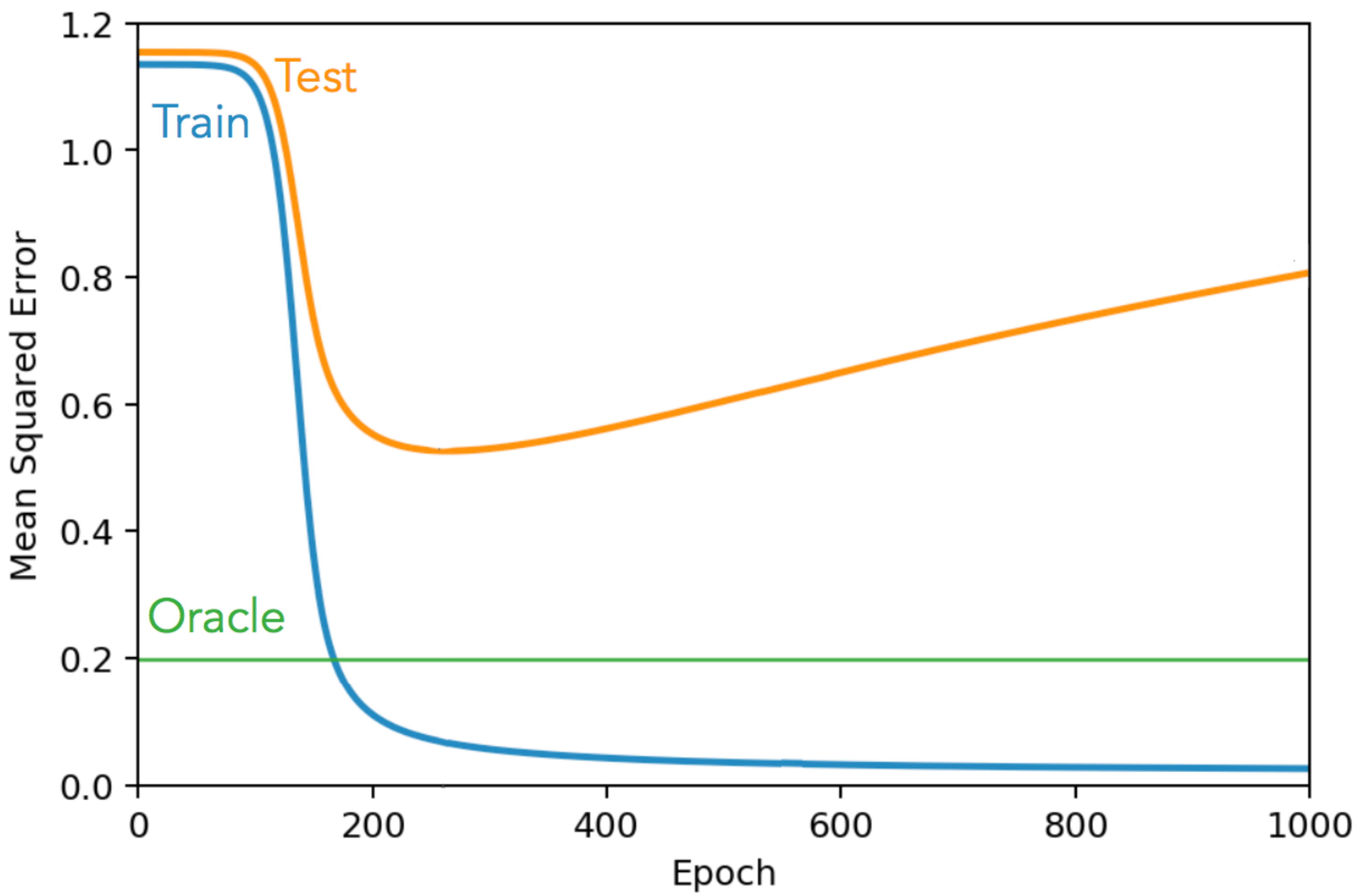
- How does training error E_t evolve?
- How does generalization error E_g evolve?



Generalization Dynamics

Introduction

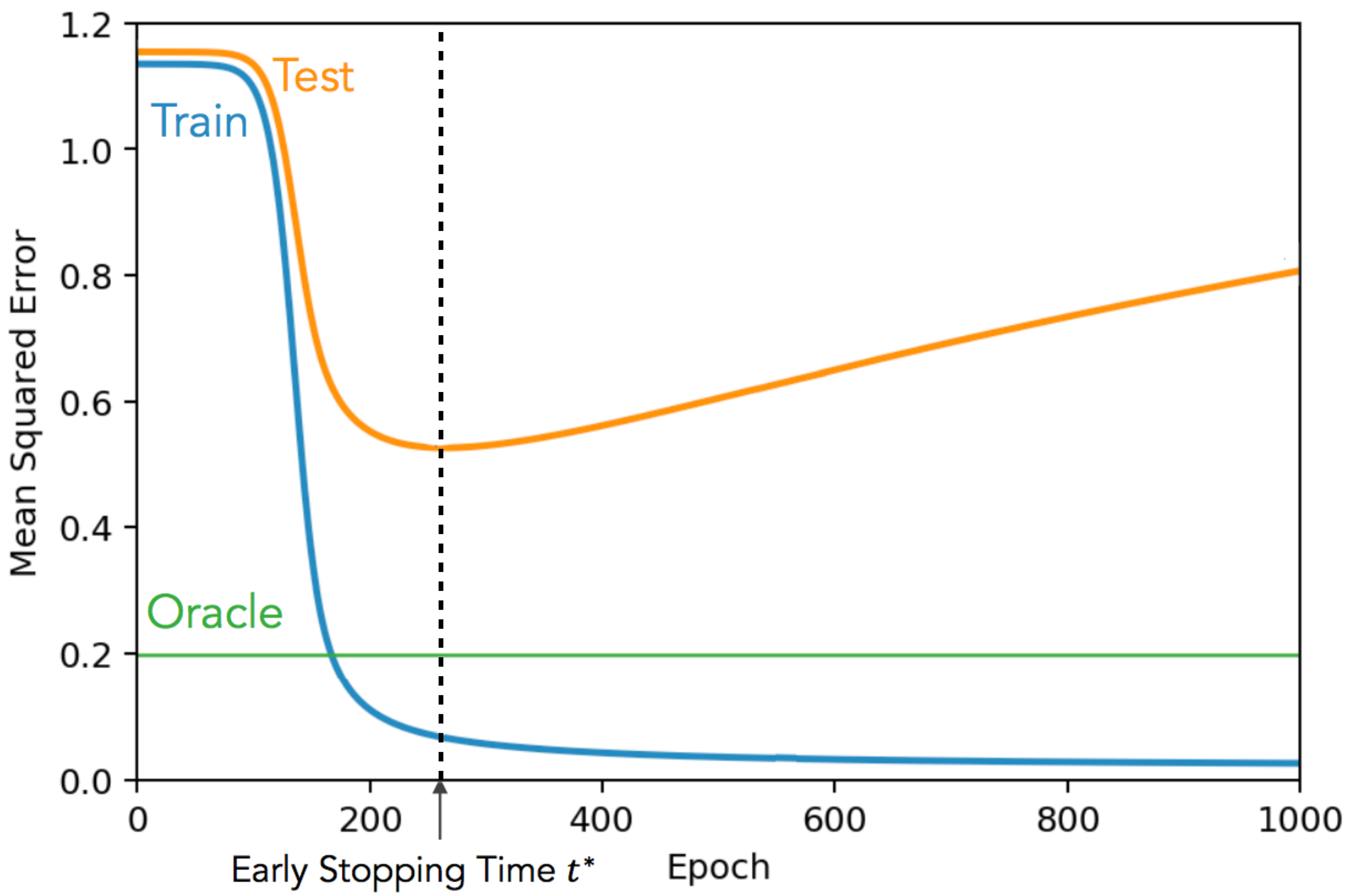
- How does training error E_t evolve?
- How does generalization error E_g evolve?
- What is the lowest generalization error we can ever get?



Generalization Dynamics

Introduction

- How does training error E_t evolve?
- How does generalization error E_g evolve?
- What is the lowest generalization error we can ever get?
- When is the optimal early stopping time?
- How does the number of parameters affect generalization?



Generalization Dynamics

Problem setup

- Teacher
- Student
- $\partial w(t)$ / ∂t
- Initial solution ⟶ final solution
  - Eliminating rotation matrices
    - Finding rotations
- Solving the equation
- Analyze results
  - $t = 0$
  - $t \to \infty$
- Optimal time

Generalization Dynamics

Problem setup

- Learning from a noisy linear teacher
- A dataset D:

$$\mathcal{D} = \{x^{\mu}, y^{\mu}\}, \mu = 1, \cdots, P$$

- Data generation process (**Teacher**):

$$y = \bar{w}X + \epsilon.$$

X is input data sampled from a gaussian distribution

- **Teacher**
- Student
- $\partial w(t)$ / ∂t
- Initial solution ⟶ final solution
  - Eliminating rotation matrices
    - Finding rotations
- Solving the equation
- Analyze results
  - $t = 0$
  - $t \to \infty$
- Optimal time

Generalization Dynamics

Problem setup

- Learning from a noisy linear teacher
- A dataset D:

$$\mathcal{D} = \{x^\mu, y^\mu\}, \mu = 1, \cdots, P$$

- Data generation process (**Teacher**):

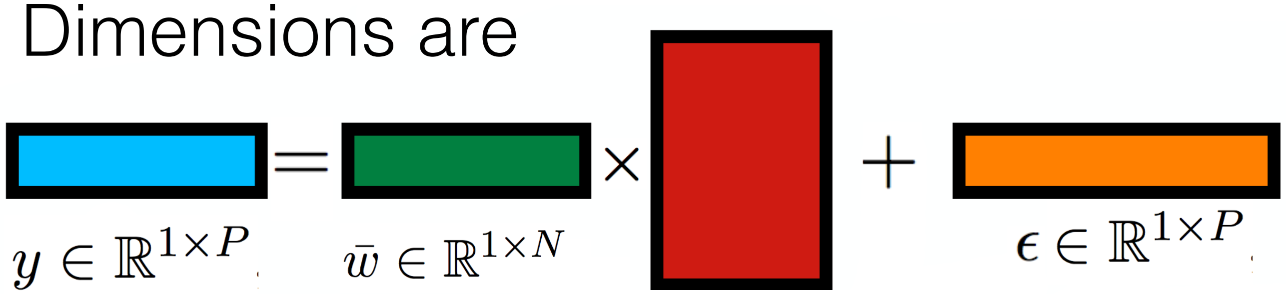
$$y = \bar{w}X + \epsilon.$$

X is input data sampled from a gaussian distribution

- $\bar{w}$ and ϵ are drawn from normal distributions.

$\bar{w} \rightarrow \sigma_w^2$, $\epsilon \rightarrow \sigma_\epsilon^2$ $\longrightarrow$ $\mathrm{SNR} \equiv \sigma_w^2/\sigma_\epsilon^2$

- Dimensions are



$y \in \mathbb{R}^{1\times P}$ = $\bar{w} \in \mathbb{R}^{1\times N}$ × $X \in \mathbb{R}^{N\times P}$ + $\epsilon \in \mathbb{R}^{1\times P}$

- **Teacher**
- Student
- $\partial w(t) / \partial t$
- Initial solution ⟶ final solution
  - Eliminating rotation matrices
    - Finding rotations
- Solving the equation
- Analyze results
  - $t = 0$
  - $t \rightarrow \infty$
- Optimal time

P ⟶ # of examples

N ⟶ # of dimensions

Generalization Dynamics

Problem setup

- Data generation process (Teacher):

$$y = \bar{w}X + \epsilon.$$

- Model (**Student**):

$$\hat{y} = wX$$

- Training Error:

$$E_t(w(t)) = \frac{1}{P}\sum_{\mu=1}^{P} \|y^{\mu} - \hat{y}^{\mu}\|_2^2,$$

- Generalization error:

$$E_g(w(t)) = \langle (y - \hat{y})^2 \rangle_{x,\epsilon}$$

- Teacher
- **Student**
- $\partial w(t) / \partial t$
- Initial solution ⟶ final solution
    - Eliminating rotation matrices
        - Finding rotations
- Solving the equation
- Analyze results
    - $t = 0$
    - $t \to \infty$
- Optimal time

$\langle \cdot \rangle$ denotes $Expectation$

Generalization Dynamics

Problem setup

- Training Error:

$$E_t(w(t)) = \frac{1}{P}\sum_{\mu=1}^{P} \|y^\mu - \hat{y}^\mu\|_2^2,$$

- Gradient descent update rule:

$$w \leftarrow w - \eta \frac{\partial E_t}{\partial w}$$

- Teacher
- Student
- $\partial w(t) / \partial t$
- Initial solution → final solution
  - Eliminating rotation matrices
    - Finding rotations
- Solving the equation
- Analyze results
  - $t = 0$
  - $t \to \infty$
- Optimal time

Generalization Dynamics

Problem setup

- Training Error:

$$E_t(w(t)) = \frac{1}{P}\sum_{\mu=1}^{P} \|y^\mu - \hat{y}^\mu\|_2^2,$$

- Gradient descent update rule:

$$w \leftarrow w - \eta \frac{\partial E_t}{\partial w}$$

- It follows that:

$$\tau \dot{w}(t) = yX^T - wXX^T$$

- Teacher
- Student
- $\partial w(t) / \partial t$
- Initial solution → final solution
  - Eliminating rotation matrices
    - Finding rotations
- Solving the equation
- Analyze results
  - $t = 0$
  - $t \to \infty$
- Optimal time

Generalization Dynamics

Exact solutions

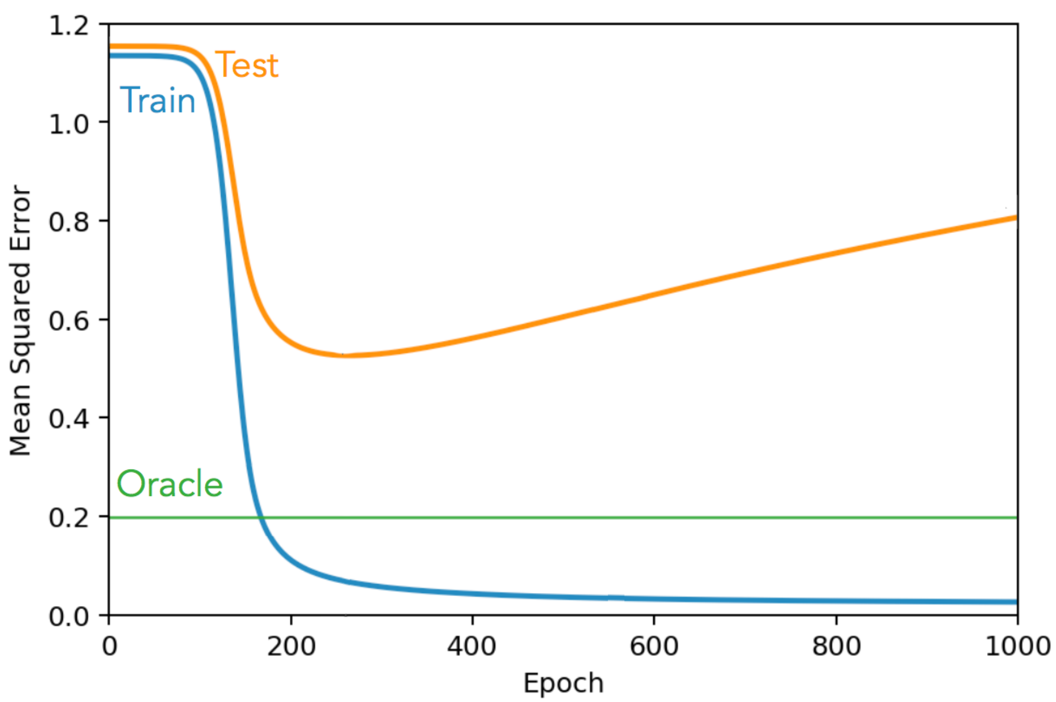
- Dynamics of $w(t)$:

$$\tau \dot{w}(t) = yX^T - wXX^T$$

- Final solution

$$w(t \to \infty) = yX^T(XX^T)^+$$

- How do we get to $w(t \to \infty)$ from $w(t=0)$?



$w(t=0)$ $w(t \to \infty)$

- Teacher
- Student
- $\partial w(t) \,/\, \partial t$
- **Initial solution ⟶ final solution**
  - Eliminating rotation matrices
    - Finding rotations
- Solving the equation
- Analyze results
  - $t = 0$
  - $t \to \infty$
- Optimal time

Generalization Dynamics

Exact solutions

- Dynamics of $w(t)$:

$$\tau \dot{w}(t) = yX^T - wXX^T$$

- Final solution

$$w(t \to \infty) = yX^T(XX^T)^+$$

- How do we get to $w(t \to \infty)$ from $w(t = 0)$?
    - **Some change of variables are needed!**
        - Input covariance matrix:

$$\Sigma^{xx} = XX^T = V\Lambda V^T$$

- Teacher
- Student
- $\partial w(t) / \partial t$
- **Initial solution ⟶ final solution**
    - Eliminating rotation matrices
        - Finding rotations
- Solving the equation
- Analyze results
    - $t = 0$
    - $t \to \infty$
- Optimal time

$$\Sigma^{xx} = XX^T = V\Lambda V^T$$

Generalization Dynamics

Exact solutions

- Dynamics of $w(t)$:

$$\tau \dot{w}(t) = yX^T - wXX^T$$

- Final solution

$$w(t \rightarrow \infty) = yX^T (XX^T)^+$$

- How do we get to $w(t \rightarrow \infty)$ from $w(t = 0)$?
    - **Some change of variables are needed!**
        - Input covariance matrix:

$$\Sigma^{xx} = XX^T = V\Lambda V^T$$

        - Input-output covariance matrix:

$$\Sigma^{yx} = yX^T = \tilde{s}V^T$$

- Teacher
- Student
- $\partial w(t) / \partial t$
- **Initial solution ⟶ final solution**
    - Eliminating rotation matrices
        - Finding rotations
- Solving the equation
- Analyze results
    - $t = 0$
    - $t \rightarrow \infty$
- Optimal time

$$\Sigma^{xx} = XX^T = V\Lambda V^T$$

$$\Sigma^{yx} = yX^T = \tilde{s}V^T$$

Generalization Dynamics

Exact solutions

- Dynamics of $w(t)$:

$$\tau \dot{w}(t) = yX^T - wXX^T$$

- Change of variable $z = wV$

$$\tau \dot{z}(t) = \tilde{s} - z\Lambda$$

- Teacher
- Student
- $\partial w(t) / \partial t$
- Initial solution → final solution
  - **Eliminating rotation matrices**
    - Finding rotations
- Solving the equation
- Analyze results
  - $t = 0$
  - $t \to \infty$
- Optimal time

$$\Sigma^{xx} = XX^T = V\Lambda V^T$$

$$\Sigma^{yx} = yX^T = \tilde{s}V^T$$

Generalization Dynamics

Exact solutions

MILA

- Dynamics of $w(t)$:

$$\tau \dot{w}(t) = yX^T - wXX^T$$

- Change of variable $z = wV$

$$\tau \dot{z}(t) = \tilde{s} - z\Lambda$$

- So:

$$\tilde{s}V^T = yX^T = \bar{w}XX^T + \epsilon X^T = \bar{z}\Lambda V^T + \tilde{\epsilon}\Lambda^{1/2}V^T$$

$$\longrightarrow \tilde{s} = \bar{z}\Lambda + \tilde{\epsilon}\Lambda^{1/2}$$

$$y = \bar{w}X + \epsilon$$
$$yX^T = \bar{w}XX^T + \epsilon X^T$$
$$XX^T = V\Lambda V^T \Rightarrow X = V\Lambda^{1/2}U^T$$
$$yX^T = \bar{w}V\Lambda V^T + \epsilon U\Lambda^{1/2}V^T$$
$$yX^T = \bar{w}V\Lambda V^T + \tilde{\epsilon}\Lambda^{1/2}V^T$$
$$\tilde{s} = \bar{z}\Lambda + \tilde{\epsilon}\Lambda^{1/2}$$

- Teacher
- Student
- $\partial w(t) / \partial t$
- Initial solution ⟶ final solution
  - Eliminating rotation matrices
    - **Finding rotations**
- Solving the equation
- Analyze results
  - $t = 0$
  - $t \rightarrow \infty$
- Optimal time

$$\Sigma^{xx} = XX^T = V\Lambda V^T$$

$$\Sigma^{yx} = yX^T = \tilde{s}V^T$$

Generalization Dynamics

Exact solutions

- Dynamics of $w(t)$:

$$\tau \dot{w}(t) = yX^T - wXX^T$$

- Change of variable $z = wV$

$$\tau \dot{z}(t) = \tilde{s} - z\Lambda$$

- So:

$$\tilde{s}V^T = yX^T = \bar{w}XX^T + \epsilon X^T = \bar{z}\Lambda V^T + \tilde{\epsilon}\Lambda^{1/2}V^T$$

$$\longrightarrow \tilde{s} = \bar{z}\Lambda + \tilde{\epsilon}\Lambda^{1/2}$$

- **The learning speed of each mode is independent of the others:**

$$\tau \dot{z}_i = (\bar{z}_i - z_i)\lambda_i + \tilde{\epsilon}_i\sqrt{\lambda_i}, \qquad i = 1, \cdots, N.$$

- Teacher
- Student
- $\partial w(t) / \partial t$
- Initial solution ⟶ final solution
  - Eliminating rotation matrices
    - **Finding rotations**
- Solving the equation
- Analyze results
  - $t = 0$
  - $t \to \infty$
- Optimal time

$$\Sigma^{xx} = XX^T = V\Lambda V^T$$

$$\Sigma^{yx} = yX^T = \tilde{s}V^T$$

Generalization Dynamics

Exact solutions

MILA

- Dynamics of $\cancel{w(t)}\ z(t)$:

$$\tau \dot{z}_i = (\bar{z}_i - z_i)\lambda_i + \tilde{\epsilon}_i \sqrt{\lambda_i}$$

- Solving the differential equation:

$$\bar{z}_i - z_i = (\bar{z}_i - z_i(0))e^{-\frac{\lambda_i t}{\tau}} - \frac{\tilde{\epsilon}_i}{\sqrt{\lambda_i}}(1 - e^{-\frac{\lambda_i t}{\tau}})$$

- Teacher
- Student
- $\partial w(t)$ / ∂t
- Initial solution ⟶ final solution
    - Eliminating rotation matrices
        - Finding rotations
- **Solving the equation**
- Analyze results
    - $t = 0$
    - $t \rightarrow \infty$
- Optimal time

Generalization Dynamics

Exact solutions

- Dynamics of ~~$w(t)$~~ $z(t)$:

$$\tau \dot{z}_i = (\bar{z}_i - z_i)\lambda_i + \tilde{\epsilon}_i \sqrt{\lambda_i}$$

- Solving the differential equation:

$$\bar{z}_i - z_i = (\bar{z}_i - z_i(0))e^{-\frac{\lambda_i t}{\tau}} - \frac{\tilde{\epsilon}_i}{\sqrt{\lambda_i}}(1 - e^{-\frac{\lambda_i t}{\tau}})$$

- Back to generalization dynamics:

$$E_g(w(t)) = \langle (y - \hat{y})^2 \rangle_{x,\epsilon}$$

- Teacher
- Student
- $\partial w(t)$ / ∂t
- Initial solution → final solution
  - Eliminating rotation matrices
    - Finding rotations
- **Solving the equation**
- Analyze results
  - $t = 0$
  - $t \to \infty$
- Optimal time

Generalization Dynamics

Exact solutions

- Dynamics of ~~$w(t)$~~ $z(t)$:

$$\tau \dot{z}_i = (\bar{z}_i - z_i)\lambda_i + \tilde{\epsilon}_i \sqrt{\lambda_i}$$

- Solving the differential equation:

$$\bar{z}_i - z_i = (\bar{z}_i - z_i(0))e^{-\frac{\lambda_i t}{\tau}} - \frac{\tilde{\epsilon}_i}{\sqrt{\lambda_i}}(1 - e^{-\frac{\lambda_i t}{\tau}})$$

- Back to generalization dynamics:

$$E_g(w(t)) = \langle (y - \hat{y})^2 \rangle_{x,\epsilon}$$

$$E_g(t) = \frac{1}{N}\sum_i \langle (\bar{z}_i - z_i)^2 \rangle + \sigma_\epsilon^2$$

$$= \frac{1}{N}\sum_i \left[(\sigma_w^2 + (\sigma_w^0)^2)e^{-\frac{2\lambda_i t}{\tau}} + \frac{\sigma_\epsilon^2}{\lambda_i}(1 - e^{-\frac{\lambda_i t}{\tau}})^2 \right] + \sigma_\epsilon^2$$

- Teacher
- Student
- $\partial w(t) / \partial t$
- Initial solution → final solution
  - Eliminating rotation matrices
    - Finding rotations
- **Solving the equation**
- Analyze results
  - $t = 0$
  - $t \rightarrow \infty$
- Optimal time

Generalization Dynamics

Technical Dive

$$E_g(t) = \frac{1}{N}\sum_i \left[(\sigma_w^2 + (\sigma_w^0)^2)e^{-\frac{2\lambda_i t}{\tau}} + \frac{\sigma_\epsilon^2}{\lambda_i}(1 - e^{-\frac{\lambda_i t}{\tau}})^2\right] + \sigma_\epsilon^2$$

initialization effect

noise effect

lower bound

How it changes over time?
How it changes over lambdas?
When overfitting happens?
Frozen subspace?

- Teacher
- Student
- $\partial w(t) / \partial t$
- Initial solution → final solution
  - Eliminating rotation matrices
    - Finding rotations
- Solving the equation
- **Analyze results**
  - $t = 0$
  - $t \to \infty$
- Optimal time

Generalization Dynamics

Technical Dive

$$E_g(t) = \frac{1}{N}\sum_i \left[(\sigma_w^2 + (\sigma_w^0)^2)e^{-\frac{2\lambda_i t}{\tau}} + \frac{\sigma_\epsilon^2}{\lambda_i}(1 - e^{-\frac{\lambda_i t}{\tau}})^2\right] + \sigma_\epsilon^2$$

initialization effect — noise effect — lower bound

- Teacher
- Student
- $\partial w(t)$ / ∂t
- Initial solution → final solution
  - Eliminating rotation matrices
    - Finding rotations
- Solving the equation
- **Analyze results**
  - $t = 0$
  - $t \to \infty$
- Optimal time

$t \to 0$ ········▸ Staying at the initialization with zero noise effect

$t \to \infty$ ········▸ Overfitting on noise

Generalization Dynamics

Technical Dive

- The eigenvalue distribution of XX^T approaches the Marchenko-Pasteur distribution

$$\rho^{\mathrm{MP}}(\lambda) = \frac{1}{2\pi}\frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} + 1_{\alpha<1}(1-\alpha)\delta(\lambda)$$

Distribution variable ⟶ λ
Distribution parameters. ⟶ $\lambda_+, \lambda_-, \alpha$

$$\lambda_\pm = (\sqrt{\alpha} \pm 1)^2$$

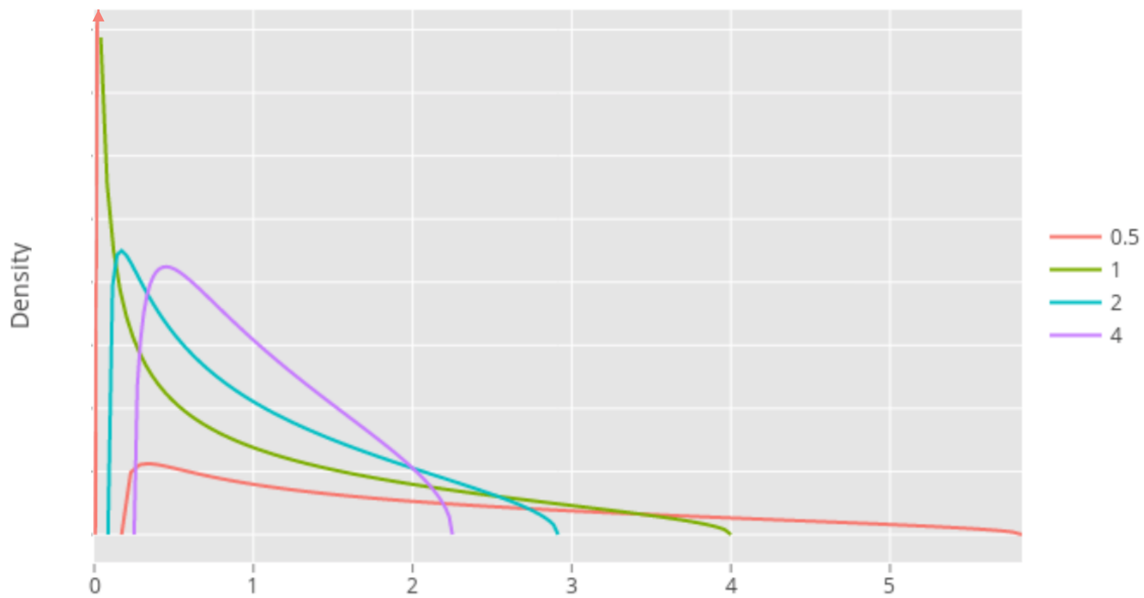
- Teacher
- Student
- $\partial w(t)$ / ∂t
- Initial solution ⟶ final solution
  - Eliminating rotation matrices
    - Finding rotations
- Solving the equation
- **Analyze results**
  - $t = 0$
  - $t \rightarrow \infty$
- Optimal time

Generalization Dynamics

Technical Dive

- The eigenvalue distribution of XX^T approaches the Marchenko-Pastur distribution

$$\rho^{\mathrm{MP}}(\lambda) = \frac{1}{2\pi}\frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} + 1_{\alpha<1}(1-\alpha)\delta(\lambda)$$



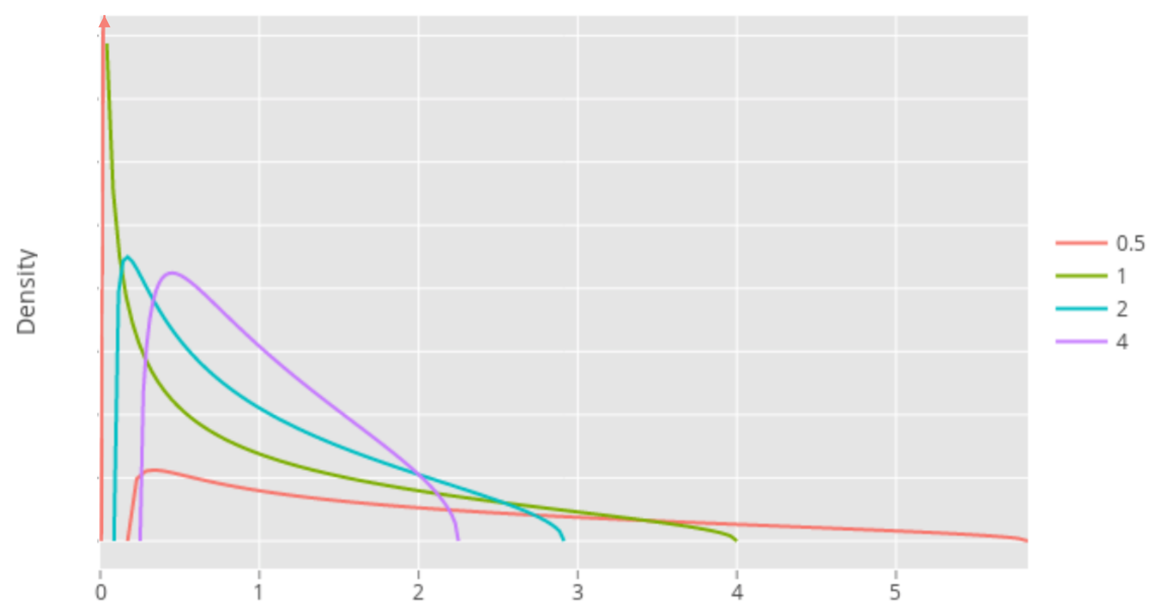
- Teacher
- Student
- $\partial w(t) \,/\, \partial t$
- Initial solution → final solution
  - Eliminating rotation matrices
    - Finding rotations
- Solving the equation
- **Analyze results**
  - $t = 0$
  - $t \to \infty$
- Optimal time

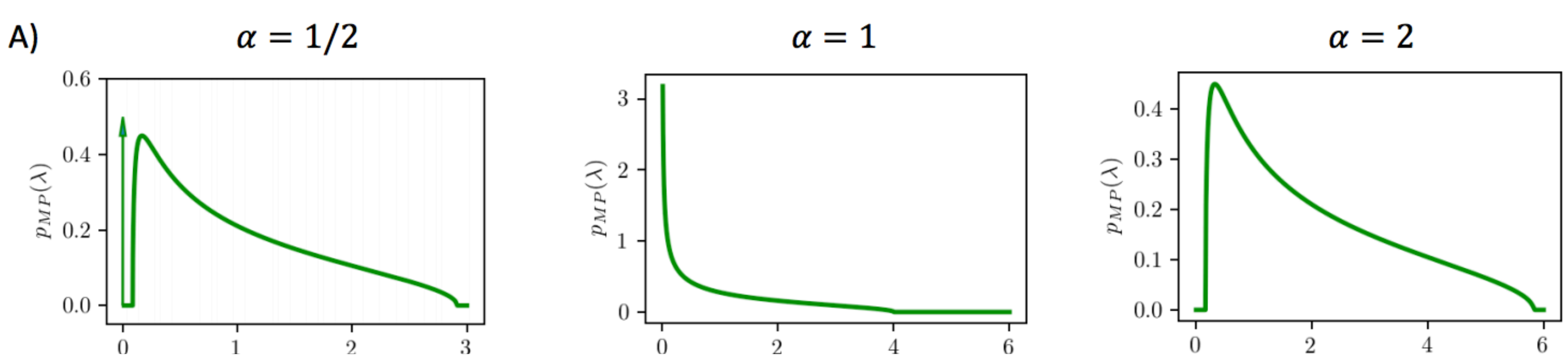
Generalization Dynamics

Technical Dive

- The eigenvalue distribution of XX^T approaches the Marchenko-Pastur distribution

$$\rho^{\mathrm{MP}}(\lambda) = \frac{1}{2\pi}\frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} + 1_{\alpha<1}(1-\alpha)\delta(\lambda)$$





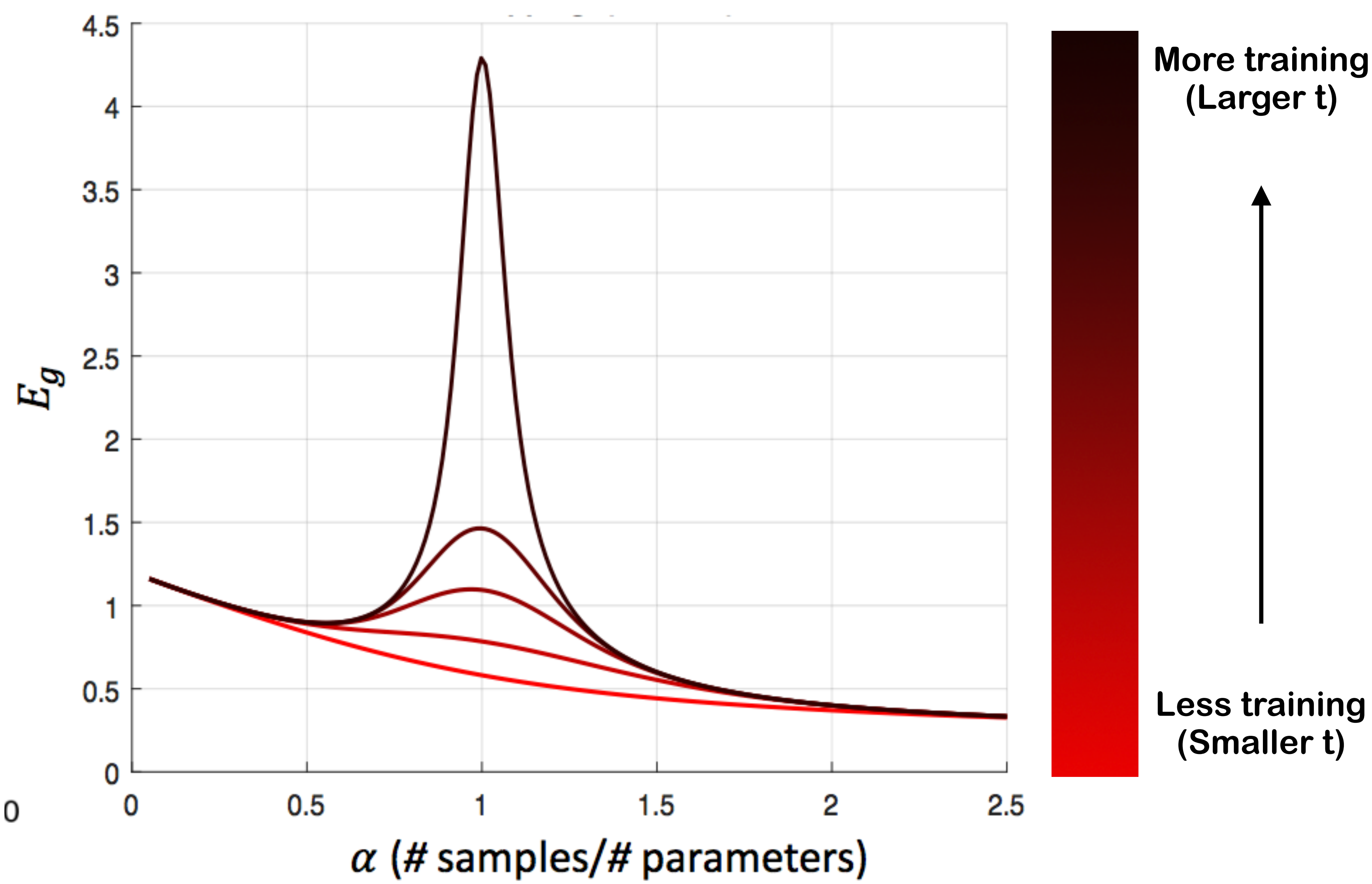
- Teacher
- Student
- $\partial w(t)$ / ∂t
- Initial solution → final solution
  - Eliminating rotation matrices
    - Finding rotations
- Solving the equation
- **Analyze results**
  - $t = 0$
  - $t \to \infty$
- Optimal time

α (# samples/# parameters)

Generalization Dynamics

Technical Dive

- The eigenvalue distribution of XX^T approaches the Marchenko-Pastur distribution



- Teacher
- Student
- $\partial w(t) / \partial t$
- Initial solution ⟶ final solution
  - Eliminating rotation matrices
    - Finding rotations
- Solving the equation
- **Analyze results**
  - $t = 0$
  - $t \rightarrow \infty$
- Optimal time

0

α (# samples/# parameters)

Generalization Dynamics

Technical Dive

$$E_g(t) = \frac{1}{N}\sum_i \left[(\sigma_w^2 + (\sigma_w^0)^2)e^{-\frac{2\lambda_i t}{\tau}} + \frac{\sigma_\epsilon^2}{\lambda_i}(1 - e^{-\frac{\lambda_i t}{\tau}})^2\right] + \sigma_\epsilon^2$$

- Bringing the distribution in

$$\frac{E_g(t)}{\sigma_w^2} = \int \rho^{\mathrm{MP}}(\lambda)\left[e^{-\frac{2\lambda t}{\tau}} + \frac{1}{\lambda \cdot \mathrm{SNR}}(1 - e^{-\frac{\lambda t}{\tau}})^2\right] d\lambda + \frac{1}{\mathrm{SNR}}$$

- Teacher
- Student
- $\partial w(t)$ / ∂t
- Initial solution ⟶ final solution
  - Eliminating rotation matrices
    - Finding rotations
- Solving the equation
- Analyze results
  - $t = 0$
  - $t \to \infty$
- **Optimal time**

Generalization Dynamics

Technical Dive

$$E_g(t) = \frac{1}{N}\sum_i \left[(\sigma_w^2 + (\sigma_w^0)^2)e^{-\frac{2\lambda_i t}{\tau}} + \frac{\sigma_\epsilon^2}{\lambda_i}(1 - e^{-\frac{\lambda_i t}{\tau}})^2\right] + \sigma_\epsilon^2$$

- Bringing the distribution in

$$\frac{E_g(t)}{\sigma_w^2} = \int \rho^{\mathrm{MP}}(\lambda)\left[e^{-\frac{2\lambda t}{\tau}} + \frac{1}{\lambda \cdot \mathrm{SNR}}(1 - e^{-\frac{\lambda t}{\tau}})^2\right] d\lambda + \frac{1}{\mathrm{SNR}}$$

- Optimal stopping time

$$t^{opt} = \frac{\tau}{\lambda}\log(\mathrm{SNR}\cdot\lambda + 1).$$

Optimal stopping time differs for different λ 's

Causes sub-optimality at early stopping

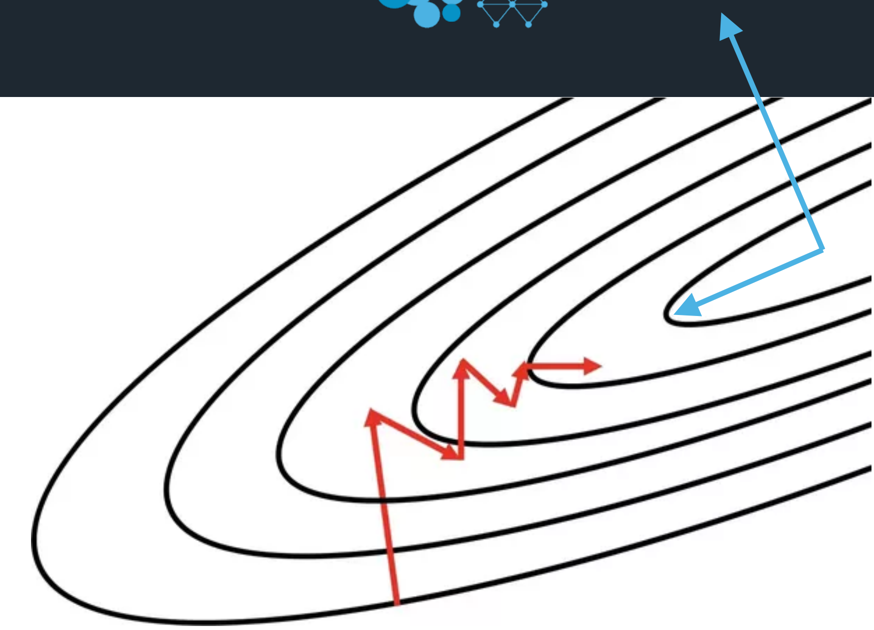
- Teacher
- Student
- $\partial w(t) / \partial t$
- Initial solution ⟶ final solution
  - Eliminating rotation matrices
    - Finding rotations
- Solving the equation
- Analyze results
  - $t = 0$
  - $t \to \infty$
- **Optimal time**

Generalization Dynamics

Our three cents   With help of Remi Tachet

$$w \leftarrow w - \eta \frac{\partial L}{\partial w}$$

$$\tau \dot{w} = -\frac{\partial L}{\partial w}$$



Generalization Dynamics

Our three cents With help of Remi Tachet

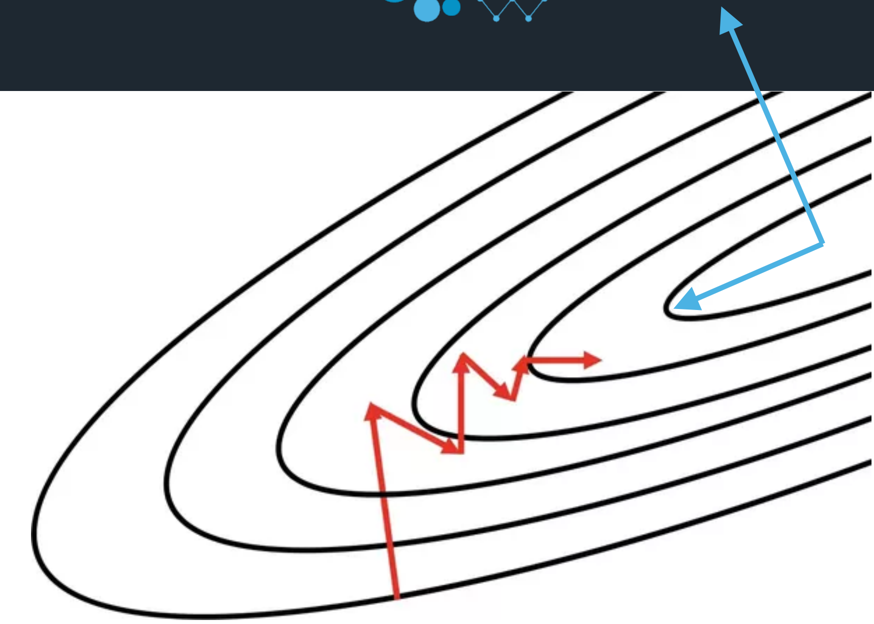
$$w \leftarrow w - \eta \frac{\partial L}{\partial w}$$

$$\tau \dot{w} = -\frac{\partial L}{\partial w}$$

Taylor expansion:

$$f(x) = f(a) + f'(a)(x - a)$$

$$f(w) = \frac{\partial L}{\partial w}$$



Generalization Dynamics

Our three cents

With help of Remi Tachet

MILA

$$w \leftarrow w - \eta \frac{\partial L}{\partial w}$$

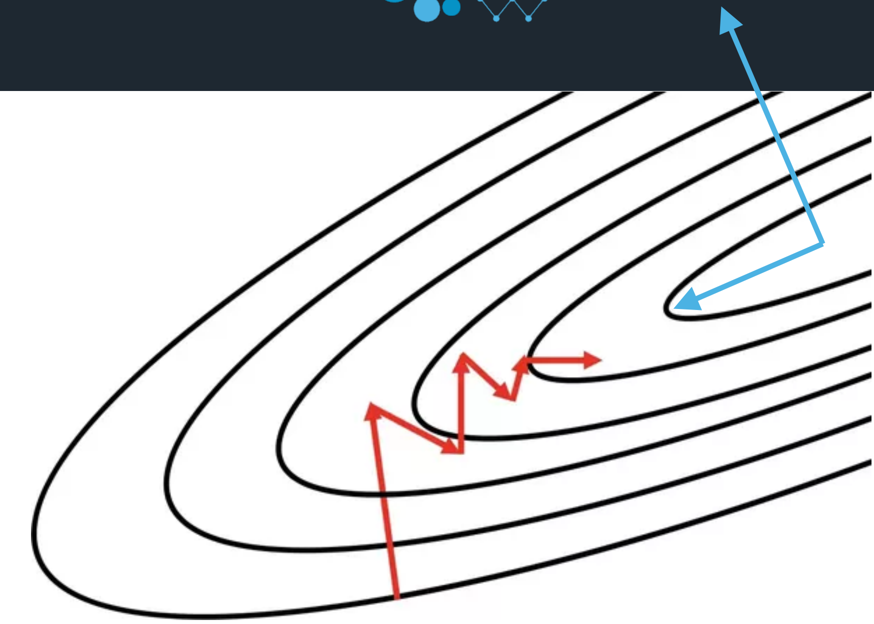
$$\tau \dot{w} = -\frac{\partial L}{\partial w}$$

Taylor expansion:

$$f(x) = f(a) + f'(a)(x-a)$$

$$f(w) = \frac{\partial L}{\partial w}$$

$$f(w) = \frac{\partial L}{\partial w}\bigg|_{w=w^*} + \frac{\partial^2 L}{\partial w^2}\bigg|_{w=w^*}(w - w^*)$$



Generalization Dynamics

Our three cents

With help of Remi Tachet

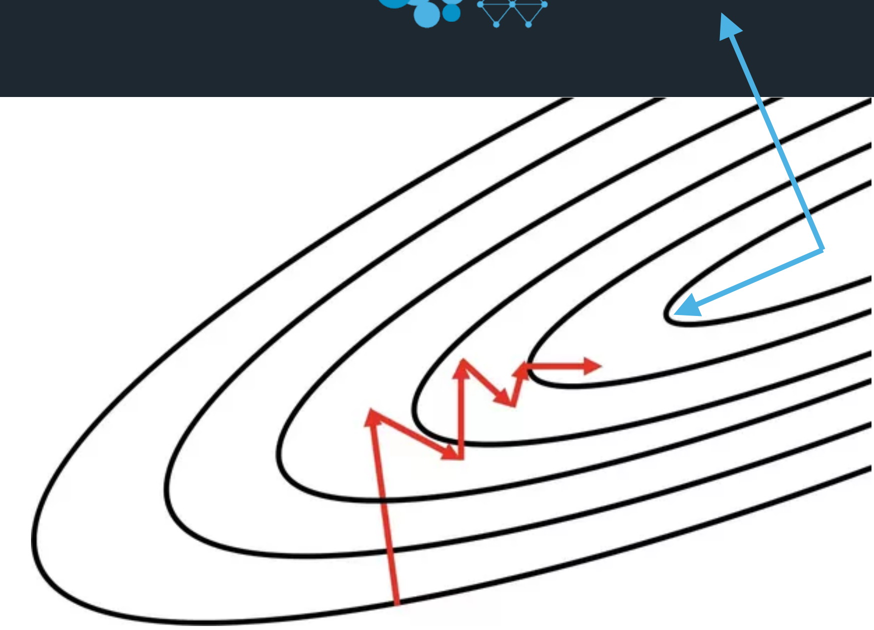
$$w \leftarrow w - \eta \frac{\partial L}{\partial w}$$

$$\tau \dot{w} = -\frac{\partial L}{\partial w}$$

Taylor expansion:

$$f(x) = f(a) + f'(a)(x-a)$$

$$f(w) = \frac{\partial L}{\partial w}$$



$$f(w) = \left.\frac{\partial L}{\partial w}\right|_{w=w^*} + \left.\frac{\partial^2 L}{\partial w^2}\right|_{w=w^*}(w - w^*)$$

$$f(w) = H(w^*)(w - w^*)$$

Generalization Dynamics

Our three cents   With help of Remi Tachet

$$w \leftarrow w - \eta \frac{\partial L}{\partial w}$$

$$\tau \dot{w} = -\frac{\partial L}{\partial w}$$

Taylor expansion:

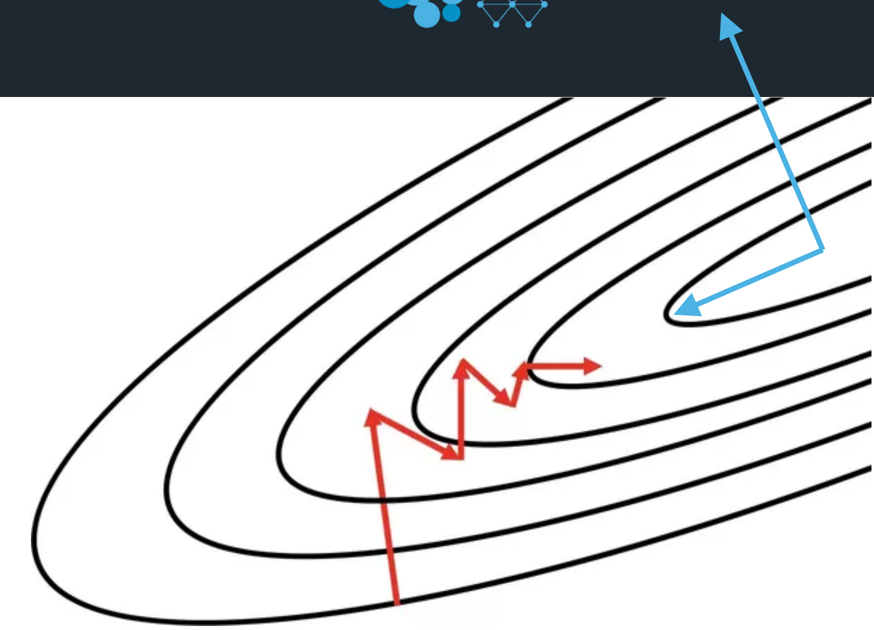
$$f(x) = f(a) + f'(a)(x-a)$$

$$f(w) = \frac{\partial L}{\partial w}$$

$$f(w) = \left.\frac{\partial L}{\partial w}\right|_{w=w^*} + \left.\frac{\partial^2 L}{\partial w^2}\right|_{w=w^*}(w - w^*)$$

$$f(w) = H(w^*)(w - w^*)$$

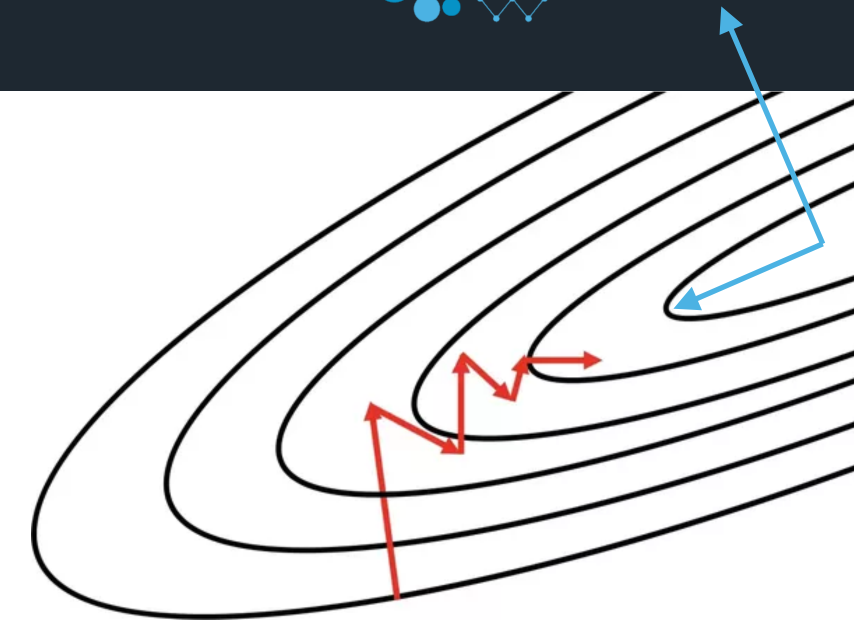
$$\dot{w} = -H(w^*)(w - w^*)$$



Generalization Dynamics

Our three cents

With help of Remi Tachet



$$w \leftarrow w - \eta \frac{\partial L}{\partial w}$$

$$\tau \dot{w} = -\frac{\partial L}{\partial w}$$

Taylor expansion:

$$f(x) = f(a) + f'(a)(x-a)$$

$$f(w) = \frac{\partial L}{\partial w}$$

$$f(w) = \frac{\partial L}{\partial w}\Big|_{w=w^*} + \frac{\partial^2 L}{\partial w^2}\Big|_{w=w^*}(w - w^*)$$

$$f(w) = H(w^*)(w - w^*)$$

$$\dot{w} = -H(w^*)(w - w^*)$$

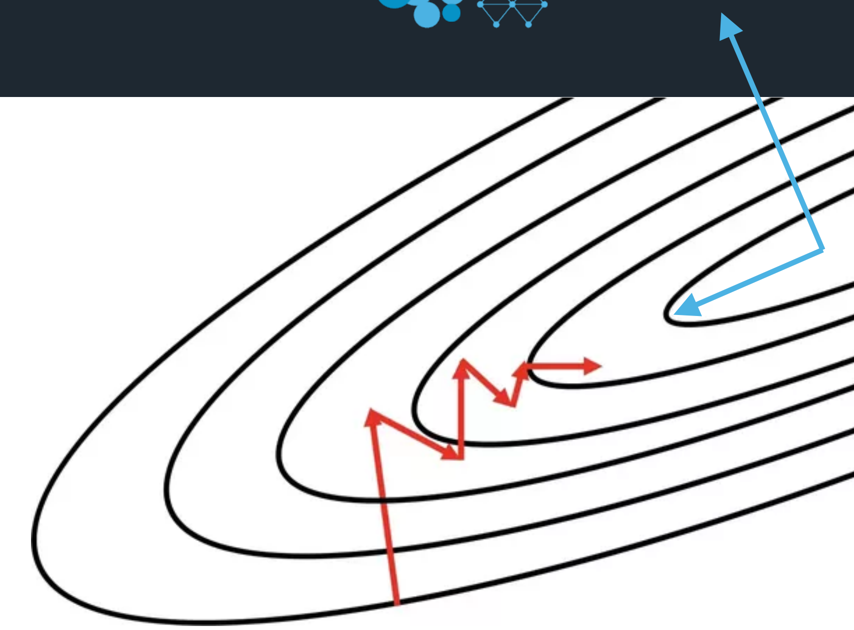
$$u = w - w^*$$

$$\frac{u'}{u} = -H(w^*)$$

Generalization Dynamics

Our three cents — With help of Remi Tachet

MILA



$$w \leftarrow w - \eta \frac{\partial L}{\partial w}$$

$$\tau \dot{w} = -\frac{\partial L}{\partial w}$$

Taylor expansion:

$$f(x) = f(a) + f'(a)(x-a)$$

$$f(w) = \frac{\partial L}{\partial w}$$

$$f(w) = \frac{\partial L}{\partial w}\Big|_{w=w^*} + \frac{\partial^2 L}{\partial w^2}\Big|_{w=w^*}(w - w^*)$$

$$f(w) = H(w^*)(w - w^*)$$

$$\dot{w} = -H(w^*)(w - w^*)$$

$$u = w - w^*$$

$$\frac{u'}{u} = -H(w^*) \longrightarrow ln(u) = -H(w^*)t + C$$

Generalization Dynamics

Our three cents

With help of Remi Tachet

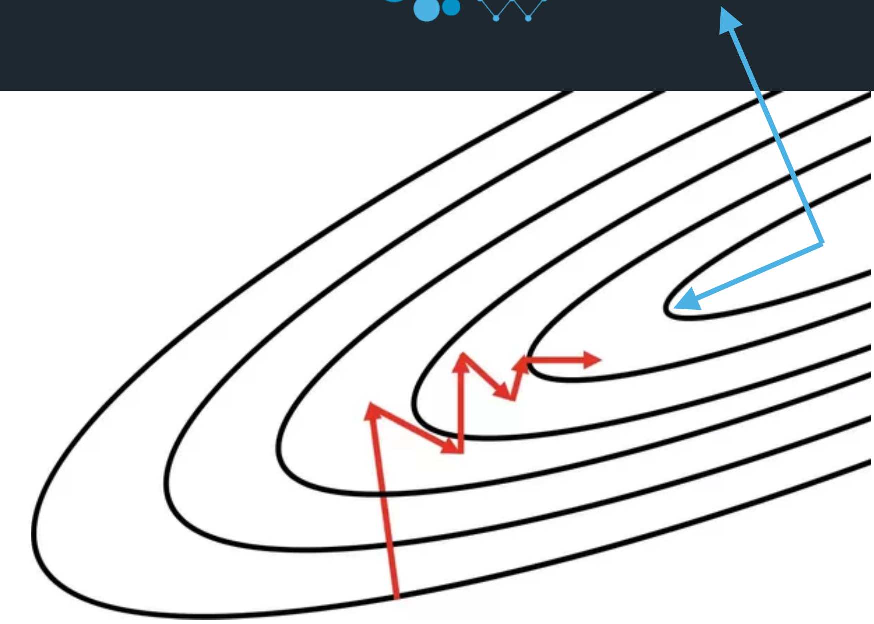
$$w \leftarrow w - \eta \frac{\partial L}{\partial w}$$

$$\tau \dot{w} = -\frac{\partial L}{\partial w}$$

Taylor expansion:

$$f(x) = f(a) + f'(a)(x-a)$$

$$f(w) = \frac{\partial L}{\partial w}$$



$$f(w) = \left.\frac{\partial L}{\partial w}\right|_{w=w^*} + \left.\frac{\partial^2 L}{\partial w^2}\right|_{w=w^*} (w - w^*)$$

$$f(w) = H(w^*)(w - w^*)$$

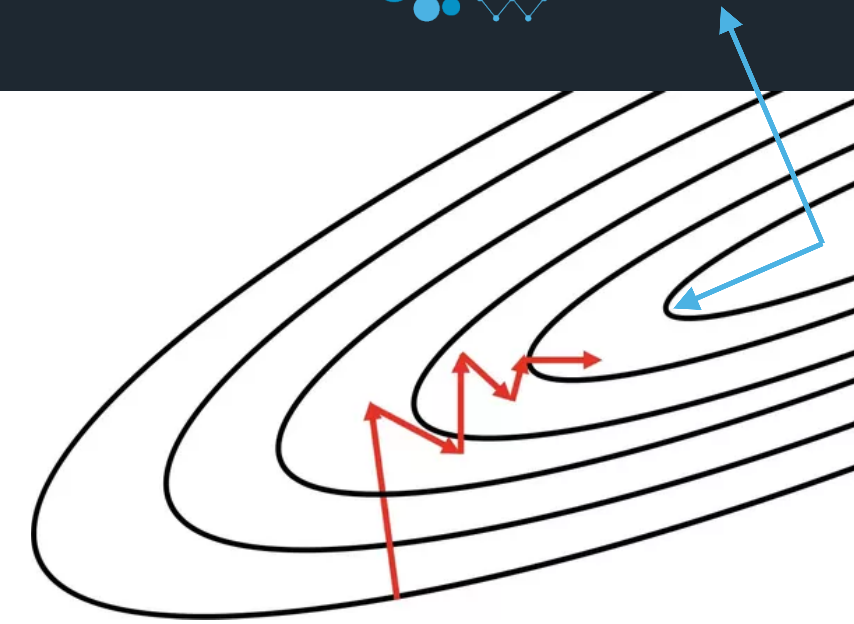
$$\dot{w} = -H(w^*)(w - w^*)$$

$$u = w - w^*$$

$$\frac{u'}{u} = -H(w^*) \longrightarrow ln(u) = -H(w^*)t + C \longrightarrow w - w^* = w(0)e^{-H(w^*)t}$$

Generalization Dynamics

Our three cents   With help of Remi Tachet



$$w \leftarrow w - \eta \frac{\partial L}{\partial w}$$

$$\tau \dot{w} = - \frac{\partial L}{\partial w}$$

Taylor expansion:

$$f(x) = f(a) + f'(a)(x-a)$$

$$f(w) = \frac{\partial L}{\partial w}$$

$$f(w) = \left.\frac{\partial L}{\partial w}\right|_{w=w^*} + \left.\frac{\partial^2 L}{\partial w^2}\right|_{w=w^*} (w - w^*)$$

$$f(w) = H(w^*)(w - w^*)$$

$$\dot{w} = -H(w^*)(w - w^*)$$

$$u = w - w^*$$

$$\frac{u'}{u} = -H(w^*) \longrightarrow ln(u) = -H(w^*)t + C \longrightarrow w - w^* = w(0)e^{-H(w^*)t}$$

$$\alpha_i = \alpha_i(0)e^{-\lambda_i t}$$

Generalization Dynamics

Related Work

- **Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data** (Gintare Karolina Dziugaite & Daniel M. Roy)

  - Concern: overfitting since #model parameters >> # available data points
  - However SGD returns solutions with low test errors on deep models.
  - Nonvacuous generalization bounds for:
    - deep stochastic neural network classifiers
    - with millions of parameters trained on only tens of thousands of examples
    - Extension of Langrod's PAC Bayes

- **Why and When Can Deep – but Not Shallow – Networks Avoid the Curse of Dimensionality: a Review** (Poggio et al)

  - Mostly focuses on power of architectures (what it can approximate and learn)
  - Studies the learning process: the unreasonable efficiency of SGD
  - Talks about generalization: over-parametrization is ok and over fitting is not that big of a problem in deep networks rather than in classical shallow networks

MILA

Merci :)