Lecture 14 Deep Learning Theory

IEMS 402 Statistical Learning

Northwestern

References

https://www.di.ens.fr/~fbach/ltfp_book.pdf

- Section 12

Neural Tangent Kernel

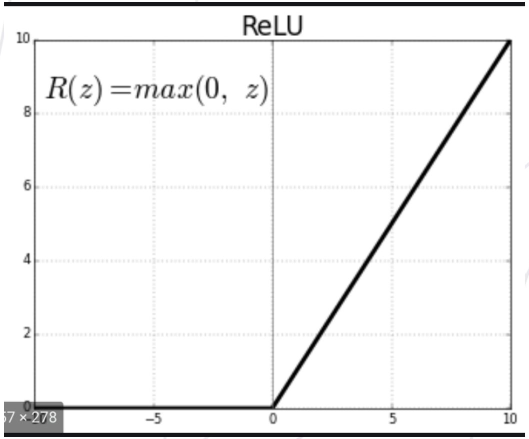
Neural Tangent Theory

Minimizing $F(w) := R(h(w))$

Consider a linearized model $\bar{F}(w) := R(h(w_0) + \nabla_w h(w_0)(w - w_0))$

lazy training the less expected situation where these two paths remain close until the algorithm is stopped.

Homogenous activation



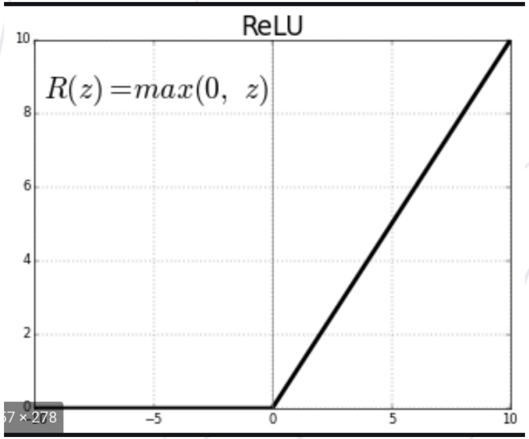
$$\text{relu}(5x) = 5\text{relu}(x)$$

Network 1 *Network 2*

$$w_2\text{relu}(\underbrace{(5w_1)}_{\tilde{w}_1}\,x) = \underbrace{(5w_2)}_{\tilde{w}_2}\,\text{relu}(w_1 x)$$

What's the thing different? $\nabla_{\tilde{w}_1} \neq \nabla_{w_1}, \nabla_{w_2} \neq \nabla_{\tilde{w}_2}$

Homogenous activation



$$\text{relu}(5x) = 5\,\text{relu}(x)$$

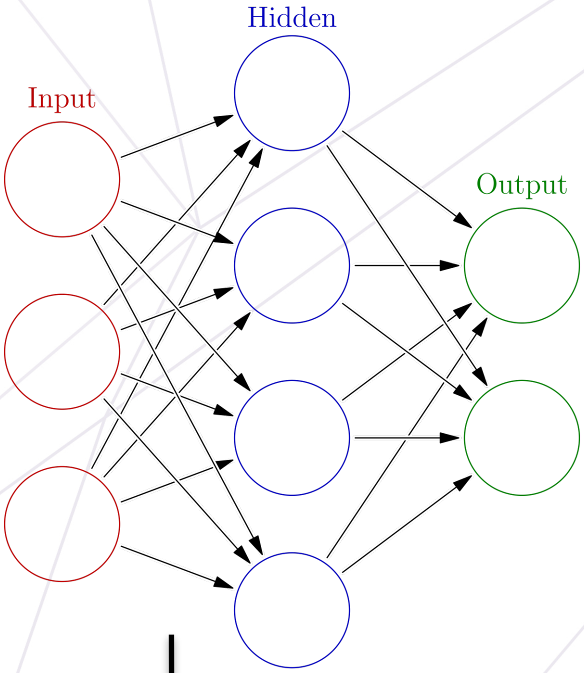
Network 1 *Network 2*

$$w_2\text{relu}(\underbrace{(5w_1)}_{\tilde{w}_1}\, x) = \underbrace{(5w_2)}_{\tilde{w}_2}\,\text{relu}(w_1 x)$$

What's the thing different? $\nabla_{\tilde{w}_1} \neq \nabla_{w_1}, \nabla_{w_2} \neq \nabla_{\tilde{w}_2}$

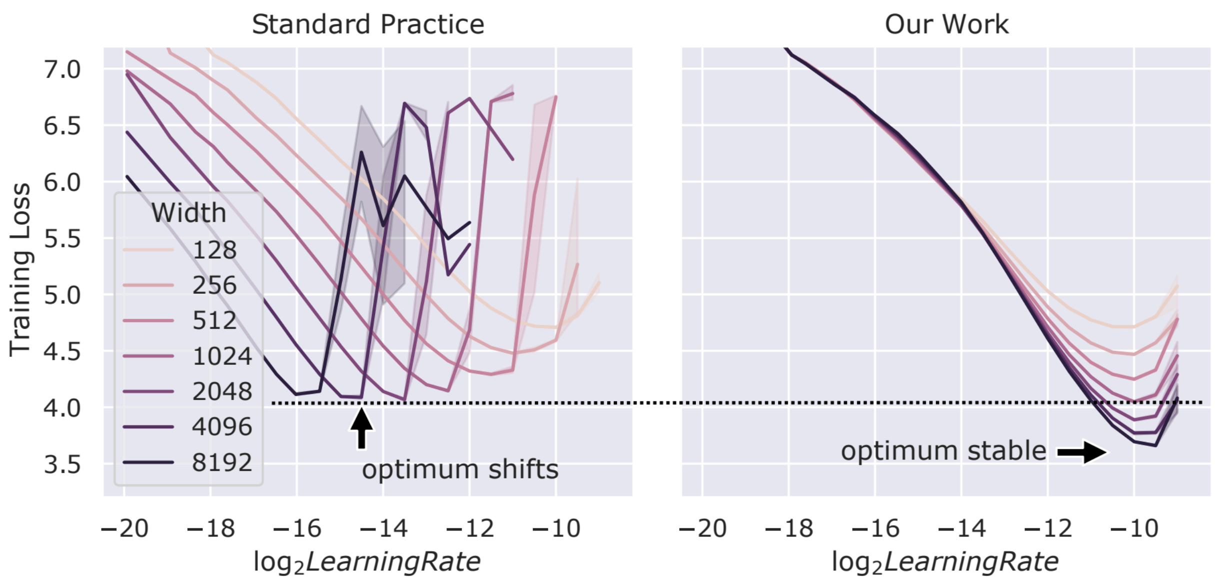
Is Adam/muon dynamics the same for two network?

Take Home Message



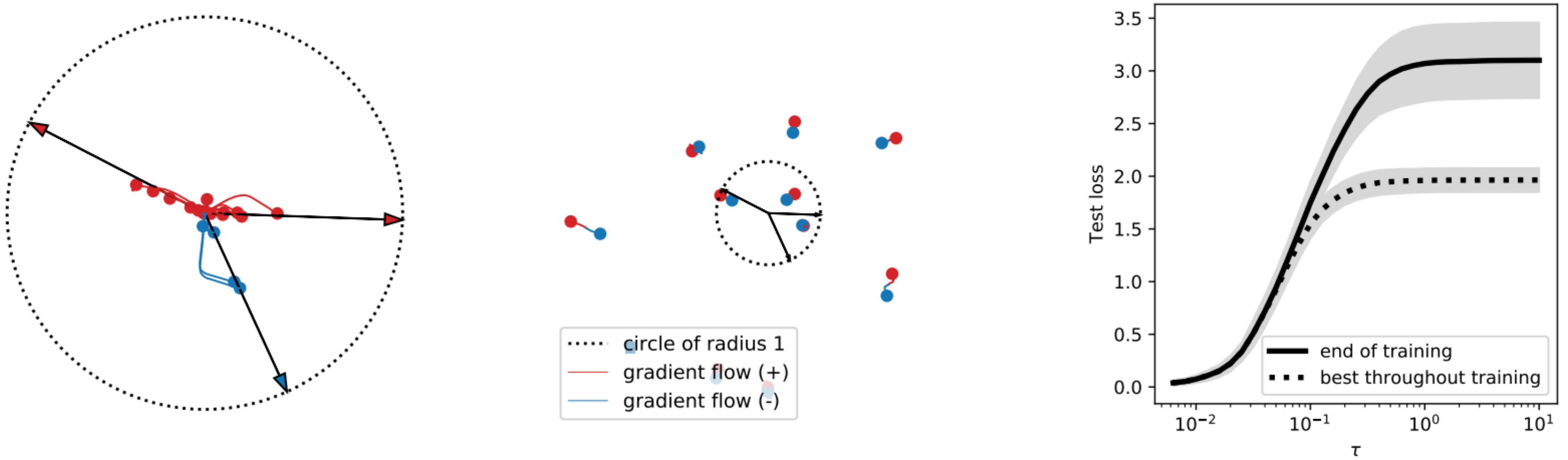
Wider network has smaller gradient on first layer

Learning rate transfer



https://arxiv.org/pdf/2203.03466

Feature Learning and Lazy Learning



(a) Non-lazy training ($\tau = 0.1$) (b) Lazy training ($\tau = 2$) (c) Generalization properties

Chizat, Lenaic, Edouard Oyallon, and Francis Bach. "On lazy training in differentiable programming." Advances in neural information processing systems 32 (2019).

When Lazy Trainning occurs?

Gradient descent $w_1 := w_0 - \eta \nabla F(w_0),$

Relative change of objective function $\Delta(F) := \frac{|F(w_1) - F(w_0)|}{F(w_0)} \approx \eta \frac{\|\nabla F(w_0)\|^2}{F(w_0)}.$

Relative change of linearization

$$\Delta(Dh) := \frac{\|Dh(w_1) - Dh(w_0)\|}{\|Dh(w_0)\|} \leq \eta \frac{\|\nabla F(w_0)\| \cdot \|D^2 h(w_0)\|}{\|Dh(w_0)\|}.$$

$$\boxed{\kappa_h(w_0) := \|h(w_0) - y^\star\| \frac{\|D^2 h(w_0)\|}{\|Dh(w_0)\|^2} \ll 1,}$$

Example: Lazy Training For Homogeneous Model

Homogeneous models. If h is q-positively homogeneous[4] then multiplying the initialization by λ is equivalent to multiplying the scale factor α by λ^q. In equation,

$$\kappa_h(\lambda w_0) = \frac{1}{\lambda^q} \| \lambda^q h(w_0) - y^\star \| \frac{\|D^2 h(w_0)\|}{\|Dh(w_0)\|^2}.$$

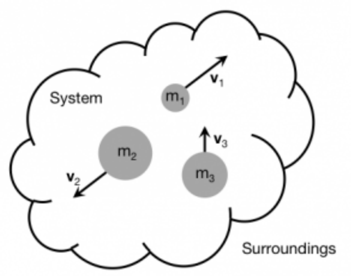
Mean Filed Theory

Mean Field Theory

$$h(x) = \frac{1}{m} \sum_{j=1}^{m} \eta_j \sigma(w_j^\top x + b_j), \longrightarrow h = \frac{1}{m} \sum_{j=1}^{m} \Psi(v_j).$$

Reformulate as probability distribution:

$$h = h(\cdot, v_1, \ldots, v_m) = \int_{\mathcal{V}} \Psi(v) d\mu(v),$$



Gradient Flow in Wasserstein Sapce

Gradient descent in weight = Gradient flow in Wasserstein space

Implicit Bias

Convergence in direction

$$F(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-y_i x_i^\top \theta)),$$

Convergence In direction!

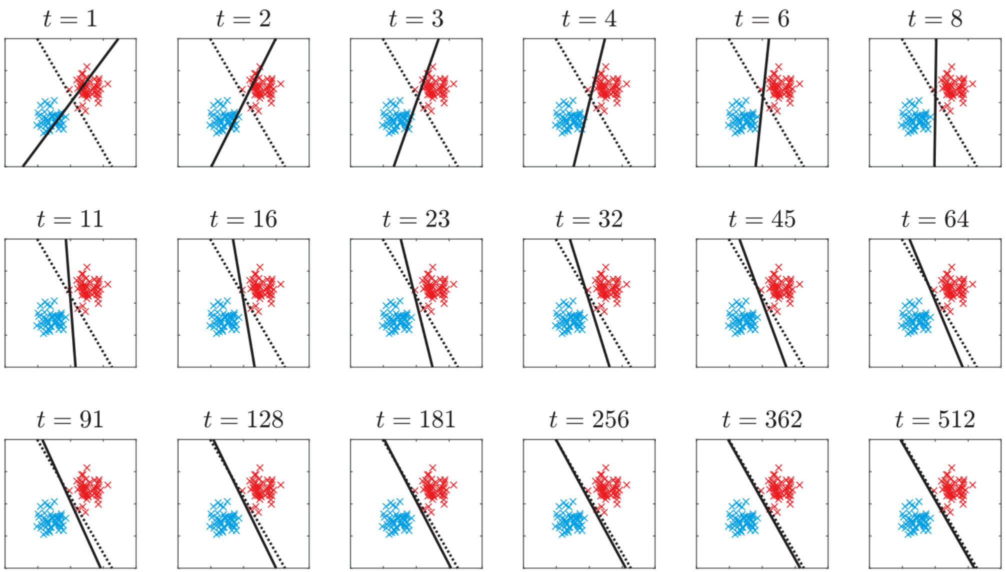
KKT condition for Largest Margin

$\min \|\theta\|_2^2$ subject to $y_i(\theta_i \cdot x) \geq 1$

SVM=Logisitic Regression

$$G'(\theta) = \frac{-\sum_{i=1}^{n} y_i x_i \exp(-y_i x_i^\top \theta)}{\sum_{i=1}^{n} \exp(-y_i x_i^\top \theta)} = -\sum_{i=1}^{n} \alpha_i y_i x_i$$

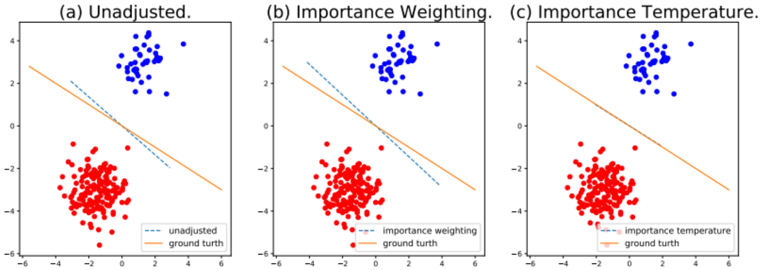
Converge to SVM solution



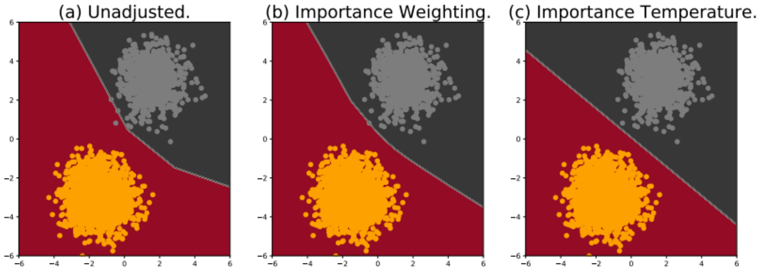
Where is the support vectors

$$F'(\theta_t) \sim -\frac{1}{n}\sum_{i\in I} y_i \exp(-\|\theta_t\|_2 y_i x_i^\top \eta) x_i.$$

Failure of Importance Weighting



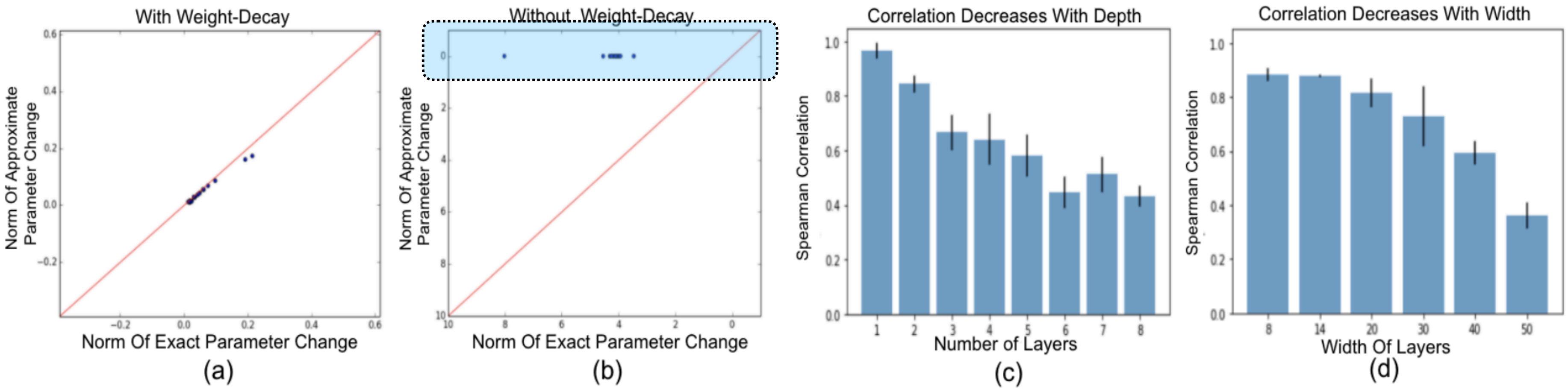
(a) Linear Model for Separable Data



(b) Multilayer Perceptron with two hidden layers of size 200

Byrd J, Lipton Z. What is the effect of importance weighting in deep learning? International conference on machine learning. PMLR, 2019: 872-881.

Failure of Influnce Function



https://arxiv.org/pdf/2006.14651