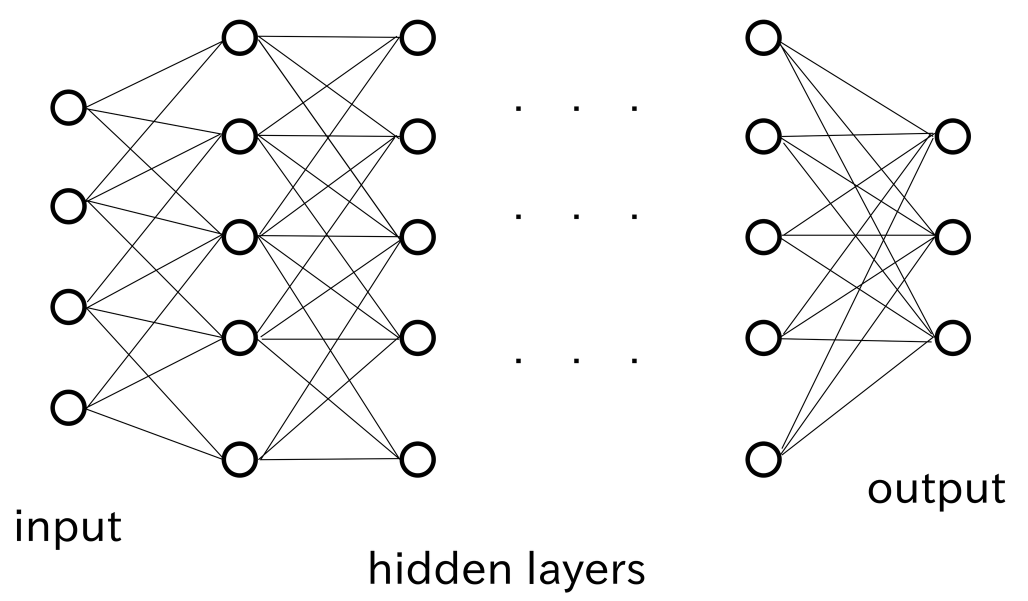
How Regularization Affects the Critical Points in Linear Networks

Amirhossein Taghvaei, Jin Kim, Prashant Mehta
Coordinated Science Laboratory, University of Illinois at Urbana-Champaign

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA

Motivation

Deep learning



Objective: Study the optimization problem in feedforward neural networks

This work: Analyze critical points of a linear network (with regularization)

Why linear network: They exhibit same behavior as nonlinear networks in learning. They are easier to analyze.

A. M. Saxe, et. al. (2013) Exact solutions to the nonlinear dynamics . . .

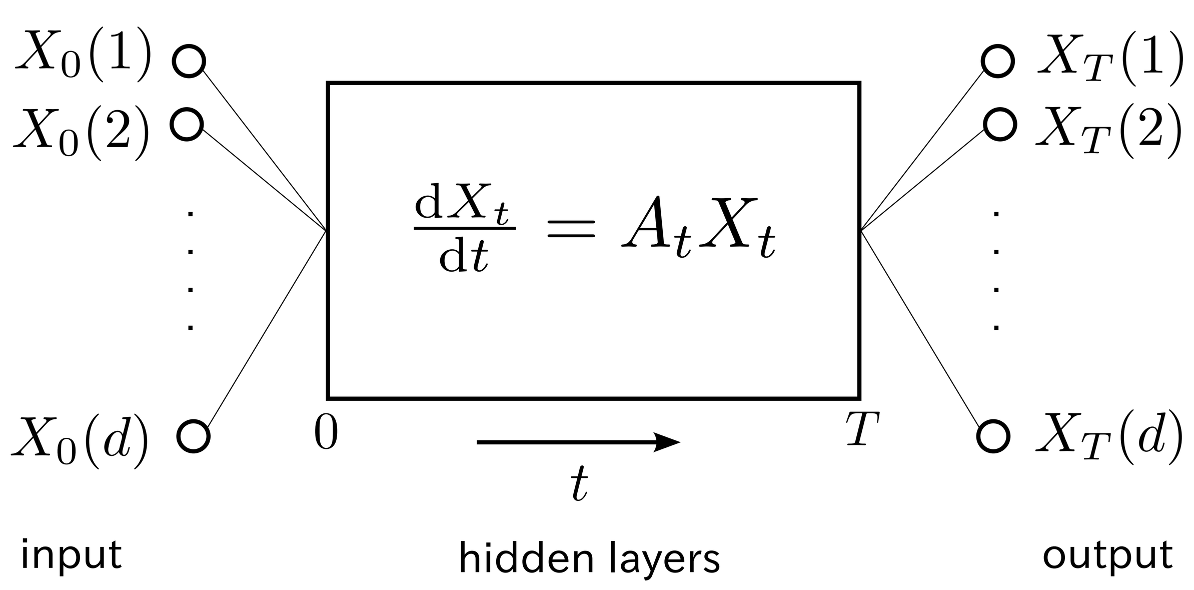
M. Hardt and T. Ma. (2016) Identity matters in deep learning

S. Gunasekar, et. al. (2017) Implicit regularization in matrix factorization

Problem setup

Network model:

Linear continuous neural network



$$X_{l+1} = \sigma(W_l X_l) \xrightarrow[\sigma \text{ is identity}]{\text{continuum limit}} \frac{dX_t}{dt} = A_t X_t$$

Data model:

Data: $(X_0, Z_0) \in \mathbb{R}^d \times \mathbb{R}^d$

Model: $Z = \underbrace{RX_0}_{\text{linear model}} + \underbrace{\xi}_{\text{noise}}$

Assumptions:

$\Sigma := \mathrm{E}[X_0 X_0^\top]$ is invertible

$\log(R)$ exists

Why continuous netowrk: Analysis is simpler and results are insightful

Optimization problem

Optimal control formulation

$$\underset{A}{\text{Minimize:}} \quad \mathrm{J}[A] = \underbrace{\frac{\lambda}{2}\int_0^T \mathrm{tr}(A_t^\top A_t)\,\mathrm{d}t}_{\text{regularization}} + \underbrace{\frac{1}{2}\,\mathrm{E}\left[|X_T - Z|^2\right]}_{\text{mean-squared loss}}$$

$$\text{Subject to:} \quad \frac{\mathrm{d}X_t}{\mathrm{d}t} = A_t X_t, \quad X_0 \sim p_0$$

$(\lambda = 0)$: No regularization.

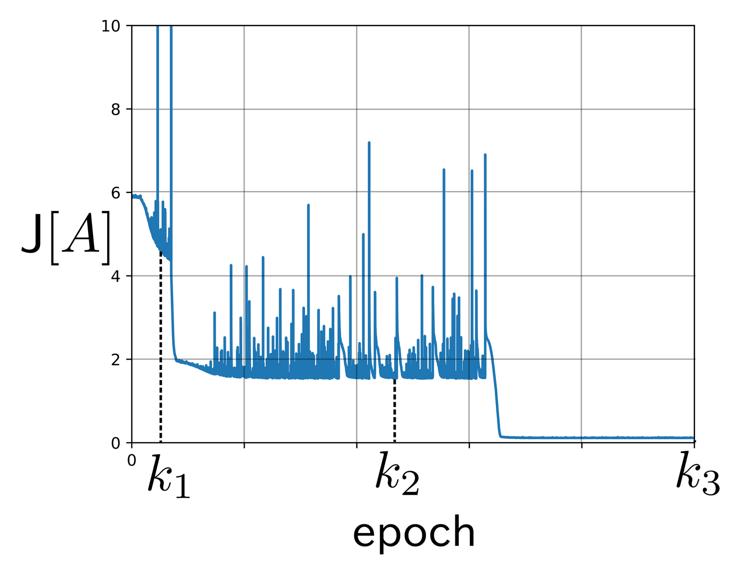
$(\lambda > 0)$: Explicit regularization

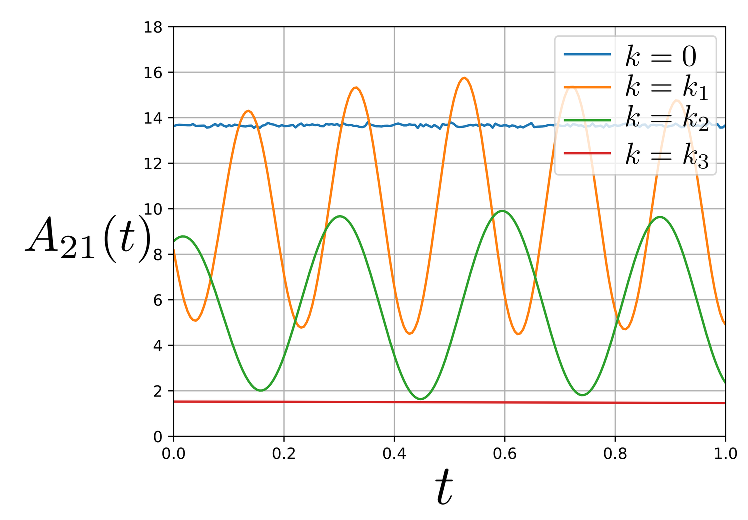
$(\lambda = 0^+)$: The limit as $\lambda \to 0$. Implicit regularization [B. Neyshabur, 2017]

Learning example

Example: R is a rotation matrix

$$Z = \underbrace{\begin{bmatrix} 0 & +1 \\ -1 & 0 \end{bmatrix}}_{R} X_0 + \xi, \quad X_0 \sim N(0, I_{2\times 2}), \quad \lambda = 0.03$$





Questions:

What are the critical points that learning get stuck at?

Is the global minimizer always constant?

Why are some critical points constant and some not?

Approach: Optimal control theory

Hamiltonian function:

$$H(x, y, B) = y^\top Bx - \frac{\lambda}{2}\mathrm{tr}(B^\top B)$$

where $x, y \in \mathbb{R}^d$ and $B \in \mathbb{R}^{d\times d}$

Pontryagin's maximum principle: A_t is the critical point iff there exists a random process $Y : [0, T] \to \mathbb{R}^d$ such that

$$\text{(Forward eq.)} \quad \frac{\mathrm{d}X_t}{\mathrm{d}t} = +\frac{\partial \mathrm{H}}{\partial y}(X_t, Y_t, A_t) = +A_t X_t, \quad X_0 \sim p_0$$

$$\text{(Backward eq.)} \quad \frac{\mathrm{d}Y_t}{\mathrm{d}t} = -\frac{\partial \mathrm{H}}{\partial x}(X_t, Y_t, A_t) = -A_t^\top Y_t, \quad Y_T = \underbrace{Z - X_T}_{\text{error}}$$

$$A_t = \underset{B \in M_d(\mathbb{R})}{\arg\max}\ \mathrm{E}[\mathrm{H}(X_t, Y_t, B)] = \frac{1}{\lambda}\mathrm{E}[Y_t X_t^\top]$$

Result: Critical points (no regularization)

Definition: Φ_t is the state transition matrix for $\frac{\mathrm{d}X_t}{\mathrm{d}t} = A_t X_t$ s.t $X_t = \Phi_t X_0$,

Proposition:

Any A_t such that $(\Phi_T - R)\Sigma = 0$ is a critical point

All critical points are global minimizers

$$\nabla \mathrm{J}[A] = 0 \quad \Leftrightarrow \quad \mathrm{J}[A] = \min_V \mathrm{J}[V] =: J^*$$

The optimality gap is upper-bounded by the gradient

$$\|\nabla \mathrm{J}[A]\|_{L^2}^2 \geq Te^{-2\int_0^T \|A_t\|_F \mathrm{d}t}\lambda_{\min}(\Sigma)(\mathrm{J}[A] - \mathrm{J}^*)$$

Result: Critical points (with regularization)

Proposition: The critical points are given by solutions to the characteristic equation:

$$(\lambda > 0): \quad \lambda \mathrm{C} = e^{T\mathrm{C}} e^{T(\mathrm{C}^\top - \mathrm{C})}(R - e^{T(\mathrm{C}-\mathrm{C}^\top)} e^{T\mathrm{C}^\top})\Sigma$$

$$(\lambda = 0^+): \quad e^{T(\mathrm{C}-\mathrm{C}^\top)} e^{T\mathrm{C}^\top} = R \quad \text{(characteristic eq.)}$$

And the weights are

$$A_t = e^{t(\mathrm{C}-\mathrm{C}^\top)} \mathrm{C} e^{-t(\mathrm{C}-\mathrm{C}^\top)}$$

Corollary:

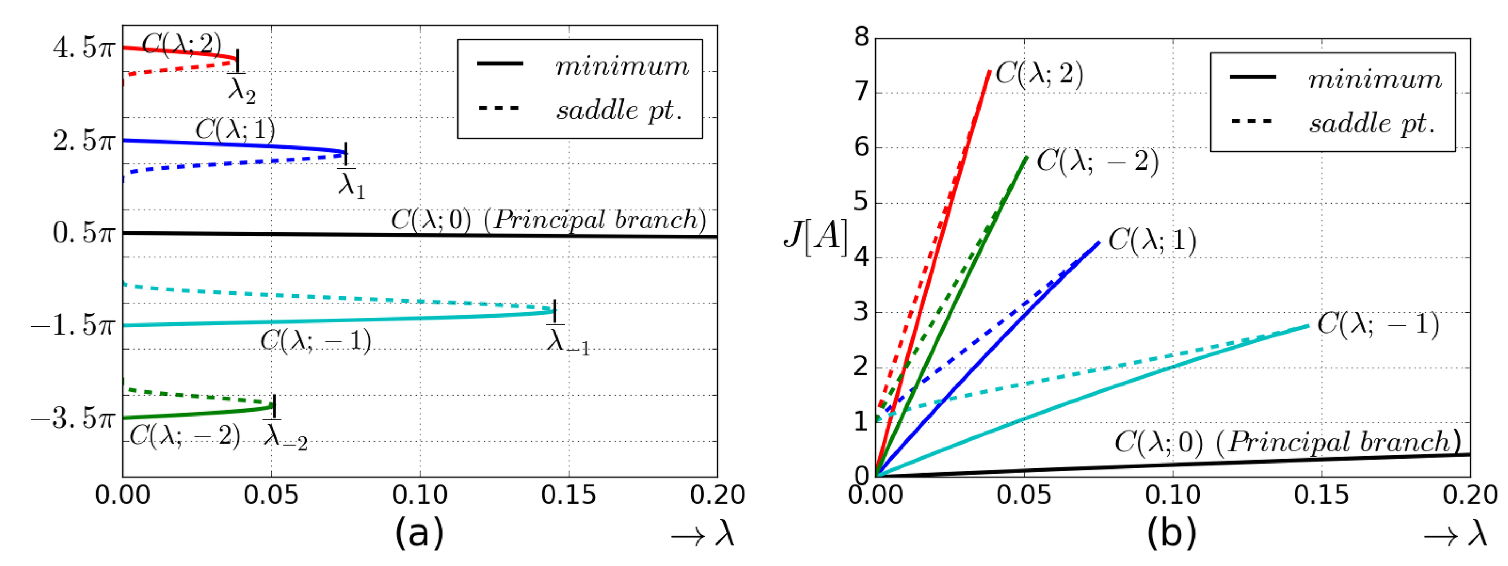
$$\mathrm{C} \text{ is normal } (\mathrm{C}^\top \mathrm{C} = \mathrm{C}\mathrm{C}^\top) \quad \Longleftrightarrow \quad A_t \text{ is constant}$$

Example

Example: R is rotation matrix

Normal critical points:

$$\mathrm{C}(\lambda; n) = \underbrace{\begin{bmatrix} 0 & -(\pi/2 + 2n\pi) \\ \pi/2 + 2n\pi & 0 \end{bmatrix}}_{\log(R)} + O(\lambda), \quad n \in \mathbb{Z}$$



Non-normal critical points: No result

Future work

Non-normal critical points

Second order analysis

Generalization error

Acknowledgment

National Science Foundation grants 1334987 and 1462773