How Regularization Affects Critical Points in Linear Network

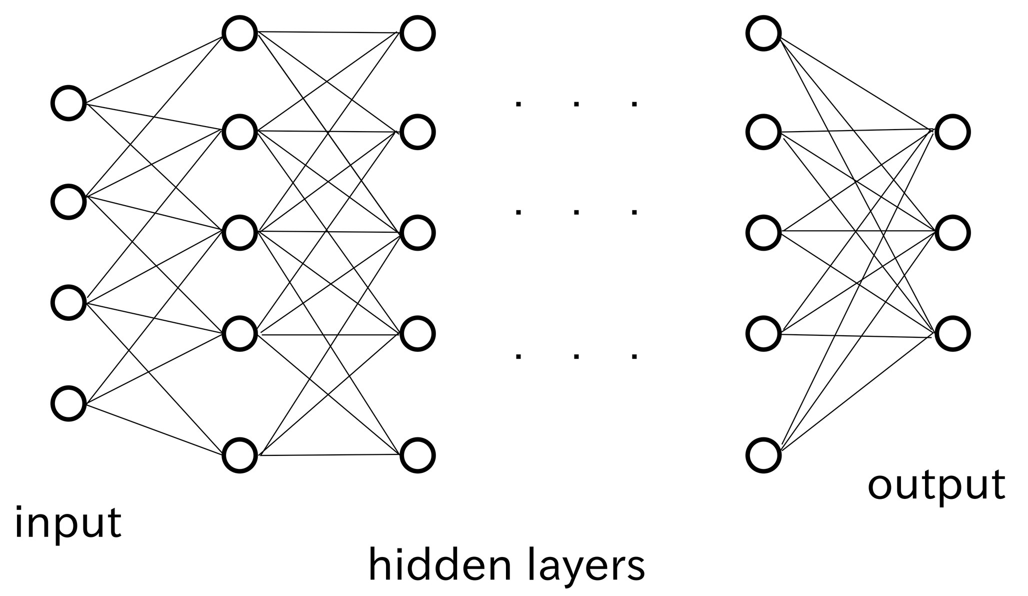
Amirhossein Taghvaei, Jin Kim, Prashant Mehta

Coordinated Science Laboratory, University of Illinois at Urbana-Champaign

Midwest Machine Learning Symposium, Chicago, 2017

Motivation

Deep learning



Objective: Analysis of the critical points of the associated non-convex optimization problem

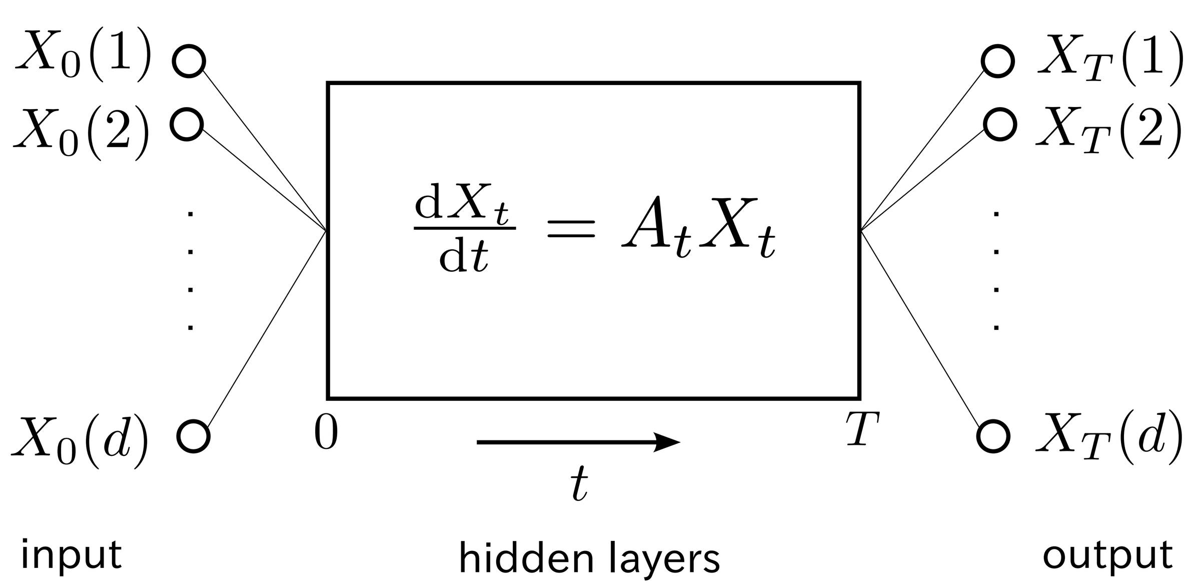
This work: Analysis of the critical points of a linear network (with regularization)

Related work:

- A. M. Saxe, et. al. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks (2013)
- M. Hardt and T. Ma. Identity matters in deep learning (2016).

Problem formulation

Linear continuous neural network



Model:

Data: $X_0 \in \mathbb{R}^d$ with $X_0 \sim p_0$

Measurement: $Z = RX_0 + \xi \in \mathbb{R}^d$

- R is a $d \times d$ matrix
- ξ is noise with mean zero
- $\Sigma := \mathrm{E}[X_0 X_0^\top]$ is invertible

Problem: Learn the linear transformation R with the linear continuous NN

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 $\lambda \to 0$. Models the dissipation in the learning.

Critical points and characteristic equation

Main result:

- $(\lambda = 0)$: Any A_t such that $\Phi_{0,T} = R$ where $\Phi_{0,t}$ is the state transition matrix corresponding to $\frac{\mathrm{d}X_t}{\mathrm{d}t} = A_t X_t$.
- $(\lambda > 0)$: $A_t = e^{t(\mathrm{C}-\mathrm{C}^\top)}\mathrm{C}e^{-t(\mathrm{C}-\mathrm{C}^\top)}$ where C is a solution of
$$\lambda \mathrm{C} = F^\top (R - F)\Sigma$$
where $F := e^{T(\mathrm{C}-\mathrm{C}^\top)}\,e^{T\mathrm{C}^\top}$.
- $(\lambda = 0^+)$: $A_t = e^{t(\mathrm{C}-\mathrm{C}^\top)}\mathrm{C}e^{-t(\mathrm{C}-\mathrm{C}^\top)}$ where C is a solution of
$$e^{T(\mathrm{C}-\mathrm{C}^\top)}\,e^{T\mathrm{C}^\top} = R$$

Hamilton's formulation

Hamiltonian function:

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

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

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

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = +\frac{\partial \mathrm{H}}{\partial y}(X_t, Y_t, A_t) = +A_t X_t, \quad X_0 \sim p_0$$
$$\frac{\mathrm{d}Y_t}{\mathrm{d}t} = -\frac{\partial \mathrm{H}}{\partial x}(X_t, Y_t, A_t) = -A_t^\top Y_t, \quad Y_T = Z - X_T$$

and A_t maximizes the expected value of the Hamiltonian

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

Backpropagation (with dissipation)

First order variation:

$$\nabla \mathrm{J}[A] := -\mathrm{E}\left[\frac{\partial \mathrm{H}}{\partial B}(X_t, Y_t, A_t)\right] = \lambda A_t - \mathrm{E}\left[Y_t X_t^\top\right]$$

where X_t and Y_t are obtained by solving the Hamilton's equations.

Stochastic gradient-descent:

$$A_t^{(k+1)} = A_t^{(k)} - \eta_k(\lambda A_t^{(k)} - Y_t^{(k)} X_t^{(k)\top}),$$

where η_k is the step-size and $X_t^{(k)}$ and $Y_t^{(k)}$ are obtained by solving the Hamilton's equations:

(Forward propagation) $\quad \frac{\mathrm{d}}{\mathrm{d}t}X_t^{(k)} = +A_t^{(k)} X_t^{(k)}, \quad$ with init. cond. $X_0^{(k)}$

(Backward propagation) $\quad \frac{\mathrm{d}}{\mathrm{d}t}Y_t^{(k)} = -A_t^{(k)\top} Y_t^{(k)}, \quad Y_T^{(k)} = \underbrace{Z^{(k)} - X_T^{(k)}}_{\text{error}}$

based on the sample $(X^{(k)}, Z^{(k)})$.

Examples

Case I: $R \in \mathbb{R}$ is a positive scalar

- $(\lambda = 0)$: $A_t = \frac{1}{T}\log(R) + B_t$ for any B_t s.t $\int_0^T B_t = 0$
- $(\lambda > 0)$: $A_t = \mathrm{C} = \frac{1}{T}\log(R) + O(\lambda)$
- $(\lambda = 0^+)$: $A_t = \mathrm{C} = \frac{1}{T}\log(R)$ (minimum norm solution)

Case II: R is a normal matrix with $\det(R) > 0$.

- $(\lambda > 0)$: All the constant solutions are
$$A_t = \mathrm{C} = \frac{1}{T}\log(R) + O(\lambda)$$
where $\log(R)$ is multi-valued

Example: $R = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$, $\log(R)_n = (\pi/2 + 2n\pi)\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$, $n \in \mathbb{Z}$

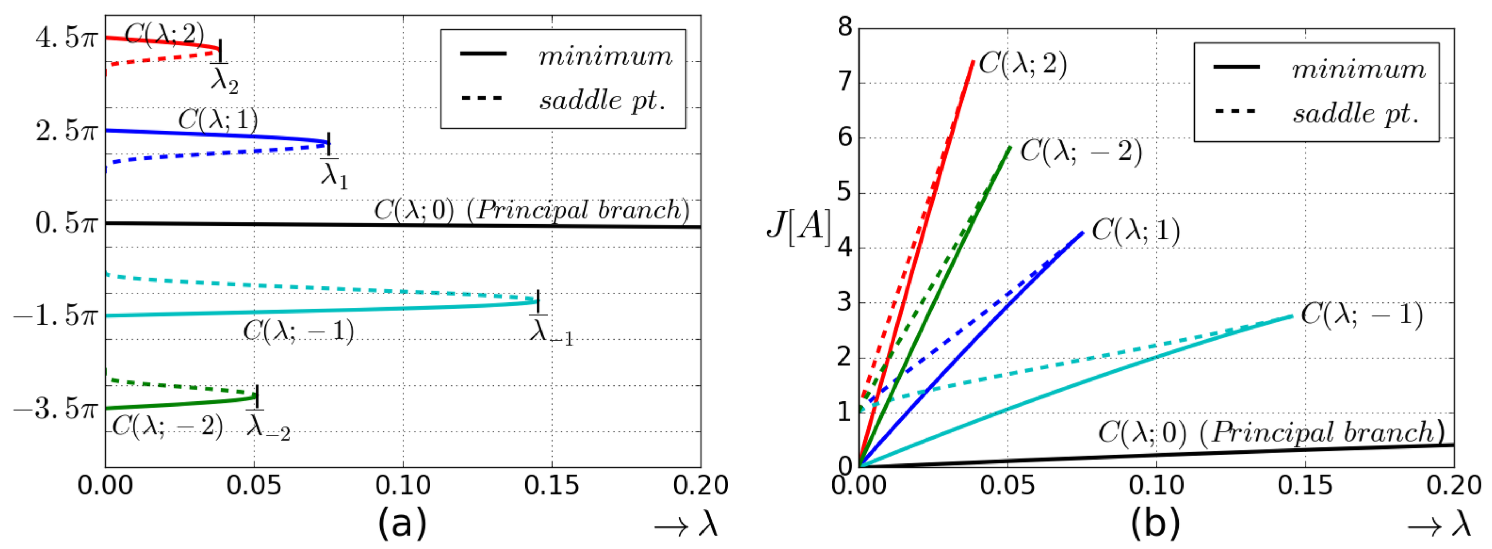


Figure : (a) Critical points (the $(2, 1)$ entry of the solution matrix $C(\lambda; n)$) is depicted for $n = 0, \pm1, \pm2$); (b) The cost $J[A]$ for these solutions.

Future work

1. Stability analysis of the critical points
2. Introducing nonlinearity to the network

Acknowledgment

National Science Foundation grants 1334987 and 1462773