Critical Points of Linear Networks

UIUC ML seminar, Fall, 2017

Amirhossein Taghvaei

Joint work with Jin W. Kim, and Prashant G. Mehta

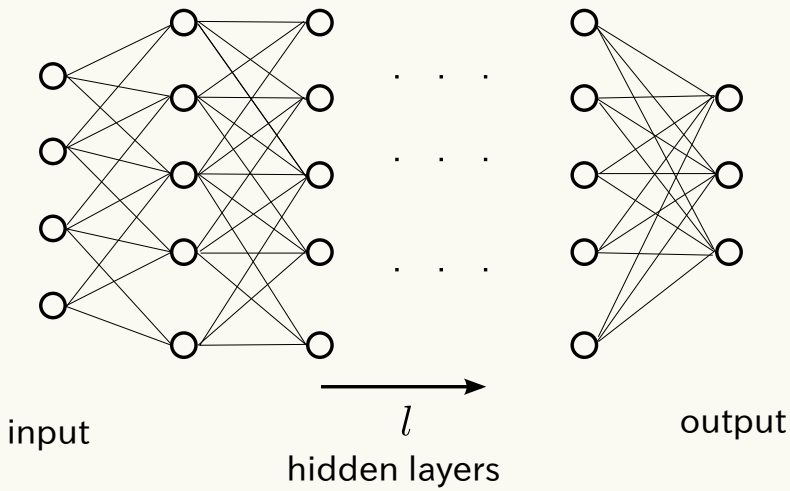
Coordinated Science Laboratory

University of Illinois at Urbana-Champaign

Sep 29, 2017

ILLINOIS

Motivation



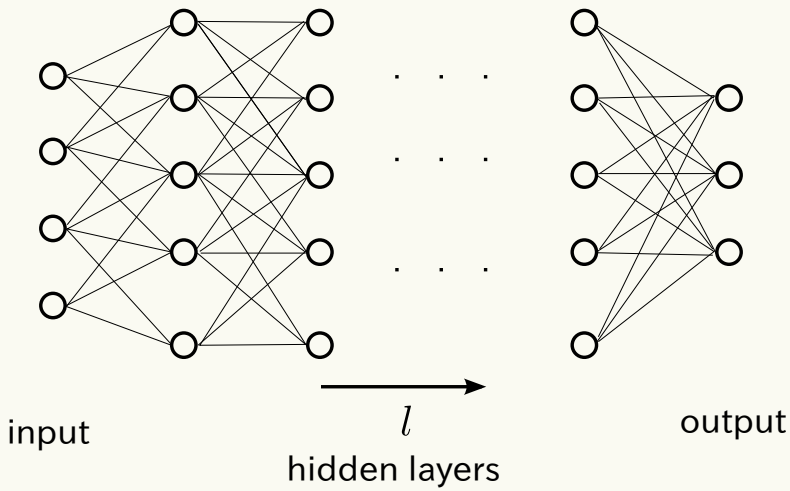
Objective: Study the optimization problem in feedforward NN

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

Related work:

- A. M. Saxe, et. al. (2013) Exact solutions to the nonlinear dynamics . . .
- M. Hardt, T. Ma. (2016) Identity matters in deep learning
- S. Gunasekar, et. al. (2017) Implicit regularization in matrix factorization

Motivation



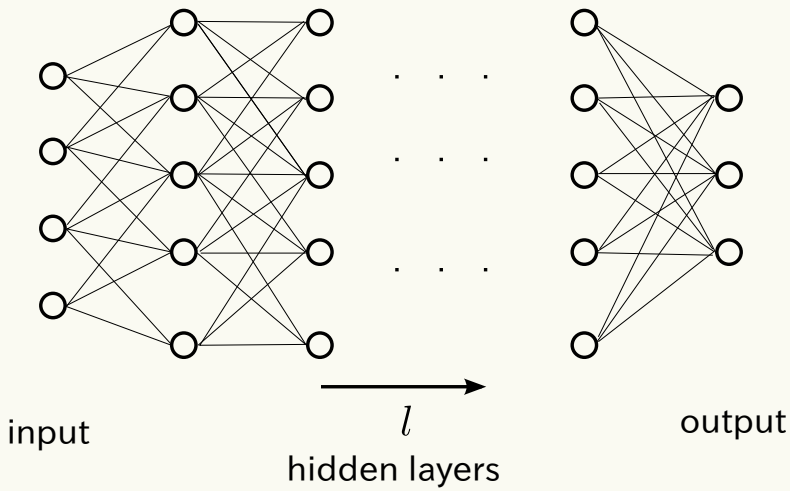
Objective: Study the optimization problem in feedforward NN

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

Related work:

- A. M. Saxe, et. al. (2013) Exact solutions to the nonlinear dynamics ...
- M. Hardt, T. Ma. (2016) Identity matters in deep learning
- S. Gunasekar, et. al. (2017) Implicit regularization in matrix factorization

Motivation



Objective: Study the optimization problem in feedforward NN

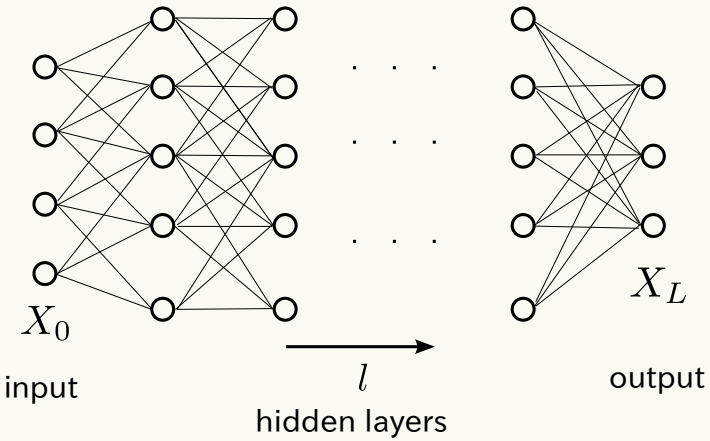
This work: Analysis of the critical points of a linear network (with regularization)

Related work:

- A. M. Saxe, et. al. (2013) Exact solutions to the nonlinear dynamics ...
- M. Hardt, T. Ma. (2016) Identity matters in deep learning
- S. Gunasekar, et. al. (2017) Implicit regularization in matrix factorization

Network model

Feedforward NN



$$X_{l+1} = \sigma(W_l X_l)$$

X_0 : input

X_L : output

W_l : weights

Linear continuous network

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = A_t X_t$$

X_0 : initial value

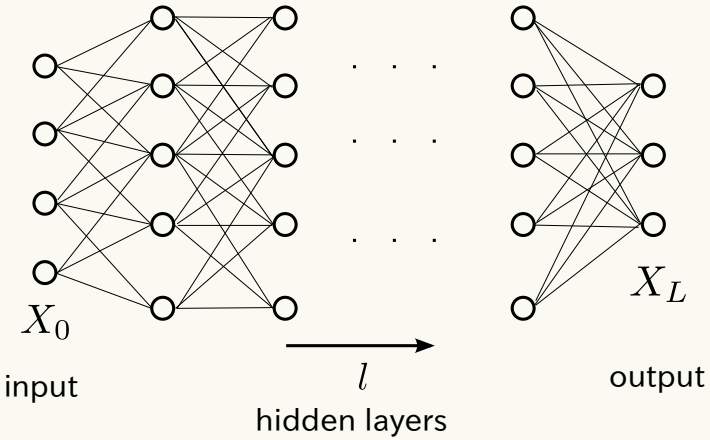
X_T : terminal value

A_t : control variable

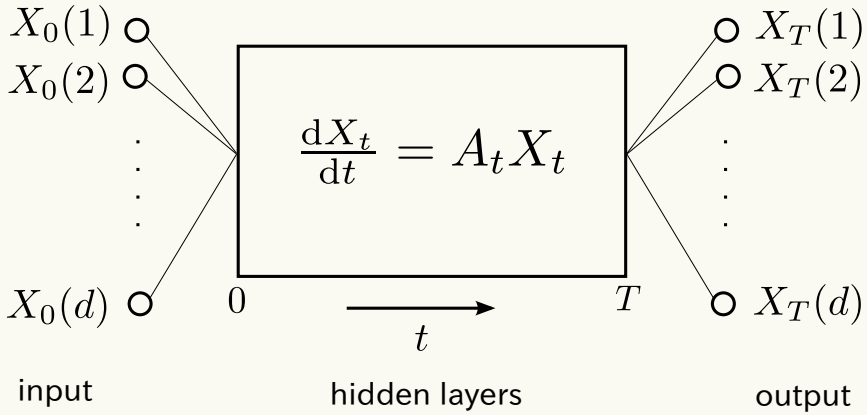
Discretization: $X_{t+\Delta t} = X_t + \Delta t A_t X_t$ (Res net)

Network model

Feedforward NN



Linear continuous network



$$X_{l+1} = \sigma(W_l X_l)$$

X_0 : input

X_L : output

W_l : weights

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = A_t X_t$$

X_0 : initial value

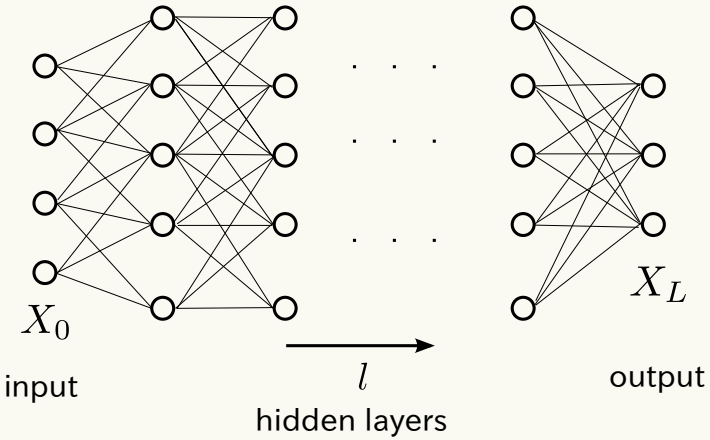
X_T : terminal value

A_t : control variable

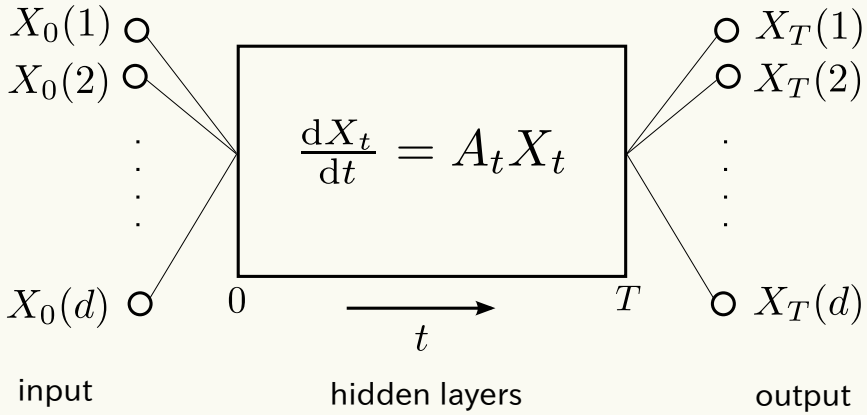
Discretization: $X_{t+\Delta t} = X_t + \Delta t A_t X_t$ (Res net)

Network model

Feedforward NN



Linear continuous network



$$X_{l+1} = \sigma(W_l X_l)$$

X_0 : input

X_L : output

W_l : weights

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = A_t X_t$$

X_0 : initial value

X_T : terminal value

A_t : control variable

Discretization: $X_{t+\Delta t} = X_t + \Delta t A_t X_t$ (Res net)

Problem formulation

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

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

Optimization problem:

$$\underset{A}{\text{Minimize:}} \quad \mathrm{J}[A] = \frac{1}{2} \underbrace{\mathsf{E}\left[|X_T - Z|^2\right]}_{\text{mean-squared loss}} + \underbrace{\lambda \|A\|_F^2}_{\text{regularization}}$$

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

- $(\lambda = 0)$: No regularization.
- $(\lambda > 0)$: Explicit regularization
- $(\lambda = 0^+)$: Implicit regularization

Problem formulation

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

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

Optimization problem:

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

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

- $(\lambda = 0)$: No regularization.
- $(\lambda > 0)$: Explicit regularization
- $(\lambda = 0^+)$: Implicit regularization

Problem formulation

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

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

Optimization problem:

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

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

- $(\lambda = 0)$: No regularization.
- $(\lambda > 0)$: Explicit regularization
- $(\lambda = 0^+)$: Implicit regularization

Outline

1. Motivation
2. Problem formulation
3. Scalar case
4. Optimal control formulation
5. Main result: Characterization of critical points.

Outline

Scalar case ($d = 1$)

$$\text{Model:} \quad Z = RX_0 + \xi$$

$$\underset{A}{\text{Minimize:}} \quad \mathsf{J}[A] = \frac{1}{2}\mathsf{E}\left[|X_T - Z|^2\right] + \frac{\lambda}{2}\int_0^T A_t^2 \, \mathrm{d}t$$

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

Proposition

Assume $R > 0$. Then the critical points are

- $(\lambda = 0)$: $A_t = \frac{1}{T}\log(R) + B_t$ s.t $\int_0^T B_t \, \mathrm{d}t = 0$
- $(\lambda > 0)$: $A_t = \mathsf{C}_\lambda$ (constant)
- $(\lambda = 0^+)$: $A_t = \mathsf{C}_0 = \frac{1}{T}\log(R)$ (unique, minimum norm)

$$\lambda \mathsf{C} = e^{T\mathsf{C}}(R - e^{T\mathsf{C}})\mathsf{E}[X_0^2]$$

Scalar case ($d = 1$)

$$\text{Model:} \quad Z = RX_0 + \xi$$

$$\underset{A}{\text{Minimize:}} \quad \mathsf{J}[A] = \frac{1}{2}\mathsf{E}\left[|X_T - Z|^2\right] + \frac{\lambda}{2}\int_0^T A_t^2 \, \mathrm{d}t$$

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

Proposition

Assume $R > 0$. Then the critical points are

- $(\lambda = 0)$: $A_t = \frac{1}{T}\log(R) + B_t$ s.t $\int_0^T B_t \, \mathrm{d}t = 0$
- $(\lambda > 0)$: $A_t = \mathsf{C}_\lambda$ (constant)
- $(\lambda = 0^+)$: $A_t = \mathsf{C}_0 = \frac{1}{T}\log(R)$ (unique, minimum norm)

$$\lambda \mathsf{C} = e^{T\mathsf{C}}(R - e^{T\mathsf{C}})\mathsf{E}[X_0^2]$$

Scalar case ($d = 1$)

$$\text{Model:} \quad Z = RX_0 + \xi$$

$$\underset{A}{\text{Minimize:}} \quad \mathsf{J}[A] = \frac{1}{2}\mathsf{E}\left[|X_T - Z|^2\right] + \frac{\lambda}{2}\int_0^T A_t^2 \, \mathrm{d}t$$

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

Proposition

Assume $R > 0$. Then the critical points are

- $(\lambda = 0)$: $A_t = \frac{1}{T}\log(R) + B_t$ s.t $\int_0^T B_t \, \mathrm{d}t = 0$
- $(\lambda > 0)$: $A_t = \mathsf{C}_\lambda$ (constant)
- $(\lambda = 0^+)$: $A_t = \mathsf{C}_0 = \frac{1}{T}\log(R)$ (unique, minimum norm)

$$\lambda \mathsf{C} = e^{T\mathsf{C}}(R - e^{T\mathsf{C}})\mathsf{E}[X_0^2]$$

Proof sketch: $X_T = e^{\int_0^T A_t \, \mathrm{d}t} X_0$ and $\int_0^T A_t^2 \, \mathrm{d}t \geq \frac{1}{T}\left(\int_0^T A_t \, \mathrm{d}t\right)^2$

Scalar case ($d = 1$)

$$\text{Model:} \quad Z = RX_0 + \xi$$

$$\underset{A}{\text{Minimize:}} \quad \mathsf{J}[A] = \frac{1}{2}\mathsf{E}\left[|X_T - Z|^2\right] + \frac{\lambda}{2}\int_0^T A_t^2 \, \mathrm{d}t$$

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

Proposition

Assume $R > 0$. Then the critical points are

- $(\lambda = 0)$: $A_t = \frac{1}{T}\log(R) + B_t$ s.t $\int_0^T B_t \, \mathrm{d}t = 0$
- $(\lambda > 0)$: $A_t = \mathsf{C}_\lambda$ (constant)
- $(\lambda = 0^+)$: $A_t = \mathsf{C}_0 = \frac{1}{T}\log(R)$ (unique, minimum norm)

$$\lambda \mathsf{C} = e^{T\mathsf{C}}(R - e^{T\mathsf{C}})\mathsf{E}[X_0^2]$$

Question: How to generalize to the vector case? Is the minimizer always a constant?

Outline

1. Motivation
2. Problem formulation
3. Scalar case
4. Optimal control formulation
5. Main result: Characterization of critical points.

Outline

Optimal control formulation

Optimal control problem:

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

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

Two appoaches:

- Dynamic programming: HJB equation
- Maximum principle: Hamilton's equations

Usual assumption: The control variable A_t is measurable w.r.t X_t

But: In this (NN) setting, A_t can not be a function of $X_t(\omega)$.

Optimal control formulation

Optimal control problem:

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

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

Two appoaches:

- Dynamic programming: HJB equation
- Maximum principle: Hamilton's equations

Usual assumption: The control variable A_t is measurable w.r.t X_t

But: In this (NN) setting, A_t can not be a function of $X_t(\omega)$.

Optimal control formulation

Optimal control problem:

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

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

Two appoaches:

- Dynamic programming: HJB equation
- Maximum principle: Hamilton's equations

Usual assumption: The control variable A_t is measurable w.r.t X_t

But: In this (NN) setting, A_t can not be a function of $X_t(\omega)$.

Optimal control formulation

Optimal control problem:

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

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

Two appoaches:

- Dynamic programming: HJB equation
- Maximum principle: Hamilton's equations

Usual assumption: The control variable A_t is measurable w.r.t X_t

But: In this (NN) setting, A_t can not be a function of $X_t(\omega)$.

Maximum principle and Hamilton's equations

Hamiltonian function:

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

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 and X_t is the corresponding trajectory. Then there exists a random process $Y : [0, T] \to \mathbb{R}^d$ such that A_t maximizes the expected value of the Hamiltonian

$$A_t = \underset{B \in M_d(\mathbb{R})}{\arg\max} \quad \mathsf{E}[\mathsf{H}(X_t, Y_t, B)]$$

and X_t, Y_t solve the Hamilton's equations

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

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

And the converse is true.

Maximum principle and Hamilton's equations

Hamiltonian function:

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

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 and X_t is the corresponding trajectory. Then there exists a random process $Y : [0, T] \to \mathbb{R}^d$ such that A_t maximizes the expected value of the Hamiltonian

$$A_t = \underset{B \in M_d(\mathbb{R})}{\arg\max} \quad \mathsf{E}[\mathsf{H}(X_t, Y_t, B)]$$

and X_t, Y_t solve the Hamilton's equations

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

And the converse is true.

Relation to backprop algorithm

First order variation:

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

Stochastic gradient-descent: Given $A_t^{(K)}$ and $(X^{(k)}, Z^{(k)})$

$$A_t^{(k+1)} = A_t^{(k)} + \eta_k \underbrace{\frac{\partial \mathsf{H}}{\partial B}(X_t^{(k)}, Y_t^{(k)}, A_t^{(k)})}_{Y_t^{(k)} X_t^{(k)\top} - \lambda A_t^{(k)}}$$

where η_k is the step-size and

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

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

[LeCun, et. al. (1988)]

Relation to backprop algorithm

First order variation:

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

Stochastic gradient-descent: Given $A_t^{(K)}$ and $(X^{(k)}, Z^{(k)})$

$$A_t^{(k+1)} = A_t^{(k)} + \eta_k \underbrace{\frac{\partial \mathsf{H}}{\partial B}(X_t^{(k)}, Y_t^{(k)}, A_t^{(k)})}_{Y_t^{(k)} X_t^{(k)\top} - \lambda A_t^{(k)}}$$

where η_k is the step-size and

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

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

[LeCun, et. al. (1988)]

Outline

1. Motivation
2. Problem formulation
3. Scalar case
4. Optimal control formulation
5. Main result: Characterization of critical points.

Outline

Vecotr case, No regularization ($\lambda = 0$)

$$\text{Model:} \quad Z = RX_0 + \xi$$

$$\underset{A}{\text{Minimize:}} \quad \mathsf{J}[A] = \frac{1}{2}\mathsf{E}\left[|X_T - Z|^2\right]$$

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

Assumption: $\log(R) \in M_d(\mathbb{R})$ exists and $\Sigma := \mathsf{E}[X_0 X_0^\top]$ is invertible

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

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

- The optimality gap is upper-bounded by the gradient

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

Vecotr case, No regularization ($\lambda = 0$)

$$\text{Model:} \quad Z = RX_0 + \xi$$

$$\underset{A}{\text{Minimize:}} \quad \mathsf{J}[A] = \frac{1}{2}\mathsf{E}\left[|X_T - Z|^2\right]$$

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

Assumption: $\log(R) \in M_d(\mathbb{R})$ exists and $\Sigma := \mathsf{E}[X_0 X_0^\top]$ is invertible

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

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

- The optimality gap is upper-bounded by the gradient

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

Vector case, with regularization ($\lambda > 0$)

$$\text{Model:} \quad Z = RX_0 + \xi$$

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

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

Proposition (main result)

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

$$\lambda \mathsf{C} = e^{T\mathsf{C}} e^{T(\mathsf{C}^\top - \mathsf{C})} (R - e^{T(\mathsf{C} - \mathsf{C}^\top)} e^{T\mathsf{C}^\top}) \Sigma$$

And the weights are

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

If R is not normal ($R^\top R \neq R R^\top$) $\Rightarrow C$ is not normal $\Rightarrow A_t$ is not constant

Vector case, with regularization ($\lambda > 0$)

$$\text{Model:} \quad Z = RX_0 + \xi$$

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

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

Proposition (main result)

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

$$\lambda \mathsf{C} = e^{T\mathsf{C}} e^{T(\mathsf{C}^\top - \mathsf{C})} (R - e^{T(\mathsf{C} - \mathsf{C}^\top)} e^{T\mathsf{C}^\top}) \Sigma$$

And the weights are

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

If R is not normal ($R^\top R \neq R R^\top$) $\Rightarrow C$ is not normal $\Rightarrow A_t$ is not constant

Vector case, with regularization ($\lambda > 0$)

$$\text{Model:} \quad Z = RX_0 + \xi$$

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

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

Proposition (main result)

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

$$\lambda \mathsf{C} = e^{T\mathsf{C}} e^{T(\mathsf{C}^\top - \mathsf{C})} (R - e^{T(\mathsf{C} - \mathsf{C}^\top)} e^{T\mathsf{C}^\top}) \Sigma$$

And the weights are

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

If R is not normal ($R^\top R \neq R R^\top$) $\Rightarrow$ C is not normal $\Rightarrow$ A_t is not constant

Analysis of the characteristic equation

Characteristic equation:

$$\lambda \mathsf{C} = e^{T\mathsf{C}} e^{T(\mathsf{C}^\top - \mathsf{C})} (R - e^{T(\mathsf{C} - \mathsf{C}^\top)} e^{T\mathsf{C}^\top}) \Sigma$$

Assumption: $\Sigma = I$ and R is normal ($RR^\top = R^\top R$)

Proposition

1. Set $\lambda = 0$. The normal solutions are

$$\mathsf{C}(0) = \frac{1}{T} \log(R)$$

2. For each solution, $\exists$ neighborhood of $\lambda = 0$ s.t the solution continue to exist

$$\mathsf{C}(\lambda) = \frac{1}{T} \log(R) - \frac{\lambda}{T^2} (RR^\top)^{-1} \log(R) + O(\lambda^2)$$

(implicit function thm)

Remark: There are non normal solutions too!

Analysis of the characteristic equation

Characteristic equation:

$$\lambda \mathsf{C} = e^{T\mathsf{C}} e^{T(\mathsf{C}^\top - \mathsf{C})} (R - e^{T(\mathsf{C} - \mathsf{C}^\top)} e^{T\mathsf{C}^\top}) \Sigma$$

Assumption: $\Sigma = I$ and R is normal ($RR^\top = R^\top R$)

Proposition

1. Set $\lambda = 0$. The normal solutions are

$$\mathsf{C}(0) = \frac{1}{T} \log(R)$$

2. For each solution, $\exists$ neighborhood of $\lambda = 0$ s.t the solution continue to exist

$$\mathsf{C}(\lambda) = \frac{1}{T} \log(R) - \frac{\lambda}{T^2} (RR^\top)^{-1} \log(R) + O(\lambda^2)$$

(implicit function thm)

Remark: There are non normal solutions too!

Analysis of the characteristic equation

Characteristic equation:

$$\lambda \mathsf{C} = e^{T\mathsf{C}} e^{T(\mathsf{C}^\top - \mathsf{C})} (R - e^{T(\mathsf{C} - \mathsf{C}^\top)} e^{T\mathsf{C}^\top}) \Sigma$$

Assumption: $\Sigma = I$ and R is normal ($RR^\top = R^\top R$)

Proposition

1. Set $\lambda = 0$. The normal solutions are

$$\mathsf{C}(0) = \frac{1}{T} \log(R)$$

2. For each solution, $\exists$ neighborhood of $\lambda = 0$ s.t the solution continue to exist

$$\mathsf{C}(\lambda) = \frac{1}{T} \log(R) - \frac{\lambda}{T^2} (RR^\top)^{-1} \log(R) + O(\lambda^2)$$

(implicit function thm)

Remark: There are non normal solutions too!

Numerical Examples

Example I: Illustrating solutions to the characteristic equation

$$R = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad \begin{cases} T & = 1 \\ \Sigma & = I \end{cases}$$

For $\lambda = 0$, infinite number of solutions which are all global minimizers exist as:

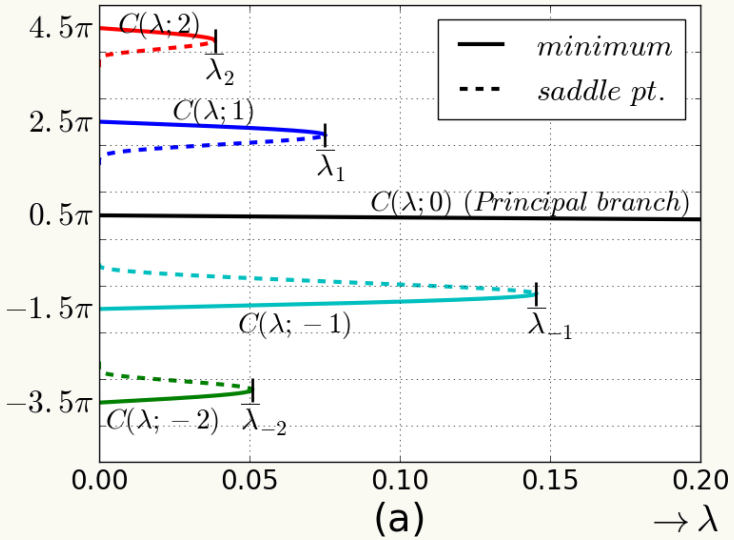
$$\log(R) = (\pi/2 + 2n\pi) \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} =: \mathsf{C}(0; n), \quad n = 0, \pm 1, \pm 2, \dots$$

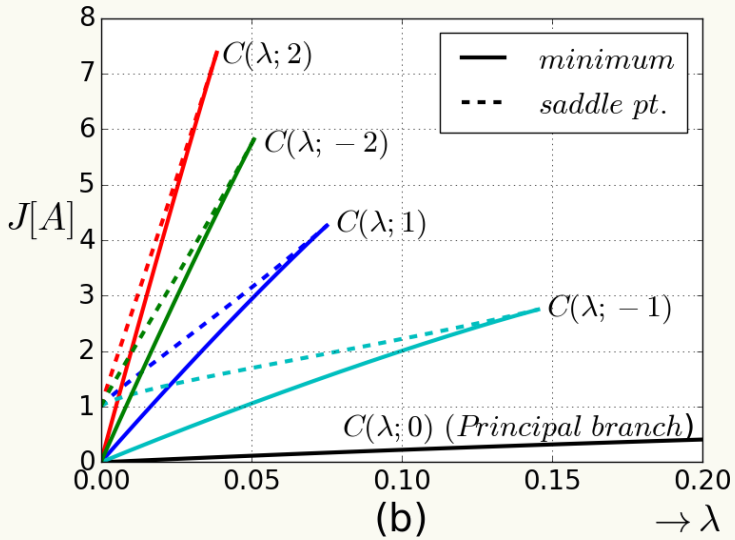
and infinite number of non-normal solutions also exist.

For $\lambda > 0$, the solution can be numerically obtained by continuing from each $\lambda = 0$ solutions.

Numerical Examples

Example I: Illustrating solutions to the characteristic equation



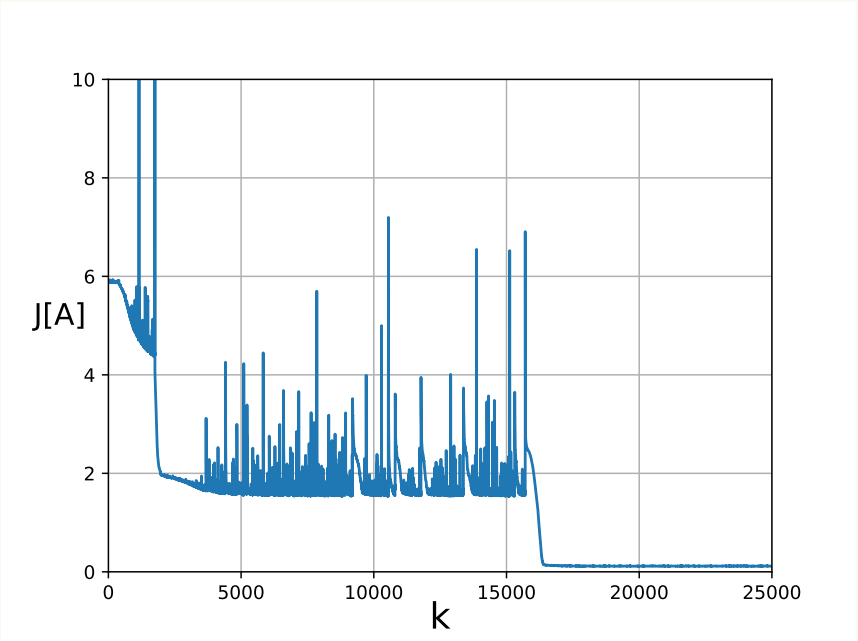


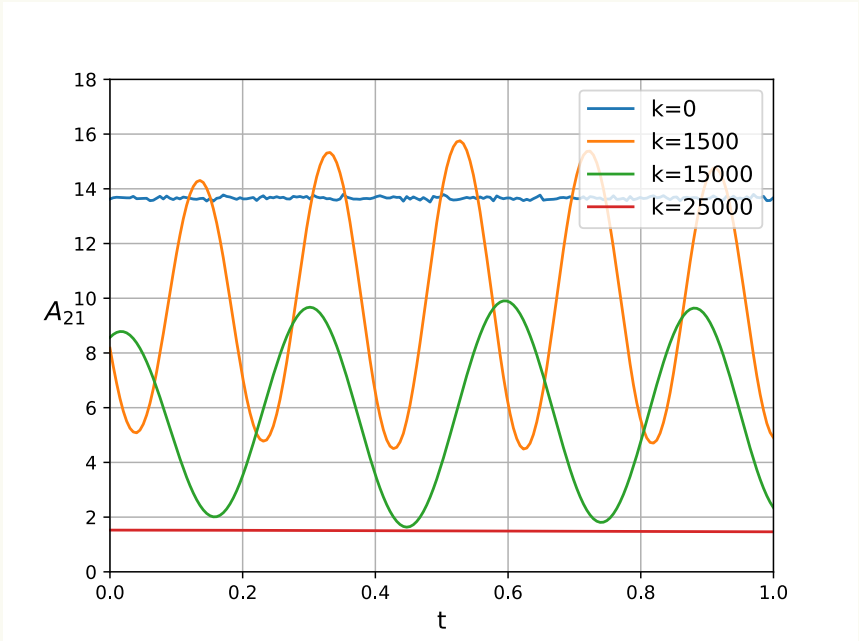
- While every stable solutions are global minimizer for $\lambda = 0$, unique global minimizer which corresponds to the principal branch arises when $\lambda > 0$.
- Local minimizers are eliminated when λ is sufficiently large

Numerical Examples

Example II: Learning

- The system initialized by A corresponding to the $\mathrm{C}(0.03, 2)$, trained with $\eta = 0.05$





- The learning method visit local minimums, but eventually converged to the global minimum corresponding to the principal branch.

Final slide

Conclusions and remarks

- (Main result) Reduction of the infinite-dimensional optimization problem to the finite-dimensional characteristic equation

$$\lambda \mathsf{C} = F^\top (R - F)\Sigma$$

- (future work) Generalization: given samples $(X_0^i, Z^i)_{i=1}^N$

$$\lambda \mathsf{C} = F^\top (R - F)\Sigma^{(N)} + F^\top Q^{(N)}$$

- (future work) Second order analysis
- (Far in the future work) Complete characterization of solutions to the characteristic eq.
- (Far in the future work) Extension to nonlinear setting

Thank you for your attention!

Questions?

Final slide

Conclusions and remarks

- (Main result) Reduction of the infinite-dimensional optimization problem to the finite-dimensional characteristic equation

$$\lambda \mathsf{C} = F^\top (R - F)\Sigma$$

- (future work) Generalization: given samples $(X_0^i, Z^i)_{i=1}^N$

$$\lambda \mathsf{C} = F^\top (R - F)\Sigma^{(N)} + F^\top Q^{(N)}$$

$$\Sigma^{(N)} := \frac{1}{N}\sum_{i=1}^{N} X_0^i {X_0^i}^\top$$

$$Q = \frac{1}{N}\sum_{i=1}^{N} \xi^i {X_0^i}^\top$$

- (future work) Second order analysis
- (Far in the future work) Complete characterization of solutions to the characteristic eq.
- (Far in the future work) Extension to nonlinear setting

Thank you for your attention!

Questions?

Final slide

Conclusions and remarks

- (Main result) Reduction of the infinite-dimensional optimization problem to the finite-dimensional characteristic equation

$$\lambda \mathsf{C} = F^\top (R - F)(\Sigma + \epsilon_1) + \epsilon_2$$

- (future work) Generalization: given samples $(X_0^i, Z^i)_{i=1}^N$

$$\lambda \mathsf{C} = F^\top (R - F)\Sigma^{(N)} + F^\top Q^{(N)}$$

$$\Sigma^{(N)} := \frac{1}{N} \sum_{i=1}^{N} X_0^i {X_0^i}^\top$$

$$Q = \frac{1}{N} \sum_{i=1}^{N} \xi^i {X_0^i}^\top$$

Generalization is related to sensitivity w.r.t ϵ_1, ϵ_2

- (future work) Second order analysis
- (Far in the future work) Complete characterization of solutions to the characteristic eq.
- (Far in the future work) Extension to nonlinear setting

Thank you for your attention!

Questions?

Final slide

Conclusions and remarks

- (Main result) Reduction of the infinite-dimensional optimization problem to the finite-dimensional characteristic equation

$$\lambda \mathsf{C} = F^\top (R - F)\Sigma$$

- (future work) Generalization: given samples $(X_0^i, Z^i)_{i=1}^N$

$$\lambda \mathsf{C} = F^\top (R - F)\Sigma^{(N)} + F^\top Q^{(N)}$$

- (future work) Second order analysis
- (Far in the future work) Complete characterization of solutions to the characteristic eq.
- (Far in the future work) Extension to nonlinear setting

Thank you for your attention!

Questions?

Final slide

Conclusions and remarks

- (Main result) Reduction of the infinite-dimensional optimization problem to the finite-dimensional characteristic equation

$$\lambda \mathsf{C} = F^\top (R - F)\Sigma$$

- (future work) Generalization: given samples $(X_0^i, Z^i)_{i=1}^N$

$$\lambda \mathsf{C} = F^\top (R - F)\Sigma^{(N)} + F^\top Q^{(N)}$$

- (future work) Second order analysis
- (Far in the future work) Complete characterization of solutions to the characteristic eq.
- (Far in the future work) Extension to nonlinear setting

Thank you for your attention!

Questions?

Final slide

Conclusions and remarks

- (Main result) Reduction of the infinite-dimensional optimization problem to the finite-dimensional characteristic equation

$$\lambda \mathsf{C} = F^\top (R - F)\Sigma$$

- (future work) Generalization: given samples $(X_0^i, Z^i)_{i=1}^N$

$$\lambda \mathsf{C} = F^\top (R - F)\Sigma^{(N)} + F^\top Q^{(N)}$$

- (future work) Second order analysis
- (Far in the future work) Complete characterization of solutions to the characteristic eq.
- (Far in the future work) Extension to nonlinear setting

Thank you for your attention!

Questions?

Final slide

Conclusions and remarks

- (Main result) Reduction of the infinite-dimensional optimization problem to the finite-dimensional characteristic equation

$$\lambda \mathsf{C} = F^\top (R - F)\Sigma$$

- (future work) Generalization: given samples $(X_0^i, Z^i)_{i=1}^N$

$$\lambda \mathsf{C} = F^\top (R - F)\Sigma^{(N)} + F^\top Q^{(N)}$$

- (future work) Second order analysis
- (Far in the future work) Complete characterization of solutions to the characteristic eq.
- (Far in the future work) Extension to nonlinear setting

Thank you for your attention!

Questions?

First order variation

$$\mathsf{J}[A] := \frac{1}{2}\mathsf{E}\left[|X_T - Z|^2\right] + \frac{\lambda}{2}\int_0^T \mathrm{tr}\left(A_t^\top A_t\right) \mathrm{d}t$$

$$A \in \mathcal{H} := L^2([0,T]; M_d(\mathbb{R}))$$

$$\mathsf{J} : \mathcal{H} \to \mathbb{R}$$

Definition: $\nabla \mathsf{J}[A] \in \mathcal{H}$ s.t

$$\langle \nabla \mathsf{J}[A], V \rangle_{L^2} = \lim_{\epsilon \to 0} \frac{\mathsf{J}(A + \epsilon V) - \mathsf{J}}{\epsilon}, \quad \forall V \in \mathcal{H}$$

Formula in terms of Hamiltonian:

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

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