Nonconvex Optimization and Model Representation with Applications in Control Theory and Machine Learning

Yue Sun

Department of Electrical and Computer Engineering
University of Washington, Seattle

November 16, 2020

Introduction - Machine learning

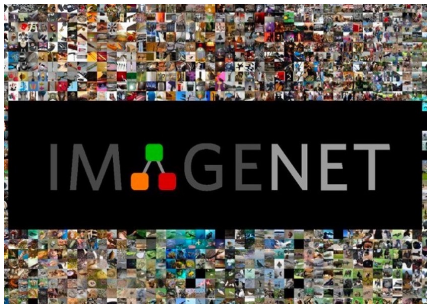
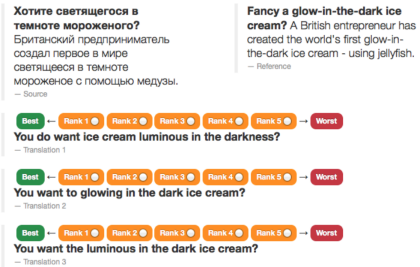
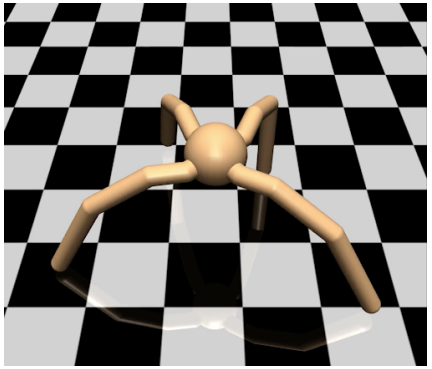


Image classification



Natural language



Control & Robotics

Game

Introduction - Machine learning

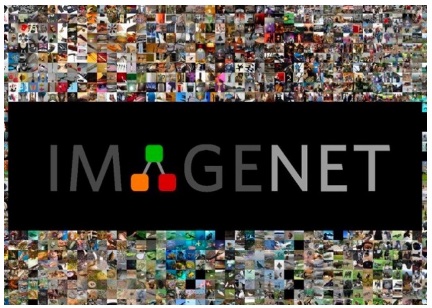
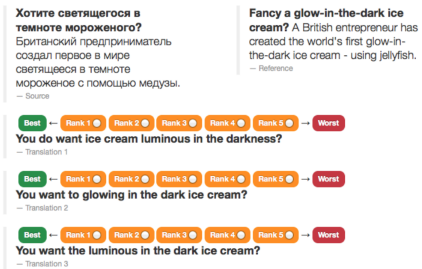
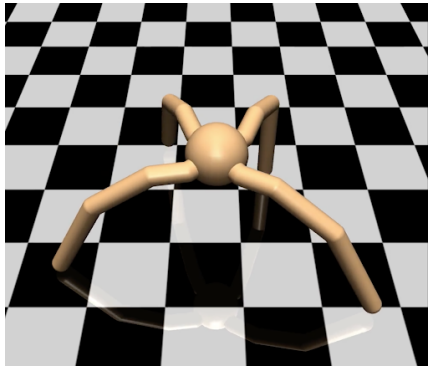


Image classification



Natural language



Control & Robotics

Game

What does ML do?

Suppose the input object is $\boldsymbol{u}$, the output/label is y.

The target is to find the function $f(\boldsymbol{u};\theta)$, parameterized by θ, that outputs y.

$$f(\boldsymbol{u};\theta) \approx y$$

Challenges

Find the function $f(\boldsymbol{u}; \theta)$ that outputs y.

$$f(\boldsymbol{u}; \theta) \approx y$$

Two challenges for efficiently learning from data.

Challenges

Find the function $f(\boldsymbol{u}; \theta)$ that outputs y.

$$f(\boldsymbol{u}; \theta) \approx y$$

Two challenges for efficiently learning from data.

1. Statistical: What model structure estimates data with good statistical guarantee (training and generalization error, sample complexity) – **Structured model**.

Challenges

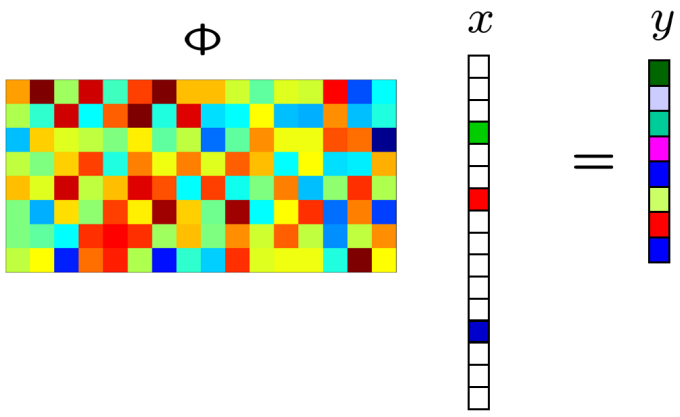
Find the function $f(\boldsymbol{u}; \theta)$ that outputs y.

$$f(\boldsymbol{u}; \theta) \approx y$$

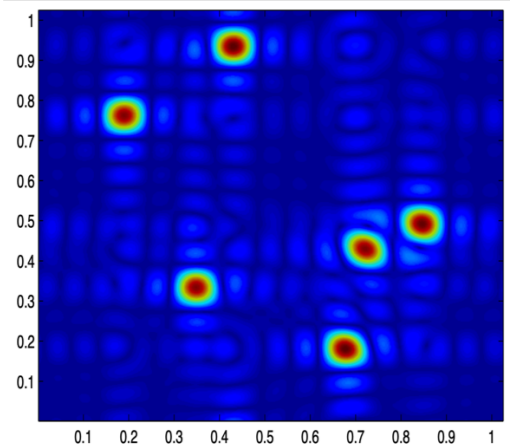
Two challenges for efficiently learning from data.

1. Statistical: What model structure estimates data with good statistical guarantee (training and generalization error, sample complexity) – **Structured model**.
2. Computational: How to find the best parameter of model quickly – **Optimization**.

Representation for machine learning

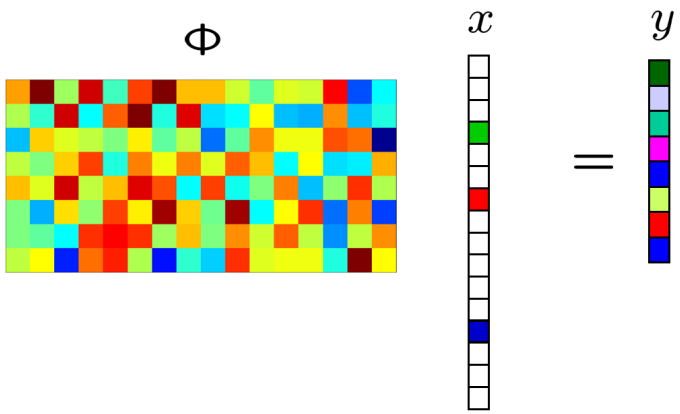


Sparse signal

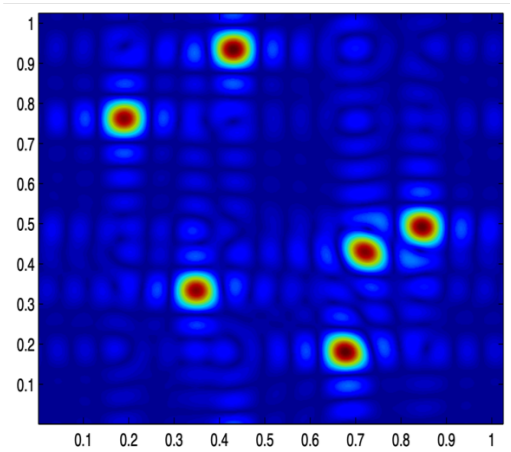


Super-resolution

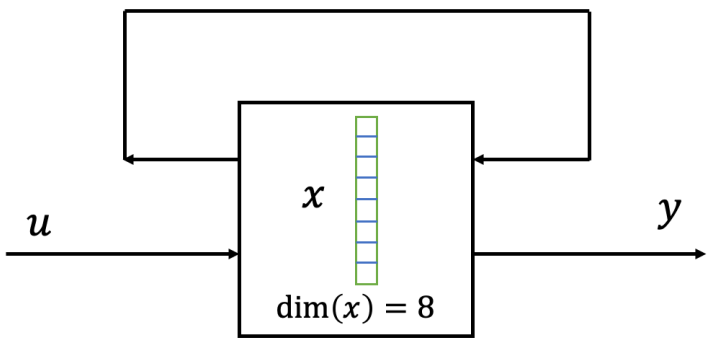
Representation for machine learning



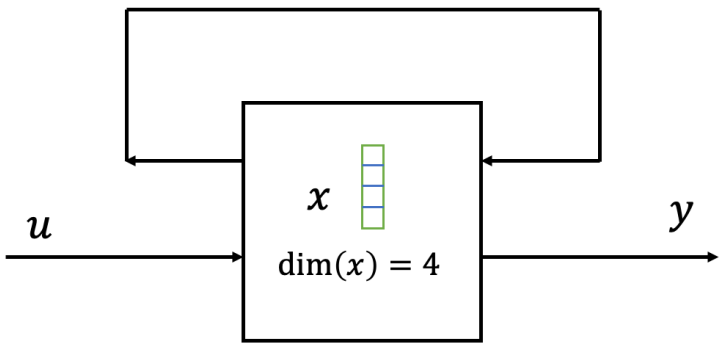
Sparse signal



Super-resolution



High order linear system



Low order linear system

Representation for machine learning

Find the function $f(\boldsymbol{u}; \theta)$ that outputs y.

$$f(\boldsymbol{u}; \theta) \approx y$$

Representation for machine learning

Find the function $f(\boldsymbol{u};\theta)$ that outputs y.

$$f(\boldsymbol{u};\theta) \approx y$$

If we have prior knowledge of θ (i.e., sparsity, low dimensionality, low order system), how to use it to make learning efficient?

1. Learning low order linear dynamical systems via nuclear norm regularization
2. (ongoing) Understanding the role of representation dimension in meta-learning

Computation for machine learning

Find the function $f(\boldsymbol{u}; \theta)$ that outputs y.

$$f(\boldsymbol{u}; \theta) \approx y$$

Computation for machine learning

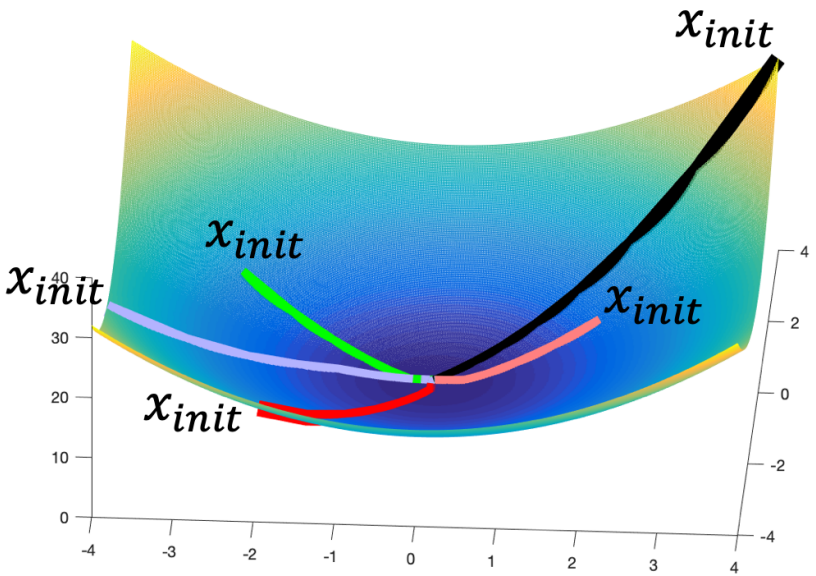
Find the function $f(\boldsymbol{u}; \theta)$ that outputs y.

$$f(\boldsymbol{u}; \theta) \approx y$$

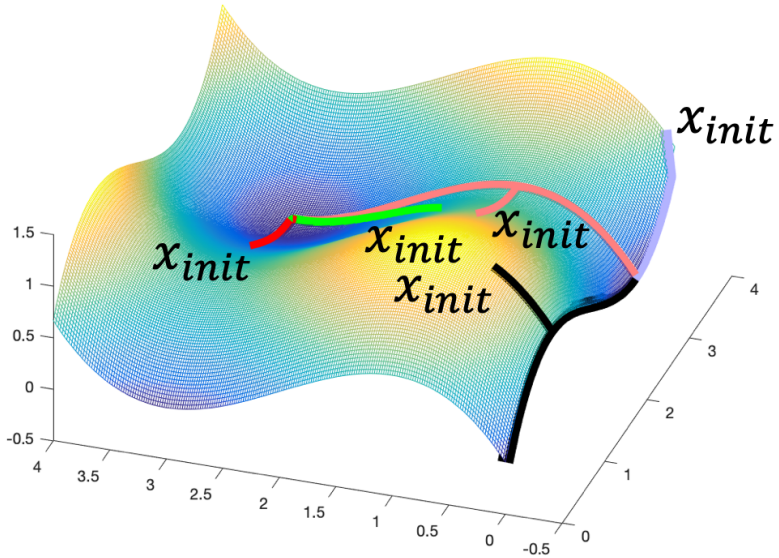
To find the best model, we solve the following optimization problem.

$$\min_{\theta} \; \mathit{distance}(f(\boldsymbol{u}; \theta), y)$$

Computation for machine learning

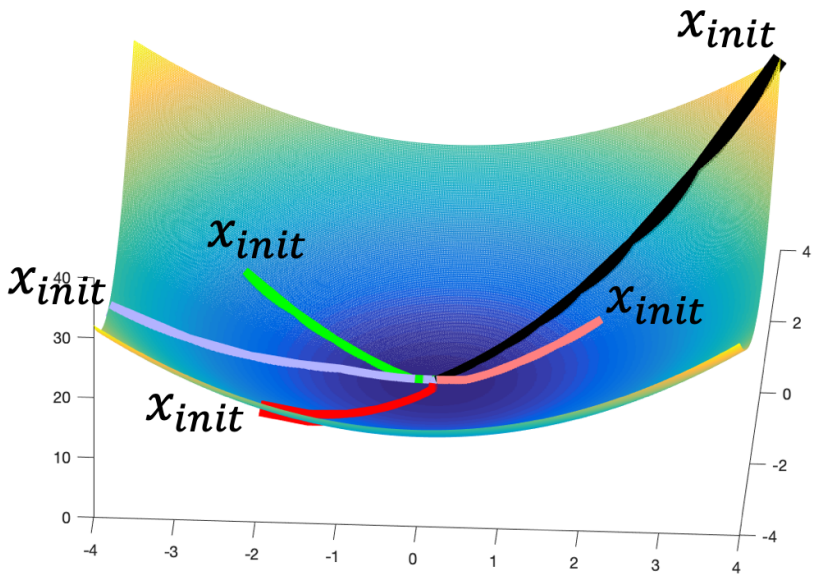


Trajectory of gradient descent on convex function

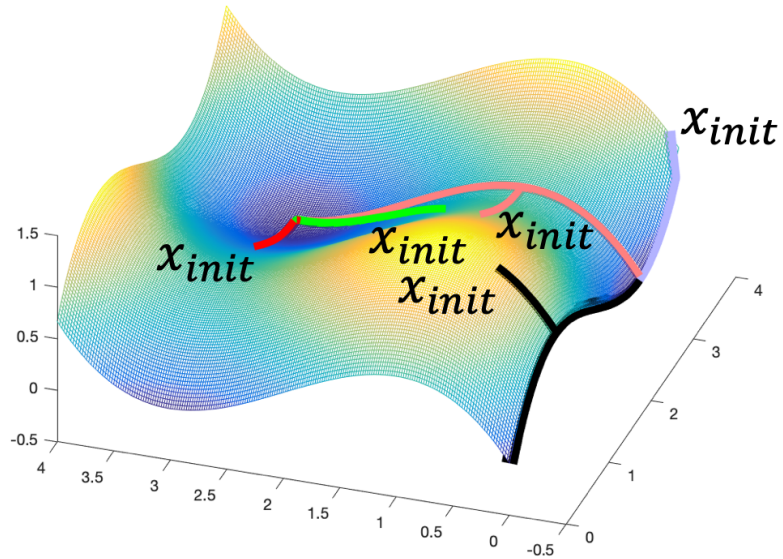


Trajectory of gradient descent on nonconvex function

Computation for machine learning



Trajectory of gradient descent on convex function



Trajectory of gradient descent on nonconvex function

Gradient descent may not converge to global minimum of nonconvex objective – What can we say about it?

1. Escaping from saddle points on Riemannian manifolds
2. Novel analysis of convergence of policy gradient descent via convexification

Overview

Overview

Order-R SISO[1] system

Let u, x, y, ξ be the system input, state, output, noise. There exists state-space model of order R, $A \in \mathbb{R}^{R\times R}$, $b \in \mathbb{R}^{R\times 1}$, $c \in \mathbb{R}^{1\times R}$

$$\begin{aligned} x(t+1) &= Ax(t) + bu(t),\ x(0) = 0, \\ y(t) &= cx(t) + \xi(t) \end{aligned}$$

Impulse response $h_i = cA^{i-1}b,\ i = 1, \ldots$.

$$R = \operatorname{rank} \begin{pmatrix} h_1 & h_2 & \cdots & h_n \\ h_2 & h_3 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ h_n & \cdots & \cdots & h_{2n-1} \end{pmatrix} = \mathcal{H}(h),\ \forall n \geq R$$

Hankel rank upper bounded by system order (Fazel, Hindi, and Boyd 2003)

[1] single-input single-output

Order-R SISO[1] system

Let u, x, y, ξ be the system input, state, output, noise. There exists state-space model of order R, $A \in \mathbb{R}^{R\times R}$, $b \in \mathbb{R}^{R\times 1}$, $c \in \mathbb{R}^{1\times R}$

$$\begin{aligned} x(t+1) &= Ax(t) + bu(t),\ x(0) = 0, \\ y(t) &= cx(t) + \xi(t) \end{aligned}$$

Impulse response $h_i = cA^{i-1}b,\ i = 1, \ldots$.

$$R = \operatorname{rank} \begin{pmatrix} h_1 & h_2 & \cdots & h_n \\ h_2 & h_3 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ h_n & \cdots & \cdots & h_{2n-1} \end{pmatrix} = \mathcal{H}(h),\ \forall n \geq R$$

Hankel rank upper bounded by system order (Fazel, Hindi, and Boyd 2003)

Goal: Recover h from u and y with unknown R.

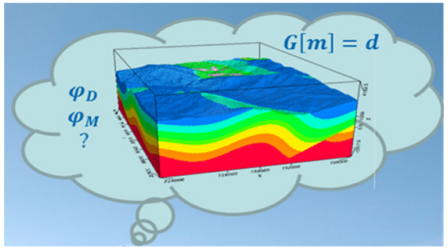
[1] single-input single-output

Low-rank Hankel matrices and applications

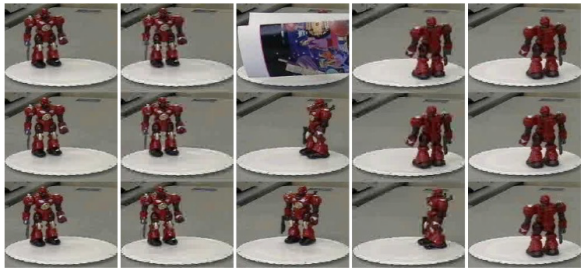
input-output LTI system ID

https://www.mathworks.com/products/sysid.html

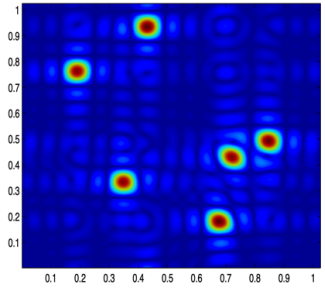
shape from moments estimation, tomography, geophysical inversion [Elad et al '04]



video inpainting [Ding et al, '07]



Super-resolution [Chen & Chi '14]



The role of Hankel nuclear norm regularization

Input-output system identification

- Classical: matrix pencil method; subspace system ID
  Ho and Kálmán 1966; Van Overschee and De Moor 1995,. . .

- More recent: regularization with Hankel nuclear norm
  Fazel, Pong, et al. 2013; Hansson, Liu, and Vandenberghe 2012; Liu, Hansson, and Vandenberghe 2013; Verhaegen and Hansson 2016; Blomberg 2016

The papers above discusses algorithm implementation: no statistical guarantee.

Goal: statistical bounds: sample complexity, error rates

Related work

Goal: statistical bounds: sample complexity, error rates.
Only recently explored for (unregularized) least squares

state observation:
Simchowitz, Mania, et al. 2018: Learning marginally stable system via LS.
Faradonbeh, Tewari, and Michailidis 2018: Unstable system, either stable or pure explosive.
Sarkar and Rakhlin 2019: Unstable system.
Dean et al. 2017: System id via LS, followed by robust control.
Dean et al. 2018: Regret analysis of online system id and robust control.
Mania, Tu, and Recht 2019: System id and controlling by nominal controller.
Sattar and Oymak 2020; Foster, Rakhlin, and Sarkar 2020: Nonlinear strictly stable system id.

Related work

Goal: statistical bounds: sample complexity, error rates.
Only recently explored for (unregularized) least squares

input-output system:
Tu et al. 2017: Learning FIR system via LS on impulse response. Designed input, $\mathcal{H}_\infty$ error analysis.
Oymak and Ozay 2018: Learning strictly stable system via LS on impulse response.
Sarkar, Rakhlin, and Dahleh 2019: Learning strictly stable system via LS on Hankel matrix.
Hazan, Singh, and C. Zhang 2017; Hazan, H. Lee, et al. 2018: Learning to predict stable system with LS and filtering.
Simchowitz, Singh, and Hazan 2020: Stable system id with LS and filtering.
Tsiamis and Pappas 2019: Kalman filtered system id.

Related work

Goal: statistical bounds: sample complexity, error rates.
Only recently explored for (unregularized) least squares:
state observation: Simchowitz, Mania, et al. 2018; Faradonbeh, Tewari, and Michailidis 2018; Sarkar and Rakhlin 2019; Dean et al. 2017, 2018; Mania, Tu, and Recht 2019; Sattar and Oymak 2020; Foster, Rakhlin, and Sarkar 2020
stable/unstable, random/designed input, sys id/control, linear/nonlinear
input-output system: Oymak and Ozay 2018; Sarkar, Rakhlin, and Dahleh 2019; Tu et al. 2017; Simchowitz, Boczar, and Recht 2019; Hazan, Singh, and C. Zhang 2017; Hazan, H. Lee, et al. 2018; Simchowitz, Singh, and Hazan 2020; Tsiamis and Pappas 2019
strictly/marginally stable, random/designed input, prediction/sysid/control, filtering

Cai et al. 2016 regularization for super-resolution, with partial results

This section: towards guarantees for regularization

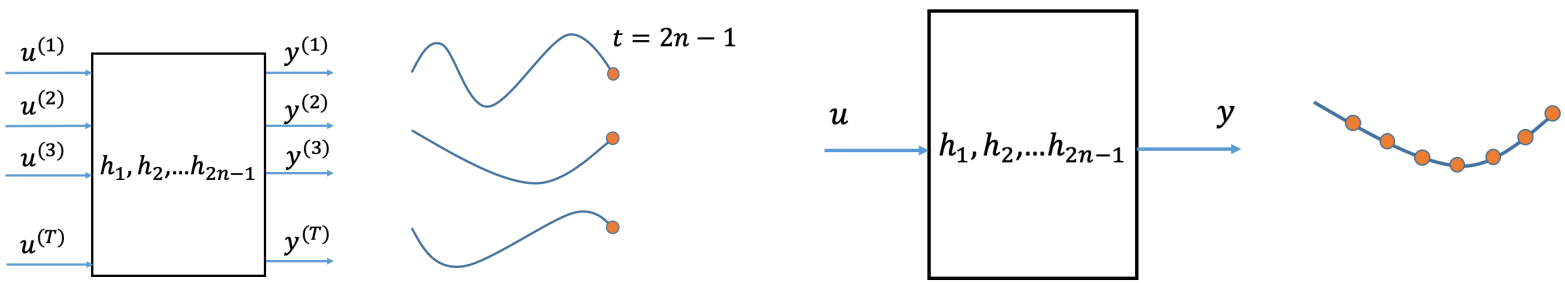
Data acquisition

The input-output mapping can also be

$$y = u * h + \xi$$

Denote **snr** $= \frac{\boldsymbol{E}(\|u\|^2)}{2n-1} \ / \ \boldsymbol{E}(\xi^2)$

Two data aqcuisition models: (a) Multi-rollout (left), and (b) single rollout (right). Sample complexity T $=$ # of dots



Nuclear norm regularization for low order system recovery

Let y, U, ξ be stacked output, input, and noise:

$$y \approx Uh$$

For multiple rollout:

$$U = \begin{bmatrix} u_1^{(1)} & u_2^{(1)} & \dots & u_{2n-1}^{(1)} \\ u_1^{(2)} & u_2^{(2)} & \dots & u_{2n-1}^{(2)} \\ \dots & \dots & \dots & \dots \\ u_1^{(T)} & u_2^{(T)} & \dots & u_{2n-1}^{(T)} \end{bmatrix}, \; h = \begin{bmatrix} h_{2n-1} \\ h_{2n-2} \\ \dots \\ h_1 \end{bmatrix}, \; y = \begin{bmatrix} y_{2n-1}^{(1)} \\ y_{2n-1}^{(2)} \\ \dots \\ y_{2n-1}^{(T)} \end{bmatrix}$$

For single rollout:

$$U = \begin{bmatrix} u_1 & u_2 & \dots & u_{2n-1} \\ u_2 & u_3 & \dots & u_{2n} \\ \dots & \dots & \dots & \dots \\ u_{T-2n+2} & u_{T-2n+3} & \dots & u_T \end{bmatrix}, \; h = \begin{bmatrix} h_{2n-1} \\ h_{2n-2} \\ \dots \\ h_1 \end{bmatrix}, \; y = \begin{bmatrix} y_{2n-1} \\ y_{2n} \\ \dots \\ y_T \end{bmatrix}$$

Nuclear norm regularization for low order system recovery

Consider the optimization problem

$$\min_{h' \in \mathbb{R}^{2n-1}} \underbrace{\|y - Uh'\|_2^2}_{\text{squared loss}} + \lambda \underbrace{\|\mathcal{H}(h')\|_*}_{\text{regularization}} \tag{1}$$

- $\mathcal{H}(h)$ denotes the $n \times n$ Hankel matrix (dimension of variable)

$$\mathcal{H}(h) := \begin{bmatrix} h_1 & h_2 & \dots & h_{n-1} & h_n \\ h_2 & h_3 & \dots & h_n & h_{n+1} \\ \dots & \dots & \dots & \dots & \dots \\ h_n & h_{n+1} & \dots & h_{2n-2} & h_{2n-1} \end{bmatrix}$$

- $\lambda = 0$, least squares; $\lambda > 0$, regularization.
- Regularizer encourages low rank Hankel structure: $R < n$.

Which error metrics?

$$\hat{h} = \operatorname{argmin}_{h'} \underbrace{\|y - Uh'\|_2^2}_{\text{squared loss}} + \lambda \underbrace{\|\mathcal{H}(h')\|_*}_{\text{regularization}}$$

How useful is $\hat{h}$?

Which error metrics?

$$\hat{h} = \operatorname{argmin}_{h'} \underbrace{\|y - Uh'\|_2^2}_{\text{squared loss}} + \lambda \underbrace{\|\mathcal{H}(h')\|_*}_{\text{regularization}}$$

How useful is $\hat{h}$?

- ℓ_2 error: $\|h - \hat{h}\|_2$
- Hankel spectral error: $\|\mathcal{H}(h) - \mathcal{H}(\hat{h})\|$
  - Useful for model selection & finding realization
  - Hankel spectral error upper bounds ℓ_2 error, and we'll prove they are on the same order.

Spectral analysis - Nuclear norm

Theorem

Let the dimension of Hankel be $n \times n$, system order be R and number of samples be T.

$$\|\mathcal{H}(h) - \mathcal{H}(\hat{h})\| \lesssim \begin{cases} \sqrt{\frac{n}{\mathbf{snr} \times T}} & \text{if} \quad T \gtrsim \min(R^2, n) \\ \sqrt{\frac{Rn}{\mathbf{snr} \times T}} & \text{if} \quad R \lesssim T \lesssim \min(R^2, n) \end{cases}$$

(Second line extended from Cai et al. 2016)

Remark:

- Minimum sample complexity is $R < n$.

Spectral analysis - Nuclear norm

Theorem

Let the dimension of Hankel be $n \times n$, system order be R and number of samples be T.

$$\|\mathcal{H}(h) - \mathcal{H}(\hat{h})\| \lesssim \begin{cases} \sqrt{\frac{n}{\mathbf{snr} \times T}} & \text{if} \quad T \gtrsim \min(R^2, n) \\ \sqrt{\frac{Rn}{\mathbf{snr} \times T}} & \text{if} \quad R \lesssim T \lesssim \min(R^2, n) \end{cases}$$

(Second line extended from Cai et al. 2016)

Remark:

- Minimum sample complexity is $R < n$.
- Oymak and Ozay 2018 obtains ℓ_2 bound $\|h - \hat{h}\|_2 \lesssim \sqrt{\frac{n}{\mathbf{snr} \times T}}$ with $O(n)$ data in unregularized least squares algorithm.

Spectral analysis - Nuclear norm

Theorem

Let the dimension of Hankel be $n \times n$, system order be R and number of samples be T.

$$\|\mathcal{H}(h) - \mathcal{H}(\hat{h})\| \lesssim \begin{cases} \sqrt{\frac{n}{\mathbf{snr} \times T}} & \text{if} \quad T \gtrsim \min(R^2, n) \\ \sqrt{\frac{Rn}{\mathbf{snr} \times T}} & \text{if} \quad R \lesssim T \lesssim \min(R^2, n) \end{cases}$$

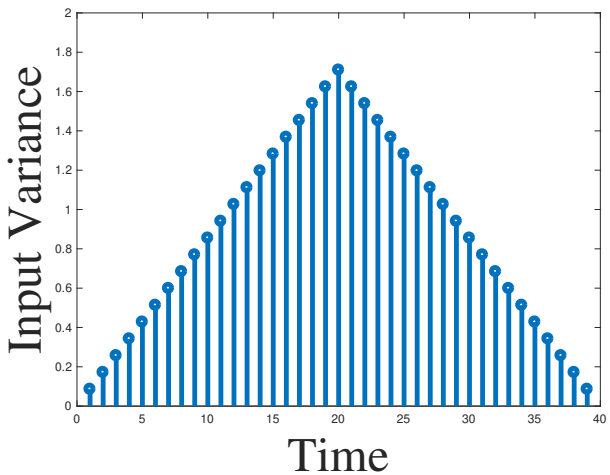
(Second line extended from Cai et al. 2016)

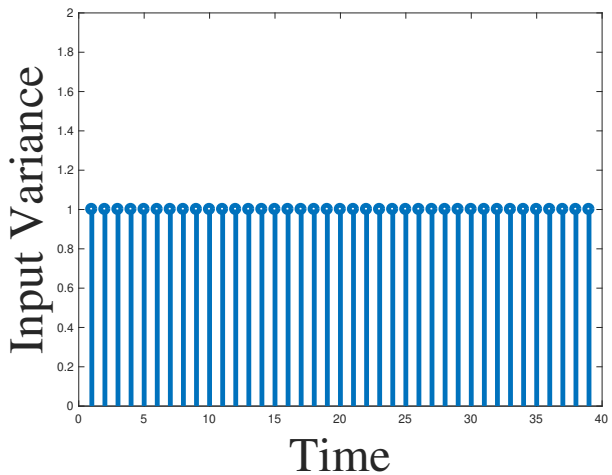
Remark:

- Minimum sample complexity is $R < n$.
- Oymak and Ozay 2018 obtains ℓ_2 bound $\|h - \hat{h}\|_2 \lesssim \sqrt{\frac{n}{\mathbf{snr} \times T}}$ with $O(n)$ data in unregularized least squares algorithm.
- Idea: Restricting the problem onto a set with small Gaussian width.

Optimal sample complexity due to shaped input

- Input with uniform variance is typically used for system id in least squares works. (Oymak and Ozay 2018; Sarkar, Rakhlin, and Dahleh 2019 etc.)
- With regularization, we have to use a shaped input for optimal sample complexity.





(a) Shaped input, recovery is guaranteed when $T \approx R$; (b) our result, i.i.d input, deterministic recovery failure $T \approx n^{1/6}$.

Optimal spectral bounds for unregularized least-squares

- Oymak and Ozay 2018 provides naive spectral error estimates.
- Sarkar, Rakhlin, and Dahleh 2019 provides suboptimal sample sizes.
- Can we get both?

Optimal spectral bounds for unregularized least-squares

- Oymak and Ozay 2018 provides naive spectral error estimates.
- Sarkar, Rakhlin, and Dahleh 2019 provides suboptimal sample sizes.
- Can we get both?

Theorem

Let the dimension of Hankel be $n \times n$ and number of samples be T. Let $\hat{h}_{LS}$ be the least squares estimate. Suppose $T \gtrsim n$, then The Hankel error obeys

$$\|\mathcal{H}(\hat{h}_{LS}) - \mathcal{H}(h)\| \lesssim \sqrt{\frac{n}{\mathbf{snr} \times T}}$$

Optimal spectral bounds for unregularized least-squares

- Oymak and Ozay 2018 provides naive spectral error estimates.
- Sarkar, Rakhlin, and Dahleh 2019 provides suboptimal sample sizes.
- Can we get both?

Theorem

Let the dimension of Hankel be $n \times n$ and number of samples be T. Let $\hat{h}_{LS}$ be the least squares estimate. Suppose $T \gtrsim n$, then The Hankel error obeys

$$\|\mathcal{H}(\hat{h}_{LS}) - \mathcal{H}(h)\| \lesssim \sqrt{\frac{n}{\textbf{snr} \times T}}$$

Takeaway: Spectral error is as good as the impulse response ℓ_2 error i.e.

$$\|\mathcal{H}(\hat{h} - h)\| \propto \|\hat{h} - h\|_2$$

Algorithmic comparison

Table 1: Hankel matrix is $n \times n$, system order is R, number of samples is T. Noise level $\sigma = 1/\sqrt{\mathbf{snr}}$. LS-IR stands for regressing impulse response with output by LS, LS-Hankel stands for regressing Hankel with output by LS.

Paper	This work	This work	[OO '18]	[SRD '19]
Sample complexity	R^2	n	n	n^2
Method	Nuc-norm	LS-IR	LS-IR	LS-Hankel
IR ℓ_2 Error	$\sigma\sqrt{n/T}$	$\sigma\sqrt{n/T}$	$\sigma\sqrt{n/T}$	$(1+\sigma)\sqrt{n/T}$
Hankel Spectral Error	$\sigma\sqrt{n/T}$	$\sigma\sqrt{n/T}$	$\sigma n/\sqrt{T}$	$(1+\sigma)\sqrt{n/T}$

Algorithmic comparison

Table 1: Hankel matrix is $n \times n$, system order is R, number of samples is T. Noise level $\sigma = 1/\sqrt{\mathbf{snr}}$. LS-IR stands for regressing impulse response with output by LS, LS-Hankel stands for regressing Hankel with output by LS.

Paper	This work	This work	[OO '18]	[SRD '19]
Sample complexity	R^2	n	n	n^2
Method	Nuc-norm	LS-IR	LS-IR	LS-Hankel
IR ℓ_2 Error	$\sigma\sqrt{n/T}$	$\sigma\sqrt{n/T}$	$\sigma\sqrt{n/T}$	$(1+\sigma)\sqrt{n/T}$
Hankel Spectral Error	$\sigma\sqrt{n/T}$	$\sigma\sqrt{n/T}$	$\sigma n/\sqrt{T}$	$(1+\sigma)\sqrt{n/T}$

- Refined bound: optimal guarantee for LS-IR

Algorithmic comparison

Table 1: Hankel matrix is $n \times n$, system order is R, number of samples is T. Noise level $\sigma = 1/\sqrt{\mathbf{snr}}$. LS-IR stands for regressing impulse response with output by LS, LS-Hankel stands for regressing Hankel with output by LS.

Paper	This work	This work	[OO '18]	[SRD '19]
Sample complexity	R^2	n	n	n^2
Method	Nuc-norm	LS-IR	LS-IR	LS-Hankel
IR ℓ_2 Error	$\sigma\sqrt{n/T}$	$\sigma\sqrt{n/T}$	$\sigma\sqrt{n/T}$	$(1+\sigma)\sqrt{n/T}$
Hankel Spectral Error	$\sigma\sqrt{n/T}$	$\sigma\sqrt{n/T}$	$\sigma n/\sqrt{T}$	$(1+\sigma)\sqrt{n/T}$

- Refined bound: optimal guarantee for LS-IR
- Good error guarantees hold for $T \gtrsim R^2$ regime.

Algorithmic comparison

Table 1: Hankel matrix is $n \times n$, system order is R, number of samples is T. Noise level $\sigma = 1/\sqrt{\mathbf{snr}}$. LS-IR stands for regressing impulse response with output by LS, LS-Hankel stands for regressing Hankel with output by LS.

Paper	This work	This work	[OO '18]	[SRD '19]
Sample complexity	R^2	n	n	n^2
Method	Nuc-norm	LS-IR	LS-IR	LS-Hankel
IR ℓ_2 Error	$\sigma\sqrt{n/T}$	$\sigma\sqrt{n/T}$	$\sigma\sqrt{n/T}$	$(1+\sigma)\sqrt{n/T}$
Hankel Spectral Error	$\sigma\sqrt{n/T}$	$\sigma\sqrt{n/T}$	$\sigma n/\sqrt{T}$	$(1+\sigma)\sqrt{n/T}$

- Refined bound: optimal guarantee for LS-IR
- Good error guarantees hold for $T \gtrsim R^2$ regime.
  - Question: Can Nuc-norm error bounds be improved to $\sigma\sqrt{R/T}$ (independent of n)?

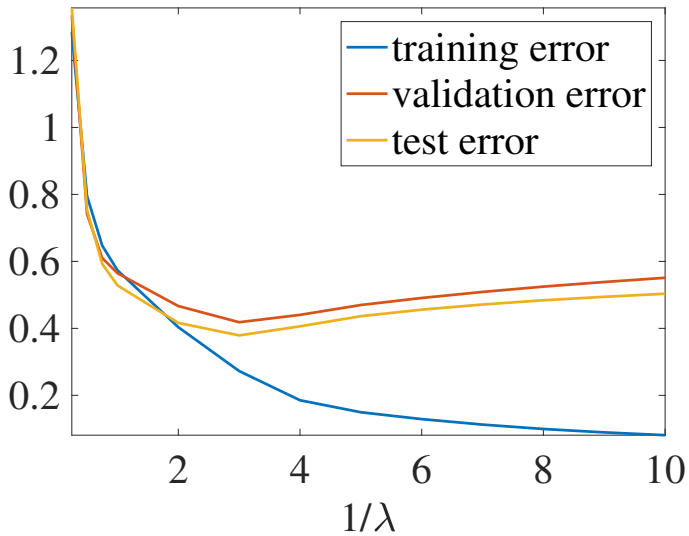
Algorithmic comparison

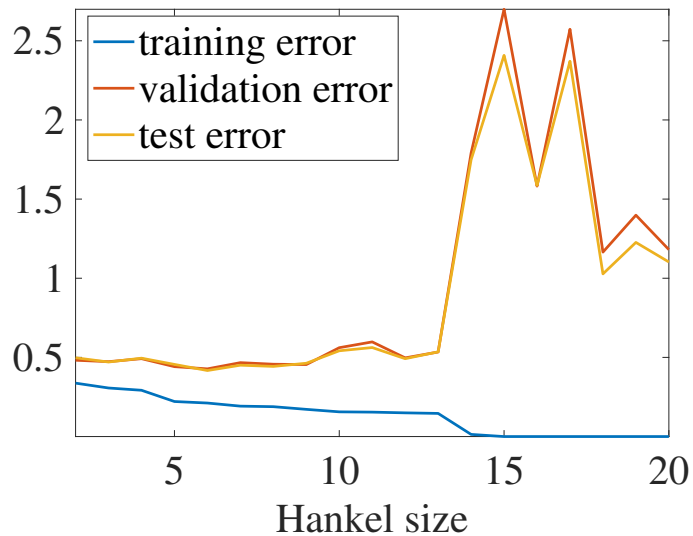
Table 1: Hankel matrix is $n \times n$, system order is R, number of samples is T. Noise level $\sigma = 1/\sqrt{\mathbf{snr}}$. LS-IR stands for regressing impulse response with output by LS, LS-Hankel stands for regressing Hankel with output by LS.

Paper	This work	This work	[OO '18]	[SRD '19]
Sample complexity	R^2	n	n	n^2
Method	Nuc-norm	LS-IR	LS-IR	LS-Hankel
IR ℓ_2 Error	$\sigma\sqrt{n/T}$	$\sigma\sqrt{n/T}$	$\sigma\sqrt{n/T}$	$(1+\sigma)\sqrt{n/T}$
Hankel Spectral Error	$\sigma\sqrt{n/T}$	$\sigma\sqrt{n/T}$	$\sigma n/\sqrt{T}$	$(1+\sigma)\sqrt{n/T}$

- Refined bound: optimal guarantee for LS-IR
- Good error guarantees hold for $T \gtrsim R^2$ regime.
  - Question: Can Nuc-norm error bounds be improved to $\sigma\sqrt{R/T}$ (independent of n)?
- If not knowing best λ, train-validation (Picking the solution with smallest validation error) ends up with same bounds.

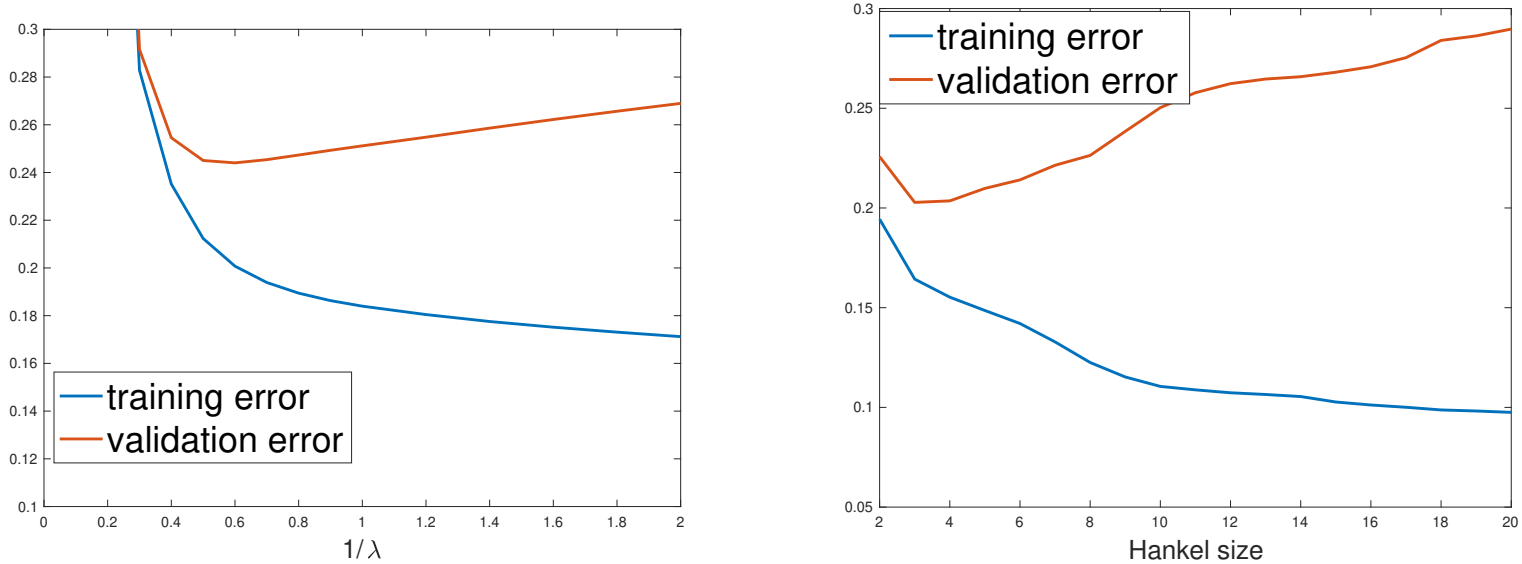
Experiments: regularization and least squares, Synthetic





System estimation for synthetic data, SNR $= 10$, assuming $n = 20$. Training data size $= 30$. (a) Training and validation error of different λ, (b) Training and validation error of different Hankel size n.

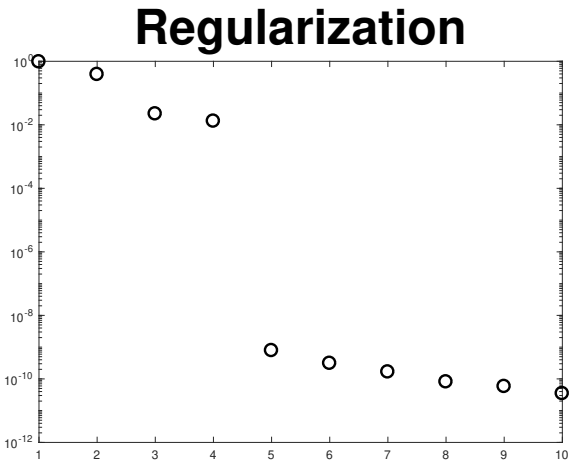
Experiments: regularization and least squares, DaISy dataset[2]

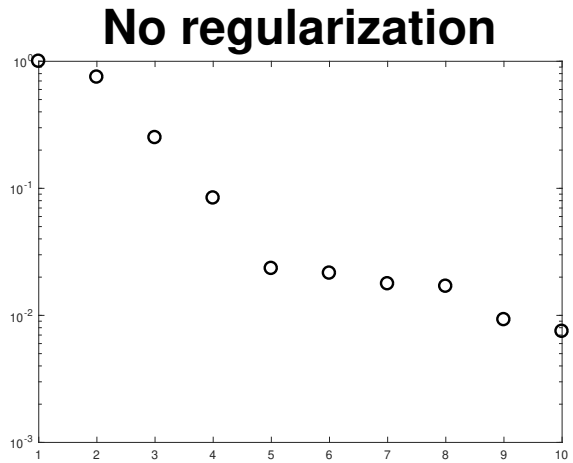


System ID for CD player arm data, assuming $n = 10$. Training data size = 200 and validation data size = 600. Data is single rollout, input is not random. (a) training and validation error of different λ, (b) training and validation error of different Hankel size n.

[2] De Moor et al. 1997

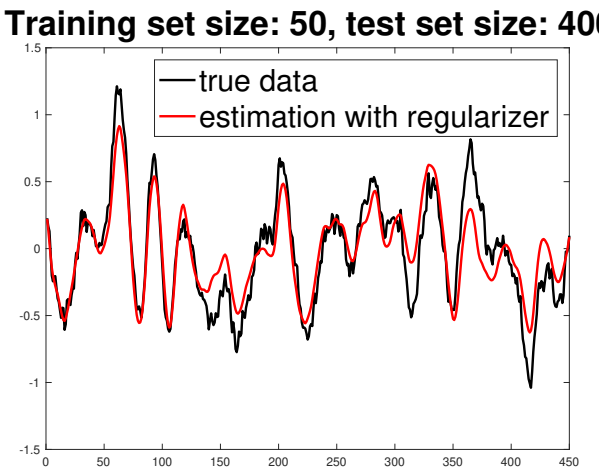
Experiments: regularization and least squares, DaISy dataset

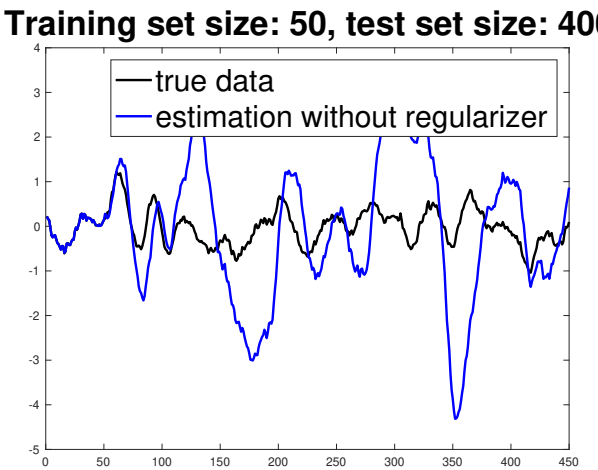




CD player arm data, $n = 10$, normalized singular value of (a) regularized (b) unregularized Hankel. We can see the low rank structure of recoverer Hankel matrix from regularization method.

Experiments: regularization and least squares, DaISy dataset





CD player arm data, (a) regularized (b) unregularized prediction, number of samples is 50 when Hankel matrix is 10×10. 400 output samples for validation.

Conclusion

- Nuclear norm regularization is practical but poorly understood
  - Simplifies model selection
  - Less sensitive to hyperparameter tuning
- New estimation error guarantees
  - Improved guarantees for nuclear norm regularization

Future: Better understanding the empirical behavior: spectral gap, sensitivity to regularization parameters, etc.
Can error bounds/statistical rates be improved?

Paper published at L4DC 2020. Coauthor: Samet Oymak, Maryam Fazel

Overview

Example of manifolds

Riemannian manifold: smooth manifold with inner product defined on tangent space.

1. Sphere. $\{x \in \mathbb{R}^d : \sum_{i=1}^d x_i^2 = r^2\}$.
2. Stiefel manifold. $\{X \in \mathbb{R}^{m \times n} : X^\top X = I\}$.
3. Grassmannian manifold. $\mathrm{Grass}(p, n)$ is set of p dimensional subspaces in $\mathbb{R}^n$.
4. Bruer-Monteiro relaxation. $\{X \in \mathbb{R}^{m \times n} : \mathrm{diag}(X^\top X) = \mathbf{1}\}$.

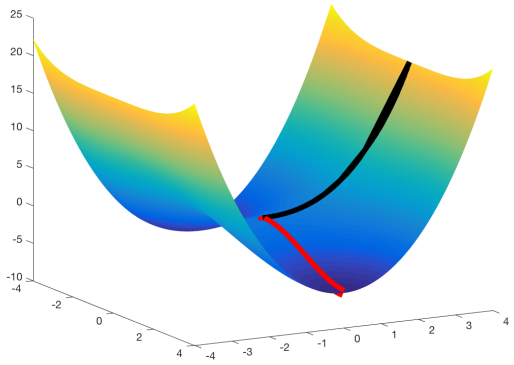
Optimization problem on manifold:
PCA (Edelman, Arias, and Smith 1998),
dictionary learning (Sun, Qu, and Wright 2017),
low rank matrix completion (Nicolas Boumal and P.-a. Absil 2011),
tensor factorization (Ishteva et al. 2011).

Manifold constrained optimization

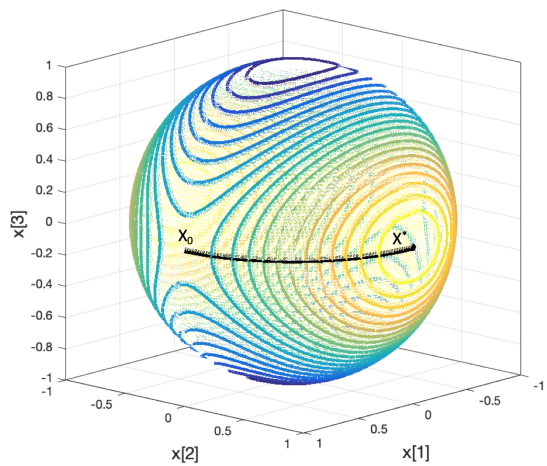
We consider the problem

$$\underset{x}{\text{minimize}} \quad f(x), \text{ subject to } \; x \in \mathcal{M}$$

Same as Euclidean space, generally we cannot find global optimum in polynomial time, so we want to find *an approximate local minimum*.



Plot of saddle point in Euclidean space.



Contour of function value on sphere.

Related work

Escaping from saddle, Euclidean space, unconstrained:
Jason D. Lee et al. 2016, 2017: Asymptotic analysis of escaping saddle.
Du et al. 2017: GD can escape saddle with exponential time.
Ge et al. 2015: Escaping from saddle via SGD.
Jin, Ge, et al. 2017: Escaping from saddle via perturbed GD.
Carmon and Duchi 2017: Escaping from saddle via cubic regularization oracle.
Jin, Netrapalli, and Jordan 2017: Escaping from saddle via perturbed heavy ball (GD with momentum)

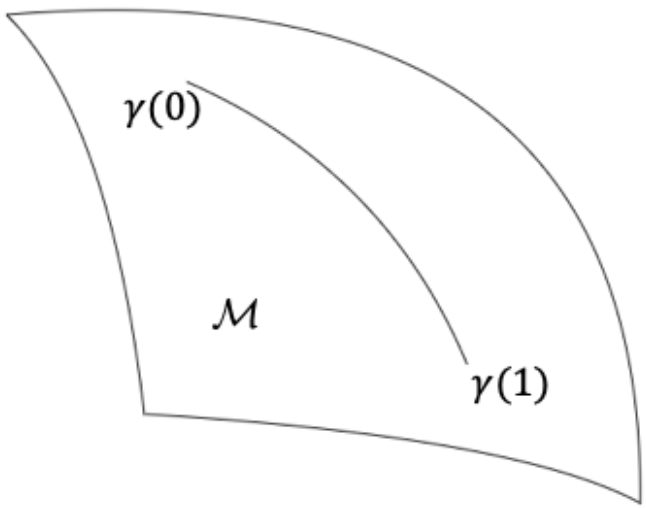
Related work

Escaping from saddle, Euclidean space, constrained: Generally NP hard.
Nouiehed, Jason D Lee, and Razaviyayn 2018; Mokhtari, Ozdaglar, and Jadbabaie 2018: Allow an oracle whose worst case complexity is exponential.
Avdiukhin, Jin, and Yaroslavtsev 2019; Lu, Zhao, et al. 2019; Lu, Razaviyayn, et al. 2019: Perturbed projected GD (the last paper with extra line search). Make extra assumptions to avoid NP-hardness.

Related work

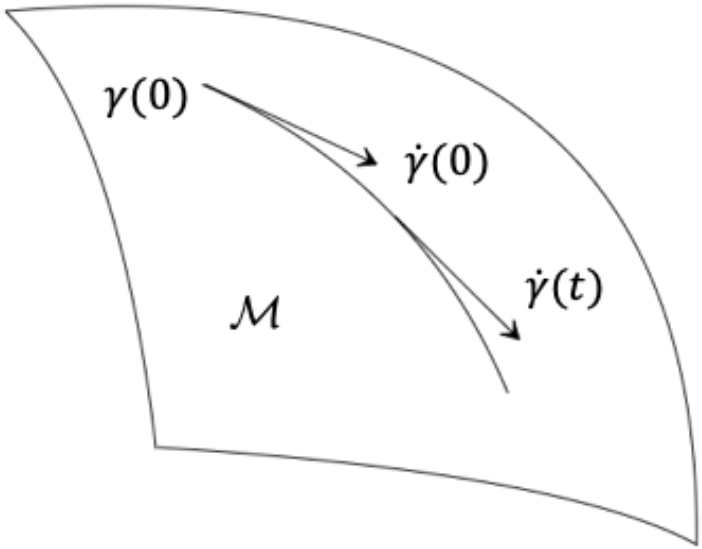
Escaping from saddle, Riemannian:
Nicolas Boumal, P.-A. Absil, and Cartis 2016; Kasai and Mishra 2018; Hu et al. 2018: Escaping saddle via Riemannian trust region.
J. Zhang and S. Zhang 2018; Agarwal et al. 2018: Escaping saddle via Riemannian cubic regularization.
Criscitiello and N. Boumal 2019 (same time as ours): Escaping from saddle via perturbed Riemannian GD.
D. Zhang and Tajbakhsh 2020 (later): Escaping from saddle via Riemannian stochastic cubic regularization.

Curve



A curve in a continuous map $\gamma(t) : \mathbb{R} \to \mathcal{M}$.
Usually $t \in [0, 1]$, where $\gamma(0)$ and $\gamma(1)$ are start and end points of the curve.

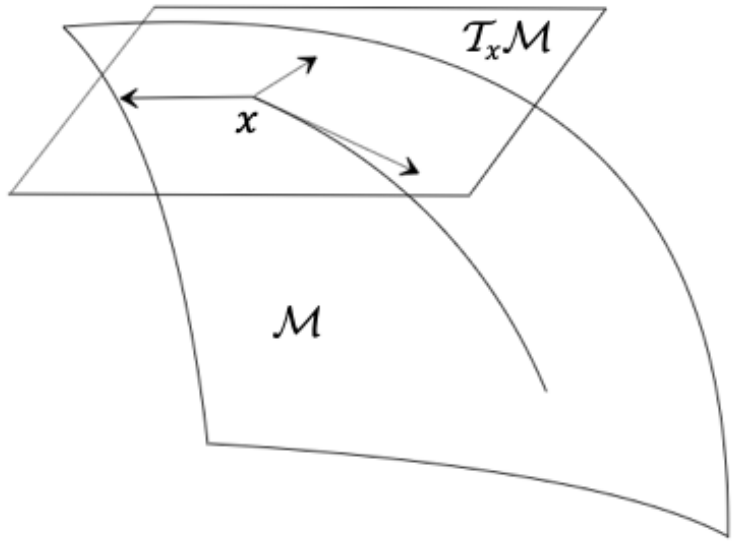
Tangent vector and tangent space



We use

$$\dot{\gamma}(t) = \lim_{\tau \to 0} \frac{\gamma(t+\tau) - \gamma(t)}{\tau}$$

as the velocity of the curve, $\dot{\gamma}(t)$ is a tangent vector at $\gamma(t) \in \mathcal{M}$.

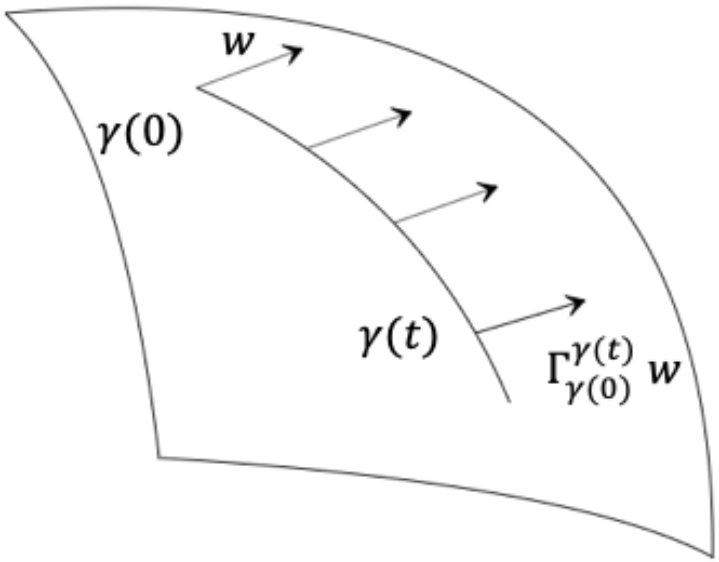


$x \in \mathcal{M}$ can be start point of many curves, and a tangent space $\mathcal{T}_x\mathcal{M}$ is the set of tangent vectors at x.

Tangent space is a metric space.

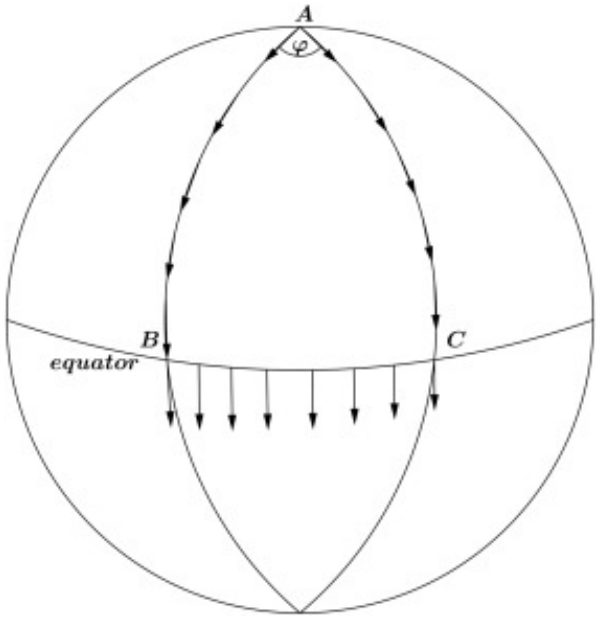
Riemannian gradient is defined in tangent space.

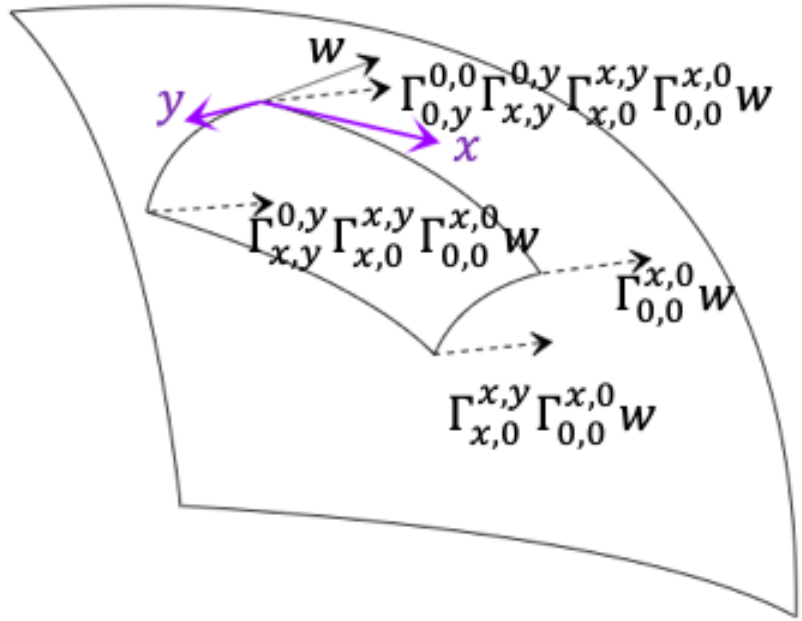
Parallel transport



The second order structure of manifold is called connection – we omit the math definition here. The parallel transport Γ transports a tangent vector w along a curve γ.

Curvature tensor

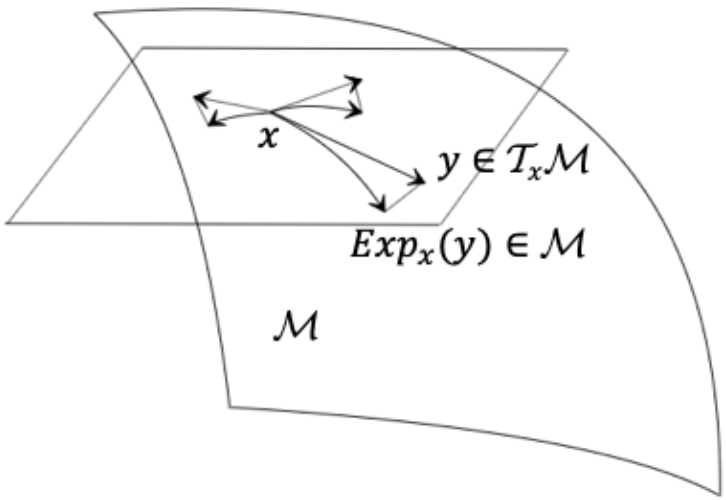




$$R(x,y)w = \lim_{t,\tau\to 0} \frac{\Gamma^{0,0}_{0,\tau y}\Gamma^{0,\tau y}_{tx,\tau y}\Gamma^{tx,\tau y}_{tx,0}\Gamma^{tx,0}_{0,0}w - w}{t\tau}$$

The curvature tensor describes how curved the manifold is. It relates to the second order structure of the manifold.

Exponential map – "projection" onto manifold



For any $x \in \mathcal{M}$, $y \in \mathcal{T}_x\mathcal{M}$ and the geodesic γ defined by y,

$$\gamma(0) = x, \ \dot{\gamma}(0) = y$$

we call the mapping $\mathrm{Exp}_x(y) : \mathcal{T}_x\mathcal{M} \to \mathcal{M}$ such that $\mathrm{Exp}_x(y) = \gamma(1)$ as exponential map.

There is a neighborhood with radius $\mathfrak{I}$ in $\mathcal{T}_x\mathcal{M}$, such that for all $y \in \mathcal{T}_x\mathcal{M}$, $\|y\| \leq \mathfrak{I}$, exponential map is a bijection/diffeomorphism.

Smooth function on Riemannian manifold

We consider the manifold constrained optimization problem

$$\underset{x}{\text{minimize}} \quad f(x), \text{ subject to } \; x \in \mathcal{M}$$

assuming the function and manifold satisfying

1. There is a finite constant β such that

$$\|\text{grad} f(y) - \Gamma_x^y \text{grad} f(x)\| \leq \beta d(x, y) \quad \text{for all } x, y \in \mathcal{M}.$$

2. There is a finite constant ρ such that

$$\|H(y) - \Gamma_x^y H(x) \Gamma_y^x\| \leq \rho d(x, y) \quad \text{for all } x, y \in \mathcal{M}.$$

3. There is a finite constant K such that

$$|R(x)[u, v]| \leq K \quad \text{for all } x \in \mathcal{M} \text{ and } u, v \in \mathcal{T}_x\mathcal{M}.$$

f may not be convex.

Riemannian gradient descent

Riemannian gradient descent:

$$x_{t+1} = \mathrm{Exp}_{x_t}(-\eta \mathrm{grad} f(x_t)),$$

Compare to projected GD in Euclidean

$$x_{t+1} = \mathcal{P}_{\mathcal{S}}(x_t - \eta \nabla f(x_t)),$$

Riemannian gradient descent

Riemannian gradient descent:

$$x_{t+1} = \mathrm{Exp}_{x_t}(-\eta \mathrm{grad} f(x_t)),$$

Compare to projected GD in Euclidean

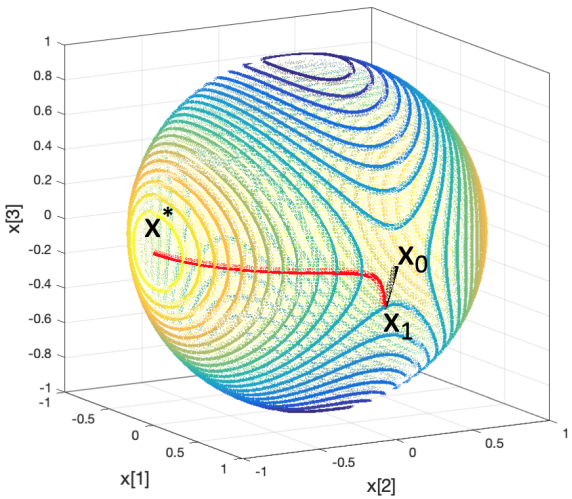
$$x_{t+1} = \mathcal{P}_{\mathcal{S}}(x_t - \eta \nabla f(x_t)),$$

There exists η such that

$$f(x_{t+1}) \leq f(x_t) - \frac{\eta}{2} \|\mathrm{grad} f(x_t)\|^2.$$

Converge to first order stationary.

Proposed algorithm for escaping saddle

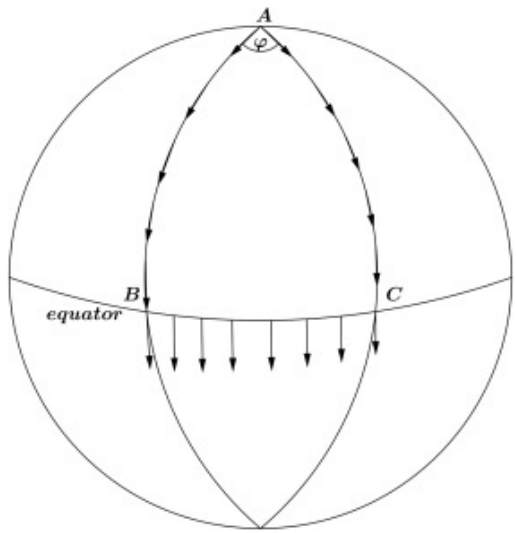


Hope to escape from saddle point and converge to an approximate local minimum.

1. At iterate x, check the norm of gradient.
2. If large: do $x^+ = \mathrm{Exp}_x(-\eta \mathrm{grad} f(x))$ to decrease function value.
3. If small: near either a saddle point or a local min. Perturb iterate by adding appropriate noise, run a few iterations.
   - 3.1 if f decreases, iterates escape saddle point (and alg continues).
   - 3.2 if f doesn't decrease: at approximate local min (alg terminates).

Difficulty of second order analysis

1. The second order curvature of function and manifold interact.
2. No universal coordinate system. We have to consider gradient in different tangent spaces.



Theorem

Theorem (Jin et al., Eucledean space)

Perturbed GD converges to a $(\epsilon, -\sqrt{\rho\epsilon})$-stationary point $(\|\nabla f\| \leq \epsilon, \lambda_{\min}(H) \geq -\sqrt{\rho\epsilon})$ with # of iterations

$$O\left(\frac{\beta(f(x_0) - f(x^*))}{\epsilon^2} \log^4\left(\frac{\beta d(f(x_0) - f(x^*))}{\epsilon^2 \delta}\right)\right)$$

We replace Hessian Lipschitz ρ by $\hat{\rho}$ as a function of ρ and K and we quantify it in the paper.

Theorem (manifold)

Perturbed RGD converges to a $(\epsilon, -\sqrt{\hat{\rho}(\rho, K)\epsilon})$-stationary point with # of iterations

$$O\left(\frac{\beta(f(x_0) - f(x^*))}{\epsilon^2} \log^4\left(\frac{\beta d(f(x_0) - f(x^*))}{\epsilon^2 \delta}\right)\right)$$

Experiment

Burer-Monteiro factorization.

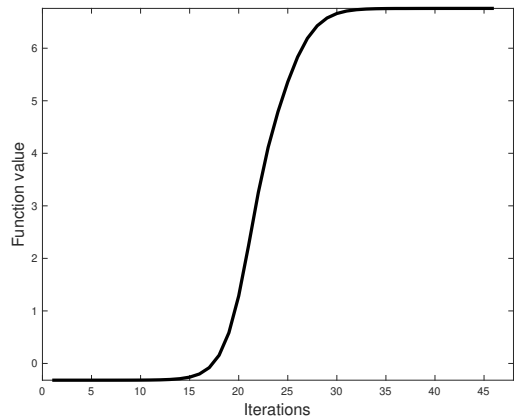
Let $A \in \mathbb{S}^{d\times d}$, the problem

$$\max_{X\in\mathbb{S}^{d\times d}} \mathbf{tr}(AX),$$
$$s.t.\ \mathrm{diag}(X) = 1, X \succeq 0, \mathrm{rank}(X) \leq r.$$

can be factorized as

$$\max_{Y\in\mathbb{R}^{d\times p}} \mathbf{tr}(AYY^\top),\ s.t.\ \mathrm{diag}(YY^\top) = 1.$$

when $r(r+1)/2 \leq d$, $p(p+1)/2 \geq d$.



Iteration versus function value.

Conclusion

- Escaping saddle with Riemannian GD is not known.
  - The second order structure of Riemannian manifold is non-trivial.
- New convergence guarantee.
  - Similar convergence rate to Euclidean space.
  - Reveals the role of Riemannian curvature.

Future: Study other first order methods on Riemannian manifold, whether accelerated GD helps with escaping, Riemannian SGD.

Papers published at ICML workshop 2018, NeurIPS 2019. Coauthor: Nicolas Flammarion, Maryam Fazel

Overview

Conclusion

- Nuclear norm regularization estimates the low-order linear system with less data.
- Perturbed Riemannian gradient descent escapes from saddle points on Riemannian manifolds, with similar behavior to Euclidean space.

Novel analysis of convergence of policy gradient descent via convexification

$$\dot{x}(t) = Ax(t) + Bu(t),\ x(0) = x_0,$$

$$loss(u(t)) = \mathbf{E}_{x_0 \sim \mathcal{N}(0,\Omega)} \int_0^\infty (x(t)^\top Q x(t) + u(t)^\top R u(t))dt$$

$$\Downarrow$$

$$\min_K \quad loss(K),$$

$$\text{s.t.,} \quad u(t) = Kx(t),\ K \text{ stabilizes}$$

$$\min_{Z,L,G} \ \mathbf{tr}(QG + ZR)$$

$$\text{s.t., } \mathcal{A}(G) + \mathcal{B}(L) + \Omega = 0,$$

$$G \succ 0,$$

$$\begin{bmatrix} Z & L^\top \\ L & G \end{bmatrix} \succeq 0$$

And $K^* = L^* G^{*-1}$.

- Model free method: estimate $\nabla loss(K)$ and run $K^+ = K - \eta \nabla loss(K)$ – Policy GD.
- Loss is nonconvex in K.
- In ML people run gradient descent on nonconvex functions.
- GD is more implementable than solving SDP (e.g., by interior point method).

Novel analysis of convergence of policy gradient descent via convexification

$$\dot{x}(t) = Ax(t) + Bu(t),\ x(0) = x_0,$$

$$loss(u(t)) = \mathbf{E}_{x_0 \sim \mathcal{N}(0,\Omega)} \int_0^\infty (x(t)^\top Q x(t) + u(t)^\top R u(t)) dt$$

$$\Downarrow$$

$$\min_K \quad loss(K),$$

$$\text{s.t.,} \quad u(t) = Kx(t),\ K \text{ stabilizes}$$

$$\min_{Z,L,G} \ \mathbf{tr}(QG + ZR)$$

$$\text{s.t., } \mathcal{A}(G) + \mathcal{B}(L) + \Omega = 0,$$

$$G \succ 0,$$

$$\begin{bmatrix} Z & L^\top \\ L & G \end{bmatrix} \succeq 0$$

And $K^* = L^* G^{*-1}$.

- Model free method: estimate $\nabla loss(K)$ and run $K^+ = K - \eta \nabla loss(K)$ – Policy GD.
- Loss is nonconvex in K.
- In ML people run gradient descent on nonconvex functions.
- GD is more implementable than solving SDP (e.g., by interior point method).

Does policy GD converge to optimal controller?

Can we prove convergence beyond LQR?

Novel analysis of convergence of policy gradient descent via convexification

$$\dot{x}(t) = Ax(t) + Bu(t),\ x(0) = x_0,$$

$$loss(u(t)) = \mathbf{E}_{x_0 \sim \mathcal{N}(0,\Omega)} \int_0^\infty (x(t)^\top Q x(t) + u(t)^\top R u(t)) dt$$

$$\Downarrow$$

$$\min_K \quad loss(K),$$

$$\text{s.t.,} \quad u(t) = Kx(t),\ K \text{ stabilizes}$$

$$\min_{Z,L,G} \ \mathbf{tr}(QG + ZR)$$

$$\text{s.t., } \mathcal{A}(G) + \mathcal{B}(L) + \Omega = 0,$$

$$G \succ 0,$$

$$\begin{bmatrix} Z & L^\top \\ L & G \end{bmatrix} \succeq 0$$

And $K^* = L^* G^{*-1}$.

Theorem (simplified)

For a few continuous time optimal control problems including LQR, minimizing $\mathcal{L}_2$ gain, $\nabla loss(K) = 0 \ \Leftrightarrow \ K = K^*$.

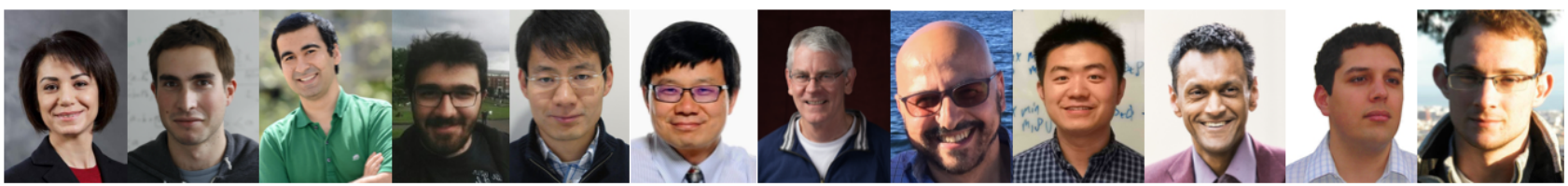
No saddle point or spurious local minimum – optimizable by GD

We propose a novel proof bridging nonconvex and convex formulations

Other works and acknowledgement

- (Ongoing project) Representation dimension and non-linearity in representation learning. Paper submitted to ICASSP 2021.
- (Internship at Google) Online learning for video coding system. Paper published at DCC 2020.

Acknowledgement:



References I

- Agarwal, Naman et al. (2018). "Adaptive regularization with cubics on manifolds". In: *arXiv preprint arXiv:1806.00065*.
- Avdiukhin, Dmitrii, Chi Jin, and Grigory Yaroslavtsev (2019). "Escaping Saddle Points with Inequality Constraints via Noisy Sticky Projected Gradient Descent". In: *11th Annual Workshop on Optimization for Machine Learning*.
- Blomberg, Niclas (2016). "On nuclear norm minimization in system identification". PhD thesis. KTH Royal Institute of Technology.
- Boumal, Nicolas and Pierre-antoine Absil (2011). "RTRMC: A Riemannian trust-region method for low-rank matrix completion". In: *Advances in neural information processing systems*, pp. 406–414.
- Boumal, Nicolas, Pierre-Antoine Absil, and Coralia Cartis (2016). "Global rates of convergence for nonconvex optimization on manifolds". In: *IMA Journal of Numerical Analysis*.

References II

- Cai, Jian-Feng et al. (2016). "Robust recovery of complex exponential signals from random Gaussian projections via low rank Hankel matrix reconstruction". In: *Applied and computational harmonic analysis* 41.2, pp. 470–490.
- Carmon, Yair and John C. Duchi (2017). "Gradient Descent Efficiently Finds the Cubic-Regularized Non-Convex Newton Step". In: *arXiv preprint arXiv:1612.00547*.
- Criscitiello, C. and N. Boumal (2019). "Efficiently escaping saddle points on manifolds". In: *arXiv preprint arXiv:1906.04321*.
- De Moor, Bart et al. (1997). "DAISY: A database for identification of systems". In: *JOURNAL A* 38, pp. 4–5.
- Dean, Sarah et al. (2017). "On the sample complexity of the linear quadratic regulator". In: *arXiv preprint arXiv:1710.01688*.
- – (2018). "Regret bounds for robust adaptive control of the linear quadratic regulator". In: *Advances in Neural Information Processing Systems*, pp. 4188–4197.

References III

- Du, Simon S et al. (2017). "Gradient descent can take exponential time to escape saddle points". In: *Advances in Neural Information Processing Systems*, pp. 1067–1077.
- Edelman, Alan, Tomás A Arias, and Steven T Smith (1998). "The geometry of algorithms with orthogonality constraints". In: *SIAM journal on Matrix Analysis and Applications* 20.2, pp. 303–353.
- Faradonbeh, Mohamad Kazem Shirani, Ambuj Tewari, and George Michailidis (2018). "Finite time identification in unstable linear systems". In: *Automatica* 96, pp. 342–353.
- Fazel, Maryam, Haitham Hindi, and Stephen P Boyd (2003). "Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices". In: *Proceedings of the 2003 American Control Conference, 2003.* Vol. 3. IEEE, pp. 2156–2162.

References IV

- Fazel, Maryam, Ting Kei Pong, et al. (2013). "Hankel matrix rank minimization with applications to system identification and realization". In: *SIAM Journal on Matrix Analysis and Applications* 34.3, pp. 946–977.
- Foster, Dylan J, Alexander Rakhlin, and Tuhin Sarkar (2020). "Learning nonlinear dynamical systems from a single trajectory". In: *arXiv preprint arXiv:2004.14681*.
- Ge, Rong et al. (2015). "Escaping from saddle points – online stochastic gradient for tensor decomposition". In: *Conference on Learning Theory*, pp. 797–842.
- Hansson, Anders, Zhang Liu, and Lieven Vandenberghe (2012). "Subspace system identification via weighted nuclear norm optimization". In: *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*. IEEE, pp. 3439–3444.

References V

- Hazan, Elad, Holden Lee, et al. (2018). "Spectral filtering for general linear dynamical systems". In: *Advances in Neural Information Processing Systems*, pp. 4634–4643.
- Hazan, Elad, Karan Singh, and Cyril Zhang (2017). "Learning linear dynamical systems via spectral filtering". In: *Advances in Neural Information Processing Systems*, pp. 6705–6715.
- Ho, BL and Rudolf E Kálmán (1966). "Effective construction of linear state-variable models from input/output functions". In: *at-Automatisierungstechnik* 14.1-12, pp. 545–548.
- Hu, J. et al. (2018). "Adaptive quadratically regularized Newton method for Riemannian optimization". In: *SIAM J. Matrix Anal. Appl.* 39.3, pp. 1181–1207.
- Ishteva, Mariya et al. (2011). "Best low multilinear rank approximation of higher-order tensors, based on the Riemannian trust-region scheme". In: *SIAM Journal on Matrix Analysis and Applications* 32.1, pp. 115–135.

References VI

- Jin, Chi, Rong Ge, et al. (2017). "How to escape saddle points efficiently". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, pp. 1724–1732.
- Jin, Chi, Praneeth Netrapalli, and Michael I. Jordan (2017). "Accelerated Gradient Descent Escapes Saddle Points Faster than Gradient Descent". In: *arXiv preprint arXiv:1711.10456*.
- Kasai, H. and B. Mishra (2018). "Inexact trust-region algorithms on Riemannian manifolds". In: *Advances in Neural Information Processing Systems 31*, pp. 4254–4265.
- Lee, Jason D. et al. (2016). "Gradient descent only converges to minimizers". In: *Conference on Learning Theory*, pp. 1246–1257.
- Lee, Jason D et al. (2017). "First-order methods almost always avoid saddle points". In: *arXiv preprint arXiv:1710.07406*.
- Liu, Zhang, Anders Hansson, and Lieven Vandenberghe (2013). "Nuclear norm system identification with missing inputs and outputs". In: *Systems & Control Letters* 62.8, pp. 605–612.

References VII

- Lu, Songtao, Meisam Razaviyayn, et al. (2019). "SNAP: Finding Approximate Second-Order Stationary Solutions Efficiently for Non-convex Linearly Constrained Problems". In: *arXiv preprint arXiv:1907.04450*.
- Lu, Songtao, Ziping Zhao, et al. (2019). "Perturbed Projected Gradient Descent Converges to Approximate Second-order Points for Bound Constrained Nonconvex Problems". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5356–5360.
- Mania, Horia, Stephen Tu, and Benjamin Recht (2019). "Certainty equivalent control of LQR is efficient". In: *arXiv preprint arXiv:1902.07826*.
- Mokhtari, Aryan, Asuman Ozdaglar, and Ali Jadbabaie (2018). "Escaping Saddle Points in Constrained Optimization". In: *arXiv preprint arXiv:1809.02162*.

References VIII

- Nouiehed, Maher, Jason D Lee, and Meisam Razaviyayn (2018). "Convergence to Second-Order Stationarity for Constrained Non-Convex Optimization". In: *arXiv preprint arXiv:1810.02024*.
- Oymak, Samet and Necmiye Ozay (2018). "Non-asymptotic identification of lti systems from a single trajectory". In: *arXiv preprint arXiv:1806.05722*.
- Sarkar, Tuhin and Alexander Rakhlin (2019). "Near optimal finite time identification of arbitrary linear dynamical systems". In: *arXiv preprint arXiv:1812.01251*.
- Sarkar, Tuhin, Alexander Rakhlin, and Munther A Dahleh (2019). "Finite-Time System Identification for Partially Observed LTI Systems of Unknown Order". In: *arXiv preprint arXiv:1902.01848*.
- Sattar, Yahya and Samet Oymak (2020). "Non-asymptotic and accurate learning of nonlinear dynamical systems". In: *arXiv preprint arXiv:2002.08538*.

References IX

- Simchowitz, Max, Ross Boczar, and Benjamin Recht (2019). "Learning Linear Dynamical Systems with Semi-Parametric Least Squares". In: *arXiv preprint arXiv:1902.00768*.
- Simchowitz, Max, Horia Mania, et al. (2018). "Learning Without Mixing: Towards A Sharp Analysis of Linear System Identification". In: *Conference On Learning Theory*, pp. 439–473.
- Simchowitz, Max, Karan Singh, and Elad Hazan (2020). "Improper learning for non-stochastic control". In: *arXiv preprint arXiv:2001.09254*.
- Sun, Ju, Qing Qu, and John Wright (2017). "Complete dictionary recovery over the sphere II: Recovery by Riemannian trust-region method". In: *IEEE Transactions on Information Theory* 63.2, pp. 885–914.
- Tsiamis, Anastasios and George J Pappas (2019). "Finite Sample Analysis of Stochastic System Identification". In: *arXiv preprint arXiv:1903.09122*.

References X

- Tu, Stephen et al. (2017). "Non-Asymptotic Analysis of Robust Control from Coarse-Grained Identification". In: *arXiv preprint arXiv:1707.04791*.
- Van Overschee, Peter and Bart De Moor (1995). "A unifying theorem for three subspace system identification algorithms". In: *Automatica* 31.12, pp. 1853–1864.
- Verhaegen, Michel and Anders Hansson (2016). "N2SID: Nuclear norm subspace identification of innovation models". In: *Automatica* 72, pp. 57–63.
- Zhang, Dewei and Sam Davanloo Tajbakhsh (2020). "Riemannian Stochastic Variance-Reduced Cubic Regularized Newton Method". In: *arXiv preprint arXiv:2010.03785*.
- Zhang, J. and S. Zhang (2018). "A Cubic Regularized Newton's Method over Riemannian Manifolds". In: *arXiv preprint arXiv:1805.05565*.