Nonconvex optimization and model representation with applications in control theory and machine learning

Yue Sun

November 18, 2020

Contents

1 Introduction **3**

2 First-order method for nonconvex optimization problem on Riemannian manifolds **5**
2.1 Introduction: escaping from saddle points on Riemannian manifolds . . . 5
2.2 Notation and Background . . . 6
2.3 Perturbed Riemannian gradient algorithm . . . 7
2.4 Main theorem: escape rate for perturbed Riemannian gradient descent . . . 8
2.5 Proof of Lemma 2 . . . 9
2.6 Proof of main theorem . . . 11
2.7 Examples . . . 12
2.8 Summary . . . 12

3 Learning linear dynamical systems via nuclear norm regularization **14**
3.1 Introduction: linear system identification from input-output data via regularized least squares . . . 14
3.2 Problem setup and algorithms . . . 16
3.3 IID inputs and the importance of input shape . . . 17
3.4 Hankel nuclear norm regularization . . . 18
3.5 Least-squares bounds . . . 19
3.6 Model selection for regularized system identification . . . 20
3.7 Experiments . . . 20
3.7.1 Experiments with synthetic data . . . 20
3.7.2 Experiments with DaISy Dataset . . . 21
3.8 Future directions . . . 21

4 Future work **25**
4.1 Novel analysis of convergence of policy gradient descent via convexification . . . 25
4.1.1 Introduction . . . 25
4.1.2 Static controller . . . 26
4.1.3 Review of convexification method for continuous LQR . . . 26
4.1.4 Main theorem . . . 27
4.1.5 Dynamic controller . . . 30
4.1.6 Conclusion and discussion . . . 32
4.2 Understanding the role of representation dimension in meta-learning . . . 32
4.2.1 Introduction . . . 32
4.2.2 Problem formulation . . . 33
4.2.3 Experiments . . . 35
4.2.4 Future directions . . . 37

A Appendix of Section 2 **45**
A.1 Taylor expansions on Riemannian manifold . . . 45
A.1.1 Taylor expansion for the gradient . . . 45
A.1.2 Taylor expansion for the function . . . 45
A.2 Linearization of the iterates in a fixed tangent space . . . 46
A.2.1 Evolution of $\mathrm{Exp}_u^{-1}(w)$ . . . 46
A.2.2 Evolution of $\mathrm{Exp}_x^{-1}(w) - \mathrm{Exp}_x^{-1}(u)$ . . . 48
A.2.3 Control of two-steps iteration . . . 49
A.3 Auxilliary lemmas . . . 51
A.4 Proof of Lemma 7 and 8 . . . 51
A.4.1 Proof of Lemma 7 . . . 52
A.4.2 Proof of Lemma 8 . . . 52
A.4.3 Proof of function value decrease at an approximate saddle point . . . 53

B Appendix of Section 3 **54**
B.1 Sample complexity for MISO and MIMO problems . . . 54
B.1.1 Proof of Theorem 10 . . . 55
B.2 Proof of error of regularized method . . . 57
B.3 Bounding Γ, where do we lose? . . . 60
B.4 Proof of suboptimal recovery guarantee with i.i.d. input . . . 61
B.5 Proof of least square spectral norm error . . . 62
B.6 Proof of model selection method . . . 63

1 Introduction

In the last decades, people have witnessed the power of machine learning models, which extracts the useful information from the data, and accomplishes variant difficult tasks based on the learnt information. A machine learning model can be as simple as a linear map, which is trained by solving a linear regression problem on features and labels. In recent years, more complicated models are specifically investigated, which perform remarkably well in many applications, such as robotics, image classification, objective detection, machine translation, recommendation systems, etc. Although these models behave well in practice, we do not have a good theoretical understanding of these methods. In this work, we aim to study two components raised in learning problems:

1. What is a good formulation of the machine learning model, that both reflects the real world phenomenon and works well with less training data required.

2. How to train the machine learning model in an computationally efficient way, so that one can find the best performing model in short time.

The first challenge means that, it is important to define the correct model for machine learning tasks, and see how the structure of models enable efficient learning. In the system identification and representation learning applications below, we hope to learn "simple" models that represents the real tasks. Both of them involves low rank structure of models, corresponding to the simplicity. We propose that, the low rank structure enables the algorithm to train on less data to retrieve the true model with small error, and a proper representation of model leads to optimal generalization guarantee, compared with the naive rank-reduction method.

Besides that, we are interested in the convergence guarantee of optimization methods for training machine learning models. Although convex optimization is well studied, during the recent trend in machine learning, people have applied gradient based algorithms for solving non-convex optimization problems, and they perform well in the empirical tasks. Thus we are interested in studying the theoretical convergence guarantee of those nonconvex optimization algorithms.

This section is a brief introduction of the following sections, and a full introduction and literature review of each theme will be specified in the corresponding sections.

- Thanks to recent boom of machine learning, people have tried to apply gradient based method to nonconvex problem, and found that they also converge fairly well despite the missing understanding of the landscape of the objective function. Thus people moved forward to study the case when a convergence guarantee can also be obtained. One line among them is strict saddle functions (Ge et al., 2015; Jin et al., 2017a), which suggests that we can find an approximate local minimum in polynomial time. Another line of work is to study the optimization method on a manifold (Absil et al., 2009b), which is generally a nonconvex optimization problem if we trivially regard the manifold as a constraint. One can combine the geometric structure of the manifold and the convex optimization algorithms to obtain the convergence guarantee of so-called Riemannian gradient methods, which is a gradient based method implemented on manifolds. However, the nonconvex optimization problem on manifolds are less studied (in fact convexity is not even well defined) before. In Section 2, we investigate the convergence rate to an approximate minimum on an Riemannian manifold, and bound the rate by the curvature constants of the function and manifold.

- Next part is the study on the system identification problem, which belongs to the model based method more usually used before. Previous works such as (Oymak & Ozay, 2018; Sarkar et al., 2019) use unregularized least squares method to regress the input-output map, however if we do not know the dimension of state space, the train-validation step (in order to find the state dimension) is required and not easy to implement. We study the benefit of using a regularizer in the system identification algorithm. We study the Hankel nuclear norm regularized problem, which encourages the simplicity of a linear system, and it saves number of required observations while the statistical rate of error is preserved and it's easy to run in practice. We study the statistical property of, and also propose a practical training-validation algorithm that tunes the regularizer efficiently.

- Finally, there are two ongoing projects being wrapped up for paper submissions. The first work applies the nonconvex optimization theories to a set of optimal control problem. Previously people establish optimal control theory on the Lyapunov theory or convex optimization theory (Dullerud & Paganini, 2013; Stengel, 1994; Rawlings et al., 2017). Motivated by recent papers (Fazel et al., 2018; Mohammadi et al., 2019; Bu et al., 2019) that directly study the nonconvex landscape of the linear quadratic regulator problem, we propose an explanation that connects convex and nonconvex analysis, and generalize nonconvex analysis to a broad range of

optimal control problems. The second work studies the role of overparametrization and dimension reduction in representation learning. It aims to uncover the principle features of the tasks, which are often low-dimensional, from limited data available for related tasks. We consider a setup where task features are approximately in a low dimensional subspace, and we do not know the subspace and its dimension. As mentioned in Kong et al. (2020b,a), A low rank approximation step is commonly used to retrieve the low dimensional space. The dimension reduction step is not necessarily optimal for generalizing to the new meta-test task. We plan to show that learning large representations where directions are weighted by their relative importance, although leading to an ill-posed overparameterized problem, can in fact be the right choice over small representations. Furthermore, our findings will reveal a double descent phenomena when the representation dimension coincides with the sample size.

2 First-order method for nonconvex optimization problem on Riemannian manifolds

In this section, we investigate the second order convergence guarantee of the first order optimization algorithms. It is known that, for solving an unconstrained optimization problem in Euclidean space, the perturbed gradient descent algorithm converges in polynomial time. In this work, we analyze the first order optimization algorithm on Riemannian manifold, and show that perturbed Riemannian gradient descent provably converges to an approximate local minimum. We give the concrete convergence rate of perturbed Riemannian gradient descent, which reveals the role of the manifold curvature with respect to the rate.

This work is published as Sun et al. (2019).

2.1 Introduction: escaping from saddle points on Riemannian manifolds

We consider minimizing a non-convex smooth function on a smooth manifold $\mathcal{M}$,

$$\min_{x \in \mathcal{M}} \quad f(x), \tag{1}$$

where $\mathcal{M}$ is a d-dimensional smooth manifold[1], and f is twice differentiable, with a Hessian that is ρ-Lipschitz (assumptions are formalized in section 4). This framework includes a wide range of fundamental problems (often non-convex), such as PCA (Edelman et al., 1998), dictionary learning (Sun et al., 2017), low rank matrix completion (Boumal & Absil, 2011), and tensor factorization (Ishteva et al., 2011). Finding the global minimum to Eq. (1) is in general NP-hard; our goal is to find an approximate second order stationary point with first order optimization methods. We are interested in first-order methods because they are extremely prevalent in machine learning, partly because computing Hessians is often too costly. It is then important to understand how first-order methods fare when applied to nonconvex problems, and there has been a wave of recent interest on this topic since (Ge et al., 2015), as reviewed below.

In the Euclidean space, it is known that with random initialization, gradient descent avoids saddle points asymptotically (Pemantle, 1990; Lee et al., 2016). Lee et al. (2017) (section 5.5) show that this is also true on smooth manifolds, although the result is expressed in terms of nonstandard manifold smoothness measures. Also, importantly, this line of work does not give quantitative rates for the algorithm's behaviour near saddle points.

Du et al. (2017) show gradient descent can be *exponentially slow* in the presence of saddle points. To alleviate this phenomenon, it is shown that for a β-gradient Lipschitz, ρ-Hessian Lipschitz function, cubic regularization (Carmon & Duchi, 2017) and perturbed gradient descent (Ge et al., 2015; Jin et al., 2017a) converges to $(\epsilon, -\sqrt{\rho\epsilon})$ local minimum[2] in polynomial time, and momentum based method accelerates (Jin et al., 2017b). Much less is known about inequality constraints: Nouiehed et al. (2018) and Mokhtari et al. (2018) discuss second order convergence for general inequality-constrained problems, where they need an NP-hard subproblem (checking the co-positivity of a matrix) to admit a polynomial time approximation algorithm. However such an approximation exists only under very restrictive assumptions. Recent works (Avdiukhin et al., 2019; Lu et al., 2019b,a) show that, when the negative curvature direction of saddle points always coordinate well with the nonlinear constraints, the perturbed projected gradient descent algorithm always converges to an approximate second order minimum. But it is unknown whether this assumption applies to the loss landscape of any well known applications.

An orthogonal line of work is optimization on Riemannian manifolds. Absil et al. (2009a) provide comprehensive background, showing how algorithms such as gradient descent, Newton and trust region methods can be implemented on Riemannian manifolds, together with asymptotic convergence guarantees to first order stationary points. Zhang & Sra (2016) provide global convergence guarantees for first order methods when optimizing geodesically convex functions. Bonnabel (2013) obtains the first asymptotic convergence result for stochastic gradient descent in this setting, which is further extended by Tripuraneni et al. (2018); Zhang et al. (2016); Khuzani & Li (2017). If the problem is non-convex, or the Riemannian Hessian is not positive definite, one can use second order methods to escape from saddle points. Boumal et al. (2016a) shows that Riemannian trust region method converges to a second order stationary point in polynomial time (see, also, Kasai & Mishra, 2018; Hu et al., 2018; Zhang & Zhang, 2018). But this method requires a Hessian oracle, whose complexity is d times more than computing gradient. In Euclidean space, trust region subproblem can be sometimes solved via a Hessian-vector product oracle, whose complexity is about the same as computing gradients. Agarwal et al. (2018) discuss its implementation on Riemannian manifolds, but not clear about the complexity and sensitivity of Hessian vector product oracle on manifold.

[1] Here d is the dimension of the manifold itself; we do not consider $\mathcal{M}$ as a submanifold of a higher dimensional space. For instance, if $\mathcal{M}$ is a 2-dimensional sphere embedded in $\mathbb{R}^3$, its dimension is $d = 2$.

[2] defined as x satisfying $\|\nabla f(x)\| \le \epsilon$, $\lambda_{\min}\nabla^2 f(x) \ge -\sqrt{\rho\epsilon}$

The study of the convergence of gradient descent for non-convex Riemannian problems is previously done only in the Euclidean space by modeling the manifold with equality constraints. Ge et al. (2015, Appendix B) prove that stochastic projected gradient descent methods converge to second order stationary points in polynomial time (here the analysis is not geometric, and depends on the algebraic representation of the equality constraints). Sun & Fazel (2018) proves perturbed projected gradient descent converges with a comparable rate to the unconstrained setting (Jin et al., 2017a) (polylog in dimension). The paper applies projections from the ambient Euclidean space to the manifold and analyzes the iterations under the Euclidean metric. This approach loses the geometric perspective enabled by Riemannian optimization, and cannot explain convergence rates in terms of inherent quantities such as the sectional curvature of the manifold.

After finishing this work, we found the recent and independent paper Criscitiello & Boumal (2019) which gives a similar convergence analysis result for a related perturbed Riemannian gradient method. We point out a few differences: (1) In Criscitiello & Boumal (2019) Lipschitz assumptions are made on the pullback map $f \circ \mathrm{Retr}$. While this makes the analysis simpler, it lumps the properties of the function and the manifold together, and the role of the manifold's curvature is not explicit. In contrast, our rates are expressed in terms of the function's smoothness parameters and the sectional curvature of the manifold separately, capturing the geometry more clearly. (2) The algorithm in Criscitiello & Boumal (2019) uses two types of iterates (some on the manifold but some taken on a tangent space), whereas all our algorithm steps are directly on the manifold, which is more natural. (3) To connect our iterations with intrinsic parameters of the manifold, we use the exponential map instead of the more general retraction used in Criscitiello & Boumal (2019). There are recent works analyzing other algorithms for escaping from saddle points on manifolds, such as cubic regularization (Agarwal et al., 2018), stochastic gradient descent (Durmus et al., 2020), stochastic variance reduced cubic regularization (Zhang & Tajbakhsh, 2020), etc.

Contributions. We provide convergence guarantees for perturbed first order Riemannian optimization methods to second-order stationary points (local minimum). We prove that as long as the function is appropriately smooth and the manifold has bounded sectional curvature, a perturbed Riemannian gradient descent algorithm escapes (an approximate) saddle points with a rate of $1/\epsilon^2$, a polylog dependence on the dimension of the manifold (hence almost dimension-free), and a polynomial dependence on the smoothness and curvature parameters. This is the first result showing such a rate for Riemannian optimization, and the first to relate the rate to geometric parameters of the manifold.

Despite analogies with the unconstrained (Euclidean) analysis and with the Riemannian optimization literature, the technical challenge in our proof goes beyond combining two lines of work: we need to analyze the interaction between the first-order method and the second order structure of the manifold to obtain second-order convergence guarantees that depend on the manifold curvature. Unlike in Euclidean space, the curvature affects the Taylor approximation of gradient steps. On the other hand, unlike in the local rate analysis in first-order Riemannian optimization, our second-order analysis requires more refined properties of the manifold structure (whereas in prior work, first order oracle makes enough progress for a local convergence rate proof, see Lemma 1), and second order algorithms such as (Boumal et al., 2016a) use second order oracles (Hessian evaluation). See section 4 for further discussion.

2.2 Notation and Background

We consider a complete[3], smooth, d dimensional Riemannian manifold $(\mathcal{M}, \mathfrak{g})$, equipped with a Riemannian metric $\mathfrak{g}$, and we denote by $\mathcal{T}_x\mathcal{M}$ its tangent space at $x \in \mathcal{M}$ (which is a vector space of dimension d). We also denote by $\mathbb{B}_x(r) = \{v \in \mathcal{T}_x\mathcal{M}, \|v\| \leq r\}$ the ball of radius r in $\mathcal{T}_x\mathcal{M}$ centered at 0. At any point $x \in \mathcal{M}$, the metric $\mathfrak{g}$ induces a natural inner product on the tangent space denoted by $\langle \cdot, \cdot \rangle : \mathcal{T}_x\mathcal{M} \times \mathcal{T}_x\mathcal{M} \to \mathbb{R}$. We also consider the Levi-Civita connection ∇ (Absil et al., 2009a, Theorem 5.3.1). The Riemannian curvature tensor is denoted by $R(x)[u, v]$ where $x \in \mathcal{M}$, $u, v \in \mathcal{T}_x\mathcal{M}$ and is defined in terms of the connection ∇ (Absil et al., 2009a, Theorem 5.3.1). The sectional curvature $K(x)[u, v]$ for $x \in \mathcal{M}$ and $u, v \in \mathcal{T}_x\mathcal{M}$ is then defined in Lee (1997, Prop. 8.8).

$$K(x)[u, v] = \frac{\langle R(x)[u, v]u, v\rangle}{\langle u, u\rangle\langle v, v\rangle - \langle u, v\rangle^2}, \ x \in \mathcal{M}, \ u, v \in \mathcal{T}_x\mathcal{M}.$$

Denote the distance (induced by the Riemannian metric) between two points in $\mathcal{M}$ by $d(x, y)$. A geodesic $\gamma : \mathbb{R} \to \mathcal{M}$ is a constant speed curve whose length is equal to $d(x, y)$, so it is the shortest path on manifold linking x and y. $\gamma_{x\to y}$ denotes the geodesic from x to y (thus $\gamma_{x\to y}(0) = x$ and $\gamma_{x\to y}(1) = y$).

[3]Since our results are local, completeness is not necessary and our results can be easily generalized, with extra assumptions on the injectivity radius.

The exponential map $\mathrm{Exp}_x(v)$ maps $v \in \mathcal{T}_x\mathcal{M}$ to $y \in \mathcal{M}$ such that there exists a geodesic γ with $\gamma(0) = x$, $\gamma(1) = y$ and $\frac{d}{dt}\gamma(0) = v$. The injectivity radius at point $x \in \mathcal{M}$ is the maximal radius r for which the exponential map is a diffeomorphism on $\mathbb{B}_x(r) \subset \mathcal{T}_x\mathcal{M}$. The injectivity radius of the manifold, denoted by $\mathfrak{I}$, is the infimum of the injectivity radii at all points. Since the manifold is complete, we have $\mathfrak{I} > 0$. When $x, y \in \mathcal{M}$ satisfies $d(x, y) \leq \mathfrak{I}$, the exponential map admits an inverse $\mathrm{Exp}_x^{-1}(y)$, which satisfies $d(x, y) = \|\mathrm{Exp}_x^{-1}(y)\|$. Parallel translation Γ_x^y denotes a the map which transports $v \in \mathcal{T}_x\mathcal{M}$ to $\Gamma_x^y v \in \mathcal{T}_y\mathcal{M}$ along $\gamma_{x\to y}$ such that the vector stays constant by satisfying a zero-acceleration condition (Lee, 1997, equation (4.13)).

For a smooth function $f : \mathcal{M} \to \mathbb{R}$, $\mathrm{grad} f(x) \in \mathcal{T}_x\mathcal{M}$ denotes the Riemannian gradient of f at $x \in \mathcal{M}$ which satisfies $\frac{d}{dt}f(\gamma(t)) = \langle \gamma'(t), \mathrm{grad} f(x)\rangle$ (see Absil et al., 2009a, Sec 3.5.1 and (3.31)). The Hessian of f is defined jointly with the Riemannian structure of the manifold. The (directional) Hessian is $H(x)[\xi_x] := \nabla_{\xi_x}\mathrm{grad} f$, and we use $H(x)[u, v] := \langle u, H(x)[v]\rangle$ as a shorthand. We call $x \in \mathcal{M}$ an $(\epsilon, -\sqrt{\rho\epsilon})$ saddle point when $\|\nabla f(x)\| \leq \epsilon$ and $\lambda_{\min}(H(x)) \leq -\sqrt{\rho\epsilon}$. We refer the interested reader to Do Carmo (2016) and Lee (1997) which provide a thorough review on these important concepts of Riemannian geometry.

2.3 Perturbed Riemannian gradient algorithm

Our main Algorithm 1 runs as follows:

1. Check the norm of the gradient: If it is large, do one step of Riemannian gradient descent, consequently the function value decreases.
2. If the norm of gradient is small, it's either an approximate saddle point or a local minimum. Perturb the variable by adding an appropriate level of noise in its tangent space, map it back to the manifold and run a few iterations.
   (a) If the function value decreases, iterates are escaping from the approximate saddle point (and the algorithm continues)
   (b) If the function value does not decrease, then it is an approximate local minimum (the algorithm terminates).

Algorithm 1 Perturbed Riemannian gradient algorithm

Require: Initial point $x_0 \in \mathcal{M}$, parameters $\beta, \rho, K, \mathfrak{I}$, accuracy ϵ, probability of success δ (parameters defined in Assumptions 1, 2, 3 and assumption of Theorem 1).
Set constants: $\hat{c} \geq 4$, $C := C(K, \beta, \rho)$ (defined in Lemma 2 and proof of Lemma 8) and $\sqrt{c_{\max}} \leq \frac{1}{56\hat{c}^2}$, $r = \frac{\sqrt{c_{\max}}}{\chi^2}\epsilon$, $\chi = 3\max\{\log(\frac{d\beta(f(x_0)-f^*)}{\hat{c}\epsilon^2\delta}), 4\}$.
Set threshold values: $f_{\text{thres}} = \frac{c_{\max}}{\chi^3}\sqrt{\frac{\epsilon^3}{\rho}}$, $g_{\text{thres}} = \frac{\sqrt{c_{\max}}}{\chi^2}\epsilon$, $t_{\text{thres}} = \frac{\chi}{c_{\max}}\frac{\beta}{\sqrt{\rho\epsilon}}$, $t_{\text{noise}} = -t_{\text{thres}} - 1$.
Set stepsize: $\eta = \frac{c_{\max}}{\beta}$.
while 1 **do**
  if $\|\mathrm{grad} f(x_t)\| \leq g_{\text{thres}}$ and $t - t_{\text{noise}} > t_{\text{thres}}$ **then**
    $t_{\text{noise}} \leftarrow t$, $\tilde{x}_t \leftarrow x_t$, $x_t \leftarrow \mathrm{Exp}_{x_t}(\xi_t)$, ξ_t uniformly sampled from $\mathbb{B}_{x_t}(r) \subset \mathcal{T}_x\mathcal{M}$.
  end if
  if $t - t_{\text{noise}} = t_{\text{thres}}$ and $f(x_t) - f(\tilde{x}_{t_{\text{noise}}}) > -f_{\text{thres}}$ **then**
    output $\tilde{x}_{t_{\text{noise}}}$
  end if
  $x_{t+1} + \leftarrow \mathrm{Exp}_{x_t}(-\min\{\eta, \frac{\mathfrak{I}}{\|\mathrm{grad} f(x_t)\|}\}\mathrm{grad} f(x_t))$.
  $t \leftarrow t + 1$.
end while

Algorithm 1 relies on the manifold's exponential map, and is useful for cases where this map is easy to compute (true for many common manifolds). We refer readers to Lee (1997, pp. 81-86) for the exponential map of sphere and hyperbolic manifolds, and Absil et al. (2009a, Example 5.4.2, 5.4.3) for the Stiefel and Grassmann manifolds. If the exponential map is not computable, the algorithm can use a retraction[4] instead, however our current analysis only covers the case of the exponential map. In Figure 1[5], we illustrate a function with saddle point on sphere, and plot the trajectory of Algorithm 1 when it is initialized at a saddle point.

[4] A retraction is a first-order approximation of the exponential map which is often easier to compute.
[5] Codes for generating figures are available at `http://students.washington.edu/yuesun/code.zip`.

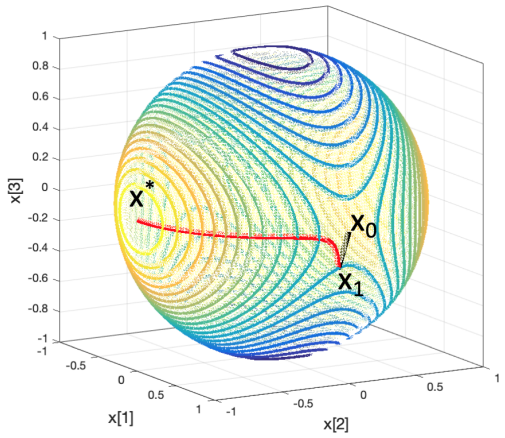


Figure 1: Function f with saddle point on a sphere. $f(x) = x_1^2 - x_2^2 + 4x_3^2$. We plot the contour of this function on unit sphere. Algorithm 1 initializes at $x_0 = [1, 0, 0]$ (a saddle point), perturbs it towards x_1 and runs Riemannian gradient descent, and terminates at $x^* = [0, -1, 0]$ (a local minimum). We amplify the first iteration to make saddle perturbation visible.

2.4 Main theorem: escape rate for perturbed Riemannian gradient descent

We now turn to our main results, beginning with our assumptions and a statement of our main theorem. We then develop a brief proof sketch.

Our main result involves two conditions on function f and one on the curvature of the manifold $\mathcal{M}$.

Assumption 1 (Lipschitz gradient)**.** *There is a finite constant β such that*

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

Assumption 2 (Lipschitz Hessian)**.** *There is a finite constant ρ such that*

$$\|H(y) - \Gamma_x^y H(x) \Gamma_y^x\|_2 \leq \rho d(x, y) \quad \textit{for all } x, y \in \mathcal{M}.$$

Assumption 3 (Bounded sectional curvature)**.** *There is a finite constant K such that*

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

K is an intrinsic parameter of the manifold capturing the curvature. We list a few examples here: (i) A sphere of radius R has a constant sectional curvature $K = 1/R^2$ (Lee, 1997, Theorem 1.9). If the radius is bigger, K is smaller which means the sphere is less curved; (ii) A hyper-bolic space H_R^n of radius R has $K = -1/R^2$ (Lee, 1997, Theorem 1.9); (iii) For sectional curvature of the Stiefel and the Grasmann manifolds, we refer readers to Rapcsák (2008, Section 5) and Wong (1968), respectively.

Note that the constant K is not directly related to the RLICQ parameter R defined by Ge et al. (2015) which first requires describing the manifold by equality constraints. Different representations of the same manifold could lead to different curvature bounds, while sectional curvature is an intrinsic property of manifold. If the manifold is a sphere $\sum_{i=1}^{d+1} x_i^2 = R^2$, then $K = 1/R^2$, but more generally there is no simple connection. The smoothness parameters we assume are natural compared to some quantity from complicated compositions (Lee et al., 2017, Section 5.5) or pullback (Zhang & Zhang, 2018; Criscitiello & Boumal, 2019). With these assumptions, the main result of this work is the following:

Theorem 1. *Under Assumptions 1,2,3, let $C(K, \beta, \rho)$ be a function defined in Lemma 2, $\hat{\rho} = \max\{\rho, C(K, \beta, \rho)\}$, if ϵ satisfies that*

$$\epsilon \leq \min\left\{ \frac{\hat{\rho}}{56 \max\{c_2(K), c_3(K)\}\eta\beta} \log\left(\frac{d\beta}{\sqrt{\hat{\rho}}\epsilon\delta}\right), \left(\frac{\mathfrak{I}\hat{\rho}}{12\hat{c}\sqrt{\eta\beta}} \log\left(\frac{d\beta}{\sqrt{\hat{\rho}}\epsilon\delta}\right)\right)^2 \right\} \tag{2}$$

where $c_2(K)$, $c_3(K)$ are defined in Lemma 4, then with probability $1 - \delta$, perturbed Riemannian gradient descent with step size $c_{\max}/\beta$ converges to a $(\epsilon, -\sqrt{\hat{\rho}}\epsilon)$-stationary point of f in

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

iterations.

Proof roadmap. For a function satisfying smoothness condition (Assumption 1 and 2), we use a local upper bound of the objective based on the third-order Taylor expansion (see supplementary material Section A for a review),

$$f(u) \leq f(x) + \langle \text{grad} f(x), \text{Exp}_x^{-1}(u) \rangle + \frac{1}{2} H(x)[\text{Exp}_x^{-1}(u), \text{Exp}_x^{-1}(u)] + \frac{\rho}{6} \|\text{Exp}_x^{-1}(u)\|^3.$$

When the norm of the gradient is large (not near a saddle), the following lemma guarantees the decrease of the objective function in one iteration.

Lemma 1. *(Boumal et al., 2018) Under Assumption 1, by choosing* $\bar{\eta} = \min\{\eta, \frac{\mathfrak{I}}{\|\text{grad} f(u)\|}\} = O(1/\beta)$*, the Riemannian gradient descent algorithm is monotonically descending,* $f(u^+) \leq f(u) - \frac{1}{2}\bar{\eta}\|\text{grad} f(u)\|^2$.

Thus our main challenge in proving the main theorem is the Riemannian gradient behaviour at an approximate saddle point:

1. Similar to the Euclidean case studied by Jin et al. (2017a), we need to bound the "thickness" of the "stuck region" where the perturbation fails. We still use a pair of hypothetical auxiliary sequences and study the "coupling" sequences. When two perturbations couple in the thinnest direction of the stuck region, their distance grows and one of them escapes from saddle point.

2. However our iterates are evolving on a manifold rather than a Euclidean space, so our strategy is to map the iterates back to an appropriate fixed tangent space where we can use the Euclidean analysis. This is done using the inverse of the exponential map and various parallel transports.

3. Several key challenges arise in doing this. Unlike Jin et al. (2017a), the structure of the manifold interacts with the local approximation of the objective function in a complicated way. On the other hand, unlike recent work on Riemannian optimization by Boumal et al. (2016a), we do not have access to a second order oracle and we need to understand how the sectional curvature and the injectivity radius (which both capture intrinsic manifold properties) affect the behavior of the first order iterates.

4. Our main contribution is to carefully investigate how the various approximation errors arising from (a) the linearization of the iteration couplings and (b) their mappings to a common tangent space can be handled on manifolds with bounded sectional curvature. We address these challenges in a sequence of lemmas (Lemmas 3 through 6) we combine to linearize the coupling iterations in a common tangent space and precisely control the approximation error. This result is formally stated in the following lemma.

Lemma 2. *Define* $\gamma = \sqrt{\rho\epsilon}$, $\kappa = \frac{\beta}{\gamma}$, *and* $\mathscr{S} = \sqrt{\eta\beta}\frac{\gamma}{\rho}\log^{-1}(\frac{d\kappa}{\delta})$. *Let us consider* x *be a* $(\epsilon, -\sqrt{\rho\epsilon})$ *saddle point, and define* $u^+ = \text{Exp}_u(-\eta\text{grad} f(u))$ *and* $w^+ = \text{Exp}_w(-\eta\text{grad} f(w))$. *Under Assumptions 1, 2, 3, if all pairwise distances between* u, w, u^+, w^+, x *are less than* $12\mathscr{S}$*, then for some explicit constant* $C(K, \rho, \beta)$ *depending only on* K, ρ, β*, there is*

$$\begin{aligned} &\|\text{Exp}_x^{-1}(w^+) - \text{Exp}_x^{-1}(u^+) - (I - \eta H(x))(\text{Exp}_x^{-1}(w) - \text{Exp}_x^{-1}(u))\| \\ &\leq C(K, \rho, \beta) d(u, w) \left(d(u, w) + d(u, x) + d(w, x)\right). \end{aligned} \tag{3}$$

The proof of this lemma includes novel contributions by strengthen known result (Lemmas 3) and also combining known inequalities in novel ways (Lemmas 4 to 6) that allow us to control all the approximation errors and arrive at the tight rate of escape for the algorithm.

2.5 Proof of Lemma 2

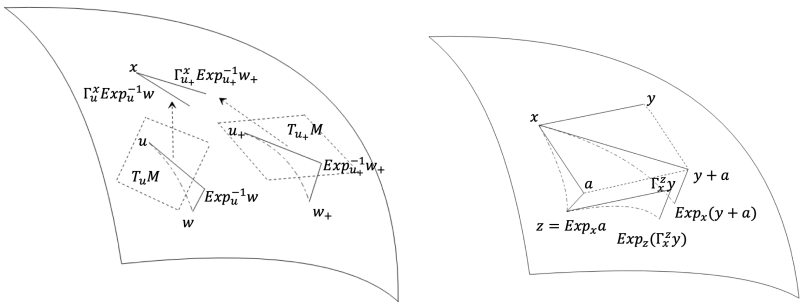
Lemma 2 controls the error of the linear approximation of the iterates when mapped in $T_x\mathcal{M}$. In this section, we assume that all points are within a region of diameter $R := 12\mathscr{S} \leq \mathfrak{I}$ (inequality follows from Eq. (2)), i.e., the distance of any two points in the following lemmas are less than R.

The proof of Lemma 2 is based on the sequence of following lemmas.

Lemma 3. *Let* $x \in \mathcal{M}$ *and* $y, a \in T_x\mathcal{M}$. *Let us denote by* $z = \text{Exp}_x(a)$ *then under Assumption 3*

$$d(\text{Exp}_x(y + a), \text{Exp}_z(\Gamma_x^z y)) \leq c_1(K) \min\{\|a\|, \|y\|\}(\|a\| + \|y\|)^2. \tag{4}$$

This lemma tightens the result of Karcher (1977, C2.3), which only shows an upper-bound $O(\|a\|(\|a\| + \|y\|)^2)$. We prove the upper-bound $O(\|y\|(\|a\| + \|y\|)^2)$ in the supplement. We also need the following lemma showing that both the exponential map and its inverse are Lipschitz.



(a) (b)

Figure 2: (a) Eq. (5). First map w and w_+ to $\mathcal{T}_u\mathcal{M}$ and $\mathcal{T}_{u_+}\mathcal{M}$, and transport the two vectors to $\mathcal{T}_x\mathcal{M}$, and get their relation. (b) Lemma 3 bounds the difference of two steps starting from x: (1) take $y + a$ step in $\mathcal{T}_x\mathcal{M}$ and map it to manifold, and (2) take a step in $\mathcal{T}_x\mathcal{M}$, map to manifold, call it z, and take $\Gamma^z_x y$ step in $\mathcal{T}_x\mathcal{M}$, and map to manifold. $\mathrm{Exp}_z(\Gamma^z_x y)$ is close to $\mathrm{Exp}_x(y + a)$.

Lemma 4. *Let $x, y, z \in M$, and the distance of each two points is no bigger than R. Then under assumption 3*

$$(1 + c_2(K)R^2)^{-1} d(y, z) \leq \|\mathrm{Exp}_x^{-1}(y) - \mathrm{Exp}_x^{-1}(z)\| \leq (1 + c_3(K)R^2) d(y, z).$$

Intuitively this lemma relates the norm of the difference of two vectors of $\mathcal{T}_x\mathcal{M}$ to the distance between the corresponding points on the manifold $\mathcal{M}$ and follows from bounds on the Hessian of the square-distance function (Sakai, 1996, Ex. 4 p. 154). The upper-bound is directly proven by Karcher (1977, Proof of Cor. 1.6), and we prove the lower-bound via Lemma 3 in the supplement.

The following contraction result is fairly classical and is proven using the Rauch comparison theorem from differential geometry (Cheeger & Ebin, 2008).

Lemma 5. *(Mangoubi et al., 2018, Lemma 1) Under Assumption 3, for $x, y \in \mathcal{M}$ and $w \in T_x\mathcal{M}$,*

$$d(\mathrm{Exp}_x(w), \mathrm{Exp}_y(\Gamma^y_x w)) \leq c_4(K) d(x, y).$$

Finally we need the following corollary of the Ambrose-Singer theorem (Ambrose & Singer, 1953).

Lemma 6. *(Karcher, 1977, Section 6) Under Assumption 3, for $x, y, z \in \mathcal{M}$ and $w \in T_x\mathcal{M}$,*

$$\|\Gamma^z_y \Gamma^y_x w - \Gamma^z_x w\| \leq c_5(K) d(x, y) d(y, z) \|w\|.$$

Lemma 3 through 6 are mainly proven in the literature, and we make up the missing part in Supplementary material Section B. Then we prove Lemma 2 in Supplementary material Section B.

The spirit of the proof is to linearize the manifold using the exponential map and its inverse, and to carefully bounds the various error terms caused by the approximation. Let us denote by $\theta = d(u, w) + d(u, x) + d(w, x)$.

1. We first show using twice Lemma 3 and Lemma 5 that

$$d(\mathrm{Exp}_u(\mathrm{Exp}_u^{-1}(w) - \eta \Gamma^u_w \mathrm{grad} f(w)), \mathrm{Exp}_u(-\eta \mathrm{grad} f(u) + \Gamma^u_{u_+} \mathrm{Exp}^{-1}_{u_+}(w_+))) = O(\theta d(u, w)).$$

2. We use Lemma 4 to linearize this iteration in $\mathcal{T}_u\mathcal{M}$ as

$$\|\Gamma^u_{u_+} \mathrm{Exp}^{-1}_{u_+}(w_+) - \mathrm{Exp}^{-1}_u(w) + \eta[\mathrm{grad} f(u) - \Gamma^u_w \mathrm{grad} f(w)]\| = O(\theta d(u, w)).$$

3. Using the Hessian Lipschitzness

$$\|\Gamma^u_{u_+} \mathrm{Exp}^{-1}_{u_+}(w_+)) - \mathrm{Exp}^{-1}_u(w) + \eta H(u) \mathrm{Exp}^{-1}_u(w)\| = O(\theta d(u, w)).$$

3. We use Lemma 6 to map to $T_x\mathcal{M}$ and the Hessian Lipschitzness to compare $H(u)$ to $H(x)$. This is an important intermediate result (see Lemma 1 in Supplementary material Section B).

$$\|\Gamma^x_{u_+} \mathrm{Exp}^{-1}_{u_+}(w_+) - \Gamma^x_u \mathrm{Exp}^{-1}_u(w) + \eta H(x) \Gamma^x_u \mathrm{Exp}^{-1}_u(w)\| = O(\theta d(u, w)). \tag{5}$$

4. We use Lemma 3 and 4 to approximate two iteration updates in $\mathcal{T}_x\mathcal{M}$.

$$\|\mathrm{Exp}_x^{-1}(w) - (\mathrm{Exp}_x^{-1}(u) + \Gamma_u^x \mathrm{Exp}_u^{-1}(w))\| \leq O(\theta d(u,w)). \tag{6}$$

And same for the u_+, w_+ pair replacing u, w.

5. Combining Eq. (5) and Eq. (6) together, we obtain

$$\|\mathrm{Exp}_x^{-1}(w^+) - \mathrm{Exp}_x^{-1}(u^+) - (I - \eta H(x))(\mathrm{Exp}_x^{-1}(w) - \mathrm{Exp}_x^{-1}(u))\| \leq O(\theta d(u,w)).$$

Now note that, the iterations u, u_+, w, w_+ of the algorithm are both on the manifold. We use $\mathrm{Exp}_x^{-1}(\cdot)$ to map them to the same tangent space at x.

Therefore we have linearized the two coupled trajectories $\mathrm{Exp}_x^{-1}(u_t)$ and $\mathrm{Exp}_x^{-1}(w_t)$ in a common tangent space, and we can modify the Euclidean escaping saddle analysis thanks to the error bound we proved in Lemma 2.

2.6 Proof of main theorem

In this section we suppose all assumptions in Section 2.4 hold. The proof strategy is to show with high probability that the function value decreases of $\mathscr{F}$ in $\mathscr{T}$ iterations at an approximate saddle point. Lemma 7 suggests that, if after a perturbation and $\mathscr{T}$ steps, the iterate is $\Omega(\mathscr{S})$ far from the approximate saddle point, then the function value decreases. If the iterates do not move far, the perturbation falls in a stuck region. Lemma 8 uses a coupling strategy, and suggests that the width of the stuck region is small in the negative eigenvector direction of the Riemannian Hessian.

Define

$$\mathscr{F} = \eta\beta\frac{\gamma^3}{\hat{\rho}^2}\log^{-3}(\frac{d\kappa}{\delta}),\ \mathscr{G} = \sqrt{\eta\beta}\frac{\gamma^2}{\hat{\rho}}\log^{-2}(\frac{d\kappa}{\delta}),\ \mathscr{T} = \frac{\log(\frac{d\kappa}{\delta})}{\eta\gamma}.$$

At an approximate saddle point $\tilde{x}$, let y be in the neighborhood of $\tilde{x}$ where $d(y,\tilde{x}) \leq \mathfrak{I}$, denote

$$\tilde{f}_y(x) := f(y) + \langle \mathrm{grad} f(y), \mathrm{Exp}_y^{-1}(\tilde{x})\rangle + \frac{1}{2}H(\tilde{x})[\mathrm{Exp}_y^{-1}(\tilde{x}), \mathrm{Exp}_y^{-1}(\tilde{x})].$$

Let $\|\mathrm{grad} f(\tilde{x})\| \leq \mathscr{G}$ and $\lambda_{\min}(H(\tilde{x})) \leq -\gamma$. We consider two iterate sequences, $u_0, u_1, ...$ and $w_0, w_1, ...$ where u_0, w_0 are two perturbations at $\tilde{x}$.

Lemma 7. *Assume Assumptions 1, 2, 3 and Eq. (2) hold. There exists a constant $c_{\max}$, $\forall \hat{c} > 3, \delta \in (0, \frac{d\kappa}{e}]$, for any u_0 with $d(\tilde{x}, u_0) \leq 2\mathscr{S}/(\kappa\log(\frac{d\kappa}{\delta}))$, $\kappa = \beta/\gamma$.*

$$T = \min\left\{\inf_t\left\{t | \tilde{f}_{u_0}(u_t) - f(u_0) \leq -3\mathscr{F}\right\}, \hat{c}\mathscr{T}\right\},$$

then $\forall \eta \leq c_{\max}/\beta$, we have $\forall 0 < t < T$, $d(u_0, u_t) \leq 3(\hat{c}\mathscr{S})$.

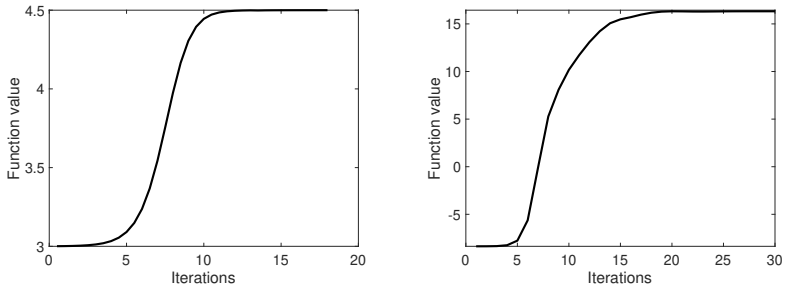
Lemma 8. *Assume Assumptions 1, 2, 3 and Eq. (2) hold. Take two points u_0 and w_0 which are perturbed from an approximate saddle point, where $d(\tilde{x}, u_0) \leq 2\mathscr{S}/(\kappa\log(\frac{d\kappa}{\delta}))$, $\mathrm{Exp}_{\tilde{x}}^{-1}(w_0) - \mathrm{Exp}_{\tilde{x}}^{-1}(u_0) = \mu r e_1$, e_1 is the smallest eigenvector[6] of $H(\tilde{x})$, $\mu \in [\delta/(2\sqrt{d}), 1]$, and the algorithm runs two sequences $\{u_t\}$ and $\{w_t\}$ starting from u_0 and w_0. Denote*

$$T = \min\left\{\inf_t\left\{t | \tilde{f}_{w_0}(w_t) - f(w_0) \leq -3\mathscr{F}\right\}, \hat{c}\mathscr{T}\right\},$$

then $\forall \eta \leq c_{\max}/l$, if $\forall 0 < t < T$, $d(\tilde{x}, u_t) \leq 3(\hat{c}\mathscr{S})$, we have $T < \hat{c}\mathscr{T}$.

We prove Lemma 7 and 8 in supplementary material Section C. We also prove, in the same section, the main theorem using the coupling strategy of Jin et al. (2017a). but with the additional difficulty of taking into consideration the effect of the Riemannian geometry (Lemma 2) and the injectivity radius.

[6] "smallest eigenvector" means the eigenvector corresponding to the smallest eigenvalue.



(a) (b)

Figure 3: (a) kPCA problem. We start from an approximate saddle point, and it converges to a local minimum (which is also global minimum). (b) Burer-Monteiro approach Plot $f(Y) = \frac{1}{2}\text{trace}(AYY^T)$ versus iterations. We start from the saddle point, and it converges to a local minimum (which is also global minimum).

2.7 Examples

kPCA. We consider the kPCA problem, where we want to find the $k \leq n$ principal eigenvectors of a symmetric matrix $H \in \mathbb{R}^{n\times n}$, as an example (Tripuraneni et al., 2018). This corresponds to

$$\min_{X\in\mathbb{R}^{n\times k}} \; -\frac{1}{2}\text{tr}(X^THX) \quad \text{subject to } X^TX = I,$$

which is an optimization problem on the Grassmann manifold defined by the constraint $X^TX = I$. If the eigenvalues of H are distinct, we denote by v_1,...,v_n the eigenvectors of H, corresponding to eigenvalues with decreasing order. Let $V^* = [v_1, ..., v_k]$ be the matrix with columns composed of the top k eigenvectors of H, then the local minimizers of the objective function are V^*G for all unitary matrices $G \in \mathbb{R}^{k\times k}$. Denote also by $V = [v_{i_1}, ..., v_{i_k}]$ the matrix with columns composed of k distinct eigenvectors, then the first order stationary points of the objective function (with Riemannian gradient being 0) are VG for all unitary matrices $G \in \mathbb{R}^{k\times k}$. In our numerical experiment, we choose H to be a diagonal matrix $H = \text{diag}(0, 1, 2, 3, 4)$ and let $k = 3$. The Euclidean basis (e_i) are an eigenbasis of H and the first order stationary points of the objective function are $[e_{i_1}, e_{i_2}, e_{i_3}]G$ with distinct basis and G being unitary. The local minimizers are $[e_3, e_4, e_5]G$. We start the iteration at $X_0 = [e_2, e_3, e_4]$ and see in Fig. 3 the algorithm converges to a local minimum.

Burer-Monteiro approach for certain low rank problems. Following Boumal et al. (2016b), we consider, for $A \in \mathbb{S}^{d\times d}$ and $r(r+1)/2 \leq d$, the problem

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

We factorize X by YY^T with an overparametrized $Y \in \mathbb{R}^{d\times p}$ and $p(p+1)/2 \geq d$. Then any local minimum of

$$\min_{Y\in\mathbb{R}^{d\times p}} \mathbf{tr}(AYY^T), \; s.t. \; \text{diag}(YY^T) = 1,$$

is a global minimum where $YY^T = X^*$ (Boumal et al., 2016b). Let $f(Y) = \frac{1}{2}\mathbf{tr}(AYY^T)$. In the experiment, we take $A \in \mathbb{R}^{100\times 20}$ being a sparse matrix that only the upper left 5×5 block is random and other entries are 0. Let the initial point $Y_0 \in \mathbb{R}^{100\times 20}$, such that $(Y_0)_{i,j} = 1$ for $5j - 4 \leq i \leq 5j$ and $(Y_0)_{i,j} = 0$ otherwise. Then Y_0 is a saddle point. We see in Fig. 3 the algorithm converges to the global optimum.

2.8 Summary

Previous works have shown that in Euclidean space, although the gradient descent can converge to an approximate second order minimum in exponential time, by simply adding a random perturbation at the stationary points, the gradient descent iteration escapes from saddle points and converges to an approximate second order minimum with provable polynomial rate. However, they require the problem being unconstrained, which does not allow a smooth manifold constraint, or the optimization problem set up in Riemannian manifolds. No result was given about the second order convergence of perturbed first order optimization methods on Riemannian manifolds, and it is unknown

how the curvature constant of the manifold contributes to the rate of escaping from saddle points. We have shown that for the constrained optimization problem of minimizing $f(x)$ subject to a manifold constraint as long as the function and the manifold are appropriately smooth, a perturbed Riemannian gradient descent algorithm will escape saddle points with a rate of order $1/\epsilon^2$ in the accuracy ϵ, polylog in manifold dimension d, and depends polynomially on the curvature and smoothness parameters.

A natural extension of our result is to consider other variants of gradient descent, such as the heavy ball method, Nesterov's acceleration, and the stochastic setting. The question is whether these algorithms with appropriate modification (with manifold constraints) would have a fast convergence to second-order stationary point (not just first-order stationary as studied in recent literature), and whether it is possible to show the relationship between convergence rate and smoothness of manifold.

3 Learning linear dynamical systems via nuclear norm regularization

In this section, we investigate the regularzation method for learning low-order linear dynamical systems from input-output data, named as system identification problem. It is known that with appropriate regularizers, the prior information of the structure learning model can be exploited. We show that, the Hankel nuclear norm regularizer reflects the low-order structure of the linear dynamical system, and the regularized least squares solver ensures small estimation error with less training data.

This work is published as Sun et al. (2020).

3.1 Introduction: linear system identification from input-output data via regularized least squares

System identification is an important topic in control theory. An accurate estimation of system dynamics is the basis of the associated control or policy decision problems in tasks varying from linear-quadratic control to deep reinforcement learning. Consider a linear time-invariant system of order R with the state-space representation

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t, \\ y_t &= Cx_t + Du_t + z_t, \end{aligned} \tag{7}$$

where $x_t \in \mathbb{R}^R$ is the state, $u_t \in \mathbb{R}^p$ is the input, $y_t \in \mathbb{R}^m$ is the output, $z_t \in \mathbb{R}^m$ is the output noise, $A \in \mathbb{R}^{R\times R}$, $B \in \mathbb{R}^{R\times p}$, $C \in \mathbb{R}^{m\times R}$, $D \in \mathbb{R}^{m\times p}$ are the system parameters, and x_0 is the initial state (we assume $x_0 = 0$). The system identification problem is finding the system parameters, given input and output observations. When $C = I$, we directly observe the state, otherwise we may obtain only partial state information. A notable line of work derives statistical bounds for system identification with limited *state* observations from a single output trajectory with a random input (Abbasi-Yadkori & Szepesvári, 2011; Simchowitz et al., 2018; Sarkar & Rakhlin, 2019). Simchowitz et al. (2018); Sarkar & Rakhlin (2019) assume that the state evolve as $x_{t+1} = Ax_t + \eta_t$ where η_t is the white noise that provides exploration of states. They observe all states and recover A by least squares operator. The main proof approach comes from (Abbasi-Yadkori et al., 2011, Thm 2,3). To contrast the two papers, Simchowitz et al. (2018) assumes the system being stable whereas Sarkar & Rakhlin (2019) does not require assumptions on the spectral radius of A.

For the hidden-state system in (7), the impulse response sequence $h_0 = D$, $h_t = CA^{t-1}B \in \mathbb{R}^{m\times p}$ for $t = 1, 2, \ldots$ (also known as the Markov parameters) uniquely identifies the end-to-end behavior of the system. The impulse response of the system has infinite length, and we let $h = [D, CB, CAB, CA^2B, \ldots, CA^{2n-3}B]^\top$ denote its first $2n-1$ entries. We also define the Hankel map $\mathcal{H} : \mathbb{R}^{m\times(2n-1)p} \to \mathbb{R}^{mn\times pn}$ as

$$H := \mathcal{H}(h) = \begin{bmatrix} h_1 & h_2 & \ldots & h_n \\ h_2 & h_3 & \ldots & h_{n+1} \\ \ldots & & & \\ h_n & h_{n+1} & \ldots & h_{2n-1} \end{bmatrix}. \tag{8}$$

If R is the system order and $n \geq R$, the Hankel matrix H is of rank R regardless of n. A practically relevant scenario is when the order R is not known in advance or may be misspecified. Specifically, we will assume that R is small, and explore the use of nuclear norm regularization to find a low-rank Hankel matrix. The notion of simplicity of a system by low-order condition (i.e., low-rank Hankel matrix) is assumed in a wide range of applications, including signal recovery of sum of complex exponentials (Cai et al., 2016; Xu et al., 2018) shape from moments estimation in tomography and geophysical inversion (Elad et al., 2004), video inpainting (Ding et al., 2007), etc.

The traditional unregularized methods include Cadzow approach (Cadzow, 1988; Gillard, 2010), matrix pencil method (Sarkar & Pereira, 1995), Ho-Kalman approach (Ho & Kálmán, 1966) and the subspace method raised in Ljung (1999); Van Overschee & De Moor (1995, 2012), further modified as frequency domain subspace method in McKelvey et al. (1996) when the inputs are single frequency signals. After obtaining an (noisy) estimate of impulse response, the algorithms reduce the rank of Hankel matrix or the order of the system impulse response. Cadzow (1988) uses alternative projections and SVD to get a low rank Hankel; Sarkar & Pereira (1995) recovers the subspace of the Hankel matrix by columns of exponent of complex numbers (the columns of Vandermonde decomposition matrix) and the order is the dimension of the subspace; Ho & Kálmán (1966) recovers system parameters A, B, C, D from a low rank approximation of Hankel matrix estimation, with the size of A, B, C, D corresponding to the system order; Van Overschee & De Moor (2012) rewrites the dynamics as the relation of input, output and state, and leverages the subspaces spanned by them,followed by recovering the system-order sized observability matrix. Recent works show that least-squares can be used to recover the Markov parameters. To identify a stable system from a

single trajectory, Oymak & Ozay (2018) estimates the Markov parameter matrix h and Sarkar et al. (2019) estimates the Hankel matrix via least-squares. The latter provides optimal Hankel spectral norm error rates, however has suboptimal sample complexity (see the table in Section 3.2). While Oymak & Ozay (2018); Sarkar et al. (2019) use random input, (Tu et al., 2017, Thm 1.1, 1.2) use impulse and single frequency signal respectively as input, both recovering system Markov parameters. These works assume (roughly) known system order, or traverse the Hankel size n to fit the system order.

There are several interesting generalizations of least squares with non-asymptotic guarantees for different goals. Hazan et al. (2018) and Simchowitz et al. (2019) introduce filtering strategies on top of least squares. The filters in Hazan et al. (2018) is the top eigenvectors of a deterministic matrix, used for output prediction in stable systems. Simchowitz et al. (2019) uses filters in frequency domain to recover the system parameters of a stable system, Tsiamis & Pappas (2019) gives a non-asymptotic analysis for learning a Kalman filter system, which can also be applied to an auto-regressive setting. As an extension, Dean et al. (2019) and Mania et al. (2019) applies system identification guarantee for further robust control, and Agarwal et al. (2019) do online control and regret analysis in adversarial setting, whose algorithm directly learns the policy end-to-end.

Nuclear norm regularization has been shown to recover an unstructured low-rank matrix in a sample-efficient way in many settings (e.g., Recht et al. (2010); Candes & Plan (2010)). The regularized methods that directly modify from subspace method are Hansson et al. (2012); Verhaegen & Hansson (2016) and Smith (2014) (frequency space), whose algorithms run nuclear norm regularization on top of it. Liu et al. (2013); Fazel et al. (2013) propose a slightly different algorithms which regress low rank matrix of output Hankel, both adding a Hankel nuclear norm regularization. Grossmann et al. (2009) specifies the regime when not all output data is collected, and runs Hankel nuclear norm regularization. Ayazoglu & Sznaier (2012) proposes a fast algorithm on solving the regularization algorithm. All above regularization works emphasize on optimization algorithms and have no statistical bounds, and more recently Cai et al. (2016) theoretically proves that a low order SISO system from multi-trajectory input-outputs can be recovered by this approach. Blomberg et al. (2015) analyses a more generic regularization problem with less concrete bound for system identification specifically, and they proposed Hankel matrix nuclear norm regularization as an example. Blomberg (2016) gives a thorough analysis on Hankel nuclear norm regularization applied in system identification, including discussion on proper error metrics, role of rank/system order in formulating the problem, implementable algorithm and selection of tuning parameters.

In this work, we study the sample complexity and estimation errors for least-squares and nuclear norm regularized estimators. Oymak & Ozay (2018) and Sarkar et al. (2019) recover the system from single rollout/trajectory of input, whereas our work, Tu et al. (2017) and Cai et al. (2016) require multiple rollouts. To ensure a standardized comparison, we define *sample complexity* to be the number of equations (equality constraints in variables h_t) used in the problem formulation. With this, we explore the following performance metrics of learning the system from a finite set of T measurements.

- **Sample complexity:** The minimum sample size T for recovering system parameters with zero error when the noise $z = 0$. This quantity is lower bounded by the system order (or the MacMillan degree in control theory), which is the smallest number of states in the system's representation. System order can be seen as the "degrees of freedom" of the model we are trying to recover.
- **Impulse Response (IR) Estimation Error:** The Frobenius norm error for the IR $\|\hat{h} - h\|_F$. Knowing the impulse response enables accurate prediction of the system output.
- **Hankel Estimation Error:** The spectral norm error of the Hankel matrix $\|\mathcal{H}(\hat{h} - h)\|$. This performance metric is particularly important for system identification as described below.

The Hankel spectral norm error is a critical quantity to control for several reasons. First, the Hankel spectral error connects to the $\mathcal{H}_\infty$ estimation of the system via classical arguments (Sanchez-Pena & Sznaier, 1998). Secondly, bounding this error allows for robustly finding balanced realizations of the system; for example, the error in reconstructing A, B, C, D via the Ho-Kalman procedure is bounded by the Hankel spectral error. Finally, it is beneficial in model selection, as a small spectral error helps distinguish the true singular values of the system from the spurious ones. Indeed, as illustrated in the experiments, the Hankel singular value gap of the solution of the regularized algorithm is more visible compared to least-squares, which helps in identifying the order of the system with a parameter λ that is easy to tune as explored in section 3.7.

Contributions. Below, we list our contributions, and contrast our results with the existing work.

• **Nuclear norm regularization** (Sec 3.4 and 3.3): For multi-input/single-output (MISO) systems, we establish sample complexity bounds for the nuclear norm regularized system identification problem, showing the required sample size grows as $O(pR\log^2 n)$, which is linear in the system order R. This result build directly on Cai et al.

(2016) which analyzed the recovery a sum-of-exponentials signal using Hankel nuclear norm (which is equivalent to SISO system identification).

Our work also establishes statistically consistent error rates on the IR and Hankel spectral errors (i.e., the estimates to the ground-truth system parameters with growing sample size). This is in contrast to the error bounds of Cai et al. (2016). Our rates are at least as good as least-squares rates; however, they apply in the small sample size regime[7] $T \lesssim pR^2 \log^2 n$[8].

Finally, Sec 3.3 shows that the weighting of the input is necessary for the sample complexity bound to reach logarithmic of n. We argue this by proving that if the input are i.i.d. Gaussian random variables, the least number of output observations to exactly recover the impulse response in the noiseless grows at least $T \gtrsim n^{1/6}$.

• **Least-squares estimator** (Sec 3.5): It is fairly straightforward to show that least-squares estimator for the impulse response h has a guaranteed error bound when $T \gtrsim np$ (c.f. Oymak & Ozay (2018)). However the bound of Oymak & Ozay (2018) is loose when it comes to Hankel spectral error. For multi-input/multi-output (MIMO) systems, we establish the *optimal spectral error bound* on the Hankel matrix. Sarkar et al. (2019) and Tu et al. (2017) also provide similar bounds, however their sample complexities are suboptimal as they require $O(n^2)$ measurements rather than $O(n)$.

• **Relating IR and Hankel errors:** Note that one can upper/lower bound the Hankel error in terms of IR error using the fact that rows of the Hankel matrix are subsets of the IR sequence. Specifically, we always have the inequality

$$\|\hat{h} - h\|_F/\sqrt{2} \leq \|\mathcal{H}(\hat{h} - h)\| \leq \sqrt{n}\|\hat{h} - h\|_F. \tag{9}$$

Observe that there is a factor of $\sqrt{n}$ difference between the left and right-hand side inequalities. One contribution of this work is that, perhaps surprisingly, we show that the left-hand side inequality is typically the tighter one and we have $\|\hat{h} - h\|_F \sim \|\mathcal{H}(\hat{h} - h)\|$.

• **Experimental performance in the single trajectory case** (Sec 3.7): Finally, we numerically explore the regularized and unregularized algorithms for system identification from single-trajectory data. This is a a case for which theoretical bounds for the regularized estimator as still not known, but we numerically explore the performance.

Our synthetic and real-data experiments (on a low-order example from the DaiSy (De Moor et al., 1997) datasets) suggest that the regularized algorithm has empirical benefits in sample complexity, error, and Hankel spectral gap, and demonstrate that the regularized algorithm is less sensitive to the choice of the tuning parameter than the least squares algorithm is to the Hankel matrix size. Another experiment compares the two least-squares approaches in Oymak & Ozay (2018) and Sarkar et al. (2019), showing that the former (which estimates the impulse response) performs substantially better than the latter (which estimates the Hankel matrix). This highlights the role of proper parameterization in system identification.

3.2 Problem setup and algorithms

Let $\|\cdot\|, \|\cdot\|_*, \|\cdot\|_F$ denote the spectral norm, nuclear norm and Frobenius norm respectively. Throughout, we work with the first $2n-1$ terms of the impulse response denoted by h. The system is excited by input u in the time interval $[0, t]$ and output y is measured at time t, i.e.,

$$y_t = \sum_{i=1}^{t} h_{t+1-i} u_i + z_t. \tag{10}$$

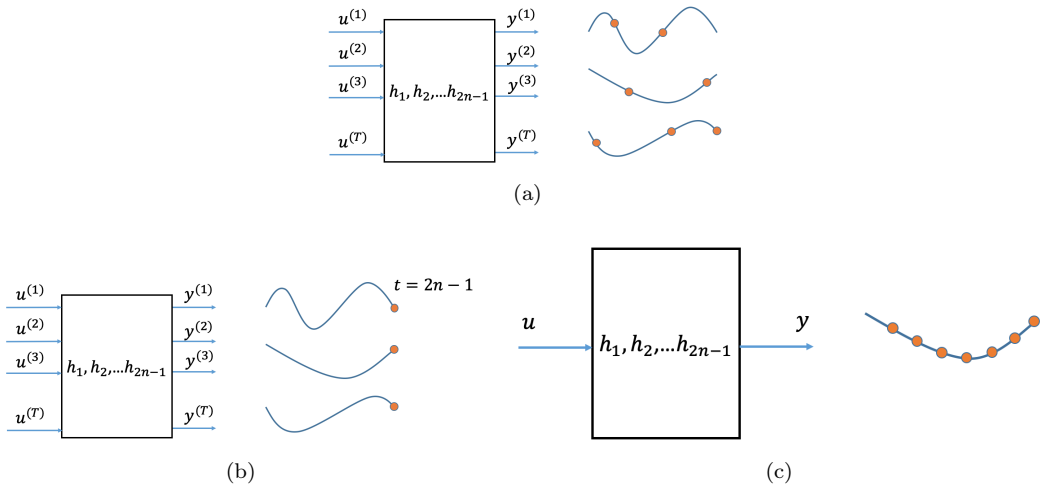
We start by describing data acquisition models. Generally there are several rounds of inputs sent into the system, and the output can be collected or thrown away at arbitrary time. In the setting that we refer to as "multi-rollout" (Figure 4(b)), for each input signal $u^{(i)}$ we take only one output measurement y_t at time $t = 2n-1$ and then the system is restarted with a new input (for example, in a chemical system experiment, or more generally in cases where measuring the output is expensive). Here the *sample complexity* is T, the number of inputs. Recent papers (e.g., Oymak & Ozay (2018) and Sarkar et al. (2019)) use the "single rollout" model (Figure 4(c)) where we apply an input signal from time 1 to $T + 2n - 1$ without restart, and collect all output from time $2n-1$; we use this model in the numerical experiments in section 3.7. We consider two estimators: the *nuclear norm regularized estimator* and the *least square estimator*. The nuclear norm regularized estimator is

$$\hat{h} = \arg\min_{h'} \quad \frac{1}{2}\|\bar{\boldsymbol{U}} h' - y\|_F^2 + \lambda \|\mathcal{H}(h')\|_*, \tag{11}$$

[7] $a \gtrsim b$ and $a \lesssim b$ stand for "there exist a constant c (that does not depend on other parameters) such that $a \geq cb$ or $a \leq cb$".

[8] We also get slightly weaker results in the regime $pR^2 \log^2 n \gtrsim T \gtrsim pR \log^2 n$.

Figure 4: (a) Arbitrary sampling on output data, and two specific data aqcuisition models: (b) Multi-rollout (left), and (c) single rollout (right).



which reduces to the (unregularized) least-squares estimator when $\leftarrow = 0$.

We would like to bound the various error metrics mentioned earlier in terms of the true system order R, the dimension of impulse response $n \gg R$, and signal to noise ratio (SNR) defined as $\mathbf{snr} = \mathbb{E}[\|u\|^2/n]/\mathbb{E}[\|z\|^2]$. The following table provides a summary and comparison of these bounds. In the table, the Hankel matrix is $n \times n$, the system order is R, and the number of samples is T, and $\sigma = 1/\sqrt{\mathbf{snr}}$ denotes the noise level. LS-IR and LS-Hankel stands for least square regression on the impulse response and on the Hankel matrix. All bounds are order-wise and hide log factors.

Paper	This work	This work	Oymak & Ozay (2018)	Sarkar et al. (2019)
Sample complexity	R	n	n	n^2
Method	Nuc-norm	LS-IR	LS-IR	LS-Hankel
IR error	see (16)	$\sigma\sqrt{n/T}$	$\sigma\sqrt{n/T}$	$(1+\sigma)\sqrt{n/T}$
Hankel spectral error	see (16)	$\sigma\sqrt{n/T}$	$\sigma n/\sqrt{T}$	$(1+\sigma)\sqrt{n/T}$

We consider a multiple rollout setup where we measure the system dynamics with T separate rollouts. For each rollout, we drive the system with an input sequence $u^{(i)} \in \mathbb{R}^{(2n-1)p}$ and measure the system output at time $2n-1$. Note that the output at time $2n-1$ is simply $h^\top u$. Define $\bar{\boldsymbol{U}} \in \mathbb{R}^{T \times (2n-1)p}$ where each row is a rollout of inputs, and let $y \in \mathbb{R}^{T \times m}$ denote the corresponding observed outputs. We consider the nuclear norm regularized problem (11). Note that the $\mathcal{H}$ operator does not preserve the Euclidean norm, so Cai et al. (2016) proposes using a normalized operator $\mathcal{G}$, where they first define the weights

$$K_j = \begin{cases} \sqrt{j}, & 1 \le j \le n \\ \sqrt{2n-j}, & n < j \le 2n-1 \end{cases}$$

and let $K = \mathrm{diag} K_j \mathbf{I}_{p\times p} \in \mathbb{R}^{(2n-1)p \times (2n-1)p}$, and define the mapping $\mathcal{G}(h) = \mathcal{H}(K^{-1}h)$. In other words, if $\beta = Kh$ then $\mathcal{G}(\beta) = \mathcal{H}(h)$. Define $\mathcal{G}^* : \mathbb{R}^{mn \times np} \to \mathbb{R}^{m \times (2n-1)p}$ as the adjoint of $\mathcal{G}$, where $[\mathcal{G}^*(M)]_i = \sum_{j+k-1=i} M_{(j)(k)}/K_i$ if we denote the j,k-th block of M (defined in (8)) by $M_{(j)(k)}$. Using this change of variable and letting $\boldsymbol{U} = \bar{\boldsymbol{U}} K^{-1}$, problem (11) can be written as

$$\hat{\beta} = \arg\min_{\beta'} \quad \frac{1}{2}\|\boldsymbol{U}\beta' - y\|_F^2 + \lambda \|\mathcal{G}(\beta')\|_*. \tag{12}$$

3.3 IID inputs and the importance of input shape

In the previous section, we applied a multi-rollout scaling K to the input, so that we can work with the weighted Hankel operator $\mathcal{G}$. A question is that: if the input is i.i.d. without scaling, what will it be? When the input is i.i.d.,

we consider the problem

$$\begin{aligned} \min_{h'} \quad & \|\mathcal{H}(h')\|_* \\ \text{s.t.,} \quad & \|\boldsymbol{U}h' - y\|_2 \leq \delta \end{aligned} \tag{13}$$

where $\mathcal{H}$ is the Hankel operator. More generally, can we modify the proof to a single rollout setting (Oymak & Ozay (2018))? In this setting, the input and output matrices are

$$\boldsymbol{U} = \begin{bmatrix} u_{2n-1}^T & u_{2n-2}^T & ... & u_1^T \\ u_{2n}^T & u_{2n-1}^T & ... & u_2^T \\ ... & & & \\ u_{2n+T-2}^T & u_{2n+T-3}^T & ... & u_T^T \end{bmatrix} \quad \text{and} \quad y = [y_{2n-1}, ..., y_{2n+T-2}] \tag{14}$$

where the energy of the input is fixed at all time and the rows of the input matrices are correlated. The main challenge is the control of restricted condition number induced by the Hankel structure. This is significantly more difficult than controlling the regular condition number (which is sufficient for least-squares proof). However, assuming one can bound the restricted isometry of an arbitrary input matrix, we have the following theorem.

Theorem 2. *We solve the problem (13) where U is arbitrary input matrix and $y = Uh + z$ where z is a noise vector. Let $\Phi = \mathcal{I}(h) \cap \mathbb{S}$ and solve (13). Assuming*

$$\min_{z \in \Phi} \|\boldsymbol{U}z\|_2 \leq \epsilon$$

holds, with probability at most

$$p_T = \exp(-O(\sqrt{T} - f(w(\Phi), \epsilon))), \tag{15}$$

for any $T \geq f(w(\Phi), \epsilon)$, one has $\|\hat{h} - h\|_F \leq O(\frac{\sigma_\xi w(\Phi)}{\epsilon\sqrt{T}})$ with probability $1 - p_T$.

A Gaussian width term $w(\Phi)$ exists in the bound, which can be computed in the previous section when the input matrix is Gaussian scaled by a matrix K, but not generally easy for arbitrary input matrices.

I.i.d. input (without shaping matrix K) is typically used for the recovery of impulse response, and Oymak & Ozay (2018) proves its optimality in terms of recovery error. We ask that, when a system is low order, is i.i.d. input without shaping in terms of sample complexity, i.e., the least number of output observations for the recovery of impulse response?

In the following theorem, we prove that, for a special case, the Gaussian width with unweighted input is polynomial in n, compared to $O(\log n)$ in weighted setting. Since the Gaussian width bound is tight with respect to sample complexity for high probability recovery (McCoy & Tropp, 2013, Thm 1), Theorem 3 indicates that the sample complexity in i.i.d. regime is larger than weighted regime.

Theorem 3. *Suppose the system impulse response is h such that $h_t = 1, \ \forall t \geq 1$, which is order 1. The Gaussian width of the set*

$$\{x \mid \|\mathcal{H}(h + x)\|_* \leq \|\mathcal{H}(h)\|_*\} \cap \mathbb{S}$$

is lower bounded by $Cn^{1/6}$ for some constant C.

Thus in the noiseless setting, the sample complexity is $T \gtrsim n^{1/6}$ to ensure an exact recovery of impulse response, which is bigger than $\log n$ dependence with input shape. The reason is that, in an $n \times n$ Hankel matrix, the entries $H_{1,1}$ and $H_{n,n}$ appears once, and $H(1, n)$ has n copies in the matrix, so that the shape of the input reverses the imbalance of the appearance of entries in the Hankel matrix. This result is rather counter-intuitive since i.i.d. inputs are often optimal for structured parameter estimation tasks (e.g. compressed sensing). Our result shows the provable benefit of input shaping.

3.4 Hankel nuclear norm regularization

To promote a low-rank Hankel matrix, we add nuclear norm regularization in our objective and solve the regularized regression problem. Here we give a finite sample analysis for the recovery of the Hankel matrix and the impulse response found via this approach. We consider a random input matrix $\bar{\boldsymbol{U}}$ and observe the corresponding noisy output vector y as in (10). We then regress y and $\bar{\boldsymbol{U}}$ such that $y = \bar{\boldsymbol{U}}h + z$ where z is the noise vector.

Theorem 4. *Consider the problem (11) in the MISO (multi-input single-output) setting (m=1, p inputs). Suppose the system is order R, $\bar{\boldsymbol{U}} \in \mathbb{R}^{T\times(2n-1)p}$, each row consists of an input rollout $u^{(i)} \in \mathbb{R}^{(2n-1)p}$, and the scaled $\boldsymbol{U} = \bar{\boldsymbol{U}}K^{-1}$ has i.i.d Gaussian entries. Let $\boldsymbol{snr} = \mathbb{E}[\|u\|^2/n]/\mathbb{E}[\|z\|^2]$ and $\sigma = 1/\sqrt{\boldsymbol{snr}}$. Let $\lambda = \sigma\sqrt{\frac{pn}{T}}\log(n)$. Then, the problem (11) returns $\hat{h}$ such that*

$$\frac{\|\hat{h}-h\|_2}{\sqrt{2}} \le \|\mathcal{H}(\hat{h}-h)\| \lesssim \begin{cases} \sqrt{\frac{np}{\boldsymbol{snr}\times T}}\log(n) & \text{if} \quad T \gtrsim \min(R^2, n) \\ \sqrt{\frac{Rnp}{\boldsymbol{snr}\times T}}\log(n) & \text{if} \quad R \lesssim T \lesssim \min(R^2, n). \end{cases} \tag{16}$$

Theorem 4 jointly bounds the impulse response and Hankel spectral errors of the system under mild conditions. We highlight the improvements that our bounds provide: (1) When the system is low order, the sample complexity T is logarithmic in n and improves upon the $O(n)$ bound of the least-squares algorithm. (2) The error rate with respect to the system parameters n, R, T is same as Oymak & Ozay (2018), Sarkar et al. (2019) and Tu et al. (2017) (e.g. compare to Theorem 17).

The regularized method also has the intrinsic advantage that it does not require knowledge of the rank or the singular values of the Hankel matrix beforehand. Numerical experiments on real data in Section 3.7 demonstrate the performance and robustness of the regularized method.

The theorem above follows by combining statistical analysis with a more general deterministic result (Theorem 5). We will state this result in terms of a restricted singular value (RSV) condition. While RSV is a common condition in sparse estimation literature, our analysis requires introducing a spectral norm variation of RSV. Given a matrix M spectral RSV over a set S is defined as follows:

$$\|M\|_S = \max_{v\in S, v\neq 0} \|\mathcal{G}(Mv)\|/\|\mathcal{G}(v)\|.$$

Theorem 5. *Consider the problem (76) in the MISO setting, where $\boldsymbol{U} \in \mathbb{R}^{T\times(2n-1)p}$. Let β denote the (weighted) impulse response of the true system which has order R, i.e., $\mathrm{rank}(\mathcal{G}(\beta)) = R$, and let $y = \boldsymbol{U}\beta + \xi$ be the measured output, where ξ is the measurement noise. Finally, denote the minimizer of (76) by $\hat{\beta}$. Define*

$$\mathcal{J}(\beta) := \left\{ v \;\middle|\; \langle v, \partial(\frac{1}{2}\|\boldsymbol{U}\beta - y\|_2^2 + \lambda\|\mathcal{G}(\beta)\|_*)\rangle \le 0 \right\}, \quad \Gamma := \|I - \boldsymbol{U}^\top\boldsymbol{U}\|_{\mathcal{J}(\beta)},$$

where $\mathcal{J}(\beta)$ is the normal cone at β, and Γ is the spectral RSV. If $\Gamma < 1$, $\hat{\beta}$ satisfies

$$\|\mathcal{G}(\hat{\beta}-\beta)\| \le \frac{\|\mathcal{G}(\boldsymbol{U}^\top\xi)\| + \lambda}{1-\Gamma}.$$

This theorem determines the generic conditions on the measurements $\boldsymbol{U}$ to ensure successful system identification. As future work, it would be desirable to extend our results to a wider range of measurement models.

3.5 Least-squares bounds

Next we consider the least-squares estimator given measurements $y = \bar{\boldsymbol{U}}h + z$. We consider the MIMO setup where $y \in \mathbb{R}^{T\times m}$ and $h \in \mathbb{R}^{(2n-1)p\times m}$. This is obtained by setting $\leftarrow= 0$ in (11) hence the estimator is given via the pseudo-inverse

$$\hat{h} := h + \bar{\boldsymbol{U}}^\dagger z = \min_{h'} \ \frac{1}{2}\|\bar{\boldsymbol{U}}h' - y\|_F^2. \tag{17}$$

The following theorem characterizes the spectral norm bound in terms of discrete Fourier transform.

Theorem 6. *Denote the discrete Fourier transform matrix by F. Denote $z_{(i)} \in \mathbb{R}^T, i = 1, ..., m$ as the noise that corresponds to the i'th coordinate of the output. The solution $\hat{h}$ of (84) obeys*

$$\|\hat{h}-h\|_F \le \|z\|_F/\sigma_{\min}(\bar{\boldsymbol{U}})$$
$$\|\mathcal{H}(\hat{h}-h)\| \le \left\| \left[\|F\bar{\boldsymbol{U}}^\dagger z_{(1)}\|_\infty, ..., \|F\bar{\boldsymbol{U}}^\dagger z_{(m)}\|_\infty\right] \right\|.$$

The next theorem bounds the error when inputs and noise are randomly generated.

Theorem 7. *Denote the solution to (84) as $\hat{h}$. Let $\bar{\boldsymbol{U}} \in \mathbb{R}^{T\times(2n-1)p}$ be input matrix obtained from multiple rollouts, with i.i.d. standard normal entries, $y \in \mathbb{R}^{T\times m}$ be the corresponding outputs and $z \in \mathbb{R}^{T\times m}$ be the noise matrix with i.i.d. $\mathcal{N}(0, \sigma_z^2)$ entries. Then the spectral norm error obeys $\|\mathcal{H}(\hat{h}-h)\| \lesssim \sigma_z\sqrt{\frac{mnp}{T}}\log(np)$.*

This theorem improves the spectral norm bound compared to Oymak & Ozay (2018) which naively bounds the spectral norm in terms of IR error using the right-hand side of (9). Instead, we show that spectral error is same as the IR error up to a log factor (when there is only output noise). Our bound also loses a log factor compared with Sarkar et al. (2019) however is applicable with much fewer samples ($O(n)$ vs $O(n^2)$). We remark that $O(\sigma_z\sqrt{np/T})$ is a tight lower bound for $\|\mathcal{H}(h-\hat{h})\|$ as well as $\|h-\hat{h}\|$ (Oymak & Ozay (2018); Arias-Castro et al. (2012)).

The proofs of the theorems above are provided in Sec B.5. As a proof sketch, we first use the fact that the spectral norm of a circulant matrix is the infinity norm of its Fourier transform. To conclude with Theorem 17, we develop probabilistic bounds on the spectrum of the Hankel error matrix which is circulant.

3.6 Model selection for regularized system identification

In Theorem 4, 5, we declared the recovery error of system impulse response with fixed parameter λ, which depends on the noise level. In practice, we do not know the noise level, thus we try a list of $\lambda \in \Lambda$ and check the validation error to do model selection. Denote the cardinality of Λ as N_λ. In Algorithm 3, we state the training and validation procedure. The guarantee of the recovery error is presented in Theorem 8.

Algorithm 2 System identification and model selection

Require: Input $\bar{\boldsymbol{U}}$ with dimension of each input p, output y with dimension of each output m, dimension of Hankel n, training data size T, list of regularization parameters Λ. $\boldsymbol{U}$ and y satisfy the assumptions in Theorem 4. Parameters a_1, a_2 such that $0 < a_1 < a_2$, c, failure probability P.
for $\lambda_i \in \Lambda$ **do**
  Solve $\hat{h}_i \leftarrow \arg\min_{h'} \ \frac{1}{2}\|\bar{\boldsymbol{U}}h' - y\|_2^2 + \lambda_i\|\mathcal{H}(h')\|_*$. Record $\hat{h}_i$.
end for
Set validation data size $T_{\text{val}} \leftarrow \max\{T, \frac{1}{c}\log\frac{N_\lambda}{P}\}$.
Request validation input $\bar{\boldsymbol{U}}_{\text{val}}$, output y_{val} that generates in the same way as training data.
$\hat{h} \leftarrow \arg\min_{\hat{h}_i} \ \frac{1}{2}\|\bar{\boldsymbol{U}}_{\text{val}}\hat{h}_i - y_{\text{val}}\|_2^2$.
return $\hat{h}$

Theorem 8. *With all assumptions in Theorem 4, suppose the optimal λ that achieves the error bound in 4 is in Λ, with probability at least $1-P$, Algorithm 3 has the error at most $\frac{a_2}{a_1}$ times of the error as in Theorem 4, i.e.,*

$$\frac{\|\hat{h}-h\|_2}{\sqrt{2}} \le \|\mathcal{H}(\hat{h}-h)\| \lesssim \begin{cases} \frac{a_2}{a_1}\sqrt{\frac{np}{\boldsymbol{snr}\times T}}\log(n) & \text{if} \quad T \gtrsim \min(R^2, n) \\ \frac{a_2}{a_1}\sqrt{\frac{Rnp}{\boldsymbol{snr}\times T}}\log(n) & \text{if} \quad R \lesssim T \lesssim \min(R^2, n). \end{cases} \tag{18}$$

Theorem 8 suggests the guarantee for the whole regularization and model selection procedure. We can see that, the amount of validation data is the same as training, which means we do not query essentially more data compared to regularized least squares itself. As a_1, a_2 are constants, Algorithm 3 achieves the same estimation error with the best regularized least squares solution (regardless of constant), so that we do not lose for not knowing the best regularization weight.

3.7 Experiments

3.7.1 Experiments with synthetic data

First we generate synthetic data of noiseless setting and compare the performance of Oymak & Ozay (2018) and Sarkar et al. (2019) in Figure 5 (a,b). We can see that, due to the constant additive error $O(\frac{1}{1-\rho}\sqrt{n/T})$ in Sarkar et al. (2019) algorithm compared to Oymak & Ozay (2018), the resulted error is larger than Oymak & Ozay (2018). Figure 5 (c,d) compares them in the setting when output noise exists and Oymak & Ozay (2018) has smaller error as well.

In this subsection, we check Theorem 4 5 via synthetic experiments, and compare with least square estimator. In the following experiment, we have a fixed strictly stable SISO linear system with order 9, the Hankel size n is initiated as 20 which exceeds the order. The input is multiple rollout, scaled i.i.d Gaussian, which means that we send in the input up to time $2n-1$, and observe the output at the end as an observation, and restart the system. The input satisfies that, after scaling by K^{-1}, $\mathbf{E}(\boldsymbol{U}^T\boldsymbol{U}) = I$, which is the assumption in Theorem 5. The observed

output can be noiseless and noisy, and the numbers of observations are 30 (undetermined for least square) and 60 (determined for least square).

We tune the regularized model by training with different weight λ of regularization, and tune the least square model by changing the size of Hankel matrix and observe the validation error. We pick the model associated with the smallest validation error at the end, and run it on test set. The size of training, validation and test set is $1:3:6$.

Noiseless, enough observations (Figure 6). When the output is noiseless and $T = 60$, we can see that both regularized and least square algorithms do well. When $\lambda \to 0$ in regularization or the size tends to 20×20 in least squares method, it almost perfectly fits the model. The singular values of the estimated Hankel are the same since we have perfect recovery.

Noisy, enough observations (Figure 7). With enough data, when the output is noisy, both regularization and least square recover the impulse response. In Figure 7, we can see that in terms of validation error, there is a best weight λ and a Hankel size n, below and above which the validation error both grow. Then we can pick the optimizer associated with the weight or size as our estimation of the system.

Noiseless, not enough observations (Figure 8). Without enough data for least square, when the output is noiseless, least square is underdetermined, even if we take the solution with the smallest ℓ_2 norm in impulse response, it suffers big error on validation and test set. However, the error of regularization method remains small and as λ getting small, the error still tends to 0. It indicates that, the solution with the least Hankel nuclear norm behaves better than least Frobenius norm solution in low sample complexity case.

Noisy, not enough observations (Figure 9). Finally we discuss the case with insufficient and noisy data. We can see that the regularized algorithm is robust to noise, where as least square algorithm remains bad.

3.7.2 Experiments with DaISy Dataset

Out experiment uses the DaISy dataset De Moor et al. (1997), where a known input signal (not random) is applied and the resulting noisy output trajectory is measured. Consider the input and output matrics as in (14), we solve the optimization problem (11) using single trajectory data. While the input model doesn't satisfy the assumptions of Theorem 5, experiments will demonstrate the advantage of regularization in terms of sample complexity, singular value gap and ease of tuning.

In our experiments, we find that with fixed Hankel size, empirically Oymak & Ozay (2018) performs better than Sarkar et al. (2019), hence we compare our method with the approach of Oymak & Ozay (2018). When necessary, to select the system order in Oymak & Ozay (2018), we simply keep running the estimated system after time $2n+T-1$, compare predictions with the true outputs, and choose the order with the smallest validation error.

With enough data for unregularized version, the algorithms perform well in both cases. The first two figures in Figure 10 show the training and validation error. The tuning parameters are weight λ and Hankel size n for the regularized and unregularized problems respectively. This step is to find the best system order by choosing the tuning parameters with the smallest validation error. The third figure in Figure 10 plots the training and validation sequence from dataset and two algorithms. We see that with sufficient sample size, the system can be recovered well. However, the validation error of regularized algorithm is more flat and λ is easier to tune compared to n.

The first two figures in Figure 11 show that the Hankel spectrums of the two algorithms have a notable difference: The system recovered by the regularized algorithm is low-order and has larger singular value gap. The last two figures in Figure 11 show the advantage of regularization with much better validation performance. As expected from our theory, the difference is most visible in small sample size (this experiment uses 50 training samples). When the number of observations T is small, regularization still returns a solution close to the true system while least-squares cannot recover the system properly.

Finally, we show that the regularization algorithm identifies a stable nonlinear system by a linearized approximation as well. We take the inverted pendulum as the experiment environment. First we use a linearized controller to stabilize it around the equilibrium, and feed in single rollout input, i.i.d. random input of dimension 1. While the dimension of the state is 4, we observe the output of dimension 1, which is the displacement of the system. Then we use the regularized and least square methods to estimate the closed loop system as a linear system and then predict. We use $T = 16$ observations for training, and set $n = 40$, which leads to an underdetermined least-squares problem. Figure 12 shows the singular values and estimated trajectory of these two methods. Despite the nonlinearity of the ground truth system, the regularized algorithm finds a linear model with order 6, while the correct order is not visible in the singular value spectrum of the unregularized least-squares.

3.8 Future directions

This work established new sample complexity and estimation error bounds for system identification. We showed that nuclear norm penalization works well with small sample size regardless of the mis-specification in the problem

(i.e. fitting impulse response with a much larger length rather than the true order). For least-squares we provide the first guarantee that is optimal in sample complexity and the Hankel spectral norm error. These results can be refined in several directions. In the proof of Theorem 5, we use a weighted version of the Hankel operator. We expect that directly computing the Gaussian width of the original Hankel operator will also lead to improvements from least square. It would also be interesting to extend the results to account for single trajectory analysis or process noise. In both cases, an accurate analysis of the regularized problem would lead to new algorithmic insights.

Figure 5: Comparison of (a) Frobenius norm (b) Hankel spectral norm error when output is noiseless, comparison of (c) impulse response Frobenius norm (d) Hankel spectral norm error when output SNR is 10 between Oymak & Ozay (2018) and Sarkar et al. (2019) with synthetic data. System is randomly generated with order 9 and Hankel $H \in \mathbb{R}^{9\times 9}$. Single trajectory and input is i.i.d. Gaussian.

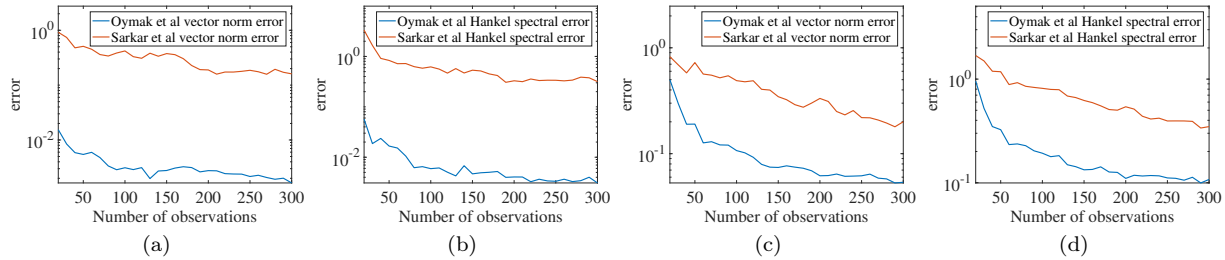


Figure 6: System estimation for synthetic data, noiseless, assuming $n = 20$. Training data size $= 60$. (a) Training and validation error of different λ, (b) Training and validation error of different Hankel size n. Singular value of (c) regularized Hankel (d) unregularized Hankel before rank truncation.

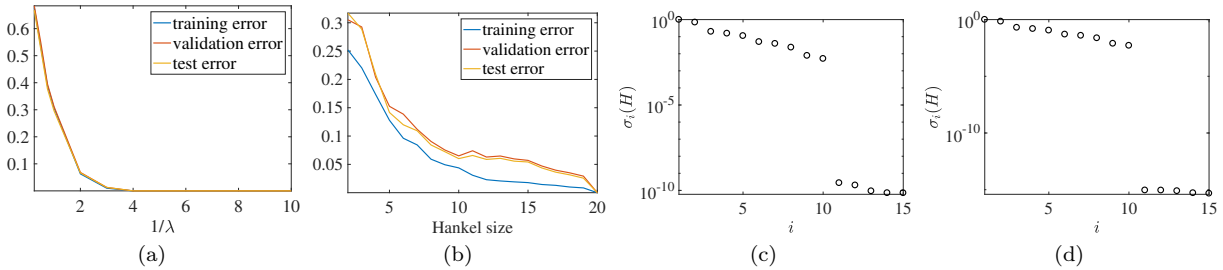


Figure 7: System estimation for synthetic data, SNR $= 10$, assuming $n = 20$. Training data size $= 60$. (a) Training and validation error of different λ, (b) Training and validation error of different Hankel size n. Singular value of (c) regularized Hankel (d) unregularized Hankel before rank truncation.

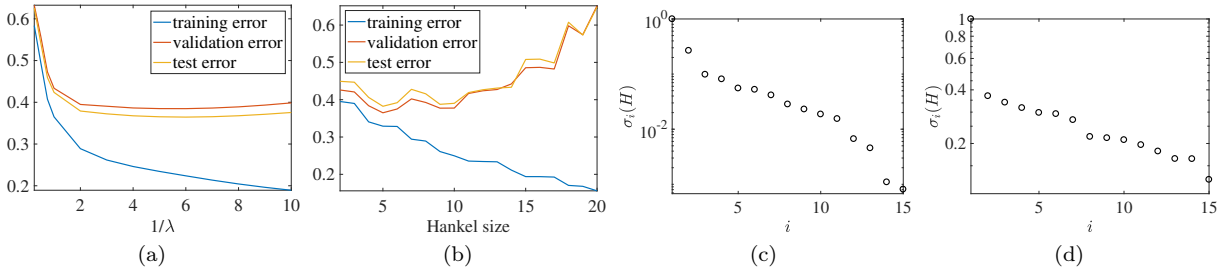


Figure 8: System estimation for synthetic data, noiseless, assuming $n = 20$. Training data size = 30. (a) Training and validation error of different λ, (b) Training and validation error of different Hankel size n. Singular value of (c) regularized Hankel (d) unregularized Hankel before rank truncation.

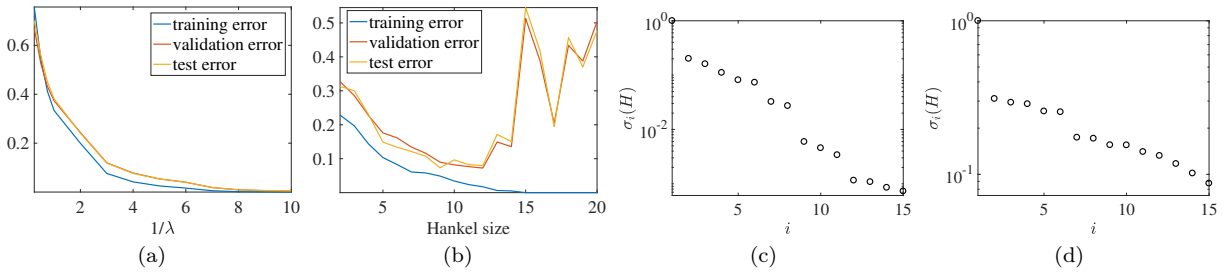


Figure 9: System estimation for synthetic data, SNR = 10, assuming $n = 20$. Training data size = 30. (a) Training and validation error of different λ, (b) Training and validation error of different Hankel size n. Singular value of (c) regularized Hankel (d) unregularized Hankel before rank truncation.

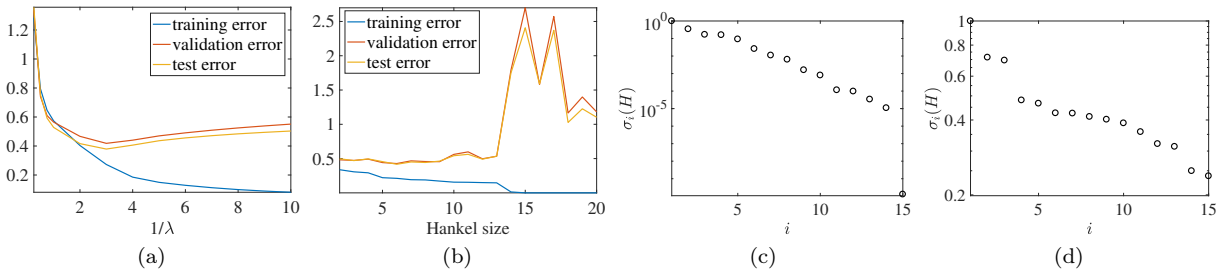


Figure 10: System identification for CD player arm data. Training data size = 200 and validation data size = 600. The first two figures are the training/validation errors of varying λ in regularized algorithm ($n = 10$), and training/validation errors of varying Hankel size n in unregularized algorithm. The last figure is the output trajectory of the true system and the recovered systems (best validation chosen for each).

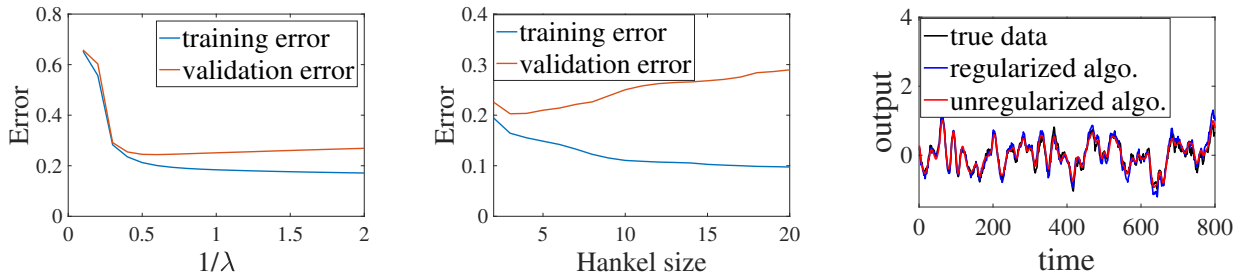


Figure 11: The upper two figures: CD player arm data, singular values of the *regularized* and *unregularized* Hankel. The lower two figures: Recovery by *regularized* and *unregularized* algorithms when Hankel matrix is 10×10. Training size is 50 and validation size is 400.

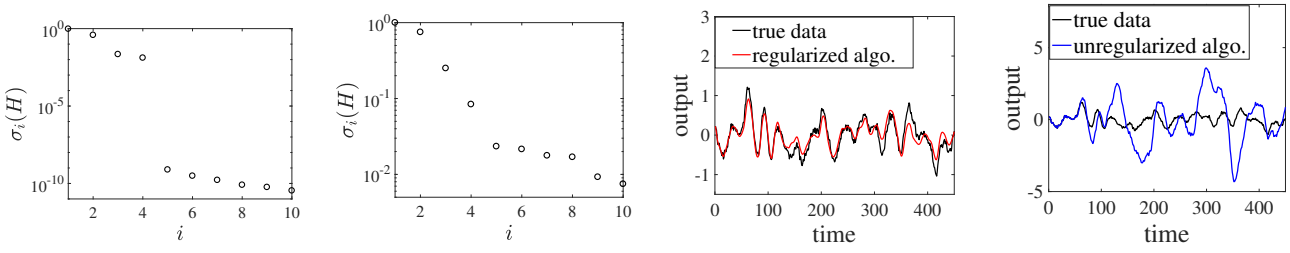
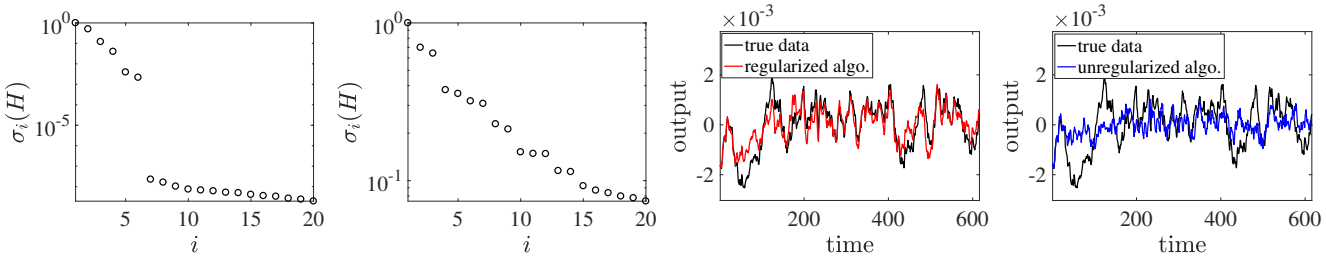


Figure 12: The upper two figures: Stabilized inverted pendulum data, singular values of the *regularized* and *unregularized* Hankel. The lower two figures: Recovery by *regularized* and *unregularized* algorithms when Hankel matrix is 40×40. Training size is 16 and validation size is 600.



4 Future work

This section discusses two ongoing projects. The first work studies the convergence guarantee of a collection of optimal control problems. Although many of them are solved by convexification algorithms, due to the recent application of nonconvex policy iteration methods in machine learning, we are interested in the nonconvex landscape of the set of optimal control problems, and how it connects with the traditional convexification approaches. The second work studies the representation dimension in mixed linear regression and meta-learning. In previous meta-learning works, a low dimensional feature subspace is learned in the training phase, and the dimension reduction step is applied when training on the new task (called meta-test task). This work proposes that, the dimension reduction step is not necessary for getting the optimal generalization guarantee on the meta-test task, where a weighted full dimensional representation also enables low generalization loss. The more detailed introduction and approaches are in the following subsections.

4.1 Novel analysis of convergence of policy gradient descent via convexification

4.1.1 Introduction

This chapter studies the optimization algorithms for a class of optimal control problems. We start from the simple linear time invariant system and state feedback control. The continuous linear system dynamics is (there is possibly a processing noise as well)

$$\dot{x} = Ax + Bu,$$

and the discrete time dynamics is

$$x(t+1) = Ax(t) + Bu(t).$$

The goal is to find the set of inputs u from time zero, that minimizes some loss form that typically depends on state and input. Probably the most famous and well studied optimal control problem is linear quadratic regulator, where we minimize a quadratic loss

$$\int_0^\infty (x(t)^T Q x(t) + u(t)^T R u(t)) dt$$

for continuous time and

$$\sum_{i=0}^{\infty} (x(t)^T Q x(t) + u(t)^T R u(t))$$

for discrete time. When the time horizon is restricted to finite time, especially in discrete case, it's easy to write the problem as convex optimization and directly solve it (Rawlings et al., 2017). For infinite time horizon, it is known that the optimal controller is linear in state, say $u(t) = Kx(t)$ for a constant K (Kalman et al., 1960). This can be extended to LQG when the output can be observed and the state can be recovered by Kalman filter (Astrom, 1971). One can solve the Riccati equation to obtain the optimal controller (Stengel, 1994; Dullerud & Paganini, 2013).

It becomes more difficult when the system is not linear, where the dynamics $Ax + Bu$ is replaced by a function of state and input. The classical solution is dynamic programming, or solving Bellman equations Bellman (1966). Recently, thanks to the development of reinforcement learning theory, this method is revisited a lot, and deep learning enables it to work really well even in the highly complicated systems. But it's still mysterious how those deep learning models work, and recently people research the better known linear systems in the hope of understanding the complex ones.

Usually people do not know the dynamics beforehand when they encounter a dynamical system, and among the optimal control algorithms, there are two major types: **model based methods**, when the system is first identified and then a controller is trained, usually used when we have a good parameterization of the system; or **model free methods**, when the controller is directly trained from the loss without characterizing the dynamics. For LQR, model based methods are largely studied. System identification dates back to Cadzow (1988); Ljung (1999) etc., and recently Simchowitz et al. (2018); Sarkar & Rakhlin (2019) gives sample complexity bounds for state-observed system, and Oymak & Ozay (2018); Sarkar et al. (2019) for output-observed system. Dean et al. (2017, 2019); Mania et al. (2019) describe the procedure of the joint system identification and optimal control based on the estimate of the system. We will give a thorough review of related work in Section 3 which is more related to model based

method. The model free method is proposed by Bradtke et al. (1994), named policy iteration, and more recently reviewed by Kakade (2002); Rajeswaran et al. (2017). It can be proven that, this approach, when applied in LQR, is a nonconvex optimization problem, and it's of interest whether we can still get a convergence guarantee for training. Fazel et al. (2018); Bu et al. (2019) proved the convergence for discrete LQR, respectively by leveraging dynamic programming or Riccati and Lyapunov equations, and obtain *gradient dominance* from the nonconvex loss landscape. Mohammadi et al. (2019) proves similar results for continuous LQR. Zhang et al. (2020) studies the convergence guarantee of gradient descent on $\mathcal{H}_2$ control with $\mathcal{H}_\infty$ constraint, and show that the gradient descent implicitly makes the controller robust.

As suggested by Mohammadi et al. (2019), we have better understanding of the convexification methods (Youla et al., 1976; Kučera, 2011; Zhou et al., 1996) by change of variables. We are interested whether it helps us understanding the easiness of solving the original nonconvex optimal control problems. Unfortunately, although Mohammadi et al. (2019) leverages the convexification as an intermediate step, their proof is still quite specific and does not show an intuitive connection. In this chapter, we will strat from building a bridge between convexification methods and nonconvex policy gradient methods, and generalize the guarantees for more optimal control problems.

4.1.2 Static controller

In this subsection, we start from the optimal static linear controller. We present a novel analysis that connects to the classical convex method, which enables us to extend from the linear quadratic regulator to a more generic set of problems.

4.1.3 Review of convexification method for continuous LQR

Convexification method is widely used in controller design problems. Define a continuous time linear time invariant system

$$\dot{x} = Ax + Bu,\ x(0) = x_0, \tag{19}$$

where x is state and u is input signal, x_0 comes from an initial distribution such that $\mathbf{E}(x_0 x_0^T) \succ 0$, one considers minimizing the LQR loss

$$\min_{u(t)} f(u(t)) := \mathbf{E}_{x_0} \int_0^\infty (x(t)^T Q x(t) + u(t)^T R u(t)) dt \tag{20}$$

where Q, R are positive definite matrices. It is known that, the input signal that minimizes the loss function $f(u)$ is a state feedback controller

$$u = K^* x = -R^{-1} B P x, \tag{21}$$

$$A^T P + PA + Q - PBR^{-1}BP = 0. \tag{22}$$

Note that once we know the state feedback controller is static, we can write loss $f(u(t))$ as $loss(K)$ which is a function of K instead, and search only in static state feedback controllers.

One approach of finding K^* is to solve the Riccati equations (21,22) to get K^*. It is also of interest of solving (20) by iterative optimization algorithms, which enables us to finish early to get smaller computational complexity, implement in noisy case or when the system parameters A, B are inexactly measured, etc. In that case, one uses a reparameterization approach Mohammadi et al. (2019):

$$loss(K) = h(X, Y) = \mathbf{trace}(QX + Y^T R Y X^{-1}). \tag{23}$$

Let $\mathcal{A}(X) = AX + XA,\ \mathcal{B}(Y) = BY + Y^T B^T$, they have the relation

$$\mathcal{A}(X) + \mathcal{B}(Y) + \Omega = 0 \tag{24}$$

where $\Omega = \mathbf{E}(x_0 x_0^T)$. One can construct a bijection from X, Y to K, P, and prove that, if we minimize $h(X, Y)$ subject to (24), the optimizer X^*, Y^* will map to the optimal K^*, and this problem is convex, so we can solve it by optimization algorithms.

Recently nonconvex optimization algorithms are widely used in machine learning, so it is also of interest whether we can run gradient algorithm in K space without reparameterization, which means that, we run the gradient flow

$$\dot{K} = -\eta \nabla_K loss(K), \tag{25}$$

can K converge to the optimal controller K^*? Mohammadi et al. (2019) answers the question by studying the mapping between the originally and reparameterized spaces, and suggests that gradient flow converges to K^* in linear rate.

This work is a generalization of Mohammadi et al. (2019). We extend the approach to a far more generic sets of problems which covers the original LQR. We will prove that, for the static state feedback controller design problems in the set of problems, gradient dominance always hold. Based on that, we argue that the nonconvex optimization problem in K space can be globally solved by gradient flow.

4.1.4 Main theorem

Theorem 9. *We consider the problems*

$$\min_{K} \quad loss(K), \tag{26a}$$

$$s.t., \quad K \text{ stabilizes}, \ K \in \mathcal{S}_K \text{ for some set } S_K \tag{26b}$$

and

$$\min_{Z,L,G} \quad f(L,G,Z), \tag{27a}$$

$$s.t., \quad (L,G,Z) \in \mathcal{S} \tag{27b}$$

We allow a special case

$$\min_{L,G} \quad f(L,G), \tag{28a}$$

$$s.t., \quad (L,G) \in \mathcal{S} \tag{28b}$$

Then we require assumptions either 1,2,3 or 1,2,4 (3 is stronger than 4 but we emphasize this important special case):

1. *The feasibility set $\mathcal{S}$ is convex.*
2. *Cost function $f(L,G,Z)$ is convex.*
3. *(a) There is a bijection between K and L,G such that $\exists Z$, $(L,G,Z) \in \mathcal{S}$. Specifically, $K = LG^{-1}$ where $G \succeq \lambda_0 I \succ 0$.*
   (b) Define $\mathscr{Z}(L,G) \in \operatorname{argmin}_Z f(L,G,Z)$ subject to $(L,G,Z) \in \mathcal{S}$ (if there are multiple minimizers we pick any one). Then when K maps to (L,G), $loss(K) = f(L,G,\mathscr{Z}(L,G))$. (If the problem is (28), $loss(K) = f(L,G)$).
4. *We can express $loss(K)$ as:*

$$loss(K) = \min_{L,G,Z} \ f(L,G,Z) \quad s.t., \ (L,G,Z) \in \mathcal{S}, \ LG^{-1} = K.$$

Define K^* as the global minimum of $loss(K)$ in feasible region. Then if we solve the problem by gradient flow

$$\dot{K} = -\eta \nabla loss(K) \tag{29}$$

then $K(t)$ converges to the global optimizer K^*, and moreover,

1. *if f is linear, the gradient satisfies*

$$\|\nabla loss(K)\| \geq C(loss(K) - loss(K^*)). \tag{30}$$

   for some constant C.

2. *if f is μ-strongly convex, the gradient satisfies*

$$\|\nabla loss(K)\| \geq C(\mu(loss(K) - loss(K^*)))^{1/2}. \tag{31}$$

   for some constant C.

Example 1. *(Assumption 1,2,3)*

1. *LQR problem can be convexified as*

$$\begin{aligned}\min_{Z,L,G}\ & f(L,G,Z) := \mathbf{tr}(QG+ZR)\\ & s.t.,\ \mathcal{A}(G)+\mathcal{B}(L)+\Omega=0,\ G\succ 0,\\ & \begin{bmatrix} Z & L^T\\ L & G\end{bmatrix}\succeq 0\end{aligned}$$

2. *Adding a robust constraint: whenever $x^TG^{-1}x\le 1$, we enforce $\|u\|\le e$. This can be generalized with a constraint*

$$\begin{bmatrix} G & L^T\\ L & eI\end{bmatrix}\succeq 0.$$

3. *Budget on total energy of input, which is the following constraint*

$$\mathbf{E}_{x_0}\int_0^\infty \|u(t)\|^2dt = \mathbf{tr}(Z)\le e_0.$$

4. *Loss: barrier function for input energy*

$$loss(K) = \mathbf{E}_{x_0}\left(\int_0^\infty (x(t)^TQx(t)+u(t)^TRu(t))dt - \sum_{i=1}^p \lambda_i\log(e_i-\int_0^\infty u_i^2(t)dt)\right)$$

$$f(L,G,Z) = \mathbf{tr}(QG+ZR) - \sum_{i=1}^p \lambda_i\log(e_i-Z_{ii}).$$

Example 2. *(Assumption 1,2,4) We consider an input output system*

$$\dot{x} = Ax+Bu+B_ww,\ y=Cx+Dw$$

and still use the state feedback controller $u=Kx$. We consider minimizing the L_2 gain of the closed loop

$$loss(K) := \sup_{\|w\|_2=1}\|y\|_2.$$

This problem can be formulated as (Boyd et al., 1994, Sec 7.5.1)

$$\begin{aligned}&\min_{L,G,\gamma}\ f(L,G,\gamma):=\gamma\\ &s.t.,\ \begin{bmatrix} AG+GA^T+BL+L^TB^T+B_wB_w^T & (CG+DL)^T\\ CG+DL & -\gamma^2I\end{bmatrix} := M(L,G,\gamma)\preceq 0,\ \gamma\ge 0.\end{aligned}$$

And $K^*=L^*G^{*-1}$. We have $loss(K^*)=f(L^*,G^*,\gamma^*)$. Note that, K^* can map to different pairs (L,G) whenever $LG^{-1}=K^*$, and γ^* associates to one pair. We can equivalently formulate

$$\begin{aligned}loss(K^*) = \min_{L,G,\gamma}\ & \gamma\\ & s.t.,\ M(L,G,\gamma)\preceq 0,\ \gamma\ge 0,\ LG^{-1}=K^*.\end{aligned}$$

The question is, can we establish the same connection at any stabilizing controller K, say,

$$loss(K) \overset{?}{=} \min_{L,G,\gamma}\ \gamma \tag{32a}$$

$$s.t.,\ M(L,G,\gamma)\preceq 0,\ \gamma\ge 0,\ LG^{-1}=K. \tag{32b}$$

Note that, the intermediate step (Boyd et al., 1994, Sec 7.5.1) is

$$loss(K) = \min_{G,\gamma}\ \gamma \tag{33a}$$

$$s.t.,\ \begin{bmatrix} (A+BK)G+G(A+BK)^T+B_wB_w^T & G^T(C+DK)^T\\ (C+DK)G & -\gamma^2I\end{bmatrix}\preceq 0,\ \gamma\ge 0. \tag{33b}$$

Denote the optimizer of (32) by $\hat{L}, \hat{G}, \hat{\gamma}$, and the optimizer of (33) by $\check{G}, \check{\gamma}$.

Note $\hat{\gamma} \leq \check{\gamma}$. If (32) is not true, $\hat{\gamma} < \check{\gamma}$, we can replace $\check{G}, \check{\gamma}$ with $\hat{G}, \hat{\gamma}$ in (33) and it's still feasible. Thus the optimality condition of $\check{G}, \check{\gamma}$ in (33) is violated, which contradicts the assumption that (32) is not true. Then we claim that (32) is true.

Proof of Theorem 9. We start by proving a simple lemma.

Lemma 9. *A linear function on a convex set $\mathcal{S}$ is gradient dominant.*

Proof. Say the function is $f(x)$, the minimum is x^*, and $x - x^* = \Delta$. Let $\nabla f(x) = g$. For any non-stationary point, $f(x) = f(x^*) + g^T\Delta$. Since $\mathcal{S}$ is a convex set, $-\Delta$ belongs to the horizon of $\mathcal{S}$ at x, so there is a direction $\frac{\Delta}{\|\Delta\|}$ such that $f(x) - f(x - t\frac{\Delta}{\|\Delta\|}) > tg^T\frac{\Delta}{\|\Delta\|}$, $t \to 0$, so the norm of projected gradient $\|\mathcal{P}_{\mathcal{S}}(\nabla f(x))\| \geq g^T\frac{\Delta}{\|\Delta\|} = \frac{f(x)-f(x^*)}{\|x-x^*\|}$. Let K^* be the optimal K and (Z^*, L^*, G^*) be the optimal point in the parameterized space. We have $loss(K^*) = f(Z^*, L^*, G^*)$. Denote u as any matrix in K space, $\mathcal{P}_u$ is projection of a vector onto direction u, then

$$\nabla loss(K)^T \nabla loss(K) \geq (\mathcal{P}_u \nabla loss(K))^T \mathcal{P}_u \nabla loss(K) = \left(\frac{\nabla loss(K)[u]}{\|u\|_F}\right)^2. \tag{34}$$

At current iteration K_t,

1. (Assumption 3) let K_t map to (L_t, G_t) and $Z_t = \mathscr{Z}(L_t, G_t)$.
2. (Assumption 4) let

$$\begin{aligned}(L_t, G_t, Z_t) = \operatorname{argmin}_{L,G,Z}\ & f(L, G, Z) \\ \text{s.t., } & (L, G, Z) \in \mathcal{S},\ G \succ 0,\ LG^{-1} = K_t.\end{aligned}$$

Note f is convex, so

$$\begin{aligned}&\nabla f(Z_t, L_t, G_t)[(Z_t, L_t, G_t) - (Z^*, L^*, G^*)] \\ \geq\ & f(Z_t, L_t, G_t) - f(Z^*, L^*, G^*) \\ =\ & f(\mathscr{Z}(L_t, G_t), L_t, G_t) - f(\mathscr{Z}(L^*, G^*), L^*, G^*) \\ =\ & loss(K_t) - loss(K^*).\end{aligned} \tag{35}$$

Now we consider the directional derivative in K space. By definition,

$$\nabla loss(K)[u] = \lim_{t \to 0^+} (loss(K + tu) - loss(K))/t.$$

Let $\Delta L = L^* - L_t$, $\Delta G = G^* - G_t$, and $u = \Delta L G_t^{-1} - L_t G_t^{-1} \Delta G G_t^{-1}$. Then

$$\begin{aligned}\nabla loss(K)[u] &= \lim_{t \to 0^+} (loss(K + tu) - loss(K))/t \\ &= \lim_{t \to 0^+} (loss(L_t G_t^{-1} + t(\Delta L G_t^{-1} - L_t G_t^{-1} \Delta G G_t^{-1})) - loss(L_t G_t^{-1}))/t \\ &= \lim_{t \to 0^+} (loss((L_t + t\Delta L)(G_t + t\Delta G)^{-1}) - loss(L_t G_t^{-1}))/t\end{aligned}$$

With assumption 3, we continue with

$$\begin{aligned}\nabla loss(K)[u] &= \lim_{t \to 0^+} (f(L_t + t\Delta L, G_t + t\Delta G, \mathscr{Z}(L_t + t\Delta L, G_t + t\Delta G)) - f(L_t, G_t, \mathscr{Z}(L_t, G_t)))/t \\ &\leq \lim_{t \to 0^+} (f(L_t + t\Delta L, G_t + t\Delta G, Z_t + t\Delta Z) - f(L_t, G_t, \mathscr{Z}(L_t, G_t)))/t \\ &= \nabla f(Z_t, L_t, G_t)[(Z^*, L^*, G^*) - (Z_t, L_t, G_t)]\end{aligned}$$

With assumption 4, we continue with

$$\begin{aligned}\nabla loss(K)[u] = \lim_{t \to 0^+} \min_{L,G,Z}\ & f(L, G, Z) - f(L_t, G_t, Z_t) \\ \text{s.t., } & (L, G, Z) \in \mathcal{S},\ G \succ 0,\ LG^{-1} = (L_t + t\Delta L)(G_t + t\Delta G)^{-1}.\end{aligned}$$

and then

$$\begin{aligned}\nabla loss(K)[u] &\leq \lim_{t\to 0^+}(f(L_t+t\Delta L, G_t+t\Delta G, Z_t+t\Delta Z) - f(L_t, G_t, \mathscr{Z}(L_t, G_t)))/t \\ &= \nabla f(Z_t, L_t, G_t)[(Z^*, L^*, G^*) - (Z_t, L_t, G_t)]\end{aligned}$$

The inequality results in that $(L_t + t\Delta L, G_t + t\Delta G, Z_t + t\Delta Z)$ is feasible so the value is bigger than the optimal value. So the final inequality holds either by Assumption 3 or 4. So

$$\nabla loss(K)[-u] \geq \nabla f(Z_t, L_t, G_t)[(Z_t, L_t, G_t) - (Z^*, L^*, G^*)] > 0.$$

Using (34) and (35), we have

$$\nabla loss(K_t)^T \nabla loss(K_t) \geq \frac{1}{\|u\|_F^2}(loss(K_t) - loss(K^*))^2$$

Note that, although f is linear, it interacts with a general convex set $\mathcal{S}$ so that we cannot get the relation between $(\Delta L, \Delta G)$ and $f(Z_t, L_t, G_t) - f(Z^*, L^*, G^*)$, or equivalently with $loss(K) - loss(K^*)$. So we cannot further cancel $(loss(K_t) - loss(K^*))^2$ with $\|u\|_F^2$ despite $u \sim (\Delta L, \Delta G) \sim K - K^*$. If $f(Z, L, G)$ is μ strongly convex, then we can restrict f in the line segment $(Z_t, L_t, G_t) - (Z^*, L^*, G^*)$ and get

$$\|\mathcal{P}_{(Z_t,L_t,G_t)-(Z^*,L^*,G^*)}\nabla f(Z_t, L_t, G_t)\| \geq \mu^{1/2}(f(Z_t, L_t, G_t) - f(Z^*, L^*, G^*))^{1/2}$$

then we have that

$$\begin{aligned}&\nabla f(Z_t, L_t, G_t)[(Z_t, L_t, G_t) - (Z^*, L^*, G^*)] \\ &= \|\mathcal{P}_{(Z_t,L_t,G_t)-(Z^*,L^*,G^*)}\nabla f(Z_t, L_t, G_t)\| \cdot \|(Z_t, L_t, G_t) - (Z^*, L^*, G^*)\| \\ &\geq \mu^{1/2}(f(Z_t, L_t, G_t) - f(Z^*, L^*, G^*))^{1/2}\|(Z_t, L_t, G_t) - (Z^*, L^*, G^*)\|.\end{aligned}$$

then

$$\begin{aligned}\nabla loss(K_t)^T \nabla loss(K_t) &\geq \frac{1}{\|u\|_F^2}(\nabla f(Z_t, L_t, G_t)[(Z_t, L_t, G_t) - (Z^*, L^*, G^*)])^2 \\ &\geq \frac{\mu\|(Z_t, L_t, G_t) - (Z^*, L^*, G^*)\|^2}{\|u\|_F^2}(f(Z_t, L_t, G_t) - f(Z^*, L^*, G^*)) \\ &= \frac{\mu(\|L^* - L_t\|^2 + \|G^* - G_t\|^2 + \|Z^* - Z_t\|^2)}{\|(L^* - L_t)G_t^{-1} - L_tG_t^{-1}(G^* - G_t)G_t^{-1}\|_F^2}(f(Z_t, L_t, G_t) - f(Z^*, L^*, G^*)) \\ &\geq \frac{\mu(\|L^* - L_t\|^2 + \|G^* - G_t\|^2)}{\|(L^* - L_t)G_t^{-1} - L_tG_t^{-1}(G^* - G_t)G_t^{-1}\|_F^2}(f(Z_t, L_t, G_t) - f(Z^*, L^*, G^*)) \\ &\geq \frac{4\mu}{(\max\{\lambda_{min}^{-1}(G_t), \lambda_{min}^{-2}(G_t)\sigma_{\max}(L_t)\})^2}(f(Z_t, L_t, G_t) - f(Z^*, L^*, G^*)).\end{aligned}$$

4.1.5 Dynamic controller

In this subsection, we extend to dynamic controller case. In the dynamic setting, quadratic invariance is required for finding the optimal controller (Lessard & Lall, 2011), and the Youla parameterization (Rotkowitz & Lall, 2005) provides a convexification method. Analogous as before, we study the space of dynamic controllers where the loss is not convex, and we prove a similar gradient dominance property as before. Hence the only stationary points are the global optimizer, which is the desired property for optimization algorithms. We start from introducing quadratic invariance.

Define subspace E satisfying strengthened quadratic invariance under G, which means for every $K_1, K_2 \in E$, $K_1GK_2 \in E$. And for $K \in S \subset E$, the series

$$\begin{aligned}(I - GK)^{-1} &= \sum_{i=1}^{n} a_i(GK)^{i-1}, \\ (I - KG)^{-1} &= \sum_{j=1}^{m} b_j(KG)^{j-1}.\end{aligned}$$

exist. Let $E\backslash S$ corresponds to infinity function value (non-stabilizing K or $I-GK$ singular). Consider S which satisfies strengthened quadratic invariance under P_{22}, and optimization problem

$$\min \quad f(K) := \frac{1}{2}\int \|P_{11}+P_{12}K(I-P_{22}K)^{-1}P_{21}\|_F^2 dw, \tag{37a}$$

$$\text{subject to} \quad K \in S \tag{37b}$$

Every stationary point satisfies

$$\nabla f(K)[D] := \mathbf{tr}(\nabla f(K)^* D) = 0, \forall D \in E \tag{38}$$

Approach 1: mapping to real case. We use jw instead of w and let $K(w) = u(w) + jv(w)$, and integrate from $-j\infty$ to $j\infty$, in this case we can make w, u, v real.

$$f(K) := \frac{1}{2}\int_{-\infty}^{\infty} \|P_{11}+P_{12}(u+jv)(I-P_{22}(u+jv))^{-1}P_{21}\|_F^2 djw$$

In this case we have to parametrize everything inside the norm by $P_1(w)+jP_2(w)$, and the norm term becomes $P_1^TP_1+P_2^TP_2$, and consider dP_1/du and so on.

Approach 2: Wirtinger derivative and parametrization. Absolute value of complex number is not differentiable, but we can use Wirtinger derivative. Observe the objective function, we see that the expression inside the norm is differentiable, and Wirtinger derivative allows chain rule. Let $K(w)$ be parametrized by a and write $K(w) := K(a; w)$. Suppose we can obtain all legit functions in E by choosing a, and we never go outside E. Then let b be a perturbation of a, and we compute directional derivative with respect to a in direction b. Let $P := P_{11} + P_{12}K(I-P_{22}K)^{-1}P_{21}$,

$$\begin{aligned}\nabla_a f(K)[b] &= (\nabla_P f(P,\bar{P}))(\nabla_K P)(\nabla_a K)[b] + (\nabla_{\bar{P}} f(P,\bar{P}))(\nabla_{\bar{K}}\bar{P})(\nabla_a \bar{K})[b] \\ &= 2\mathscr{R}\left\{\mathbf{tr}\left(\int \left((P_{11}+P_{12}K(I-P_{22}K)^{-1}P_{21})^*\right.\right.\right. \\ &\qquad \left.\left.\left. P_{12}(\nabla_a K[b](I-P_{22}K)^{-1} + K(I-P_{22}K)^{-1}P_{22}\nabla_a K[b](I-P_{22}K)^{-1})P_{21}\right) dw\right)\right\}.\end{aligned}$$

Here $\mathscr{R}$ denotes real part. Here the question is how to parametrize E, and how to make sure that $\nabla_a(K) \neq 0$ in order not to induce new saddle points.

Approach 3: Derivative with respect to K. In parametrization case, we used $(\nabla_a K)[b]$ and $(\nabla_a \bar{K})[b]$. However, if it's possible not to do the last layer derivative, just make $(\nabla_a K)[b] = dK$, and we can formally take directional derivative with respect to K in direction dK. Call $dK := D$, and

$$\nabla f(K)[D] = \mathbf{tr}\left(\int \left((P_{11}+P_{12}K(I-P_{22}K)^{-1}P_{21})^*\right.\right. \tag{39a}$$

$$\left.\left. P_{12}(D(I-P_{22}K)^{-1} + K(I-P_{22}K)^{-1}P_{22}D(I-P_{22}K)^{-1})P_{21}\right) dw\right) \tag{39b}$$

$$= \mathbf{tr}\left(\int \left((P_{11}+P_{12}K(I-P_{22}K)^{-1}P_{21})^*\right.\right. \tag{39c}$$

$$\left.\left. P_{12}(I+K(I-P_{22}K)^{-1}P_{22})D(I-P_{22}K)^{-1}P_{21}\right) dw\right) \tag{39d}$$

$$= \mathbf{tr}\left(\int \left(P_{11}+P_{12}K(I-P_{22}K)^{-1}P_{21}\right)^* P_{12}(I-KP_{22})^{-1}D(I-P_{22}K)^{-1}P_{21}\right) dw\Bigg). \tag{39e}$$

Note that we do not have real part compare to the second approach, since $\nabla_a K$ and $\nabla_a \bar{K}$ both exist (and they are conjugate), but $\nabla_K \bar{K} = 0$.

Denote

$$\begin{aligned}\bar{E} &:= (I-KP_{22})^{-1}E(I-P_{22}K)^{-1} \\ &= \left\{(I-KP_{22})^{-1}D(I-P_{22}K)^{-1} : D \in E\right\}.\end{aligned}$$

Now we prove $\bar{E} = E$. If $\bar{D} \in \bar{E}$,

$$\begin{aligned}\bar{D} &= (I - KP_{22})^{-1} D (I - P_{22}K)^{-1} \\ &= (\sum_{j=1}^{m} b_j (KP_{22})^{j-1}) D (\sum_{i=1}^{n} a_i (P_{22}K)^{i-1}).\end{aligned}$$

Every term in the sum is in E and E is a subspace, so $\bar{D} \in E$, $\bar{E} \subseteq E$. If $D \in E$,

$$D = (I - KP_{22})\bar{D}(I - P_{22}K). \tag{42}$$

Now

$$\begin{aligned}\bar{D}P_{22}K &= (I - KP_{22})^{-1} D (I - P_{22}K)^{-1} P_{22}K \\ &= (I - KP_{22})^{-1} D (\sum_{i=1}^{n} a_i (P_{22}K)^{i-1}) P_{22}K \\ &= (I - KP_{22})^{-1} D P_{22}K (\sum_{i=1}^{n} a_i (P_{22}K)^{i-1}) \\ &= (I - KP_{22})^{-1} D P_{22}K (I - P_{22}K)^{-1}.\end{aligned}$$

Note $DP_{22}K \in E$, so $\bar{D}P_{22}K \in \bar{E}$. Similarly $KP_{22}\bar{D}, KP_{22}\bar{D}P_{22}K \in \bar{E}$, so $D \in \bar{E}$, $E \subseteq \bar{E}$. So $E = \bar{E}$.

This suggests that, we can reparametrize $D \leftarrow (I - KP_{22})\hat{D}(I - P_{22}K)$, such that if I traverse all $\hat{D} \in E$, it's equivalent to traverse all $D \in E$. Change variable and set (39e)=0,

$$\int \mathbf{tr}\left((P_{11} + P_{12}K(I - P_{22}K)^{-1}P_{21})^* P_{12}\hat{D}P_{21}\right) dw = 0, \ \ \forall \hat{D} \in E, \tag{44}$$

by replacing $(I - KP_{22})^{-1} D (I - P_{22}K)^{-1}$ with $\hat{D}$. Define subspace Z such that

$$\mathcal{Z} = \left\{ Z : \int \mathbf{tr}(Z^* P_{12} D P_{21}) dw = 0, \ \ \forall D \in E \right\}. \tag{45}$$

Define the inner product by $\langle A(w), B(w) \rangle = \int \mathbf{tr}(A(w)^* B(w)) dw$ and define projection $\mathcal{P}_{\mathcal{Z}}$ with respect to this inner product, we have $P_{12}K(I - P_{22}K)^{-1}P_{21} \perp \mathcal{Z}$, $P_{12}DP_{21} \perp \mathcal{Z}$, (44) means $P_{11} + P_{12}K(I - P_{22}K)^{-1}P_{21} = \mathcal{P}_{\mathcal{Z}} P_{11}$. This means $\forall K \in S$ such that $\nabla f(K)[D] = 0, \forall D \in E$, i.e., $K \in S$ is a stationary point, $f(K) = \frac{1}{2}\int \|\mathcal{P}_{\mathcal{Z}} P_{11}\|_F^2 dw$ is the same function value, so K has to be the global minimizer.

4.1.6 Conclusion and discussion

In this section, we discussed the nonconvex landscape of some classical optimal control problems. The simple LQR problem can be solved by Riccati equations, and the more general ones can be solved by convexification, a change of variable method to make an equivalent convex problem. The nonconvex landscape of the original problems are studied in Fazel et al. (2018); Mohammadi et al. (2019); Bu et al. (2019); Zhang et al. (2020), however the application of the analysis are restricted to a couple of problems and the proof technique does not bridge the nonconvex problem and convexification methods. This section sheds light on the interaction between original and convexified problems via a concise proof, and this approach can be generalized to a big collection of control problems with different objectives, constraints, both static and dynamic cases, etc.

There are still exciting directions to pursue. Firstly, the convergence rate of gradient descent on the class of problems, is not as good as the analysis designed for every specific problems, such as Fazel et al. (2018); Mohammadi et al. (2019); Bu et al. (2019); Zhang et al. (2020), so it is hoped to improve in at least a few regimes. Secondly, we need to find a functional optimization algorithm for the dynamical controller setting that both theoretically converges and practically works. Finally, we will do experiments and make the section in the form of academic paper for submission.

4.2 Understanding the role of representation dimension in meta-learning

4.2.1 Introduction

Meta-learning aims to uncover the principle features of the tasks, which are often low-dimensional, from limited data available for related tasks. We consider a setup where task features are approximately in a low dimensional subspace,

and we do not know the subspace and its dimension beforehand. A low rank approximation step is commonly used to retrieve the low dimensional space. When the tasks are only approximately low dimensional, the dimension reduction step is not necessarily optimal for generalizing to the new meta-test task.

This part aims to provide a theoretical comparison of more traditional low-dimensional representations, obtained by low-rank truncation, and high-dimensional representations which might actually lead to an ill-posed overparameterized problem when learning a new task from limited data. We will show that learning large representations where directions are weighted by their relative importance can in fact be the right choice over small representations. Furthermore, we will reveal a double descent phenomena when the representation dimension coincides with the sample size. Extensive experiments corroborate the benefit of large and weighted representations over utilizing simple low-dimensional subspaces.

4.2.2 Problem formulation

The meta-learning algorithm consists of two phases, meta-training and meta-testing. In the meta-training phase, we adopt a mixed linear regression setup to learn the low dimensional feature space. In the meta-testing phase, we use the feature space to learn the feature of a new task.

In the first phase, there are a few tasks, each associated with its own parameters, called features. One accesses batches of data, each of whom is collected from a task, but we do not know which task it comes from. We make this setup more precise using the following definitions.

Definition 1. ***Task.*** *A task is associated with parameter $\beta \in \mathbb{R}^d$. For any parameter β, one can generate data $(\boldsymbol{x}, y)$ from it. Here $\boldsymbol{x} \in \mathbb{R}^d$, $y \in \mathbb{R}$. $\boldsymbol{x}$, y satisfy $y = \boldsymbol{x}^\top \beta$. We assume there are in total K tasks in the training dataset.*

Definition 2. ***Task distribution.*** *All the training tasks β_k for $k = 1, ..., K$ are generated i.i.d. from a Gaussian distribution $\mathcal{N}(0, \boldsymbol{\Sigma})$. For simplicity of notation, we assume all β_k are distinct*[9]*. We denote the set of tasks as $\mathcal{S}^{\text{task}} = \{\beta_k \mid k = 1, ..., K\}$.*

Definition 3. ***Meta-training data and batches.*** *There are in total n batches of data, and $n \geq K$. The j-th batch is denoted as $\mathcal{S}_j$ and let $t_j := |\mathcal{S}_j|$. Denote $\mathcal{S}_j = \{(\boldsymbol{x}_{i,j}, y_{i,j}) \mid i = 1, ..., t_j\}$ and $\mathcal{S} = \{(\boldsymbol{x}_{i,j}, y_{i,j}) \mid j = 1, ..., n,\ i = 1, ..., t_j\}$. The data from the j-th batch is generated from a task $\theta_j \in \mathbb{R}^d$, which means $y_{i,j} = \boldsymbol{x}_{i,j}^\top \theta_j$.*

We have defined the set of tasks $\mathcal{S}^{\text{task}}$ in Def. 2. We assume that $\theta_j \in \mathcal{S}^{\text{task}}$, i.e., $\theta_j = \beta_k$ for some number $k \in [1, K]$. However, we do not know the correspondence between j and k.

For a fixed k, β_k can correspond to multiple batches. Let b_k denote the number of batches corresponding to β_k, defined as

$$b_k = |\{j \mid \theta_j = \beta_k\}|, \ k = 1, ..., K.$$

Note that $\sum_{k=1}^K b_k = n$ since there are n batches in total.

The number of tasks is K, and the number of batches is n. We use β_k to denote the feature of the k-th task and θ_j to denote the feature corresponding to the j-th batch. To explain the relation of β_k and θ_j, we list two examples.

1. The feature vector of each batch of data is independently generated from $\mathcal{N}(0, \boldsymbol{\Sigma})$. In this case $k = n$, and we can define $\beta_j = \theta_j$, $j = 1, ..., n$. $b_k = 1$ for all k from 1 to K.
2. The set of task features $\mathcal{S}^{\text{task}}$ is generated independently from $\boldsymbol{\Sigma}$. The task corresponding to the data of the j-th batch, is randomly chosen from $\mathcal{S}^{\text{task}}$, with probability $p_1, ..., p_K$. In expectation of $p_1, ..., p_K$, we count how many times the k-th task appears among all batches of data, and it follows

$$\boldsymbol{E}_{\boldsymbol{P}(\theta_j = \beta_k) = p_k} b_k = n p_k.$$

Definition 4. ***Data distribution.*** *All $\boldsymbol{x}_{i,j} \in \mathbb{R}^d$ are generated i.i.d. from distribution $\mathcal{N}(0, \boldsymbol{\Sigma}_x)$ and $y_{i,j} = \boldsymbol{x}_{i,j}^\top \theta_j$. For simplicity, we discuss $\boldsymbol{\Sigma}_x = I$.*

Remark 1. *If $\boldsymbol{\Sigma}_x \neq I$ and we know $\boldsymbol{\Sigma}_x$ when running the algorithm, we can consider*

$$\begin{aligned} y_{i,j} &= \boldsymbol{x}_{i,j}^\top \theta_j \\ &= (\boldsymbol{\Sigma}_x^{-1/2} \boldsymbol{x}_{i,j})^\top \boldsymbol{\Sigma}_x^{1/2} \theta_j. \end{aligned}$$

Then the distribution of $\boldsymbol{\Sigma}_x^{-1/2} \boldsymbol{x}_{i,j}$ is standard normal. We can use the same method to estimate $\boldsymbol{\Sigma}^{1/2} \theta_j$.

[9] That happens almost surely.

Let the eigendecomposition of $\boldsymbol{\Sigma}$ be $\boldsymbol{\Sigma} = \boldsymbol{U}\Lambda\boldsymbol{U}^\top$. We study the case that Λ is spiked with size r, say $\lambda_1, ..., \lambda_r \gg \lambda_{r+1}, ..., \lambda_d$.

Definition 5. ***Effective dimension.*** *(Informal) When the eigenvalue of Σ satisfies $\lambda_r \gg \lambda_{r+1}$, we can use r as the effective dimension.*

The effective dimension is informally defined, as we do not quantify $\lambda_r \gg \lambda_{r+1}$ exactly. We define effective dimension because when we fix any r, we will bound the expected generalization error (Def. 9) by the magnitude of $\lambda_{r+1}, ..., \lambda_d$. When λ_{r+1} is small, the expected generalization error is guaranteed to be small.

When the dimension of the span of features is low, Kong et al. (2020b) performs a dimension reduction algorithm to find the low dimensional subspace that the features span. This is done by selecting the top eigenvectors of the feature covariance estimator.

Definition 6. ***Moment estimator of feature covariance.*** *The covariance (of all β_j) estimator is*

$$\hat{\boldsymbol{M}} = \sum_{j=1}^{n} \frac{2}{t_j^2} \left[(\sum_{i=1}^{t_j/2} y_{i,j}\boldsymbol{x}_{i,j})(\sum_{i=t_j/2+1}^{t_j} y_{i,j}\boldsymbol{x}_{i,j})^\top + (\sum_{i=t_j/2+1}^{t_j} y_{i,j}\boldsymbol{x}_{i,j})(\sum_{i=1}^{t_j/2} y_{i,j}\boldsymbol{x}_{i,j})^\top \right] \tag{46a}$$

When $\boldsymbol{\Sigma}_x = I$, the mean of the subGaussian matrix random variable $\hat{\boldsymbol{M}}$ is

$$\boldsymbol{M} = \boldsymbol{E}_{\boldsymbol{x}\sim\mathcal{N}(0,\boldsymbol{\Sigma}_x)}(\hat{\boldsymbol{M}}) = \frac{1}{n}\sum_{k=1}^{K} b_k \beta_k \beta_k^\top$$

If we consider the randomness of β_k, then $\boldsymbol{E}_{\beta_1,...,\beta_K\sim\mathcal{N}(0,\boldsymbol{\Sigma})}(\boldsymbol{M}) = \boldsymbol{\Sigma}$.

We mention two lemmas that are straightforward corollaries of Kong et al. (2020b).

Lemma 10. *((Kong et al., 2020b, Lemma 5.1, A.9)) Let $\epsilon > 0$ be the estimation error, $t_j \geq \underline{t} \geq 2$ for all $j = 1, ..., n$, let $\delta \in (0, 1)$ be failure probability, $n\underline{t} \gtrsim d\max\{\epsilon^{-2}, \epsilon^{-1}\log(nd/\delta)\}\log^2(nd/\delta)$.[10] With probability $1-\delta$, we have $\|\hat{\boldsymbol{M}} - \boldsymbol{M}\| \leq \epsilon d$.*

Lemma 11. *((Kong et al., 2020b, Lemma A.10)) Let $\epsilon > 0$, $\delta \in (0, 1)$ be failure probability, $n \gtrsim \log^3(K/\delta)/\epsilon^2$. With probability $1-\delta$, we have $\|\boldsymbol{\Sigma} - \boldsymbol{M}\| \leq \epsilon d$.*

Throughout the work, we start from the analysis for the case $\hat{\boldsymbol{M}} = \boldsymbol{\Sigma}$, and generalize to the case where $\|\hat{\boldsymbol{M}} - \boldsymbol{\Sigma}\|$ is bounded by a positive number, which is justified using the lemmas above.

Eigenvalue truncation. We do the R-eigenvalue truncation to $\hat{\boldsymbol{M}}$ to get a low dimensional truncation. Let $\hat{\boldsymbol{U}}\hat{\Lambda}\hat{\boldsymbol{U}}^\top$ be the eigen-decomposition of $\hat{\boldsymbol{M}}$. Denote $\hat{\lambda}_j$ as the jth eigenvalue of $\hat{\Lambda}$. Let $\hat{\boldsymbol{U}}_R$ be the first R columns of $\hat{\boldsymbol{U}}$, and the R-eigenvalue truncation is $\hat{\boldsymbol{M}}_R = \hat{\boldsymbol{U}}_R\text{diag}(\hat{\lambda}_1, ..., \hat{\lambda}_R)\hat{\boldsymbol{U}}_R^\top$. Note that eigenvalues can be negative. When $R < r$, the recovery is incorrect since a big eigenvalue is truncated, which means an important feature is missing. So we only consider $r \leq R \leq d$.

Definition 7. ***Representation dimension.*** *We call R, the rank at which we do eigenvalue truncation, as representation dimension.*

The moment estimator is utilized for learning the feature of new task, which is generated independently from a common distribution with the same approximately low rank covariance matrix.

Definition 8. ***Meta-test task.*** *The meta-test task corresponds to a vector β, which is generated from $\mathcal{N}(0, \boldsymbol{\Sigma})$ independently from the training tasks. There are t samples from the dataset. With $i = 1, ..., t$, $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^d$ are i.i.d. standard normal and we have $\tilde{y}_i = \tilde{\boldsymbol{x}}_i^\top\beta$. We use $\mathcal{S}_{\text{MT}} = \{(\tilde{\boldsymbol{x}}_i, \tilde{y}_i) \mid i = 1, ..., t\}$ to denote the mata-training dataset.*

The amount of data is not sufficient, $t \in [r, d]$. We first project $\boldsymbol{x}$ onto a R dimensional subspace. Let $\boldsymbol{A} \in \mathbb{R}^{R\times R}$, we regress the map $\tilde{y}_i = (\boldsymbol{A}\hat{\boldsymbol{U}}_R^\top\tilde{\boldsymbol{x}}_i)^\top\alpha$ to find $\alpha \in \mathbb{R}^R$. Let $\boldsymbol{X} \in \mathbb{R}^{t\times d}$ and the i-th row is $\tilde{\boldsymbol{x}}_i^\top$, $\boldsymbol{y} = [\tilde{y}_1, ..., \tilde{y}_t]^\top$, then we estimate

$$\hat{\alpha} = (\boldsymbol{X}\boldsymbol{B})^\dagger\boldsymbol{y}.$$

We study the generalization guarantee.

[10] The notations $\gtrsim$, $\lesssim$, $\approx$ mean comparison regardless of constant factors.

Definition 9. ***Generalization error and expected generalization error.*** *For any vector $\alpha \in \mathbb{R}^R$, we define the generalization error as*

$$\begin{aligned}\mathcal{L}(\alpha) &= \boldsymbol{E}_{\boldsymbol{x}\sim\mathcal{N}(0,\boldsymbol{\Sigma}_x),y=\boldsymbol{x}^\top\beta}(\alpha^\top \boldsymbol{A}\hat{\boldsymbol{U}}_R^\top \boldsymbol{x} - y)^2 \\ &= \boldsymbol{E}_{\boldsymbol{x}\sim\mathcal{N}(0,\boldsymbol{\Sigma}_x)}((\hat{\boldsymbol{U}}_R\boldsymbol{A}^\top\alpha - \beta)^\top \boldsymbol{x})^2\end{aligned}$$

The expected generalization error is

$$\boldsymbol{E}_{\beta\sim\mathcal{N}(0,\boldsymbol{\Sigma})}\mathcal{L}(\alpha)$$

Note that, if α is the solution of a linear regression problem in meta-test phase, with the meta-test data dependent on β, e.g., the least squares solution $\hat{\alpha}$, then $\mathcal{L}(\hat{\alpha})$ implicitly depends on β.

Algorithm 3 Representation learning

Require: Dataset $\mathcal{S}$ with n batches, j-th batch has size t_j. Dimension d, number of tasks K. The meta-test task dataset $\mathcal{S}_{\mathrm{MT}}$ with batch size t. Representation dimension R.
Compute $\hat{\boldsymbol{M}}$ by (46).
R-eigenvalue truncation:
$\hat{\boldsymbol{M}}_R \leftarrow \hat{\boldsymbol{U}}_R\mathrm{diag}(\hat{\Lambda}_{1,1},...,\hat{\Lambda}_{R,R})\hat{\mathbf{V}}_R^\top$.
$\boldsymbol{A} \leftarrow \sqrt{\max\{0,\hat{\Lambda}_R\}}$
$\boldsymbol{B} \leftarrow \hat{\boldsymbol{U}}_R\boldsymbol{A}^\top$, $\boldsymbol{y} \leftarrow [y_1,...,y_t]^\top$.
$\hat{\alpha} \leftarrow (\boldsymbol{X}\boldsymbol{B})^\dagger\boldsymbol{y}$.
return $\hat{\alpha}$

The representation learning algorithm is detailed in Algorithm 3. As a summary, it involves the following steps:

1. Construct the moment estimator $\hat{\boldsymbol{M}}$ from the training data. $\hat{\boldsymbol{M}}$ estimates the covariance matrix $\boldsymbol{\Sigma}$ of the tasks β.
2. Execute dimension reduction by applying truncated eigen-decompostion to moment estimator $\hat{\boldsymbol{M}}$, with truncation level R. We will study different choices of R. Specifically, when $R = d$, we do not reduce the dimension of $\hat{\boldsymbol{M}}$.
3. Construct a weighting matrix $\boldsymbol{A}$, the shaping matrix $\boldsymbol{B}$ and the shaped data matrix $\boldsymbol{X}\boldsymbol{B}$, used in the next linear regression step.
4. Solve linear regression with $\boldsymbol{X}\boldsymbol{B}$ and $\boldsymbol{y}$.

We study the following choices of parameters and the corresponding generalization error.

1. Choice of R: $R = r$, $R = t$, $R = d$.
2. Choice of $\boldsymbol{A}$ when $R = d$: $\boldsymbol{A} = I_R$, $\boldsymbol{A} = \sqrt{\hat{\Lambda}_R}$.

When $R \le t$, different choice of $\boldsymbol{A}$ leads to the same least squares solution (whenever $\boldsymbol{A}$ is invertible). In the overparametrized case when $R > t$, we adopt the smallest ℓ_2 norm solution, which depends on the choice of $\boldsymbol{A}$.

We will show and plan to quantify the following observation.

1. When $R = d$ and $\boldsymbol{A} = \sqrt{\max\{0,\hat{\Lambda}\}}$, the generalization error is small.
2. When $R = t$, the generalization error is big.
3. When $R = r$, the generalization error is small.

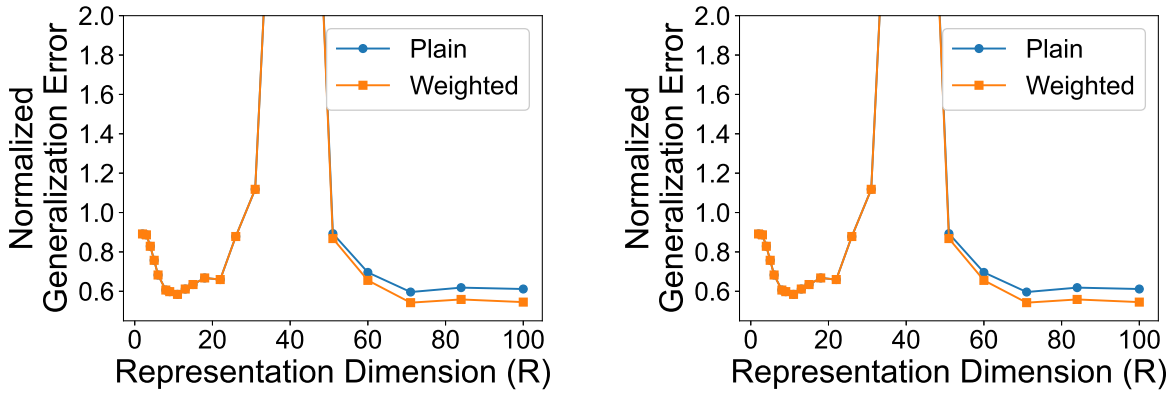
4.2.3 Experiments

In this section, we will illustrate the double descent phenomenon by simulation and an experiment on the MNIST dataset. When we choose the representation dimension $R = d, r$ to make overparametrized or overdetermined, we should see small generalization error, compared to $R = t$ when we should see big error. We will run the algorithm with the representation dimension ranging, and plot their empirical generalization error.

The dimension of the problem is d, the effective dimension of the features is r, which means the feature vector $\beta \in \mathbb{R}^d$ is generated from $\mathcal{N}(0,\boldsymbol{\Sigma})$, and λ_{r+1}, the $r+1$-th eigenvalue of $\boldsymbol{\Sigma}$, is much smaller than the r-th eigenvalue λ_r. The number of meta-test data is t, and $r < t < d$. The representation dimension, at which we perform dimension reduction, is R.

We considered three cases. The baseline we compare with is $R = r$, where we take the top r eigenvectors of covariance estimator for meta-test, and we give an exact (up to constant constant) theoretical error bound. When $R = d$, we solve an overparametrized linear regression. The error is as small as when $R = r$. When solving the linear regression at $R = t$, the error is much bigger than the other two cases.

We will plot the empirical generalization error versus different choice of R. Set $r = 10$, and the 11th eigenvalue of $\mathbf{\Sigma}$ is either 0 or very small. In Fig. 13, we first choose $\mathbf{\Sigma} = [I_{10}, 0; 0, 0]$, an exact low dimensional feature, which means the covariance is exactly rank-r. We first generate 50 tasks for estimating the covariance of β, and get the matrix $\hat{\boldsymbol{M}}$. Note that $\hat{\boldsymbol{M}}$ is not necessarily equal to $\mathbf{\Sigma}$. Then we range R from 2 to 100, and train it on meta-test task, which is also generated from $\mathbf{\Sigma}$. After training we test it with the new data generated i.i.d. from the same task, and record the generalization error. We repeat meta-test for 50 times and report the average generalization error. We plot two lines with different $\boldsymbol{A}$. The "weighted" case corresponds to $\boldsymbol{A} = \sqrt{\max\{\hat{\Lambda}_R, 0\}}$, and the "plain" case corresponds to $\boldsymbol{A} = I$. The solutions are same in overdetermined case, but with overparametrization the solutions are different. In the second figure we plot the case when $\mathbf{\Sigma} = [I_{10}, 0; 0, 0.01 \cdot I_{90}]$, the other settings are the same.



(a) (b)

Figure 13: Synthetic data, error of meta-test with different representation dimension. Left, $\mathbf{\Sigma} = [I_{10}, 0; 0, 0]$; Right, $\mathbf{\Sigma} = [I_{10}, 0; 0, 0.01 \cdot I_{90}]$. "Plain" and "Weighted" corresponds to $\boldsymbol{A} = I$ or $\boldsymbol{A} = \sqrt{\max\{\hat{\Lambda}_R, 0\}}$. Small error when $R = r$ or $R = d$, big error when $R = t$.

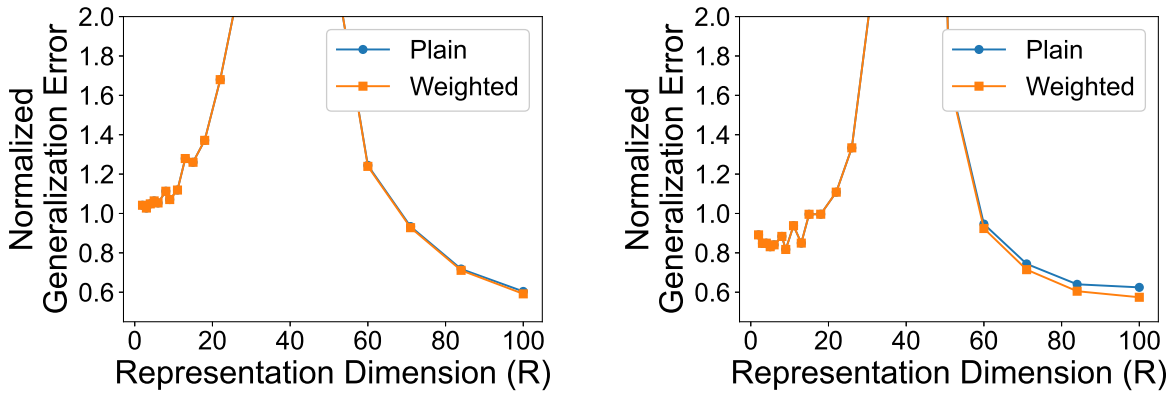
In Figure 13, the generalization error when $R = r$ and $R = d$ are both small, because the covariance is approximately low rank. In Figure 14, we choose the covariance of task as $\mathbf{\Sigma} = \mathrm{diag}(\mathbf{1}_{10}, 0.2 \cdot \mathbf{1}_{10}, 0.1 \cdot \mathbf{1}_{80})$, or $\mathbf{\Sigma}$ such that $\mathbf{\Sigma}_{j,j} = (1 + 9j/100)^{-3}$, whose eigenvalues decrease fast but has no clear gaps in between. We can still see in the figure that the generalization error is small on two sides $R = r$ or $R = d$ and explodes when $R = t$. Also the generalization error of $R = d$ case is slightly smaller than $R = r$. This is due to that the tail of the eigenvalues is heavier than the settings in Figure 13, thus the eigenvalue truncation causes more error.

We also run linear regression on MNIST dataset to classify the images of numbers 0 to 9. Each image is of size 28×28 and we flatten it to $\mathbb{R}^{784}$. Let $\boldsymbol{x}_{i,j}$ be the i-th image in the j-th class, where $j = 0, ..., 9$. We take the images of number 3 to 9, each 3000 samples for estimating the feature covariance. Different from mixed linear regression, we split the images of each class into two disjoint batches, each consisting of 1500 samples. Let $t_j = 3000$ for all j, The covariance estimator is

$$\hat{\boldsymbol{M}} = \sum_{j=3}^{9} \frac{2}{t_j^2} \left[(\sum_{i=1}^{t_j/2} \boldsymbol{x}_{i,j})(\sum_{i=t_j/2+1}^{t_j} \boldsymbol{x}_{i,j})^\top + (\sum_{i=t_j/2+1}^{t_j} \boldsymbol{x}_{i,j})(\sum_{i=1}^{t_j/2} \boldsymbol{x}_{i,j})^\top \right]$$

Then we record the eigenvalues of $\hat{\boldsymbol{M}}$ and truncate $\hat{\boldsymbol{M}}$ to different ranks for further use. The meta-test data contains 125 images of number 0 to 2, and we will learn a classifier for them. We apply one-hot encoding to the labels. If the image is 0, the label is transformed to $[1, 0, 0]$, similarly for $1, 2$. Note that the label is 3 dimensional, but we can still run linear regression. Let $\boldsymbol{y} \in \mathbb{R}^{125 \times 3}$ where each row is a one-hot encoded label. The solution is

$$\hat{\alpha} = \mathrm{argmin}_{\alpha \in \mathbb{R}^{R \times 3}} \|\boldsymbol{X}\boldsymbol{B}\alpha - \boldsymbol{y}\|^2$$



(a) (b)

Figure 14: Synthetic data, error of meta-test with different representation dimension. Left, $\mathbf{\Sigma} = \text{diag}(\mathbf{1}_{10}, 0.2 \cdot \mathbf{1}_{10}, 0.1 \cdot \mathbf{1}_{80})$; Right, $\mathbf{\Sigma} = \text{diag}((1 + 9j/100)^{-3})$. Small error when $R = r$ or $R = d$, big error when $R = t$.

The empirical generalization error is plotted in Figure 15. Surprisingly, the generalization error without weighting is smaller than that with weighting. We believe the reason is that, the MNIST dataset do not satisfy the assumptions in our theorem. The images of each class is clustered about a point rather than zero centered Gaussian random vectors, and the feature space of numbers 3 to 9 might not cover the feature space of numbers 0 to 2. Although the real data does not have the desired statistical propertied above, it still suffices to validate the double descent theory. The generalization error is low when the rank is small but above 10 (the representation dimension is above the effective dimension). With R becoming bigger, the generalization error increases but drops again when $R > t$ and gets closer to 784. We also see that the error is lower when $R = 784$ compared to any low-dimensional representation.

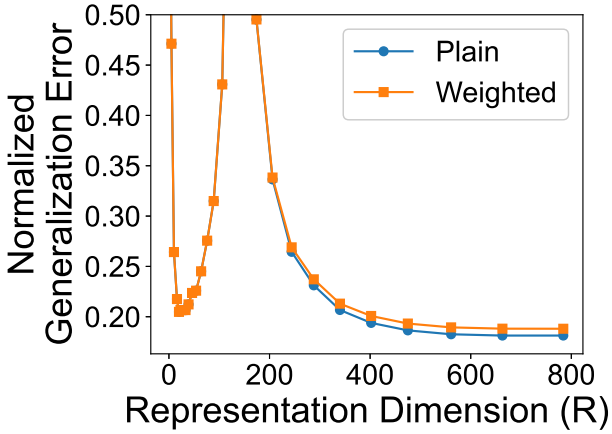


Figure 15: MNIST data, error of meta-test with different representation dimension. Training data: images of 3-9; Test data: images of 0-2.

4.2.4 Future directions

In this part, we plan to further investigate the role of representation dimension in meta-learning tasks, and seeking alternative algorithms besides the dimension-reduced least squares, such as the weighted overparametrized least squares mentioned above. Beyonds that, we plan to study the following ideas within the scope of meta-learning and model representation.

1. **Nonlinear tasks.** Existing theoretical results on representation learning are restricted to linear tasks, such

as Kong et al. (2020b,a); Hastie et al. (2019), whereas many important machine learning tasks are nonlinear. An example is classification, where we learn a mapping from objectives to discrete labels, which is usually nonlinear. If one can learn the model that classify between a few tasks, the feature space used in this model can be retrieved and transferred to a classifier for other objectives. We are interested whether the idea of identifying low dimension representation of objectives can be applied to nonlinear tasks, and how the low-dimensional representation benefits with generalization.

2. **Distribution of data.** It is usually assumed that the data $\boldsymbol{x}$ is distributed as i.i.d. standard normal vectors, which ensures that the moment estimator retrieves the low dimensional feature space. However, the real data is usually not Gaussian, for example, in classification tasks, the data are clustered, and the Gaussian mixture model is preferable. It is also desired that, the covariance of the data aligns with the feature subspace, which hopefully helps with learning the subspace. We plan to study how the distribution of data affects the estimation error with finite training samples.

3. **Application in reinforcement learning.** We are interested in the application of representation learning setup in reinforcement learning framework. One interesting connection is hierarchical reinforcement learning (HRL). In HRL, the overall task is difficult, which requires training a complicated policy. However, the big task can be split into a sequence of easy tasks, which allows us to train a simple policy for each specific task. It means that, after an appropriate splitting operation, the search space of downstream policy is smaller, which corresponds to the low dimensional feature space of meta-test phase in our setup. We plan to study the connection between HRL and meta-learning, and show that the HRL policy can be learned efficiently using similar ideas.

References

Abbasi-Yadkori, Y. and Szepesvári, C. Regret bounds for the adaptive control of linear quadratic systems. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 1–26, 2011.

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Online least squares estimation with self-normalized processes: An application to bandit problems. *arXiv preprint arXiv:1102.2670*, 2011.

Absil, P.-A., Mahony, R., and Sepulchre, R. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009a.

Absil, P.-A., Mahony, R., and Sepulchre, R. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009b.

Agarwal, N., Boumal, N., Bullins, B., and Cartis, C. Adaptive regularization with cubics on manifolds. *arXiv preprint arXiv:1806.00065*, 2018.

Agarwal, N., Bullins, B., Hazan, E., Kakade, S. M., and Singh, K. Online control with adversarial disturbances. *arXiv preprint arXiv:1902.08721*, 2019.

Ambrose, W. and Singer, I. M. A theorem on holonomy. *Transactions of the American Mathematical Society*, 75(3): 428–443, 1953.

Arias-Castro, E., Candes, E. J., and Davenport, M. A. On the fundamental limits of adaptive sensing. *IEEE Transactions on Information Theory*, 59(1):472–481, 2012.

Astrom, K. J. *Introduction to stochastic control theory*. Elsevier, 1971.

Avdiukhin, D., Jin, C., and Yaroslavtsev, G. Escaping saddle points with inequality constraints via noisy sticky projected gradient descent. In *11th Annual Workshop on Optimization for Machine Learning*, 2019.

Ayazoglu, M. and Sznaier, M. An algorithm for fast constrained nuclear norm minimization and applications to systems identification. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pp. 3469–3475. IEEE, 2012.

Banerjee, A., Chen, S., Fazayeli, F., and Sivakumar, V. Estimation with norm regularization. In *Advances in Neural Information Processing Systems*, pp. 1556–1564, 2014.

Bellman, R. Dynamic programming. *Science*, 153(3731):34–37, 1966.

Blomberg, N. *On nuclear norm minimization in system identification*. PhD thesis, KTH Royal Institute of Technology, 2016.

Blomberg, N., Rojas, C. R., and Wahlberg, B. Approximate regularization paths for nuclear norm minimization using singular value bounds–with implementation and extended appendix. *arXiv preprint arXiv:1504.05208*, 2015.

Bonnabel, S. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9): 2217–2229, 2013.

Boumal, N. and Absil, P.-a. Rtrmc: A riemannian trust-region method for low-rank matrix completion. In *Advances in neural information processing systems*, pp. 406–414, 2011.

Boumal, N., Absil, P.-A., and Cartis, C. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 2016a.

Boumal, N., Voroninski, V., and Bandeira, A. The non-convex burer-monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, pp. 2757–2765, 2016b.

Boumal, N., Absil, P.-A., and Cartis, C. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, pp. drx080, 2018. doi: 10.1093/imanum/drx080. URL `http://dx.doi.org/10.1093/imanum/drx080`.

Boyd, S., El Ghaoui, L., Feron, E., and Balakrishnan, V. *Linear matrix inequalities in system and control theory*. SIAM, 1994.

Bradtke, S. J., Ydstie, B. E., and Barto, A. G. Adaptive linear quadratic control using policy iteration. In *Proceedings of 1994 American Control Conference-ACC'94*, volume 3, pp. 3475–3479. IEEE, 1994.

Bu, J., Mesbahi, A., Fazel, M., and Mesbahi, M. Lqr through the lens of first order methods: Discrete-time case. *arXiv preprint arXiv:1907.08921*, 2019.

Cadzow, J. A. Signal enhancement-a composite property mapping algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(1):49–62, 1988.

Cai, J.-F., Qu, X., Xu, W., and Ye, G.-B. Robust recovery of complex exponential signals from random gaussian projections via low rank hankel matrix reconstruction. *Applied and computational harmonic analysis*, 41(2): 470–490, 2016.

Candes, E. J. and Plan, Y. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

Carmon, Y. and Duchi, J. C. Gradient descent efficiently finds the cubic-regularized non-convex newton step. *arXiv preprint arXiv:1612.00547*, 2017.

Cheeger, J. and Ebin, D. G. *Comparison Theorems in Riemannian Geometry*. AMS Chelsea Publishing, Providence, RI, 2008.

Criscitiello, C. and Boumal, N. Efficiently escaping saddle points on manifolds. *arXiv preprint arXiv:1906.04321*, 2019.

De Moor, B., De Gersem, P., De Schutter, B., and Favoreel, W. Daisy: A database for identification of systems. *JOURNAL A*, 38:4–5, 1997.

Dean, S., Mania, H., Matni, N., Recht, B., and Tu, S. On the sample complexity of the linear quadratic regulator. *arXiv preprint arXiv:1710.01688*, 2017.

Dean, S., Tu, S., Matni, N., and Recht, B. Safely learning to control the constrained linear quadratic regulator. In *2019 American Control Conference (ACC)*, pp. 5582–5588. IEEE, 2019.

Ding, T., Sznaier, M., and Camps, O. I. A rank minimization approach to video inpainting. In *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8. IEEE, 2007.

Do Carmo, M. P. *Differential Geometry of Curves and Surfaces*. Courier Dover Publications, 2016.

Du, S. S., Jin, C., Lee, J. D., Jordan, M. I., Singh, A., and Poczos, B. Gradient descent can take exponential time to escape saddle points. In *Advances in Neural Information Processing Systems*, pp. 1067–1077, 2017.

Dullerud, G. E. and Paganini, F. *A course in robust control theory: a convex approach*, volume 36. Springer Science & Business Media, 2013.

Durmus, A., Jiménez, P., Moulines, É., Said, S., and Wai, H.-T. Convergence analysis of riemannian stochastic approximation schemes. *arXiv preprint arXiv:2005.13284*, 2020.

Edelman, A., Arias, T. A., and Smith, S. T. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Elad, M., Milanfar, P., and Golub, G. H. Shape from moments-an estimation theory perspective. *IEEE Transactions on Signal Processing*, 52(7):1814–1829, 2004.

Fazel, M., Pong, T. K., Sun, D., and Tseng, P. Hankel matrix rank minimization with applications to system identification and realization. *SIAM Journal on Matrix Analysis and Applications*, 34(3):946–977, 2013.

Fazel, M., Ge, R., Kakade, S. M., and Mesbahi, M. Global convergence of policy gradient methods for linearized control problems. *arXiv preprint arXiv:1801.05039*, 2018.

Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points – online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pp. 797–842, 2015.

Gillard, J. Cadzow's basic algorithm, alternating projections and singular spectrum analysis. *Statistics and its Interface*, 3(3):335–343, 2010.

Gordon, Y. On milman's inequality and random subspaces which escape through a mesh in r n. In *Geometric aspects of functional analysis*, pp. 84–106. Springer, 1988.

Grossmann, C., Jones, C. N., and Morari, M. System identification with missing data via nuclear norm regularization. In *2009 European Control Conference (ECC)*, pp. 448–453. IEEE, 2009.

Hansson, A., Liu, Z., and Vandenberghe, L. Subspace system identification via weighted nuclear norm optimization. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pp. 3439–3444. IEEE, 2012.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Hazan, E., Lee, H., Singh, K., Zhang, C., and Zhang, Y. Spectral filtering for general linear dynamical systems. In *Advances in Neural Information Processing Systems*, pp. 4634–4643, 2018.

Ho, B. and Kálmán, R. E. Effective construction of linear state-variable models from input/output functions. *at-Automatisierungstechnik*, 14(1-12):545–548, 1966.

Hu, J., Milzarek, A., Wen, Z., and Yuan, Y. Adaptive quadratically regularized Newton method for Riemannian optimization. *SIAM J. Matrix Anal. Appl.*, 39(3):1181–1207, 2018.

Ishteva, M., Absil, P.-A., Van Huffel, S., and De Lathauwer, L. Best low multilinear rank approximation of higher-order tensors, based on the riemannian trust-region scheme. *SIAM Journal on Matrix Analysis and Applications*, 32(1):115–135, 2011.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1724–1732. JMLR. org, 2017a.

Jin, C., Netrapalli, P., and Jordan, M. I. Accelerated gradient descent escapes saddle points faster than gradient descent. *arXiv preprint arXiv:1711.10456*, 2017b.

Kakade, S. M. A natural policy gradient. In *Advances in neural information processing systems*, pp. 1531–1538, 2002.

Kalman, R. E. et al. Contributions to the theory of optimal control. *Bol. soc. mat. mexicana*, 5(2):102–119, 1960.

Karcher, H. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541, 1977.

Kasai, H. and Mishra, B. Inexact trust-region algorithms on riemannian manifolds. In *Advances in Neural Information Processing Systems 31*, pp. 4254–4265. 2018.

Khuzani, M. B. and Li, N. Stochastic primal-dual method on riemannian manifolds with bounded sectional curvature. *arXiv preprint arXiv:1703.08167*, 2017.

Kong, W., Somani, R., Kakade, S., and Oh, S. Robust meta-learning for mixed linear regression with small batches. *arXiv preprint arXiv:2006.09702*, 2020a.

Kong, W., Somani, R., Song, Z., Kakade, S., and Oh, S. Meta-learning for mixed linear regression. *arXiv preprint arXiv:2002.08936*, 2020b.

Krahmer, F., Mendelson, S., and Rauhut, H. Suprema of chaos processes and the restricted isometry property. *Communications on Pure and Applied Mathematics*, 67(11):1877–1904, 2014.

Kučera, V. A method to teach the parameterization of all stabilizing controllers. *IFAC Proceedings Volumes*, 44(1): 6355–6360, 2011.

Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. Gradient descent only converges to minimizers. *Conference on Learning Theory*, pp. 1246–1257, 2016.

Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., and Recht, B. First-order methods almost always avoid saddle points. *arXiv preprint arXiv:1710.07406*, 2017.

Lee, J. M. *Riemannian manifolds : an introduction to curvature*. Graduate texts in mathematics ; 176. Springer, New York, 1997. ISBN 9780387227269.

Lessard, L. and Lall, S. Quadratic invariance is necessary and sufficient for convexity. In *Proceedings of the 2011 American Control Conference*, pp. 5360–5362. IEEE, 2011.

Liu, Z., Hansson, A., and Vandenberghe, L. Nuclear norm system identification with missing inputs and outputs. *Systems & Control Letters*, 62(8):605–612, 2013.

Ljung, L. System identification: theory for the user. *PTR Prentice Hall, Upper Saddle River, NJ*, pp. 1–14, 1999.

Lu, S., Razaviyayn, M., Yang, B., Huang, K., and Hong, M. Snap: Finding approximate second-order stationary solutions efficiently for non-convex linearly constrained problems. *arXiv preprint arXiv:1907.04450*, 2019a.

Lu, S., Zhao, Z., Huang, K., and Hong, M. 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)*, pp. 5356–5360. IEEE, 2019b.

Mangoubi, O., Smith, A., et al. Rapid mixing of geodesic walks on manifolds with positive curvature. *The Annals of Applied Probability*, 28(4):2501–2543, 2018.

Mania, H., Tu, S., and Recht, B. Certainty equivalent control of lqr is efficient. *arXiv preprint arXiv:1902.07826*, 2019.

McCoy, M. B. and Tropp, J. A. The achievable performance of convex demixing. *arXiv preprint arXiv:1309.7478*, 2013.

McKelvey, T., Akçay, H., and Ljung, L. Subspace-based multivariable system identification from frequency response data. *IEEE Transactions on Automatic Control*, 41(7):960–979, 1996.

Mohammadi, H., Zare, A., Soltanolkotabi, M., and Jovanović, M. R. Convergence and sample complexity of gradient methods for the model-free linear quadratic regulator problem. *arXiv preprint arXiv:1912.11899*, 2019.

Mokhtari, A., Ozdaglar, A., and Jadbabaie, A. Escaping saddle points in constrained optimization. *arXiv preprint arXiv:1809.02162*, 2018.

Nouiehed, M., Lee, J. D., and Razaviyayn, M. Convergence to second-order stationarity for constrained non-convex optimization. *arXiv preprint arXiv:1810.02024*, 2018.

Oymak, S. and Ozay, N. Non-asymptotic identification of lti systems from a single trajectory. *arXiv preprint arXiv:1806.05722*, 2018.

Oymak, S., Thrampoulidis, C., and Hassibi, B. Simple bounds for noisy linear inverse problems with exact side information. *arXiv preprint arXiv:1312.0641*, 2013.

Pemantle, R. Nonconvergence to unstable points in urn models and stochastic approximations. *The Annals of Probability*, pp. 698–712, 1990.

Rajeswaran, A., Lowrey, K., Todorov, E. V., and Kakade, S. M. Towards generalization and simplicity in continuous control. In *Advances in Neural Information Processing Systems*, pp. 6550–6561, 2017.

Rapcsák, T. Sectional curvatures in nonlinear optimization. *Journal of Global Optimization*, 40(1-3):375–388, 2008.

Rawlings, J. B., Mayne, D. Q., and Diehl, M. *Model predictive control: theory, computation, and design*, volume 2. Nob Hill Publishing Madison, WI, 2017.

Recht, B., Fazel, M., and Parrilo, P. A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Rotkowitz, M. and Lall, S. A characterization of convex problems in decentralized control. *IEEE transactions on Automatic Control*, 50(12):1984–1996, 2005.

Sakai, T. *Riemannian Geometry*, volume 149 of *Translations of Mathematical Monographs*. American Mathematical Society, 1996.

Sanchez-Pena, R. S. and Sznaier, M. *Robust systems theory and applications*. Wiley-Interscience, 1998.

Sarkar, T. and Rakhlin, A. Near optimal finite time identification of arbitrary linear dynamical systems. *arXiv preprint arXiv:1812.01251*, 2019.

Sarkar, T., Rakhlin, A., and Dahleh, M. A. Finite-time system identification for partially observed lti systems of unknown order. *arXiv preprint arXiv:1902.01848*, 2019.

Sarkar, T. K. and Pereira, O. Using the matrix pencil method to estimate the parameters of a sum of complex exponentials. *IEEE Antennas and Propagation Magazine*, 37(1):48–55, 1995.

Simchowitz, M., Mania, H., Tu, S., Jordan, M. I., and Recht, B. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pp. 439–473, 2018.

Simchowitz, M., Boczar, R., and Recht, B. Learning linear dynamical systems with semi-parametric least squares. *arXiv preprint arXiv:1902.00768*, 2019.

Smith, R. S. Frequency domain subspace identification using nuclear norm minimization and hankel matrix realizations. *IEEE Transactions on Automatic Control*, 59(11):2886–2896, 2014.

Stengel, R. F. *Optimal control and estimation*. Courier Corporation, 1994.

Sun, J., Qu, Q., and Wright, J. Complete dictionary recovery over the sphere ii: Recovery by riemannian trust-region method. *IEEE Transactions on Information Theory*, 63(2):885–914, 2017.

Sun, Y. and Fazel, M. Escaping saddle points efficiently in equality-constrained optimization problems. In *Workshop on Modern Trends in Nonconvex Optimization for Machine Learning, International Conference on Machine Learning*, 2018.

Sun, Y., Flammarion, N., and Fazel, M. Escaping from saddle points on riemannian manifolds. In *Advances in Neural Information Processing Systems*, pp. 7276–7286, 2019.

Sun, Y., Oymak, S., and Fazel, M. Finite sample system identification: Optimal rates and the role of regularization. In *Learning for Dynamics and Control*, pp. 16–25. PMLR, 2020.

Tripuraneni, N., Flammarion, N., Bach, F., and Jordan, M. I. Averaging Stochastic Gradient Descent on Riemannian Manifolds. *arXiv preprint arXiv:1802.09128*, 2018.

Tsiamis, A. and Pappas, G. J. Finite sample analysis of stochastic system identification. *arXiv preprint arXiv:1903.09122*, 2019.

Tu, L. W. *Differential geometry : connections, curvature, and characteristic classes*. Graduate texts in mathematics ; 275. Springer, Cham, Switzerland, 2017. ISBN 9783319550848.

Tu, S., Boczar, R., Packard, A., and Recht, B. Non-asymptotic analysis of robust control from coarse-grained identification. *arXiv preprint arXiv:1707.04791*, 2017.

Van Overschee, P. and De Moor, B. A unifying theorem for three subspace system identification algorithms. *Automatica*, 31(12):1853–1864, 1995.

Van Overschee, P. and De Moor, B. *Subspace identification for linear systems: Theory–Implementation–Applications*. Springer Science & Business Media, 2012.

Verhaegen, M. and Hansson, A. N2sid: Nuclear norm subspace identification of innovation models. *Automatica*, 72: 57–63, 2016.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Wong, Y.-c. Sectional curvatures of Grassmann manifolds. *Proc. Nat. Acad. Sci. U.S.A.*, 60:75–79, 1968.

Xu, W., Yi, J., Dasgupta, S., Cai, J.-F., Jacob, M., and Cho, M. Sep] ration-free super-resolution from compressed measurements is possible: an orthonormal atomic norm minimization approach. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 76–80. IEEE, 2018.

Youla, D., Jabr, H., and Bongiorno, J. Modern wiener-hopf design of optimal controllers–part ii: The multivariable case. *IEEE Transactions on Automatic Control*, 21(3):319–338, 1976.

Zhang, D. and Tajbakhsh, S. D. Riemannian stochastic variance-reduced cubic regularized newton method. *arXiv preprint arXiv:2010.03785*, 2020.

Zhang, H. and Sra, S. First-order methods for geodesically convex optimization. *arXiv:1602.06053*, 2016. *Preprint.*

Zhang, H., Reddi, S. J., and Sra, S. Riemannian svrg: fast stochastic optimization on riemannian manifolds. In *Advances in Neural Information Processing Systems*, pp. 4592–4600, 2016.

Zhang, J. and Zhang, S. A cubic regularized newton's method over riemannian manifolds. *arXiv preprint arXiv:1805.05565*, 2018.

Zhang, K., Hu, B., and Basar, T. Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence. In *Learning for Dynamics and Control*, pp. 179–190, 2020.

Zhou, K., Doyle, J. C., Glover, K., et al. *Robust and optimal control*, volume 40. Prentice hall New Jersey, 1996.

A Appendix of Section 2

Throughout the section we assume that the objective function and the manifold are smooth. Here we list the assumptions that are used in the following lemmas.

Assumption 1 (Lipschitz gradient)**.** *There is a finite constant β such that*

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

Assumption 2 (Lipschitz Hessian)**.** *There is a finite constant ρ such that*

$$\|H(y) - \Gamma_x^y H(x) \Gamma_y^x\|_2 \leq \rho d(x,y) \quad \textit{for all } x, y \in \mathcal{M}.$$

Assumption 3 (Bounded sectional curvature)**.** *There is a finite constant K such that*

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

A.1 Taylor expansions on Riemannian manifold

We provide here the Taylor expansion for functions and gradients of functions defined on a Riemannian manifold.

A.1.1 Taylor expansion for the gradient

For any point $x \in \mathcal{M}$ and $z \in \mathcal{M}$ be a point in the neighborhood of x where the geodesic $\gamma_{x\to z}$ is defined.

$$\begin{aligned}\Gamma_z^x(\mathrm{grad} f(z)) &= \mathrm{grad} f(x) + \nabla_{\gamma'_{x\to z}(0)} \mathrm{grad} f + \textstyle\int_0^1 (\Gamma^x_{\gamma_{x\to z}(\tau)} \nabla_{\gamma'_{x\to z}(\tau)} \mathrm{grad} f - \nabla_{\gamma'_{x\to z}(0)} \mathrm{grad} f) dx_\tau \\ &= \mathrm{grad} f(x) + \nabla_{\gamma'_{x\to z}(0)} \mathrm{grad} f + \Delta(z), \end{aligned} \tag{47}$$

where $\Delta(z) := \int_0^1 (\Gamma^x_{\gamma_{x\to z}(\tau)} \nabla_{\gamma'_{x\to z}(\tau)} \mathrm{grad} f - \nabla_{\gamma'_{x\to z}(0)} \mathrm{grad} f) d\tau$. The Taylor approximation in Eq. (47) is proven by Absil et al. (2009a, Lemma 7.4.7).

A.1.2 Taylor expansion for the function

Taylor expansion of the gradient enables us to approximate the iterations of the main algorithm, but obtaining the convergence rate of the algorithm requires proving that the function value decreases following the iterations. We need to give the Taylor expansion of f with the parallel translated gradient on LHS of Eq. (47). To simplify the notation, let γ denote the $\gamma_{x\to z}$.

$$f(z) - f(x) = \int_0^1 \frac{d}{d\tau} f(\gamma(\tau)) d\tau \tag{48a}$$

$$= \int_0^1 \langle \gamma'(\tau), \mathrm{grad} f(\gamma(\tau)) \rangle d\tau \tag{48b}$$

$$= \int_0^1 \langle \Gamma^x_{\gamma(\tau)} \gamma'(\tau), \Gamma^x_{\gamma(\tau)} \mathrm{grad} f(\gamma(\tau)) \rangle d\tau \tag{48c}$$

$$= \int_0^1 \langle \gamma'(0), \Gamma^0_{\gamma(\tau)} \mathrm{grad} f(\gamma(\tau)) \rangle d\tau \tag{48d}$$

$$= \int_0^1 \langle \gamma'(0), \mathrm{grad} f(x) + \nabla_{\tau\gamma'(0)} \mathrm{grad} f + \Delta(\gamma(\tau)) \rangle d\tau \tag{48e}$$

$$= \langle \gamma'(0), \mathrm{grad} f(x) + \tfrac{1}{2} \nabla_{\gamma'(0)} \mathrm{grad} f + \bar{\Delta}(z) \rangle. \tag{48f}$$

$\Delta(z)$ is defined in Eq. (47). $\bar{\Delta}(z) = \int_0^1 \Delta(\gamma(\tau)) d\tau$. The second line is just rewriting by definition. Eq. (48c) means the parallel translation preserves the inner product (Tu, 2017, Prop. 14.16). Eq. (48d) uses $\Gamma^x_{\gamma(t)} \gamma'(t) = \gamma'(0)$, meaning that the velocity stays constant along a geodesic (Absil et al., 2009a, (5.23)). Eq. (48e) uses Eq. (47). In Euclidean space, the Taylor expansion is

$$f(z) - f(x) = \langle z, \nabla f(x) + \nabla^2 f(x) z + \int_0^1 (\nabla^2 f(\tau z) - \nabla^2 f(x)) z d\tau \rangle. \tag{49}$$

Compare Eq. (48) and Eq. (49), z is replaced by $\gamma'(0) := \gamma'_{x\to z}(0)$ and τz is replaced by $\tau\gamma'_{x\to z}(0)$ or $\gamma_{x\to z}(\tau)$.

Now we have

$$f(u_t) = f(x) + \langle \gamma'(0), \mathrm{grad} f(x) \rangle + \frac{1}{2} H(x)[\gamma'(0), \gamma'(0)] + \langle \gamma'(0), \bar{\Delta}(u_t) \rangle.$$

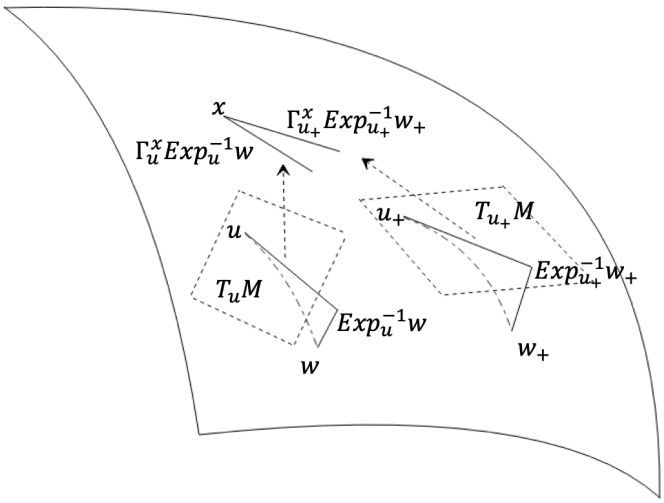


Figure 16: Lemma 12. First map w and w_+ to $\mathcal{T}_u\mathcal{M}$ and $\mathcal{T}_{u_+}\mathcal{M}$, and transport the two vectors to $\mathcal{T}_x\mathcal{M}$, and get their relation.

A.2 Linearization of the iterates in a fixed tangent space

In this section we linearize the progress of the iterates of our algorithm in a fixed tangent space $\mathcal{T}_x\mathcal{M}$. We always assume here that all points are within a region of diameter $R := 12\mathscr{S} \leq \mathfrak{I}$. In the course of the proof we need several auxilliary lemmas which are stated in the last two subsections of this section.

A.2.1 Evolution of $\mathrm{Exp}_u^{-1}(w)$

We first consider the evolution of $\mathrm{Exp}_u^{-1}(w)$ in a fixed tangent space $\mathcal{T}_x\mathcal{M}$. We show in the following lemma that it approximately follows a linear reccursion.

Lemma 12. *Define $\gamma = \sqrt{\hat{\rho}\epsilon}$, $\kappa = \frac{\beta}{\gamma}$, and $\mathscr{S} = \sqrt{\eta\beta}\frac{\gamma}{\hat{\rho}}\log^{-1}(\frac{d\kappa}{\delta})$. Let us consider x be a $(\epsilon, -\sqrt{\hat{\rho}\epsilon})$ saddle point, and define $u^+ = \mathrm{Exp}_u(-\eta \mathrm{grad} f(u))$ and $w^+ = \mathrm{Exp}_w(-\eta \mathrm{grad} f(w))$. Under Assumptions 1, 2, 3, if all pairwise distances between u, w, u^+, w^+, x are less than $12\mathscr{S}$, then for some explicit constant $C_1(K, \rho, \beta)$ depending only on K, ρ, β, there is*

$$\begin{aligned}&\|\Gamma^x_{u^+}\mathrm{Exp}^{-1}_{u^+}(w^+) - (I - \eta H(x))\Gamma^x_u \mathrm{Exp}^{-1}_u(w)\| \\ &\leq C_1(K,\rho,\beta) d(u,w)\left(d(u,w) + d(u,x) + d(w,x)\right).\end{aligned}$$

for some explicit function C_1.

This lemma is illustrated in Fig. 16.

Proof. Denote $-\eta \mathrm{grad} f(u) = v_u$, $-\eta \mathrm{grad} f(w) = v_w$. v is a smooth map. We first prove the following claim.

Claim 1.

$$d(u_+, w_+) \leq c_6(K) d(u,w),$$

where $c_6(K) = c_4(K) + 1 + c_2(K)R^2$.

To show this, note that

$$d(u_+, w_+) \leq d(u_+, \tilde{w}_+) + d(\tilde{w}_+, w_+),$$

and using Lemma 5 with $\tilde{w}_+ = \mathrm{Exp}_w(\Gamma^w_u v_u)$,

$$\begin{aligned}d(\tilde{w}_+, w_+) &= d(\mathrm{Exp}_w(v_w), \mathrm{Exp}_w(\Gamma^w_u v_u)) \\ &\leq (1 + c_2(K)R^2)\|v_w - \Gamma^w_u v_u\| \\ &\leq \beta(1 + c_2(K)R^2) d(u,w).\end{aligned}$$

Using Lemma 5,

$$d(\tilde{w}_+, u_+) \leq c_4(K) d(u, w). \tag{50}$$

Adding the two inequalities proves the claim.

We use now Lemma 3 between (u, w, u_+, w_+) in two different ways. First let us use it for $a = \mathrm{Exp}_u^{-1}(w)$ and $y = \Gamma_w^u v_w$. We obtain:

$$d(w_+, \mathrm{Exp}_u(\mathrm{Exp}_u^{-1}(w) + \Gamma_w^u v_w)) \leq c_1(K) d(u, w)(d(u, w)^2 + \|v_w\|^2). \tag{51}$$

Then we use it for $a = \mathrm{Exp}_u^{-1}(v_u)$ and $y = \Gamma_{u_+}^u \mathrm{Exp}_{u_+}^{-1}(w_+)$ which yields

$$\begin{aligned}
&d(w_+, \mathrm{Exp}_u(v_u + \Gamma_{u_+}^u \mathrm{Exp}_{u_+}^{-1}(w_+))) \\
&\leq c_1(K) d(u_+, w_+)(d(u_+, w_+)^2 + \|v_u\|^2) \\
&\leq c_1(K) c_5(K, \|v_u\|, \|v_w\|) d(u, w) \cdot \Big[c_5(K, \|v_u\|, \|v_w\|)^2 d(u, w)^2 + \|v_u\|^2 \Big].
\end{aligned}$$

Using the triangular inequality we have

$$\begin{aligned}
&d(\mathrm{Exp}_u(\mathrm{Exp}_u^{-1}(w) + \Gamma_w^u v_w), \mathrm{Exp}_u(v_u + \Gamma_{u_+}^u \mathrm{Exp}_{u_+}^{-1}(w_+))) \\
&\leq d(w_+, \mathrm{Exp}_u(\mathrm{Exp}_u^{-1}(w) + \Gamma_w^u v_w)) + d(w_+, \mathrm{Exp}_u(v_u + \Gamma_{u_+}^u \mathrm{Exp}_{u_+}^{-1}(w_+))) \\
&\leq c_7 d(u, w)
\end{aligned}$$

with c_7 defined as

$$c_7 = c_1(K) c_6(K) \cdot [c_5(K, \|v_u\|, \|v_w\|)^2 d(u, w)^2 + \|v_u\|^2 + \|v_w\|^2 \Big].$$

We use again Lemma 4,

$$\|\Gamma_{u_+}^u \mathrm{Exp}_{u_+}^{-1}(w_+)) - \mathrm{Exp}_u^{-1}(w) - [v_u - \Gamma_w^u v_w]\| \leq (1 + c_3(K) R^2) \cdot c_7 d(u, w).$$

Therefore we have linearized the iterate in $T_u\mathcal{M}$. We should see how to transport it back to $T_x\mathcal{M}$. With Lemma 6 we have

$$\|[\Gamma_u^x \Gamma_{u_+}^u - \Gamma_{u_+}^x] \mathrm{Exp}_{u_+}^{-1}(w_+))\| = c_5(K) d(u, x) d(u_+, w_+) \|v_u\|.$$

Note v_u and v_w are $-\eta \mathrm{grad} f(u)$ and $-\eta \mathrm{grad} f(w)$, we define $\nabla v(x)$ the gradient of v, i.e., $-\eta H$. Using Hessian Lipschitz,

$$\begin{aligned}
&\|v_u - \Gamma_w^u v_w + \eta H(u) \mathrm{Exp}_u^{-1}(w)\| \\
&= \|v_u - \Gamma_w^u v_w - \nabla v(u) \mathrm{Exp}_u^{-1}(w)\| \\
&\leq \rho d(u, w)^2,
\end{aligned}$$

and

$$\|\nabla v(u) \mathrm{Exp}_u^{-1}(w) - \Gamma_x^u \nabla v(x) \Gamma_u^x \mathrm{Exp}_u^{-1}(w)\| \leq \rho d(u, w) d(u, x).$$

So we have

$$\begin{aligned}
&\|\Gamma_{u_+}^x \mathrm{Exp}_{u_+}^{-1}(w_+) - (I + \nabla v(x)) \Gamma_u^x \mathrm{Exp}_u^{-1}(w)\| \\
&\leq c_7 d(u, w) + \rho d(u, w)(d(u, w) + d(u, x)) + c_5(K) d(u, x) d(u_+, w_+) \|v_u\| := D_1
\end{aligned} \tag{52}$$

□

A.2.2 Evolution of $\mathrm{Exp}_x^{-1}(w) - \mathrm{Exp}_x^{-1}(u)$

We consider now the evolution of $\mathrm{Exp}_x^{-1}(w) - \mathrm{Exp}_x^{-1}(u)$ in the fixed tangent space $\mathcal{T}_x\mathcal{M}$. We show in the following lemma that it also approximately follows a linear iteration.

Lemma 2. *Define $\gamma = \sqrt{\hat{\rho}\epsilon}$, $\kappa = \frac{\beta}{\gamma}$, and $\mathscr{S} = \sqrt{\eta\beta}\frac{\gamma}{\hat{\rho}}\log^{-1}(\frac{d\kappa}{\delta})$. Let us consider x be a $(\epsilon, -\sqrt{\hat{\rho}\epsilon})$ saddle point, and define $u^+ = \mathrm{Exp}_u(-\eta \mathrm{grad} f(u))$ and $w^+ = \mathrm{Exp}_w(-\eta \mathrm{grad} f(w))$. Under Assumptions 1, 2, 3, if all pairwise distances between u, w, u^+, w^+, x are less than $12\mathscr{S}$, then for some explicit constant $C(K, \rho, \beta)$ depending only on K, ρ, β, there is*

$$\begin{aligned} &\|\mathrm{Exp}_x^{-1}(w^+) - \mathrm{Exp}_x^{-1}(u^+) - (I - \eta H(x))(\mathrm{Exp}_x^{-1}(w) - \mathrm{Exp}_x^{-1}(u))\| \\ &\le C(K,\rho,\beta) d(u,w)\left(d(u,w) + d(u,x) + d(w,x)\right). \end{aligned} \tag{53}$$

This lemma controls the error of the linear approximation of the iterates hen mapped in $\mathcal{T}_x\mathcal{M}$ and largely follows from Lemma 12.

Proof. We have that

$$w = \mathrm{Exp}_x(\mathrm{Exp}_x^{-1}(w)) \tag{54}$$

$$= \mathrm{Exp}_u(\mathrm{Exp}_u^{-1}(w)). \tag{55}$$

Use Eq. (55), let $a = \mathrm{Exp}_x^{-1}(u)$ and $v = \Gamma_u^x \mathrm{Exp}_u^{-1}(w)$, Lemma 3 suggests that

$$\begin{aligned} &d(\mathrm{Exp}_u(\mathrm{Exp}_u^{-1}(w)), \mathrm{Exp}_x(\mathrm{Exp}_x^{-1}(u) + \Gamma_u^x \mathrm{Exp}_u^{-1}(w))) \\ &\le c_1(K)\|\mathrm{Exp}_u^{-1}(w)\|(\|\mathrm{Exp}_u^{-1}(w)\| + \|\mathrm{Exp}_x^{-1}(u)\|)^2. \end{aligned}$$

Compare with Eq. (54), we have

$$\begin{aligned} &d(\mathrm{Exp}_x(\mathrm{Exp}_x^{-1}(w)), \mathrm{Exp}_x(\mathrm{Exp}_x^{-1}(u) + \Gamma_u^x \mathrm{Exp}_u^{-1}(w))) \\ &\le c_1(K)\|\mathrm{Exp}_u^{-1}(w)\|(\|\mathrm{Exp}_u^{-1}(w)\| + \|\mathrm{Exp}_x^{-1}(u)\|)^2 \\ &:= D. \end{aligned} \tag{56}$$

Denote the quantity above by D. Now use Lemma 4

$$\|\mathrm{Exp}_x^{-1}(w) - (\mathrm{Exp}_x^{-1}(u) + \Gamma_u^x \mathrm{Exp}_u^{-1}(w))\| \le (1 + c_3(K)R^2)D.$$

Analogously

$$\|\mathrm{Exp}_x^{-1}(w_+) - (\mathrm{Exp}_x^{-1}(u_+) + \Gamma_{u_+}^x \mathrm{Exp}_{u_+}^{-1}(w_+))\| \le (1 + c_3(K)R^2)D_+$$

where

$$D_+ = c_1(K)\|\mathrm{Exp}_{u_+}^{-1}(w_+)\|(\|\mathrm{Exp}_{u_+}^{-1}(w_+)\| + \|\mathrm{Exp}_x^{-1}(u_+)\|)^2 \tag{57}$$

And we can compare $\Gamma_u^x \mathrm{Exp}_u^{-1}(w)$ and $\Gamma_{u_+}^x \mathrm{Exp}_{u_+}^{-1}(w_+)$ using Eq. (52). In the end we have

$$\begin{aligned} &\|\mathrm{Exp}_x^{-1}(w^+) - \mathrm{Exp}_x^{-1}(u^+) - (I - \eta H(x))(\mathrm{Exp}_x^{-1}(w) - \mathrm{Exp}_x^{-1}(u))\| \\ &\le \|\mathrm{Exp}_x^{-1}(w_+) - (\mathrm{Exp}_x^{-1}(u_+) + \Gamma_{u_+}^x \mathrm{Exp}_{u_+}^{-1}(w_+))\| \\ &\quad + \|\mathrm{Exp}_x^{-1}(w) - (\mathrm{Exp}_x^{-1}(u) + \Gamma_u^x \mathrm{Exp}_u^{-1}(w))\| \\ &\quad + \|\Gamma_{u_+}^x \mathrm{Exp}_{u_+}^{-1}(w_+) - \Gamma_u^x \mathrm{Exp}_u^{-1}(w) - \nabla v(x)\Gamma_u^x \mathrm{Exp}_u^{-1}(w)\| \\ &\quad + \|\nabla v(x)(\Gamma_u^x \mathrm{Exp}_u^{-1}(w) - (\mathrm{Exp}_x^{-1}(w) - \mathrm{Exp}_x^{-1}(u)))\| \\ &\le (1 + c_3(K)R^2)(D_+ + D) + D_1 + \eta\|H(x)\|D. \end{aligned}$$

D, D_+ and D_1 are defined in Eq. (56), Eq. (57) and Eq. (52), they are all order $d(u,w)\big(d(u,w) + d(u,x) + d(w,x)\big)$ so we get the correct order in Eq. (3). □

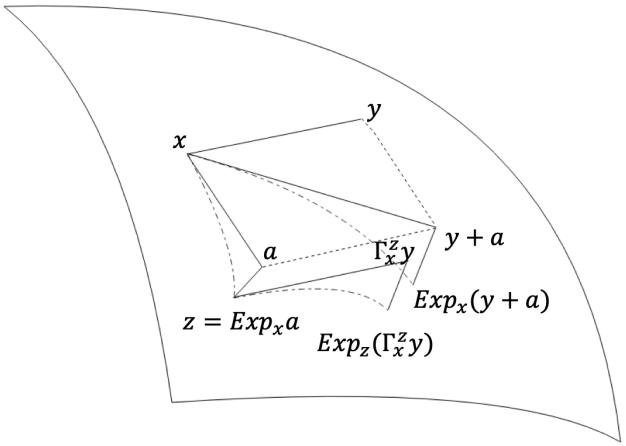


Figure 17: Lemma 3 bounds the difference of two steps starting from x: (1) take $y + a$ step in $\mathcal{T}_x\mathcal{M}$ and map it to manifold, and (2) take a step in $\mathcal{T}_x\mathcal{M}$, map to manifold, call it z, and take $\Gamma_x^z y$ step in $\mathcal{T}_x\mathcal{M}$, and map to manifold. $\mathrm{Exp}_z(\Gamma_x^z y)$ is close to $\mathrm{Exp}_x(y + a)$.

A.2.3 Control of two-steps iteration

In the following lemma we control the distance between the point obtained after moving along the sum of two vectors in the tangent space, and the point obtained after moving a first time along the first vector and then a second time along the transport of the second vector. This is illustrated in Fig. 17.

Lemma 3. *Let $x \in \mathcal{M}$ and $y, a \in T_x\mathcal{M}$. Let us denote by $z = \mathrm{Exp}_x(a)$ then under Assumption 3*

$$d(\mathrm{Exp}_x(y + a), \mathrm{Exp}_z(\Gamma_x^z y)) \leq c_1(K) \min\{\|a\|, \|y\|\}(\|a\| + \|y\|)^2. \tag{58}$$

This lemma which is crucial in the proofs of Lemma 2 and Lemma 12 tightens the result of Karcher (1977, C2.3), which only shows an upper-bound $O(\|a\|(\|a\| + \|y\|)^2)$.

Proof. We adapt the proof of Karcher (1977, Eq. (C2.3) in App C2.2), the only difference being that we bound more carefully the initial normal component. We restate here the whole proof for completeness.

Let $x \in \mathcal{M}$ and $y, a \in T_x\mathcal{M}$. We denote by $\gamma(t) = \mathrm{Exp}_x(ta)$. We want to compare the point $\mathrm{Exp}_x(r(y + a))$ and $\mathrm{Exp}_\gamma(1)(\Gamma_x^{\gamma(1)y})$. These two points , for a fixed r are joined by the curve

$$t \mapsto c(r, t) = \mathrm{Exp}_{\gamma(t)}(r\Gamma_x^{\gamma(t)}(y + (1 - t)a)).$$

We note that $\frac{d}{dt}c(r, t)$ is a Jacobi field along the geodesic $r \mapsto c(r, t)$, which we denote by $J_t(r)$. We importantly remark that the length of the geodesic $r \mapsto c(r, t)$ is bounded as $\|\frac{d}{dr}c(r, t)\| \leq \|y + (1 - t)a\|$. We denote this quantity by $\rho_t = \|y + (1 - t)a\|$. The initial condition of the Jacobi field J_t are given by:

$$J_t(0) = \frac{d}{dt}\gamma(t) = \Gamma_x^{\gamma(t)}a$$
$$\frac{D}{dr}J_t(0) = \frac{D}{dr}\Gamma_x^{\gamma(t)}(y + (1 - t)a) = -\Gamma_x^{\gamma(t)}a.$$

These two vectors are linearly dependent and it is therefore possible to apply Karcher (1977, Proposition A6) to bound J_t^{norm}. Moreover, following Karcher (1977, App A0.3), the tangential component of the Jacobi field is known explicitly, independent of the metric, by

$$J_t^{\tan}(r) = \left(J_t^{\tan}(0) + r\frac{D}{dr}J_t^{\tan}(0)\right)\frac{d}{dr}c(r, t)$$

where the initial conditions of the tangential component of the Jacobi fields are given by $J_t^{\tan}(0) = \langle J_t(0), \frac{\frac{d}{dr}c(r,t)}{\|\frac{d}{dr}c(r,t)\|}\rangle$ and $\frac{D}{dr}J_t^{\tan}(0) = \langle \frac{D}{dr}J_t(0), \frac{\frac{d}{dr}c(r,t)}{\|\frac{d}{dr}c(r,t)\|}\rangle = -J_t^{\tan}(0)$. Therefore

$$J_t^{\tan}(r) = (1 - r)J_t^{\tan}(0)\frac{d}{dr}c(r, t),$$

and $J_t^{\tan}(1) = 0$.

We estimate now the distance $d(\mathrm{Exp}_x(y+a), \mathrm{Exp}_z(\Gamma_x^z y))$ by the length of the curve $t \mapsto c(r,t)$ as follows:

$$d(\mathrm{Exp}_x(y+a), \mathrm{Exp}_z(\Gamma_x^z y)) \leq \int_0^1 \|\frac{d}{dt} c(1,t)\| dt = \int_0^1 \|J_t^{\mathrm{norm}}(1)\| dt,$$

where we use crucially that $J_t^{tan}(1) = 0$.

We utilize (Karcher, 1977, Proposition A.6) to bound $\|J_t^{\mathrm{norm}}(1)\|$ as

$$\|J_t^{\mathrm{norm}}(1)\| \leq \|J_t^{\mathrm{norm}}(0)\|(\cosh(\sqrt{K}\rho_t) - \frac{\sinh(\sqrt{K}\rho_t)}{\sqrt{K}\rho_t})$$

using (Karcher, 1977, Equation (A6.3)) with $\kappa = 0$, $f_\kappa(1) = 0$ and recalling that the geodesics $r \mapsto c(r,t)$ have length ρ_t.

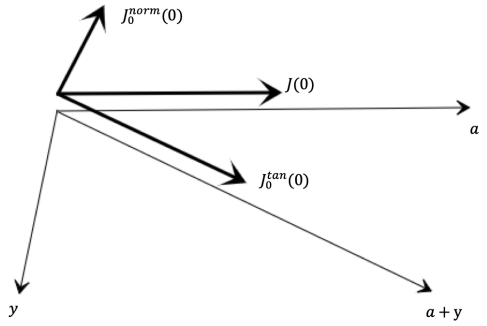


Figure 18: Figure for Lemma 3.

In particular for small value $\|a\| + \|y\|$ we have for some constant $c_1(K)$,

$$\|J_t^{\mathrm{norm}}(1)\| \leq \|J_t^{\mathrm{norm}}(0)\| c_1(K) \rho_t^2.$$

We bound $\|J_t^{\mathrm{norm}}(0)\|$ now. This is the main difference with the original proof of Karcher (1977) who directly bounded $\|J_t^{\mathrm{norm}}(0)\| \leq \|J_t(0)\| = \|a\|$ and $\rho_t \leq \|a\| + \|y\|$. Therefore his proof does not lead to the correct dependence in $\|y\|$.

We have $J_t^0 = \Gamma_x^{\gamma(t)} a$, and the tangential component (velocity of $r \to c(r,t)$) is in the $\Gamma_x^{\gamma(t)}(y + (1-t)a)$ direction. Let $\tilde{z} = \Gamma_x^{\gamma(t)}(y + (1-t)a)$ and $\mathcal{P}_{\tilde{z}^\perp}$ and $\mathcal{P}_{a^\perp}$ denote the projection onto orthogonal complement of $\tilde{z}$ and a.

$$\begin{aligned}
\|J_t^{\mathrm{norm}}(0)\|^2 &= \|\mathcal{P}_{\tilde{z}^\perp}(a)\|^2 \\
&= \|a\|^2 - \frac{(a^T\tilde{z})^2}{\|\tilde{z}\|^2} \\
&= \frac{\|a\|^2}{\|\tilde{z}\|^2}\left(\|\tilde{z}\|^2 - \frac{(a^T\tilde{z})^2}{\|\tilde{z}\|^2}\right) \\
&\leq \frac{\|a\|^2}{\|\tilde{z}\|^2}\|\mathcal{P}_{a^\perp}(\Gamma_x^{\gamma(t)}(y + (1-t)a))\|^2 \\
&\leq \frac{\|a\|^2}{\|\tilde{z}\|^2}\|\mathcal{P}_{a^\perp}(\Gamma_x^{\gamma(t)}((1-t)a)) + \mathcal{P}_{a^\perp}(\Gamma_x^{\gamma(t)} y)\|^2 \\
&= \frac{\|a\|^2}{\|\tilde{z}\|^2}\|\mathcal{P}_{a^\perp}(\Gamma_x^{\gamma(t)} y)\|^2 \\
&\leq \frac{\|a\|^2\|y\|^2}{\|\tilde{z}\|^2}.
\end{aligned}$$

So

$$\begin{aligned}
\|J_t^{\mathrm{norm}}(1)\| &\leq \|J_t^{\mathrm{norm}}(0)\| c_1(K)\rho_t^2 \\
&\leq \frac{\|a\| \cdot \|y\|}{\|\tilde{z}\|} c_1(K)\|\tilde{z}\|^2 \\
&\leq c_1(K)\|a\| \cdot \|y\|(\|a\| + \|y\|),
\end{aligned}$$

and

$$d(\mathrm{Exp}_x(y+a), \mathrm{Exp}_z(\Gamma_x^z y)) \leq c_1(K)\|a\| \cdot \|y\|(\|a\| + \|y\|).$$

□

A.3 Auxilliary lemmas

In the proofs of Lemma 12 and Lemma 2 we needed numerous auxiliary lemmas we are stating here.

We needed the following lemma which shows that both the exponential map and its inverse are Lipschitz.

Lemma 4. *Let $x, y, z \in M$, and the distance of each two points is no bigger than R. Then under Assumption 3*

$$(1 + c_2(K)R^2)^{-1} d(y,z) \leq \|\mathrm{Exp}_x^{-1}(y) - \mathrm{Exp}_x^{-1}(z)\| \leq (1 + c_3(K)R^2) d(y,z).$$

Intuitively this lemma relates the norm of the difference of two vectors of $\mathcal{T}_x\mathcal{M}$ to the distance between the corresponding points on the manifold $\mathcal{M}$ and follows from bounds on the Hessian of the square-distance function (Sakai, 1996, Ex. 4 p. 154).

Proof. The upper-bound is directly proven in Karcher (1977, Proof of Cor. 1.6), and we prove the lower-bound via Lemma 3 in the supplement. Let $b = \mathrm{Exp}_y(\Gamma_x^y(\mathrm{Exp}_x^{-1}(z) - \mathrm{Exp}_x^{-1}(y)))$. Using $d(y,b) = \|\mathrm{Exp}_y^{-1}(b)\|$ and Lemma 3,

$$\begin{aligned} d(y,z) &\leq d(y,b) + d(b, \mathrm{Exp}_x(\mathrm{Exp}_x^{-1}(z))) \\ &\leq \|\mathrm{Exp}_x^{-1}(y) - \mathrm{Exp}_x^{-1}(z)\| \\ &\quad + c_1(K)\|\mathrm{Exp}_x^{-1}(y) - \mathrm{Exp}_x^{-1}(z)\|(\|\mathrm{Exp}_x^{-1}(y) - \mathrm{Exp}_x^{-1}(z)\| + \|\mathrm{Exp}_x^{-1}(y)\|)^2 \end{aligned}$$

□

The following contraction result is fairly classical and is proven using the Rauch comparison theorem from differential geometry (Cheeger & Ebin, 2008).

Lemma 5. *(Mangoubi et al., 2018, Lemma 1) Under Assumption 3, for $x, y \in \mathcal{M}$ and $w \in T_x\mathcal{M}$,*

$$d(\mathrm{Exp}_x(w), \mathrm{Exp}_y(\Gamma_x^y w)) \leq c_4(K) d(x,y).$$

Eventually we need the following corollary of the famous Ambrose-Singer holonomy theorem (Ambrose & Singer, 1953).

Lemma 6. *(Karcher, 1977, Section 6) Under Assumption 3, for $x, y, z \in \mathcal{M}$ and $w \in T_x\mathcal{M}$,*

$$\|\Gamma_y^z \Gamma_x^y w - \Gamma_x^z w\| \leq c_5(K) d(x,y) d(y,z) \|w\|.$$

A.4 Proof of Lemma 7 and 8

In this section we prove two important lemmas from which the proof of our main result mainly comes out. Then we show, in the last subsection, how to combine them to prove this main result.

Lemma 7. *Assume Assumptions 1, 2, 3 hold, and*

$$\epsilon \leq \min\left\{ \frac{\hat{\rho}}{56 \max\{c_2(K), c_3(K)\}\eta\beta} \log\left(\frac{d\beta}{\sqrt{\hat{\rho}\epsilon}\delta}\right), \left(\frac{\mathfrak{I}\hat{\rho}}{12\hat{c}\sqrt{\eta\beta}} \log\left(\frac{d\beta}{\sqrt{\hat{\rho}\epsilon}\delta}\right)\right)^2 \right\} \tag{60}$$

from the main theorem. There exists a constant $c_{\max}$, $\forall \hat{c} > 3, \delta \in (0, \frac{d\kappa}{e}]$, for any u_0 with $d(\tilde{x}, u_0) \leq 2\mathscr{S}/(\kappa \log(\frac{d\kappa}{\delta}))$, $\kappa = \beta/\gamma$.

$$T = \min\left\{ \inf_t \left\{ t | \tilde{f}_{u_0}(u_t) - f(u_0) \leq -3\mathscr{F} \right\}, \hat{c}\mathscr{T} \right\},$$

then $\forall \eta \leq c_{\max}/\beta$, we have $\forall 0 < t < T$, $d(u_0, u_t) \leq 3(\hat{c}\mathscr{S})$.

Lemma 8. *Assume Assumptions 1, 2, 3 and Eq. (60) hold. Take two points u_0 and w_0 which are perturbed from approximate saddle point, where $d(\tilde{x}, u_0) \leq 2\mathscr{S}/(\kappa \log(\frac{d\kappa}{\delta}))$, $\mathrm{Exp}_{\tilde{x}}^{-1}(w_0) - \mathrm{Exp}_{\tilde{x}}^{-1}(u_0) = \mu r e_1$, e_1 is the smallest eigenvector[11] of $H(\tilde{x})$, $\mu \in [\delta/(2\sqrt{d}), 1]$, and the algorithm runs two sequences $\{u_t\}$ and $\{w_t\}$ starting from u_0 and w_0. Denote*

$$T = \min\left\{ \inf_t \left\{ t | \tilde{f}_{w_0}(w_t) - f(w_0) \leq -3\mathscr{F} \right\}, \hat{c}\mathscr{T} \right\},$$

then $\forall \eta \leq c_{\max}/l$, if $\forall 0 < t < T$, $d(\tilde{x}, u_t) \leq 3(\hat{c}\mathscr{S})$, we have $T < \hat{c}\mathscr{T}$.

[11] "smallest eigenvector" means the eigenvector corresponding to the smallest eigenvalue.

A.4.1 Proof of Lemma 7

Suppose $f(u_{t+1}) - f(u_t) \leq -\frac{\eta}{2}\|\text{grad} f(u_t)\|^2$.

$$\begin{aligned}
d(u_{\hat{c}\mathscr{T}}, u_0)^2 &\leq (\sum_0^{\hat{c}\mathscr{T}-1} d(u_{t+1}, u_t))^2 \\
&\leq \hat{c}\mathscr{T} \sum_0^{\hat{c}\mathscr{T}-1} d(u_{t+1}, u_t)^2 \\
&\leq \eta^2 \hat{c}\mathscr{T} \sum_0^{\hat{c}\mathscr{T}-1} \|\text{grad} f(u_t)\|^2 \\
&\leq 2\eta\hat{c}\mathscr{T} \sum_0^{\hat{c}\mathscr{T}-1} f(u_t) - f(u_{t+1}) \\
&= 2\eta\hat{c}\mathscr{T}(f(u_0) - f(u_{\hat{c}\mathscr{T}})) \\
&\leq 6\eta\hat{c}\mathscr{T}\mathscr{F} = 6\hat{c}\mathscr{S}^2.
\end{aligned}$$

A.4.2 Proof of Lemma 8

Note that, for any points inside a region with diameter R, under the assumption of Lemma 8, we have $\max\{c_2(K), c_3(K)\}R^2 \leq 1/2$.

Define $v_t = \text{Exp}_{\tilde{x}}^{-1}(w_t) - \text{Exp}_{\tilde{x}}^{-1}(u_t)$, let $v_0 = e_1$ be the smallest eigenvector of $H(\tilde{x})$, then let $\hat{y}_{2,t}$ be a unit vector, we have

$$\begin{aligned}
v_{t+1} = (I - \eta H(\tilde{x}))v_t + C(K, \rho, \beta)d(u_t, w_t) \\
\cdot (d(u_t, \tilde{x}) + d(w_t, \tilde{x}) + d(\tilde{x}, u_0))\hat{y}_{2,t}.
\end{aligned} \tag{62}$$

Let $C := C(K, \rho, \beta)$. Suppose lemma 8 is false, then $0 \leq t \leq T$, $d(u_t, \tilde{x}) \leq 3\hat{c}\mathscr{S}$, $d(w_t, \tilde{x}) \leq 3\hat{c}\mathscr{S}$, so $d(u_t, w_t) \leq 6\hat{c}\mathscr{S}$, and the norm of the last term in Eq. (62) is smaller than $14\eta C\hat{c}\mathscr{S}\|v_t\|$.

Lemma 4 indicates that

$$\|v_t\| \in [1/2, 2] \cdot d(u_t, w_t) = [3/2, 6] \cdot \hat{c}\mathscr{S}. \tag{63}$$

Let ψ_t be the norm of v_t projected onto e_1, the smallest eigenvector of $H(0)$, and φ_t be the norm of v_t projected onto the remaining subspace. Then Eq. (62) is

$$\begin{aligned}
\psi_{t+1} &\geq (1 + \eta\gamma)\psi_t - \mu\sqrt{\psi_t^2 + \phi_t^2}, \\
\phi_{t+1} &\leq (1 + \eta\gamma)\phi_t + \mu\sqrt{\psi_t^2 + \phi_t^2}.
\end{aligned}$$

Prove that for all $t \leq T$, $\phi_t \leq 4\mu t\psi_t$. Assume it is true for t, we have

$$\begin{aligned}
4\mu(t+1)\psi_{t+1} &\geq 4\mu(t+1) \cdot \left((1 + \eta\gamma)\psi_t - \mu\sqrt{\psi_t^2 + \phi_t^2}\right), \\
\phi_{t+1} &\leq 4\mu t(1 + \eta\gamma)\phi_t + \mu\sqrt{\psi_t^2 + \phi_t^2}.
\end{aligned}$$

So we only need to show that

$$(1 + 4\mu(t+1))\sqrt{\psi_t^2 + \phi_t^2} \leq (1 + \eta\gamma)\psi_t.$$

By choosing $\sqrt{c_{\max}} \leq \frac{1}{56\hat{c}^2}$ and $\eta \leq c_{\max}/\beta$, we have

$$4\mu(t+1) \leq 4\mu T \leq 4\eta C\mathscr{S} \cdot 14\hat{c}^2\mathscr{T} = 56\hat{c}^2\frac{C}{\hat{\rho}}\sqrt{\eta\beta} \leq 1.$$

This gives

$$4(1 + \eta\gamma)\psi_t \geq 2\sqrt{2\psi_t^2} \geq (1 + 4\mu(t+1))\sqrt{\psi_t^2 + \phi_t^2}.$$

Now we know $\phi_t \le 4\mu t\psi_t \le \psi_t$, so $\psi_{t+1} \ge (1+\eta\gamma)\psi_t - \sqrt{2}\mu\psi_t$, and

$$\mu = 14\hat{c}\eta C\mathscr{S} \le 14\hat{c}\sqrt{c_{\max}}\eta\gamma C\log^{-1}(\frac{d\kappa}{\delta})/\hat{\rho} \le \eta\gamma/2,$$

so $\psi_{t+1} \ge (1+\eta\gamma/2)\psi_t$.

We also know that $\|v_t\| \le 6\hat{c}\mathscr{S}$ for all $t \le T$ from Eq. (63), so

$$\begin{aligned}
6\hat{c}\mathscr{S} &\ge \|v_t\| \ge \psi_t \ge (1+\eta\gamma/2)^t\psi_0 \\
&= (1+\eta\gamma/2)^t\frac{\mathscr{S}}{\kappa}\log^{-1}(\frac{d\kappa}{\delta}) \\
&\ge (1+\eta\gamma/2)^t\frac{\delta\mathscr{S}}{2\sqrt{d}\kappa}\log^{-1}(\frac{d\kappa}{\delta}).
\end{aligned}$$

This implies

$$\begin{aligned}
T &< \frac{\log(12\frac{\kappa\sqrt{d}}{\delta}\hat{c}\log(\frac{d\kappa}{\delta}))}{2\log(1+\eta\gamma/2)} \\
&\le \frac{\log(12\frac{\kappa\sqrt{d}}{\delta}\hat{c}\log(\frac{d\kappa}{\delta}))}{\eta\gamma} \\
&\le (2+\log(12\hat{c}))\mathscr{T}.
\end{aligned}$$

By choosing $\hat{c}$ such that $2+\log(12\hat{c}) < \hat{c}$, we have $T \le \hat{c}\mathscr{T}$, which finishes the proof.

A.4.3 Proof of function value decrease at an approximate saddle point

With Lemma 7 and 8 proved, we can lower bound the function value in $O(\mathscr{T})$ iterations decrease by $\Omega(\mathscr{F})$, thus match the convergence rate in the main theorem. Let $T' := \inf_t\left\{t|\tilde{f}_{u_0}(u_t) - f(u_0) \le -3\mathscr{F}\right\}$. Let $\check{}$ denote the operator $\mathrm{Exp}_{u_0}^{-1}(\cdot)$. If $T' \le T$,

$$\begin{aligned}
&f(u_{T'}) - f(u_0) \\
&\le \nabla f(u_0)^T(u_{T'} - u_0) + \frac{1}{2}H(u_0)[\check{u}_{T'} - u_0, \check{u}_{T'} - u_0] \\
&\quad + \frac{\rho}{6}\|\check{u}_{T'} - u_0\|^3 \\
&\le \tilde{f}_{u_0}(u_t) - f(u_0) + \frac{\rho}{2}d(u_0,\tilde{x})\|\check{u}_{T'} - u_0\|^2 \\
&\le -3\mathscr{F} + O(\rho\mathscr{S}^3) \le -2.5\mathscr{F}.
\end{aligned}$$

If $T' > T$, then $\inf_t\left\{t|\tilde{f}_{w_0}(w_t) - f(w_0) \le -3\mathscr{F}\right\} \le T$, and we know $f(w_T) - f(w_0) \le -2.5\mathscr{F}$.

Remark 2. *What is left is bounding the volume of the stuck region, to get the probability of getting out of the stuck region by the perturbation. The procedure is the same as in Jin et al. (2017a). We sample from a unit ball in $\mathcal{T}_x\mathcal{M}$, where x is the approximate saddle point. In Lemma 7 and 8, we study the inverse exponential map at the approximate saddle point x, and the coupling difference between $\mathrm{Exp}_x^{-1}(w)$ and $\mathrm{Exp}_x^{-1}(u)$. The iterates we study and the noise are all in the tangent space $\mathcal{T}_x\mathcal{M}$ which is a Euclidean space, so the probability bound is same as the one in Jin et al. (2017a).*

B Appendix of Section 3

B.1 Sample complexity for MISO and MIMO problems

This section establishes sample complexity bounds for MISO and MIMO systems. The technical argument builds on Cai et al. (2016) and extends their results from from SISO case to MISO. We consider recovering a MISO system impulse response. The system is given in (7), with output size $m = 1$ and the system is order R. For multi-rollout case, we only observe the output at time $2n-1$, and let $u_{2n} = 0$, we have

$$y_{2n-1} = \sum_{i=1}^{2n-2} CA^{2n-2-i}Bu_i + Du_{2n-1}. \tag{68}$$

Denote the impulse response by $h \in \mathbb{R}^{p(2n-1)}$, which is a block vector

$$h = \begin{bmatrix} h^{(1)} \\ h^{(2)} \\ ... \\ h^{(2n-1)} \end{bmatrix}$$

where each block $h^{(i)} \in \mathbb{R}^p$. $\beta \in \mathbb{R}^{p(2n-1)}$ is a weighted version of h, with weights

$$\beta^{(i)} = K_i h^{(i)}$$

and

$$\beta = \begin{bmatrix} \beta^{(1)} \\ \beta^{(2)} \\ ... \\ \beta^{(2n-1)} \end{bmatrix}$$

Define the reweighted Hankel map for the same h by

$$\mathcal{G}(\beta) = \begin{bmatrix} \beta^{(1)}/K_1 & \beta^{(2)}/K_2 & \beta^{(3)}/K_3 & ... \\ \beta^{(2)}/K_2 & \beta^{(3)}/K_3 & \beta^{(4)}/K_2 & ... \\ ... & & & \end{bmatrix}^T \in \mathbb{R}^{n \times pn}$$

and $\mathcal{G}^*$ is the adjoint of $\mathcal{G}$. We define each rollout input $u_1, ..., u_{2n-1}$ as independent Gaussian vectors with

$$u_i \sim \mathcal{N}(0, K_i^2 \mathbf{I}) \tag{69}$$

Now let $\boldsymbol{U} \in \mathbb{R}^{T \times p(2n-1)}$, each entry is iid standard Gaussian. Let $y \in \mathbb{R}^T$ be the concatenation of outputs

$$y = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_T \end{bmatrix}$$

where $y_i \in \mathbb{R}^m$ is defined in (68). We consider the question

$$\begin{aligned} \min_{\beta'} \quad & \|\mathcal{G}(\beta')\|_* \\ \text{s.t.}, \quad & \|\boldsymbol{U}\beta' - y\|_2 \le \delta \end{aligned} \tag{70}$$

where the norm of overall (state and output) noise is bounded by δ. We will present the following theorem, which generalizes the result of Cai et al. (2016) from SISO case to MISO case.

Theorem 10. *Let β be the true impulse response. If $T = \Omega((\sqrt{pR}\log n + \epsilon)^2)$ is the number of output observations, C is some constant, the solution $\hat{\beta}$ to (70) satisfies $\|\beta - \hat{\beta}\|_2 \le 2\delta/\epsilon$ with probability*

$$1 - \exp\left(-\frac{1}{2}(\sqrt{T-1} - C(\sqrt{pR}\log n + \epsilon) - \epsilon)^2\right).$$

When the system output is $y = \boldsymbol{U}\beta + z$ and z is i.i.d. Gaussian noise with variance σ_z^2, we have that $\|\beta - \hat{\beta}\|_2 \lesssim (\sqrt{pR} + \epsilon)\sigma_z \log n$ with probability ((Oymak et al., 2013, Thm 1))

$$1 - 6\exp\left(-\frac{1}{2}(\sqrt{T-1} - C(\sqrt{pR}\log n + \epsilon) - \epsilon)^2\right).$$

This theorem says that when the input dimension is p, the sample complexity is $O(\sqrt{pR}\log n)$. The proof strongly depends on the following lemma (Cai et al. (2016); Gordon (1988)):

Lemma 9. *Define the Gaussian width*[12]

$$w(S) := E_g(\sup_{\gamma \in S} \gamma^T g)$$

where g is standard Gaussian vector of size p. Let $\Phi = \mathcal{I}(\beta) \cap \mathbb{S}$ where $\mathbb{S}$ is unit sphere. We have

$$P(\min_{z \in \Phi} \|\boldsymbol{U}z\|_2 < \epsilon) \leq \exp\left(-\frac{1}{2}(\sqrt{T-1} - w(\Phi) - \epsilon)^2\right).$$

We will present the proof in Appendix B.1.1.

MIMO. For MIMO case, we say output size is m. We take each channel of output as a system of at most order R, and solve m problems

$$\begin{aligned} \boldsymbol{P}_i : \min_{\beta_i} \quad & \|\mathcal{G}(\beta_i)\|_* \\ \text{s.t.,} \quad & \|\boldsymbol{U}x_i - y_i\|_2 \leq \delta, \\ y_i \in \mathbb{R}^T & \text{ is the } i\text{th output.} \end{aligned}$$

and for each problem we have failure probability equal to (72), which means the total failure probability is

$$m\exp\left(-\frac{1}{2}(\sqrt{T-1} - w(\Phi) - \epsilon)^2\right)$$

so we need $T = O((\sqrt{pR}\log n + \log(m) + \epsilon)^2)$. Let the solution to those optimization problems be $[x_1^*, ..., x_m^*]$, and the true impulse response be $[\hat{x}_1, ..., \hat{x}_m]$, then $\|[x_1^*, ..., x_m^*] - [\hat{x}_1, ..., \hat{x}_m]\|_F \leq \sqrt{m}\delta/\epsilon$ with probability

$$1 - \exp\left(-\frac{1}{2}(\sqrt{T-1} - w(\Phi) - \epsilon)^2\right)$$

Another way is that, for each rollout of input data, the output is m dimensional, but we take 1 channel of output from the observation and throw away other $m-1$ output. And we uniformly pick among channels and get T observations for each channel, and in total mT observations/input rollouts. In this case, when the sample complexity is $m\sqrt{pR}\log n$ (m times of before), we can recover the impulse response with Frobenius norm $\sqrt{m}\delta/\epsilon$ with probability

$$1 - \exp\left(-\frac{1}{2}(\sqrt{T/m-1} - w(\Phi) - \epsilon)^2\right)$$

B.1.1 Proof of Theorem 10

Theorem 11. *Let β be the true impulse response. If $T = \Omega((\sqrt{pR}\log(n) + \epsilon)^2)$ is the number of output observations, C is some constant, the solution $\hat{\beta}$ to (70) satisfies $\|\beta - \hat{\beta}\|_2 \leq 2\delta/\epsilon$ with probability*

$$1 - \exp\left(-\frac{1}{2}(\sqrt{T-1} - C(\sqrt{pR}\log(n) + \epsilon) - \epsilon)^2\right).$$

Proof. Let $\mathcal{I}(\beta)$ be the descent cone of $\|\mathcal{G}(\beta)\|_*$ at β, we have the following lemma:

[12]The Gaussian width of the normal cone of (76) and (70) are different up to a constant Banerjee et al. (2014).

Lemma 10. *Assume*

$$\min_{z\in\mathcal{I}(\beta)} \frac{\|\boldsymbol{U}z\|_2}{\|z\|_2} \geq \epsilon,$$

then $\|\beta - \hat{\beta}\|_2 \leq 2\delta/\epsilon$.

(Proof omitted) To prove Theorem 11, we only need lower bound LHS with Lemma 10. The following lemma gives the probability that LHS is lower bounded.

Lemma 11. *Define the Gaussian width*

$$w(S) := E_g(\sup_{\gamma\in S} \gamma^T g) \tag{71}$$

where g *is standard Gaussian vector of size* p. *Let* $\Phi = \mathcal{I}(\beta) \cap \mathbb{S}$ *where* $\mathbb{S}$ *is unit sphere. We have*

$$P(\min_{z\in\Phi} \|\boldsymbol{U}z\|_2 < \epsilon) \leq \exp\left(-\frac{1}{2}(\sqrt{T-1} - w(\Phi) - \epsilon)^2\right). \tag{72}$$

Now we need to study $w(\Phi)$.

Lemma 12. *(Cai et al. (2016) eq. (17)) Let* $\mathcal{I}^*(\beta)$ *be the dual cone of* $\mathcal{I}(\beta)$, *then*

$$w(\Phi) \leq E(\min_{\gamma\in\mathcal{I}^*(\beta)} \|g - \gamma\|_2). \tag{73}$$

Note that $\mathcal{I}^*(\beta)$ is just the cone of subgradient of $\mathcal{G}(\beta)$, so it can be written as

$$\mathcal{I}^*(\beta) = \{\mathcal{G}^*(V_1V_2^T + W) | V_1^T W = 0, WV_2 = 0, \|W\| \leq 1\}$$

where $\mathcal{G}(\beta) = V_1\Sigma V_2^T$ is the SVD of $\mathcal{G}(\beta)$. So

$$\min_{\gamma\in\mathcal{I}^*(\hat{x})} \|g - \gamma\|_2 = \min_{\lambda,W} \|\lambda\mathcal{G}^*(V_1V_2^T + W) - g\|_2.$$

For RHS, we have

$$\begin{aligned}
\|\lambda\mathcal{G}^*(V_1V_2^T + W) - g\|_2 &= \|\lambda\mathcal{G}\mathcal{G}^*(V_1V_2^T + W) - \mathcal{G}(g)\|_F \\
&= \|\lambda(V_1V_2^T + W) - \mathcal{G}(g)\|_F + \|\lambda(I - \mathcal{G}\mathcal{G}^*)(V_1V_2^T + W)\|_F \\
&\leq \|\lambda(V_1V_2^T + W) - \mathcal{G}(g)\|_F.
\end{aligned}$$

Let $\mathcal{P}_W$ be projection operator onto subspace spanned by W, i.e.,

$$\{W | V_1^T W = 0, WV_2 = 0\}$$

and $\mathcal{P}_V$ be projection onto its orthogonal complement. Choose $\lambda = \|\mathcal{P}_W(\mathcal{G}(g))\|$ and $W = \mathcal{P}_W(\mathcal{G}(g))/\lambda$.

$$\begin{aligned}
\|\lambda(V_1V_2^T + W) - \mathcal{G}(g)\|_F &= \|\mathcal{G}(g) - \mathcal{P}_W(\mathcal{G}(g)) - \|\mathcal{P}_W(\mathcal{G}(g))\|V_1V_2^T\|_F \\
&\leq \|\mathcal{P}_V(\mathcal{G}(g)) - \|\mathcal{P}_W(\mathcal{G}(g))\|V_1V_2^T\|_F \\
&\leq \|\mathcal{P}_V(\mathcal{G}(g))\|_F + \|\mathcal{P}_W(\mathcal{G}(g))\|\|V_1V_2^T\|_F \\
&= \|\mathcal{P}_V(\mathcal{G}(g))\|_F + \sqrt{R}\|\mathcal{P}_W(\mathcal{G}(g))\| \\
&= \|\mathcal{P}_V(\mathcal{G}(g))\|_F + \sqrt{R}\|\mathcal{G}(g)\|.
\end{aligned}$$

Bound the first term by (note V_1 and V_2 span R dimensional space, so $V_1 \in \mathbb{R}^{n\times R}$ and $V_2 \in \mathbb{R}^{pn\times R}$)

$$\begin{aligned}
\|\mathcal{P}_V(\mathcal{G}(g))\|_F &= \|V_1V_1^T\mathcal{G}(g) + (I - V_1V_1^T)\mathcal{G}(g)V_2V_2^T\|_F \\
&\leq \|V_1V_1^T\mathcal{G}(g)\|_F + \|\mathcal{G}(g)V_2V_2^T\|_F \\
&\leq 2\sqrt{R}\|\mathcal{G}(g)\|.
\end{aligned}$$

we get

$$\begin{aligned} w(\Phi) &\le E(\min_{\lambda,W} \|\lambda\mathcal{G}^*(V_1V_2^T + W) - g\|_2) \\ &\le E(\|\lambda\mathcal{G}^*(V_1V_2^T + W) - g\|_2)\big|_{\lambda=\|\mathcal{P}_W(\mathcal{G}(g))\|, W=\mathcal{P}_W(\mathcal{G}(g))/\lambda} \\ &\le 3\sqrt{R}\|\mathcal{G}(g)\|. \end{aligned}$$

We know that, if $p = 1$, then $E\|\mathcal{G}(g)\| = O(\log(n))$. For general p, let

$$g^{(i)} = [g_1^{(i)}, ..., g_p^{(i)}]^T,$$

we rearrange the matrix as

$$\begin{aligned} \bar{\mathcal{G}}(g) &= \begin{bmatrix} \begin{bmatrix} g_1^{(1)} & g_1^{(2)}/\sqrt{2} & ... \\ g_1^{(2)}/\sqrt{2} & g_1^{(3)}/\sqrt{3} & ... \\ ... & & \end{bmatrix} & \begin{bmatrix} g_2^{(1)} & g_2^{(2)}/\sqrt{2} & ... \\ g_2^{(2)}/\sqrt{2} & g_2^{(3)}/\sqrt{3} & ... \\ ... & & \end{bmatrix} & ... \end{bmatrix} \\ &= [G_1, ..., G_p] \end{aligned}$$

where expectation of operator norm of each block is $\log(n)$. Then (note v below also has a block structure $[v^{(1)}; ...; v^{(n)}]$)

$$\begin{aligned} \|\bar{\mathcal{G}}(g)\| &= \max_{u,v} \frac{u^T\bar{\mathcal{G}}(g)v}{\|u\|\|v\|} \\ &= \max_{u,v^1,...,v^p} \sum_{i=1}^{p} \frac{u^T G_i v^{(i)}}{\|u\|\|v\|} \\ &\le \max_{v^1,...,v^p} O(\log(n)) \frac{\sum_{i=1}^p \|v^{(i)}\|}{\sqrt{\sum_{i=1}^p \|v^{(i)}\|^2}} \\ &\le O(\sqrt{p}\log(n)). \end{aligned}$$

And $\|\bar{\mathcal{G}}(g)\| = \|\mathcal{G}(g)\|$. So we have $\|\mathcal{G}(g)\| = \sqrt{p}\log(n)$. So $w(\Phi) = C\sqrt{pR}\log(n)$. Get back to (72), we want the probability be smaller than 1, and we get

$$\sqrt{T-1} - C\sqrt{pR}\log n - \epsilon > 0$$

thus $T = O((\sqrt{pR}\log(n) + \epsilon)^2)$.

At the end, we give a different version of Theorem 11. Theorem 11 in Cai et al. (2016) works for the any noise with bounded norm. Here we consider the iid Gaussian noise, and use the result in Oymak et al. (2013), we have the following theorem.

Theorem 12. *Let the system output $y = \boldsymbol{U}\beta + z$ where $\boldsymbol{U}$ entries are iid Gaussian $\mathcal{N}(0, 1/T)$, β is the true system parameter and $z \sim \mathcal{N}(0, \sigma_z^2)$. Then (70) recovers $\hat{\beta}$ with error $\|\hat{\beta} - \beta\|_2 \le w(\Phi)\|z\|_2/\sqrt{T} \lesssim \sqrt{pR}\sigma_z \log n$ with high probability.*

Remark 3. *Since the power of $\boldsymbol{U}$ is n times of that of $\bar{\boldsymbol{U}}$ and the variance of $\boldsymbol{U}$ is $1/T$, $\sigma_z = \sqrt{n/T}\sigma$, we have $\|\hat{h} - h\|_2 \le \|\hat{\beta} - \beta\|_2 \lesssim \sqrt{\frac{pnR}{T}}\sigma \log n$.*

□

B.2 Proof of error of regularized method

Theorem 13. *Consider problem*

$$\hat{h} = \arg\min_{h'} \quad \frac{1}{2}\|\bar{\boldsymbol{U}}h' - y\|_F^2 + \lambda\|\mathcal{H}(h')\|_*. \tag{74}$$

in the MISO (multi-input single-output) setting (m=1, p inputs), the system is order R, $\bar{\boldsymbol{U}} \in \mathbb{R}^{T\times(2n-1)p}$, each row consisting an input rollout $u^{(i)} \in \mathbb{R}^{(2n-1)p}$, and the scaled $\boldsymbol{U}$ has i.i.d Gaussian entries. Let $\boldsymbol{snr} = \mathbb{E}[\|u\|^2/n]/\mathbb{E}[\|z\|^2]$ and $\sigma = 1/\sqrt{\boldsymbol{snr}}$. Let $\lambda = \sigma\sqrt{\frac{pn}{T}}\log(n)$, (74) returns $\hat{h}$ such that

$$\|\hat{h} - h\|_2 \lesssim \|\mathcal{H}(\hat{h} - h)\| \lesssim \begin{cases} \sqrt{\frac{np}{\boldsymbol{snr}\times T}}\log(n) & \text{if} \quad T \gtrsim \min(R^2, n) \\ \sqrt{\frac{Rnp}{\boldsymbol{snr}\times T}}\log(n) & \text{if} \quad R \lesssim T \lesssim \min(R^2, n). \end{cases} \tag{75}$$

We will prove the first case of (75). The second case is a direct application of Theorem 12.

Theorem 14. *We study the problem*

$$\min_{\beta'} \quad \frac{1}{2}\|\boldsymbol{U}\hat{\beta}' - y\|^2 + \lambda\|\mathcal{G}(\hat{\beta}')\|_*, \tag{76}$$

in the MISO (multi-input single-output) setting (m=1, p inputs), where $\boldsymbol{U} \in \mathbb{R}^{T\times(2n-1)p}$. Let β denote the (weighted) impulse response of the true system which has order R, i.e., $\operatorname{rank}(\mathcal{G}(\beta)) = R$, and let $y = \boldsymbol{U}\beta + \xi$ be the measured output, where ξ is the measurement noise. Finally, denote the minimizer of (76) by $\hat{\beta}$. Define

$$\mathcal{J}(\beta) := \left\{ v \;\middle|\; \langle v, \partial(\frac{1}{2}\|\boldsymbol{U}^T\beta - y\|^2 + \lambda\|\mathcal{G}(\beta)\|_*)\rangle \le 0 \right\},$$
$$\Gamma := \|I - \boldsymbol{U}^T\boldsymbol{U}\|_{\mathcal{J}(\beta)},$$

$\mathcal{J}(\beta)$ is the normal cone at β, and Γ is the spectral RSV. If $\Gamma < 1$, $\hat{\beta}$ satisfies

$$\|\mathcal{G}(\hat{\beta} - \beta)\| \le \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\| + \lambda}{1 - \Gamma}.$$

Lemma 13. *Suppose $\xi \sim \mathcal{N}(0, \sigma_\xi I)$, $T \lesssim pR^2\log^2 n$, and $\boldsymbol{U}$ has iid Gaussian entries with $\mathbf{E}(\boldsymbol{U}^\top\boldsymbol{U}) = 1$. Then, we have that $\mathbf{E}(\Gamma) < 0.5$, and $P(\Gamma < 0.5) \ge 1 - O(R\log n\sqrt{p/T})$. In this case $\|\mathcal{G}(\hat{\beta} - \beta)\| \lesssim \sigma_\xi\sqrt{p}\log n$.*

Remark 4. *To be consistent with the main theorem, we need to find the relation between σ_ξ and SNR, or σ. We do the following computation: (1) $\mathcal{G}(\hat{\beta} - \beta) = \mathcal{H}(\hat{h} - h)$, so we are bounding the Hankel spectral norm error here; (2) Each column of the input is unit norm, so each input is $\mathcal{N}(0, 1/T)$, and the average power of input is $1/T$; (3) Because of the scaling matrix K, the actual input of $\bar{\boldsymbol{U}}$ is n times the power of entries in $\boldsymbol{U}$. With all above discussion, we have $\sigma_\xi = \sigma\sqrt{n/T}$, which results in $\|\mathcal{G}(\hat{\beta} - \beta)\| \lesssim \sqrt{\frac{np}{T}}\sigma\log n$.*

Proof. Now we bound $\|\mathcal{G}(\hat{\beta} - \beta)\|$ by partitioning it to $\|\mathcal{G}(I - \boldsymbol{U}^T\boldsymbol{U})(\hat{\beta} - \beta)\|$ and $\|\mathcal{G}(\boldsymbol{U}^T\boldsymbol{U}(\hat{\beta} - \beta))\|$. We have

$$\begin{aligned} \|\mathcal{G}(I - \boldsymbol{U}^T\boldsymbol{U})(\hat{\beta} - \beta)\| &= \|\mathcal{G}(I - \boldsymbol{U}^T\boldsymbol{U})\mathcal{G}^*\mathcal{G}(\hat{\beta} - \beta)\| \\ &\le \|\mathcal{G}(I - \boldsymbol{U}^T\boldsymbol{U})\mathcal{G}^*\|_{2,\mathcal{G}\mathcal{J}(\beta)}\|\mathcal{G}(\hat{\beta} - \beta)\| \\ &= \Gamma\|\mathcal{G}(\hat{\beta} - \beta)\|. \end{aligned} \tag{77}$$

And then we also have

$$\begin{aligned} \|\mathcal{G}(\boldsymbol{U}^T\boldsymbol{U}(\hat{\beta} - \beta))\| &= \|\mathcal{G}\boldsymbol{U}^T(\boldsymbol{U}\hat{\beta} - y + \xi)\| \\ &\le \|\mathcal{G}\boldsymbol{U}^T(\boldsymbol{U}\hat{\beta} - y)\| + \|\mathcal{G}(\boldsymbol{U}^T\xi)\|. \end{aligned}$$

Since $\hat{\beta}$ is the optimizer, we have

$$\boldsymbol{U}^T(\boldsymbol{U}\hat{\beta} - y) + \lambda\mathcal{G}^*(\hat{V}_1\hat{V}_2^T + \hat{W}) = 0,$$

where $\mathcal{G}(\hat{\beta}) = \hat{V}_1\hat{\Sigma}\hat{V}_2^T$ is the SVD of $\mathcal{G}(\hat{\beta})$, $\hat{W} \in \mathbb{R}^{n\times n}$ where $\hat{V}_1^T\hat{W} = 0, \hat{W}\hat{V}_2 = 0, \|\hat{W}\| \le 1$. We have

$$\|\mathcal{G}(\boldsymbol{U}^T\boldsymbol{U}(\hat{\beta} - \beta))\| \le \|\mathcal{G}(\boldsymbol{U}^T\xi)\| + \lambda. \tag{78}$$

Combining (77) and (78), we have

$$\begin{aligned} \|\mathcal{G}(\hat{\beta} - \beta)\| &\le \|\mathcal{G}(I - \boldsymbol{U}^T\boldsymbol{U})(\hat{\beta} - \beta)\| + \|\mathcal{G}(\boldsymbol{U}^T\boldsymbol{U}(\hat{\beta} - \beta))\| \\ &\le \Gamma\|\mathcal{G}(\hat{\beta} - \beta)\| + \|\mathcal{G}(\boldsymbol{U}^T\xi)\| + \lambda \end{aligned}$$

or equivalently,

$$\|\mathcal{G}(\hat{\beta}-\beta)\| \leq \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|+\lambda}{1-\Gamma}, \quad \Gamma = \|\mathcal{G}(I-\boldsymbol{U}^T\boldsymbol{U})\mathcal{G}^*\|_{2,\mathcal{G}\mathcal{J}(\beta)}.$$

Bounding Γ. Denote the SVD of $\mathcal{G}(\beta) = V_1\Sigma V_2^T$. Denote projection operators $\mathcal{P}_V(M) = V_1V_1^TM + MV_2V_2^T - V_1V_1^TMV_2V_2^T$ and $\mathcal{P}_W(M) = M - \mathcal{P}_V(M)$. First we prove some side results for later use. From optimality of $\hat{\beta}$, we have

$$\begin{aligned}
& \frac{1}{2}\|y-\boldsymbol{U}\hat{\beta}\|^2 + \lambda\|\mathcal{G}\hat{\beta}\|_* \leq \frac{1}{2}\|y-\boldsymbol{U}\beta\|^2 + \lambda\|\mathcal{G}\beta\|_* = \frac{1}{2}\|\xi\|^2 + \lambda\|\mathcal{G}\beta\|_* \\
\Rightarrow \quad & \frac{1}{2}\|\boldsymbol{U}\beta+\xi-\boldsymbol{U}\hat{\beta}\|^2 + \lambda\|\mathcal{G}\hat{\beta}\|_* \leq \frac{1}{2}\|\xi\|^2 + \lambda\|\mathcal{G}\beta\|_* \\
\Rightarrow \quad & \frac{1}{2}\|\boldsymbol{U}(\beta-\hat{\beta})\|^2 + \xi^T\boldsymbol{U}(\beta-\hat{\beta}) + \lambda\|\mathcal{G}\hat{\beta}\|_* \leq \lambda\|\mathcal{G}\beta\|_* \\
\Rightarrow \quad & \lambda\|\mathcal{G}\hat{\beta}\|_* \leq \lambda\|\mathcal{G}\beta\|_* + \xi^T\boldsymbol{U}(\hat{\beta}-\beta) \\
\Rightarrow \quad & \|\mathcal{G}\hat{\beta}\|_* - \|\mathcal{G}\beta\|_* \leq \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|}{\lambda}\|\mathcal{G}(\hat{\beta}-\beta)\|_*
\end{aligned} \tag{79}$$

(79) is an important result to note, and following that,

$$\begin{aligned}
& \|\mathcal{G}\hat{\beta}\|_* - \|\mathcal{G}\beta\|_* \leq \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|}{\lambda}\|\mathcal{G}(\hat{\beta}-\beta)\|_* \\
\Rightarrow \quad & \langle \mathcal{G}(\hat{\beta}-\beta), V_1V_2^T + W\rangle \leq \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|}{\lambda}\|\mathcal{G}(\hat{\beta}-\beta)\|_* \\
\Rightarrow \quad & \|\mathcal{P}_W\mathcal{G}(\hat{\beta}-\beta)\|_* \leq -\langle \mathcal{G}(\hat{\beta}-\beta), V_1V_2^T\rangle + \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|}{\lambda}\|\mathcal{G}(\hat{\beta}-\beta)\|_* \\
\Rightarrow \quad & \|\mathcal{P}_W\mathcal{G}(\hat{\beta}-\beta)\|_* \leq \|\mathcal{P}_V\mathcal{G}(\hat{\beta}-\beta)\|_* + \frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|}{\lambda}(\|\mathcal{P}_V\mathcal{G}(\hat{\beta}-\beta)\|_* + \|\mathcal{P}_W\mathcal{G}(\hat{\beta}-\beta)\|_*) \\
\Rightarrow \quad & \|\mathcal{P}_W\mathcal{G}(\hat{\beta}-\beta)\|_* \leq \frac{1+\frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|}{\lambda}}{1-\frac{\|\mathcal{G}(\boldsymbol{U}^T\xi)\|}{\lambda}}\|\mathcal{P}_V\mathcal{G}(\hat{\beta}-\beta)\|_*
\end{aligned} \tag{80}$$

Let $\boldsymbol{U}$ be iid Gaussian matrix with scaling $\boldsymbol{E}(\boldsymbol{U}^T\boldsymbol{U}) = I$. Here we need to study the Gaussian width of the normal cone $w(\mathcal{J}(\beta))$ of (76). Banerjee et al. (2014) proves that, if (79) is true, and $\lambda \geq 2\|\mathcal{G}(\boldsymbol{U}^T\xi)\|$, then the Gaussian width of this set (intersecting with unit ball) is less than 3 times of Gaussian width of $\{\hat{\beta} : \|\mathcal{G}(\hat{\beta})\|_* \leq \|\mathcal{G}(\beta)\|_*\}$, which is $O(\sqrt{R}\log n)$ Cai et al. (2016).
A simple bound is that, let $\delta = \hat{\beta}-\beta$, Γ can be replaced by

$$\max \|\mathcal{G}((I-\boldsymbol{U}^T\boldsymbol{U})\delta)\|/\|\mathcal{G}(\delta)\|$$

subject to $\hat{\beta} \in \mathcal{J}(\beta)$. With (80), we have $\|\mathcal{P}_W\mathcal{G}(\delta)\|_* \leq 3\|\mathcal{P}_V\mathcal{G}(\delta)\|_*$.
Denote $\sigma = \|\mathcal{G}(\delta)\|$, we know that $\sigma \geq \max\{\|\mathcal{P}_W\mathcal{G}(\delta)\|, \|\mathcal{P}_V\mathcal{G}(\delta)\|\}$ and $\|\mathcal{P}_V\mathcal{G}(\delta)\| \geq \|\mathcal{P}_V\mathcal{G}(\delta)\|_*/(2R)$. And simple algebra gives that

$$\max_{0<\sigma_i<\sigma, \sum_i \sigma = S} \sum_i \sigma_i^2 \leq S\sigma.$$

So let σ_i be singular values of $\mathcal{P}_V\mathcal{G}(\delta)$ or $\mathcal{P}_W\mathcal{G}(\delta)$, and $S = \|\mathcal{P}_V\mathcal{G}(\delta)\|_*$ or $\|\mathcal{P}_W\mathcal{G}(\delta)\|_*$,

$$\frac{\sigma}{\|\mathcal{P}_V\mathcal{G}(\delta)\|_F} \geq \sqrt{\frac{\|\mathcal{P}_V\mathcal{G}(\delta)\|_*}{2R\|\mathcal{P}_V\mathcal{G}(\delta)\|_*}} \geq \sqrt{1/2R}$$

$$\frac{\sigma}{\|\mathcal{P}_W\mathcal{G}(\delta)\|_F} \geq \sqrt{\frac{\|\mathcal{P}_V\mathcal{G}(\delta)\|_*}{2R\|\mathcal{P}_W\mathcal{G}(\delta)\|_*}} \geq \sqrt{1/6R}$$

the second last inequality comes from (80). Thus if $\|(I-\boldsymbol{U}^T\boldsymbol{U})\delta\| = O(1/\sqrt{R})\|\delta\|$, in other words, $\|\mathcal{G}((I-\boldsymbol{U}^T\boldsymbol{U})\delta)\|_F = O(1/\sqrt{R})\|\mathcal{G}(\delta)\|_F$, whenever δ in normal cone, we have

$$\|\mathcal{G}((I-\boldsymbol{U}^T\boldsymbol{U})\delta)\| \leq \|\mathcal{G}((I-\boldsymbol{U}^T\boldsymbol{U})\delta)\|_F \leq O(1/\sqrt{R})\|\mathcal{G}(\delta)\|_F \leq \|\mathcal{G}(\delta)\| \tag{81}$$

so $\Gamma < 1$. To get this, we need $\sqrt{T}/w(\mathcal{J}(\beta)) = O(\sqrt{R})$ where $T = O(pR^2 \log^2 n)$ (Vershynin, 2018, Thm 9.1.1), still not tight in R, but $O(\min\{n, R^2 \log^2 n\})$ is as good as Oymak & Ozay (2018) and better than Sarkar et al. (2019), which are $O(n)$ and $O(n^2)$ correspondingly. (Vershynin, 2018, Thm 9.1.1) is a bound in expectation, but it naively turns into high probability bound since $\Gamma \geq 0$. □

B.3 Bounding Γ, where do we lose?

The previous proof is not tight here.

$$\underbrace{\|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\| \leq \|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\|_F}_{\text{not tight}} \leq O(1/\sqrt{R})\|\mathcal{G}(\delta)\|_F \leq \|\mathcal{G}(\delta)\| \tag{82}$$

If we can show that, for all δ in the tangent cone (thus independent of $\boldsymbol{U}$), $\|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\| = O(1/\sqrt{R})\|\mathcal{G}((I - \boldsymbol{U}^T\boldsymbol{U})\delta)\|_F$ for $\boldsymbol{U} \in \mathbb{R}^{O(R\log^2 n)\times n}$, then we can get the correct sample complexity. The difficulty is that, we do not know the distribution of $(I - \boldsymbol{U}^T\boldsymbol{U})\delta$. Let $M = I - \boldsymbol{U}^T\boldsymbol{U}$ and $g := M\delta$. Let $\tilde{g}$ be a Gaussian vector with same mean and covariance as g that will be studied later. We know that $g_i = \sum M_{ij}\delta_j$. Let $z_{ij} = U_{:,i}^T U_{:,j}$, u, v denote standard Gaussian vectors of dimension T, we have (the last equation: $i \neq j$)

$$\begin{aligned}
E((1 - z_{ii}^2)^2) &= E((1 - \frac{1}{T}u^T u)^2) \\
&= 1 - \frac{2}{T}\sum_{i=1}^{T} E(u_i^2) + \frac{1}{T^2}(\sum_{i=1}^{T} E(u_i^4) + \sum_{i\neq j}^{T} E(u_i^2 u_j^2)) = \frac{2}{T}. \\
E(z_{ij}^2) &= E((\frac{1}{T}u^T v)^2) \\
&= \frac{1}{T^2}E(\sum u_i^2 v_i^2) = \frac{1}{T}. \\
E(g_i) &= 0, \\
E(g_i^2) &= E((\sum M_{ij}\delta_j)^2) \\
&= \delta_i^2 E((1 - z_{ii}^2)^2) + \sum_{j\neq i}\delta_j^2 E(z_{ij}^2) + \sum_{j\neq k}\delta_j\delta_k E(M_{ij}M_{ik}) \\
&\leq \frac{1}{T}(\delta_i^2 + \|\delta\|^2). \\
E(g_i g_j) &= E((\sum M_{ik}\delta_k)(\sum M_{jl}\delta_l)) \\
&= \delta_i\delta_j E(M_{ij}M_{ji}) \\
&= \frac{1}{T}\delta_i\delta_j.
\end{aligned}$$

So

$$Cov(g) = \frac{1}{T}(\|\delta\|^2 I + \delta\delta^T).$$

The problem is that g is not Gaussian so even we know mean and variance it's still hard to deal with. Let's study Gaussian first. If $\tilde{g} = \tilde{g}_1 + \check{g}_2\delta$ where $\tilde{g}_1 \sim \mathcal{N}(0, \frac{\|\delta\|^2}{T}I)$ and $\check{g}_2 \sim \mathcal{N}(0, 1/T)$, then we have

$$\begin{aligned}
E(\|\mathcal{G}(\tilde{g})\|) &\leq E(\|\mathcal{G}(\tilde{g}_1)\|) + E(|\check{g}_2|\|\mathcal{G}(\delta)\|) \\
&\leq \frac{1}{\sqrt{T}}(\|\delta\|\frac{\log n}{\sqrt{n}} + \|\mathcal{G}(\delta)\|) \\
&\leq \frac{1}{\sqrt{T}}(\underbrace{\frac{\sqrt{R}\log n}{\sqrt{n}}}_{\text{proven before}} + 1)\|\mathcal{G}(\delta)\| \\
&\leq \frac{2}{\sqrt{T}}\|\mathcal{G}(\delta)\|.
\end{aligned}$$

If we have

$$P(\|\mathcal{G}(\tilde{g})\| > \alpha E(\|\mathcal{G}(\tilde{g})\|)) \le \psi(\alpha),$$

then let $\alpha = \sqrt{T}/2$, we have

$$P(\|\mathcal{G}(\tilde{g})\| > E(\|\mathcal{G}(\delta)\|)) \le \psi(\sqrt{T}/2)$$

We hope that $\psi(\alpha) = \exp(-O(\alpha^2))$ or $\log(\psi(\alpha)) = -O(\alpha^2)$. Then with a set of Gaussian width $\sqrt{R}\log n$, we use a union bound and have (if we ignore the difference between g and $\tilde{g}$)

$$P(\max_{\delta} \|\mathcal{G}(g)\| > \|\mathcal{G}(\delta)\|) \le \psi(\sqrt{T}/2)\exp(O(R\log^2 n)) = \exp(O(R\log^2 n) + \log(\psi(\sqrt{T}/2))).$$

So if the derivation of a Gaussian vector can be applied to a non-Gaussian $g = (I - \boldsymbol{U}^T\boldsymbol{U})\delta$ with the same mean and variance, and $\|\mathcal{G}(g)\|$ is a subGaussian random variable, then we can get the tight bound.

B.4 Proof of suboptimal recovery guarantee with i.i.d. input

Theorem 15. *Suppose the system impulse response is h such that $h_t = 1,\ \forall t \ge 1$, which is order 1. The Gaussian width of the set*

$$\{x \mid \|\mathcal{H}(h + x)\|_* \le \|\mathcal{H}(h)\|_*\} \cap \mathbb{S}$$

is lower bounded by $Cn^{1/6}$ for some constant C.

Proof. We consider the Gaussian width $w(\Phi)$ defined in this specific case.

Let $V = \frac{1}{n}\mathbf{1}\mathbf{1}^T$, and

$$\mathcal{I}^*(h) = \{\mathcal{H}^*(V + W) | \mathbf{1}^T W = 0, W\mathbf{1} = 0, \|W\| \le 1\}$$

we have[13]

$$w(\Phi) = E(\min_{\lambda, W} \|\lambda\mathcal{H}^*(V + W) - g\|_2).$$

In the instance, $V = \frac{1}{n}\mathbf{1}\mathbf{1}^T$. and we take W such that $\|W\| \le 1$ and $W\mathbf{1} = W^T\mathbf{1} = 0$.

First, we note that

$$\begin{aligned}
&E(\min_{\lambda, W} \|\lambda\mathcal{H}^*(V + W) - g\|_2) \\
&= \frac{1}{2}\left(E(\min_{\lambda, W} \|\lambda\mathcal{H}^*(V + W) - g\|_2 \mid \mathbf{1}^T g \le 0)\right. \\
&\quad \left. + E(\min_{\lambda, W} \|\lambda\mathcal{H}^*(V + W) - g\|_2 \mid \mathbf{1}^T g > 0)\right) \\
&\ge \frac{1}{2}E(\min_{\lambda, W} \|\lambda\mathcal{H}^*(V + W) - g\|_2 \mid \mathbf{1}^T g \le 0). \qquad (83)
\end{aligned}$$

Proof strategy: Based on the previous derivation, we focus on the case when $\mathbf{1}^T g \le 0$. Denote $z = \lambda\mathcal{H}^*(V+W) - g$, and the vector $z_{1:k}$ is the first 1 to k entries of z. Then we prove that

$$(1)\ \lambda \le \|z\|_2/\sqrt{n}, \quad (2)\ \|z_{1:1/\lambda}\|_2 = \Omega(\lambda^{-1/2}).$$

Then we have

$$\|z\|_2 = \Omega(\|z_{1:1/\lambda}\|_2) \overset{2}{=} \Omega(\lambda^{-1/2}) \overset{1}{=} \Omega((\|z\|_2/\sqrt{n})^{-1/2})$$

which suggests $\|z\|_2 = \Omega(n^{1/6})$.

[13] We slightly change the definition of Gaussian width. We refer readers to (McCoy & Tropp, 2013, Thm 1). It is known to be as tight and the probability of failure is order constant if the number of measurements is smaller than order square of the quantity.

Lemma 14. *Let g be a standard Gaussian vector of size $2n-1$ conditioned on $\mathbf{1}^T g \le 0$. Let $z = \lambda \mathcal{H}^*(V+W) - g$ where $V = \frac{1}{n}\mathbf{1}\mathbf{1}^T$, and $W\mathbf{1} = W^T\mathbf{1} = 0$, $\|W\| \le 1$. Then we have that $\lambda \le \|z\|_2/\sqrt{n}$.*

We observe that $\mathbf{1}^T\mathcal{H}^*(X)$ is the summation of every entry in X for any matrix X. Thus $\mathbf{1}^T\mathcal{H}^*(W) = 0$ since $W\mathbf{1} = 0$. Conditioned on $\mathbf{1}^T g \le 0$, we have

$$\mathbf{1}^T(\lambda\mathcal{H}^*(V+W) - g) \ge \lambda \mathbf{1}^T\mathcal{H}^*(V) = \lambda n.$$

And so that $\|\lambda\mathcal{H}^*(V+W) - g\|_2 \ge \lambda\sqrt{n}$. Then $\|z\|_2/\sqrt{n} \ge \lambda$, we have proven the first point.

Lemma 15. *Let g be a standard Gaussian vector of size $2n-1$ conditioned on $\mathbf{1}^T g \le 0$. Let $z = \lambda \mathcal{H}^*(V+W) - g$ where $V = \frac{1}{n}\mathbf{1}\mathbf{1}^T$, and $W\mathbf{1} = W^T\mathbf{1} = 0$, $\|W\| \le 1$. Let the vector $z_{1:k}$ is the first 1 to k entries of z.Then we have that $\|z_{1:1/\lambda}\|_2 = \Omega(\lambda^{-1/2})$.*

If $\|z\|_2 \le \sqrt{n}$, we observe $z_{1:\sqrt{n}/\|z\|_2}$. When $i \le \sqrt{n}/\|z\|_2$, the i-th entry of $\mathcal{H}^*(V+W)$, denoted as $(\mathcal{H}^*(V+W))_i$, is summation of $2i$ terns in V and W. Since these two matrices have bounded spectral norm 1, then every entry of V is $1/n$ and every entry of W is no bigger than 1. So

$$\begin{aligned} z_i &= \lambda(\mathcal{H}^*(V+W))_i - g_i \\ &\in \pm(1+1/n)i\lambda - g_i \\ &\in \pm\frac{(1+1/n)i\|z\|_2}{\sqrt{n}} - g_i. \end{aligned}$$

Thus

$$\begin{aligned} \|z_{1:\sqrt{n}/\|z\|_2}\|_2 &\ge -\frac{(1+1/n)\|z\|_2}{\sqrt{n}}\|[1,2,...,\sqrt{n}/\|z\|_2]\|_2 + \|g_{1:\sqrt{n}/\|z\|_2}\|_2 \\ &\ge -\frac{(1+1/n)n^{1/4}}{\sqrt{3}\|z\|_2^{1/2}} + \frac{n^{1/4}}{\|z\|_2^{1/2}}. \end{aligned}$$

Note that the first term is smaller than the second, so we have

$$\|z_{1:\sqrt{n}/\|z\|_2}\|_2 \ge C_1\frac{n^{1/4}}{\|z\|_2^{1/2}}$$

for some constant C_1. Note this is the norm of a part of z, which is smaller than the norm of z, so we have

$$\frac{C_1 n^{1/4}}{\|z\|_2^{1/2}} \le \|z\|_2$$

so that $\|z\|_2 = \Omega(n^{1/6})$, and we have bounded the quantity (83). □

B.5 Proof of least square spectral norm error

Theorem 16. *Denote the discrete Fourier transform matrix by F. Denote $z_{(i)} \in \mathbb{R}^T, i = 1,...,m$ as the noise that corresponds to each dimension of output. The solution $\hat{h}$ of*

$$\hat{h} := h + \bar{\boldsymbol{U}}^\dagger z = \min_{h'} \ \frac{1}{2}\|\bar{\boldsymbol{U}}h' - y\|_F^2. \tag{84}$$

obeys

$$\begin{aligned} \|\hat{h} - h\|_F &\le \|z\|_F/\sigma_{\min}(\bar{\boldsymbol{U}}) \\ \|\mathcal{H}(\hat{h} - h)\| &\le \left\| \left[\|F\bar{\boldsymbol{U}}^\dagger z_{(1)}\|_\infty, ..., \|F\bar{\boldsymbol{U}}^\dagger z_{(m)}\|_\infty\right] \right\|. \end{aligned}$$

Proof. First we clarify the notation here. In regularization part, we only consider the MISO system, whereas we can prove the bound for MIMO system as well in least square. Here we assume the input is p dimension and output is m dimension, at each time. For the notation in (84), $\bar{\boldsymbol{U}} \in \mathbb{R}^{T\times(2n-1)p}$, whose each row is the input in a time interval of length $2n-1$. The impulse response is $h \in \mathbb{R}^{(2n-1)p\times m}$ and output and noise are $y, z \in \mathbb{R}^{T\times m}$, where each column

corresponds to one channel of the output. Each row of y is an output observation at a single time point. $z_{(i)} \in \mathbb{R}^T$ is a column of the noise, meaning one channel of the noise contaminating all observations at this channel.

(84) has close form solution and we have $\|\hat{h} - h\| = \|\bar{\boldsymbol{U}}^\dagger z\| \le \|z\|/\sigma_{\min}(\bar{\boldsymbol{U}})$. To get the error bound in Hankel matrix, we can denote $\bar{z} = \bar{\boldsymbol{U}}^\dagger z = (\bar{\boldsymbol{U}}^T\bar{\boldsymbol{U}})^{-1}\bar{\boldsymbol{U}}^T z$, and

$$H_{\bar{z}} = \begin{bmatrix} \bar{z}_1 & \bar{z}_2 & \dots & \bar{z}_{2n-1} \\ \bar{z}_2 & \bar{z}_3 & \dots & \bar{z}_1 \\ \dots & & & \\ \bar{z}_{2n-1} & \bar{z}_1 & \dots & \bar{z}_{2n-2} \end{bmatrix}.$$

If $m = 1$, $\bar{z} \in \mathbb{R}^{(2n-1)p}$ is a vector (Krahmer et al., 2014, Section 4) proves that

$$H_{\bar{z}} = F^{-1}\mathrm{diag} F\bar{z} F.$$

So the spectral norm error is bounded by $\|\mathrm{diag} F\bar{z}\|_2 = \|F\bar{z}\|_\infty$.

If $m > 1$, all columns of z are independent, so $H_{\bar{z}}$ can be seen as concatenation of m independent noise matrices where each satisfies the previous derivation. □

Theorem 17. *Denote the solution to (84) as $\hat{h}$. Let $\bar{\boldsymbol{U}} \in \mathbb{R}^{T\times(2n-1)p}$ is multiple rollout input, where every entry is i.i.d. Gaussian random variable, y be the corresponding output and z is i.i.d. Gaussian matrix with each entry has mean 0 and variance σ_z, then the spectral norm error is $\|\mathcal{H}(\hat{h} - h)\| \lesssim \sigma_z\sqrt{\frac{mnp}{T}}\log(np)$.*

Proof. We use Theorem 16. First let $m = 1$. The covariance of $F\bar{z} = F\bar{\boldsymbol{U}}^\dagger z$ is $F(\bar{\boldsymbol{U}}^T\bar{\boldsymbol{U}})^{-1}F^T$. If $T = \tilde{\Omega}(n)$, it's proven Vershynin (2018) that $\frac{TI}{2} \preceq \bar{\boldsymbol{U}}^T\bar{\boldsymbol{U}} \preceq \frac{3TI}{2}$ then $\frac{n}{2T}I \preceq F(\bar{\boldsymbol{U}}^T\boldsymbol{U})^{-1}F^T \preceq \frac{3n}{2T}I$. So $\|F\bar{z}\|_\infty$ should scale as $O(\sigma_z\sqrt{\frac{n}{T}}\log n)$. So $\|\mathcal{H}(\bar{z})\|_2 \le \|H_{\bar{z}}\|_2 \le \|F\bar{z}\|_\infty = O(\sigma_z\sqrt{\frac{n}{T}}\log n)$. If $m > 1$, then by concatenation we simply bound the specral norm by m times MISO case. When $m > 1$, with previous discussion of concatenation, and each submatrix to be concatenated has the same distribution, so the spectral norm error is at most $\sqrt{m}$ times larger. □

B.6 Proof of model selection method

Theorem 18. *With all assumptions in Theorem 4, suppose the optimal λ that achieves the error bound in 4 is in Λ, with probability at least $1 - P$, Algorithm 3 has the error at most $\frac{a_2}{a_1}$ times of the error as in Theorem 4, i.e.,*

$$\frac{\|\hat{h} - h\|_2}{\sqrt{2}} \le \|\mathcal{H}(\hat{h} - h)\| \lesssim \begin{cases} \frac{a_2}{a_1}\sqrt{\frac{np}{\boldsymbol{snr}\times T}}\log(n) & \text{if} \quad T \gtrsim \min(R^2, n) \\ \frac{a_2}{a_1}\sqrt{\frac{Rnp}{\boldsymbol{snr}\times T}}\log(n) & \text{if} \quad R \lesssim T \lesssim \min(R^2, n). \end{cases} \tag{85}$$

Proof. We use the change of variable as (76). Let $\hat{\beta}(\lambda)$ be the estimator associated with a certain regularization parameter λ. We choose the solution

$$\hat{\beta} = \mathrm{argmin}_{\hat{\beta}(\lambda)}\|\boldsymbol{U}_{\mathrm{val}}\hat{\beta}(\lambda) - y_{\mathrm{val}}\|_2^2 \tag{86}$$

Denote the noise in validation data as ξ_{val}. We have that

$$\|\boldsymbol{U}_{\mathrm{val}}\hat{\beta} - y_{\mathrm{val}}\|_2^2 = \|\boldsymbol{U}_{\mathrm{val}}(\hat{\beta} - \beta) - \xi_{\mathrm{val}}\|_2^2 \tag{87}$$

$$= \|\boldsymbol{U}_{\mathrm{val}}(\hat{\beta} - \beta)\|_2^2 + \|\xi_{\mathrm{val}}\|_2^2 - 2\xi_{\mathrm{val}}^\top\boldsymbol{U}_{\mathrm{val}}(\hat{\beta} - \beta) \tag{88}$$

In this formulation, $\|\xi_{\mathrm{val}}\|_2^2$ in (88) is regarded as fixed among all problems, and we study the other two terms. Since $\boldsymbol{U}$ is normalized that each entry is i.i.d. $\mathcal{N}(0, 1/T_{\mathrm{val}})$, we have $\boldsymbol{E}\|\boldsymbol{U}_{\mathrm{val}}(\hat{\beta} - \beta)\|_2^2 = \|\hat{\beta} - \beta\|_2^2$.

The quantity $\xi_{\mathrm{val}}^\top\boldsymbol{U}_{\mathrm{val}}(\hat{\beta} - \beta)$ is zero mean. And

$$\boldsymbol{U}_{\mathrm{val}}(\hat{\beta} - \beta) \sim \mathcal{N}(0, \frac{\|\hat{\beta} - \beta\|_2^2}{T_{\mathrm{val}}}I) \tag{89}$$

So the variance of its inner product with ξ_{val} is $\sigma_{\xi_{\mathrm{val}}}^2\|\hat{\beta} - \beta\|_2^2/T_{\mathrm{val}}$ (the distribution of the inner product is sub-Gaussian). We know that

$$\|\hat{\beta} - \beta\|_2 \approx \sqrt{\frac{R\log^2 n}{T}}\|\xi\|_2 = \sqrt{\frac{R\log^2 n}{T}}\sqrt{T_{\mathrm{val}}}\sigma_{\xi_{\mathrm{val}}}. \tag{90}$$

If $T_{\text{val}} \geq T$, we have that $\|\hat{\beta} - \beta\|_2 \gtrsim \sigma_{\xi_{\text{val}}}$.

Suppose the number of validated parameters λ is N_λ and we need to choose the size of validation data. With different validation data size T_{val}, the variance of $\|\boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta)\|_2^2$ decreases with rate $1/T_{\text{val}}$.

We fix factors a_1, a_2, such that with high probability, for all choices of λ, $\|\boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta)\|_2^2 - 2\xi_{\text{val}}^\top \boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta)$ is in the set $(a_1\|\hat{\beta} - \beta\|_2^2, a_2\|\hat{\beta} - \beta\|_2^2)$. The terms are sub-Gaussian, so there exists a constant c (depends quadratically on a_1, a_2) such that for every choice of λ,

$$\mathbf{Pr}\left(\left|\|\boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta)\|_2^2 - 2\xi_{\text{val}}^\top \boldsymbol{U}_{\text{val}}(\hat{\beta} - \beta)\right| \notin (a_1, a_2) \cdot \|\hat{\beta} - \beta\|_2^2\right) < \exp(-cT_{\text{val}}). \tag{91}$$

We choose probability P that any of the event in (91) happens, and solve for

$$N_\lambda \exp(-cT_{\text{val}}) < P, \tag{92}$$

where the factor N_λ comes from the union bound that all N_λ validations corresponding to λ_i succeed. Thus we set

$$T_{\text{val}} = \max\{T, \frac{1}{c}\log\frac{N_\lambda}{P}\} \tag{93}$$

so that (85) holds with probability $1 - P$. □