arXiv:2304.01708v1 [stat.ML] 4 Apr 2023

高维线性高斯系统的学习与集中现象：不变子空间方法

Muhammad Abdullah Naeem

摘要—**本文研究具有各向同性高斯噪声的稳定线性系统的两个时刻实现之间相关性的非渐近界。因此，通过从一条子轨迹中采样并利用*Talagrand* 不等式，我们证明，奖励的经验均值以高概率集中于稳态奖励附近（在线性反馈策略下闭环系统稳定时，动力系统会达到混合状态）。与“谱半径越大，样本之间的相关性越强”这一普遍看法相反，**系统特征值的代数重数与几何重数之间的巨大差异，会导致与系统转移矩阵相关的大型不变子空间**；系统一旦进入这一大型不变子空间，就会在一段时间内远离原点，随后接近以原点为中心的单位球。在那里，各向同性高斯噪声很可能使其逃离当前所在的不变子空间，从而在张$\mathbb{R}^n$成的不同不变子空间之间形成瓶颈，具体而言：在大型不变子空间中初始化的系统将长时间滞留其中，其滞留时间随不变子空间的维度呈对数线性增长，并与特征值幅度倒数的对数成反比。在通过单条轨迹对系统转移矩阵进行普通最小二乘法估计的问题中，当与大型不变子空间相关的转移矩阵的谱是发散的，而小型不变子空间对应稳定特征值时，这一现象甚至更加明显。我们的分析首次为连续高维状态空间上的随机动力系统的学习与集中现象的复杂性提供了具有可解释性的几何解释；它揭示了高维环境中的意外现象，并表明，无论问题属于系统辨识还是策略评估，实践者都应避免采用朴素的各向同性高斯激励。应根据可能遇到的不变子空间的最坏情况大小，选择预处理方法或不同类型的激励，从而确保在最少的时间步内对状态空间进行“充分”探索。**

I. 引言

过去十年间，我们见证了控制任务中学习的样本复杂度分析迅猛发展。无论是学习与某个控制策略相对应的价值函数问题（例如参见 [1]），还是像 ([2]) 中那样进行系统辨识，相关分析都严重依赖繁琐的概率与解析方法，却几乎无法提供解释或几何洞见。因此，我们目前对通过单条轨迹学习动力系统的理解仍存在诸多不确定性；正如我们稍后将要讨论的，即使是对稳定线性高斯系统（LG）的深入理解也尚付阙如。本文将彻底解决采样复杂度问题，阐明谱半径的作用，并回应近期关于大噪声可能有益的推测 [3], [4] 和 [5],[6]。事实上， [7], 最近指出，动力系统轨迹上相距较远的两个样本之间的相关性衰减与谱隙（泛函分析现象）有关，而与有限维矩阵的谱半径大小无关。我们对他们的结果给出有限维解释，并且全文都假设底层状态空间为高维。

为厘清这些模糊之处，本文研究两个简单问题。假设某个未知的 LG 系统在某个稳定策略 π 下运行，并且已达到平稳状态。理想情况下，我们希望知道平稳分布下的期望奖励，但实际只能获得奖励的时间平均值。与从平稳分布中独立同分布抽取的奖励相比，LG 系统中具有时间相关性的奖励时间平均值对期望奖励的近似效果如何？事实证明，这个简单问题的答案解释了引言中的所有疑问。其次，我们研究如何通过动力系统的单条轨迹，利用普通最小二乘法（OLS）估计进行系统辨识。

据作者所知，我们首次通过将原始状态转移矩阵直和分解到其不变子空间上，分析了集中现象与系统辨识。结合 Talagrand 集中不等式以及子空间上的高斯投影，我们得出一个有别于常见观点的结论：系统辨识与集中现象的主要问题在于维数较大的不变子空间，以及与之相关的特征值绝对值较大。如果不变子空间的维数较大，且激励为各向同性高斯分布，那么以压倒性概率，大部分激励信号将落在这个大维数不变子空间中（高斯测度集中现象）。更复杂的是，如果与该大维数不变子空间相关的谱不稳定，或处于稳定与不稳定的临界边缘，那么投影后的状态转移算子作用于初始激励时，会使新的实现远离原点，同时仍将其限制在这个大不变子空间内。因此，信号始终无法探索较小的不变子空间，导致学习到的行为不准确。大维数不变子空间源于状态转移矩阵相关特征值的代数重数与几何重数不一致。与我们的方法不同，大多数系统辨识研究侧重于构造谱随仿真轨迹长度变化的界，或将系统初始化于原点并

作者谨此感谢 Yuansi Chen 博士提供的宝贵见解与反馈。
Muhammad Abdullah Naeem 任职于美国北卡罗来纳州达勒姆市 27708 邮编的杜克大学电气与计算机工程系，电子邮箱：muhammad.abdullah.naeem@duke.edu

证明其结果适用于低维模拟。正如我们将在高维下普通最小二乘法的仿真结果中所展示的，将随机初始激励投影到大不变子空间中会导致估计错误。

当我们研究时间平均向空间平均（由线性高斯系统趋于混合的分布定义；只有当系统矩阵的谱半径严格小于 1 时，这种分布才存在）的集中现象时，也能明显观察到这一现象。即使相关马尔可夫链已经混合至其平稳分布，也只有经过精心选取的子轨迹的时间平均会集中于空间平均附近，因为轨迹一旦进入高维不变子空间，就会持续远离原点（同时仍处于上述不变子空间内），直到其首次收缩命中时间；此时轨迹已朝原点移动，而各向同性高斯激励会使其跳出并进入另一个不变子空间，于是轨迹似乎重新生成。尽管从相关样本构成的子轨迹中采样，与从平稳分布中采集独立同分布样本，可得到相似的集中性结果，但该子轨迹是通过选取连续的实现生成的，其间隔为：在所有线性无关的不变子空间中取最大值，间隔随子空间维数呈对数线性增长，并与相应特征值模长倒数的对数成反比。

本文的组织结构如下。第二节介绍符号和预备知识。第三节针对一般动力系统，发展了 Talagrand 不等式的两种张量化方法，这些方法将是理解稳定线性随机动力系统的集中现象以及发散系统中普通最小二乘法不一致性的核心。第四节介绍稳定LG系统的一些基本事实，并提出利用 Gelfand公式从子轨迹中进行采样的方法。第五节简要介绍与状态转移矩阵相关的不变子空间，并给出通过从子轨迹中采样获得的集中现象的非渐近界。第六节的 A 小节首先简要介绍基于单条轨迹的普通最小二乘问题，随后通过对张量化 Talagrand 不等式的直接应用，得出各向同性高斯分布激励下发散系统中普通最小二乘法的不一致性。B 小节提出一种等周方法，用于直观说明高维情况下可能出现的问题，以及这些问题如何导致错误的普通最小二乘估计。第七节给出仿真结果。我们表明，与关于正则系统中普通最小二乘法一致性的现有认识相反，当不变子空间的大部分对应于一个爆炸特征值时，高维正则系统中的普通最小二乘法会失效。最后，我们在第八节总结全文并讨论未来的研究方向。

II. 符号与预备知识

1) 符号：我们使用 $I_n \in \mathbb{R}^{n\times n}$ 表示 n 维单位矩阵。对于随机变量 x 和 y，$Cov(x,\ y)$ 表示协方差。 $\mathcal{B}_\alpha^n := \{x \in \mathbb{R}^n :\ \|x\| := \|x\|_2 \leqslant \alpha\}$ 是 $\mathbb{R}^n$ 中的 α-球。同样， $\mathcal{S}^{p-1} := \{x \in \mathbb{R}^p : \|x\|_2 = 1\}$ 是 $\mathbb{R}^p$ 中的单位球面。 $\chi_{\{\}}()$ 是示性函数函数，而 $\rho(A)$、 $\|A\|_2$、 $\|A\|_F$、 $det(A)$、 $tr(A)$ 和 $\sigma(A)$ 分别表示 A 的谱半径、矩阵 2-范数、弗罗贝尼乌斯范数、行列式、迹和特征值集合（谱）。对于正定矩阵 A，最大和最小特征值分别记为 $\lambda_{max}(A)$ 和 $\lambda_{min}(A)$。对于每个矩形矩阵 $\mathbb{X} \in \mathbb{R}^{n\times N}$，其奇异值为 $\sigma_1(\mathbb{X}) \geqslant \sigma_2(\mathbb{X}),\ .....,\ \sigma_n(\mathbb{X}) \geqslant 0$，其中不失一般性地假设 $N > n$。最重要的是最大奇异值 $\sigma_1(\mathbb{X}) := \sup_{a\in\mathcal{S}^{N-1}} \|\mathbb{X}a\|$ 和最小奇异值 $\sigma_n(\mathbb{X}) := \inf_{a\in\mathcal{S}^{N-1}} \|\mathbb{X}a\|$。矩阵 $\mathbb{X}$ 的条件数是最大奇异值与最小奇异值之比，记为 $\kappa(\mathbb{X}) = \frac{\sigma_1(\mathbb{X})}{\sigma_n(\mathbb{X})}$。如果 $\mathbb{X}$ 的像空间的张成空间是 $\mathbb{R}^n$，更紧致地写作 $Im(\mathbb{X}) = \mathbb{R}^n$，则最小奇异值等于矩阵 $\mathbb{X}$ 的逆矩阵的范数的倒数，即 $\sigma_n(\mathbb{X}) = \frac{1}{\|\mathbb{X}^{-1}\|}$。函数 g: $\mathbb{R}^n \to \mathbb{R}^p$ 在常数 L 下是利普希茨的，即对于每个 $x,\ y \in \mathbb{R}^n$， $\|g(x)-g(y)\| \leqslant L\|x-y\|$。

一个序列 $\{a(N)\}_{N\in\mathbb{N}} \in \mathcal{O}(N)$，若其关于 N 至多线性增长（这并不限于渐近结果）。 $\mathcal{O}(1)$ 将用于表示与底层状态空间的大小或迭代次数无关的量。在 $\mathcal{X}$（连续空间）上的概率测度空间记为 $\mathcal{P}(\mathcal{X})$，其博雷尔子集空间记为 $\mathbb{B}\big(\mathcal{P}(\mathcal{X})\big)$。对于函数 r 和 $\mu \in \mathcal{P}(\mathcal{X})$，我们用 $< r >_\mu$ 表示 r 关于 μ 的期望。最后，对于集合 $\mathcal{K} \subseteq \{1,\ ..... M\}$，其补集为 $\mathcal{K}^{\complement} := \{1,\ ..... M\}\backslash\mathcal{K}$。

在度量空间 $(\mathcal{X},\ d)$ 上，对于 $\mu,\ \nu \in \mathcal{P}(\mathcal{X})$，我们定义 $p \in [1,\ \infty)$ 阶 Wasserstein度量如下：

$$\mathcal{W}_d^p(\nu,\mu) = \left(\inf_{(X,Y)\in\Gamma(\nu,\mu)} \mathbb{E}\ d^p(X,Y)\right)^{\frac{1}{p}}; \qquad (1)$$

这里， $\Gamma(\nu,\ \mu) \in P(\mathcal{X}^2)$，且 $(X,\ Y) \in \Gamma(\nu,\ \mu)$ 表示随机变量 $(X,\ Y)$ 在 $P(\mathcal{X}^2)$ 上服从某些概率分布，其边缘分布分别为 ν 和 μ。比较 $\mathcal{X}$ 上两个概率分布的另一种方法是使用相对熵，其定义如下：

$$Ent(v||u) = \begin{cases} \int \log\left(\frac{d\nu}{d\mu}\right) d\nu, & \text{if } \nu << \mu, \\ +\infty, & \text{otherwise.} \end{cases} \qquad (2)$$

在介绍用于推导相关随机变量集中现象的数学框架之前，我们先介绍本文后续将使用的以下结果。

a) Talagrand 不等式或传输-熵不等式：考虑度量空间 $(\mathcal{X},\ d)$ 以及参考概率测度 $\mu \in P(\mathcal{X})$。那么，如果对于某个 $C > 0$，对所有 $\nu \in P(\mathcal{X})$ 都有如下关系，我们称 μ 满足 $\mathcal{T}_1^d(C)$，或简洁地记作 $\mu \in \mathcal{T}_1^d(C)$：

$$\mathcal{W}_d(\mu,\nu) \leqslant \sqrt{2CEnt(\nu||\mu)}. \qquad (3)$$

定理1 ([8])。 μ 满足 $\mathcal{T}_1^d(C)$ 当且仅当对于任意

利普希茨函数 f 满足 $<f>_\mu := \mathbb{E}_\mu f$，有

$$\int e^{\lambda(f-<f>_\mu)}d\mu \leqslant \exp(\frac{\lambda^2}{2}C\|f\|^2_{L(d)}), \tag{4}$$
$$where \quad \|f\|_{L(d)} := \sup_{x\neq y}\frac{|f(x)-f(y)|}{d(x,y)}.$$

注 2. *(4)*结合马尔可夫不等式可知，若从 $\mu \in \mathcal{T}_1^d(C)$ 中采样 x，则

$$\mathbb{P}\left[\left|r(x)- < r >_\mu \right| > \epsilon\right] \leqslant 2\exp\left(-\frac{\epsilon^2}{2C\|r\|^2_{L(d)}}\right). \tag{5}$$

III. 通过张量化将集中现象推广到相关随机变量

在某个依赖于状态的策略 π 作用下，我们考虑如下形式的闭环随机动力系统

$$x_{k+1} = F\big(x_k, \pi(x_k), \epsilon_k\big), \quad \text{with } \epsilon_k \quad i.i.d,^1 \tag{6}$$

其中，对所有 $k \in \mathbb{N}$ 和 F， $x_k \in \mathbb{R}^n$：$\mathbb{R}^n \times \mathbb{R}^n \times\mathbb{R}^n \longrightarrow \mathbb{R}^n$。对于遍历平均的集中现象问题，我们假设转移核在 Wasserstein度量W_d 下，关于某个距离函数 d 收敛到某个平稳分布 μ_∞。该随机动力系统可以视为具有分布 $\mu^N \in P(\mathcal{X}^N)$ 和 $P^m(x, \mathcal{B})$ 的马尔可夫链 $x^N := (x_i)_{i=1}^N$，对所有属于 $\mathcal{X}$ 的博雷尔子集 $\mathcal{B}$ 均成立。我们可以将度量 d 扩展到 $\mathcal{X}^N$，如下所示

$$d_{(N)}(x^N, y^N) = \sum_{i=1}^N d(x_i, y_i). \tag{7}$$

若 $\mu^N \in \mathcal{T}_1^{d_{(N)}}(\mathcal{O}(N))$ 和 r $\|r\|_{L(d)} \leqslant$是 1-利普希茨的，即，则 $\Phi(x^N) := \frac{1}{N}\sum_{i=1}^N r(x_i)$满足 $\|\Phi\|_{L(d_{(N)})} \leqslant \frac{1}{N}$；将这些结果代入 (4)，可得—

$$\mu^N\left[\left|\frac{1}{N}\sum_{i=1}^N r(x_i) - \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^N r(x_i)\right)\right| > \epsilon\right] \tag{8}$$
$$\leqslant 2\exp\left(-\frac{N\epsilon^2}{2C}\right).$$

A. 压缩性与一致传输常数

由 (7) 不难想到，对于马尔可夫链的过程级分布，T-E 何时至多随维度线性增长（样本内项）？ 充分条件（例如参见 [9], [10]）为

$$(i) \quad P(x,\cdot) \in T_1^d(C), \text{ for all } x \in \mathcal{X}, \text{and some } C > 0, \tag{9}$$
$$(ii) \quad \mathcal{W}_d(P(x,\cdot), P(y,\cdot)) \leqslant \hat{\lambda} d(x,y), \text{ for all } (x,y) \in \mathcal{X}^2 \tag{10}$$

ˆ以及一些 $\lambda \in [0,1)$。

性质（9）通常称为一致传输常数的存在性，而（10）表示.....的压缩性

[1]为简洁起见，从现在起我们将在状态更新方程中省略对 π 的引用，因为状态依赖策略意味着存在某个函数 G，使得 $F(x_k, \pi(x_k), \epsilon_k) = G(x_k, \epsilon_k)$。

马尔可夫链在 Wasserstein 度量下的压缩性 /瓦瑟斯坦意义下的谱隙。现在，以下结果成立。

定理 3。若(9) 和(10) 成立，则来自马尔可夫链 $(x_1, \ldots, x_N)$ 的样本的过程级分布（记为 $Law\ x_1, \ldots, x_N$），并以 μ^N 表示）满足 (
$T_1^{d_{(N)}}\left(\frac{CN}{(1-\hat{\lambda})^2}\right)$, *for all* $N \in \mathbb{N}$.

证明。详细证明见 [9] 的定理 2.5。 □

a) 相关性衰减：结合条件 (9) 和 (10)，并对 Bobkov-Gotze 对偶形式 (4) 两边出现的较小 λ（阶数不超过 λ^2）进行泰勒展开，可得对所有 $x \in \mathcal{X}$，均有ˆ

$$|Cov_{P_x}[f(x_n), f(x_{n+k})]| \leqslant \frac{\lambda^k}{1-\hat{\lambda}^2}C\|f\|^2_{L(d)}. \tag{11}$$

我们将在第 VI 节中大量依赖另一种张量化过程，以便更好地理解数据矩阵奇异值的偏差不等式。该过程采用受 ℓ^2 启发的 $\mathcal{X}^N$ 上的度量，如 [11], 所示，即

$$d^2_{(N)}(x^N, y^N) := \sqrt{\sum_{i=1}^N d^2(x_i, y_i)}. \tag{12}$$

IV. 稳定动力系统与近似独立子轨迹

A. 从不变测度中独立采样

所考虑的马尔可夫链是一个 n 维、带有各向同性噪声的线性高斯系统：

$$x_{t+1} = Ax_t + w_t, \quad \rho(A) < 1 \quad \text{and i.i.d} \quad w_t \sim \mathcal{N}(0, \mathcal{I}_n). \tag{13}$$

它收敛至平稳分布 $\mu_\infty \sim \mathcal{N}(0, P_\infty)$，其中可控格拉姆矩阵 P_∞ 是下列李雅普诺夫方程的唯一正定解：

$$A^T P_\infty A - P_\infty + I_n = 0. \tag{14}$$

$$\mathbb{P}\Big[\big|\frac{1}{N}\sum_{i=1}^N r(x_i) - \mu_\infty(r)\big| > \epsilon\Big] \leqslant 2\exp\left(-\frac{N\epsilon^2}{2\lambda_{max}(P_\infty)}\right). \tag{15}$$

B. 从线性高斯系统的子轨迹中采样

在线性系统领域，控制系统的稳定性对应于 $\rho(A) < 1$（临界稳定对应于 $\rho(A) \geqslant 1$，而当 $\rho(A) > 1$时，系统发散）；并且实分析中的一个经典结果（Gelfand 公式）表明，对于所有 $\rho \in (\rho(A), 1)$ 和 $k \in \mathbb{N}$，存在一个有限正常数L_ρ，使得 $\|A^k\| \leqslant L_\rho \rho^k$。由于 $L_\rho \rho$ 可能非常大，朴素的 Wasserstein 压缩性条件（10）并不成立。为了将前述结果推广到谱半径情形，我们将首次收缩命中时间 定义为：

$$\hat{k} := \min\{k \in \mathbb{N} : \|A^k\| < 1\}, \tag{16}$$

ˆ 即使转移核在瓦瑟斯坦意义下具有压缩性所需的最小自然数，其中第 k 步转移核 $P(x_{\hat{k}} = \cdot|x_0)$。

因此，我们不考虑原始轨迹$(x_0, x_1, x_2, \ldots)$而考虑子轨迹

$(x_0,\ x_{\hat{k}},\ x_{2\hat{k}},\ \ldots)$ 其LG动力学经过了修改:

$$x_{\hat{k}(i+1)} = A^{\hat{k}} x_{\hat{k}(i)} + s_{\hat{k}(i)} \tag{17}$$

其中 $s_{\hat{k}(i)} \sim \mathcal{N}(0\ \Sigma_{\hat{k}})$ $\Sigma_{\hat{k}} = \sum_{l=0}^{k-1}[A^l][A^l]^T$,, $\forall i \in \mathbb{N}$ 独立同分布且满足:。注意到 $\Sigma_{\hat{k}}$ 是正定的: 因此,

$$(i)\quad P^{\hat{k}}(x,\cdot) \in T_1^d\left(\|\Sigma_{\hat{k}}^{\frac{1}{2}}\|^2\right), \qquad \text{for all } x \in \mathcal{X}. \tag{18}$$

$$(ii)\quad \mathcal{W}_d(P^{\hat{k}}(x,\cdot), P^{\hat{k}}(y,\cdot)) \leqslant \lambda_{\hat{k}} d(x,y), \tag{19}$$

对于所有 $(x,y) \in \mathcal{X}^2$ 和 $\lambda_{\hat{k}} := \|A^k\| \in [0,1)$。 $\bar{\mu}_{\hat{k}}^N$: $Law(x_{\hat{k}(1)}\ \ldots\ x_{\hat{k}(N)}) \in T_1^{d_{(N)}}\left(\frac{\|\Sigma_{\hat{k}}^{\frac{1}{2}}\|^2 N}{(1-\|A^{\hat{k}}\|)^2}\right)$ – 3 ,,并且根据,

如果我们从 $x_0 \sim \mu_\infty$ 开始, 则有以下集中现象:

$$\mathbb{P}\Big[\big|\frac{1}{N}\sum_{i=1}^{N} r(x_{\hat{k}(i)}) - \mu_\infty(r)\big| > \epsilon\Big] \leqslant 2\exp\left(-\frac{N\epsilon^2[1-\|A^{\hat{k}}\|]^2}{2\lambda_{\max}(P_\infty)}\right) \tag{20}$$

$$= 2\exp\left(-\frac{N\epsilon^2[1-\|A^{\hat{k}}\|]^2\lambda_{\min}(P_\infty^{-1})}{2}\right). \tag{21}$$

因此, 与来自 μ_∞ 的独立同分布样本相比, 虽然在时间上相关但分布相同的样本 (每个样本单独服从 μ_∞), 间隔 k 倍时间抽取, 可以实现良好的集中现象, 但我们需要长度为 Nk 的轨迹。注意, 随着 $k \to \infty$, 平稳链集中性会变为独立同分布集中现象。关于非平稳情形的详细分析, 请参见 [12]。为了界定首次收缩命中时间, 我们需要理解:

V. 与非对称算子特征值问题相关的不变子空间结构

线性算子 A 的特征值位置或大小只能提供其性质的部分信息 (为便于阐述, 本文假设 A 没有非平凡零空间)。事实上, 了解 A 等同于了解其不变子空间 (参见例如 [13])。粗略地说, 特征值的代数重数由矩阵的行列式决定。

$$det(zI - A) = \prod_{i=1}^{K}(z-\lambda_i)^{m_i}, \tag{22}$$

其中 λ_i 互不相同, 重数为 m_i。当 $dim[N(A-\lambda_i I)] < m_i$ 时, 情况会变得复杂, 由此产生不变子空间 (由多个线性无关向量张成)。因此, 状态空间可表示为 A–不变子空间的直和分解。

$$\mathbb{R}^n = M_{\phi(1)} \oplus M_{\phi(2)} \oplus \ldots \oplus M_{\phi(L)} \tag{23}$$

以及相应的正交投影 $[E_{\phi(i)}]_{i=1}^L$, 使单位矩阵可写为:

$$I_n = E_{\phi(1)} \oplus E_{\phi(2)} \oplus \ldots \oplus E_{\phi(L)} \tag{24}$$

其中, ϕ 是从 $\{1,\ldots,L\}$ 到 $\sigma(A)$的满射。ϕ 是双射, 当且仅当特征向量张成 $\mathbb{R}^n$。当 $\sigma(A)$的某个元素的代数重数与几何重数之间存在差距时, 考虑不变子空间 M_λ, 其中某个 $\lambda \in \sigma(A)$的代数重数为 λ, 但只有一个线性无关的特征向量 v_1, 使得 $Av_1 = \lambda v_1$。因此, 我们递归地生成广义特征向量 $v_2, v_3, \ldots, v_{|B_\lambda|}$, 满足 $(A-\lambda I)v_2 = v_1$、$(A-\lambda I)v_3 = v_2$, 依此类推。我们有以下第 k 步迭代:

$$\begin{aligned}
A^k v_1 &= \lambda^k v_1 \\
A^k v_2 &= \lambda^k v_2 + \binom{k}{1}\lambda^{k-1} v_1 \\
A^k v_3 &= \lambda^k v_3 + \binom{k}{1}\lambda^{k-1} v_2 + \binom{k}{2}\lambda^{k-2} v_1 \\
&\ldots = \ldots \\
A^k v_{|B_\lambda|} &= \lambda^k v_{|B_\lambda|} + \binom{k}{1}\lambda^{k-1} v_{|B_\lambda|-1} + \ldots \\
&+ \binom{k}{|B_\lambda|-1} v_1
\end{aligned} \tag{25}$$

定理 4. 尽管从前面的迭代中可以明显看出这一点, 我们仍可对矩阵 A 作用于不变子空间 M_λ所对应的 k –th 次迭代的范数给出严格的上下界, 具体如下:

$$|\lambda|^k \sum_{m=0}^{|B_\lambda|-1} \frac{1}{|\lambda|^m} \leqslant \|A_{M_\lambda}^k\|_2 \leqslant |\lambda|^k k^{|B_\lambda|} \sum_{m=0}^{|B_\lambda|-1} \frac{1}{|\lambda|^m}, \tag{26}$$

其中 $A_{M_\lambda}^k := A^k E_\lambda$

证明。 该结果可通过对 [14] 中给出的界作简单变换得到。

$$\begin{aligned}
|\lambda|^k \sum_{m=0}^{|B_\lambda|-1} \frac{1}{|\lambda|^m} \leqslant \|A_{M_\lambda}^k\|_2 &= \sum_{m=0}^{|B_\lambda|-1} \binom{k}{m} |\lambda|^{k-m} \\
&\leqslant |\lambda|^k k^{|B_\lambda|} \sum_{m=0}^{|B_\lambda|-1} \frac{1}{|\lambda|^m}
\end{aligned} \tag{27}$$

在第一个和第二个不等式中, 我们使用了这一事实: $1 \leqslant \binom{k}{m} \leqslant k^{|B_\lambda|}$ 对于 $m \in [0,\ \ldots\ldots,\ B_{|\lambda|}-1]$

□

注 5。 此外, 若 $|\lambda| \in (0,1)$, 则:

$$\|A_{M_\lambda}^k\|_2 \leqslant k^{|B_\lambda|}|B_\lambda||\lambda|^{k+1-|B_\lambda|} \tag{28}$$

and

$$k \geqslant \frac{\ln(|B_\lambda|)}{\ln(\frac{1}{|\lambda|})} + \frac{|B_\lambda|\ln(k)}{\ln(\frac{1}{|\lambda|})} + (|B_\lambda|-1) \tag{29}$$

对于 $\|A_{M_\lambda}^k\|_2 < 1$ 已足够。

现在, 我们可以对如何选择子轨迹, 使时间平均值围绕其空间平均值呈现尖锐的集中现象这一问题, 给出确切的分析性结论, 正如 IV-B 小节中所提出的。

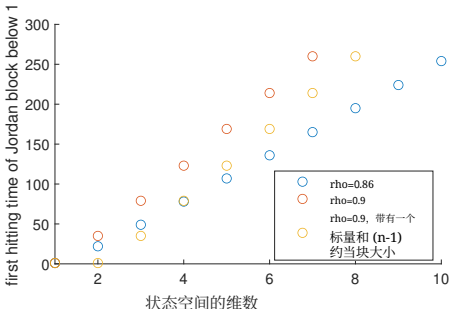


图 1. $n \times n$ 约当形的首次收缩命中时间。情形 1：单个特征值为 0.86 的约当块。情形 2：单个特征值为 0.9 的约当块。情形 3：$(n-1) \times (n-1)$ 一个特征值为 0.9 的约当块，以及一个特征值为 0.9 的单块。

定理 6. 算子 A 限制在不变子空间 $M_{\phi(i)}$ 上时，其首次收缩命中时间为 $\mathcal{O}\left(\begin{smallmatrix} |B_{\phi(i)}| \ln |B_{\phi(i)}| \\ \left(\frac{1}{|\phi(i)|}\right) \end{smallmatrix}\right) \ln$。因此，线性算子 A 的收缩命中时间，是其所有不变子空间上的最差收缩命中时间。

$$\hat{k} = \min\left[k \in \mathbb{N} : k \geqslant \max_{i \in 1,\ldots,L} \left(\frac{4|B_{\phi(i)}| \ln |B_{\phi(i)}|}{\ln \frac{1}{|\lambda_{\phi(i)}|}}\right)\right], \quad (30)$$

图 *1* 所示的仿真也验证了这一点。

备注 7. 由于首次收缩命中时间的一个非常宽松的上界为 $\left(\begin{smallmatrix} n \ln n \\ \left(\frac{1}{\rho(A)}\right) \end{smallmatrix}\right) \ln$，系统辨识领域的大多数文献在研究稳定系统时都将其用作选择块大小的开销（例如，见 *[3]* 和 *[2]*）。

由于 $\|A^k\|$ 在瓦瑟斯坦意义下是收缩的，我们得到子轨迹 (17) 指数收敛到平稳分布：

定理 8。 $\lambda_{\hat{k}} := \|A^k\|_2 < 1$ 且对所有 $m \in \mathbb{N}$

$$\begin{aligned} W_d(P_x^{\hat{k}(m)}, \mu_\infty) &\leqslant \lambda_{\hat{k}}^m W_d(P_x^{\hat{k}}, \mu_\infty) \\ &\leqslant \lambda_{\hat{k}}^m \sqrt{\lambda_{\hat{k}} \|x\| + Tr\left([\sqrt{\Sigma_{\hat{k}}} - \sqrt{P_\infty}]^2\right)}. \end{aligned} \quad (31)$$

证明。只需证明，对于所有 x, y 属于 $\mathbb{R}^n$，$W_d(P_x^k, P_y^k) \leqslant \lambda_{\hat{k}} \|x - y\|$，参见例如 [15]。给定任意 $x \in \mathbb{R}^n$，我们可以将其写成直和 $x = \sum_{i=1}^L E_{\phi(i)} x$。利用正交性：$E_{\phi(i)} E_{\phi(j)} = 0$ 对于 $i \neq j$，我们还有 $\|x\|_2 = \sum_{i=1}^L \|E_{\phi(i)} x\|_2$ $\|A^{\hat{k}} x\|_2 =$。现在，$\|A^{\hat{k}} \sum_{i=1}^L E_{\phi(i)} x\|_2 \leqslant \sum_{i=1}^L \|A^{\hat{k}}_{M_{\phi(i)}}\|_2 \|E_{\phi(i)} x\|_2 \leqslant \lambda_{\hat{k}} \|x\|_2$ 其中最后一个不等式由假设以及式（29）中给出的各个块的第一个收缩命中时间推出。因此，对于所有 x, y 属于 $\mathbb{R}^n$，都有 $W_d(P_x^k, P_y^k) \leqslant \lambda_{\hat{k}} \|x - y\|$，结论可由高斯分布之间 Wasserstein距离关于其均值和协方差的一般表达式得出，参见例如 [16]。 □

VI. 基于单条轨迹的系统辨识

A. 普通最小二乘法

本节分析从单条观测轨迹（如 [17], [3], [2]）对 $(x_0, x_1, \ldots, x_N)$ 进行普通最小二乘估计系统转移矩阵 A 的问题，其中该轨迹满足：

$$x_{t+1} = Ax_t + \eta_t, \quad \text{where } \eta_t \sim \mathcal{N}(0, \mathcal{I}_n). \quad (32)$$

普通最小二乘法的解为：

$$\hat{A} = \arg \min_{B \in \mathbb{R}^{n \times n}} \sum_{t=0}^{N-1} \|x_{t+1} - Bx_t\|. \quad (33)$$

设 $\mathbb{X}_+ = [x_1, x_2, \ldots\ldots, x_N]$ 和 $\mathbb{X}_- = [x_0, x_1, \ldots\ldots, x_{(N-1)}]$，以及噪声协变量 $E = [\eta_0, \eta_1, \ldots\ldots, \eta_{N-1}]$，则最小二乘解及误差的闭式表达式为：

$$\hat{A} = \mathbb{X}_+ \mathbb{X}_-^T (\mathbb{X}_- \mathbb{X}_-^T)^{-1} \quad (34)$$

$$\|A - \hat{A}\| = \|E\mathbb{X}_-^T (\mathbb{X}_- \mathbb{X}_-^T)^{-1}\| \quad (35)$$

误差可以被上界限定为：

$$\begin{aligned} \|A - \hat{A}\| = \|E\mathbb{X}_-^T (\mathbb{X}_- \mathbb{X}_-^T)^{-1}\| &\leqslant \|E\| \sigma_1(\mathbb{X}_-) \frac{1}{\sigma_n^2(\mathbb{X}_-)} \\ &= \frac{\sigma_1(E) \kappa(\mathbb{X}_-)}{\sigma_n(\mathbb{X}_-)}, \end{aligned} \quad (36)$$

其中，回想一下，$\kappa(\mathbb{X}_-)$ 是 $\mathbb{X}_-$的条件数。在随机矩阵理论中有一个众所周知的结果（例如参见 [18]）：如果 n 和 N 同时增大，同时保持其比值 $\frac{n}{N} = \gamma \in (0, 1)$不变，那么：

$$\sigma_1(E) \sim \sqrt{N} + \sqrt{n}, \quad (37)$$

其中，$\sim$ 表示下文定理 12 的讨论中解释的典型行为。因此，我们剩下的任务是界定数据矩阵 $\mathbb{X}_-$ 的奇异值（该矩阵包含相关随机变量；但在此我们将看到 Talagrand 不等式的全部威力）

定理 9。 数据矩阵的所有奇异值 $[\sigma_k(\mathbb{X}_-)]_{k=1}^n$ 满足以下集中界

1) *if* $\|A\|_2 < 1$*:*

$$\begin{aligned} &\mathbb{P}\left[\left|\sigma_k(\mathbb{X}_-) - \mathbb{E}\sigma_k(\mathbb{X}_-)\right| > \epsilon\right] \\ &\leqslant 2\exp\left(-\epsilon^2[1 - \|A\|]^2\right) \end{aligned}$$

2) 若 $\|A\|_2 = 1$：

$$\begin{aligned} &\mathbb{P}\left[\left|\sigma_k(\mathbb{X}_-) - \mathbb{E}\sigma_k(\mathbb{X}_-)\right| > \epsilon\right] \\ &\leqslant 2\exp\left(-\frac{\epsilon^2[e-1]}{N(N+1)}\right) \end{aligned}$$

3) $\|A\|_2 > 1$：

$$\begin{aligned} &\mathbb{P}\left[\left|\sigma_k(\mathbb{X}_-) - \mathbb{E}\sigma_k(\mathbb{X}_-)\right| > \epsilon\right] \\ &\leqslant 2\exp\left(-\frac{\epsilon^2[N-1]}{\|A\|^N e(N+1)}\right) \end{aligned}$$

证明。 证明思路源于对相关协变量的 Talagrand 不等式进行张量化。注意，对于任意 $a \in S^{N-1}$:

$$(x_0, x_1, \ldots, x_{N-1}) \mapsto \mathbb{X}_- a \mapsto \|\mathbb{X}_- a\|_2 \tag{38}$$

在度量 $d^2_{(N)}$下，这是一个从 $(\mathbb{R}^n)^N$ 到 $\mathbb{R}$ 的 1-利普希茨映射。由于 S^{N-1} 是紧致的，下确界和上确界均可取到，因此最小和最大奇异值都是 1-利普希茨的（根据它们的极小极大刻画／Courant–Fischer 定理，中间奇异值也如此）。令 $\mu^N := Law(x_0,\ x_1,\ \ldots,\ x_{N-1})$

- if $\|A\|_2 < 1$:

$$\mu^N \in T_1^{d^2_{(N)}}\left(\frac{1}{[1-\|A\|_2]^2}\right) \tag{39}$$

- 若 $\|A\|_2 = 1$:

$$\mu^N \in T_1^{d^2_{(N)}}\left(\frac{N(N+1)}{e-1}\right) \tag{40}$$

- 若 $\|A\|_2 > 1$:

$$\mu^N \in T_1^{d^2_{(N)}}\left(\frac{\|A\|^N e(N+1)}{N-1}\right) \tag{41}$$

见 [19] 中的命题 4.1，结论随即成立。 □

注 10. 由于情形 $\|A\|_2 > 1$ 包含发散系统，且方差似乎会随样本数量增加而恶化；因此，最小奇异值不会集中，普通最小二乘法是不一致的。还要注意，对稳定系统的子轨迹进行张量化如何能得到更好的集中估计，如

$$\mu^N_{\hat{k}} = Law(x_{\hat{k}(1)} \ \ldots \ x_{\hat{k}(N)}) \in :\ , T_1^{d^2_{(N)}}\left(\frac{\|\Sigma^{\frac{1}{2}}_{\hat{k}}\|^2}{(1-\|A^{\hat{k}}\|)^2}\right)$$

, 从而其奇异值满足与维数无关的集中不等式。

现在，我们通过等周方面的推理，直观地解释为什么普通最小二乘法在发散系统中是不一致的。

B. 各向同性高斯投影

备注 11. 本小节仅用于说明，使用了“约等于”之类的符号，但我们并未对此作出论证，因为这里的结果本质上属于等周性质；我们计划在后续工作中研究这些结果，不过即便现在，它们也能直观地解释高维动力系统的学习过程中可能出现的问题。

定理 12。 设 γ_n 是 $\mathbb{R}^n$ 中的各向同性高斯分布，且 $S \subset \mathbb{R}^n$ 是一个 $k-$ 维子空间。给定 $x \in \mathbb{R}^n$，令 x_S表示将 x 投影到 S上的结果。那么对于任意 $\delta \in (0,1)$

$$\gamma_n\left(x \in \mathbb{R}^n : \frac{\|x_S\|}{\|x\|} \geqslant (1-\delta)^{-1}\sqrt{\frac{k}{n}}\right) \leqslant e^{-\frac{\delta^2 k}{4}} + e^{-\frac{\delta^2 n}{4}}$$
$$\gamma_n\left(x \in \mathbb{R}^n : \frac{\|x_S\|}{\|x\|} \leqslant (1-\delta)\sqrt{\frac{k}{n}}\right) \leqslant e^{-\frac{\delta^2 k}{4}} + e^{-\frac{\delta^2 n}{4}}, \tag{42}$$

参见 *[20]*中的引理 *3.2*。

也就是说，投影到一个 $k-$ 维子空间上的向量范数与原向量范数之比通常为（∼） $\sqrt{\frac{k}{n}}$ 。这一观察的显著优势在于，我们可以将其应用于A-不变子空间，从而直观地理解不变子空间之间的瓶颈。严格来说，如果 E_λ 是到 $A-$不变子空间 M_λ 上的投影，那么:

$$\frac{\|E_\lambda \eta_0\|}{\|\eta_0\|} \sim \sqrt{\frac{|B_\lambda|}{n}} \tag{43}$$

随着底层状态空间规模增大，不变子空间的规模也随之增大；在保持 $n - |B_\lambda| = \mathcal{O}(1)$ 的同时，我们几乎可以确定，几乎所有激励信号 η_0 都位于 M_λ 内（由式（42）可得）。

备注 13. （42）中的估计是 *Johnson-Lindenstrauss*（*JL*）引理的核心。该引理是高维欧几里得空间中用于降维、规避维数灾难的强大工具，例如参见 *[21]*第 *1* 章引理 *2.1*。

在给定 $x_0 = 0$ 的条件下，我们可以将信号的 $N-th$ 实现表示为:

$$x_N = \sum_{t=1}^{N} A^{N-t}\eta_{t-1}. \tag{44}$$

对于爆炸情形，即 $\lambda \in \sigma(A)$ 满足 $|\lambda| > 1$，若 $\eta_0 \sim \mathcal{N}(0,\ I_n)$，则以压倒性概率有如下范数界

$$|\lambda|^{k-|B_\lambda|+1}\left(\frac{|B_\lambda|^{\frac{3}{2}}}{n^{\frac{1}{2}}}\right) \lesssim \frac{\|A^k E_\lambda \eta_0\|}{\|\eta_0\|} \lesssim k^{|B_\lambda|}|\lambda|^k\left(\frac{|B_\lambda|^{\frac{3}{2}}}{n^{\frac{1}{2}}}\right) \tag{45}$$

由于 $A^k E_\lambda = A^k_{M_\lambda} E_\lambda$，信号 $A^k E_\lambda \eta_0$ 完全位于子空间 M_λ 内，并且（45）可以解释为：如果块大小非常大且对应的特征值是发散的，那么我们的动力系统的实现将主要位于 M_λ 内，因此，任何实际可行的算法都无法学习该系统在状态空间其他部分的行为。

假设 $\mathbb{R}^n = M_\lambda \oplus M_\lambda^\perp$，其中 $\sigma(A_{M_\lambda^\perp})$ 稳定且最大特征值为 $\lambda^\perp \in (0,\ 1)$。由于 $n - |B_\lambda| = \mathcal{O}(1)$ 意味着 $|B_\lambda^\perp| = \mathcal{O}(1)$ ，显然可知，即使 N 的值很小，

$$\frac{\|A^N E_\lambda^\perp \eta_0\|}{\|A^N E_\lambda \eta_0\|} \approx 0. \tag{46}$$

VII. 普通最小二乘法的模拟结果

$A \in \mathbb{R}^{50\times 50}$ 仅有两个不同的特征值： $\lambda_1 = 1.5$块大小为 $|B_{\lambda_1}| = 47$ ，以及 $\lambda_2 = -0.5$ 块大小为 $|B_{\lambda_2}| = 3$，采用直和分解:

$$A = A_{\lambda_1} \oplus A_{\lambda_2}. \tag{47}$$

同样，状态空间可以表示为两个 A-不变子空间的直和分解:

$$\mathbb{R}^n = M_{\lambda_1} \oplus M_{\lambda_2}. \tag{48}$$

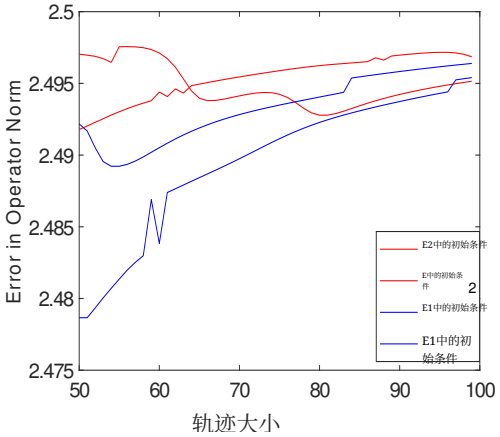


图2. 各向同性激励下正则爆炸系统的普通最小二乘法

我们可以通过计算 $[M,\ D] = eig(A)$ 来得到与两个Jordan块相关的两个不变子空间的张成空间，其中 $M = [M_{\lambda_1}\ M_{\lambda_2}]$，并分别通过应用投影算子 $E_1 = M_{\lambda_1}(M_{\lambda_1}^T M_{\lambda_1})^\dagger M_{\lambda_1}^T$ 和 $E_2 = M_{\lambda_2}(M_{\lambda_2}^T M_{\lambda_2})^\dagger M_{\lambda_2}^T$ 对 M_{λ_1} 和 M_{λ_2} 进行正交投影；$\dagger$ 用于表示伪逆。在图 2 中，我们进行了四次不同的仿真，并记录了当轨迹长度从 50 增加到 100 时的算子范数误差。蓝色曲线对应的初始条件是由 E_1 定义的正交投影，该投影作用于随机采样的各向同性高斯分布；类似地，红色曲线对应的初始条件是由 E_2 定义的正交投影，该投影作用于随机采样的各向同性高斯分布。正如发散系统最小奇异值方差的 Talagrand 不等式定理（9）所表明的，普通最小二乘法是不一致的。

注14. 案例研究中考虑的系统是完全正则的（如 *[22]* 中所定义），但最小二乘估计仍然不正确（高斯测度集中现象正在发挥作用）

VIII. 结论与未来工作

本文首先研究了高维情形下，采用各向同性高斯分布作为激励的稳定线性系统的两个不同时间实现之间的相关性。我们采用了一种新颖的方法：分析时不仅依据特征值的大小，还通过状态转移矩阵相关的不变子空间大小信息，考虑与算子相关的 几何结构 。这不仅使我们获得了几何层面的洞见，还通过从间隔较小的子轨迹中采样，提高了集中性结果；此前人们认为，两个近乎不相关的样本之间的间隔要大得多。在论文后半部分，我们结合这些几何洞见和 Talagrand 不等式，分析了以各向同性高斯分布为激励信号的爆炸系统中普通最小二乘法的不一致性问题。结果表明，高维动力系统的系统辨识问题与测度集中现象以及高维中的等周不等式密不可分。我们通过运用与高斯测度在大子空间上的投影集中性相关的结果，揭示了采用各向同性激励时普通最小二乘法存在的问题。未来，我们计划借助高维几何工具将这些结果形式化。

参考文献

[1] S. Tu 和 B. Recht，“线性二次调节器的最小二乘时序差分学习”，载于国际机器学习会议，2018年，第5005–5014页。[2] A. Tsiamis 和 G. J. Pappas，“ 线性系统难以学习 ”，载于 2021 年 IEEE 第 60 届决策与控制会议（CDC）。IEEE，2021年，第2903–2910页。[3] M. Simchowitz、H. Mania、S. Tu、M. I. Jordan 和 B. Recht，“无混合学习：迈向线性系统辨识的精确分析”，载于学习理论会议，2018年，第439–473页。[4] S. Oymak 和 N. Ozay，“重新审视基于 Ho–Kalman 的系统辨识：鲁棒性与有限样本分析”，《IEEE自动控制汇刊》，第67卷，第4期，第1914–1928页，2021年。[5] A. Tsiamis 和 G. J. Pappas，“具有对数遗憾的卡尔曼滤波器在线学习”，《IEEE自动控制汇刊》，2022年。[6] A. Tsiamis、I. Ziemann、N. Matni 和 G. J. Pappas，“控制的统计学习理论：有限样本视角”，arXiv预印本 arXiv:2209.05423，2022年。[7] M. A. Naeem，“随机动力系统的集中现象：一种算子理论方法”，arXiv预印本 arXiv:2212.03670，2022年。[8] S. G. Bobkov 和 F. G¨otze，“与对数Sobolev不等式相关的指数可积性与输运代价”，《泛函分析杂志》，第163卷，第1期，第1–28页，1999年。[9] H. Djellout、A. Guillin、L. Wu 等，“输运代价-信息不等式及其在随机动力系统与扩散过程中的应用”，《概率年刊》，第32卷，第3B期，第2702–2732页，2004年。[10] F. Bolley 和 C. Villani，“加权 csisz´ar-kullback-pinsker 不等式及其在输运不等式中的应用”，载于《图卢兹科学学院年刊：数学》，第14卷，第3期，2005年，第331–352页。[11] F. Malrieu，“若干非线性 PDE 的对数 Sobolev不等式”，《随机过程及其应用》，第95卷，第1期，第109–132页，2001年。[12] M. A. Naeem 和 M. Pajic，“连续状态空间动力系统的输运不等式、Lyapunov稳定性与采样”，arXiv预印本 arXiv:2205.12448，2022年。[13] T. Kato，《线性算子扰动理论》。Springer Science & Business Media，2013年，第132卷。[14] J. A. Tropp，“矩阵谱半径公式的初等证明”，2001年。[15] M. Hairer 和 J. C. Mattingly，“再谈马尔可夫链的 Harris 遍历定理”，载于《随机分析、随机场及其应用研讨会VI》。Springer，2011年，第109–117页。[16] C. R. Givens 和 R. M. Shortt，“一类概率分布的 Wasserstein 度量”，《密歇根数学杂志》，第31卷，第2期，第231–240页，1984年。[17] T. Sarkar 和 A. Rakhlin，“任意线性动力系统的近乎最优有限时间辨识”，载于国际机器学习会议。PMLR，2019年，第5610–5618页。[18] M. Rudelson 和 R. Vershynin，“随机矩形矩阵的最小奇异值”，《纯粹与应用数学通讯：库朗数学科学研究所发行的期刊》，第62卷，第12期，第1707–1739页，2009年。[19] G. Blower 和 F. Bolley，“乘积空间上的集中不等式及其在马尔可夫过程中的应用”，arXiv预印本 math/0505536，2005年。[20] A. Barvinok，“数学710：测度集中”，讲义，2005年。[21] L. Ambrosio，“最优输运、几何与泛函不等式”，（无标题）。[22] M. K. S. Faradonbeh、A. Tewari 和 G. Michailidis，“不稳定线性系统的有限时间辨识”，《自动化》，第96卷，第342–353页，2018年。