arXiv:2601.22200v1 [q-fin.ST] 29 Jan 2026

自适应良性过拟合（ABO）：面向非平稳时间序列在线学习的过参数化RLS

Luis Ontaneda
Mijares 伦敦大学学院
luis.ont.mij@gmail.com

Nick Firoozye 计算机科学系，伦敦大学学院
n.firoozye@ucl.ac.uk

摘要

过参数化模型近期挑战了传统学习理论：它们在插值极限之外展现出更强的泛化能力，这种现象称为良性过拟合。本文提出自适应良性过拟合（*ABO*），通过一种基于正交—三角更新的数值稳定形式，将递归最小二乘（RLS）框架扩展至这一领域。本文提出一种基于 QR 分解的指数加权递归最小二乘（QR–EWRLS）算法，将随机傅里叶特征映射与遗忘因子正则化相结合，从而实现非平稳条件下的在线自适应。正交分解避免了协方差形式的递归最小二乘所伴随的数值发散，同时保留了对不断变化的数据分布的适应能力。在非线性合成时间序列上的实验表明，所提方法能够维持有界残差和稳定的条件数，同时再现过参数化模型所特有的双下降行为。将该方法用于外汇和电力需求预测的结果表明，ABO 的精度很高（与基线核方法相当），同时速度提升了 20% 至 40%。这些结果在稳定的在线学习框架中，为自适应滤波、核近似与良性过拟合建立了统一的联系。

1 引言

自适应滤波，或等价地说在线回归，是现代信号处理、控制和序贯预测的基础。随着现代系统日益运行于高维、非平稳环境中，经典的稳定性—偏差权衡已不再适用；即使不发生过拟合，也有可能实现插值。本文从过参数化的视角重新审视递归最小二乘（RLS）框架，将自适应滤波与近期的良性过拟合理论联系起来，从而为非平稳时间序列的大型特征模型在线推断提供一种快速实用的方法。

1.1 良性过拟合

“良性过拟合”这一现象——即模型即使完美插值训练数据，也能实现良好的泛化——重塑了我们对高维学习中泛化的认识。相关的双下降 行为，即测试误差在超过插值阈值后再次下降，最早在深度神经网络中被观察到 [4, 5, 39]，随后在线性模型和核模型中得到了形式化研究 [15, 24]。即使在简单的最小二乘回归中，只要协方差矩阵满足较弱的谱条件，过参数化也能带来较低的风险（以均方误差衡量的测试集误差）。

在线性回归中，良性过拟合通常通过最小范数（亦称无岭回归）估计器进行分析

$$\hat{\beta} = \lim_{\lambda \to 0} \arg\min_{\beta} \Big(\sum_{t} \|y_t - X_t \beta\|^2 + \lambda \|\beta\|^2 \Big) \tag{1}$$

它等价于

$$\hat{\beta} = \arg\min_{\beta} \|\beta\|^2 \quad \text{s.t. } y = X\beta \tag{2}$$

当 X 为 $N \times D$ 高斯独立同分布随机矩阵，且 $N, D \to \infty$为 $N/D \to \gamma < 1$ 时，只要 $X^\top X$ 的谱保持良好性质[1]，测试误差便可能随着 D 增大而降低。这一现象表明，过参数化可以充当隐式正则化器，以谱平滑取代显式收缩。

尽管良性过拟合的分析最初是在静态情境下建立的，但类似的行为自然会出现在自适应学习和在线学习中。递归最小二乘（RLS）及其指数加权变体（EWRLS） [17, 28] 会随着新数据的到来持续更新回归系数，实际上执行着岭回归的流式类比。它们的稳定性，以及在过参数化情况下的泛化能力，仍是推动本研究的关键问题。

这些发现处于非平稳时间序列的背景下，或许可以解释为什么增加模型宽度、降低有效条件数能够改善泛化，尤其是在噪声方向足够广泛、可以吸收与主成分正交的方差时。

1.2 从核回归到神经切线核

在介绍神经网络和随机特征的类比方法之前，不妨先回顾一下：最小二乘估计可以视为采用线性核的核回归。将这一概念扩展到非线性特征映射后，便得到核方法；其泛化由核矩阵的特征值谱决定，而非原始设计矩阵 [16, 26]。这一核视角

[1] 具有相当数量的大奇异值以及任意多的小奇异值，但不趋于零的奇异
ar 值

构成了神经切线核（*NTK*）及其通过随机傅里叶特征（*RFFs*）实现的高效近似方法的基础。

1.3 从神经切线核到随机特征

神经切线核（NTK）框架 [22] 将无限宽度极限下深度神经网络训练的梯度下降动力学解释为再生核希尔伯特空间（RKHS）中带有特定数据依赖核的核回归。因此，过参数化神经网络的收敛与泛化性质可通过相应NTK的谱性质进行分析。这一观点在深度学习与经典核理论之间架起了直接桥梁，解释了过参数化神经网络为何表现得如同数据依赖核机器。

然而，核方法因需要计算并求逆大型核矩阵，随数据集规模的扩展性较差；一种计算上可行的类似方法由随机傅里叶特征 （RFFs） [27], 提出，它通过源自B¨ochner定理的随机正弦基来近似平移不变核。给定特征 z（x）$\in \mathbb{R}^D$，回归按如下方式进行

$$\min_{\beta} \|y - Z\beta\|^2 + \lambda\|\beta\|^2$$

其中 Z 具有与样本量无关的固定宽度。因此，随机傅里叶特征在保持计算高效的同时，保留了核方法的大部分表达能力。对过参数化区域中随机傅里叶特征的研究揭示了其与NTK及深度网络的相似之处，包括良性过拟合和双重下降现象。

1.4 随机傅里叶特征中的双下降

随机傅里叶特征可视为具有三角激活函数的两层神经网络；ReLU 和 Leaky-ReLU 等其他非线性函数也同样可行 [32]。其计算简单，再加上无限宽网络的通用逼近保证 [8, 21], ，使随机傅里叶特征成为研究插值区域和非经典区域行为的理想场景。尤其是，随机傅里叶特征也表现出与过参数化神经网络中相同的双下降 现象 [4, 15]。

为统一记号，最小范数插值器为

$$\hat{y} = X\hat{\beta}, \qquad \hat{\beta} = (X^\top X)^\dagger X^\top y$$

一个便于统一理解的视角来自岭回归:

$$\begin{aligned}\hat{\beta} &= \lim_{\lambda \to 0^+} (X^\top X + \lambda I)^{-1} X^\top y \\ &= X^\dagger y\end{aligned}$$

此极限总是返回 Moore–Penrose最小范数解，因此涵盖这两种情形:

- **经典区域**（$D \le N$）： $X^\top X$ 是非奇异的，且 $X^\dagger =$ （$X^\top X$）$^{-1}X^\top$
- **过参数化区域**（$D > N$）： $X^\top X$ 是奇异的，因此需要使用伪逆。

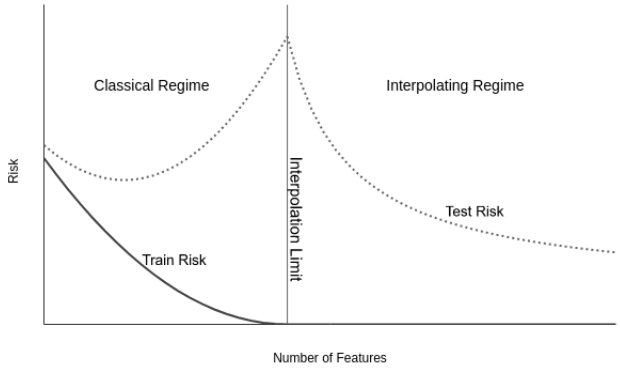

图1. 随机傅里叶特征中双下降现象的示意图。

在这两种情况下，基本恒等式

$$(X^\top X)^\dagger X^\top = X^\dagger$$

确保估计器的形式一致。

从经验上看，随机傅里叶特征呈现出典型的双下降曲线：

- 训练误差在插值阈值之前单调下降；
- 测试误差呈现经典的 U 形，在插值处达到峰值；
- 超过插值阈值后，随着模型过参数化程度不断提高，测试误差再次下降。

双下降现象存在于许多随机特征构造中。尽管在宽区域中，伪逆计算的成本会更高，但它也揭示了大型特征集如何对估计器起到正则化作用。泛化性能与有效条件数 密切相关，具体而言，$X^\top X$的经验分析和理论分析 [19] 表明，测试误差恰好在插值时有效条件数达到最大值时出现峰值。随着特征数量增加，有效条件数通常会降低，这解释了过参数化区域中测试误差的改善。

并非所有情形都会出现双下降。渐近理论 [4] 要求满足特定的谱性质——少数占主导地位的奇异值和大量中等大小的奇异值。如果这些性质不成立（例如，加入极高的确定性频率），性能可能会下降。特征加权 [19] 或通过随机傅里叶特征参数 σ 进行带宽控制可以缓解这一影响：较小的 σ 会带来更平滑的表示，而较大的 σ 则会产生高度振荡的特征，需要更多样本才能实现泛化。因此， σ 起到偏差–方差超参数的作用。

最后，也可以通过随机傅里叶特征的傅里叶分析起源来理解其行为。设$f(t)$为一个平方可积函数，其傅里叶变换为

$$\hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int f(t)\, e^{i\omega t}\, dt, \tag{3}$$

$$f(t) = \frac{1}{\sqrt{2\pi}} \int \hat{f}(\omega)\, e^{-i\omega t}\, d\omega \tag{4}$$

为了将这一点与核函数联系起来，请注意，平移不变核 $k(x_t - x_s) = g(t - s)$ 完全由其滞后域轮廓 $g(\tau)$ 刻画，该轮廓扮演着一个平稳“信号”的角色，定义

在时滞上。根据博赫纳定理，

$$g(\tau) = \int p(\omega)\, e^{i\omega\tau}\, d\omega$$

其中 $p(\omega)$ 是核函数的谱密度。对于径向基函数网络的情况，$\omega \sim \mathcal{N}(0, \gamma^2)$。我们注意到，在一般情况下，从随机傅里叶特征中采样时，γ 的选择是一种正则化形式，可防止生成的插值函数出现不必要的高频振荡（例如参见 [19]）。

随机　傅里叶特征通过蒙特卡洛采样近似该积分。定义

$$\varphi(x) = \sqrt{\frac{2}{m}} \left[\cos(\omega_1^\top x + b_1),\, \ldots,\, \cos(\omega_m^\top x + b_m)\right]^\top$$

其中 $\omega_j \sim p$ 且 $b_j \sim$ 服从 $\mathrm{Unif}[0, 2\pi]$。

Then

$$k(x_t - x_s) \approx \varphi(x_t)^\top \varphi(x_s)$$

因此，随机傅里叶特征可以解释为对核函数时滞域信号的傅里叶表示进行随机离散化。

1.5 在线学习与非平稳性

许多现实世界系统，例如金融、生理或环境过程，本质上都是非平稳的。指数加权递归最小二乘 (EWRLS) 通过遗忘因子 $\lambda < 1$ 对较旧的数据赋予较低权重，以适应此类动态，从而能够跟踪时变相关性和结构突变。

本研究的动机在于，金融市场中常见的非平稳时间序列需要高效的预测方法。通信、声学和信号处理领域广泛使用的非平稳信号分析与分解经典工具包括：

- **短时傅里叶变换 (STFT)：** 加窗傅里叶变换，

$$\hat{f}_\lambda(t, \omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi_\lambda(t - s) f(s) e^{-i\omega s}\, ds,$$

  其中，窗函数 $\phi_\lambda(r) = \phi(\lambda r)$ 通常为标准差为 $1/\lambda$的高斯函数，不过单边指数核 $\phi_\lambda(r) = e^{-\lambda r}$ 对于 $r > 0$ 也同样可行 [10, 33]。

- **小波变换：** 或多尺度分析，

$$W_\psi[f](a, b) = \int_{-\infty}^{\infty} \psi^*\left(\frac{x - b}{a}\right) f(x)\, dx,$$

  其中，$\psi(\cdot)$ 是具有紧支撑的母小波，* 表示复共轭。

- **维格纳变换：** 或维格纳准概率分布，

$$W_f(t, \omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f^*(t + s)\, f(t - s)\, e^{-i\omega s}\, ds$$

这些方法都通过紧支撑基函数或加窗，在时间或频率上引入局部化。其中，短时傅里叶变换（STFT）可以说是最常用的方法，应用领域包括声音和音乐的谱分析（它们显然具有不断变化的波形）。我们主要关注指数加权短时傅里叶变换，因为加入新信息后，它易于更新。

1.6 随机傅里叶特征的短时傅里叶变换与指数加权递归最小二乘

为将随机傅里叶特征（RFF）与短时傅里叶分析联系起来，考虑我们目标函数中的指数加权核函数项：

$$e^{-\lambda(t-s)}k(x_t - x_s) = e^{-\lambda\tau}\int p(\omega)e^{i\omega^\top(x_t-x_s)}d\omega$$

其中 $\tau = t - s$ 是时滞。遗忘因子 $e^{-\lambda\tau}$ 充当作用于时滞的单侧指数窗，以与指数加权短时傅里叶变换（STFT）相同的方式调制核函数的每个傅里叶分量。将其代入指数加权目标函数，可得如下形式：

$$\hat{\beta}_t = \arg\min_{\beta}\sum_{s\le t} e^{-\lambda(t-s)}\left\|y_s - \varphi(x_s)^\top\beta\right\|^2$$

该问题可通过指数加权递归最小二乘（EWRLS）更新，采用常规的在线方法求解。在此框架中，RLS 算法充当动态谱估计器，其中 RFF 权重 β_t 对应于信号谱分量的瞬时幅值，这些分量通过随机化基底解析得到。

1.6.1 计算复杂度与谱反卷积

这些方法在计算阶数上存在显著差异。传统的分块式短时傅里叶变换实现利用快速傅里叶变换（FFT），使每帧的复杂度达到 $O(M \log M)$；而递归变体的复杂度可达 $O(M)$ [34]。相比之下，所提出的 ABO 框架每次更新的复杂度为 $O(ND)$，其中 N 是滑动窗口长度， D 是随机特征的数量。

不同于标准的协方差形式的递归最小二乘，其计算复杂度为 $O(D^2)$，这是由于需要维护并求逆 $D \times D$ 自相关矩阵 [17, 28], ；ABO 中基于 QR 分解的形式利用滑动窗口的结构特性，使计算复杂度与特征维度保持线性关系 D。具体而言，通过在秩亏情形下高效更新最小范数最小二乘解 [30], ，更新过程仅限于活跃观测的 N 维子空间，因此只需执行 $O(ND)$ 次运算。这使算法能够有效扩展至实现良性过拟合所需的过参数化区域。

1.6.2 时域之外的扩展

与在数学上局限于时间或空间维度的短时傅里叶变换不同，EWRLS-RFF 将谱分析推广至任意高维特征空间。由于随机傅里叶特征能够为任意平移不变核函数提供显式映射，因此这一自适应框架可应用于图信号或多维传感器阵列等抽象输入。在这些情形下，该算法能够跟踪特征空间中的局部结构，而传统的基于网格的快速傅里叶变换方法从根本上无法适用。

1.7 通过QR更新实现稳定性

采用秩一更新与降阶的标准递归最小二乘方法，已知会因遗忘因子而出现数值不稳定性 [17, 28], 。对于过参数化模型，必须调整更新与降阶更新，以考虑奇异情形（即增秩更新与降秩降阶更新）。我们将更新与降阶更新公式的细节

列入附录 A，仅为完整起见。尽管如此，在过参数化区域中，这种递归最小二乘的秩一形式同样存在数值不稳定性。协方差形式的递归最小二乘的这种脆弱性促使我们将递推重新表述为正交三角（QR）更新，从而在序贯自适应下保持数值稳定性 [3, 18]。

格拉姆矩阵（$X^\top X$）† 的逆协方差，或在欠定区域（$D < N$）下的 Moore–Penrose 伪逆，使用 Sherman–Morrison 公式递归更新。尽管这种方法效率很高，但数值上较为脆弱：每次秩一修正都涉及近乎共线的外积之差，这可能破坏半正定性，并使协方差估计的条件数在每次迭代中实际上翻倍 [13, 18]。

因此，在有限精度算术下，协方差矩阵可能偏离对称性或变为不定矩阵，尤其是在病态或过参数化的设置中。平方根或基于 QR 的变体通过维护正交三角分解，并利用吉文斯旋转或豪斯霍尔德旋转进行更新，来解决这些问题 [2, 13], ，确保舍入误差保持有界。

1.8 贡献

基于这些见解，我们提出了自适应良性过拟合（*ABO*），这是一种自适应、过参数化的指数加权递归最小二乘框架，将随机傅里叶特征与指数加权 QR 更新相结合。我们的贡献有两方面：

- 我们提出了一种基于吉文斯旋转、用于秩一更新与降阶更新的数值稳定递归方案。通过显式维护 Moore-Penrose伪逆，该算法在秩亏区域中仍有明确的定义，从而确保在过参数化情形下得到最小范数插值器。

- 在非线性合成数据、金融数据和能源时间序列上的实证评估表明，该方法能够实现有界残差和数值稳定性，同时成功复现良性过拟合所特有的双下降行为。

- 我们证明，ABO 利用滑动窗口结构和伪逆性质，使每次更新的复杂度达到 $O(ND)$。这种关于特征维度的线性缩放 D （其中$D \gg N$）相比标准递归最小二乘的 $O(D^2)$ 开销有显著改进 [17]，从而实现高维在线学习。

这一表述弥合了良性过拟合理论分析与在线学习实用递归算法之间的差距，为非平稳环境中的自适应建模提供了统一框架。接下来，我们将介绍形式模型和算法表述，把递归最小二乘（RLS）扩展到过参数化和秩亏情形，同时在顺序更新过程中保持数值稳定性。

1.9 相关工作

经典的递归最小二乘（RLS）算法已在自适应滤波和信号处理领域得到广泛研究 [17, 28]。引入指数加权后，这些算法跟踪非平稳信号的能力有所提升 [3], ，但标准协方差形式仍容易出现数值不稳定性，因为它需要反复更新矩阵，即使采用标准岭惩罚也不例外[2]。

[2]随着样本量增长，指数加权递归最小二乘中的标准岭惩罚通常会变得极小，因为这有利于更新。因此，其稳定性在一定程度上受到限制。

后来，人们开发了平方根和基于 QR 的变体，以提高数值稳健性 [2,13,18]，但这些分析通常假设数据满秩且良态。

解决高维自适应滤波固有的病态性，传统上依赖显式正则化，尤其是通过稀疏性约束。受 LASSO 启发，零吸引递归最小二乘（ZA-RLS）算法及其重加权变体会对权重向量 w_n施加 ℓ_1 范数惩罚（或近似的 ℓ_0 或 计数 “范数”惩罚），有效地将“不显著”的系数压至零，以降低方差 [9, 20]。采用自适应投影次梯度法（APSM）等基于投影的方法，将权重约束在凸集 [38]内，也能实现类似的稳定性保证。

与其他用于自适应滤波器的正则化方法不同，我们研究过参数化所产生的隐式正则化，即良性过拟合， [4, 5, 15]。我们并非通过稀疏性或投影来人为限制模型容量，而是将特征空间（通过随机傅里叶特征扩展，见 [24, 27, 31, 36] ）扩展到插值阈值之上。我们注意到，尽管针对秩亏情形的研究很少，但该领域已有一些关于 Cholesky 方法和秩分解方法的探讨 [29, 30]。

正如近期“良性过拟合”研究的发现所示，我们的方法允许“噪声”能量无害地分布在高维空间的正交方向上，从而稳定递归最小二乘更新，而无需采用 ZA-RLS 的稀疏结构假设。

本研究通过构建一种基于 QR 的指数加权递归最小二乘算法，弥合了这些领域之间的差距。该算法在过参数化区域仍能保持数值稳定性，同时保留实时应用所需的递归结构。

2 系统模型

附录 A 中讨论的协方差更新方法的局限性，凸显了对一种数值稳定的递归公式的需求；这种公式在满秩和秩亏区域中都能可靠运行。为此，我们采用正交–三角（QR）分解重新表述 RLS 问题，从而避免对格拉姆矩阵进行显式求逆，并在滑动窗口更新过程中保持数值稳定性。

在每个时间步 t，我们观测一个原始协变量 $x_t \in \mathbb{R}^d$ 和一个标量响应变量 $y_t \in \mathbb{R}$。随机傅里叶特征提供了一种非线性映射

$$z_t = \varphi(x_t) \in \mathbb{R}^D,$$

并在 RFF 设计矩阵上执行回归。在长度为 N 的滚动窗口内，设计矩阵和响应向量为

$$Z_t = \begin{bmatrix} z_{t-N+1}^\top \\ \vdots \\ z_t^\top \end{bmatrix} \in \mathbb{R}^{N\times D}, \qquad Y_t = \begin{bmatrix} y_{t-N+1} \\ \vdots \\ y_t \end{bmatrix} \in \mathbb{R}^N.$$

经典最小二乘需要构造格拉姆矩阵 $A_t = Z_t^\top Z_t$，当 $D > N$ 时该矩阵可能是奇异的，并且按秩一方式更新时数值不稳定。相反，QR 递归最小二乘维护一个正交三角分解

$$Z_t = Q_t R_t, \tag{5}$$

其中 $Q_t \in \mathbb{R}^{N\times N}$ 是正交的，$R_t \in \mathbb{R}^{N\times D}$ 是上梯形的。随着新数据（z_{t+1}，y_{t+1}）到达并移除最早的一行，使用吉文斯旋转更新因子（Q_t，R_t），从而避免显式矩阵求逆，并确保在满秩和秩亏情况下的数值稳定性。

2.1 通过 QR 求解最小二乘

最小二乘估计求解

$$\min_\beta \ \|Z_t\beta - Y_t\|_2^2$$

利用 QR 分解 (5) 以及 Q_t 的正交性，目标函数变为

$$\|Z_t\beta - Y_t\|_2^2 = \|Q_t^\top (Z_t\beta - Y_t)\|_2^2 \tag{6}$$

$$= \|R_t\beta - Q_t^\top Y_t\|_2^2 \tag{7}$$

因此，最小二乘问题归结为求解三角方程组

$$R_t\beta = Q_t^\top Y_t$$

当 R_t 具有满列秩（$D \le N$）时，解可通过回代法得到：

$$\beta_t = R_t^{-1} Q_t^\top Y_t \tag{8}$$

在过参数化区域（$D > N$）中，R_t 秩亏，最小范数解通过将 R_t^{-1}替换为其 Moore–Penrose 伪逆得到：

$$\beta_t = R_t^\dagger Q_t^\top Y_t \tag{9}$$

该表达式在两种区域中均成立，并避免构造格拉姆矩阵 $Z_t^\top Z_t$，从而确保数值稳定性。

注 1（最小范数与良性过拟合）。在过参数化区域（$D > N$）中，公式（9）给出了唯一的最小-ℓ_2-范数插值解：

$$\beta_t = \arg\min_{\beta} \|\beta\|_2 \quad \text{s.t.} \quad Z_t\beta = Y_t, \tag{10}$$

等价于 Moore-Penrose 解 $\beta_t = Z_t^\dagger Y_t$。在这种设定下，泛化由最小范数约束所带来的隐式正则化决定，而非显式收缩。随着 Z_t 的零空间不断扩展，估计器会将方差分散到许多尾部特征值上——这一谱性质对于良性过拟合至关重要。这一机制与在线学习中观察到的双下降行为一致，也证明了对伪逆解采用稳定递归更新的合理性。

2.2 随机傅里叶特征

为引入非线性，同时保留易于处理的线性模型，我们使用随机傅里叶特征 （RFF）[27]将每个原始观测$x_t \in \mathbb{R}^d$ 映射到更高维的随机特征空间。

设 $A \in \mathbb{R}^{d\times D}$的列 $\omega_1, \ldots, \omega_D$独立地从 $p(\omega)$ 中抽取，且令 $b \in \mathbb{R}^D$包含从 Unif$[0, 2\pi]$中采样的独立同分布相位。随机特征映射为

$$z_t = z(x_t) = \sqrt{\frac{2}{D}} \cos(A^\top x_t + b) \in \mathbb{R}^D \tag{11}$$

其中余弦按元素应用。为使记号更加紧凑，我们采用这一表述，尽管许多人会同时包含 cos 和 sin，或采用新颖的采样方案，以确保表示更加高效。[3]

在长度为 N 的滚动窗口上，RFF 设计矩阵为

$$Z_t = \begin{bmatrix} z_{t-N+1}^\top \\ \vdots \\ z_t^\top \end{bmatrix} \in \mathbb{R}^{N\times D}$$

它构成了以下各节所述的基于 QR 的递归最小二乘算法的输入。维度 D 是固定的，并且独立于样本量进行选择，从而能够在满秩和过参数化区域中实现高效更新。

2.3 基于吉文斯旋转的 QR-RLS 更新

2.3.1 吉文斯旋转

吉文斯（平面）旋转 $G(i, j; c, s) \in \mathbb{R}^{N\times N}$ 除第 i 行和第 j 列之外均为单位矩阵，其中

$$G(i, j; c, s)_{[i,j],[i,j]} = \begin{bmatrix} c & s \\ -s & c \end{bmatrix}, \qquad c^2 + s^2 = 1$$

给定同一列中的主元 x 和次对角元素 y，我们选择

$$r = \sqrt{x^2 + y^2}, \qquad c = \frac{x}{r}, \qquad s = \frac{y}{r}$$

[3]众所周知，通过基于 *Orthogonal Random Features*、*Quasi-Monte Carlo Random Features*、*Structured Orthogonal Random Features and Fast Food* 的采样方法等，可以提高效率。

使得

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} r \\ 0 \end{bmatrix}$$

依次应用一系列此类旋转 $\{G_\ell\}$，以消去增广矩阵中的次对角线元素；这些旋转的乘积构成正交因子：

$$Q_t^\top = G_1^\top G_2^\top \cdots G_k^\top$$

由此得到 QR关系

$$Q_t^\top Z_t = R_t$$

其中 R_t 是更新后的上梯形因子。

2.3.2 观测更新

当新的特征向量 $z_t^\top$到达时，我们首先构造增广矩阵

$$\bar{R}_t = \begin{bmatrix} R_{t-1} \\ z_t^\top \end{bmatrix} \in \mathbb{R}^{(N+1)\times D}$$

这相当于将新的一行附加到窗口化设计矩阵 Z_{t-1}中。该矩阵不再是上梯形：其最后一行包含必须消去的次对角线元素。

应用一系列吉文斯旋转 G_t，将这些次对角元素置零：

$$G_t^\top \begin{bmatrix} R_{t-1} \\ z_t^\top \end{bmatrix} = \begin{bmatrix} R_t \\ 0^\top \end{bmatrix} \tag{12}$$

得到更新后的三角因子 $R_t \in \mathbb{R}^{N\times D}$。相应的正交因子更新为

$$Q_t = \begin{bmatrix} Q_{t-1} & 0 \\ 0 & 1 \end{bmatrix} G_t$$

使得增广设计矩阵满足

$$\begin{bmatrix} Z_{t-1} \\ z_t^\top \end{bmatrix} = Q_t \begin{bmatrix} R_t \\ 0^\top \end{bmatrix}$$

最小二乘估计随后为

$$\beta_t = R_t^\dagger Q_t^\top Y_t$$

无需从头重新计算 $R_t^\dagger$，而是通过吉文斯变换更新伪逆：

$$R_t^\dagger = \left(G_t^\top \begin{bmatrix} R_{t-1} \\ z_t^\top \end{bmatrix} \right)^\dagger = \begin{bmatrix} R_{t-1} \\ z_t^\top \end{bmatrix}^\dagger G_t$$

这由 $(GA)^\dagger = A^\dagger G^\top$ 对任意正交矩阵 G 得出。该递推避免了显式求逆，并在在线更新过程中保持数值稳定性。

2.3.3 伪逆更新（Greville/Cline）及其与 QR-R 的联系 LS

¯由上一节，我们得到行增广矩阵 R_t 及其 Moore–Penrose 伪逆 $\bar{R}_t^{\dagger}$ 。

定义：

$$
\begin{aligned}
c &= (I - R_{t-1}^{\dagger} R_{t-1})\, z_t \\
h &= z_t^{\top} R_{t-1}^{\dagger} \\
c^{\dagger} &= \frac{c}{\|c\|_2^2}
\end{aligned}
$$

则伪逆的 Greville/Cline 更新为

$$
\bar{R}_t^{\dagger} = \begin{bmatrix} R_{t-1}^{\dagger} - bh & b \end{bmatrix}, \qquad b = \begin{cases} c^{\dagger} & \text{if } c \neq 0, \\ \dfrac{R_{t-1}^{\dagger} h^{\top}}{1 + hh^{\top}} & \text{if } c = 0 \end{cases} \tag{13}
$$

该更新涵盖列空间完整和列相关情形 [7]。不同于通常受限于超定系统（$D \leq N$）的标准 QRD-RLS，ABO 利用 Greville/Cline 更新，在 $D > N$ 状态下维持稳定的最小范数解。

将变换后的右端项定义为

$$
w_t = Q_t^{\top} Y_t \tag{14}
$$

该项可通过以下方式递归更新：

$$
w_t = G_t^{\top} \begin{bmatrix} w_{t-1} \\ y_t \end{bmatrix}
$$

根据 Q_t的正交性，我们有

$$
\big(G_t^{\top} \begin{bmatrix} R_{t-1} \\ z_t^{\top} \end{bmatrix}\big)^{\dagger} = R_t^{\dagger} \iff \begin{bmatrix} R_{t-1} \\ z_t^{\top} \end{bmatrix}^{\dagger} G_t = R_t^{\dagger}
$$

因此，我们将 G_t 同时应用于三角因子和变换后的右端项。根据正交性， $G_t^{\top} G_t = I$，所以

$$
\begin{aligned}
\beta_t &= \begin{bmatrix} R_{t-1} \\ z_t^{\top} \end{bmatrix}^{\dagger} G_t G_t^{\top} \begin{bmatrix} w_{t-1} \\ y_t \end{bmatrix} \\
&= \begin{bmatrix} R_{t-1} \\ z_t^{\top} \end{bmatrix}^{\dagger} \begin{bmatrix} w_{t-1} \\ y_t \end{bmatrix}
\end{aligned}
$$

2.3.4 权重更新

$$
\begin{aligned}
\beta_t &= R_t^{\dagger} Q_t^{\top} y_t && (15) \\
&= \begin{bmatrix} R_{t-1}^{\dagger} - bz_t^{\top} R_{t-1}^{\dagger} & b \end{bmatrix} G_t G_t^{\top} \begin{bmatrix} w_{t-1} \\ y_t \end{bmatrix} && (16) \\
&= R_{t-1}^{\dagger} w_{t-1} - bz_t^{\top} R_{t-1}^{\dagger} w_{t-1} + by_t && (17) \\
&= \beta_{t-1} + b(y_t - z_t^{\top} \beta_{t-1}) && (18)
\end{aligned}
$$

向量 b 起到卡尔曼增益的作用；其表达式取决于新行 z_t 是增加 R_{t-1} （$c \neq 0$）的秩，还是位于其列空间（$c = 0$）中。

2.3.5 观测降阶

为实现滚动窗口，必须从 QR 因子中移除最早观测的贡献。令 G_{t-N} 表示吉文斯旋转，它可撤销先前施加于 Q_{t-1} [11] 第一行的正交变换。将 Q_{t-1} 分块为

$$Q_{t-1} = \begin{bmatrix} q_{t-N}^\top \\ Q_{t-1}^{(2:N)} \end{bmatrix}$$

其中，$q_{t-N}^\top$ 是与被丢弃的观测对应的行。选择降阶旋转 G_{t-N}，使得

$$G_{t-N}^\top q_{t-N} = \alpha e_1, \qquad \alpha \in \{\pm 1\}$$

从而得到

$$Q_{t-1} G_{t-N} = \begin{bmatrix} \alpha & 0 \\ 0 & Q_t \end{bmatrix}$$

如果 $\alpha = -1$，则对 G_{t-N} 的符号取反，以保持定向。

将 $G_{t-N}^\top$ 应用于三角因子，会将 R_{t-1} 的第一行恢复为旋转前的值$z_{t-N}^\top$。因此

$$G_{t-N}^\top R_{t-1} = \begin{bmatrix} z_{t-N}^\top \\ R_t \end{bmatrix}$$

减去秩一项 $e_1 z_{t-N}^\top$ 即可移除这一恢复的行：

$$G_{t-N}^\top R_{t-1} - e_1 z_{t-N}^\top = \begin{bmatrix} 0 \\ R_t \end{bmatrix}$$

丢弃零行后，剩余块就是降阶后的三角因子 R_t。

为便于后续使用，定义中间矩阵

$$\hat{R}_t := R_{t-1} - G_{t-N} e_1 z_{t-N}^\top$$

使得

$$G_{t-N}^\top \hat{R}_t = \begin{bmatrix} 0 \\ R_t \end{bmatrix}$$

由于 G_{t-N} 正交，广义逆恒等式 $(GA)^\dagger = A^\dagger G^\top$ 适用 [6], ，因此得到

$$(G_{t-N}^\top \hat{R}_t)^\dagger = \hat{R}_t^\dagger G_{t-N}$$

这一关系将用于表示降阶后的伪逆，并将降阶更新与后续的权重更新联系起来。

2.3.6 利用广义和公式进行伪逆降阶更新

为了应用 [6, 7], 广义逆和公式，引入

$$h = z_{t-N}^\top R_{t-1}^\dagger, \qquad k = R_{t-1}^\dagger G_{t-N} e_1,$$
$$u = (I - R_{t-1} R_{t-1}^\dagger) G_{t-N} e_1, \qquad v = z_{t-N}^\top (I - R_{t-1}^\dagger R_{t-1})$$

对于降阶更新，被移除的行$z_{t-N}^\top$同时属于 R_{t-1}的行空间和列空间，因此

$$u = 0, \qquad v = 0$$

奇异情形:

$$\hat{R}_t^\dagger = R_{t-1}^\dagger - kk^\dagger R_{t-1}^\dagger - R_{t-1}^\dagger h^\dagger h + (k^\dagger R_{t-1}^\dagger h^\dagger)\, kh \tag{19}$$

非奇异情形:

$$\hat{R}_t^\dagger = R_{t-1}^{-1} + \frac{R_{t-1}^{-1} G_{t-N} e_1 z_{t-N}^\top R_{t-1}^{-1}}{1 - z_{t-N}^\top R_{t-1}^{-1} G_{t-N} e_1} \tag{20}$$

在这两种情形下，降阶更新后的伪逆 $R_t^\dagger$ 可通过右乘正交旋转得到:

$$\begin{bmatrix} 0 & R_t^\dagger \end{bmatrix} = \hat{R}_t^\dagger G_{t-N}$$

由此可得到一种完全递归的伪逆降阶更新方法，兼容满秩和秩亏情形。

2.3.7 权重降级更新

为计算迭代权重降阶更新，我们首先注意到， Y_{t-1} 的更新需要移除其第一个元素，这与从特征矩阵中移除第一个观测值 z_{t-N} 类似。由于 y_{t-N} 是一个标量，这一操作可以表示为

$$G_{t-N}^\top Q_{t-1}^\top \begin{bmatrix} y_{t-N} \\ Y_t \end{bmatrix} - \begin{bmatrix} y_{t-N} \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ Q_t^\top Y_t \end{bmatrix}$$

其中 $y_{t-N} = e_1^\top Y_{t-1}$ 表示向量 Y_{t-1} 的第一个元素。这个关系可以重新整理为 s

$$\begin{bmatrix} 0 \\ Q_t^\top Y_t \end{bmatrix} = G_{t-N}^\top Q_{t-1}^\top Y_{t-1} - e_1 y_{t-N}$$

因此，回归权重的完整降阶更新表达式为

$$\begin{aligned} \beta_t &= R_t^\dagger Q_t^\top Y_t \\ &= \left(R_{t-1} - G_{t-N} e_1 z_{t-N}^\top\right)^\dagger G_{t-N} \left(G_{t-N}^\top Q_{t-1}^\top Y_{t-1} - e_1 y_{t-N}\right) \\ &= \left(R_{t-1} - G_{t-N} e_1 z_{t-N}^\top\right)^\dagger \left(Q_{t-1}^\top Y_{t-1} - G_{t-N} e_1 y_{t-N}\right) \end{aligned}$$

最后一步中，根据系统运行的情形——也就是说，特征矩阵是超定还是欠定——代入相应的伪逆或逆矩阵更新公式。

奇异情形:

要应用广义伪逆降阶恒等式，我们需要满足标准值域条件 [6]:

$$G_{t-N} e_1 \in \mathrm{Range}(R_{t-1}), \qquad z_{t-N} \in \mathrm{Range}(R_{t-1}^\top)$$

由于 $G_{t-N} e_1$ 和 z_{t-N} 都是向量，这些条件确保相应的斜投影为零:

$$\left(I - R_{t-1} R_{t-1}^\dagger\right) G_{t-N} e_1 = 0, \qquad z_{t-N}^\top (I - R_{t-1}^\dagger R_{t-1}) = 0^\top$$

在这些假设下，增广系统的秩保持不变：

$$\rho(R_{t-1}) = \rho(\hat{R}_t)$$

因此，降阶更新不会改变 R_{t-1}的列空间。

考虑分块矩阵

$$\hat{R}_{t-1} = \begin{bmatrix} R_{t-1} & G_{t-N}e_1 \\ z_{t-N}^\top & D \end{bmatrix}$$

其舒尔补为 $D = G_{t-N}e_1 R_{t-1}^\dagger z_{t-N}^\top$。根据 [14], ，我们令 $D = 1$ ，以避免引入额外的尺度因子。由此可直接得到归一化

$$1 = z_{t-N}^\top R_{t-1}^\dagger G_{t-N}e_1$$

由于 $kk^\dagger k = k$，矩阵 $kk^\dagger$ 是投影到由向量 $k = R_{t-1}^\dagger G_{t-N}e_1$张成的子空间上的正交投影算子。

由于模型在窗口内插值拟合数据，响应向量 Y_{t-1} 属于 Z_{t-1}的行空间。等价地，估计器 β_{t-1} 满足

$$z_{t-N}^\top \beta_{t-1} = y_{t-N}$$

现在我们推导降阶更新后的系数向量 β_t。

$$\begin{aligned}
\beta_t &= R_t^\dagger Q_t^\top Y_t \\
&= (R_{t-1}^\dagger - kk^\dagger R_{t-1}^\dagger - R_{t-1}^\dagger h^\dagger h + (k^\dagger R_{t-1}^\dagger h^\dagger)\, kh) \\
&\quad (Q_{t-1}^\top Y_{t-1} - G_{t-N}e_1 y_{t-N}) \\
&= \beta_{t-1} - kk^\dagger \beta_{t-1} - R_{t-1}^\dagger h^\dagger z_{t-N}^\top \beta_{t-1} \\
&\quad + (k^\dagger R_{t-1}^\dagger h^\dagger) k z_{t-N}^\top \beta_{t-1} \\
&\quad - k y_{t-N} + kk^\dagger k y_{t-N} \\
&\quad + R_{t-1}^\dagger h^\dagger z_{t-N}^\top R_{t-1}^\dagger G_{t-N}e_1 y_{t-N} \\
&\quad - (k^\dagger R_{t-1}^\dagger h^\dagger) k z_{t-N}^\top R_{t-1}^\dagger G_{t-N}e_1 y_{t-N} \\
&= \beta_{t-1} - kk^\dagger \beta_{t-1} \\
&\quad + (R_{t-1}^\dagger h^\dagger - (k^\dagger R_{t-1}^\dagger h^\dagger)k)(y_{t-N} - z_{t-N}^\top \beta_{t-1}) \\
&\quad - k y_{t-N} + k y_{t-N} \\
\beta_t &= \beta_{t-1} - kk^\dagger \beta_{t-1}
\end{aligned} \tag{21}$$

因此，该更新是将 β_{t-1} 正交投影到与 k 正交的子空间上：

$$\beta_t = (I - kk^\dagger)\, \beta_{t-1}$$

由于 $kk^\dagger$ 是投影到 Range(k) 上的投影算子，降阶更新后的参数满足

$$k^\top \beta_t = 0$$

确保与被丢弃行相关的信息从估计值中彻底移除。

非奇异情形:

$$
\begin{aligned}
\beta_t &= R_t^{\dagger} Q_t^{\top} Y_t \\
&= (R_{t-1}^{\dagger} + \frac{R_{t-1}^{\dagger} G_{t-N} e_1 z_{t-N}^{\top} R_{t-1}^{\dagger}}{(1 - z_{t-N}^{\top} R_{t-1}^{\dagger} G_{t-N} e_1)}) \\
&\quad (Q_{t-1}^{\top} Y_{t-1} - G_{t-N} e_1 y_{t-N}) \\
&= \beta_{t-1} + \frac{R_{t-1}^{\dagger} G_{t-N} e_1 z_{t-N}^{\top} \beta_{t-1}}{(1 - z_{t-N}^{\top} R_{t-1}^{\dagger} G_{t-N} e_1)} \\
&\quad - y_{t-N} R_{t-1}^{\dagger} G_{t-N} e_1 (1 + \frac{z_{t-N}^{\top} R_{t-1}^{\dagger} G_{t-N} e_1}{(1 - z_{t-N}^{\top} R_{t-1}^{\dagger} G_{t-N} e_1)}) \\
&= \beta_{t-1} + \frac{R_{t-1}^{\dagger} G_{t-N} e_1 z_{t-N}^{\top} \beta_{t-1}}{(1 - z_{t-N}^{\top} R_{t-1}^{\dagger} G_{t-N} e_1)} \\
&\quad - \frac{y_{t-N} R_{t-1}^{\dagger} G_{t-N} e_1}{(1 - z_{t-N}^{\top} R_{t-1}^{\dagger} G_{t-N} e_1)} \\
\beta_t &= \beta_{t-1} - \frac{R_{t-1}^{\dagger} G_{t-N} e_1}{(1 - z_{t-N}^{\top} R_{t-1}^{\dagger} G_{t-N} e_1)} (y_{t-N} - z_{t-N}^{\top} \beta_{t-1})
\end{aligned}
\tag{22}
$$

在这一公式中，卡尔曼增益处于其标准位置，作为新息项 $(y_{t-N} - z_{t-N}^{\top}\beta_{t-1})$ 的乘法因子，从而按预测误差的大小更新估计值。

整体递归过程总结于算法 **1** 中，该算法整合了指数加权更新和上述基于 QR 的正交化。

2.3.8 遗忘因子与正则化

在递归最小二乘代价函数中引入遗忘因子 $\lambda \in (0, 1]$，使算法能够通过对过去的观测值进行指数降权来适应非平稳信号。这一机制也起到隐式正则化的作用，改善协方差矩阵 [3, 28] 的数值条件性。

前述推导对应于特殊情况 $\lambda = 1$，此时所有样本的权重相同。然而在实践中，遗忘因子通过对角加权矩阵引入

$$\Lambda_t = \mathrm{diag}(\sqrt{\lambda^{N-1}}, \sqrt{\lambda^{N-2}}, \ldots, \sqrt{\lambda^0}),$$

从而得到指数加权最小二乘问题

$$\min_{\beta} \|\Lambda_t (Z_t \beta_t - Y_t)\|_2^2.$$

通过 QR 分解可得

$$\beta = R_{t-1}^{\dagger} Q_t^{\top} \Lambda_t Y_t$$

为求得　一种带指数遗忘的递归 QR 更新，假设在时间 $t-1$

$$\Lambda_{t-1} Z_{t-1} = Q_{t-1} R_{t-1}$$

引入新的观测值 z_t，遗忘因子为 $\lambda \in (0,1]$ 得到

$$\Lambda_t Z_t = \begin{bmatrix} \sqrt{\lambda}\,\Lambda_{t-1} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} Z_{t-1} \\ z_t^\top \end{bmatrix} = \begin{bmatrix} \sqrt{\lambda}\,Q_{t-1}R_{t-1} \\ z_t^\top \end{bmatrix}$$

提出正交块 diag(Q_{t-1},1) 后，得到

$$\Lambda_t Z_t = \begin{bmatrix} Q_{t-1} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{\lambda}\,R_{t-1} \\ z_t^\top \end{bmatrix} \tag{23}$$

因此，更新可简化为对式 (23) 中堆叠矩阵进行 QR 分解。

加权响应向量的更新类似地表示为

$$\Lambda_t Y_t = \begin{bmatrix} \sqrt{\lambda}\,\Lambda_{t-1} Y_{t-1} \\ y_t \end{bmatrix}$$

相应地，递归分解更新变为

$$R_t = \begin{bmatrix} \sqrt{\lambda}\,R_{t-1} \\ z_t^\top \end{bmatrix}$$

When 更新伪逆矩阵时，遗忘因子通过缩放被吸收 g

$$R_{t-1}^\dagger \leftarrow \frac{1}{\sqrt{\lambda}} R_{t-1}^\dagger$$

然后执行式 (13) 中的降阶步骤。此调整确保逆矩阵能够递归地反映指数加权。详细证明见附录B。

也可以通过对角加载引入正则化，它会将正规方程修改为

$$(Z^\top Z + \delta I)\beta = Z^\top Y$$

从而将奇异或病态系统转化为适定系统 [17, 26]。对角加载（岭回归）通过移动谱来提高稳定性，而 ABO 则通过正交分解本身实现稳定性，使模型能够利用整个零空间吸收方差，而不会引入由 δI 带来的偏差。在我们的设定中，我们有意避免这种显式正则化，因为我们的目标是研究过参数化情形，在这些情形下，模型可以有意地插值数据；而岭回归正则化则完全不允许这样做。

2.4 QR–EWRLS算法

现在总结所提出的基于 QR 的指数加权递归最小二乘（QR–EWRLS）流程。算法 1 展示了完整的更新周期。以下定理证明，在理想化的算术条件下，该算法计算得到的是最小范数的指数加权滑动窗口最小二乘解。

算法 1. 通过 QR-EWRLS 实现自适应良性过拟合（ABO）

要求： 数据流 $\{(x_t,\ y_t)\}$，窗口大小 N，遗忘因子 λ，RFF维度 D

确保： 在线预测 $\{\hat{y}_t\}$，系数向量 $\{\beta_t\}$1: **初始化：** 2: 计算初始 $Z_0 \in \mathbb{R}^{N\times D}$ 和 $Y_0 \in \mathbb{R}^N$

3: 执行 QR 分解：$Z_0 = Q_0 R_0$ ▷ [(5)]

4: 求解 $R_0\beta_0 = Q_0^\top Y_0$ 以得到最小范数 β_0 ▷ [(9)]

5: **对于** $t = N+1,\ N+2,$ **.....执行**6: **步骤 1：RFF 映射** 7: $z_t \leftarrow \sqrt{\frac{2}{D}}\cos(A^\top x_t + b)$8: **步骤 2：非平稳缩放**9: $R_{t-1} \leftarrow \sqrt{\lambda}R_{t-1}$, $Y_{t-1} \leftarrow \sqrt{\lambda}Y_{t-1}-$ ▷ [(11)] ▷ [§2.3.8]

10: **步骤 3：观测更新（吉文斯旋转）** [(12)]

11: $\bar{R}_t \leftarrow \begin{bmatrix} R_{t-1} \\ z_t^\top \end{bmatrix}$, $Y_t \leftarrow \begin{bmatrix} Y_{t-1} \\ y_t \end{bmatrix}$12: 应用 G_t 使得 $G_t^\top \bar{R}_t \leftarrow \begin{bmatrix} R_t \\ \mathbf{0}^\top \end{bmatrix}$13: $Q_t \leftarrow \begin{bmatrix} Q_{t-1} & 0 \\ 0 & 1 \end{bmatrix} G_t$14: **步骤 4：伪逆更新**

15: $h \leftarrow z_t^\top R_{t-1}^\dagger$16: **如果** $D \le N$ **则**

17: $\bar{R}_t^\dagger \leftarrow \begin{bmatrix} R_{t-1}^\dagger - \frac{R_{t-1}^\dagger h^\top}{1+hh^\top} h & \frac{R_{t-1}^\dagger h^\top}{1+hh^\top} \end{bmatrix}$18: **否则** ▷ 经典区域 [(13)]

19: $\bar{R}_t^\dagger \leftarrow \begin{bmatrix} R_{t-1}^\dagger - c^\dagger h & c^\dagger \end{bmatrix}$ ▷过参数化区域 [(13)]

20: $\bar{R}_t^\dagger G_t \leftarrow R_t^\dagger$ 21: **步骤 5：回归权重** ▷ 最小范数解

22: **如果** $D \le N$ **则**

23: 24: **否则**$\beta_t \leftarrow \beta_{t-1} + \frac{R_{t-1}^\dagger h^\top}{1+hh^\top}(y_t - z_t^\top\beta_{t-1})$ ▷ 经典区域 [(15)]

25: $\beta_t \leftarrow \beta_{t-1} + c^\dagger(y_t - z_t^\top\beta_{t-1})$ 26: **步骤 6：滑动窗口降阶更新** ▷过参数化区域 [(15)]

27: 确定旋转 G_{t-N} 以隔离 Q_{t-1}的第一行 28: 更新 R_t和 Y_t以移除 $(z_{t-N},\ y_{t-N})$ ▷ [§2.3.5]

29: 将 R_t 保持为上梯形 30: **步骤 7：伪逆降阶更新**

31: $h \leftarrow z_{t-N}^\top R_{t-1}^\dagger$, $k \leftarrow R_{t-1}^\dagger G_{t-N} e_1$32: **如果** $D \le N$ **则** ▷ 经典区域 [(20)]

33: $\hat{} R_t^\dagger \leftarrow R_{t-1}^{-1} + \frac{k z_{t-N}^\top R_{t-1}^{-1}}{1 - z_{t-N}^\top k}$34: **否则** ▷过参数化区域 [(19)]

35: $\hat{} R_t^\dagger \leftarrow R_{t-1}^\dagger - kk^\dagger R_{t-1}^\dagger - R_{t-1}^\dagger h^\dagger h + (k^\dagger R_{t-1}^\dagger h^\dagger)\ kh$

36: $\begin{bmatrix} 0 & R_t^\dagger \end{bmatrix} \leftarrow \hat{} R_t^\dagger G_{t-N}$37: **步骤8：权重降阶更新**38: **如果** $D \le N$ **则** ▷经典区域[(22)]

39: $\beta_t \leftarrow \beta_{t-1} - \frac{k}{(1 - z_{t-N}^\top k)}(y_{t-N} - z_{t-N}^\top\beta_{t-1})$

40: **否则** ▷过参数化区域 [(21)]

41: $\beta_t \leftarrow \beta_{t-1} - kk^\dagger\beta_{t-1}$42: **步骤9：预测**43: $\hat{y}_t \leftarrow z_t^\top\beta_t$ ▷ 预测下一个值

ˆ定理 1 (与最小范数指数加权最小二乘的一致性)。 *Let β_t denote the unique minimum-ℓ_2-norm solution of the exponentially weighted sliding-window least-squares problem at time t,*

$$\hat{\beta}_t = \arg\min_{\beta} \ \|\beta\|_2 \quad s.t. \quad \beta \in \arg\min_{\gamma} \sum_{i=0}^{N-1} \lambda^i \|y_{t-i} - z_{t-i}^\top \gamma\|_2^2 \tag{24}$$

Let β_t be the estimate produced by Algorithm 1. Assume exact arithmetic, admissible rank-preserving updates anddowndates, and initialization $\beta_0 = \hat{}\beta_0$. Then, for all $t \geq 0$, the algorithmic iterate coincides with the minimum-normsolution,

$$\beta_t = \hat{\beta}_t.$$

Proof. 证明通过对 t进行归纳法得出。
基本情形（$t = 0$）： 算法 1 通过对初始窗口化设计矩阵进行显式批量 QR 分解来初始化 β_0 。因此，得到的伪逆解与最小-ℓ_2-范数解 ˆ β_0一致。
I归纳步骤： 假设 $\beta_{t-1} =$ $\hat{\beta}_{t-1}$。从 $t-1$ 到 t 的更新包含两个操作 .
首先，通过 $\sqrt{\lambda}$ 进行缩放并按 $(z_t,\ y_t)$进行行增广，恰好等价于将加权观测 $\lambda^0 z_t z_t^\top$加入指数加权正规方程。根据 Greville 伪逆更新的性质，得到的中间估计值仍然是增广系统的最小范数解。

其次，降阶步骤会应用正交变换，以分离最早的加权观测 $\lambda^N z_{t-N} z_{t-N}^\top$ 的贡献，随后进行伪逆降阶更新。在满足可采纳性假设的条件下，此操作在代数上等价于从正规方程中移除该项，同时保持最小范数性质。

由于这两个步骤准确实现了指数加权滑动窗口目标中各项的添加与移除，因此更新后的估计值满足 $\beta_t = \hat{}\beta_t$。

□

3 实验

3.1 合成时间序列生成

遵循 [23], ，我们模拟一个定义如下的非线性自回归过程

$$x_t = \frac{2x_{t-1}}{1 + 0.8x_{t-1}^2} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{U}(-1, 1), \tag{25}$$

以 $x_0 \sim \mathcal{U}(-1,\ 1)$ 初始化。该过程呈现有界混沌动力学，常用于评估非线性预测方法。根据模拟序列 $\{x_t\}_{t=1}^{10500}$，我们构造滞后特征向量

$$\mathbf{x}_t = (x_{t-1}, x_{t-2}, \ldots, x_{t-7}), \tag{26}$$

这些变量用于预测当前值 x_t。随后，我们将这 7 个时滞变量转换为 *Random FourierFeatures (RFF)* 个所需维度的特征，从而实现非线性特征映射，同时保留平移不变核函数所诱导的内积结构。因此，模型利用前 7 个时滞变量经随机傅里叶特征变换后的表示来预测 x_t 。

3.2 数据预处理

对所有特征进行标准化，以确保不同尺度间的可比性，并近似满足同方差性，即方差在时间上保持恒定 [12]。标准化采用以下方式进行：

$$z_{score} = \frac{x - \mu}{\sigma}, \tag{27}$$

其中， μ 和 σ 分别表示每个特征的均值和标准差。变换后，所有特征均满足 $\mu = 0$ 和 $\sigma = 1$。

随着新观测值的引入，我们递归地进行标准化，以确保整个数据流 [16]中的缩放保持一致。

3.3 稳定性分析

为评估所提出更新的数值稳定性，我们进行了两项互补分析：（i）经验残差跟踪；（ii）系统矩阵条件性。

3.3.1 经验残差误差

经验残差量化了顺序更新引入的累积数值误差。在每次迭代t中，我们将残差定义为

$$r_t = \|y_t - z_t^\top \beta_t\|_2,$$

并跟踪 $\{r_t\}_{t=1}^N$ 随时间变化的均值和标准差。稳定更新表现为有界残差，不会随着 t增加而发散。

我们报告以下诊断指标：

- **平均残差：** $\bar{r} = \frac{1}{T}\sum_t r_t$
- **残差方差：** $\sigma_r^2 = \frac{1}{T-1}\sum_t (r_t - \bar{r})^2$

r_t 较低的方差表明对数值漂移具有鲁棒性。

3.3.2 条件数的演变

条件数 可作为设计矩阵或相关矩阵数值敏感性的替代指标。在每一步 我们计算

$$\kappa_t = \mathrm{cond}(R_t) = \frac{\sigma_{\max}(R_t)}{\sigma_{\min}(R_t)},$$

其中 $\sigma_{\max}$ 和 $\sigma_{\min}$ 分别是 R_t的最大和最小奇异值 . 良态矩阵满足 $\kappa_t \ll 10^{10}$，而 κ_t 的急剧增大则表明正交性丧失或条件性变差。

3.4 不使用遗忘因子时的结果（$\lambda = 1$）

在本次实验中，我们将遗忘因子固定为 $\lambda = 1$，这对应于经典递归最小二乘（RLS）形式，即对所有过去样本赋予相同权重。其余超参数如下：

表 1. 实验超参数。

参数	值
更新次数	10000
窗口大小	20
核函数方差（随机傅里叶特征）	1

该设置作为基线，用于评估我们提出的更新方法在不采用指数加权时的稳定性和残差动态。

表 2 汇总了不同随机傅里叶特征（RFF）维度下残差和条件数的均值与方差，我们将这些维度记为 D。在本表及以下各表中，我们用粗体标出在所有特征维度中使测试残差方差和均方误差最小的数值。正如预期，模型在插值区域附近表现出越来越明显的数值不稳定性：对于中等大小的特征维度（例如 $D = 16$），残差方差和条件数均急剧上升；随后，随着模型变得过参数化，模型趋于稳定。最优值出现在边界情形，并且很可能会继续单调下降。

表 2. $\lambda = 1$ 的残差均值和方差

$\log_2(D)$	训练残差		测试残差	
	Mean	方差	Mean	方差
1	0.8121	0.9808	1.0008	1.4941
2	0.7061	0.8256	1.0966	2.0205
3	0.4556	0.4147	1.3103	3.5529
4	0.0813	0.0211	3.1712	69.2871
I*	0.0000	0.0000	819.825	1.2×10^9
5	0.0000	0.0000	1.5744	8.2697
6	0.0000	0.0000	0.7797	1.2136
7	0.0000	0.0000	0.6197	0.6885
8	0.0000	0.0000	0.6013	0.6142
9	0.0000	0.0000	0.5749	0.5378
10	0.0000	0.0000	0.5534	0.4957
11	0.0000	0.0000	0.5547	0.4982
12	0.0000	0.0000	0.5452	0.4809
13	0.0000	0.0000	0.5447	0.4806
14	0.0000	0.0000	**0.5420**	**0.4738**

* 插值极限对应于 $D = N = 20$，即 $\log_2(20) \approx 4.32$。

3.5 遗忘因子下的结果 ($\lambda = 0.9$)

现在，我们引入遗忘因子 $\lambda = 0.9$，其对应的有效记忆长度为

$$N_{\text{eff}} = \frac{1}{1-\lambda} = 10$$

这种方式会按指数衰减降低较早观测值的权重，以适应非平稳动力学 [17, 28]。其余所有参数均与上一个实验相同。

表3表明，$\lambda = 0.9$ 的残差统计与使用 $\lambda = 1$得到的结果几乎相同。在过参数化区域（$D \gg N$）中，解由基于QR的伪逆的隐式最小范数偏置所支配，从而使指数重加权对稳态预测性能基本无关紧要。

表 3. $\lambda = 0.9$ 残差的均值和方差

$\log_2(D)$	训练残差		测试残差	
	Mean	方差	Mean	方差
1	0.6267	0.6399	0.9899	1.5632
2	0.4337	0.3549	1.1232	2.2689
3	0.1727	0.0756	1.3308	3.8850
4	0.0135	0.0008	3.3350	73.7217
I*	0.0000	0.0000	78534.6	1.02×10^{13}
5	0.0000	0.0000	1.5744	8.2697
6	0.0000	0.0000	0.7797	1.2136
7	0.0000	0.0000	0.6197	0.6885
8	0.0000	0.0000	0.6013	0.6142
9	0.0000	0.0000	0.5749	0.5378
10	0.0000	0.0000	0.5534	0.4957
11	0.0000	0.0000	0.5547	0.4982
12	0.0000	0.0000	0.5452	0.4809
13	0.0000	0.0000	0.5447	0.4806
14	0.0000	0.0000	**0.5420**	**0.4738**

3.6 条件数分析

表4报告了无遗忘因子情形（$\lambda = 1$）下的条件数统计。随着特征数量增加并接近插值极限（$D \approx N$），条件数显著增大，表明系统矩阵的数值条件性正在恶化。这一行为与插值阈值附近最小二乘更新的已知敏感性一致，此时数据中的微小扰动可能导致估计参数 [13, 18]出现大幅变化。

表 4. $\lambda = 1$ 的条件数均值和方差

$\log_2(D)$	**Mean**	**方差**
1	1.2801	0.0387
2	1.7820	0.0966
3	4.0124	1.1380
4	21.5274	127.6520
I*	∞	NA
5	15.4765	50.3692
6	7.8396	12.6857
7	5.9778	7.5213
8	5.4786	6.5071
9	5.3213	6.7122
10	5.1483	6.1914
11	5.1902	6.3840
12	5.0775	6.2375
13	5.0659	6.0952
14	5.0563	6.1619

3.7 双下降

在此图中，随着随机特征维度 D 接近插值阈值，模型的测试误差显著上升，随后在过参数化区域内下降。这种行为是双下降现象 [5, 15], 的典型特征，反映了从欠参数化到精确插值的转变。本实验对插值阈值进行了明确采样。虚线表示理论插值点 $I^* \approx 4.32$，对应于随机特征数量与窗口大小有效相等的情况。观测到的测试误差峰值出现在该阈值附近，证实泛化性能的下降与插值区域有关。

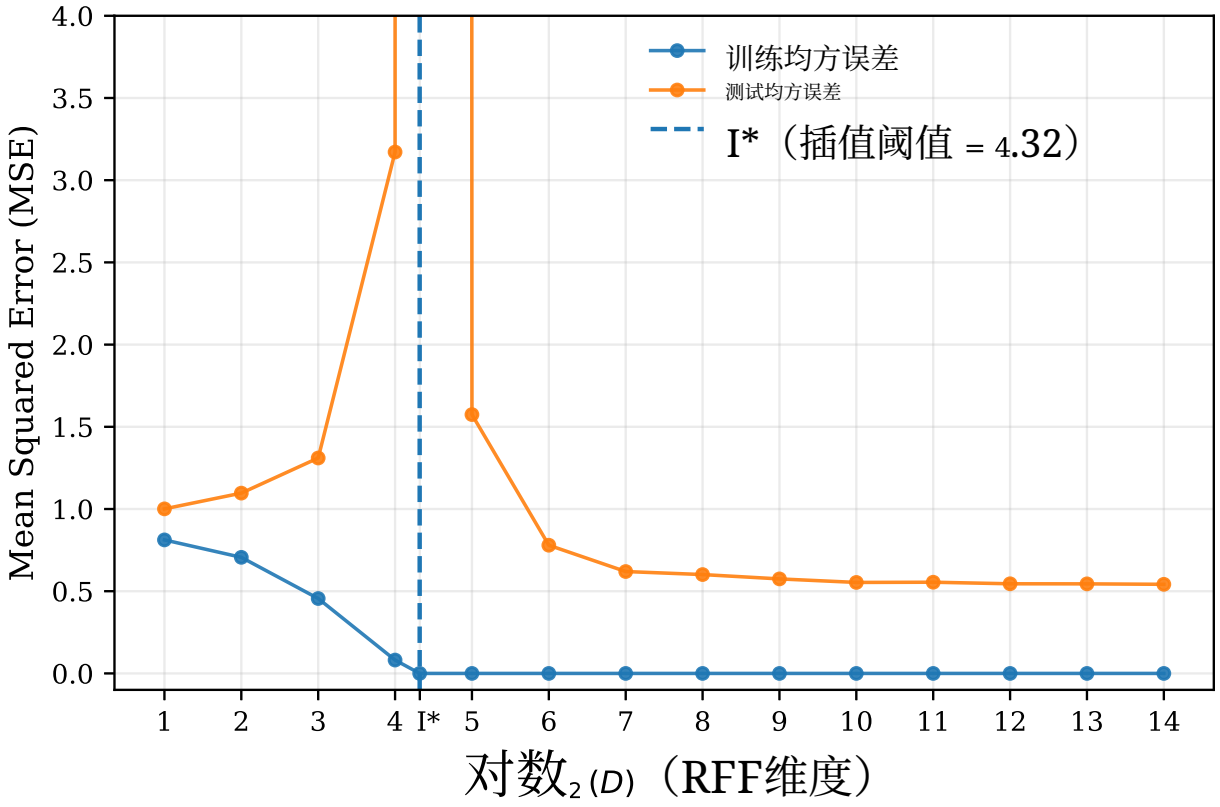

图 2. 测试误差的双下降行为

3.8 计算速度

使用 Google Benchmark (v1.9.4) 对计算时间进行评估，测试时链接了优化后的 BLAS 和 LAPACK 程序，并采用单线程执行。基准测试在配备 32 GB 内存的第 13 代 Intel® CoreTM i7-1370P @ 1.90 GHz 上进行（Windows 11、WSL2 Ubuntu 22.04 LTS），并禁用了 CPU 频率缩放和 ASLR，以确保计时结果可复现。表 5 报告了十次重复测试中 1,000 个连续预测–更新步骤的均值、标准差和变异系数 (CV)。

实证结果表明，总运行时间随RFF维度 D增加，范围从 $D = 2^1$ 处的6.11毫秒到 $D = 2^{14}$处的约11.97秒。在值域的大部分范围内，相对于 D 的缩放保持近似线性，这与基于QR的表述在每个更新步骤上的理论 $O(ND)$复杂度一致。虽然变异系数在测试值域的边界处保持较低，但在 $D = 2^9$ 与 $D = 2^{13}$ 之间观察到方差的短暂上升（变异系数峰值达36.64%），这可能是由于矩阵维度缩放时的内存层次结构效应或缓存未命中所致。尽管如此，在 $D = 2^{14}$ 处变异系数回落至4.20%的低水平，证实了所提出的递归实现在规模扩展时的**数值稳定性** 与**可预测的计算性能** 。

表 5. ABO RLS 递归最小二乘更新+预测运行时间与 RFF维度 D （1000 次顺序更新；Google Benchmark；重复 10 次）。

$\log_2(D)$	平均 CPU 时间（毫秒）	标准差（毫秒）	变异系数（%）
1	6.11	0.19	3.14
2	6.25	0.05	0.79
3	7.59	0.30	3.95
4	9.22	0.18	1.99
5	13.25	0.19	1.46
6	19.79	0.44	2.22
7	36.00	2.11	5.86
8	64.30	1.14	1.77
9	438.02	160.49	36.64
10	887.13	292.02	32.92
11	1495.53	290.62	19.43
12	2699.60	492.79	18.25
13	5508.04	832.17	15.11
14	11 971.30	502.58	4.20

4 结果与基线比较

我们在真实世界的非平稳时间序列数据上评估所提出的自适应良性过拟合方法（ABO/ QR–EWRLS）。

所有实验均严格遵循在线预quential评估协议。我们首先使用八个验证折进行超参数选择，然后在五个互不重叠的测试折上评估性能。
对于每个折，模型都使用一个大小等于所选窗口长度的批次进行初始化，随后逐个观测值依次进行更新和预测。

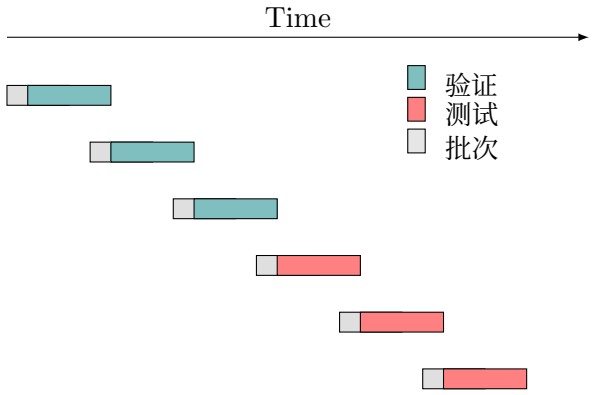

图3。采用重叠窗口进行向前滚动验证，测试窗口之间严格不重叠，且每个测试窗口之前均进行批次初始化

4.1 真实世界数据上的基线性能

作为比较，我们报告窗口化的基于QR分解的递归最小二乘（QRD-RLS）算法 [25] 和窗口化的RBF核递归最小二乘法 [37] 的结果。这两种方法在自适应滤波领域已有充分研究，并被广泛用作窗口化和在线学习场景中的数值稳定基线。

4.1.1 欧元/美元高频预测

欧元/美元数据集由高频外汇数据构成，时间跨度为 24/11/2025 至 02/01/2026，采样频率为每分钟一次，共约 16,800 个连续观测值。数据来源于 Dukas Copy Market Data 数据库。[4]

实验采用滚动折评估协议，在严格的在线环境下进行。每折包含 960 个用于验证的观测值和 1920 个用于测试的观测值，相当于每折约 16 小时的市场活动。模型输入由滞后观测值构成，并采用有效的窗口化更新方案进行学习。

使用 Optuna 在 8 个验证折上选择超参数 [1]。最终性能在 5 个互不重叠、未用于调参的测试折上报告。所有报告的结果均为各测试折的平均值。

4.1.2 电力负荷预测

此外，我们还在电力负荷图数据集 [35]，上评估了所提出的方法。该数据集包含 2011年至2014年期间葡萄牙电网的电力消耗测量数据，采样频率为 15 分钟。

实验采用严格的在线评估协议，使用滚动折进行，其中每折包含 672 个用于验证的观测值和 1344 个用于测试的观测值，相当于每折约一周的数据。验证和测试均采用相同的滚动折结构，且每折内的片段互不重叠。实验使用滞后输入和固定大小的窗口化学习方案，在验证片段上使用 Optuna 选择超参数，并在对应的测试片段上报告最终性能。

与高频金融场景相比，该数据集呈现出显著的日周期性和周周期性，且噪声水平大幅降低，为评估预测性能和数值稳定性提供了互补基准。

4.2 超参数选择

超参数通过使用 Optuna [1]实现的贝叶斯优化进行选择。对于每种方法，搜索空间和试验次数均已固定 $a\ priori$，并在所有折数中保持一致。优化在独立于测试数据的验证集上进行，最终报告的结果对应于在留出测试折上评估的最佳验证配置。

所有模型均在相同的预测任务上进行评估，并采用一致的预测起始索引，从而确保性能差异完全源于模型结构和学习动态。

对于两个数据集，所有模型的滞后阶数均固定为 $L = 20$ 。对于 ABO 和 KRLS-RBF，需调优窗口大小和核带宽；而对于 QRD-RLS，仅优化窗口大小，这反映了线性递归最小二乘中不存在核参数。相应的 Optuna 搜索空间汇总于表 6。

表 6. Optuna 超参数搜索空间。

数据集	模型	折数	L	调优参数
欧元/美元（1 分钟）	ABO	$\{0, 1, ..., 8\}$	20	W, σ
	KRLS-RBF	$\{0, 1, ..., 8\}$	20	W, σ
	QRD-RLS	$\{0, 1, ..., 8\}$	20	W
电力（15分钟）	ABO	$\{0, 1, ..., 8\}$	20	W, σ
	KRLS-RBF	$\{0, 1, ..., 8\}$	20	W, σ
	QRD-RLS	$\{0, 1, ..., 8\}$	20	W

5 结果

5.1 实验设置

实验在固定滞后阶数 $L = 20$ 和遗忘因子 $\lambda = 1$ 的条件下进行。核函数超参数（σ）和窗口大小（W）通过 Optuna 进行优化。为确保数值稳定性，QRD-RLS 和 KRLS-RBF 使用

[4]https://www.dukascopy.com/swiss/english/marketwatch/historical/

对角正则化参数 10^{-2}。除非另有说明，ABO 的随机傅里叶特征（RFF）维度固定为 $D = 8192$ （$\log_2 m = 13$）。

5.2 欧元/美元预测结果

由于噪声水平高且时间结构较弱，欧元/美元数据集是一项具有挑战性的任务。性能指标见表7。

在该情形下，非线性表示至关重要；ABO与KRLS-RBF在残差均方误差（ResMSE）和残差方差（ResVAR）方面均显著优于线性QRD-RLS基线。ABO在无需显式核表示的情况下保持稳定的在线更新，同时取得与KRLS-RBF相当的预测性能。

表7. 欧元/美元平均预测性能与相对更新时间（$L = 20,\ D = 8192$）。

模型	W	σ	残差均方误差	残差方差	时间（×ABO）
ABO（本文方法）	21	8.0	1.3587	1331.63	1.00
KRLS-RBF	421	0.31	1.3559	1335.34	1.70
QRD-RLS	272	NA	2.3294	2494.01	0.02

5.3 电力负荷预测结果

与金融数据相比，电力负荷序列呈现出更强的时间结构和更低的噪声。结果汇总于表 8。

在这种结构化环境下，QRD-RLS 的残差均方误差和残差方差最低，凸显了线性自适应滤波的有效性。ABO 的预测准确度与 KRLS-RBF 不相上下，但由于无需显式进行核函数计算，其计算性能更优。

表 8. 平均电力负荷预测性能和相对更新时间（$L = 20,\ D = 8192$）。

模型	W	σ	残差均方误差	残差方差	时间（×ABO）
ABO（我们的方法）	21	6.5	1.0096	10.92	1.00
KRLS-RBF	761	0.32	1.0110	10.87	1.28
QRD-RLS	272	NA	0.9521	9.06	1.9×10^{-3}

5.4 对随机特征维度的敏感性

表 9 和表 10 单独考察了随机特征维度 D 对 ABO 的影响。在两个数据集上，将 D从 2048 增加到 8192，预测准确度均有所提升，但幅度不大，这与随机傅里叶特征核近似理论相符。然而，这一提升使端到端实际运行时间增加约 $6\times$，反映了基于随机特征的自适应滤波每次更新的计算复杂度为 $O(ND)$。

表 9. D 对 ABO 的影响（欧元/美元，$L = 20,\ W = 21,\ \sigma = 8.0$）。

D	残差均方误差	残差方差	T_{e2e} (ms)	时间(×2048)
2048	1.3711	1339.80	9.11	1.00
8192	1.3587	1331.63	55.32	6.08

表10. D 对 ABO 的影响（电力，$L = 20,\ W = 21,\ \sigma = 6.5$）。

D	残差均方误差	残差方差	T_{e2e} (毫秒)	时间(×2048)
2048	1.0189	10.86	8.52	1.00
8192	1.0096	10.92	52.29	6.13

6 讨论

实验结果验证了 ABO 框架能够有效地将“良性过拟合”现象扩展到递归的非平稳场景。接下来的小节将分析驱动这些观测结果的数学机制。

6.1 隐式正则化与最小范数解

在我们的随机傅里叶特征实验中观察到的“双下降”行为证实，QR-EWRLS更新执行了隐式正则化。正如备注 2.1 所述，当 $D > N$ 时，估计器收敛到唯一的最小-ℓ_2-范数解。这一性质在非平稳环境中至关重要；它确保方差分布在特征谱的“尾部”。因此，只有当高频RFF分量显著有助于降低残差时，它们才会被激活，从而使模型能够插值信号，同时保持对噪声的鲁棒性——这正是良性过拟合的标志。

6.2 欠定区域中的数值稳定性

过参数化RLS面临的核心挑战是格拉姆矩阵的奇异性。在标准协方差形式的递归最小二乘中，当 $D > N$ 时，更新 P_t 会变得病态，导致灾难性的发散。通过利用吉文斯旋转进行正交三角更新，ABO 直接维护因子 R_t。即使系统严重欠定，R 因子的条件数仍保持有界。通过 Greville/Cline方法进行的 Moore-Penrose伪逆更新可确保权重向量 β_t 保持稳定，并避免标准求逆中常见的零空间放大。

6.3 跟踪与过参数化

在非平稳环境中，遗忘因子 λ 与特征维度 D之间存在固有的权衡。我们的结果表明，过参数化不会降低跟踪速度；相反，增加的自由度使滤波器能够更灵活地响应状态转移。QR更新的 $O(ND)$ 复杂度确保这种容量提升不会带来过高的计算成本，从而能够在高维空间中实现实时自适应。

6.4 并行可扩展性与集成平均

RFF生成的解耦特性使其能够自然扩展至分布式架构。由于在递归最小二乘更新之前，各个特征分量彼此独立，因此可以将特征空间划分到多个处理核心上。在独立核心上训练子模型，随后对估计器取平均——类似于在线自助聚合（装袋法）——能够进一步降低方差。这种分布式方法不仅降低了单次更新延迟，还能提高估计器对经验时间序列中常见的重尾噪声和异常值的稳定性。

7 结论

本文提出了 *Adaptive Benign Overfitting*（ABO），这是一个连接现代过参数化理论与经典自适应滤波器设计的框架。通过利用随机傅里叶特征（RFF）映射，将递归最小二乘（RLS）算法扩展到 $D \gg N$ 领域，我们证明了“良性过拟合”——传统上在批次设置下进行分析——能够有效应用于非平稳环境中的在线学习。

我们的主要技术贡献是开发了一种基于 QR 分解的指数加权递归最小二乘（QR-EWRLS）更新机制。与标准的协方差形式的递归最小二乘不同，后者在秩亏情形下容易出现数值不稳定性；我们的正交三角形式通过递归更新 Moore–Penrose 伪逆，确保了数值完整性。我们通过实验证实，即使特征维度显著超过观测值数量，该方法仍能复现“双下降”泛化曲线，同时保持条件数稳定。

此外，随机傅里叶特征映射的解耦特性表明，其在计算加速方面具有巨大潜力。未来的工作将研究并行化集成架构：将高维特征空间的子段分配到各个独立核心上进行处理，再通过自助聚合（装袋法）或方差缩减平均法进行汇总。这种分布式方法不仅能进一步降低单次更新延迟，还有望增强估计器在重尾噪声环境下的鲁棒性。

A 秩一更新与降阶更新

在得出最终的 QR–EWRLS 公式之前，我们曾尝试采用经典的协方差形式递归最小二乘（RLS）架构，其中格拉姆矩阵定义为 $A = X^\top X$。尽管这种公式在概念上直观明了，但众所周知，在长序列或回归量近乎共线的情况下，由于 A 的病态性，它容易出现数值不稳定性。

为同时处理满秩和秩亏情形，我们采用了统一的递归策略：对于奇异矩阵，使用 Campbell 和 Meyer 的广义矩阵求和恒等式 [6]；对于非奇异情形，则使用 Sherman–Morrison 公式进行相应的秩一更新。该公式能够显式处理这两种情形，并为后文介绍的稳定 QR 方法奠定理论基础。

A.1 Sherman–Morrison 更新与降阶更新公式

为保持滚动窗口长度固定，必须移除最旧样本 x_{t-N} 的贡献。在非奇异情形下，Sherman–Morrison 公式给出了逆矩阵秩一更新的经典形式：

$$(A_{t-1} + x_t x_t^\top)^{-1} = A_{t-1}^{-1} - \frac{A_{t-1}^{-1} x_t x_t^\top A_{t-1}^{-1}}{1 + x_t^\top A_{t-1}^{-1} x_t}.$$

类似地，应用逆矩阵秩一降阶更新可得

$$(A_{t-1} - x_{t-N} x_{t-N}^\top)^{-1} = A_{t-1}^{-1} + \frac{A_{t-1}^{-1} x_{t-N} x_{t-N}^\top A_{t-1}^{-1}}{1 - x_{t-N}^\top A_{t-1}^{-1} x_{t-N}}.$$

A.2 奇异情形下的增秩更新与降秩降阶更新

当 A_{t-1} 变为奇异时，经典的逆更新不再适用。在这种情况下，可以使用 Campbell 和 Meyer 的广义逆恒等式：

$$c_t = x_t^\top (I - A_{t-1}^\dagger A_{t-1})$$

l由此得到伪逆的 Sherman–Morrison 型增秩更新公式：

$$(A_{t-1} + x_t x_t^\top)^\dagger = A_{t-1}^\dagger - A_{t-1}^\dagger x_t (c_t^\dagger)^\top - c_t^\dagger x_t^\top A_{t-1}^\dagger + \frac{c_t^\dagger (c_t^\dagger)^\top}{1 + x_t^\top A_{t-1}^\dagger x_t}$$

对于降阶更新，我们定义辅助量

$$h = x_{t-N}^\top A_{t-1}^\dagger, \qquad k = A_{t-1}^\dagger x_{t-N},$$
$$u = (I - A_{t-1} A_{t-1}^\dagger) x_{t-N}, \qquad v = x_{t-N}^\top (I - A_{t-1}^\dagger A_{t-1}),$$

l由此得到伪逆的 Sherman–Morison 型降秩降阶公式 e:

$$(A_{t-1} - x_{t-N} x_{t-N}^\top)^\dagger = A_{t-1}^\dagger - k k^\dagger A_{t-1}^\dagger - A_{t-1}^\dagger h^\dagger h + (k^\dagger A_{t-1}^\dagger h^\dagger) k h$$

尽管这一公式能够正确处理秩亏和满秩两种情形，但在实际应用中，它对舍入误差和条件性非常敏感，尤其是在长序列中。Staub [30] 报告了递归更新最小二乘伪逆时出现的类似不稳定性

$$A^\dagger = C^\top (C C^\top)^{-1} (B B^\top)^{-1} B^\top$$

其中，在秩分解中显式使用格拉姆矩阵会放大数值误差。这些局限最终促使我们在经典区域和插值区域采用基于 QR 的正交–三角更新，而在非经典情形下采用正交–梯形更新。我们将在第 2 节对此进行说明。QR 更新方法通过保持正交性并避免显式矩阵求逆来确保数值稳定性。

B 遗忘因子缩放的证明

我们证明，在更新之前用遗忘因子 $\lambda^{-1/2}$ 对 $A_{t-1}^{\dagger}$ 进行缩放，可以得到$A_t^{\dagger}$ 的正确递推计算。证明通过展开$A_t^{\dagger}A_t$ 并利用残差分量 c_t的正交性得出。

$$A_t = \begin{bmatrix} \sqrt{\lambda}A_{t-1} \\ a_t \end{bmatrix}, \qquad A_t^{\dagger} = \begin{bmatrix} B_{t-1} & b_t \end{bmatrix}$$

展开乘积 $A_t^{\dagger}A_t$:

$$A_t^{\dagger}A_t = \sqrt{\lambda}B_{t-1}A_{t-1} + b_t a_t,$$
$$A_t^{\dagger}A_t A_{t-1}^{\dagger} = \sqrt{\lambda}B_{t-1}A_{t-1}A_{t-1}^{\dagger} + b_t a_t A_{t-1}^{\dagger}$$

由此可得:

$$A_{t-1}^{\dagger} = \sqrt{\lambda}B_{t-1} + b_t a_t A_{t-1}^{\dagger}$$
$$B_{t-1} = \frac{1}{\sqrt{\lambda}}\left(A_{t-1}^{\dagger} - b_t a_t A_{t-1}^{\dagger}\right)$$

代入 $A_t^{\dagger}$ 可得:

$$A_t^{\dagger} = \begin{bmatrix} \frac{1}{\sqrt{\lambda}}\left(A_{t-1}^{\dagger} - b_t a_t A_{t-1}^{\dagger}\right) & b_t \end{bmatrix} \tag{28}$$

定义 c_t 为 a_t 在 A_{t-1}的列空间的正交补方向上的分量:

$$c_t = a_t - a_t A_{t-1}^{\dagger}A_{t-1}, \tag{29}$$
$$c_t = a_t(I - A_{t-1}^{\dagger}A_{t-1}). \tag{30}$$

将 (30) 代入 (28) 得:

$$A_t^{\dagger}A_t = A_{t-1}^{\dagger}A_{t-1} - b_t a_t A_{t-1}^{\dagger}A_{t-1} + b_t a_t$$
$$A_t^{\dagger}A_t A_{t-1}^{\dagger} = A_{t-1}^{\dagger}A_{t-1}A_{t-1}^{\dagger} + b_t c_t A_{t-1}^{\dagger}$$

因此,

$$A_{t-1}^{\dagger} = A_{t-1}^{\dagger} + b_t c_t A_{t-1}^{\dagger}$$
$$c_t A_{t-1}^{\dagger} = 0$$

这表明 c_t 与 $A_{t-1}^{\dagger}$ 的行空间正交，因此也与 A_{t-1} 的列空间正交。由于 $c_t^{\dagger} = \frac{c_t^{\top}}{\|c_t\|^2}$ 是 $c_t^{\top}$ 的标量倍，我们有 $c_t^{\dagger}c_t = 1$。

——

$$c_t c_t^{\dagger} = a_t c_t^{\dagger} - a_t A_{t-1}^{\dagger}A_{t-1}c_t^{\dagger}$$
$$c_t^{\dagger}c_t = a_t c_t^{\dagger}$$
$$a_t c_t^{\dagger} = 1$$

定义:

$$P_t = A_{t-1}^{\dagger}A_{t-1} + c_t^{\dagger}c_t$$

则:

$$a_t P_t = a_t A_{t-1}^{\dagger}A_{t-1} + a_t c_t^{\dagger}c_t$$
$$a_t P_t = a_t A_{t-1}^{\dagger}A_{t-1} + c_t$$
$$a_t P_t = a_t$$

此外,

$$A_{t-1}P_t = A_{t-1}$$

因此， P_t 是 A_t的右单位元:

$$P_t = A_t^{\dagger}A_t$$

Now,

$$
\begin{aligned}
c_t^\dagger c_t &= P_t - A_{t-1}^\dagger A_{t-1} \\
c_t^\dagger c_t &= A_{t-1}^\dagger A_{t-1} - b_t a_t A_{t-1}^\dagger A_{t-1} \\
&\quad + b_t a_t - A_{t-1}^\dagger A_{t-1} \\
c_t^\dagger c_t &= b_t c_t \\
c_t^\dagger &= b_t
\end{aligned}
$$

代回后得到:

$$
A_t^\dagger = \begin{bmatrix} \frac{1}{\sqrt{\lambda}}\left(A_{t-1}^\dagger - c_t^\dagger a_t A_{t-1}^\dagger\right) & c_t^\dagger \end{bmatrix}
$$

情况： $c_t = 0$

当 $c_t = 0$、 $a_t = a_t A_{t-1}^\dagger A_{t-1}$ 时，也就是说， a_t 位于 A_{t-1} 的行空间中。令 $d_t = a_t A_{t-1}^\dagger$ ，并定义:

$$
T = A_{t-1} A_{t-1}^\dagger - A_{t-1} b_t d_t.
$$

由于 T 是对称的，因此 $A_{t-1} b_t d_t$ 也必须是对称的。因此，对于某个标量 h，有 $A_{t-1} b_t = h d_t^\top$。现在我们有

$$
A_t A_t^\dagger = \begin{bmatrix} A_{t-1} A_{t-1}^\dagger - h d_t^\top d_t & \frac{1}{\sqrt{\lambda}}(d_t - h d_t d_t^\top d_t) \\ \sqrt{\lambda} h d_t^\top & h d_t^\top d_t \end{bmatrix}
$$

根据 $A_t A_t^\dagger$ 的对称性，我们得到:

$$
\begin{aligned}
\sqrt{\lambda} h (d_t^\top)^\top &= \frac{1}{\sqrt{\lambda}}(d_t - h d_t d_t^\top d_t) \\
d_t &= h d_t d_t^\top d_t + \lambda h d_t \\
d_t &= \lambda h (\frac{1}{\lambda} d_t d_t^\top d_t + d_t) \\
h &= \frac{\frac{1}{\lambda}}{1 + \frac{1}{\lambda} d_t d_t^\top}
\end{aligned}
$$

因此,

$$
\begin{aligned}
b_t &= \frac{\frac{1}{\lambda} A_{t-1}^\dagger d_t^\top}{1 + \frac{1}{\lambda} d_t d_t^\top} \\
A_t^\dagger &= \begin{bmatrix} \frac{1}{\sqrt{\lambda}} \left(A_{t-1}^\dagger - \frac{\frac{1}{\lambda} A_{t-1}^\dagger d_t^\top d_t}{1 + \frac{1}{\lambda} d_t d_t^\top} \right) & \frac{\frac{1}{\lambda} A_{t-1}^\dagger d_t^\top}{1 + \frac{1}{\lambda} d_t d_t^\top} \end{bmatrix}
\end{aligned}
$$

这证明，在更新之前先用遗忘因子对 $R_{t-1}^\dagger$ 进行缩放，能够得到 $R_t^\dagger$ 的正确递归形式。

引理 1 *Let R_t denote the upper-trapezoidal factor obtained at time t after the update step but before the downdate*
(加权降阶一致性)。 ,

corresponding to the exponentially weighted observations $\{z_{t-i}\}_{i=0}^N$. Then the application of the downdate rotation G_{t-N} followed by the pseudoinverse downdate step computes the Moore–Penrose pseudoinverse of the matrix obtained by N 了缩放因子为 $\sqrt{\lambda}$ 的缩放操作。因此，它对因子 R_t 的有效贡献为

Proof. 在时间 t，观测值 z_{t-N} 已连续经历 N 了缩放因子为 $\sqrt{\lambda}$ 的缩放操作。因此，它对因子 R_t 的有效贡献为

$$
v = \lambda^{N/2} z_{t-N}
$$

通过将对应行与标准基向量 e_1对齐，构造降阶旋转 G_{t-N} 以分离这一贡献。在隐式-Q 表述中，可得

$$
G_{t-N}^\top R_t = \begin{bmatrix} v^\top \\ \widetilde{R}_t \end{bmatrix} = \begin{bmatrix} \sqrt{\lambda^N}\, z_{t-N}^\top \\ \widetilde{R}_t \end{bmatrix}
$$

对于秩一移除，应用广义 Moore–Penrose 伪逆降阶公式，在代数上等价于从对应的格拉姆矩阵中减去外积 $vv^\top$：

$$
\begin{aligned}
R_t^\top R_t - vv^\top &= \sum_{i=0}^{N} \lambda^i z_{t-i} z_{t-i}^\top - \lambda^N z_{t-N} z_{t-N}^\top \\
&= \sum_{i=0}^{N-1} \lambda^i z_{t-i} z_{t-i}^\top
\end{aligned}
$$

该表达式恰好是与指数加权滑动窗口 $\{z_{t-i}\}_{i=0}^{N-1}$ 相关的格拉姆矩阵。因此，移除旋转因子中单独的一行，在代数上等价于从底层最小二乘目标函数中移除最早的加权观测。

□

致谢

作者谨向 Konstaninos Ntetsikas 和 Andrew Ching Hoe Lee 致谢，感谢他们认真完成小组项目；该项目为本文后续的改进奠定了基础。Nick Firoozye 也感谢 Fauziah Ariff 在这项工作漫长的酝酿过程中，以及在准备本文稿期间给予的真知灼见和支持。

参考文献

[1] Takuya Akiba、Shotaro Sano、Toshihiko Yanase、Takeru Ohta与Masanori Koyama。Optuna：下一代超参数优化框架。载于 *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 2019，。[2] S. Thomas Alexander与Avinash L. Ghimikar。一种基于逆 QR 分解的递归最小二乘滤波方法。*IEEE Transactions on Signal Processing*，41(1):20，1993年。[3] 编Apolin´ario Jr，Jo˜ao A. *QRD–RLS Adaptive Filtering*。施普林格，2009年。[4] Peter L Bartlett、Philip M Long、G´abor Lugosi与Alexander Tsigler。线性回归中的良性过拟合。*Proceedings of the National Academy of Sciences*，117(48):30063–30070，2020年。[5] Mikhail Belkin、Daniel Hsu、Siyuan Ma与Soumik Mandal。调和现代机器学习实践与经典偏差-方差权衡。*Proceedings of the National Academy of Sciences* 116 32 15849 15854, ，（）：–2019。[6] Stephen L. Campbell 与 Carl D. Meyer。 *Generalized Inverses of Linear Transformations*。工业与应用数学学会，2009年。[7] Randall E. Cline。分块矩阵广义逆的表示。*Journal of the Society for Industrial and Applied Mathematics* 12 3 588 600, 1964，（）：–。[8] George Cybenko。用 S 形函数的叠加进行逼近。*Mathematics of Control Signals* ，，：–。*and Systems* 2 303 314, 1989 [9] Ender M. Eksioglu与A. Korhan Tanc。带凸正则化的RLS算法。*IEEE Signal Processing Letters* 18 8 470 473, 2011，（）：–。[10] Patrick Flandrin。 *Time-Frequency/Time-Scale Analysis*。学术出版社，圣迭戈，1998 年。[11] Gene H. Golub 与 Charles F. Van Loan。 *Matrix Computations*。约翰·霍普金斯大学出版社，巴尔的摩，MD，第3版，1996年。[12] Damodar N. Gujarati与Dawn C. Porter。 *Basic Econometrics*。麦格劳-希尔，新加坡，2003年。[13] Sven Hammarling与Craig Lucas。更新QR分解与最小二乘问题。2008年。[14] Robert Hartwig。秩分解与Moore-Penrose逆。*Industrial Mathematics*，26，1976年。[15] Trevor Hastie、Andrea Montanari、Saharon Rosset与Ryan J. Tibshirani。高维无岭最小二乘插值中的意外结果。 *Annals of Statistics*，50(2):949–986，2022年。[16] Trevor Hastie、Robert Tibshirani与Jerome H. Friedman。 *The Elements of Statistical Learning: Data Mining, Inference and Prediction* 2 2009，。施普林格，第nd版，。[17] Simon Haykin。 *Adaptive Filter Theory*。普伦蒂斯·霍尔，第4版，2002年。

[18] Nicholas J. Higham。 *Accuracy and Stability of Numerical Algorithms*。工业与应用数学学会，第2版，2002年。[19] David Hogg与Soledad Villar。高频插值中的特征加权与双下降。 *arXiv preprint arXiv:2305.01624* 2023 ，。

[20] Xia Hong、Junbin Gao与Sheng Chen。零吸引递归最小二乘算法。 *IEEE Transactions on Vehicular Technology* 66 1 213 221, 2017 ，

（）：–。[21] Kurt Hornik。多层前馈网络的逼近能力。 *Neural Networks* 4 2 251 257, ，1991

（）：–。[22] Arthur Jacot、Franck Gabriel与Cl´ement Hongler。神经切线核：神经网络中的收敛与泛化。载于 *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)* 2018 ，。

[23] Holger Kantz与Thomas Schreiber。 *Nonlinear Time Series: Analysis, Methods and Applications*。剑桥大学出版社，2004年。[24] Song Mei与Andrea Montanari。随机特征回归的泛化误差：精确渐近分析与双下降曲线。 *Communications on Pure and Applied Mathematics*，75(4):667–766，2022年。[25] C.-T. Pan与R.J. Plemmons。基于逆分解的最小二乘修正：并行意义。*Journal of Computational and Applied Mathematics*，27(1):109–127，1989年。数值线性代数并行算法专刊。[26] Jos´e C. Pr´ıncipe、Weifeng Liu与Simon Haykin。 *Kernel Adaptive Filtering: A Comprehensive Introduction*。约翰·威立父子出版公司，2011年。[27] Ali Rahimi与Benjamin Recht。面向大规模核机器的随机特征。载于 *Advances in Neural Information Processing Systems (NeurIPS)* 2007 ，。

[28] Ali H. Sayed。 *Fundamentals of Adaptive Filtering*。约翰·威立父子出版公司，2003年。[29] Enrique Sentana。半正定矩阵对称分解的正秩一修正的计量经济学应用。 *Spanish Economic Review*，1(1):79–90，1999年。[30] Ruben Staub与Stephan N. Steinmann。面向秩亏矩阵的高效递归最小二乘求解器。*Applied Mathematics and Computation*，399:125996，2021年。[31] Qian Su与Dino Sejdinovic。良性过拟合与噪声特征。 *Journal of the American Statistical Association* 118 541 164 177, 2023 ，

（）：–。[32] Yitong Sun、Anna Gilbert与Ambuj Tewari。论随机ReLU特征的逼近性质。 *arXiv preprint arXiv:1810.04374* 2018 ，。

[33] Sa˘so Toma˘zi˘c。论采用单边指数窗的短时傅里叶变换。*Signal Processing*，55 2 141 148，1996年（）：–。[34] Sa˘so Toma˘zi˘c。论采用单边指数窗的短时傅里叶变换。 *Signal Processing*，55 2 141 148，1996年（）：–。[35]Artur Trindade。ElectricityLoadDiagrams20112014。UCI机器学习资料库，2015年。DOI: https://doi.org/10.24432/C58C86。[36] Alexander Tsigler与Peter L Bartlett。岭回归中的良性过拟合。 *The Journal of Machine Learning Research* 24 1 5202 5277, 2023 ，

（）：–。[37] S. Van Vaerenbergh、J. Via与I. Santamaria。一种滑动窗口核递归最小二乘算法及其在非线性信道辨识中的应用。载于 *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings* 5, 2006 ，卷，

页V–V，。[38] Isao Yamada、Nobuhiko Ogura与Masahiro Yukawa。自适应投影次梯度法及其加速技术。 *IFAC Proceedings Volumes*，37(12):639–644，2004年。IFAC控制与信号处理中的自适应与学习研讨会（ALCOSP 04）与IFAC周期控制系统研讨会（PSYCO 04），横滨，日本，2004年8月30日至9月1日。[39] Chiyuan Zhang、Samy Bengio、Moritz Hardt、Benjamin Recht与Oriol Vinyals。理解深度学习需要重新思考泛化。载于 *Proceedings of the International Conference on Learning Representations (ICLR)* 2017 ，。