arXiv:2305.02544v1 [cs.LG] 4 May 2023

抗异常值主成分分析的近线性时间算法与流式算法

Ilias Diakonikolas[†]威斯康星大学麦迪逊分校 ilias@cs.wisc.edu

Daniel M. Kane[‡]加州大学圣迭戈分校 dakane@cs.ucsd.edu

Ankit Pensia[§]威斯康星大学麦迪逊分校 ankitp@cs.wisc.edu

Thanasis Pittas[¶]威斯康星大学麦迪逊分校 pittas@wisc.edu

2023年5月5日

摘要

我们研究主成分分析（PCA）：给定来自分布的 $\mathbb{R}^d$ 数据集，任务是找到一个单位向量 v，使该分布沿 v投影后的方差近似最大化。尽管这是一项经典任务，但只要数据中含有少量异常值，标准估计量就会严重失效，这促使人们研究鲁棒主成分分析。近期研究提出了用于鲁棒主成分分析的计算高效算法，但这些算法要么需要超线性时间，要么误差保证不够理想。我们的主要贡献是提出一种用于鲁棒主成分分析的近线性时间算法，并具有近乎最优的误差保证。我们还提出了一种用于鲁棒主成分分析的单遍流式算法，其内存使用量关于维度近乎线性。

*作者按姓氏字母顺序排列。

[†]由美国国家科学基金会中型项目资助 CCF-2107079、美国国家科学基金会项目资助 CCF-1652862（CAREER）、斯隆研究奖学金以及 DARPA Learning with LessLabels（LwLL）资助提供支持。

[‡]由美国国家科学基金会中型项目资助 CCF-2107547、美国国家科学基金会项目资助 CCF-1553288（CAREER）以及斯隆研究奖学金提供支持。[§]由美国国家科学基金会项目资助 CCF-1652862（CAREER），以及美国国家科学基金会资助 CCF-1841190 和 CCF-2011255 提供支持。

[¶]由美国国家科学基金会中型项目资助 CCF-2107079 提供支持。

目录

1 引言 **1**

1.1 我们的结果.1 1.2 我们的技术.3 1.3 相关工作.4

2 预备知识 **5**

2.1 杂项事实.5 2.2 内点稳定性条件.6 2.3 过滤.8

3 近线性时间的 RobustPCA 9

3.1 高级证书引理.9 3.1.1 断言 3.4 的证明： Y 的方差界.12 3.2 势函数乘法递减.14 3.3 改进运行时间对 ϵ 的依赖.16 3.4 定理 3.1 的证明.17 3.4.1 势函数较小意味着满足停止条件.18 3.4.2 势函数乘法下降的形式化证明.19 3.4.3 综合所有内容.23 3.4.4 运行时间分析.26

4 流式算法 RobustPCA 2 的算法 **6**

4.1 建立条件 4.2.294.2 定理 4.1 的证明概要.324.2.1 位复杂度.32

5 结论 **33**

A 预备知识 ： 省略事实 38

A.1 信息论误差.38 A.2 两种停止条件的比较.38 A.3 关于稳定性的省略证明：引理 2.14 的证明.39 A.4 筛选.42

B 第 4 节中省略的证明 **44**

B.1 条目 1 和 2.45 B.2 条目 4.46 B.3 条目 3.47

1 引言

主成分分析（PCA）是降维和数据可视化中的核心子程序。PCA 通常用于识别数据中方差较大的方向，这些方向被认为是数据中信息量最大的部分。在经典设定中，我们观测到来自具有协方差 $\mathbf{\Sigma}$ 的（次高斯）分布、位于 $\mathbb{R}^d$ 中的 n 个独立同分布点的集合 S。然后，经典估计量（例如，S 的经验协方差的主特征向量）会输出一个方向 v，满足 $v^\top \mathbf{\Sigma} v$ 随着 n 增大而趋近于 $\|\mathbf{\Sigma}\|_{\mathrm{op}}$。重要的是，这些估计量可以快速计算——也就是说，它们的运行时间近乎线性。[1]

然而，获取独立同分布样本通常是不现实的假设，数据中可能包含少量异常值，具体形式化如下：

定义 1.1（强污染模型）。给定参数 $0 < \epsilon < 1/2$ 和分布类 $\mathcal{D}$，强对手按如下方式操作：算法指定样本数量 n，随后对手从某个 $D \in \mathcal{D}$ 中抽取一组 n 个独立同分布样本，在检查这些样本后，删除其中至多 ϵn 个样本，并将其替换为任意点。最终得到的集合作为学习算法的输入。我们称一个集合 ϵ-污染集，如果它由上述过程生成。

遗憾的是，异常值会使经典估计量的结果发生偏差，因为它们会人为增大低方差方向上的方差，从而使这些估计量的输出变得不可靠。[2] 为了解决这一问题，鲁棒算法必须能够从数据中有效移除异常值，同时在高维情况下保持计算效率。

近期的研究提出了用于鲁棒主成分分析的计算高效算法 [JLT20，KSKO20]。这些研究属于算法鲁棒统计的范畴，其目标是开发计算高效的算法，用于高维统计估计，并能抵御一定比例的异常值，参见例如 [DKK+16、LRV16、DK19、DK23]。然而，这些多项式时间算法 [JLT20、KSKO20] 要么（i）实现近乎最优的误差，但运行时间超线性（随 nd^2变化）；要么（ii）以近线性时间运行，但误差并非最优，且需要额外的假设。随着现代数据集在样本量和维度上不断扩大，对具有近乎最优误差保证的近线性时间算法和流式算法的需求日益增长。我们的工作通过开发满足这些需求的鲁棒主成分分析算法，回应了这一需求。

1.1 我们的结果

本文的主要结果是一种近线性时间算法，可从遭到破坏的高维数据集中识别一个方差较大的方向。下面我们针对次高斯分布陈述结果（定义 2.1），并指出，对于更一般的“稳定”分布类（定义 2.12），同样的结果也成立。

定理 *1.2*（近线性时间鲁棒主成分分析）。设 $\epsilon \in (0, c)$，其中 $c > 0$ 是一个较小的常数。设 D 是一个零均值、（未知）协方差矩阵为 $\mathbf{\Sigma}$ 的次高斯分布。对于足够大的 $C > 0$，设 S 是在强污染模型（定义 *1.1*）下，从 D 中抽取的 $n \geq Cd/\epsilon^2$ 个样本构成的一个经过 ϵ 污染的集合。存在一种算法，以 S 和 ϵ 作为输入，在 $\widetilde{O}(nd/\epsilon^2)$ 时间内运行，并以至少 *0.99* 的概率输出一个单位向量 u，使得 $u^\top \mathbf{\Sigma} u \geq (1 - O(\epsilon \log(1/\epsilon)))\|\mathbf{\Sigma}\|_{\mathrm{op}}$。

[1]所谓“近线性时间”算法，是指运行时间随输入规模 nd 线性增长（至多相差对数因子）。[2]朴素剪枝和随机投影等朴素技术也容易受到异常值的影响。

我们注意到，我们的算法的误差保证近乎最优，仅差对数因子（参见附录 A.1）。与我们工作最为密切相关的工作是 [JLT20，定理 2]，而我们的算法在三个重要方面对其进行了改进：

- （近乎最优的误差）定理 1.2 的输出达到了对 $(1 - O(\epsilon\log(1/\epsilon)))$的近似$\|\mathbf{\Sigma}\|_{\mathrm{op}}$，而 [JLT20] 的结果则达到了次优的近似 $(1 - O(\sqrt{\epsilon}\log(1/\epsilon)\log d))$。[3]

- （对 ϵ 运行时间的依赖）我们的算法运行时间为 $\widetilde{O}(nd/\epsilon^2)$，相比之下，[JLT20]，的运行时间为 $\widetilde{O}\left(nd/\epsilon^{4.5}+ndm/\epsilon^{1.5}\right)$，其中 m 是下文讨论的一个参数，属于 $[2, d]$。因此，我们的算法在所有情况下的运行时间都优于它们。

- （特征值间隔）[JLT20] 算法要求作出如下假设：m第 $\widetilde{}$ 个最大 $\mathbf{\Sigma}$ 特征值小于 $\|\mathbf{\Sigma}\|_{\mathrm{op}}$ 的 $(1 - \Theta(\epsilon))$，其中 $m \in [2, d]$，该项出现在运行时间中。因此，[JLT20] 只有在 $m=\mathrm{polylog}(d/\epsilon)$ 时才能实现近线性时间。定理 1.2 没有这一限制。

最后需要指出的是，即使不存在污染，$\widetilde{}\Omega(d/\epsilon^2)$ 个样本对于 $(1-O(\epsilon\log(1/\epsilon))$-近似 $\|\mathbf{\Sigma}\|_{\mathrm{op}}$也是必要的。此外，即使不存在污染，实现这一保证的标准 Lanczos 算法的运行时间为 $\widetilde{O}\left(\frac{nd}{\sqrt{\epsilon}}+\frac{1}{\epsilon}\right)$ [SV14，定理 10。1] 因此，我们的结果显著降低了主成分分析中鲁棒性的代价，使其更加接近干净数据设置。

流式算法

现代数据科学应用的分布式特性以及大规模数据集，往往要求完整数据集不能存储在内存中。在这种情况下，下面定义的流式模型是一个更符合实际的设定：

定义 1.3 (单遍流式模型)。设 S 是一个固定集合。在单遍流式模型中，S 的元素会逐个呈现给算法，并且算法只能对这些点进行一遍扫描。

在流式模型中，算法还需要优化其使用的内存量。我们仍然考虑数据中的污染。这意味着上文的输入 S 并非来自某个分布的独立同分布样本，而是来自一个“污染”分布，具体形式化定义如下：

定义 1.4 (全变差污染)。给定参数 $\epsilon \in (0, 1/2)$ 和一个分布类 $\mathcal{D}$，对手指定一个分布 D'，使得存在 $D \in \mathcal{D}$ 满足 $d_{\mathrm{TV}}(D, D') \le \epsilon$。然后，算法从 D' 中抽取独立同分布样本。我们称分布 D' 是分布 D 的一个在全变差距离下的 ϵ- 污染版本。

此前所有鲁棒主成分分析算法都需要将整个数据集存储在内存中，因此空间使用量至少为 $\widetilde{}\Omega(d^2/\epsilon^2)$。[4] 基于 定理 1.2，我们提出了首个用于鲁棒 $\widetilde{}$主成分分析的算法，其仅使用 $O_\epsilon(d)$ 内存。

定理 *1.5*（鲁棒主成分分析的流式算法；非正式版本）。设 $\epsilon \in (0, c)$，其中 $c > 0$ 是一个较小的常数。设 D 是一个零均值、协方差矩阵 $\mathbf{\Sigma}$ 未知的次高斯分布。设 P 是 D 的一个经过 ϵ 污染的版本，且二者的全变差距离为 ϵ（定义 *1.4*）。存在一个单遍流式算法，给定 d/ϵ，该算法从 P 中读取 *poly(* u *)* 个样本，并以 *0.99* 的概率返回一个单位向量 $u^\top \mathbf{\Sigma} u \ge$，使得 $1 - O(\epsilon log(1/\epsilon))\|\mathbf{\Sigma}\|_{\mathrm{op}}$。此外，该算法使用 $(d/\epsilon) \cdot polylog(d/\epsilon)$ 的内存，并在 $(nd/\epsilon^2)polylog(d/\epsilon)$ 时间内运行。

[3]特别是，后一个结果要求 $\epsilon \to 0$ 以便 $d \to \infty$ 获得非平凡的保证。[4]在实验中，内存使用量一直被强调为稳健估计中的关键瓶颈；参见 [DKK+17]。

注意，该算法的渐近误差近乎最优，且其内存使用量相对于 d （输出的大小）近乎线性。关于位复杂度的进一步讨论，请参阅第 4.2.1 节。

1.2 我们的方法

我们的方法是将 [JLT20、KSKO20]提出的鲁棒主成分分析算法的“简单”版本，与 [DKPP22]提出的快速均值估计算法结合起来（另见 [DKK+22b]）。这个“简单”算法的核心是计算经验协方差矩阵的最大特征值对应的特征向量 v，然后将样本投影到 v方向，并估计 v方向上的真实协方差。如果该估计值接近经验协方差，那么结合误差不可能在任何方向上大幅降低经验协方差这一事实，就可以推出 v 接近真实协方差矩阵的一个主特征向量（参见引理 2.15）。否则，差异必定是由少量污染造成的，而这些污染可以通过算法过滤掉。随后重复过滤异常值的过程，直到找到一个近似主特征向量。

这个简单算法的问题在于运行时间。尽管每个过滤步骤都可以在近线性时间内完成，但无法保证过滤步骤的数量很少。完全可能出现这样的情况：每个过滤步骤只移除一小部分污染，而这些污染在主特征向量方向上的投影非常大 v。为了解决这个问题，我们借鉴 [DKPP22]的思路，将 v 定义为 $\mathbf{\Sigma}_t$ （第 t步的经验协方差矩阵）的若干个最大特征值对应特征向量的随机线性组合，即令其为 $\mathbf{\Sigma}_t^p w$ （而非经验协方差矩阵的主特征向量），其中 w 是随机高斯向量， p 是一个足够大的整数。这样可以防止对手通过使某个污染与特定的 v正交来“隐藏”它。相反，任何对$\mathbf{\Sigma}_t$ 的某个最大特征值有显著贡献的污染，都有较大概率被过滤掉。我们的方法是跟踪势函数$\mathrm{tr}(\mathbf{\Sigma}_t^{2p+1})$，并证明：（i）势函数值较小，足以保证可以从经验协方差矩阵中恢复出一个解向量；（ii）在无法得到这一保证时，我们可以将势函数按乘法因子（$1-\Omega(\epsilon)$）降低。这一方法面临两个主要挑战。

第一个关键挑战与前面提到的“认证”有关：在不对 $\mathbf{\Sigma}$ 的谱作任何限制的情况下获得最优误差（回想一下，[JLT20] 的近线性时间结果对 $\mathbf{\Sigma}$ 的谱施加了限制）。一种自然的停止条件是：当前的 $\mathbf{\Sigma}_t$ 与 $\mathbf{\Sigma}$ 在适当意义下相差不超过 $(1\pm\epsilon)$ 倍。 [JLT20] 使用 Schatten-p 范数来比较 $\mathbf{\Sigma}$ 和 $\mathbf{\Sigma}_t$，其中 p 足够大。然而，一个简单的例子（例 3.2）表明，即使如此，任何依赖于 $\mathbf{\Sigma}_t$ 的确定性算法都必须花费 nd^2 时间。我们的双管齐下方案是：(i) 对鲁棒主成分分析进行白盒分析，以施加更强的停止条件；(ii) 输出 $\mathbf{\Sigma}_t$ 的随机主特征向量（参见第 3.1 节）。

其次，与 [DKPP22] 中同，我们无法完全实现 $\|\mathbf{\Sigma}\|_{\mathrm{op}}=1$，朴素过滤可以保证我们在 $t=0$ 处的初始 $\mathbf{\Sigma}_t$ 的特征值至多为 poly(d)。[5] 因此，我们的势函数初值为 $d^{O(p)}$，而 需要进行 $O(p/\epsilon)$ 轮过滤（每轮耗时 $O(pnd)$，因为我们需要对 $\mathbf{\Sigma}$ 进行 p 次矩阵 - 向量乘积），总运行时间为 $\widetilde{O}(ndp^2/\epsilon)$。现在，我们的算法要求 p 至少为 $\log(d)/\epsilon$，以区分相差 $(1+\epsilon)$ 倍的特征值；相比之下，[DKPP22], 的结果只需 $p=O(\log(d))$，因此运行时间为 $\widetilde{O}(nd/\epsilon^3)$。

这种质上是由于 的选择导致的。假设 $\|\mathbf{\Sigma}\|_{\mathrm{op}}=1$，我们 p 可以 $\widetilde{O}(nd)$ 运行时间内实现 $O(1/\epsilon^2)$ 次迭代。注意到特征值之间如此细微的差异，只有在我们算法的最后一个阶段（“认证阶段”）才会变得重要。在此之前，我们可以采用较小的 p。更正式地说，对于任意 $p'\geq\log d$，我们仍然可以降低势函数

[5]由于所有量都随 $\|\mathbf{\Sigma}\|_{\mathrm{op}}$ 缩放，为了简化表述，我们采用这种归一化。

，直到满足停止条件为止。尽管这已无法再保证我们得到的是一个好的解，但我们可以据此将势函数上界限定为 poly(d)。因此，我们分多个阶段运行该方法：在阶段 k，我们取 $p_k = 2^k \log d$，并降低势函数，直到其变为d^C，这需要 $\tilde{O}(ndp_k/\epsilon)$ 时间。到达 $p = \log(d)/\epsilon$ 时，由于该阶段开始时的势函数也至多为 d^C，算法最终会在 $\tilde{O}(nd/\epsilon^2)$ 时间内终止。

最后，正如 [DKPP22], 上述算法可以转换为单遍流式算法。由于我们使用少量过滤器，其中每个过滤器都会移除所有 x 具有$|v^\top x| > T$（对于某些经过仔细选择的 v 和 T）的样本，因此我们可以通过仅存储先前创建的过滤器的 v'和 T'以及必要计算所需的最少临时变量来实现低内存占用。

1.3 相关工作

我们的工作处于算法鲁棒统计领域，该领域旨在为高维估计问题开发能够抵御异常值的计算高效算法。自 [DKK+16、LRV16], 发表以来，针对许多任务的计算高效鲁棒算法相继问世，包括均值估计 [KSS18, CDG19, DHL19, DL22, DKP20, HLZ20, HR21, CTBJ22, ZJS21], 稀疏估计 [BDLS17, DKK+19b, CDK+22, DKK+22a, DKLP22], 协方差估计 [CDGW19], 线性回归 [KKM18, DKS19, PJL20, CAT+20]、聚类与列表可解码学习 [DKS18, HL18, KSS18, KKK19, LM21, BDJ+22], 以及随机优化 [DKK+19a, PSBR20]。关于这一领域的概述，读者可参阅近期出版的专著 [DK23] 。

针对高维鲁棒估计任务的近线性时间算法研究，始于 [CDG19] 在均值估计领域的工作。自该工作以来，研究者在多种背景下取得了许多改进和推广。与本文最相关的一系列工作是关于高维鲁棒主成分分析的研究，始于 [XCM13]。虽然这项早期工作未能在多项式时间内获得与维度无关的误差保证，但随后 ϵ 提出了误差保证对 [JLT20、KSKO20]近乎最优依赖关系的多项式时间算法。需要指出的是，这些算法的运行时间至少为$\Omega(nd^2)$。此外，[JLT20] 提出了一种用于鲁棒主成分分析的近线性时间算法；然而，该算法仅在某些情况下具有近线性运行时间，而且误差保证对 ϵ 的依赖并非最优，且依赖于维度。最后，我们的流式算法受到 [DKPP22], 的启发；他们提出了首批用于高维鲁棒均值估计及相关任务的流式算法。

我们指出，鲁棒主成分分析与$\widehat{}$算子范数下的鲁棒协方差估计问题密切相关，在该问题中，算法需要在给定一个受污染的 $\mathbf{\Sigma}$ 样本集合后输出一个估计量$\widehat{\quad}$，使得 $\|\mathbf{\Sigma} - \mathbf{\Sigma}\|_{\mathrm{op}}$ 足够小。容易看出， $\mathbf{\Sigma}$ 的最大特征值对应的特征向量满足鲁棒主成分分析的保证。在本段中，我们只考虑高斯数据，且 ϵ是一个足够小的绝对常数。尽管算子范数下鲁棒协方差估计的样本复杂度$\tilde{}$仍为 $O(d)$, [DKS17，定理 15] 表明，除非使用$\Omega(d^2)$ 个样本，否则该问题在计算上是困难的（在统计查询模型中）。使用$\Omega(d^2)$ 个样本，可以在相对弗罗贝尼乌斯范数下估计协方差（因此也能在算子范数下估计）；参见 [DKS17] 中的讨论。因此，与算子范数下的鲁棒协方差估计相比，鲁棒主成分分析在计算和统计上都更容易，同时仍然具有实用价值。

最后，我们强调，本研究（以及 [JLT20、KSKO20]）的关注点与其他使用类似术语的研究 [CLMW11、MMA18、BJSY22], 在一定程度上是正交的，因为鲁棒性的定义有显著差异。

2 预备知识

符号约定 我们记 $[n] := \{1, \ldots, n\}$。对于向量 v，我们分别用 $\|v\|_2$ 和 $\|v\|_\infty$ 表示其 ℓ_2 范数和无穷范数。矩阵用粗体大写字母表示。用 I 表示单位矩阵。对于两个矩阵 A 和 B，用 tr(A) 表示 A 的迹，并用 $\langle A, B\rangle := \mathrm{tr}(A^\top B)$ 表示迹内积。对于矩阵 A （其中 $p \geq 1$），我们用 $\|A\|_F$、$\|A\|_{op}$、$\|A\|_p$ 分别表示其 Frobenius 范数、算子范数和 Schatten p-范数。特别地，若 A 是特征值为 $\lambda_1, \ldots, \lambda_d$ 的 $d \times d$ 对称矩阵，则 $\|A\|_p := (\sum_i |\lambda_i|^p)^{1/p}$。若方阵 A 是半正定的（positive semidefinite），则称 A 为 PSD，并记作 $A \succeq 0$；此时对所有 $x \in \mathbb{R}^d$，均有 $x^\top A x \geq 0$。若 $B - A$ 为半正定矩阵，则记作 $A \preceq B$。若存在绝对普适常数 $C > 0$ 使得 $a \leq Cb$，则记作 $a \lesssim b$。

对于定义域 $\mathcal{X}$ 上的分布 D 和权重函数 $w : \mathcal{X} \to \mathbb{R}_+$，我们用 D_w 表示定义在 $\mathbb{R}^d$ 上、概率密度函数为 $D_w(x) := D(x)w(x)/\int D(x)w(x)$ 的分布。我们用 $\mathbf{\Sigma}_D$ 表示 D 的二阶矩：$= \mathbf{E}_{X\sim D}[XX^\top]$。我们将反复用到以下事实：$\mathbf{E}_{X\sim D}[\|\mathbf{U}X\|_2^2] = \langle \mathbf{U}^\top \mathbf{U}\ \mathbf{\Sigma}_D\rangle$。

我们采用以下次高斯性定义，该定义也曾用于 [JLT20]。

定义 2.1。对于参数 $r \geq 1$，若分布 D 定义在 $\mathbb{R}^d$ 上，均值为零且协方差为$\mathbf{\Sigma}$，则称其为 r-次高斯分布；对于所有单位向量 v 和 $t > 0$,

$$\mathop{\mathbf{E}}_{X\sim D}[\exp(tv^\top X)] \leq \exp\left(t^2 r(v^\top \mathbf{\Sigma} v)/2\right)$$

注意，这种次高斯性定义比 [Ver18，定义 3.4。1] 更强，因为方向 v 上的次高斯代理参数会随 $v^\top \mathbf{\Sigma} v$ 缩放。

2.1 杂项事实

现在我们汇集一些将在本文中用到的事实。

线性代数事实 首先，我们指出半正定矩阵的迹内积具有以下有用性质：

事实 2.2. 若$\mathbf{A}$、$\mathbf{B}$、$\mathbf{C}$ 是对称的 $d \times d$ 矩阵，且$\mathbf{A} \succeq 0$，并且$\mathbf{B} \preceq \mathbf{C}$，那么$\mathrm{tr}(\mathbf{AB}) \leq \mathrm{tr}(\mathbf{AC})$。

以下事实的证明可见，例如 [JLT20]。

事实 2.3. 设$\mathbf{A}$和$\mathbf{B}$ 为半正定矩阵，且$\mathbf{B} \preceq \mathbf{A}$ 以及 $p \in \mathbb{N}$。则，$\mathrm{tr}\left(\mathbf{B}^p\right) \leq \mathrm{tr}\left(\mathbf{A}^{p-1}\mathbf{B}\right)$.

下面的事实表明，半正定矩阵在同时左乘和右乘后仍然是半正定的：

事实 2.4. 如果 $\mathbf{A} \in \mathbb{R}^{d\times d}$ *is a PSD matrix, f*或任意 *o ther* $\mathbf{B} \in \mathbb{R}^{d\times d}$ 它 *h olds* $\mathbf{BAB}^\top \succeq 0$。证明。对于任意向量 x，记 $y := \mathbf{B}^\top x$，则有 $x^\top(\mathbf{BAB}^\top)x = y^\top \mathbf{A} y \geq 0$。 □

我们将用以下事实来比较$\mathbf{A}^m$ 对于某个 $m \geq 1$ 以及$\mathbf{A}$ 沿某些方向的情形：

事实 2.5。设$\mathbf{A}$ 是一个半正定矩阵。那么对于任意单位向量 x 和 $m \in \mathbb{N}$，$x^\top \mathbf{A}^m x \geq (x^\top \mathbf{A} x)^m$。证明。由谱分解，$\mathbf{A}$ 等于 $\sum_i \lambda_i u_i u_i^\top$，其中 $\lambda_i \geq 0$，且u_i为正交归一向量。此外，$\mathbf{A}^m = \sum_i \lambda_i^m u_i u_i^\top$。对于任意单位向量 x，u_i的正交归一性意味着$\sum_i (u_i^\top x)^2 = 1$。

现在定义非负随机变量 Z，它以 $(u_i^\top x)^2$, 的概率取值为 λ_i；这是一个有效的概率分配，因为它非负且总和为 1。在此定义中，我们有 $\mathbf{E}[Z] = \sum_i \lambda_i (u_i^\top x)^2 = x^\top \mathbf{A} x$。此外，$\mathbf{E}[Z^m] = \sum_i \lambda_i^m (u_i^\top x)^2 = x^\top (\sum_i \lambda_i^m u_i u_i^\top) x = x^\top \mathbf{A}^m x$。因此，所需结果等价于指出，对于非负随机变量 Z，有 $\mathbf{E}[Z^m] \geq (\mathbf{E}[Z])^m$；这由詹森不等式可得。 □

现在我们将注意力转向 Schatten 范数：对于对称矩阵$\mathbf{A}$，有$\|\mathbf{A}\|_2 = \|\mathbf{A}\|_{\mathrm{F}}$，$\|\mathbf{A}\|_\infty = \|\mathbf{A}\|_{\mathrm{op}}$；对于半正定矩阵 $\mathbf{A}$，$\|\mathbf{A}\|_1 = \mathrm{tr}(\mathbf{A})$。Schatten 范数对于迹内积满足 Hölder 不等式：

事实 2.6 (Schatten 范数的 Hölder 不等式)。设$\mathbf{A}$和$\mathbf{B}$是两个维度为 $m \times n$ 的矩阵。令 $p \geq 1$。定义 $q = \frac{p}{p-1}$：若 $p > 1$，则 $q = \infty$；否则 $q = \infty$。于是$\langle \mathbf{A}、\mathbf{B} \rangle \leq \|\mathbf{A}\|_p \|\mathbf{B}\|_q$。特别地，$\langle \mathbf{A}、\mathbf{B} \rangle \leq \|\mathbf{A}\|_{\mathrm{F}} \|\mathbf{B}\|_{\mathrm{F}}$，且 $\langle \mathbf{A}、\mathbf{B} \rangle \leq \|\mathbf{A}\|_{\mathrm{op}} \|\mathbf{B}\|_1$。

此外，Schatten-p 范数随 p 增大而减小，并且当 p 较大时彼此接近。

事实 2.7。若 $\mathbf{A} \in \mathbb{R}^{d \times d}$ 是对称的且 $p \geq 1$，则 $\mathbf{A}$ 的 *Schatten* 范数满足以下关系：$\|\mathbf{A}\|_{p+1} \leq \|\mathbf{A}\|_p \leq \|\mathbf{A}\|_{p+1} d^{p(p+1)}$.

特别地，我们将经常使用以下特殊情形：

事实 2.8. 如果 $\mathbf{A}$ 是一个半正定矩阵，$\|\mathbf{A}\|_{\mathrm{F}}^2 \leq \mathrm{tr}(\mathbf{A})^2$.

证明。注意 $\|\mathbf{A}\|_{\mathrm{F}} = \|\mathbf{A}\|_2$ 和 $\mathrm{tr}(\mathbf{A}) = \|\mathbf{A}\|_1$ 因为 $\mathbf{A}$ 是半正定的。然后由事实 2.7可得结果。 □

概率事实高斯分布的二次多项式将在我们的分析中发挥重要作用，即 $z^\top \mathbf{A} z$ 对于 $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$，我们将反复使用以下事实。

事实 2.9。对于任意对称的 $d \times d$ 矩阵 $\mathbf{A}$，我们有 $\mathrm{Var}_{z \sim \mathcal{N}(0,\mathbf{I})}[z^\top \mathbf{A} z] = 2\|\mathbf{A}\|_{\mathrm{F}}^2$。如果 $\mathbf{A}$ 是半正定矩阵，那么对于任意 $\beta > 0$，有 $\mathbf{Pr}_{z \sim \mathcal{N}(0,\mathbf{I})}[z^\top \mathbf{A} z \geq \beta \mathrm{tr}(\mathbf{A})] \geq 1 - \sqrt{e\beta}$。

对于论文的流式处理部分，我们将使用以下关于经验二阶矩矩阵集中性的结果：

事实 *2.10*（参见例如 [*Ver*10]）。考虑定义在 $\mathbb{R}^d$ 上的一个分布 D，其支撑集包含在一个以原点为中心、半径为 R 的 ℓ_2-球内。令 $\mathbf{\Sigma}$ 为其二阶矩矩阵，$\mathbf{\Sigma}_N = (1/N) \sum_{i=1}^N X_i X_i^\top$ 为使用 N 个独立同分布样本 $X_i \sim D$ 得到的经验二阶矩矩阵。存在一个常数 C，使得对于任意 $0 < \epsilon < 1$ 和 $0 < \tau < 1$，若 $N > C\epsilon^{-2} \|\mathbf{\Sigma}\|_2^{-1} R^2 \log(d/\tau)$，则我们有 $\|\mathbf{\Sigma} - \mathbf{\Sigma}_N\|_{\mathrm{op}} \leq \epsilon \|\mathbf{\Sigma}\|_{\mathrm{op}}$，且其成立的概率至少为 $1 - \tau$。

最后，我们记录幂迭代的保证：

事实 2.11（幂迭代）。对于任意半正定矩阵 $\mathbf{A} \in \mathbb{R}^{d \times d}$ 以及 $\epsilon, \delta \in (0,1)$，若 $p > \frac{C}{\epsilon} \log(d/(\epsilon\delta))$ 对于足够大的常数 C成立，且 $u := \mathbf{A}^p z$ 对于 $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 成立，则

$$\mathbf{Pr}\left[u^\top \mathbf{A} u / \|u\|_2^2 \geq (1-\epsilon)\|\mathbf{A}\|_{\mathrm{op}}\right] \geq 1 - \delta \,.$$

2.2 内点稳定性条件

回顾一下，在我们的符号约定中，我们用 $\mathbf{\Sigma}_D = \mathbf{E}_{X \sim D}[XX^\top]$ 表示分布 D的二阶矩，而 $D_w(x) := D(x)w(x)/\int D(x)w(x)$表示由 $w(x)$重新加权后的分布 D。我们的算法将依赖于以下条件。

定义 *2.12*（稳定性条件）。设 $0 < \epsilon < 1/2$ 和 $\epsilon \leq \gamma < 1$。如果对于每个权重函数 $w: \mathbb{R}^d \to [0,1]$，满足 $E_{X \sim G}[w(X)] \geq 1 - \epsilon$，加权二阶矩矩阵 $\mathbf{\Sigma}_{G_w} := E_{X \sim G}[w(X) XX^\top]/E_{X \sim G}[w(X)]$，均满足 $(1-\gamma)\mathbf{\Sigma} \preceq \mathbf{\Sigma}_{G_w} \preceq (1+\gamma)\mathbf{\Sigma}$，则称分布 G 在 $\mathbb{R}^d$ 上相对于半正定矩阵 $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$ 是 (ϵ, γ)-稳定的。

特别地，二阶矩矩阵在删除 ϵ 比例的样本后仍保持稳定。以下几点需要说明：(i) 上述定义适用于均值为零的分布；不失一般性，可以作此假设，因为我们总能使用成对差值进行处理；(ii) 从协方差为 $\mathbf{\Sigma}$ 的 $O(1)$-次高斯分布（参见定义 2.1）中抽取 $n = Cd/(\epsilon^2 \log(1/\epsilon))$ 个独立同分布样本后，在这些样本上定义的均匀分布以高概率是 (ϵ, γ)-稳定的，且 $\gamma = O(\epsilon \log(1/\epsilon))$ [JLT20，引理 11]，；(iii) 如定理 1.2 所述，该算法以来自次高斯分布的、经 ϵ 污染的样本集作为输入。然而，为方便符号表示，我们将输入视为分布 $P = (1-\epsilon)G + \epsilon B$，其中 G 是（未知的）稳定分布，而 B 是任意分布。P 表示在经 ϵ 污染的样本集上的均匀分布，G 表示其余内点部分，而 B 表示异常值部分。[6] 我们强调，算法并不知道哪些样本是异常值。

我们的算法将迭代地移除点，因此我们将使用二元权重 $w(x) \in \{0, 1\}$ 来区分保留的点（$w(x) = 1$）和已移除的点（$w(x) = 0$）。在算法运行过程中，我们将确保权重满足以下设定：

设定 2.13。令 $0 < 20\epsilon \le \gamma < \gamma_0$ 为一个较小的常数 γ_0。令 P 为定义在 $\mathbb{R}^d$ 上的混合分布 $P = (1-\epsilon)G + \epsilon B$，其中 G 是关于半正定 $d \times d$ 矩阵 $\mathbf{\Sigma}$ 的 $(20\epsilon, \gamma)$-稳定分布，而 $\mathbf{B}$ 是任意的。令 $w : \mathbb{R}^d \to \{0, 1\}$ 为一个权重函数，满足 $\mathbf{E}_{X \sim G}[w(X)] \ge 1 - 3\epsilon$。

由于我们将使用点的投影来滤除异常值，因此需要以下稳定性推论，涉及这些投影（以及经过阈值处理的投影）的平均值、分位数及其稳健估计量。

引理 *2.14*。在设定 *2.13* 中，定义如下：*(i)* $g(x) := \|Ux\|_2^2$，对于某个 $U \in \mathbb{R}^{m \times d}$，*(ii)* L 为在 P_w（P 经 w 重新加权）下，$g(x)$ 的前 3ϵ-分位数，并且 $S := \{x : g(x) > L\}$，*(iii)* $\widehat{\sigma} := E_{X \sim P}[w(X)g(X)\mathbb{1}(g(X) \le L)]$。那么，

(i) $\left| \mathop{\mathbf{E}}_{X \sim G}[w(X)g(X)] - \left\langle \mathbf{U}^\top \mathbf{U}, \mathbf{\Sigma} \right\rangle \right| \le 2\gamma \left\langle \mathbf{U}^\top \mathbf{U}, \mathbf{\Sigma} \right\rangle,$

(ii) $\mathbf{E}_{X \sim G}[w(X)g(X)\mathbb{1}\{X \in S\}] \le 2.35\gamma \left\langle \mathbf{U}^\top \mathbf{U}, \mathbf{\Sigma} \right\rangle,$

(iii) $L \le 1.65(\gamma/\epsilon)\langle \mathbf{U}^\top \mathbf{U}, \mathbf{\Sigma} \rangle,$

(iv) $\left| \widehat{\sigma} - \left\langle \mathbf{U}^\top \mathbf{U}, \mathbf{\Sigma} \right\rangle \right| \le 4\gamma \left\langle \mathbf{U}^\top \mathbf{U}, \mathbf{\Sigma} \right\rangle,$

(v) $\mathbf{Pr}_{X \sim G}\left[\|X\|_2^2 > 2(d/\epsilon)\|\mathbf{\Sigma}\|_{\mathrm{op}} \right] \le \epsilon.$

如前所述，我们的目标是不断滤除点，直到数据集的经验二阶矩 u 所对应的最大特征值对应的特征向量 $\mathbf{\Sigma}_{P_w}$ 能够近似最大化真实方差。与以往的研究一样，检测这一情况的一种简单方法是比较沿 u 方向的经验方差与沿相同方向的真实方差（尽管后者未知，但可以利用上述条目 4 中的方法轻松地进行稳健估计）。以下结论隐含于 [XCM13、JLT20、KSKO20]：

引理 2.15（基本证书引理）。考虑 设定 *2.13*。令 u 为一个满足以下条件的单位向量：$u^\top \mathbf{\Sigma}_{P_w} u \ge (1 - O(\gamma))\|\mathbf{\Sigma}_{P_w}\|_{\mathrm{op}}$。如果 $u^\top \mathbf{\Sigma} u \ge (1 - O(\gamma)) u^\top \mathbf{\Sigma}_{P_w} u$，那么我们有 $u^\top \mathbf{\Sigma} u / \|u\|_2^2 \ge (1 - O(\gamma))\|\mathbf{\Sigma}\|_{\mathrm{op}}$。

为完整起见，其证明可见于附录 A.3，其中该引理被重新表述为引理 A.7。

[6] 更准确地说，(iii) 中提到的断言需要证明；例如，参见 [DKPP22] .

2.3 过滤

现在我们回顾算法鲁棒统计文献中的标准过滤流程。该滤波器为每个点使用评分 $\tau(x)$，用于表示该点相对于分布的异常程度 x。给定一个（已知的）上界 T，用于表示仅在内点上的平均分数$\mathbf{E}_{X\sim G}[w(X)\tau(X)]$,，如果所有（被破坏的）点的平均分数$\mathbf{E}_{X\sim P}[w(X)\tau(X)]$,远大于 T，那么具有较大分数的点很可能是异常值。因此，算法 1 会移除所有分数大于某个（随机）阈值的点。[7] 我们还注意到以下事实：

（i）无需为每个 w（x）存储一个比特 x，我们可以存储对过滤器更简洁的描述，即仅存储 r_ℓ的第 4 行，并在需要时计算 w（x）。这一修改适用于流式设定。

（ii）由于每次我们都会移除所有满足 τ（x）$> r_\ell$和 $r_\ell \sim \mathcal{U}$（$[0, r_{\ell-1}]$）的点，因此 τ（x）的范围（按期望）会减半，从而算法 1按期望经过 O（log(max$_x \tau$（x）/ min$_x \tau$（x））步后终止。

以下结果表明，算法 1按期望移除的异常值多于内点。

算法 1 HardThresholdingFilter

$\sim$**1**：输入：分布 $P = (1-\epsilon)G + \epsilon B$、权重 w、分数 τ、参数 T、 R。**2**：$w_0(x) \leftarrow w(x)$，以及 $r_0 \leftarrow R$、 $\ell \leftarrow 1$。**3**：当 $\mathrm{E}_{X\sim P}[w(X)\tau(X)] > \frac{5}{2}\widehat{T}$时 **4**：抽取 $r_\ell \sim \mathcal{U}([0, r_{\ell-1}])$。**5**：$w_{\ell+1}(x) \leftarrow w_\ell(x) \cdot \mathbb{1}(\tau(x) > r_\ell)$。**6**：$\ell \leftarrow \ell + 1$。**7**：返回 $w_\ell(x)$。

$\widehat{}$**引理 2。16** T $1-\epsilon$ $_{X\sim G}[w\ X\ \tau\ X\ ] < T$ T(过滤保证)。 设满足$\widehat{}()\,\mathbf{E}()()$，且满足 $|T - T| < T/5$。记 F为 HardThresholdingFilter 的随机性（即所使用的随机阈值集合 $r_1, r_2 \ldots$）。令 w' 为返回的权重函数。那么，

(i) $\mathbf{E}_{X\sim P}[w'(X)\tau(X)] \le 3T$几乎必然成立，并且

(ii) $\mathbf{E}_F[\epsilon\,\mathbf{E}_{X\sim B}[w(X) - w'(X)]] > (1-\epsilon)\,\mathbf{E}_{X\sim G}[w(X) - w'(X)]]$.

证明引理 2.16 见附录 A.4。我们的主算法将反复调用HardThresholdingFilter，并使用适当的 τ（x）及其他参数。从技术角度来看，如果引理 2.16 的第二部分（即从异常值中移除的质量多于从内点中移除的质量）是确定性成立的，那么无论调用滤波器多少次，我们都有$\mathbf{E}_{X\sim G}[w$（X）$] \ge 1 - \epsilon$ 成立。这意味着设定 2.13将在整个主算法过程中保持成立。然而，引理 2.16 的第 (ii) 部分仅在期望意义下（相对于 F）成立，但仍可通过鞅论证证明，一个放宽后的条件$\mathbf{E}_{X\sim G}[w$（X）$] \ge 1 - 3\epsilon$以高概率在整个主算法过程中保持成立；更多细节见附录 A.4（引理 A.9）。初次阅读时，可以简单地认为设定 2.13 始终成立。

[7]因此，对于每个点 x, $\mathbb{P}$（x 被移除）$\propto \tau$（x）。

3 近线性时间的鲁棒主成分分析

组织结构 本节证明定理 3.1，这是对定理 1.2 的更一般化版本，后者在第 1.1 节中陈述。为简化表述，第 3.1 节 和第 3.2 节将重点建立定理 3.1的一个较弱版本，其运行时间为 $\widetilde{O}\left(nd/\gamma^3\right)$ （其中 γ 是稳定性参数；回顾一下，对于次高斯分布，有 $\gamma = O\left(\epsilon\log(1/\epsilon)\right)$ ）。尽管如此，在运行时间和误差保证两方面，这仍优于先前工作，同时有效展示了我们方法的关键要点。接下来，在第 3.3 节中，我们将概述一种改进方法，将运行时间降至$\widetilde{O}\left(nd/\gamma^2\right)$ 。最后，在第 3.4 节中，我们将给出定理 3.1 的形式化证明及全部细节。

现在，对我们的高斯分布的主要结果，该结果确立了 定理 1.2 因为 $\gamma =$ $\widetilde{O}(\epsilon)$

定理 *3.1.* 设 γ_0 是一个充分小的正常数。设 $d > 2$ 是一个整数，且$\epsilon, \gamma \in (0, 1)$，满足 $0 < 20\epsilon < \gamma < \gamma_0$。设 P是定义在包含 n 个点的集合上的均匀分布，该分布可分解为 $P = (1-\epsilon)G + \epsilon B$，其中 G 是关于半正定矩阵 $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$的$(20\epsilon, \gamma)$-稳定分布（定义 *2.12*）。存在一种算法，以 P, ϵ, γ为输入，运行时间为 $O(\frac{nd}{\gamma^2}\log^4(d/\epsilon))$，并且以至少 *0.99* 的概率输出一个单位向量u，使得 $u^\top \mathbf{\Sigma} u \geq (1 - O(\gamma))\|\mathbf{\Sigma}\|_{\mathrm{op}}$。

如前文所述，在 第 1.2 节和 第 2节中，我们运行迭代算法，在每次迭代t中，为每个点 x分配一个评分，并利用这些分数移除点。这些评分通常是点沿某个方向的投影，其中异常值的投影远大于内点。在满足稳定性条件时，只要经验（被污染的）方差在某个方向上比真实方差大 $(1 + C\gamma)$ 倍，我们就能可靠地滤除异常值。

由于每一轮过滤都会降低方差，我们看到 $\mathbf{\Sigma}_t$ 应随 t 减小（经过适当归一化后）。正如先前工作中指出的，使用涉及样本在经验二阶矩 $\mathbf{\Sigma}_t$ 的最大特征值对应的特征向量 v_t 上的投影的评分，不一定能取得良好进展，因为过滤只会沿最大方差的单一（固定）方向移除点（详情参见 第 1.2 节 ）。相反，需要同时（平均而言）沿许多这样的高方差方向进行过滤，我们通过沿方向 $v_t = \mathbf{M}_t z$ （其中$z \sim \mathcal{N}$取自$(\mathbf{0}, \mathbf{I})$ ）以及沿方向 $\mathbf{M}_t = \mathbf{\Sigma}_t^p$ 进行过滤来实现，其中整数 $p = {}^{C}$ 很大，且满足$\log(d/\gamma)\,\gamma$。

为了跟踪算法的进展，我们使用势函数 $\phi_t := \mathrm{tr}(\mathbf{\Sigma}_t^{2p+1})$。该势函数跟踪所有较大特征值的贡献（而不仅仅是最大的特征值），并且已被先前关于鲁棒均值估计的工作所采用。在我们的设定中，我们将证明：(I) 足够小的 ϕ_t （例如，$\phi_t \leq \|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p+1}/\mathrm{poly}(d^p)$）意味着 $\mathbf{\Sigma}$ 和 $\mathbf{\Sigma}_t$ 共享主特征子空间，且不对 $\mathbf{\Sigma}$的谱作任何限制；这进而可以保证生成一个良好的解向量；以及 (II) 如果 $\mathbf{\Sigma}$ 和 $\mathbf{\Sigma}_t$ 的谱不一致，那么我们可以（平均而言）将 ϕ_t 乘法式地降低。具体而言，我们将证明， $\phi_{t+1} \leq (1-\gamma)\phi_t$ 在平均意义下成立。结合简单的剪枝始终可以确保 $\phi_0 \leq \mathrm{poly}(d^p)\|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p+1}$这一事实，我们只需经过 $t = O(p\log(d)/\gamma)$轮，就可以得到$\phi_t < \|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p+1}/\mathrm{poly}(d)$。

现在我们解释上述 (I) 和 (II)：第 3.1 节将“共享主特征子空间”的概念形式化，而第 3.2 节概述势函数的下降过程。随后在第 3.4 节给出形式化证明。

3.1 高级证书引理

我们首先概述 [JLT20] 的算法框架。具体而言，我们重点指出其框架中移除特征值间隔条件的障碍，以及我们提出的改进方案；该方案依赖于将“共享主特征子空间”的概念形式化。

粗略地说，他们的迭代算法会在某次迭代 t 处停止，使得 $\|\mathbf{\Sigma}_t\|_p^p \le (1+C\gamma)^p\|\mathbf{\Sigma}\|_p^p$，其中 γ 是稳定性参数。[8] 然后，他们的算法会尝试 m 个 $\mathbf{\Sigma}_t$ 的顶部特征向量，其中 $m \in \mathbb{N}$ 是某个参数，并返回能够取得最佳 v 值的（归一化）特征向量$v^\top \mathbf{\Sigma} v$ （该值可以估计到较小误差）；参见 引理 2.15。具体而言，最后一步需要 $\sim O_\gamma(ndm)$ 时间。然而，正如以下例子所示（取 $r = d/2$ 为 $m = \Omega(d)$），这种方法无法实现近线性时间算法。

例 3.2 （Schatten p-范数停止条件的困难示例）。设 $\mathbf{\Sigma}$ 是一个秩为 r的半正定投影矩阵，那么 $\mathbf{\Sigma}_t = \mathbf{I}$ 满足条件 $\|\mathbf{\Sigma}_t\|_p^p \le (1+C\gamma)^p\|\mathbf{\Sigma}\|_p^p$，只要 $p \gtrsim \log(d/r)/\gamma$。然而，任何从 $\mathbf{\Sigma}_t$的特征向量中识别出 $\mathbf{\Sigma}$ 的良好方向的确定性算法，都至少需要 $m = \Omega(d-r)$。

因此， [JLT20] 额外要求最大特征值与 m中第 $\mathbf{\Sigma}$ 大特征值之间存在间隔，其中 m。因此，他们的结果及其证书版本本质上都受限于对 m的依赖；参见其中的命题 4。

出于技术原因，我们对算法作了两项关键修改（此外，我们还通过换元，改用 $2p+1$ 而不是 p）。第一项修改是强化停止条件：我们不再以降低 Schatten 范数为目标，因为这等价于 $\langle \mathbf{\Sigma}_t, \mathbf{\Sigma}_t^{2p}\rangle \le (1+C\gamma)^{2p+1}\langle \mathbf{\Sigma}, \mathbf{\Sigma}^{2p}\rangle$；而是采用更强的停止条件[9]，即达到 $\langle \mathbf{\Sigma}_t, \mathbf{\Sigma}_t^{2p}\rangle \le (1+C\gamma)\langle \mathbf{\Sigma}, \mathbf{\Sigma}_t^{2p}\rangle$。也就是说，只要在 $\mathbf{\Sigma}_t^p$、 $\langle \mathbf{\Sigma}_t, \mathbf{\Sigma}_t^{2p}\rangle$方向上的经验方差与真实方差 $\langle \mathbf{\Sigma}, \mathbf{\Sigma}_t^{2p}\rangle$相差至多一个 $(1+C\gamma)$ 因子（参见 引理 2.14，其中 $\mathbf{U} = \mathbf{\Sigma}_t^p$），我们就停止。注意，这个 $(1+C\gamma)$ 因子是比较稳健方差与经验方差时能够达到的最佳因子。之所以能够实现这一更强的停止条件，是因为在满足该条件之前，我们的算法会沿着 $\mathbf{\Sigma}_t^p$ 移除异常值，从而使势函数按乘法因子降低（这将在下一小节讨论）。

不过，来自 例 3.2 的困难示例仍满足这一更强条件：必须取 $m = \Theta(d\gamma)$ 若 $r = d/(1+\gamma)$。因此，任何确定性算法都需要 $m = \Omega(nd^2\gamma)$的时间，这相当于 d的平方级时间。我们的第二个主要见解是对算法的输出进行随机化。也就是说，尽管确定性算法需要尝试 $\Omega(\gamma d)$ 个 $\mathbf{\Sigma}_t$的最大特征值对应的特征向量，才能找到 $\mathbf{\Sigma}$的高方差方向，但从 $\mathbf{\Sigma}_t$ 的谱中随机抽取一个最大特征值对应的特征向量，以较大的常数概率就足够了。具体而言，我们抽取一个高斯样本，并按照幂迭代的方式反复将其与 $\mathbf{\Sigma}_t$ 相乘。形式化地，我们证明如下结论：

引理 *3.3*。设 C 为充分大的常数。设 $0 < 20\epsilon \le \gamma < \gamma_0$，其中 γ_0 为充分小的绝对常数。设 $P = (1-\epsilon)G + \epsilon B$，其中 G 是关于 $\mathbf{\Sigma}$ 的 $(20\epsilon, \gamma)$ -稳定分布。设 $w\colon \mathbb{R}^d \to [0,1]$，且 $E_{X\sim G}[w(X)] \ge 1-3\epsilon$，$p > C\log(d/\gamma)/\gamma$，$M := (E_{X\sim P}[w(X)XX^\top])^p$，并假设 $\langle \mathbf{\Sigma}, M^2\rangle \ge (1-250\gamma)\langle \mathbf{\Sigma}_{P_w}, M^2\rangle$。若 $u := Mz$，其中随机向量 $z \sim \mathcal{N}(0, I)$，则以至少 *0.9* 的概率（针对 z 的随机选取），有：

(i) $u^\top \mathbf{\Sigma}_{P_w} u / \|u\|_2^2 \ge (1-\gamma)\|\mathbf{\Sigma}_{P_w}\|_{\mathrm{op}}$,

(ii) $u^\top \mathbf{\Sigma} u > (1-O(\gamma))u^\top \mathbf{\Sigma}_{P_w} u$.

上述结果表明，只要 $\langle \mathbf{\Sigma}, \mathbf{M}^2\rangle \ge (1-250\gamma)\langle \mathbf{\Sigma}_{P_w}, \mathbf{M}^2\rangle$ 对于足够大的 p，我们就可以使用算法 2 得到一个向量 u，使其满足引理 2.15 中的“基本证书”条件，因此它是一个良好的解。因此，我们的主算法将调用 算法 2来输出一个向量。

[8]在稳定性下比较范数时，$(1+C\gamma)$ 是必要的。 [9]请参见 附录 A.2 了解为何这一条件更强。

算法 2 SampleTopEigenvector

1：输入：分布 P、权重 w、参数 ϵ, γ, δ。 2：$y \leftarrow \mathbf{\Sigma}_{P_w}^{p'} g$对 $p' = \frac{C}{\gamma}\log\left(\frac{d}{\gamma\delta}\right)$ 和 $g \sim \mathcal{N}(0, \mathrm{I})$。 3：$\widehat{r} \leftarrow y^\top \mathbf{\Sigma}_{P_w} y / \|y\|_2^2$。 4：令 $\mathrm{M} := (\mathrm{E}_{X\sim P}[w(X)XX^\top])^p$， ▷ cf. Fact 2.11 对于 $p = C\log(d/\gamma)\,\gamma$。 5：$u \leftarrow \mathrm{M}z$，对于 $z \sim \mathcal{N}(0, \mathrm{I})$。 6：找到 $\widehat{\sigma}_u$，使得 $|\widehat{\sigma}_u - u^\top \mathbf{\Sigma} u| \le 4\gamma u^\top \mathbf{\Sigma} u$。 7：如果 $\widehat{\sigma}_u \ge (1 - C\gamma) u^\top \mathbf{\Sigma}_{P_w} u$ 且 $\frac{u^\top \mathbf{\Sigma}_{P_w} u}{\|u\|_2^2} \ge (1-\gamma)\widehat{r}$，则 8：返回 $u/\|u\|_2$。 ▷ 参见 引理 2.15

9：**结束 if**

引理 *3.3* 的证明。结论中关于 $u^\top \mathbf{\Sigma}_{P_w} u / \|u\|_2^2 \ge (1-\gamma)\,\|\mathbf{\Sigma}_{P_w}\|_{\mathrm{op}}$ 的部分可直接由幂迭代的保证推出（事实 2.11 失败概率小于0。001）。

我们现在着重证明 $u^\top \mathbf{\Sigma} u > (1 - O(\gamma)) u^\top \mathbf{\Sigma}_{P_w} u$。令 $\mathrm{A} := \mathbf{\Sigma} - (1 - 250\gamma)\mathbf{\Sigma}_{P_w}$ 以高概率成立。我们分析随机变量 $Y := z^\top \mathrm{MAM} z$，其中 $z \sim \mathcal{N}(0, \mathrm{I})$。假设 $\langle \mathbf{\Sigma}, \mathrm{M}^2 \rangle \ge (1 - 250\gamma)\langle \mathbf{\Sigma}_{P_w}, \mathrm{M}^2 \rangle$ 意味着 $\langle \mathrm{M}^2, \Sigma - (1 - 250\gamma)\mathbf{\Sigma}_{P_w} \rangle = \langle \mathrm{M}^2, \mathrm{A} \rangle \ge 0$。注意到 $\mathrm{E}[Y] = \mathrm{tr}(\mathrm{MAM}) = \langle \mathrm{M}^2, \mathrm{A} \rangle$，我们有 $\mathrm{E}[Y] \ge 0$。因此，如果 Y 的方差不太大，那么以较高概率，Y 不会太负。

我们在 Y 的第 3.1.1 节中证明以下关于方差的技术结果。

断言 3.4（Y 的方差）。 $\mathrm{Var}(Y) \lesssim \gamma^2 \|\mathbf{\Sigma}_{P_w}\|_{\mathrm{op}}^2 \|\mathrm{M}\|_{\mathrm{F}}^4$。

断言 3.4 的证明概要。由事实 2.9 可知：$\mathrm{Var}[Y] = 2\|\mathrm{MAM}\|_{\mathrm{F}}^2$，因此我们需要证明 $\|\mathrm{MAM}\|_{\mathrm{F}} \lesssim \gamma \|\mathbf{\Sigma}_{P_w}\|_{\mathrm{op}} \|\mathrm{M}\|_{\mathrm{F}}^2$。回顾 A 的定义。我们可以将 $\mathbf{\Sigma}_{P_w}$ 分解为内点的贡献（由于稳定性，该贡献与 $\mathbf{\Sigma}$ 非常接近）以及异常值带来的贡献 $\mathbf{\Sigma}_B := \mathrm{E}_{X\sim B}[w(X)XX^\top]$。粗略地说，我们有 $\mathbf{\Sigma}_{P_w} \approx (1-\epsilon)\mathbf{\Sigma} + \epsilon \mathbf{\Sigma}_B$，这意味着 $\mathrm{A} \approx \gamma \mathbf{\Sigma} + \epsilon \mathbf{\Sigma}_B$ 如 $\epsilon \le \gamma$。

根据三角不等式，$\|\mathrm{MAM}\|_{\mathrm{F}} \le \gamma \|\mathrm{M}\mathbf{\Sigma}\mathrm{M}\|_{\mathrm{F}} + \epsilon \|\mathrm{M}\mathbf{\Sigma}_B\mathrm{M}\|_{\mathrm{F}}$。利用 $\mathbf{\Sigma}$ 和 M 的半正定性，可以很容易地对第一项进行如下上界估计：$\|\mathrm{M}\mathbf{\Sigma}\mathrm{M}\|_{\mathrm{F}} \lesssim \|\mathbf{\Sigma}\|_{\mathrm{op}} \|\mathrm{M}^2\|_{\mathrm{F}} \lesssim \|\mathbf{\Sigma}_{P_w}\|_{\mathrm{op}} \|\mathrm{M}\|_{\mathrm{F}}^2$，由事实 2.7 和稳定性可得。第二项的处理则更复杂。利用条件 $\langle \mathbf{\Sigma}, \mathrm{M}^2 \rangle \ge (1 - O(\gamma))\langle \mathbf{\Sigma}_{P_w}, \mathrm{M}^2 \rangle$中 $\mathbf{\Sigma}_{P_w}$ 的分解，我们得到 $\langle \mathbf{\Sigma}_B, \mathrm{M}^2 \rangle \lesssim (\gamma/\epsilon)\langle \mathbf{\Sigma}, \mathrm{M}^2 \rangle$。最后，由于 $\mathbf{\Sigma}_B$ 是半正定矩阵，我们得到 $\epsilon \|\mathrm{M}\mathbf{\Sigma}_B\mathrm{M}\|_{\mathrm{F}} \lesssim \epsilon\,\mathrm{tr}(\mathrm{M}\mathbf{\Sigma}_B\mathrm{M}) = \epsilon \langle \mathbf{\Sigma}_B, \mathrm{M}^2 \rangle \lesssim \gamma \langle \mathbf{\Sigma}, \mathrm{M}^2 \rangle \le \gamma \|\mathbf{\Sigma}\|_{\mathrm{op}} \|\mathrm{M}\|_{\mathrm{F}}^2$，从而得到所需结果。 □

利用 断言 3.4 以及切比雪夫不等式，我们可得，以至少 0.999 的 Y 满足以下下界：

$$Y \ge \mathbf{E}[Y] - \sqrt{1000\,\mathrm{Var}_{z\sim\mathcal{N}(0,\mathbf{I})}[Y]} = -O\left(\gamma \|\mathbf{M}\|_{\mathrm{F}}^2 \|\mathbf{\Sigma}_{P_w}\|_{\mathrm{op}}\right). \tag{1}$$

我们还需要一个中间结果。对于向量 $u := \mathbf{M}z$，当 $z \sim \mathcal{N}(0, \mathbf{I})$ 时，应用 事实 2.9，令 $\mathbf{A} = \mathbf{M}^2$ 且 $\beta = 10^{-4}/e$，可得

$$\mathbf{Pr}\left[\|\mathbf{M}\|_{\mathrm{F}}^2 / \|u\|_2^2 \le 1/\beta\right] = \mathop{\mathbf{Pr}}_{z\sim\mathcal{N}(0,\mathbf{I})}\left[z^\top \mathbf{M}^2 z \ge \beta\,\mathrm{tr}(\mathbf{M}^2)\right] \ge 1 - \sqrt{e\beta} = 0.99\,. \tag{2}$$

现在我们可以完成引理 3.3的证明。接下来，我们假设以下事件成立：(i) $\|\mathbf{M}\|_{\mathrm{F}}^2 / \|u\|_2^2 \le 1/\beta \le 10^5$，(ii) 方程（1）中的不等式成立，以及 (iii) 引理第一部分的结论成立。由于这些事件各自发生的概率均为0.99

因此，三个事件的交集发生的概率至少为0.97，这是由并集界得出的。回想一下，$Y = z^\top \mathbf{MAM}z = u^\top \mathbf{\Sigma} u - (1-250\gamma) u^\top \mathbf{\Sigma}_{P_w} u$，并将 $\|u\|_2^2$ 方程（1）的两边除以，我们得到：

$$\begin{aligned}
\frac{u^\top \mathbf{\Sigma} u}{\|u\|_2^2} &\geq (1-250\gamma)\frac{u^\top \mathbf{\Sigma}_{P_w} u}{\|u\|_2^2} - O\left(\gamma \frac{\|\mathbf{M}\|_{\mathrm{F}}^2}{\|u\|_2^2}\|\mathbf{\Sigma}_{P_w}\|_{\mathrm{op}}\right) \\
&\geq (1-250\gamma)\frac{u^\top \mathbf{\Sigma}_{P_w} u}{\|u\|_2^2} - O\left(\gamma \|\mathbf{\Sigma}_{P_w}\|_{\mathrm{op}}\right) \quad \text{(using that the event from Equation (2) holds)} \\
&\geq (1-250\gamma)\frac{u^\top \mathbf{\Sigma}_{P_w} u}{\|u\|_2^2} - O\left(\frac{\gamma}{1-\gamma}\frac{u^\top \mathbf{\Sigma}_{P_w} u}{\|u\|_2^2}\right) \qquad (u^\top \mathbf{\Sigma}_{P_w} u/\|u\|_2^2 \geq (1-\gamma)\|\mathbf{\Sigma}_{P_w}\|_{\mathrm{op}}) \\
&= (1-O(\gamma))\frac{u^\top \mathbf{\Sigma}_{P_w} u}{\|u\|_2^2}\,. \qquad (\gamma < 1/20)
\end{aligned}$$

因此，我们证明了引理 3.3的第二条结论以0.97的概率成立。对于足够大的常数 C，第一条结论以 0.99 的概率成立。因此，根据并集界，两条结论同时成立的概率至少为0.90。至此，引理 3.3得证。□

3.1.1 断言 3.4的证明：关于 Y方差的界

现在我们证明以下关于方差界的结果，该结果此前略去未证。

断言 3.4 （Y 的方差）。$\mathrm{Var}(Y) \lesssim \gamma^2 \|\mathbf{\Sigma}_{P_w}\|_{\mathrm{op}}^2 \|\mathbf{M}\|_{\mathrm{F}}^4$。

证明。由于 $Y = z^\top \mathbf{MAM}z$，事实 2.9 可推出$\mathrm{Var}[Y] = 2\|\mathbf{MAM}\|_{\mathrm{F}}^2$。因此，在证明的剩余部分，我们将给出$\|\mathbf{MAM}\|_{\mathrm{F}}$ 的上界。

由于 $P = (1-\epsilon)G + \epsilon B$，我们有 $\mathbf{\Sigma}_{P_w} = (1-\epsilon)\mathbf{\Sigma}_{P_w}^G + \epsilon \mathbf{\Sigma}_{P_w}^B$，其分量为$\mathbf{\Sigma}_{P_w}^G := \mathbf{E}_{X\sim G}[w(X)XX^\top]/\mathbf{E}_{X\sim P}[w(X)]$，和 $\mathbf{\Sigma}_{P_w}^B := \mathbf{E}_{X\sim B}[w(X)XX^\top]/\mathbf{E}_{X\sim P}[w(X)]$。利用这一分解，我们得到$\mathbf{A}$的如下表达式：

$$\begin{aligned}
\mathbf{A} &= \mathbf{\Sigma} - (1-250\gamma)\left((1-\epsilon)\mathbf{\Sigma}_{P_w}^G + \epsilon \mathbf{\Sigma}_{P_w}^B\right) \\
&= \left(\mathbf{\Sigma} - \mathbf{\Sigma}_{P_w}^G(1-250\gamma-\epsilon+250\gamma\epsilon)\right) - \epsilon(1-250\gamma)\mathbf{\Sigma}_{P_w}^B\,.
\end{aligned}$$

为简化符号约定，我们定义 $\mathbf{\Delta} := \mathbf{\Sigma} - \mathbf{\Sigma}_{P_w}^G(1-250\gamma-\epsilon+250\gamma\epsilon)$，以及 $\tilde{\epsilon} := \epsilon(1-250\gamma)$。因此有 $\mathbf{A} = \mathbf{\Delta} - \tilde{\epsilon}\mathbf{\Sigma}_{P_w}^B$。我们将分别对 $\|\mathbf{M\Delta M}\|_{\mathrm{F}}$ 和 $\|\mathbf{M\Sigma}_{P_w}^B\mathbf{M}\|_{\mathrm{F}}$ 给出上界。以下中间结果可由分布G的稳定性直接推出，因而会很有用：

命题 3.5。我们有 *(i)* $0 \preceq \mathbf{\Delta} \preceq 270\gamma\mathbf{\Sigma}_{P_w}$，*(ii)* $\mathbf{\Sigma} - (1-\epsilon)\mathbf{\Sigma}_{P_w}^G \preceq 1.2\gamma\mathbf{\Sigma}$，以及 *(iii)* $\mathbf{\Sigma} \preceq (1/(1-1.2\gamma))\mathbf{\Sigma}_{P_w}$。

证明。我们首先证明 $\mathbf{\Delta}$ 的半正定性。利用稳定性，我们先得到 $\mathbf{\Sigma}_{P_w}^G$ 的以下上界。

$$(1-4\epsilon)\mathbf{\Sigma}_{P_w}^G \preceq (1-\epsilon)(1-3\epsilon)\mathbf{\Sigma}_{P_w}^G \preceq \frac{\mathbf{E}_{X\sim P}[w(X)]}{\mathbf{E}_{X\sim G}[w(X)]}\mathbf{\Sigma}_{P_w}^G = \mathop{\mathbf{E}}_{X\sim G_w}[XX^\top] \preceq (1+\gamma)\mathbf{\Sigma}\,, \tag{3}$$

其中，第二个不等式在分母中使用了 $w(x) \leq 1$，并在分子中使用了假设 $\mathbf{E}_{X\sim P}[w(X)] \geq (1-\epsilon)\mathbf{E}_{X\sim G}[w(X)] \geq (1-\epsilon)(1-3\epsilon)$；最后一个不等式使用了 G 的稳定性。重新整理上述不等式，可得：

$$\mathbf{\Sigma}_{P_w}^G \preceq \frac{1+\gamma}{1-4\epsilon}\mathbf{\Sigma} \preceq \frac{1}{1-\epsilon-250\gamma+250\epsilon\gamma}\mathbf{\Sigma}\,,$$

其中，第一步使用了 公式（3），第二步对于 $\epsilon < \gamma/20$成立。将该不等式代入 $\mathbf{\Delta}$ 的定义可得 $\mathbf{\Delta} \succeq 0$。

现在我们证明 $\mathbf{\Delta} \preceq 270\gamma\mathbf{\Sigma}_{P_w}$。首先，我们注意到

$$\begin{aligned}\mathbf{\Sigma}_{P_w} &\succeq (1-\epsilon)\mathbf{\Sigma}^G_{P_w} = (1-\epsilon)\frac{\mathbf{E}_{X\sim G}[w(X)]}{\mathbf{E}_{X\sim P}[w(X)]}\mathop{\mathbf{E}}_{X\sim G_w}[XX^\top] \\ &\succeq (1-\epsilon)(1-3\epsilon)(1-\gamma)\mathbf{\Sigma} \succeq (1-1.2\gamma)\mathbf{\Sigma}\,, \end{aligned} \tag{4}$$

其中，第一个不等式使用了分解 $\mathbf{\Sigma}_{P_w} = (1-\epsilon)\mathbf{\Sigma}^G_{P_w} + \epsilon\mathbf{\Sigma}^B_{P_w}$，第二个不等式是改写，第三个不等式使用了 G 的稳定性以及 $\mathbf{E}_{X\sim G}[w(X)] \geq 1-3\epsilon$，最后一个不等式则利用了 $\epsilon < \gamma/20$。现在我们可以完成证明：

$$\begin{aligned}\mathbf{\Delta} &= \mathbf{\Sigma} - (1-\epsilon-250\gamma+250\gamma\epsilon)\mathbf{\Sigma}^G_{P_w} \\ &\preceq \mathbf{\Sigma} - (1-\epsilon-250\gamma+250\gamma\epsilon)(1-3\epsilon)(1-\gamma)\mathbf{\Sigma} && \text{(using middle part of Equation (4))} \\ &\preceq \mathbf{\Sigma} - (1-\epsilon-250\gamma+250\gamma\epsilon)(1-1.15\gamma)\mathbf{\Sigma} && (\epsilon \leq \gamma/20) \\ &\preceq 252\gamma\mathbf{\Sigma} && \text{(using } \epsilon \leq \gamma/20 \text{ and } \gamma \leq 1/20) \\ &\preceq \frac{252\gamma}{1-1.2\gamma}\mathbf{\Sigma}_{P_w} \preceq 270\gamma\mathbf{\Sigma}_{P_w}\,. && \text{(using Equation (4) and } \gamma \leq 1/20)\end{aligned}$$

将此与 $\mathbf{\Delta} \succeq 0$ 结合，即完成了断言第 (i) 部分的证明。断言第 (ii) 部分，即 $\mathbf{\Sigma} - (1-\epsilon)\mathbf{\Sigma}^G_{P_w} \preceq 1.2\gamma\mathbf{\Sigma}$，经过一些整理后可从公式（4）的中间部分得到。第 (iii) 部分也可直接由公式（4）推出。

□

现在我们可以给 Y 的方差作上界估计：利用 事实 2.9，可得

$$\mathrm{Var}[Y] = 2\|\mathbf{M}\mathbf{\Delta}\mathbf{M}\|_{\mathrm{F}}^2 \leq 4\|\mathbf{M}\mathbf{\Delta}\mathbf{M}\|_{\mathrm{F}}^2 + 4\tilde{\epsilon}^2\left\|\mathbf{M}\mathbf{\Sigma}^B_{P_w}\mathbf{M}\right\|_{\mathrm{F}}^2\,, \tag{5}$$

其中最后一个不等式来自三角不等式，且$(a+b)^2 \leq 2a+2b$。我们分别对每一项进行上界估计。利用命题 3.5，第一个方差项可作如下估计：

$$\begin{aligned}\|\mathbf{M}\mathbf{\Delta}\mathbf{M}\|_{\mathrm{F}}^2 &= \mathrm{tr}(\mathbf{M}^2\mathbf{\Delta}\mathbf{M}^2\mathbf{\Delta}) \text{（根据 } \|\mathbf{A}\|_{\mathrm{F}}^2 = \langle\mathbf{A},\,\mathbf{A}\rangle \text{ 对称矩阵 } \mathbf{A} \text{ 以及迹的循环性质）} \\ &\lesssim \gamma\mathrm{tr}(\mathbf{M}^2\mathbf{\Delta}\mathbf{M}^2\mathbf{\Sigma}_{P_w}) && \text{(using Claim 3.5 and Fact 2.2)} \\ &= \gamma\mathrm{tr}(\mathbf{M}^2\mathbf{\Sigma}_{P_w}\mathbf{M}^2\mathbf{\Delta}) && \text{(cyclic property of trace)} \\ &\lesssim \gamma^2\mathrm{tr}(\mathbf{\Sigma}_{P_w}\mathbf{M}^2\mathbf{\Sigma}_{P_w}\mathbf{M}^2) && \text{(using Claim 3.5 and Fact 2.2)} \\ &\leq \gamma^2\|\mathbf{\Sigma}_{P_w}\|_{\mathrm{op}}^2\mathrm{tr}(\mathbf{M}^4) && \text{(using Fact 2.2)} \\ &= \gamma^2\|\mathbf{\Sigma}_{P_w}\|_{\mathrm{op}}^2\|\mathbf{M}\|_4^4 \\ &\leq \gamma^2\|\mathbf{\Sigma}_{P_w}\|_{\mathrm{op}}^2\|\mathbf{M}\|_2^4 && \text{(using } \|\mathbf{A}\|_{q+1} \leq \|\mathbf{A}\|_q) \\ &= \gamma^2\|\mathbf{\Sigma}_{P_w}\|_{\mathrm{op}}^2\|\mathbf{M}\|_{\mathrm{F}}^4\,, && (6)\end{aligned}$$

上文第一次应用事实 2.2 时，需要满足$\mathbf{M}^2\mathbf{\Delta}\mathbf{M}^2 \succeq 0$，而这一条件确实成立，因为我们已在 $\mathbf{\Delta} \succeq 0$ 中命题 3.5 证明了这一点，因此事实 2.4 蕴含$\mathbf{M}^2\mathbf{\Delta}\mathbf{M}^2 \succeq 0$。第二次应用事实 2.2时也使用了类似的论证。

接下来需要对公式（5）的第二项进行界定。为此，我们有：

$$\begin{aligned}\epsilon\left\langle\mathbf{\Sigma}^B_{P_w}, \mathbf{M}^2\right\rangle &= \left\langle\mathbf{\Sigma}_{P_w} - (1-\epsilon)\mathbf{\Sigma}^G_{P_w}, \mathbf{M}^2\right\rangle \\ &\leq \left((1+O(\gamma))\left\langle\mathbf{\Sigma}, \mathbf{M}^2\right\rangle - (1-\epsilon)\left\langle\mathbf{\Sigma}^G_{P_w}, \mathbf{M}^2\right\rangle\right)\end{aligned}$$

$$= \left(\left\langle \mathbf{\Sigma} - (1-\epsilon)\mathbf{\Sigma}_{P_w}^G, \mathbf{M}^2 \right\rangle + O(\gamma)\left\langle \mathbf{\Sigma}, \mathbf{M}^2 \right\rangle\right)$$
$$\lesssim \gamma \left\langle \mathbf{\Sigma}, \mathbf{M}^2 \right\rangle , \tag{7}$$

其中，第一行使用了分解 $\mathbf{\Sigma}_{P_w} = (1-\epsilon)\mathbf{\Sigma}_{P_w}^G + \epsilon\mathbf{\Sigma}_{P_w}^B$，第二行使用了我们的假设 $\langle \mathbf{\Sigma}_{P_w}, \mathbf{M}^2\rangle \le \langle \mathbf{\Sigma}, \mathbf{M}^2\rangle/(1-250\gamma) \le (1+O(\gamma))\langle \mathbf{\Sigma}, \mathbf{M}^2\rangle$，第四行则根据命题$\mathbf{\Sigma} - (1-\epsilon)\mathbf{\Sigma}_{P_w}^G \preceq 1.2\gamma\mathbf{\Sigma} \preceq 2\gamma\mathbf{\Sigma}$ 得出命题 3.5。利用公式（7），我们最终可以对公式（5）的第二项给出上界：

$$\begin{aligned}
\tilde{\epsilon}^2 \left\|\mathbf{M}\mathbf{\Sigma}_{P_w}^B\mathbf{M}\right\|_{\mathrm{F}}^2 &\le \tilde{\epsilon}^2 \mathrm{tr}\left(\mathbf{M}\mathbf{\Sigma}_{P_w}^B\mathbf{M}\right)^2 && \text{(using Fact 2.8)}\\
&= \tilde{\epsilon}^2 \left\langle \mathbf{\Sigma}_{P_w}^B, \mathbf{M}^2 \right\rangle^2 \\
&\le \epsilon^2 \left\langle \mathbf{\Sigma}_{P_w}^B, \mathbf{M}^2 \right\rangle^2 && (\tilde{\epsilon} \le \epsilon)\\
&\lesssim \gamma^2 \left\langle \mathbf{\Sigma}, \mathbf{M}^2 \right\rangle^2 && \text{(by Equation (7))}\\
&\le \frac{\gamma^2}{1-1.2\gamma} \cdot \left\langle \mathbf{\Sigma}_{P_w}, \mathbf{M}^2 \right\rangle^2 && \text{(Claim 3.5 and Fact 2.2)}\\
&\lesssim \gamma^2 \|\mathbf{\Sigma}_{P_w}\|_{\mathrm{op}}^2 \mathrm{tr}(\mathbf{M}^2)^2 && (8)\\
&= \gamma^2 \|\mathbf{\Sigma}_{P_w}\|_{\mathrm{op}}^2 \|\mathbf{M}\|_{\mathrm{F}}^4 , && (9)
\end{aligned}$$

其中，第一步利用了 $\mathbf{\Sigma}_{P_w}^B$ 是半正定矩阵，因此 $\mathbf{M}\mathbf{\Sigma}_{P_w}^B\mathbf{M}$ 也是半正定的。综合上述结果，我们证明了 $\mathrm{Var}[Y] \lesssim \gamma^2\|\mathbf{\Sigma}_{P_w}\|_{\mathrm{op}}^2\|\mathbf{M}\|_{\mathrm{F}}^4$。 □

3.2 乘法式降低势函数

现在我们来描述算法 3，这是实现定理 1.2的主要组成部分。我们沿用算法 3中定义的符号约定。回想一下，作为输入给出的分布 P 就是（已损坏）数据集上的均匀分布，并假设其形式为 $P = (1-\epsilon)G + \epsilon B$ 其中 G 是（$C\epsilon$，γ）-稳定的（参见定义 2.12下方的注释）。

在本节中，我们将通过在每一轮 t 中跟踪势函数 $\phi_t := \mathrm{tr}(\mathbf{B}_t^{2p+1})$ 来量化算法的进展（其中 $\mathbf{B}_t$ 是 $\mathbf{\Sigma}_t$ 的缩放版本，定义于第 8 行）。[10] 如前所述，算法 3 的正确性可概括为以下论点：(i) 每当条件 $\langle \mathbf{\Sigma}, \mathbf{M}_t^2\rangle \ge (1-C\gamma)\langle \mathbf{\Sigma}_t, \mathbf{M}_t^2\rangle$ 为真时，我们可以使用 SampleTopEigenvector 输出一个良好解；(ii) 如果该条件为假，那么 ϕ_{t+1} 平均而言会以乘法方式减少 $\phi_{t+1} \le (1-\gamma)\phi_t$；(iii) 如果 $\phi_t \le \|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p+1}/\mathrm{poly}(d)$，则 (i) 中的条件必然为真。

我们已经在上一小节中证明了 (i)，并将在第 3.4.1 节 （引理 3.8）中陈述并证明 (iii)。在本节中，我们将重点以以下形式建立 (ii)：引理 3.6。

在证明引理 3.6 之前，我们先概述这些断言如何推出运行时间为 $\widetilde{O}(nd/\gamma^3)$[11] 4 $\phi_0 \le d/\epsilon^{O(p)}\|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p+1}$ $t_{\mathrm{end}} = (\,)$ 的定理 $O(^2 d/\epsilon)/\gamma^2$ t_{end} $\phi_{t_{\mathrm{end}}} < 1-\gamma^{t_{\mathrm{end}}}\phi_0 \le \|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p+1}/$ d 1.2 的一个版本：根据第 行的剪枝，$(\,)$。因此，存在 $(\log(\,))$，使得经过 轮后，$(\,)$ $\mathrm{poly}(\,)$，从而 SampleTopEigenvector 将输出一个良好的解。现在，循环的每次迭代都可以在 $O(ndp)$ 时间内实现：计算 $\mathbf{M}_t z$ 需要将经验二阶矩矩阵与 z 相乘 p 次，每次乘法都可以在 $O(nd)$ 时间内完成，因为 $\mathbf{\Sigma}_t z = \sum_x x(x^\top z)$。因此，总运行时间为 $\widetilde{O}(t_{\mathrm{end}} \cdot ndp) = \widetilde{O}(nd/\gamma^3)$。详细证明见第 3.4 节。

[10]在前面的章节中，为简便起见，我们使用了 $\mathrm{tr}(\mathbf{\Sigma}_t^{2p+1})$。由于我们的形式化证明需要 ϕ_t 随着 t自然递减，因此我们使用未归一化的二阶矩 $\mathbf{B}_t$，其满足 $\mathbf{B}_{t+1} \preceq \mathbf{B}_t$。 [11]朴素剪枝只会移除少量内点；见 引理 2.14。(v)

算法 3 少量迭代中的鲁棒主成分分析

1：输入： P, ϵ, γ。 2：令 $p = {}^{C}\log(d/\gamma)\,\gamma, t_{\text{end}} = {}^{C}\log^2(d/\epsilon)\,\gamma^2$，其中 C 足够大。 3：求估计量 $\widehat{\sigma}_{\text{op}} \in (0.8\|\mathbf{\Sigma}\|_{\text{op}}, 2d\|\mathbf{\Sigma}\|_{\text{op}})$。 ▷ 参见引理 2.14.(iv)，其中 $\mathrm{U} = \mathrm{I}$ 4：初始化 $w_1 \mathbb{1}(x) = \mathbb{1}\left(\|x\|_2^2 \le 10\widehat{\sigma}_{\text{op}}(d/\epsilon)\right)$。 5：对于 $t = 1, \ldots, t_{\text{end}}$，执行 6：调用 SampleTopEigenvector $(P, w_t, \epsilon, \gamma, \frac{1}{t_{\text{end}}})$。 7：令 P_t 为由 w_t 加权的 P 的分布： $P(x)w_t(x)/\mathrm{E}_{X\sim P}[w_t(X)]$。 8：令 $\mathrm{B}_t := \mathrm{E}_{X\sim P}[w_t(X)XX^\top]$，并令 $\mathrm{M}_t := \mathrm{B}_t^p$。 ▷ M_t 无需显式计算。 9：令 $g_t(x) := \|\mathrm{M}_t x\|_2^2$。 10：$v_t \leftarrow \mathrm{M}_t z_t$，其中 $z_t \sim \mathcal{N}(0, \mathrm{I})$。 11：令 $f_t(x) = (v_t^\top x)^2$。 12：令 L_t 为 $f_t(\cdot)$ 在 P_t 下的 3ϵ 分位数。 13：$L_t \leftarrow \max\{L_t, (0.1/d)\widehat{\sigma}_{\text{op}}\|v_t\|_2^2\}$ [12] 14：令 $\tau_t(x) = f_t(x)\mathbb{1}(f_t(x) > L_t)$ 15：求满足 $|\widehat{\sigma}_t - v_t^\top \mathbf{\Sigma} v_t| \le 4\gamma v_t^\top \mathbf{\Sigma} v_t$ 的 $\widehat{\sigma}_t$。（例如，$\widehat{\sigma}_t := \mathrm{E}_{X\sim P}[w_t(X)f_t(X)\mathbb{1}(f_t(X) \le L_t)]$）。 ▷ 2 14 $= v_t^\top$参见引理 . .(iv)，其中 U 16：$T_t \leftarrow 2.35\gamma\widehat{\sigma}_t$。 17：$w_{k,t+1} \leftarrow$ HardThresholdingFilter(P, w_t, τ_t, T_t, R)。 18：结束循环 19：返回 FAIL。

引理 3.6 (非正式)。 在 设定 *2.13* 中，若 $\langle \mathbf{\Sigma}, \mathbf{M}_t^2 \rangle \ge (1 - C\gamma)\langle \mathbf{\Sigma}_t, \mathbf{M}_t^2 \rangle$，则 $\mathbf{E}_t[\phi_{t+1}] \le (1-\gamma)\phi_t$是算法 *3*中的，其中 $\mathbf{E}_t$ 是截至第 t次迭代的条件期望。

引理 *3.6* 的证明概要引理 *3.6*。 为简单起见，我们假设算法（第 $f_t(x)$ 行）使用的分数 $g_t(x)$ 与 $g_t(x) = \|\mathbf{M}_t x\|_2^2$ 相同。注意，对所有 n 点计算 $g_t(x) = \|\mathbf{M}_t x\|_2^2$ 的计算成本会很高。这就是我们使用一维随机投影 $f_t(x)$的原因，它们是 $g_t(x)$的无偏估计量。使用这些估计量的完整证明要复杂得多，详见第 3.4.2 节。

根据我们的定义，我们计算第 $(t+1)$ 步中势函数的减少量：

$$
\begin{aligned}
\phi_{t+1} &= \operatorname{tr}\left(\mathbf{B}_{t+1}^{2p+1}\right) \le \operatorname{tr}\left(\mathbf{B}_t^p \mathbf{B}_{t+1} \mathbf{B}_t^p\right) = \operatorname{tr}(\mathbf{M}_t \mathbf{B}_{t+1} \mathbf{M}_t) \\
&= \mathop{\mathbf{E}}_{X\sim P}[w_{t+1}(X)g_t(X)] \\
&= (1-\epsilon) \mathop{\mathbf{E}}_{X\sim G}[w_{t+1}(X)g_t(X)] + \epsilon \mathop{\mathbf{E}}_{X\sim B}[w_{t+1}(X)g_t(X)] \\
&\le \mathop{\mathbf{E}}_{X\sim G}[w_t(X)g_t(X)] + \epsilon \mathop{\mathbf{E}}_{X\sim B}[w_{t+1}(X)g_t(X)] \,, \qquad (10)
\end{aligned}
$$

其中，第一个不等式利用了 $\mathrm{B}_{t+1} \preceq \mathrm{B}_t$ 在半正定序下递减这一性质（以及事实 2.3），最后一个不等式利用了 $w_{t+1} \le w_t$。我们证明上述右端为 $(1 + O(\gamma))\langle \mathbf{\Sigma}, \mathbf{M}_t^2 \rangle$：公式（10）中的第一项对应内点，利用稳定性（引理 2.14.(i)），可将其上界为 $(1 + 2\gamma)\langle \mathbf{\Sigma}, \mathbf{M}_t^2 \rangle$。对于公式（10）中的第二项，我们使用 $\tau_t(x) \le g_t(x) + L_t \le g_t(x) + 1.65(\gamma/\epsilon)\langle \mathbf{\Sigma}, \mathbf{M}_t^2 \rangle$，其中最后一个不等式利用了引理 2.14.(iii)。此外，过滤可确保

[12] 这确保了 $\tau(x)$ 在其非零时有一个下界。为何需要这样做，请参见 第 2.3 节 中的第二个要点。

其中 $\epsilon \mathbf{E}_{X\sim B}[w_{t+1}(X)\tau_t(X)] \le 2.35\gamma\langle \mathbf{\Sigma}, \mathbf{M}_t^2\rangle$，其中我们使用 引理 2.16，取 $T = 2.35\gamma\langle \mathbf{\Sigma}, \mathbf{M}_t^2\rangle$，选择此 T 的依据是引理 2.14。(ii)。结合这些界，我们得到

$$\phi_{t+1} \le (1 + O(\gamma))\langle \mathbf{\Sigma}, \mathbf{M}_t^2\rangle \,. \tag{11}$$

现在，利用我们的假设 $\langle \mathbf{\Sigma}, \mathbf{M}_t^2\rangle < (1 - C\gamma)\langle \mathbf{\Sigma}_t, \mathbf{M}_t^2\rangle$，我们如下完成证明：

$$\begin{aligned}\phi_{t+1} &< (1+O(\gamma))\langle \mathbf{\Sigma}, \mathbf{M}_t^2\rangle < (1+O(\gamma))(1-C\gamma)\langle \mathbf{\Sigma}_t, \mathbf{M}_t^2\rangle \\ &\le \frac{(1 + O(\gamma))(1 - C\gamma)}{\mathbf{E}_{X\sim P}[w_t(X)]}\langle \mathbf{B}_t, \mathbf{M}_t^2\rangle \le (1 - \gamma)\phi_t \,,\end{aligned}$$

利用 $C \gg 1$，$\langle \mathbf{B}_t$、$\mathbf{M}_t^2\rangle = \phi_t$和 $\mathbf{E}_{X\sim P}[w_t\ (X)\ ] \ge 1-O\ (\epsilon)$，因为过滤不会从内点中删除超过$O$（$\epsilon$）质量的部分（参见下文关于引理 2.16的讨论）。 □

上述概述展示了论证的主要思路，但相当简略；形式化证明见第 3.4.2 节。我们注意到，引理 3.6 对所有 $p \ge \log d$ 均成立；然而，引理 3.3 要求 p 较大。

3.3 改善运行时间对 ϵ 的依赖

我们首先从高层次描述这项改进，然后再转向第 3.4 节中的形式化证明。在这个较简单的算法版本中，运行时间较长的主要原因是：(i) 初始势函数为$\mathrm{poly}((d/\epsilon)^p)\|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p+1}$；(ii) p 的值为 $\frac{C}{\gamma}\log(\frac{d}{\gamma})$；(iii) 势函数每次仅减少$(1-\gamma)$ 倍；(iv) 算法在 $\phi_t \le \|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p+1}/\mathrm{poly}(d)$时终止。由于 (ii)–(iv) 很可能是必要的，我们对算法进行修改，以确保初始势函数显著减小：$\mathrm{poly}((d/\epsilon)^{\log(d)})\|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p+1}$。如果我们能以较低成本确保这一点，算法只需$\mathrm{polylog}(d/\epsilon)/\gamma$ 轮即可，从而使运行时间减少一个 $1/\gamma$ 因子。

其思路是：虽然确实需要 (ii)，但只有当算法即将给出最终解时才需要它。[13] 假设我们以 $p = \log d$ 运行算法 3，此时初始势函数确实为 $\mathrm{poly}((d/\epsilon)^{\log(d)})\|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p+1}$。在第 3.2 节中，我们保证势函数持续减少 $(1-\gamma)$，直到条件 $\langle \mathbf{\Sigma}, \mathbf{M}_t^2\rangle \ge (1 - C\gamma)\langle \mathbf{\Sigma}_t, \mathbf{M}_t^2\rangle$ 被触发。然而，即使该条件被触发，我们也不知道算法 2 是否会如 $p \ll \log(d/\gamma)/\gamma$ 那样成功（参见引理 3.3）。尽管如此，当该条件被触发时，我们会发现最终势函数已经大幅减小：$\phi_t = \langle \mathrm{B}_t, \mathbf{M}_t^2\rangle \le \langle \mathbf{\Sigma}_t, \mathbf{M}_t^2\rangle \le (1 + C\gamma)\langle \mathbf{\Sigma}, \mathbf{M}_t^2\rangle \le (1 + C\gamma)\|\mathbf{\Sigma}\|_{\mathrm{op}}\|\mathrm{B}_t\|_{2p}^{2p}$，其中第二步利用了该条件不成立这一事实。利用事实 2.7 建立 $\|\mathrm{B}_t\|_{2p}^{2p}$ 与 $\phi_t = \|\mathrm{B}_t\|_{2p+1}^{2p+1}$ 之间的关系，可得 $\phi_t \le \mathrm{poly}(d/\epsilon)\|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p+1}$。

到目前为止，我们已经证明，从 $\phi_0 \le \mathrm{poly}((d/\epsilon)^{\log(d)})\|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p+1}$开始，可以将 ϕ_t 降至 $\phi_t \le \mathrm{poly}(d)\|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p+1}$，这需要 $t \asymp \log^2(d/\epsilon)/\gamma$ 轮。然后，我们可以将 p 的值加倍至 $p' = 2p$，并重新启动计数器 t。新的势函数 ϕ_0' 将利用 事实 2.7 进行如下上界估计：

$$\phi_0' = \|\mathbf{B}_t\|_{2p'+1}^{2p'+1} \le \|\mathbf{B}_t\|_{2p+1}^{2p'+1} = (\phi_t)^{\frac{2p'+1}{2p+1}} = (\phi_t)^{\frac{4p+1}{2p+1}},$$

也就是说，它仍然是 $\mathrm{poly}(d)\|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p'+1}$。我们反复运行相同的过程，直到 p 达到其最终值 $\Theta(\log^2(d/\gamma)/\gamma)$，此时前两节的分析便适用，从而返回一个良好的向量。这就得到 算法 4中概述的嵌套循环算法。该算法将实现定理 3.1。

[13] 引理 3.3 (证书引理)要求 $p > C\log(d/\gamma)/\gamma$。

算法 4 改进运行时间的 RobustPCA

1: 输入： P, ϵ, γ。 2: 令 C, C' 为足够大的绝对常数，且其比值 C/C' 足够大。 3: 令 $k_{\mathrm{end}} := \log((\log(d/\gamma)/\log(d))/\gamma)$，以及 $t_{\mathrm{end}} := {}^{C}\log^2(d/\epsilon)\,\gamma$。 4: 求估计量 $\widehat{\sigma}_{\mathrm{op}} \in (0.8\|\mathbf{\Sigma}\|_{\mathrm{op}}, 2d\|\mathbf{\Sigma}\|_{\mathrm{op}})$。 ▷ 参见 $\mathrm{U} = \mathrm{I}$ 时的引理 2.14 的第 (iv) 项。 5: 初始化 $w_{1,1}(x) = \mathbb{1}\left(\|x\|_2^2 \le 10\widehat{\sigma}_{\mathrm{op}}(d/\epsilon)\right)$。 6: 对 $k = 1, \ldots, k_{\mathrm{end}}$ 执行 7: 令 $p_k = 2^{k-1}p$，其中 $p = C'\log(d)$。 ▷ p_k 的取值范围从 $C'\log(d)$ 到 $C'\log(d/\gamma)/\gamma$。 8: 对 $t = 1, \ldots, t_{\mathrm{end}}$ 执行 9: 调用 SampleTopEigenvector $(P, w_{k,t}, \epsilon, \gamma, 1/(k_{\mathrm{end}} \cdot t_{\mathrm{end}}))$。 ▷ 参见算法 2。 10: 令 $P_{k,t}$ 为按 $w_{k,t}$ 加权后的 P 的分布： $P(x)w_{k,t}(x)/\mathrm{E}_{X\sim P}[w_{k,t}(X)]$。 11: 令 $\mathrm{B}_{k,t} := \mathrm{E}_{X\sim P}[w_{k,t}(X)XX^\top]$，并令 $\mathrm{M}_{k,t} := \mathrm{B}_k^{p_k}, t$。 ▷ 无需显式计算 $\mathrm{M}_{k,t}$。 12: 令 $g_{k,t}(x) := \|\mathrm{M}_{k,t}x\|_{2^n}^2$ 13: $v_{k,t} \leftarrow \mathrm{M}_{k,t}z_{k,t}$，其中 $z_{k,t} \sim \mathcal{N}(0, \mathrm{I})$。 14: 令 $f_{k,t}(x) = (v_{k,t}^\top x)^2$。 15: 令 $L_{k,t}$ 为在 $P_{k,t}$ 下 $f_{k,t}(\cdot)$ 的 3ϵ 分位数。 16: $L_{k,t} \leftarrow \max\{L_{k,t}, (0.1/d)\widehat{\sigma}_{\mathrm{op}}\|v_{k,t}\|_2^2\}$ 17: 令 $\tau_{k,t}(x) = f_{k,t}(x)\mathbb{1}(f_{k,t}(x) > L_{k,t})$ 18: 求 $\widehat{\sigma}_{k,t}$，使得 $|\widehat{\sigma}_{k,t} - v_{k,t}^\top \mathbf{\Sigma} v_{k,t}| \le 4\gamma v_{k,t}^\top \mathbf{\Sigma} v_{k,t}$（例如，$\widehat{\sigma}_{k,t} := \mathrm{E}_{X\sim P}[w_{k,t}(X)f_{k,t}(X)\mathbb{1}(f_{k,t}(X) \le L_{k,t})]$）。 ▷ 参见 $\mathrm{U} = v_{k,t}^\top$ 时的引理 2.14 的第 (iv) 项。 19: $T_{k,t} \leftarrow 2.35\gamma\widehat{\sigma}_{k,t}$。 20: $w_{k,t+1} \leftarrow \mathrm{HardThresholdingFilter}(P, w_{k,t}, \tau_{k,t}, T_{k,t}, R, 0)$。 ▷ 参见算法 4。 21: 结束循环 22: 设 $w_{k+1,0} \leftarrow w_{k,t+1}$。 23: 结束循环 24: 返回 FAIL。

3.4 定理 3.1 的证明

现在，我们给出定理 3.1 的形式化证明，并补充前面概要中省略的细节，例如，算法计算出的分数 $f_t(x)$ 与 $g_t(x)$ 并不一致这一事实。我们还将结合第 3.3 节 中关于改进运行时间的论证。

如前所述，证明包含三个主要断言：(i) 当条件 $\langle \mathbf{\Sigma}, \mathbf{M}_t^2 \rangle \ge (1 - C\gamma)\langle \mathbf{\Sigma}_t, \mathbf{M}_t^2 \rangle$ 成立时（只要 p 足够大），我们就能通过SampleTopEigenvector 输出一个良好的解；(ii) 若该条件不成立，则 ϕ_{t+1} 平均而言会以乘法方式下降 $\phi_{t+1} \le (1 - \gamma)\phi_t$；(iii) 若 $\phi_t \le \|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p+1}/\mathrm{poly}(d)$，则 (i) 中的条件必然成立。第一部分已在引理 3.3中证明；最后一部分是由稳定性直接推出的结论，证明见第 3.4.1 节 (引理 3.8)；对应于 (ii) 的结果则在第 3.4.2 节中证明。随后，我们在第 3.4.3 节中将这些部分合并，完成定理的证明。详细的运行时间分析见第 3.4.4 节。

在符号约定方面，我们将沿用算法 4中定义的符号约定。特别是，许多相关量会带有两个下标，例如 $\mathbf{\Sigma}_{k,\,t}$：这里 k 表示外层循环， ℓ 表示内层循环。

对于形式化证明，我们将用到这样一个结论：在某个绝对常数概率下，分数$f_{k,\,t}(x)$是分数 $g_{k,\,t}(x)$的常数因子近似（我们在前一小节的证明草图中省略了这一细节）

）。这由事实 2.9可得。此外，为了使本节的分析更灵活，并更容易将其调整到下一节的流式设定中，我们将对 $f_t(\cdot)$ 采用较弱的近似，其中还包括一个加性误差项（该加性误差将用于处理流式处理部分中对$\mathbf{M}$ 的近似）。

假设 3.7。 向量v_k的随机选取t 在第13 行的算法 4 中满足：对于定义域中的每个点 x ，

$$\Pr_{v_{k,t}}\left[f_{k,t}(x) > \frac{g_{k,t}(x)}{10} - 0.01\frac{\gamma}{\epsilon}\|\mathbf{M}_{k,t}\|_{\mathrm{F}}^2\|\mathbf{\Sigma}\|_{\mathrm{op}}\right] \geq 0.4\ .$$

我们再次强调，事实 2.9 意味着假设 3.7 对于算法 4成立。

3.4.1 小势函数意味着停止条件成立

本节将正式证明以下结果：当势函数 $\phi := \mathrm{tr}(\mathbf{B}^{2p+1})$ 小于 $\|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p+1}/\mathrm{poly}(d/\epsilon)$，且 p 足够大时，前一个引理的条件便成立。注意，下一个引理结论中提到的因子 $(1-2\gamma)^{2p}$ 确实大于 $1/\mathrm{poly}(d/\epsilon)$，因为 $p = O(\log(d/\gamma)/\gamma)$。

引理 3.8。设 $0 < 20\epsilon \leq \gamma < 1/20$，且 p 为正整数。设 $P = (1-\epsilon)G + \epsilon B$，其中 G 是相对于 $\mathbf{\Sigma}$ 的 $(20\epsilon, \gamma)$-稳定分布；再设 $w: \mathbb{R}^d \to [0,1]$，且 $E_{X\sim G}[w(X)] \geq 1-3\epsilon$。定义 $B := E_{X\sim P}[w(X)XX^\top]$，$M := B^p$，以及 $\phi := tr(B^{2p+1})$。若 $\phi \leq E_{X\sim P}[w(X)]\frac{(1-2\gamma)^{2p}}{1-250\gamma}\|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p+1}$，则$\langle \mathbf{\Sigma}, M^2\rangle \geq (1 - 250\gamma)\langle \mathbf{\Sigma}_{P_w}, M^2\rangle$。

证明。 我们断言，要使该陈述成立，只需证明

$$\langle \mathbf{M}^2, \mathbf{\Sigma}\rangle \geq (1-2\gamma)^{2p}\|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p+1}\ . \tag{12}$$

这确实足够，因为如果我们已有 公式（12） ，那么只要它成立 $\phi \leq \mathbf{E}_{X\sim P}[w(X)]\frac{(1-2\gamma)^{2p}}{1-250\gamma}\|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p+1}$，就有

$$\langle \mathbf{M}^2, \mathbf{\Sigma}_{P_w}\rangle = \frac{\phi}{\mathbf{E}_{X\sim P}[w(X)]} \leq \frac{(1-2\gamma)^{2p}}{1-250\gamma}\|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p+1} \leq \frac{1}{1-250\gamma}\langle \mathbf{M}^2, \mathbf{\Sigma}\rangle\ ,$$

其中，第一步使用了定义 $\phi = \mathrm{tr}(\mathbf{B}^{2p+1}) = \langle \mathbf{M}^2, \mathbf{B}\rangle = \mathbf{E}_{X\sim P}[w(X)]\langle \mathbf{M}^2, \mathbf{\Sigma}_{P_w}\rangle$，第二步使用了 $\phi \leq \mathbf{E}_{X\sim P}[w(X)]\frac{(1-2\gamma)^{2p}}{1-250\gamma}\|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p+1}$，最后一步使用了 公式（12）。

在补充材料中，我们证明 公式（ 12 $\mathbf{\Sigma} =$ ）。
$\sum_{i=1}^d \lambda_i u_i u_i^\top$ $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ $\langle\ ^2\ \mathbf{\Sigma}\rangle$ $\langle\ ^2\ \lambda_1 u_1 u_1^\top\rangle$
考虑如下谱分解。我们将使用以下方法给出 $\mathbf{M}$，的下界：$\mathbf{M}$，如下：

$$\langle \mathbf{M}^2, \mathbf{\Sigma}\rangle = \left\langle \mathbf{M}^2, \sum_{i=1}^d \lambda_i u_i u_i^\top \right\rangle = \sum_{i=1}^d \lambda_i u_i^\top \mathbf{M}^2 u_i \geq \lambda_1 u_1^\top \mathbf{M}^2 u_1 = \|\mathbf{\Sigma}\|_{\mathrm{op}} u_1^\top \mathbf{M}^2 u_1\ . \tag{13}$$

在证明的剩余部分中，我们将证明 $u_1^\top \mathbf{M}^2 u_1 \geq (1-2\gamma)^{2p}\|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p+1}$，从而完成证明。我们先在下文的半正定意义下给出 $\mathbf{B}$ 的下界 $\mathbf{\Sigma}$：

$$\begin{aligned}\mathbf{B} &= \mathop{\mathbf{E}}_{X\sim P}[w(X)XX^\top] \succeq (1-\epsilon)\mathop{\mathbf{E}}_{X\sim G}[w(X)]\frac{\mathbf{E}_{X\sim G}[w(X)XX^\top]}{\mathbf{E}_{X\sim G}[w(X)]} \\ &\succeq (1-\epsilon)(1-3\epsilon)(1-\gamma)\mathbf{\Sigma} \succeq (1-2\gamma)\mathbf{\Sigma}\ ,\end{aligned} \tag{14}$$

其中，第一行用到了 $P=(1-\epsilon)G+\epsilon B$，最后一行用到了 $\mathbf{E}_{X\sim G}[w(X)]\geq 1-3\epsilon$、归一化分布 G_w的二阶矩稳定性条件，而最后一步用到了 $\epsilon<\gamma/20$。由于 $\mathbf{M}=\mathbf{B}^p$，且 $\mathbf{B}$ 至少为 $\mathbf{\Sigma}$，我们得到 $u_1^\top\mathbf{M}^2u_1$的如下下界：

$$u_1^\top\mathbf{M}^2u_1=u_1^\top\mathbf{B}^{2p}u_1\geq(u_1^\top\mathbf{B}u_1)^{2p}\geq((1-2\gamma)u_1^\top\mathbf{\Sigma}u_1)^{2p}=(1-2\gamma)^{2p}\|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p},$$

其中，第一个不等式用到了事实 2.5，第二个不等式用到了公式（14）。 □

3.4.2 势函数乘法递减的形式化证明

本节将给出一个形式化证明，证明以下结论：如果条件 $\langle\mathbf{\Sigma},\mathbf{M}_{k,t}^2\rangle\geq(1-250\gamma)\langle\mathbf{\Sigma}_{k,t},\mathbf{M}_{k,t}^2\rangle$ 不成立，那么势函数的期望值会乘法递减；回顾一下，(k,t) 表示正在分析的迭代的双重索引。首先，我们证明引理 3.9，它分析了在前几次迭代条件下，算法内层循环的单次迭代。然后，我们陈述推论 3.11，说明整个循环结束时，势函数的期望值会减少$(1-\gamma)^t$ 倍。

下面的引理 3.9 指出，如果对于每次先前的迭代 (k',t')，都有 $\langle\mathbf{\Sigma},\mathbf{M}_{k',t'}^2\rangle<(1-250\gamma)\langle\mathbf{\Sigma}_{k',t'},\mathbf{M}_{k',t'}^2\rangle$（这对应于引理陈述中的事件 $\mathcal{E}_{k,t}^{(2)}$），[14] 并且异常值所占比例仍为 $O(\epsilon)$（如事件 $\mathcal{E}_{k,t}^{(1)}$ 中所述，），[15] 那么我们有 $\phi_{k,t+1}\leq(1-\gamma)\phi_{k,t}$，且该结论在期望意义下成立。此处的期望是针对随机方向 $v_{k,t}$ 的随机性取的（参见 RobustPCA 的第 13 行）；这些方向用于定义评分 $f_{k,t}$。第 3.2 节的证明概要作了简化，假设 $f_{k,t}$ 始终等于其期望值，因此我们在那里确定性地证明了 $\phi_{k,t+1}\leq(1-\gamma)\phi_{k,t}$；但正如我们将在下文的形式化证明中所示，这一结论是在对 $f_{k,t}$ 的随机性取期望后成立的。HardThresholdingFilter 的随机性（在引理陈述中记作 $\mathcal{F}_{k,t}$）与此结论无关（它只影响算法的运行时间，详见第 3.4.4 节），因此，无论如何对$\mathcal{F}_{k,t}$进行条件化，势函数的期望值都会乘法递减（公式（15））。

引理 3.9。考虑算法 4中的符号约定，其中输入分布为混合分布$P=(1-\epsilon)G+\epsilon B$，其中 G 是相对于 $\mathbf{\Sigma}$ 满足 $0<20\epsilon\leq\gamma<\gamma_0$的$(20\epsilon,\gamma)$稳定分布，且正常数 γ_0充分小。作出假设 3.7。令 $\mathcal{E}_{k,t}=\mathcal{E}_{k,t}^{(1)}\cap\mathcal{E}_{k,t}^{(2)}$ 表示以下两个事件的交集：

(i) $\mathcal{E}_{k,t}^{(1)}$：$(1-\epsilon)\,\mathbf{E}_{X\sim G}[1-w_{k',t'}(X)]+\epsilon\,\mathbf{E}_{X\sim B}[w_{k',t'}(X)]\leq 3\epsilon$，对每次迭代$(k',t')$，只要它发生在$(k,t)$之前（含该次迭代）。

(ii) $\mathcal{E}_{k,t}^{(2)}$：$\langle\mathbf{\Sigma},\mathbf{M}_{k',t'}^2\rangle<(1-250\gamma)\langle\mathbf{\Sigma}_{k',t'},\mathbf{M}_{k',t'}^2\rangle$ 对于从(k',t') 到（含）$(k,1)$ 的每次迭代，直至(k,t)。

定义势函数 $\phi_{k,t}:=tr(B_{k,t}^{2p_k+1},)$。令 $F_{k,t}$ 为 *RobustPCA* 在第 *(k,t)* 次循环中 *HardThresh- oldingFilter* 使用的随机性（即 $F_{k,t}$ 是该滤波器使用的随机变量 $r_1,r_2,\ldots$ 的集合），并令 $v_{k,t}$ 为 *Robust- PCA*（算法 *4*）第 *13* 行中使用的随机向量。此外，定义 $\mathcal{F}_{k,t}=\{F_{k',t'}:\ k'\leq k-1,t'\leq t_{\mathrm{end}}\}\cup\{F_{k,t'}:\ t'\leq t\}$，即截至（并包括）第 *(k,t)* 次迭代的所有过滤器历史记录。对 $v_{k,t}$ 也类似地定义 $\mathcal{V}_{k,t}$。

[14]注意，事件 $\mathcal{E}_{k,t}^{(2)}$ 意味着算法尚未结束（因为我们知道，事件首次发生时，$\langle\mathbf{\Sigma},\mathbf{M}_{k,t}^2\rangle\geq(1-250\gamma)\langle\mathbf{\Sigma}_{k,t},\mathbf{M}_{k,t}^2\rangle$，根据引理 3.3，算法将以高概率停止）。[15]稍后我们将证明，该事件确实会在整个算法过程中持续成立，作为一个不变条件。

那么，对于对 $\mathcal{F}_{k,\,t}$、$\mathcal{V}_{k,\,t-1}$ 的每一种条件化，以下结论均成立：

$$\mathop{\mathbf{E}}_{v_{k,t}}\left[\phi_{k,t+1}\mathbb{1}(\mathcal{E}_{k,t}) \mid \mathcal{F}_{k,t}, \mathcal{V}_{k,t-1}\right] \le (1-\gamma)\phi_{k,t}\mathbb{1}(\mathcal{E}_{k,t-1})\,. \tag{15}$$

证明。我们分析 $(k,\ t)$- 次迭代中的 RobustPCA，即索引为 k 的外层循环以及索引为 t的内层循环。我们将此前发生的一切视为已知，只关注当前内层循环轮次中由 $v_{k,\,t}$带来的随机性会产生什么影响。具体而言，期望是以过去的历史记录$\mathcal{F}_{k,\,t}$，$\mathcal{V}_{k,\,t-1}$ 为条件的，如 公式（15） 所示；为方便符号表达，我们省略了对这些条件的明确写出。

当 $\mathbf{1}(\mathcal{E}_{k,\,t}) = 0$ 时，引理的结论显然成立。因此，在证明的其余部分，我们分析 $\mathbf{1}(\mathcal{E}_{k,\,t}) = 1$ 的情形。

如果点 x 是满的，我们称其为「满点」；若 $f_{t,\,k}(x) > g_{k,\,t}(x)/10 - 0.01(\gamma/\epsilon)\|\mathbf{M}_{k,\,t}\|_{\mathrm{F}}^2\|\mathbf{\Sigma}\|_{\mathrm{op}}$，否则称其为空点。注意，$x$ 为满意味着

$$\begin{aligned}\tau_{k,t}(x) &\ge f_{k,t}(x) - L_{k,t} \ge \frac{g_{k,t}(x)}{10} - L_{k,t} - 0.01\frac{\gamma}{\epsilon}\|\mathbf{M}_{k,t}\|_{\mathrm{F}}^2\|\mathbf{\Sigma}\|_{\mathrm{op}} \\ &\ge \frac{g_{k,t}(x)}{10} - 1.65\frac{\gamma}{\epsilon}\|v_{k,t}\|_2^2\|\mathbf{\Sigma}\|_{\mathrm{op}} - 0.01\frac{\gamma}{\epsilon}\|\mathbf{M}_{k,t}\|_{\mathrm{F}}^2\|\mathbf{\Sigma}\|_{\mathrm{op}}\,,\end{aligned} \tag{16}$$

其中第一个不等式使用了 $\tau_{k\text{ 的定义},\,t}$，最后一个不等式使用了 条目 (iii) 的引理 2.14。

现在，我们使用引理 A.8 来分析通过 HardThresholdingFilter（算法 8）进行过滤的效果：我们将引理应用于 $T = 2.35\gamma v_{k}^{\top}{}_{,\,t}\mathbf{\Sigma}v_{k,\,t}$、$\widehat{T} = 2.35\gamma\widehat{\sigma}_{k,\,t}$ 和 $\delta = 0$；现在说明选择这些参数的理由。引理 A.8 的第一个要求是 $(1-\epsilon)\,\mathrm{E}_{X\sim G}\left[w_{k,t}(x)\tau_{k,t}(x)\right] < T$，引理 2.14 的条目 (ii) 满足了这一要求（具体而言，是使用 $\mathrm{U} = v_{k}^{\top}{}_{,t}$ 的特殊情形）。对于其第二个要求 $\widehat{T} > T/5$，我们有 $|\widehat{T} - T| = 2.35\gamma|\widehat{\sigma}_{k,t} - v_{k}^{\top}{}_{,t}\mathbf{\Sigma}v_{k,t}| \le 2.35\gamma\cdot 4\gamma\cdot v_{k}^{\top}{}_{,t}\mathbf{\Sigma}v_{k,t} \le (1/5)\cdot 2.35\gamma v_{k}^{\top}{}_{,t}\mathbf{\Sigma}v_{k,t} = T/5$，其中第二步使用了引理 2.14 的第 (iv) 项（取 $\mathrm{U} = v_{k}^{\top}{}_{,t}$），最后一个不等式则利用了 $\gamma < 1/20$。因此，该引理适用，并可推出<lemma>

$$\mathop{\mathbf{E}}_{X\sim P}[w_{k,t+1}(X)\tau_{k,t}(X)] \le 3T \le 7.1\gamma\|v_{k,t}\|_2^2\|\mathbf{\Sigma}\|_{\mathrm{op}}\,, \tag{17}$$

以概率 1 成立。因此，对于完整的坏点，其在当前迭代经过滤波器后更新的权重满足：

$$\begin{aligned}\epsilon\mathop{\mathbf{E}}_{X\sim B}[w_{k,t+1}(X)g_{k,t}(X)\mathbb{1}(X\text{ full})] &\le 10\epsilon\mathop{\mathbf{E}}_{X\sim B}[w_{k,t+1}(X)\tau_{k,t}(X)] + 16.5\gamma\|v_{k,t}\|_2^2\|\mathbf{\Sigma}\|_{\mathrm{op}} + 0.1\gamma\,\|\mathbf{M}_{k,t}\|_{\mathrm{F}}^2 \\ &< 88\gamma\|v_{k,t}\|_2^2\|\mathbf{\Sigma}\|_{\mathrm{op}} + 0.1\gamma\,\|\mathbf{M}_{k,t}\|_{\mathrm{F}}^2\,\|\mathbf{\Sigma}\|_{\mathrm{op}}\,.\end{aligned} \tag{18}$$

其中，第一步使用 公式 (16)，第二步使用 公式 (17)。

根据假设 3.7，每个点以 0.4 的概率是满的。现在，我们将对 $v_{k,\,t}$ 取期望（再次提醒读者，该期望是在给定过去历史$\mathcal{F}_{k,\,t}$，$\mathcal{V}_{k,\,t-1}$ 的条件下计算的，如公式（15）所示；为方便符号表达，我们省略了这些条件）。

$$\begin{aligned}&\mathop{\mathbf{E}}_{v_{k,t}}\left[\epsilon\mathop{\mathbf{E}}_{X\sim B}[w_{k,t+1}(X)g_{k,t}(X)]\right] \\ &= \mathop{\mathbf{E}}_{v_{k,t}}\left[\epsilon\mathop{\mathbf{E}}_{X\sim B}[w_{k,t+1}(X)g_{k,t}(X)\mathbb{1}(X\text{ full})]\right] + \mathop{\mathbf{E}}_{v_{k,t}}\left[\epsilon\mathop{\mathbf{E}}_{X\sim B}[w_{k,t+1}(X)g_{k,t}(X)\mathbb{1}(X\text{ empty})]\right] \\ &\le \mathop{\mathbf{E}}_{v_{k,t}}\left[\epsilon\mathop{\mathbf{E}}_{X\sim B}[w_{k,t+1}(X)g_{k,t}(X)\mathbb{1}(X\text{ full})]\right] + \epsilon\mathop{\mathbf{E}}_{X\sim B}\left[w_{k,t}(X)g_{k,t}(X)\mathop{\mathbf{E}}_{v_{k,t}}[\mathbb{1}(X\text{ empty})]\right]\end{aligned}$$

$$\leq \mathop{\mathbf{E}}_{v_{k,t}}\left[\left(88\gamma\|v_{k,t}\|_2^2\|\mathbf{\Sigma}\|_{\mathrm{op}}+0.1\gamma\,\|\mathbf{M}_{k,t}\|_{\mathrm{F}}^2\,\|\mathbf{\Sigma}\|_{\mathrm{op}}\right)\right]+\epsilon\mathop{\mathbf{E}}_{X\sim B}[w_{k,t}(X)g_{k,t}(X)(0.6)]$$
$$=88.1\gamma\|\mathbf{M}_{k,t}\|_{\mathrm{F}}^2\|\mathbf{\Sigma}\|_{\mathrm{op}}+0.6\epsilon\mathop{\mathbf{E}}_{X\sim B}[w_{k,t}(X)g_{k,t}(X)]\,, \tag{19}$$

其中，第二步对第二项使用了 $w_{k,\,t+1}\leq w_{k,\,t}$，第三步使用了方程（18）以及点为空的概率至多为 0.6，最后一步用到了$\mathbf{E}_{v_{k,\,t}}\left[\|v_{k,t}\|_2^2\right]=\mathbf{E}_{z_{k,\,t}\sim\mathcal{N}(0,\,\mathbf{I})}[z_{k,t}^\top\mathbf{M}_{k,\,t}^2z_{k,\,t}]=\mathrm{tr}(\mathbf{M}_{k,\,t}^2)=\|\mathbf{M}_{k,\,t}\|_{\mathrm{F}}^2$。

现在，我们使用以下一系列不等式来给势函数的期望值一个上界，具体说明如下：[16]

$$\mathop{\mathbf{E}}_{v_{k,t+1}^t}\left[\text{翻译}\left(\mathbf{B}_k^{2p_k+1},\right)\right]\leq\mathbf{E}_{v_k}\left[\text{tr}\left(\mathbf{B}_{k,\,t}^{p_k}\mathbf{B}_{k,\,t+1}\mathbf{B}_{k,\,t}^{p_k}\right)\right] \tag{20}$$
$$=\mathop{\mathbf{E}}_{v_{k,t}}\left[\text{tr}\left(\mathbf{M}_{k,t}\mathbf{B}_{k,t+1}\mathbf{M}_{k,t}\right)\right]$$
$$=\mathbf{E}^{v_{k,t}}\left[\mathbf{E}_{X\sim P}[w_{k,t+1}(X)g_{k,t}(X)]\right] \tag{21}$$
$$=\mathop{\mathbf{E}}_{v_{k,t}}\left[(1-\epsilon)\mathop{\mathbf{E}}_{X\sim G}[w_{k,t+1}(X)g_{k,t}(X)]+\epsilon\mathop{\mathbf{E}}_{X\sim B}[w_{k,t+1}(X)g_{k,t}(X)]\right] \tag{22}$$
$$\leq(1-\epsilon)\mathop{\mathbf{E}}_{X\sim G}[w_{k,t}(X)g_{k,t}(X)]+\mathop{\mathbf{E}}_{v_{k,t}}\left[\epsilon\mathop{\mathbf{E}}_{X\sim B}[w_{k,t+1}(X)g_{k,t}(X)]\right] \tag{23}$$
$$\leq(1+2\gamma)\left\langle\mathbf{\Sigma},\mathbf{M}_{k,t}^2\right\rangle+89\gamma\|\mathbf{M}_{k,t}\|_{\mathrm{F}}^2\|\mathbf{\Sigma}\|_{\mathrm{op}}+0.6\epsilon\mathop{\mathbf{E}}_{X\sim B}[w_{k,t}(X)g_{k,t}(X)] \tag{24}$$
$$=(1+2\gamma)\left\langle\mathbf{\Sigma},\mathbf{M}_{k\,\,,t}^2\right\rangle+89\gamma\|\mathbf{M}_{k,t}\|_{\mathrm{F}}^2\|\mathbf{\Sigma}\|_{\mathrm{op}}$$
$$+0.6\left(\mathbf{E}_{X\sim P}[w_{k,t}(X)g_{k,\,t}(X)]-(1-\epsilon)\,\mathbf{E}_{X\sim G}[w_{k,t}(X)g_{k,\,t}(X)]\right) \tag{25}$$
$$=(1+2\gamma)\left\langle\mathbf{\Sigma},\mathbf{M}_{k\,\,,t}^2\right\rangle+89\gamma\|\mathbf{M}_{k,t}\|_{\mathrm{F}}^2\|\mathbf{\Sigma}\|_{\mathrm{op}}$$
$$+0.6\left(\phi_{k,t}-(1-\epsilon)\mathop{\mathbf{E}}_{X\sim G}[w_{k,t}(X)g_{k,t}(X)]\right) \tag{26}$$
$$\leq(1+2\gamma)\left\langle\mathbf{\Sigma},\mathbf{M}_{k,t}^2\right\rangle+89\gamma\|\mathbf{M}_{k,t}\|_{\mathrm{F}}^2\|\mathbf{\Sigma}\|_{\mathrm{op}}+0.6\phi_{k,t}-0.6\,(1-3\gamma)\left\langle\mathbf{\Sigma},\mathbf{M}_{k,t}^2\right\rangle \tag{27}$$
$$=0.4\langle\mathbf{\Sigma},\mathbf{M}_{k,t}^2\rangle+3.8\gamma\langle\mathbf{\Sigma},\mathbf{M}_{k,t}^2\rangle+89\gamma\|\mathbf{M}_{k,t}\|_{\mathrm{F}}^2\|\mathbf{\Sigma}\|_{\mathrm{op}}+0.6\phi_{k,t}$$
$$\leq0.4\langle\mathbf{\Sigma},\mathbf{M}_{k,t}^2\rangle+93\gamma\|\mathbf{M}_{k,t}\|_{\mathrm{F}}^2\|\mathbf{\Sigma}\|_{\mathrm{op}}+0.6\phi_{k,t}\,, \tag{28}$$

其中，公式 (20) 利用了 $\mathbf{B}_{k,t+1}\preceq\mathbf{B}_{k,t}$，以及事实 2.3 和迹算子的循环性质；公式 (21) 利用了 $\mathbf{B}_{k,t}=\mathbf{E}_{X\sim P}[w_{k,t}(X)]\mathbf{\Sigma}_{k,t}=\mathbf{E}_{X\sim P}[w_{k,t}(X)XX^\top]$ 的定义；公式 (22) 利用了 $P=(1-\epsilon)G+\epsilon B$；对于第一项，公式 (23) 利用了 $w_{k,t+1}(x)\leq w_{k,t}(x)$；对于第一项，公式 (24) 利用了 $1-\epsilon\leq1$ 和引理 2.14 的第 (i) 项，对于第二项则利用了公式 (19)；公式 (25) 利用了 $P=(1-\epsilon)G+\epsilon B$；公式 (26) 利用了 $\mathbf{B}_{k,t}$ 的定义，得到 $\mathbf{E}_{X\sim P}[w_{k,t}(X)g_{k,t}(X)]=\mathbf{E}_{X\sim P}[w_{k,t}(X)\mathrm{tr}(\mathbf{M}_{k,t}^2XX^\top)]=\mathrm{tr}(\mathbf{M}_{k,t}^2\mathbf{E}_{X\sim P}[w_{k,t}(X)XX^\top])=\mathrm{tr}(\mathbf{B}_{k\,t}^{2p_k+1},)=\phi_{k,t}$。现在我们解释最后三个步骤。从公式 (27) 开始，回想一下，我们已对事件 $(1-\epsilon)\,\mathbf{E}_{X\sim G}[1-w_{k,t}(X)]+\epsilon\mathbf{E}_{X\sim B}[w_{k,t}(X)]\leq2.9\epsilon$ 进行条件化。因此，我们使用引理 2.14 的第 (i) 项，得到 $(1-\epsilon)\,\mathbf{E}_{X\sim G}[w_{k,t}(X)g_{k,t}(X)]\geq(1-\epsilon)(1-2\gamma)\langle\mathbf{\Sigma},\mathbf{M}_{k,t}^2\rangle\geq(1-3\gamma)\langle\mathbf{\Sigma},\mathbf{M}_{k,t}^2\rangle$（其中最后一个不等式利用了 $\epsilon<\gamma/20$）。最后，公式 (28) 将项 $\gamma\langle\mathbf{\Sigma},\mathbf{M}_{k,t}^2\rangle$ 上界为 $\gamma\|\mathbf{\Sigma}\|_{\mathrm{op}}\|\mathbf{M}\|_{\mathrm{F}}^2$。现在我们将把 $\langle\mathbf{\Sigma},\mathbf{M}_{k,t}^2\rangle$ 和 $\|\mathbf{\Sigma}\|_{\mathrm{op}}\|\mathbf{M}\|_{\mathrm{F}}^2$ 都与 $\phi_{k,t}$ 联系起来。

[16]回顾一下，我们处于 $\mathbf{1}(\mathcal{E}_{k,\,t-1})=1$ 的情形，因此不再显式地在期望中写出指示函数。

我们先考虑项 $\langle \mathbf{\Sigma}, \mathbf{M}_{k,t}^2 \rangle$：由于事件 $\mathcal{E}_{k\,t}^{(2)}$，成立，我们有

$$\langle \mathbf{\Sigma}, \mathbf{M}_{k,t}^2 \rangle < (1-250\gamma)\langle \mathbf{\Sigma}_{k,t}, \mathbf{M}_{k,t}^2 \rangle = \frac{1-250\gamma}{\mathbf{E}_{X\sim P}[w_{k,t}(X)]} \langle \mathbf{B}_{k,t}, \mathbf{M}_{k,t}^2 \rangle \tag{29}$$

$$\leq \frac{1-250\gamma}{(1-\epsilon)(1-3\epsilon)} \phi_{k,t} \leq (1-240\gamma)\phi_{k,t} \,, \tag{30}$$

其中，第一个不等式使用了事件 $\mathcal{E}_{k\,t}^{(2)}$ 的定义，接下来的步骤利用了 $\mathbf{E}_{X\sim P}[w_{k,t}(X)] \geq (1-\epsilon)\,\mathbf{E}_{X\sim G}[w_{k,t}(X)] \geq (1-\epsilon)(1-3\epsilon)$，其中此处的最后一个不等式使用了事件 $\mathcal{E}_{k\,t}^{(1)}$ 的定义。

现在我们陈述 述以下将 $\phi_{k,\ t}$ 与 $\|\mathbf{\Sigma}\|_{\mathrm{op}}\|\mathbf{M}_{k,\ t}\|_{\mathrm{F}}^2$ 相关联的结果，该结果对充分大 e:

命题 3.10。 若对于充分大的常数 C， $p_k \geq C\log(d)$，且事件 $\mathcal{E}_{k,\ t}$ 成立，则

$$\mathrm{tr}\left(\mathbf{B}_{k,t}^{2p_k}\right)\|\mathbf{\Sigma}\|_{\mathrm{op}} \leq 1.01 \cdot \mathrm{tr}\left(\mathbf{B}_{k,t}^{2p_k+1}\right) \,.$$

证明。 我们将利用 Schatten p-范数之间的关系，并利用稳定性将 $\mathbf{\Sigma}$与 $\mathbf{\Sigma}_t$ 联系起来，从而证明该结果，如下所示:

$$\begin{aligned}
\mathrm{tr}\left(\mathbf{B}_{k,t}^{2p_k}\right)\|\mathbf{\Sigma}\|_{\mathrm{op}} &< (1+3\gamma)\mathrm{tr}\left(\mathbf{B}_{k,t}^{2p_k}\right)\|\mathbf{B}_{k,t}\|_{\mathrm{op}} && \text{(using stability as in Equation (50))} \\
&= (1+3\gamma)\|\mathbf{B}_{k,t}\|_{2p_k}^{2p_k}\|\mathbf{B}_{k,t}\|_{\infty} \\
&\leq (1+3\gamma)\|\mathbf{B}_{k,t}\|_{2p_k}^{2p_k+1} && (\|\mathbf{A}\|_{\infty} \leq \|\mathbf{A}\|_q \ \forall q) \\
&\leq (1+3\gamma)\left(d^{\frac{1}{2p_k(2p_k+1)}}\|\mathbf{B}_{k,t}\|_{2p_k+1}\right)^{2p_k+1} && \text{(by Fact 2.7)} \\
&= (1+3\gamma)d^{1/2p_k}\|\mathbf{B}_{k,t}\|_{2p_k+1}^{2p_k+1} \\
&\leq (1+3\gamma)\cdot 1.001 \cdot \mathrm{tr}\left(\mathbf{B}_{k,t}^{2p_k+1}\right) && (p_k \geq C\log(d) \text{ for } C \text{ large enough}) \\
&\leq 1.01 \cdot \mathrm{tr}\left(\mathbf{B}_{k,t}^{2p_k+1}\right) \,. && (\gamma < \gamma_0 \text{ for a sufficiently small } \gamma_0)
\end{aligned}$$

命题 3.10的证明至此完成。 □

现在我们可以利用公式 (28)、公式 (29)和命题 3.10对右端进行上界估计:

$$\begin{aligned}
0.4\langle \mathbf{\Sigma}, \mathbf{M}_{k,t}^2 \rangle + 93\gamma\|\mathbf{M}_{k,t}\|_{\mathrm{F}}^2\|\mathbf{\Sigma}\|_{\mathrm{op}} + 0.6\phi_{k,t} &\leq 0.4(1-240\gamma)\phi_{k,t} + 93\cdot 1.01\cdot\gamma\phi_{k,t} + 0.6\phi_{k,t} \\
&\leq (1-\gamma)\phi_{k,t} \,.
\end{aligned}$$

证明至此完成。

□

推论 3.11。 在引理 *3.9*的背景下，假设$\mathbf{E}_{\mathcal{V}_{k-1,\ t_{\mathrm{end}}}}[\phi_{k,1} \mid \mathcal{F}_{k-1,t_{\mathrm{end}}}] \leq R$。那么，对于每一个形式为 $\mathcal{F}_{k-1,\ t_{\mathrm{end}}} \cup \{F_{k,t'}: \ \ t' \leq t-1\}$ 的条件化 $\mathcal{F}_{k,\ t-1}$ （即，对于滤波器截至第$(k, t-1)$次迭代的每一个条件化，只要其在前几次迭代中截至第$(k-1, t_{\mathrm{end}})$次迭代均与 $\mathcal{F}_{k-1,\ t_{\mathrm{end}}}$ 一致），我们有:

$$\mathop{\mathbf{E}}_{\mathcal{V}_{k,t-1}}[\phi_{k,t}\mathbb{1}(\mathcal{E}_{k,t-1}) \mid \mathcal{F}_{k,t-1}] \leq (1-\gamma)^{t-1}R \,.$$

证明。 我们通过对 t 进行归纳来证明这一点。基例 $t=1$ 根据假设显然成立。现在，我们假设该断言对于索引 (k,t) 成立，并将证明它对于索引 $(k,t+1)$仍然成立。也就是说，假设 $\mathbf{E}_{v_{k,t-1}}[\phi_{k,t}\mathbb{1}(\mathcal{E}_{k,t-1}) \mid \mathcal{F}_{k,t-1}] \le (1-\gamma)^{t-1}R$。那么，根据 引理 3.9 ，我们知道

$$\mathop{\mathbf{E}}_{v_{k,t}}[\phi_{k,t+1}\mathbb{1}(\mathcal{E}_{k,t}) \mid \mathcal{F}_{k,t}, \mathcal{V}_{k,t-1}] \le (1-\gamma)\phi_{k,t}\mathbb{1}(\mathcal{E}_{k,t-1}) .$$

对两边关于 $\mathcal{V}_{k,t-1}$ 取期望，得到

$$\mathop{\mathbf{E}}_{\mathcal{V}_{k,t}}[\phi_{k,t+1}\mathbb{1}(\mathcal{E}_{k,t}) \mid \mathcal{F}_{k,t}] \le (1-\gamma)\mathop{\mathbf{E}}_{\mathcal{V}_{k,t-1}}[\phi_{k,t}\mathbb{1}(\mathcal{E}_{k,t-1}) \mid \mathcal{F}_{k,t-1}] \le (1-\gamma)^t R .$$

□

3.4.3 综合所有内容

现在，我们将前面各小节中的要素整合起来，以完成 定理 3.1 的证明。

定理 *3.1* 的证明。只需证明该定理的结论以一个较小的正常数概率成立，因为通过重复运行算法，并选择使 u 最大的输出 $u^\top \mathbf{\Sigma} u$ （更准确地说，是选择方差的估计量；该估计量可通过 引理 2.14 的 第 (iv) 项 获得），即可任意提高这一概率。为完整起见，我们在第 3.4.4 节给出运行时间的界。

首先，注意如果算法通过 SampleTopEigenvector 返回一个向量，那么所得向量以高概率满足所需的保证。我们采用与引理 3.9 陈述中相同的符号约定。我们将通过对 k 进行归纳证明：在外层循环任意一次迭代开始时（即，对每个 $k \le k_{\text{end}}$ 和 $t=1$），都有 $\mathrm{E}_{\mathcal{V}_{k-1},t_{\text{end}}}[\phi_{k,1}\mathbb{1}(\mathcal{E}^{(1)}_{k-1},t_{\text{end}})] \le (d/\epsilon)^{C_2}\log^d\|\mathbf{\Sigma}\|^{2p_k+1}\mathrm{op}$，其中 C_2 是充分大的正常数。算法开始时（$k=1$），由于第 5 行，每个点的范数为 $\|x\|_2^2 \le 10(d/\epsilon)\hat{\sigma}_{\text{op}} \le (20(d^2/\epsilon))\|\mathbf{\Sigma}\|_{\text{op}}$，势函数满足以下界

$$\begin{aligned}\phi_{1,1} &\le d\|\mathbf{\Sigma}_{k,t}\|_{\text{op}}^{2p+1} \le \left(20(d^2/\epsilon)\|\mathbf{\Sigma}\|_{\text{op}}\right)^{2p+1} \le (d/\epsilon)^{10p}\|\mathbf{\Sigma}\|_{\text{op}}^{2p+1} \\ &= (d/\epsilon)^{10C'\log d}\|\mathbf{\Sigma}\|_{\text{op}}^{2p_1+1} \le (d/\epsilon)^{C_2\log d}\|\mathbf{\Sigma}\|_{\text{op}}^{2p_1+1} .\end{aligned}$$

根据归纳假设，设所需结论对某个$k \ge 1$及其之前（包括该值）均成立。我们先建立$\phi_{k,\ t_{\text{end}}+1}$ 与 $\phi_{k+1,\ 1}$ 之间的如下关系：

$$\begin{aligned}\phi_{k+1,1} &= \mathrm{tr}\left(\mathbf{B}_{k,t_{\text{end}}+1}^{2p_{k+1}+1}\right) = \|\mathbf{B}_{k,t_{\text{end}}+1}\|_{2p_{k+1}+1}^{2p_{k+1}+1} \le \|\mathbf{B}_{k,t_{\text{end}}+1}\|_{2p_k+1}^{2p_{k+1}+1} \\ &= \mathrm{tr}\left(\mathbf{B}_{k,t_{\text{end}}+1}^{2p_k+1}\right)^{(2p_{k+1}+1)/(2p_k+1)} = \phi_{k,t_{\text{end}+1}}^{(2p_{k+1}+1)/(2p_k+1)} . \qquad (31)\end{aligned}$$

根据归纳假设，在外层循环的第 k次迭代开始时，有 $\mathrm{E}_{\mathcal{V}_{k-1},t_{\text{end}}}[\phi_{k,1}\mathbb{1}(\mathcal{E}^{(1)}_{k-1},t_{\text{end}})] \le (d/\epsilon)^{C_2}\log^d\|\mathbf{\Sigma}\|^{2p_k+1}\mathrm{op}$。在外层循环的第 k次迭代中，我们将证明，在内层循环结束时（即在 $t=t_{\text{end}}+1$），势函数会小得多。具体而言，我们将在下文证明 $\mathrm{E}_{\mathcal{V}_{k,t_{\text{end}}}}\left[\phi_{k,t_{\text{end}}+1}\mathbb{1}(\mathcal{E}^{(1)}_{k},t_{\text{end}})\right] \le (d/\epsilon)^{0.1C_2}\log^d\|\mathbf{\Sigma}\|^{2p_{k-1}+1}\mathrm{op}$。注意，在算法的第 k次迭代中，可能出现两种情况：(i) 内层循环的所有 t_{end} 次迭代中，条件$\langle \mathbf{\Sigma}_t, \mathrm{M}_{k,t}^2\rangle < (1-250\gamma)\langle \mathbf{\Sigma}_{k,t}, \mathrm{M}_{k,t}^2\rangle$ 均成立，这将使势函数减少 $(1-\gamma)^{t_{\text{end}}}$；或者该条件对某个 t不成立，而这本身就意味着

势函数很小。回想一下，$\mathcal{E}_{k\,t}^{(2)}$，对应这两种情况。因此，我们有如下分解：

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathcal{V}_{k,t_{\mathrm{end}}}}\left[\phi_{k,t_{\mathrm{end}}+1}\mathbb{1}(\mathcal{E}_{k,t_{\mathrm{end}}}^{(1)})\right] &= \mathop{\mathbf{E}}_{\mathcal{V}_{k,t_{\mathrm{end}}}}\left[\phi_{k,t_{\mathrm{end}}+1}\mathbb{1}(\mathcal{E}_{k,t_{\mathrm{end}}}^{(1)})\mathbb{1}(\mathcal{E}_{k,t_{\mathrm{end}}}^{(2)})\right] \\ &\quad + \mathop{\mathbf{E}}_{\mathcal{V}_{k,t_{\mathrm{end}}}}\left[\phi_{k,t_{\mathrm{end}}+1}\mathbb{1}(\mathcal{E}_{k,t_{\mathrm{end}}}^{(1)})\mathbb{1}\left(\overline{\mathcal{E}_{k,t_{\mathrm{end}}}^{(2)}}\right)\right] \qquad (32)\end{aligned}$$

公式（32）中的第一项可根据推论 3.11界定如下：

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathcal{V}_{k,t_{\mathrm{end}}}}\left[\phi_{k,t_{\mathrm{end}}+1}\mathbb{1}(\mathcal{E}_{k,t_{\mathrm{end}}}^{(1)})\mathbb{1}(\mathcal{E}_{k,t_{\mathrm{end}}}^{(2)})\right] &\le (1-\gamma)^{t_{\mathrm{end}}+1}(d/\epsilon)^{C_2\log d}\|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p_k+1} \qquad (33)\\ &\le e^{-0.1\cdot t_{\mathrm{end}}\cdot\gamma}(d/\epsilon)^{C_2\log d}\|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p_k+1}, \qquad (34)\end{aligned}$$

对于足够大的 C，其小于 $(d/\epsilon)^{0.05C_2\log^d}$，因为 $t_{\mathrm{end}} := C\log^2(d/\epsilon)/\gamma$（注意，我们选取的 C 远大于 C_2）。

现在我们关注公式（32）中的第二项，并证明若 $\mathcal{E}_{k\,t}^{(2)}$，t_{end}不成立，则势函数将为 $\mathrm{poly}(d/\epsilon)\|\mathbf{\Sigma}\|^{2p_k+1}_{\mathrm{op}}$。更正式地说，若对于某个 $(k,\ t)$，有 $\langle\mathbf{\Sigma},\mathbf{M}_{k,\ t}^2\rangle \ge (1-250\gamma)\langle\mathbf{\Sigma}_{k,t},\mathbf{M}_{k,\ t}^2\rangle$，则存在 c 使得

$$\begin{aligned}\phi_{k,t} = \|\mathbf{B}_{k,t}\|_{2p_k+1}^{2p_k+1} = \langle\mathbf{B}_{k,t},\mathbf{M}_{k,t}^2\rangle &\le \langle\mathbf{\Sigma}_{k,t},\mathbf{M}_{k,t}^2\rangle \le \frac{1}{1-250\gamma}\langle\mathbf{\Sigma},\mathbf{M}_{k,t}^2\rangle \\ &\le e^{c\gamma}\|\mathbf{\Sigma}\|_{\mathrm{op}}\|\mathbf{B}_{k,t}\|_{2p_k}^{2p_k} \qquad (\text{using } \gamma \ll 1)\\ &\le e^{c\gamma}\|\mathbf{\Sigma}\|_{\mathrm{op}}\|\mathbf{B}_{k,t}\|_{2p_k+1}^{2p_k} d^{\frac{1}{2p_k+1}}. \qquad (\text{using Fact 2.7})\end{aligned}$$

重新整理上述不等式可得 $\|\mathbf{B}_{k,t}\|_{2p_k+1}^{2p_k+1} \le e^{c\gamma p_k}d\|\mathbf{\Sigma}\|^{2p_k+1}_{\mathrm{op}} \le (d/\gamma)^{O(1)}\|\mathbf{\Sigma}\|^{2p_k+1}_{\mathrm{op}}$，其中最后一个不等式用到了对所有 k 都有 $p_k = O(\log(d/\gamma)/\gamma)$。结合公式（32）和公式（34），可得 $\mathbf{E}_{\mathcal{V}_{k,t_{\mathrm{end}}}}\left[\phi_{k,t_{\mathrm{end}}+1}\mathbb{1}(\mathcal{E}_{k}^{(1)},_{t_{\mathrm{end}}})\right] \le (d/\epsilon)^{0.1C_2\log^d}\|\mathbf{\Sigma}\|^{2p_k+1}_{\mathrm{op}}$。将其与公式（31）以及观察结果 $(2p_{k+1}+1)/(2p_k+1) \le 2$ 结合起来，根据 p_k 的定义（算法 4 第 8 行），我们得出

$$\mathop{\mathbf{E}}_{\mathcal{V}_{k,t_{\mathrm{end}}}}\left[\phi_{k+1,1}\mathbb{1}(\mathcal{E}_{k,t_{\mathrm{end}}}^{(1)})\right] \le \|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p_{k+1}+1}(d/\epsilon)^{C_2\log d}.$$

到目前为止，我们已经证明 $\mathbf{E}_{\mathcal{V}_{k-1,\ t_{\mathrm{end}}}}[\phi_{k,1}\mathbb{1}(\mathcal{E}_{k-1,\ t_{\mathrm{end}}}^{(1)})] \le \|\mathbf{\Sigma}\|^{2p_k+1}_{\mathrm{op}}(d/\epsilon)^{C_2\log^d}$ 对于 $k=1,\ \ldots\ldots,\ k_{\mathrm{end}}$。接下来需要分析最后一次迭代 $k=k_{\mathrm{end}}$。为此，我们再次利用 推论 3.11 得到如下结果：

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathcal{V}_{k_{\mathrm{end}},t_{\mathrm{end}}}}[\phi_{k_{\mathrm{end}},t_{\mathrm{end}}+1}\mathbb{1}(\mathcal{E}_{k_{\mathrm{end}},t_{\mathrm{end}}})] &\le (1-\gamma)^{t_{\mathrm{end}}+1}(d/\epsilon)^{C_2\log d}\|\mathbf{\Sigma}\|_{\mathrm{op}}^{1+2p_{k_{\mathrm{end}}}} \\ &\le (d/\epsilon)^{-100C'}\|\mathbf{\Sigma}\|_{\mathrm{op}}^{1+2p_{k_{\mathrm{end}}}}, \qquad (35)\end{aligned}$$

其中最后一步利用了 $t_{\mathrm{end}} := C\log^2(d/\epsilon)/\gamma$ 对充分大的 C 成立，并且注意到常数 C 相对于常数 C'（参见第 2 行）。

回顾 $\mathcal{E}_{k\ \text{事件的定义},\ t}$ 该定义在引理 3.9 的陈述中给出，是以下两个事件的交集：

1. $\mathcal{E}_{k\,t}^{(1)}$，：对于每次迭代 $(1-\epsilon\mathbf{E}_{X\sim G}[1-w_{k',t'}(X)]+\epsilon\,\mathbf{E}_{X\sim B}[w_{k',t'}(X)]\le 3\epsilon)$，包括 $(k',\ t')$ 和之前的迭代，均有 $(k,\ t)$。

2. $\mathcal{E}^{(2)}_{k\ t}$, : $\langle \mathbf{\Sigma}, \mathbf{M}^2_{k',\ t'} \rangle <$ $(1-250\gamma)$ $\langle \mathbf{\Sigma}_{k',t'}, \mathbf{M}^2_{k',\ t'} \rangle$ ，对于从（k'， t'）到（k，1）（包括两者）之间的每次迭代，均有（k， t）。

根据 引理 A.9 可得 $\mathbf{Pr}[\mathcal{E}^{(1)}_{k_{\text{end}},t_{\text{end}}}] > 0.2$。因此

$$\mathop{\mathbf{E}}_{\mathcal{V}_{k,t_{\text{end}}}}\left[\phi_{k_{\text{end}},t_{\text{end}}+1}\mathbb{1}(\mathcal{E}^{(2)}_{k_{\text{end}},t_{\text{end}}}) \mid \mathcal{E}^{(1)}_{k_{\text{end}},t_{\text{end}}}\right] = \frac{\mathbf{E}_{\mathcal{V}_{k,t_{\text{end}}}}\left[\phi_{k_{\text{end}},t_{\text{end}}+1}\mathbb{1}(\mathcal{E}_{k_{\text{end}},t_{\text{end}}})\right]}{\mathbf{Pr}[\mathcal{E}^{(1)}_{k_{\text{end}},t_{\text{end}}}]} \le 5(d/\epsilon)^{-100C'}\|\mathbf{\Sigma}\|_{\text{op}}^{1+2p_{k_{\text{end}}}}$$

因此，我们有

$$\begin{aligned}
&\mathbf{Pr}\left[\phi_{k_{\text{end}},t_{\text{end}}+1}\mathbb{1}(\mathcal{E}^{(2)}_{k_{\text{end}},t_{\text{end}}}) > 500(d/\epsilon)^{-100C'}\|\mathbf{\Sigma}\|_{\text{op}}^{1+2p_{k_{\text{end}}}} \text{ or } \mathbb{1}(\mathcal{E}^{(1)}_{k_{\text{end}},t_{\text{end}}}) = 0\right] \\
&\le \mathbf{Pr}[\mathbb{1}(\mathcal{E}^{(1)}_{k_{\text{end}},t_{\text{end}}}) = 0] + \mathbf{Pr}\left[\phi_{k_{\text{end}},t_{\text{end}}+1}\mathbb{1}(\mathcal{E}^{(2)}_{k_{\text{end}},t_{\text{end}}}) > 500(d/\epsilon)^{-100C'}\|\mathbf{\Sigma}\|_{\text{op}}^{1+2p_{k_{\text{end}}}} \mid \mathcal{E}^{(1)}_{k_{\text{end}},t_{\text{end}}}\right] \\
&\le 0.2 + 1/100\ , \qquad (36)
\end{aligned}$$

其中最后一步使用了马尔可夫不等式。

令 t^* 为算法 4 的第 k_{end} 次外层循环迭代期间满足$\langle \mathbf{\Sigma}, \mathbf{M}^2_{k_{\text{end}},t} \rangle \ge (1-250\gamma)\langle \mathbf{\Sigma}_{k_{\text{end}},t}, \mathbf{M}^2_{k_{\text{end}},t} \rangle$ 的首次时刻（或等价地，满足 $\mathbf{1}(\mathcal{E}^{(2)}_{k_{\text{end}},t}) = 0$）。下面我们论证 $t^* \le t_{\text{end}}$。我们通过反证法证明这一点。假设 $\mathbf{1}(\mathcal{E}^{(2)}_{k_{\text{end}},t_{\text{end}}}) = 1$。

在 $\phi_{k_{\text{end}},t_{\text{end}}+1}\mathbb{1}(\mathcal{E}^{(2)}_{k_{\text{end}},t_{\text{end}}}) < 500(d/\epsilon)^{-100C'}$ 和 $\mathbb{1}(\mathcal{E}^{(1)}_{k_{\text{end}},t_{\text{end}}}) = 1$ 同时发生的事件中（根据 公式 (36)，这两个事件同时成立的概率至少为 0。79），我们有

$$\begin{aligned}
\phi_{k_{\text{end}},t_{\text{end}}+1} &= \phi_{k_{\text{end}},t_{\text{end}}+1}\mathbb{1}(\mathcal{E}^{(2)}_{k_{\text{end}},t_{\text{end}}}) \\
&< 500(d/\epsilon)^{-100C'}\|\mathbf{\Sigma}\|_{\text{op}}^{2p_{k_{\text{end}}}+1} \\
&< 0.5(1-2\gamma)^{2p_{k_{\text{end}}}}\|\mathbf{\Sigma}\|_{\text{op}}^{2p_{k_{\text{end}}}+1} \\
&\le \mathop{\mathbf{E}}_{X\sim P}[w_{k,t}(X)](1-2\gamma)^{2p_{k_{\text{end}}}}\frac{1}{1-250\gamma}\|\mathbf{\Sigma}\|_{\text{op}}^{2p_{k_{\text{end}}}+1}\ , \qquad (37)
\end{aligned}$$

其中，第一行使用了假设 $\mathbf{1}(\mathcal{E}^{(2)}_{k_{\text{end}},\ t_{\text{end}}}) = 1$，第三行使用了 $p_{k_{\text{end}}} = C'\log(d/\gamma)/\gamma$，第四行则使用了 $\mathbf{E}_{X\sim P}[w_{k,t}(X)] \ge 1/2$，因为事件 $\mathcal{E}^{(1)}_{k_{\text{end}},\ t_{\text{end}}}$ 成立。

引理 3.8 和 公式 (37) 意味着 $\langle \mathbf{\Sigma}, \mathbf{M}^2_{k_{\text{end}},\ t} \rangle \ge (1-250\gamma)\langle \mathbf{\Sigma}_{k_{\text{end}},t}, \mathbf{M}^2_{k_{\text{end}},\ t} \rangle$，由此得出矛盾。因此，$t^* \le t_{\text{end}}$。

为完成证明，还需证明：如果算法此前尚未以一个良好的解终止，那么在第（k_{end}， t^*）次迭代期间，算法将返回 $\mathbf{\Sigma}$ 的最大特征值的良好近似并终止。为此，我们使用引理 3.3（该引理适用，是因为我们对事件 $\mathcal{E}_{k_{\text{end}},\ t_{\text{end}}}$ 进行了条件化，而如前所见，该事件以常数概率发生，因此其要求 $\mathrm{E}_{X\sim G}[w_{k,t}(X)] \ge 1-3\epsilon$ 得以满足）。该引理的结论意味着：当 $t = t^*$ 时，SampleTopEigenvector 的输出满足 $u^\top \mathbf{\Sigma} u > (1-O(\gamma))\ u^\top \mathbf{\Sigma}_{k,\ t} u$ 且 $\frac{u^\top \mathbf{\Sigma}_{k,\ t} u}{\|u\|_2^2} \ge (1-\gamma)\|\mathbf{\Sigma}_{k,\ t}\|_{\text{op}}$ 的概率为 0.9。对此事件进行条件化，并注意到第 3 行的估计量 $\widehat{r}_t$ 以高概率在所有迭代中满足 $\widehat{r}_t \ge (1-\gamma)\|\mathbf{\Sigma}\|_{\text{op}}$，因此第 7 行的检查将被触发；根据引理 A.7，算法将返回一个向量 u，使得 $u^\top \mathbf{\Sigma} u/\|u\|_2^2 \ge (1-O(\gamma))\|\mathbf{\Sigma}\|_{\text{op}}$。

3.4.4 运行时间分析

设输入分布 P 为 $\mathbb{R}^d$ 中 n 个点上的均匀分布。外层循环重复 k_{end} 次，内层循环重复 t_{end} 次，其中 $k_{\mathrm{end}} = O\left(\log\left(1/\gamma\right)\right)$，且 $t_{\mathrm{end}} = O\left(\log^2\left(d/\epsilon\right)/\gamma\right)$。

在每次循环中，运行时间由以下部分决定：计算 $v_{k,t}$ in 13 可通过从 $z_{k,t}$ 开始并反复将其与 $\mathrm{B}_{k,t}$ 相乘，在 $O(ndp_k)$ 时间内完成（将向量 z 与二阶矩矩阵 $\sum_x xx^\top$ 相乘可在 $O(nd)$ 时间内完成：先计算括号内的内积 $\sum_x x(x^\top z)$，再计算所得向量的平均值）。算法 2 第 3 行的幂迭代估计器运行时间为 $O(\frac{nd}{\gamma}\log(d\cdot k_{\mathrm{end}}\cdot t_{\mathrm{end}}/\gamma))$（且最多调用 $k_{\mathrm{end}}\cdot t_{\mathrm{end}}$ 次）。接下来分析 HardThresholdingFilter（算法 8）的运行时间：未被阈值处理置零的分数 $\tau_{k,t}(x)$ 满足以下上界和下界：$\mathrm{poly}(\epsilon/d)\|\mathbf{\Sigma}\|_{\mathrm{op}}\|v_{k,t}\|_2^2 \le 0.1(\widehat{\sigma}_{\mathrm{op}}/d)\|v_{k,t}\|_2^2 \le \tau_{k,t}(x) \le \|x\|_2^2\|v_{k,t}\|_2^2 \le 2\widehat{\sigma}_{\mathrm{op}}(d^4/\epsilon)\|v_{k,t}\|_2^2 \le \mathrm{poly}(d/\epsilon)\|\mathbf{\Sigma}\|_{\mathrm{op}}\|v_{k,t}\|_2^2$，其中用到了第 5 行的剪枝、第 16 行的定义以及第 6 行的朴素估计器。由于 HardThresholdingFilter 在期望意义下会将 $\tau(x)$ 的最大值减半，因此它在终止前执行的过滤步骤数的期望为 $O(\log(d/\epsilon))$；据此，根据马尔可夫不等式，所有 $k_{\mathrm{end}}\cdot t_{\mathrm{end}}$ 次 HardThresholdingFilter 执行对运行时间的贡献以较高的常数概率为 $O(nd\cdot k_{\mathrm{end}}t_{\mathrm{end}}\log(d/\epsilon))$。

□

因此，总运行时间为

$$
\begin{aligned}
T &= O\left(\frac{nd}{\gamma}k_{\mathrm{end}}t_{\mathrm{end}}\log(d\cdot k_{\mathrm{end}}\cdot t_{\mathrm{end}}/\gamma)\right) + O(nd\cdot k_{\mathrm{end}}t_{\mathrm{end}}\log(d/\epsilon)) + O\left(nd\cdot t_{\mathrm{end}}\sum_{k=1}^{k_{\mathrm{end}}}p_k\right)\\
&= O\left(\frac{nd}{\gamma^2}\log(1/\gamma)\log^2(d/\epsilon)\log\left(\frac{d}{\gamma}\log(d/\epsilon)\right) + \frac{nd}{\gamma}\log(1/\gamma)\log^3(d/\epsilon) + \frac{nd}{\gamma^2}\log^2(d/\epsilon)\log(d/\gamma)\right)\\
&= O\left(\frac{nd}{\gamma^2}\log^4(d/\epsilon)\right)\ .
\end{aligned}
$$

4 鲁棒主成分分析的流式算法

组织结构 本节介绍我们近线性时间鲁棒主成分分析算法的流式版本。我们在 定理 4.1 中给出该算法的表述，并讨论它与前述情形的差异，以及证明技术所需的调整。这带来了一组额外的确定性条件，列于条件 4.2中，并在 第 4.1 节和 附录 B中予以证明。

我们采用标准的单遍流式模型。在该模型中，算法不是处理一个固定的数据集，而是以在线方式从分布中抽取样本。设 G 是一个 $(20\epsilon,\gamma)$-稳定分布，它将作为内点的底层分布。设“污染”分布为 P，并假定它是 ϵ-污染的 G 全变差版本（参见 定义 1.4）。注意，对 ϵ前面的常数作少许调整后，我们也可以等价地使用混合表示 $P = (1-\epsilon)G + \epsilon B$（参见 定义 2.12之后的讨论）。与前一节不同，前一节中的输入数据集基本上可以存放在“免费”的内存中；而现在的数据访问模型是：从 n 中独立同分布地抽取样本，并逐个提供给算法（定义 1.3）。

该算法仍然可以维护本地内存，但存储在本地内存中的任何内容都会计入其空间复杂度。目标仍然是找到一个单位向量 u，使得 $u^\top\mathbf{\Sigma}u \ge (1-O(\gamma))\|\mathbf{\Sigma}\|_{\mathrm{op}}$。

定理 *4.1.* 设整数 $d > 2$ 和实数 $0 < 20\epsilon < \gamma < \gamma_0$，其中 γ_0 足够小。设 G 是关于半正定矩阵 $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ 的 $(20\epsilon, \gamma)$- 稳定分布（定义 *2.12* ），且 $r \geq 1$ 是满足 $Pr_{X\sim G}[\|X\|_2 > r\sqrt{d}\|\mathbf{\Sigma}\|_{\mathrm{op}}] \leq \epsilon$ 的半径。设 P 是满足 $d_{\mathrm{TV}}(P, G) \leq \epsilon$ 的分布。存在一种算法，以 ϵ, γ, r 作为输入，并使用一个数据流.....

$$n \lesssim \left(\frac{r^2d^2}{\gamma^5} + \frac{1}{\epsilon\gamma}\right) \mathrm{polylog}(d/\epsilon) .$$

来自 P（参见 定义 *1.3*），需要额外使用内存来存储 O（（$d/\gamma + 1/\epsilon$）polylog(d/ϵ））个实数（或在字 *RAM* 模型下，位复杂度为（d/γ^2）polylog(d/ϵ）），运行时间为O（$\frac{nd}{\gamma^2}$ polylog(d/ϵ），且以至少 0.99 的概率输出一个向量 u，使得 $u^\top\mathbf{\Sigma}u \geq$（$1 - O$（$\gamma$）） $\|\mathbf{\Sigma}\|_{\mathrm{op}}$。

实现上述定理的算法概述见算法 6 和 7。上述定理针对次高斯分布的特例中，若 $\gamma = O(\epsilon$ log$(1/\epsilon))$ 且 $r = O(\sqrt{\log(1/\epsilon)})$，则样本复杂度为 $O((d^2/\epsilon^5)$polylog$(d/\epsilon))$。此外，定理 4.1 中所述的内存使用量是按算法内存中需要存储的实数个数计算的。关于如何以有界精度（即在标准字 RAM 模型下）实现我们的算法，请参见第 4.2.1 节：使用 (d/γ)polylog(d/ϵ) 个寄存器，每个寄存器的大小为 $(1/\gamma)$polylog(d/ϵ) 位（总位复杂度为 (d/γ^2)polylog(d/ϵ)）。

流式设定中的差异 受 [DKPP22]，中技术的启发，我们发现上一节的算法在一定程度上已经适用于流式设定：所用的过滤器形式为 $1(v^\top x > L)$，其紧凑表示只需占用 $O(d)$ 空间（只需存储向量 v 和阈值 L）。基于势函数的分析还表明，我们至多会创建 O(polylog$(d/\epsilon)/\gamma)$ 个这样的过滤器。剩下需要调整的是，应对没有固定数据集可供遍历这一事实。新算法见算法 6。关于符号约定，量 P_k、$_t$、$\mathbf{\Sigma}_k$、$_t$、B_k、$_t$、M_k、$_t$，以及评分函数 g_k、$_t(x) = \|\mathrm{M}_k$、$_t x\|_2^2$、f_k、$_t(x) = (v_k^\top$、$_t x)^2$、τ_k、$_t(x) = f_k$、$_t(x)1(f_k$、$_t(x) > L_k$、$_t)$ 都是总体层面的量，即算法无法获知它们。不过，算法可以通过抽取样本并构造估计量来近似这些量（估计量的具体形式需要能够在流式设定中轻松计算，后文将对此进行讨论）。我们用“帽子”$\widehat{\ }$ $\widehat{\ }$ $\widehat{\ }$表示相应的基于样本的近似，即 $\mathbf{\Sigma}_k$、$_t$、B_k、$_t$、M_k、$_t$，以及由此得到的评分 $\widehat{g}_k$、$_t(x) = \|\widehat{\mathrm{M}}_k$、$_t x\|_2^2$、$\widehat{f}_k$、$_t(x) = (\widehat{v}_k^\top$、$_t x)^2$、$\widehat{\tau}_k$、$_t(x) = \widehat{f}_k$、$_t(x)1(\widehat{f}_k$、$_t(x) > \widehat{L}_k$、$_t)$。我们采用的方法是：如果基于样本的量与其对应的总体层面量足够接近，那么上一节的正确性证明（其中涉及总体层面的量）仍然适用。只需检查上一节的证明即可确认，所有界中都有足够的余量，允许在总体层面的量与其基于样本的近似之间进行转换。本节不再重复所有正确性证明，因为唯一的变化只涉及某些不等式中使用的常数。相反，我们主要着重推导获得足够精确近似所需的样本复杂度。

基于样本的估计量 上一节算法的许多步骤都涉及将向量 x 与$\mathbf{M}_{k,\ t} := \mathbf{B}_k^{p_k}$、$_t$相乘；例如，评分 $g_{k,\ t}(x)$涉及$\mathbf{M}_{k,\ t}x$的范数。具体做法是从 x 开始，反复与$\mathbf{B}_{k,\ t}$相乘。新算法将不再使用$\mathbf{B}_{k,\ t}$ 而是$\widehat{\ }$使用经验矩矩阵$\mathbf{B}_{k,\ t}$。正如上一节所述，将经验二阶矩矩阵与向量相乘，可以在遍历样本一次的过程中，以线性时间和 $O(d)$ 的内存使用量完成。然而，由于算法无法存储

$\widehat{\quad}$ $\widehat{\quad}$该矩阵，因此每次都会反复与 x 新估计量 $\mathbf{B}_{k,\,t,\,1}$，……，$\mathbf{B}_{k,\,t,\,p_k}$ 相乘。因此，结果将呈现为 $\widehat{\mathbf{M}}_{k,\,t}z$ 其中 $\widehat{\mathbf{M}}_{k,\,t} = \prod_{\ell=1}^{p_k} \widehat{\mathbf{B}}_{k,\,t,\,\ell}$（更多细节另见 算法 5）。

算法 5 从小批量得到的$\mathbf{B}_{k\,,\,t}^{p}$ 估计量。

$\frown$1：从 P 中抽取一批 S_0 $\tilde{n}$ 个样本，并令估计量为 $W_{k,t} = \mathbf{E}_{X\sim\mathcal{U}(S_0)}[w_{k,t}(X)]$。2：抽取 p 批次 $S_1, \ldots, S_p$，每批包含 $\tilde{n}$ 个样本，分别从 $P_{k,t}$ 中抽取。3：对于 $\ell \in [p]$，执行 4：令 $\widehat{\mathbf{\Sigma}}_{k,t,\ell} = \frac{1}{\tilde{n}} \sum_{x\in S_\ell} xx^T$。5：令 $\widehat{\mathrm{B}}_{k,t,\ell} = \widehat{W}_{k,t}^2 \widehat{\mathbf{\Sigma}}_{k,t,\ell}$。6：结束循环 7：返回 $\widehat{\mathbf{M}}_{k,t,\ell} = \prod_{\ell=1}^{p} \widehat{\mathrm{B}}_{k,t,\ell}$。

此外，我们还需要两种新的内存高效估计量，以便在流式模型中运行。首先，我们需要一个用于估计底层分布分位数的估计量，以替代旧算法中的第 15 行。这些分位数将通过经验分位数计算。其次，我们还需要一种内存高效的方法，在 HardThresholdingFilter(算法 8) 内部评估停止条件，因为停止条件 $\mathbf{E}_{X\sim P}[w(x)\tau(x)]$, 是一个总体级数量；为了调整我们旧算法中的第 18行，也需要一个类似的估计量。至此，我们就完成了简短的非正式概述。

形式上，我们汇总了关于 条件 4.2 中估计量所需的所有保证。补充一个小的技术说明，条件 4.2 中的条件只需在以下情况成立时满足$\mathbf{E}_{X\sim P}[w_{k,t}(X)] \geq 1 - 3\epsilon$ 以及 $\|\mathbf{B}_{k,t}\|_{\mathrm{op}} \geq 0.5\|\mathbf{\Sigma}\|_{\mathrm{op}}$。由于我们会滤除大部分异常值，因此可以假设第一个条件在算法运行过程中始终成立。第二个条件成立，是因为我们还知道，若 $\|\mathbf{B}_{k,t}\|_{\mathrm{op}} \geq 0.5\|\mathbf{\Sigma}\|_{\mathrm{op}}$ 被违反，则势函数足够小，因此算法必定已经终止（参见 引理 3.8）。该条件以高概率在算法运行过程中成立，我们将在 第 4.1 节 中证明这一点。

条件 4.2（算法 6的条件）。在算法 6的情境下，假设以下条件成立。对于每个 $k \in [k_{\mathrm{end}}]$ 和 $t \in [t_{\mathrm{end}}]$，如果 $\mathbf{E}_{X\sim P}[w_{k,t}(X)] \geq 1 - 3\epsilon$ 和 $\|\mathbf{B}_{k,\,t}\|_{\mathrm{op}} \geq 0.5\|\mathbf{\Sigma}\|_{\mathrm{op}}$，那么：

1. $\widehat{g}_{k,t}(x) \geq 0.5 g_{k,t}(x) - 0.01(\gamma/\epsilon)\|\mathbf{M}_{k,t}\|_{\mathrm{F}}^2\|\mathbf{\Sigma}\|_{\mathrm{op}}$. 2. $\left|\|\widehat{\mathbf{M}}_{k,t}\|_{\mathrm{F}}^2 - \|\mathbf{M}_{k,t}\|_{\mathrm{F}}^2\right| \leq 0.01\|\mathbf{M}_{k,t}\|_{\mathrm{F}}^2$. 3. 第 16 行的估计量 $L_{k,t}$ 满足 $\Pr_{X\sim P_{k,t}}[\widehat{g}_{k,t}(X) > L_{k,t}] \in (3.99\epsilon, \{\mathrm{v28})$。

4. 回顾参数 r，将其定义为满足 $\mathbf{Pr}_{X\sim G}[\|X\|_2 > r\sqrt{d}\|\mathbf{\Sigma}\|_{\mathrm{op}}] \leq \epsilon$ 的半径。对于任意权重函数 $w\colon \mathbb{R}^d \to [0,1]$，算法都可以使用一个估计量 $\widehat{F}$ 来估计量 $F_{k,\,t} := \mathbf{E}_{X\sim P}[w(X)\, \widehat{f}_{k,\,t}(X)]$，其准确度为 $|\widehat{F} - F_{k,\,t}| \leq 0.01\gamma F_{k,\,t} + \frac{0.01\gamma}{dr^2}\|\widehat{v}_{k,\,t}\|_2^2\|\mathbf{\Sigma}_{k,\,t}\|_{\mathrm{op}}$，总共调用 $t_{\mathrm{end}}k_{\mathrm{end}}\log^5(d/\epsilon)$ 次。

5. 算法 7 第7 行中的估计量 $\widehat{r}_t$ 满足 $\widehat{r}_t \geq (1-\gamma)\|\mathbf{\Sigma}_{k,\,t}\|_{\mathrm{op}}$。

6. 每次执行第 10 行的算法 7 时，$u^\top \mathbf{\Sigma}_{k,t} u/\|u\|_2^2 \geq (1-\gamma)\|\mathbf{\Sigma}_{k,t}\|_{\mathrm{op}}$发生的概率为 0.9。

特别地，条件 4.2 的第 1 项表明，我们在证明上一节主定理时使用的假设 3.7在当前设定下对算法 6 成立。第 2 项为

算法 6 流式模型中的 RobustPCA

1：输入： P, ϵ, γ。 2：令 C, C' 为足够大的绝对常数，且其比值 C/C' 足够大。 3：令估计量 R 满足 $|\Pr_{X\sim G}[\|X\|_2 \geq R] - \epsilon| \leq 2\epsilon$（注意 $R \leq r\sqrt{d\|\mathbf{\Sigma}\|_{\mathrm{op}}}$） 4：初始化 w_1,1(x) $=$ 1($\|x\|_2 \leq R$)。 5：令 $k_{\mathrm{end}} := \log((\log(d/\gamma)/\log(d))/\gamma)$，以及 $t_{\mathrm{end}} := C\log^2(d/\epsilon)\,\gamma$。 6：求估计量 $\widehat{\sigma}_{\mathrm{op}} \in (0.8\|\mathbf{\Sigma}\|_{\mathrm{op}}, 2d\|\mathbf{\Sigma}\|_{\mathrm{op}})$。▷参见 U $=$ I 时的引理 2.14 第 (iv) 项。 7：对于 $k = 1, \ldots, k_{\mathrm{end}}$，执行 8：令 $p_k = 2^{k-1}p$，其中 $p = C'\log(d)$。▷ p_k 的取值范围从 $C'\log(d)$ 到 $C'\log(d/\gamma)/\gamma$。 9：对于 $t = 1, \ldots, t_{\mathrm{end}}$，执行 10：令 $P_{k,t}$ 为按 w_k 加权后的 P 分布： $P(x)w_{k,t}(x)/\,\mathrm{E}_{X\sim P}[w_{k,t}(X)]$。 11：令 $\mathrm{B}_{k,t} := \mathrm{E}_{X\sim P}[w_{k,t}(X)XX^\top]$，并令 $\mathrm{M}_{k,t} := \mathrm{B}_k^{p_k}, t$。 12：令 $g_{k,t}(x) := \|\mathrm{M}_{k,t}x\|_2^2$。 13：令 $\mathrm{M}_{k,t}$ 为 $\mathrm{M}_{k,t}$ 的基于样本的版本，如算法 5 中所定义。 14：$\widehat{v}_{k,t} \leftarrow \mathrm{M}_{k,t}z_{k,t}$，其中 $z_{k,t} \sim \mathcal{N}(0, \mathrm{I})$。 15：令 $\widehat{f}_{k,t}(x) = (\widehat{v}_k^\top, t^x)^2$。

$L_{k\,t}$ 3ϵ $f_{k\,t}\cdot$ $P_{k\,t}$ | $_{X\sim P_{k\,t}}[f_{k,t}\ X\ >$ 为 $\widehat{L}_{k\,t}] - 3\epsilon| \leq 0.01\epsilon$ $\widehat{L}_{k\,t} \leftarrow$ $\{\widehat{L}_{k,t}, \frac{0.1}{d}\widehat{\sigma}_{\mathrm{op}}\|v_{k,t}\|_2^2\}$

16：令 在 $\widehat{\tau}_k$ 下的 t、x 的 $= f_k$ 分位数估计量，使得 Pr t，$x(f_k)$，然后执行 t、max。 17：令 x,> $L_k(t)$ P,$w_k(t)$1($\epsilon, \gamma(1/)$ $k_{\mathrm{end}} \cdot t_{\mathrm{end}}$▷)。 18：调用 SampleTopEigenvector ($\widehat{\sigma}_{k,t}$,$|\widehat{\sigma}_{k,t} - \widehat{v}_k^\top$, $_t\mathbf{\Sigma}\widehat{v}_{k,t}| \leq 4\gamma\widehat{v}_k^\top(_t\mathbf{\Sigma}\widehat{v}_k)$)。 $_t$参见算法 7。 19：令 $\widehat{\sigma}_k$ 满足 $_t$,= ,$_{X\sim P}[w_{k,t}$,X,$f_{k,t}$,$_t$。 X 20：（例如，$f_{k,t}$:XE$\leq L_k(t)$]▷ $(\widehat{\sigma}'_k)$1($_t$,$|\widehat{\sigma}'_k(_t - \widehat{\sigma}_k)\ _t| \leq 0.01\widehat{\sigma}_{k,t} + \frac{0.01\gamma}{r^2 d}\|\widehat{v}_k\|_t\|_2^2\|\mathbf{\Sigma}_k)$ 。 $_t\|_{\mathrm{op}}$参见引理 2.14 的第 (iv) 项。 21：求估计量 T_k，使得 $_t = 2.35\gamma\widehat{\sigma}'_{k,t}$,$w_{k,t+1} \leftarrow$,P,w_k。 $_t$22：令 $\tau_{k,t}$, $\widehat{T}_k$。 23：$_t$, RHardThresholdingFilter($\frac{0.1\gamma}{r^2 d}\|\widehat{v}_{k,t}\|_2^2\|\mathbf{\Sigma}_{k,t}$,$\widehat{\mathbf{\Sigma}}_{k,t}$,$\mathbf{\Sigma}_{k,t}$▷ ,$w_{k+1,0} \leftarrow w_{k,t+1}$)，其中 , 是 , 的基于样本的版本。参见算法 8。 24：结束循环 25：置 ,,。 26：结束循环 27：返回 FAIL。

与我们先前分析中的方程 方程 (16) 相关。为使其适用于当前情形，$v_{k,\ t}$ 将替换为 $\widehat{v}_{k,\ t}$ 并且对 $\widehat{v}_{k,\ t}$ 取期望后，其范数将变为 $\|\widehat{\mathbf{M}}_{k,\ t}\|_{\mathrm{F}}^2$从而可以将其与 $\|\mathbf{M}_{k,\ t}\|_{\mathrm{F}}^2$联系起来。条目 3负责分位数估计。条目 4 是我们将用于在 HardThresholdingFilter 中评估停止条件，并修改旧算法第 18 行的估计量。条目 5 和6 是在无法访问目标矩阵 $\mathbf{\Sigma}_{k,\ t}$ 而是从一个随机高斯向量开始，并将其与新的估计量反复相乘时，对幂迭代保证所做的调整 $\mathbf{\Sigma}_{k,\ t}$。

4.1 建立条件 4.2

在本节中，我们重点说明条件 4.2 在样本数足够多时以高概率成立。关于条目 1、2 和4 的证明与 [DKPP22] 非常相似，因此放在附录 B中。条目 3 是 Chernoff-Hoeffding 界的一个基本应用，也在附录中给出。因此，本节重点证明条目 5 和6，它们都与幂

算法 7 SampleTopEigenvector 2

1：输入：分布 P，**权重** w，**参数** ϵ, γ, δ。 **2：令** $\mathbf{M} := (\mathbf{E}_{X\sim P}[w(X)XX^\top])^p$，**其中** $p = C\log(d/\gamma)/\gamma$。 **3：** $\widehat{r} \leftarrow 0$。 **4：对于** ~~$j \in [\log(1/\delta)]$~~，**执行 5：令** $\widehat{\mathbf{M}}$ **为一个基于样本的估计量，用算法 5 计算得到** $\mathbf{\Sigma}^p_{P_w}$。 **6：** $y \leftarrow \mathbf{M}' \cdot g$ **对于** $g \sim \mathcal{N}(0, \mathbf{I})$。 **7：** $\widehat{r} \leftarrow \max(\widehat{r} \leftarrow, y^\top \widehat{\mathbf{\Sigma}}_{P_w} y/\|y\|_2^2)$。 **8：结束循环**

▷ 参见 引理 4.3

9: 令 $\widehat{\mathbf{M}}$ 为使用算法 5 计算得到的 $\mathbf{\Sigma}^p_{P_w}$ 的基于样本的估计量。10: $u \leftarrow \mathbf{M}z$ 对于 $z \sim \mathcal{N}(0, \mathbf{I})$。 11: 令 $\widehat{\sigma}_u$ 为任意满足 $|\widehat{\sigma}_u - u^\top \mathbf{\Sigma} u| \le 4\gamma u^\top \mathbf{\Sigma} u$ 的近似，例如 $\widehat{\sigma}_u := \mathrm{E}_{X\sim P}[w(X)(u^\top X)^2 \mathbb{1}((u^\top X)^2 \le Q)]$, ，其中 Q 与 $(u^\top X)^2$ 在 P_w 下的 3ϵ-分位数之间的距离不超过 0.01ϵ。12: 求满足 $|\widehat{\sigma}'_u - \widehat{\sigma}_u| \le 0.01\gamma\widehat{\sigma}_u + 0.01\gamma\|u\|_2^2\|\mathbf{\Sigma}\|_{\mathrm{op}}$ 的估计量 $\widehat{\sigma}'_u$。13: 令 $\mathbf{\Sigma}_{P_w}$ 为 $\mathbf{\Sigma}_{P_w}$ 的基于样本的版本。 14: 如果 $\widehat{\sigma}'_u \ge (1 - C\gamma)u^\top \widehat{\mathbf{\Sigma}}_{P_w} u$ 且 $u^\top \widehat{\mathbf{\Sigma}}_{k, \frac{t}{\|u\|_2^2}} u \ge (1 - \gamma)\widehat{r}_t$，则 15: 返回 $u/\|u\|_2$。

▷ 参见 引理 2.15

16: **结束 if**

迭代方法，如下文所述。

对于标准幂迭代方法，我们知道，对于一个 $d \times d$ 半正定矩阵 A，如果 $y = \mathrm{A}^p z$ 对于较大的 p 成立，且 $z \sim \mathcal{N}(0, \mathrm{I})$，那么以高概率，$y^\top \mathrm{A} y/\|y\|_2^2$ 接近 $\|\mathrm{A}\|_{\mathrm{op}}$。形式化地，我们有 $y^\top \mathrm{A} y/\|y\|_2^2 \ge (1 - \gamma)\|\mathrm{A}\|_{\mathrm{op}}$，其中 $p = C\log(d/\gamma)/\gamma$；参见事实 2.11。然而，在我们的分析中，我们只能计算 $y = \mathrm{A}z$，其中 $\|\mathrm{A} - \mathrm{A}^p\|_{\mathrm{op}} \le \delta\|\mathrm{A}\|_{\mathrm{op}}$，且 δ 较小；这里，δ 表示采样误差。尽管只能访问 $\widehat{\mathrm{A}}$，我们仍希望 $y^\top \mathrm{A} y/\|y\|_2^2$ 至少为 $(1 - O(\gamma))\|\mathrm{A}\|_{\mathrm{op}}$。通过幂迭代，这对 $\delta = 0$ 显然成立；我们想了解 δ 可以有多大而使其仍然成立，因为较大的 δ 会带来较小的样本复杂度。我们在下文的引理 4.3 中陈述结果。

有了引理 4.3 后，条目 5 和条目 6 将按如下方式得出。对于条目 6，我们使用 $\widehat{\mathbf{M}}$ 下面的引理，其中 $\mathrm{A} = \mathrm{B}_{k,t}$，且 $\mathrm{A} = \mathrm{M}_{k,t}$；根据引理 B.1，可保证 $\|\mathrm{A} - \mathrm{A}^p\|_{\mathrm{op}} \le \delta\|\mathrm{A}^p\|_{\mathrm{op}} \le \delta\|\mathrm{A}^p\|_{\mathrm{F}}$。条目 5 也随之得出，因为我们可以通过重复该过程并返回使 $y\,\mathrm{A} y^\top$ 最大化的向量 $y/\|y\|_2^2$，将成功概率从 0.9 提升到任意接近 1。现在陈述并证明下面的引理：

$\widehat{\mathbf{M}}$ **引理 4.3。** *设 $\delta > 0$, $\gamma \in (0, 1/2)$, $p \in \mathbb{N}$ 且 $d \times d$ A、A 为半正定矩阵，满足 $\|A - A^p\|_{\mathrm{op}} \le \delta\|A^p\|_{\mathrm{F}}$。设 $z \sim \mathcal{N}(0, I)$，且 $y = Az$。如果 $p > C\log(d/\gamma)/\gamma$ 足够大的常数，且 $\delta < c\gamma/\sqrt{d}$ 为足够小的正常数，那么以至少 0.9 的概率，有 $y^\top A \frac{y}{\|y\|_2^2} \ge (1 - O(\gamma))\|A\|_{\mathrm{op}}$.*

$\widehat{\mathbf{M}}$ **证明。** 记 $\mathbf{\Delta} := \mathbf{A} - \mathbf{A}^p$，其算子范数至多为 $\delta\|\mathbf{A}^p\|_{\mathrm{F}}$。我们首先分析随机变量 $U = z^\top (\mathbf{A}^p\mathbf{\Delta} + \mathbf{\Delta}\mathbf{A}^p) z$。然后

$$\mathbf{E}[U] = \mathrm{tr}\,(\mathbf{A}^p\mathbf{\Delta} + \mathbf{\Delta}\mathbf{A}^p) = 2\langle \mathbf{A}^p, \mathbf{\Delta}\rangle \le 2\|\mathbf{A}^p\|_{\mathrm{F}}\|\mathbf{\Delta}\|_{\mathrm{F}} \le 2\sqrt{d}\|\mathbf{A}^p\|_{\mathrm{F}}\|\mathbf{\Delta}\|_{\mathrm{op}} \le 2\delta\sqrt{d}\|\mathbf{A}^p\|_{\mathrm{F}}^2,$$

因为 $\|\mathbf{\Delta}\|_{\mathrm{op}} \le \delta\|\mathbf{A}\|_{\mathrm{F}}$。由于 $\mathbf{A}^p\mathbf{\Delta} + \mathbf{\Delta}\mathbf{A}^p$ 是对称的，根据 事实 2.9，U 的方差满足 $\mathrm{Var}(U) \lesssim \|\mathbf{A}^p\mathbf{\Delta} + \mathbf{\Delta}\mathbf{A}^p\|_{\mathrm{F}}^2 \lesssim \|\mathbf{\Delta}\|_{\mathrm{op}}^2\|\mathbf{A}^p\|_{\mathrm{F}}^2 \le \delta^2\|\mathbf{A}^p\|_{\mathrm{F}}^4$。通过应用切比雪夫不等式，我们

有至少 0.999 的概率

$$|U| = |z^\top \left(\mathbf{A}^p \mathbf{\Delta} + \mathbf{\Delta} \mathbf{A}^p\right) z| \lesssim \delta \sqrt{d} \|\mathbf{A}^p\|_{\mathrm{F}}^2. \tag{38}$$

通过类似的论证，我们得到该结论成立的概率至少为 0.999，

$$|z^\top \left(\mathbf{A}^{p+1} \mathbf{\Delta} + \mathbf{\Delta} \mathbf{A}^{p+1}\right) z| \lesssim \delta \sqrt{d} \|\mathbf{A}\|_{\mathrm{op}} \|\mathbf{A}^p\|_{\mathrm{F}}^2. \tag{39}$$

现在令 $U' = z^\top \mathbf{A}^{2p} z$。那么 $\mathbf{E}[U'] = \mathrm{tr}(\mathbf{A}^{2p}) = \|\mathbf{A}^p\|_{\mathrm{F}}^2$。根据高斯反集中性（事实 2.9），$U'$的标准差至多为 $O(\|\mathbf{A}^p\|_{\mathrm{F}})$，由此可得 U'以高概率成立的上界。结合事实 2.9，我们得到该结论成立的概率为 0.999：

$$z^\top \mathbf{A}^{2p} z \asymp \|\mathbf{A}^p\|_{\mathrm{F}}^2 . \tag{40}$$

类似地，我们以高概率得到 $\|z\|_2^2 \lesssim d$。结合方程 (38)至(40)，我们得到该结论成立的概率至少为 0.99：

$$\begin{aligned}
\|y\|_2^2 &= z^\top \widehat{\mathbf{A}}^2 z = z^\top \left(\mathbf{A}^p + \mathbf{\Delta}\right)^2 z \\
&= z^\top \mathbf{A}^{2p} z + z^\top \left(\mathbf{A}^p \mathbf{\Delta} + \mathbf{\Delta} \mathbf{A}^p\right) z + z^\top \mathbf{\Delta}^2 z \\
&\le z^\top \mathbf{A}^{2p} z + C\delta \sqrt{d} \|\mathbf{A}^p\|_{\mathrm{F}}^2 + Cd \|\mathbf{\Delta}\|_{\mathrm{op}}^2 \\
&\le z^\top \mathbf{A}^{2p} z + C \|\mathbf{A}^p\|_{\mathrm{F}}^2 \left(\delta \sqrt{d} + d\delta^2\right) \\
&\le z^\top \mathbf{A}^{2p} z + 2C\delta \sqrt{d} \|\mathbf{A}^p\|_{\mathrm{F}}^2 .
\end{aligned} \tag{41}$$

Si类似地，我们得到以下下界，该下界成立的概率至少为 0.99：

$$\|y\|_2^2 \ge z^\top \mathbf{A}^{2p} z - C\delta \sqrt{d} \|\mathbf{A}^p\|_{\mathrm{F}}^2 . \tag{42}$$

根据 事实 2.11，标准的幂迭代表明，以 0.999 的概率，我们有：

$$\frac{z^\top \mathbf{A}^{2p+1} z}{z^\top \mathbf{A}^{2p} z} \ge (1-\gamma) \|\mathbf{A}\|_{\mathrm{op}} . \tag{43}$$

在所有这些事件的交集上，由并集界可知其成立的概率至少为 0.9，因此我们得到 $y^\top \mathbf{A} y$ 的以下下界：

$$\begin{aligned}
y^\top \mathbf{A} y &= z^\top \widehat{\mathbf{A}} \mathbf{A} \widehat{\mathbf{A}} z = z^\top \left(\mathbf{A}^p + \mathbf{\Delta}\right) \mathbf{A} \left(\mathbf{A}^p + \mathbf{\Delta}\right) z \\
&= z^\top \mathbf{A}^{2p+1} z + z^\top (\mathbf{A}^{p+1} \mathbf{\Delta} + \mathbf{\Delta} \mathbf{A}^{p+1}) z + z^\top (\mathbf{\Delta} \mathbf{A} \mathbf{\Delta}) z \\
&\qquad \text{(using Equation (39) and } z^\top \mathbf{\Delta} \mathbf{A} \mathbf{\Delta} z \ge 0) \\
&\ge z^\top \mathbf{A}^{2p+1} z - C\delta \sqrt{d} \|\mathbf{A}\|_{\mathrm{op}} \|\mathbf{A}^p\|_{\mathrm{F}}^2 \\
&\ge (1-\gamma) \|\mathbf{A}\|_{\mathrm{op}} z^\top \mathbf{A}^{2p} z - C\delta \sqrt{d} \|\mathbf{A}\|_{\mathrm{op}} \|\mathbf{A}^p\|_{\mathrm{F}}^2 \qquad \text{(using Equation (43))} \\
&\ge (1-\gamma) \|\mathbf{A}\|_{\mathrm{op}} \left(z^\top \mathbf{A}^{2p} z - 3C\delta \sqrt{d} \|\mathbf{A}^p\|_{\mathrm{F}}^2\right) \qquad (\gamma < 2/3) \\
&\ge (1-\gamma) \|\mathbf{A}\|_{\mathrm{op}} \left(\|y\|^2 - 5C\delta \sqrt{d} \|\mathbf{A}^p\|_{\mathrm{F}}^2\right) . \qquad \text{(using Equation (41))}
\end{aligned}$$

最后，两边同时除以 $\|y\|_2^2$，得到

$$\frac{y^\top \mathbf{A} y}{\|y\|_2^2} \ge (1-\gamma) \|\mathbf{A}\|_{\mathrm{op}} \left(1 - 5C\delta \sqrt{d} \|\mathbf{A}^p\|_{\mathrm{F}}^2 \frac{1}{\|y\|_2^2}\right).$$

现在我们将证明，在上述事件的交集上，$5C\delta\sqrt{d}\|\mathbf{A}^p\|_{\mathrm{F}}^2 \frac{1}{\|y\|_2^2} \lesssim \gamma$，从而完成证明：

$$\begin{aligned}
\delta\sqrt{d}\|\mathbf{A}^p\|_{\mathrm{F}}^2 \frac{1}{\|y\|_2^2} &\leq \frac{\delta\sqrt{d}\|\mathbf{A}^p\|_{\mathrm{F}}^2}{z^\top \mathbf{A}^{2p} z - 2C\delta\sqrt{d}\|\mathbf{A}^p\|_{\mathrm{F}}^2} && \text{(using Equation (42))} \\
&\lesssim \frac{\delta\sqrt{d}\|\mathbf{A}^p\|_{\mathrm{F}}^2}{\|\mathbf{A}^p\|_{\mathrm{F}}^2 - C'\delta\sqrt{d}\|\mathbf{A}^p\|_{\mathrm{F}}^2} && \text{(using Equation (40))} \\
&= \frac{\delta\sqrt{d}}{1 - C'\delta\sqrt{d}} \lesssim \delta\sqrt{d} \lesssim \gamma. && \text{(using } \delta < c\gamma/\sqrt{d}\text{)}
\end{aligned}$$

□

4.2 定理4.1 的证明概要

算法 6 的正确性证明 与第 3 节中的证明相同，只需对界中的常数作少量调整，以考虑使用我们的基于样本的估计量。因此，这里我们仅推导样本复杂度、内存和运行时间界。

样本复杂度：考虑算法 6 的第 (k,t) 次迭代（即外层循环的第 k 次迭代和内层循环的第 t 次迭代）。为了确保满足条件 4.2 的所有部分，我们使用引理 B.1，其中 $\delta = \frac{0.01}{\sqrt{d}}\min\left(\frac{\sqrt{\gamma/\epsilon}}{r}, \gamma\right)$，并将失败概率 τ设为关于 ϵ/d 的一个足够大的多项式，使得以高概率所有估计量在所有迭代中同时都准确。因此，每个估计量 $\widehat{\mathrm{B}}_{k,t,\ell}$ 使用 $(r^2 p_k^2 d/\delta^2)\mathrm{polylog}(d/\epsilon)$ 个样本。我们在第 13 行使用 p_k 个此类估计量，因此需要 $(r^2 p_k^3 d/\delta^2)\mathrm{polylog}(d/\epsilon)$ 个样本。此外，对于第 16 行的估计量，我们使用引理 B.4（同样令 $\tau = \mathrm{poly}(\epsilon/d)$），这会使样本复杂度额外增加 $(1/\epsilon)\mathrm{polylog}(d/\epsilon)$。我们还调用引理 B.3 的估计量，它使用 $O((r^4 d^2/\gamma^2)\mathrm{polylog}(d/\epsilon))$ 个样本。因此，算法 6 的第 (k,t) 次迭代期间使用的样本数为$n_{k,t} := O\left(r^2 p_k^3 d/\delta^2 \mathrm{polylog}(d/\epsilon) + (1/\epsilon)\mathrm{polylog}(d/\epsilon) + (r^4 d^2/\gamma^2)\mathrm{polylog}(d/\epsilon)\right)$，总样本复杂度为

$$\begin{aligned}
n = \sum_{k=1}^{k_{\mathrm{end}}}\sum_{t=1}^{t_{\mathrm{end}}} n_{k,t} &\lesssim t_{\mathrm{end}}\left(\sum_{k=1}^{k_{\mathrm{end}}} p_k^3\right)\frac{r^2 d}{\delta^2}\mathrm{polylog}(d/\epsilon) + \frac{t_{\mathrm{end}} k_{\mathrm{end}}}{\epsilon}\mathrm{polylog}(d/\epsilon) + t_{\mathrm{end}} k_{\mathrm{end}}\frac{r^4 d^2}{\gamma^2}\mathrm{polylog}(d/\epsilon) \\
&\lesssim \frac{r^2 d^2 \max(r^2\epsilon, 1)}{\gamma^5}\mathrm{polylog}(d/\epsilon) + \frac{1}{\epsilon\gamma}\mathrm{polylog}(d/\epsilon) + \frac{r^4 d^2}{\gamma^3}\mathrm{polylog}(d/\epsilon)\,.
\end{aligned}$$

内存使用量：算法 6 中第 16 行的估计量使用 $O((1/\epsilon)\mathrm{polylog}(d/\epsilon))$ 内存，而且下次执行第 16 行时，可以释放并重复使用这部分内存。HardThresholdingFilter 创建的过滤器由一个向量和一个一维阈值构成，因此可以存储在 $O(d)$ 内存中。每次调用 HardThresholdingFilter 时，创建的过滤器数量至多为 $O(\log(d/\epsilon))$，共调用 $t_{\mathrm{end}} \cdot k_{\mathrm{end}}$ 次。因此，存储所有这些过滤器所用的内存为 $O(t_{\mathrm{end}} \cdot k_{\mathrm{end}} \cdot d\log(d/\epsilon)) = O((d/\gamma)\mathrm{polylog}(d/\epsilon))$。算法 6 中执行的其他操作，要么是将经验二阶矩矩阵与向量相乘（如前所述，这可以在 $O(d)$ 内存中实现），要么是更基本的操作。

运行时间：这里同样适用第 3.4.4 节 中的运行时间分析。

4.2.1 位复杂度

我们简要讨论如何使用 $(d/\gamma)\mathrm{polylog}(d/\epsilon)$ 个寄存器来实现我们的算法，其中每个寄存器的大小为 $(1/\gamma)$ $\mathrm{polylog}(d/\epsilon)$ 位，从而使其能够在有界精度下运行（即在标准字 RAM 模型中）。我们假设

除去质量为 ϵ 的内点分布外，所有内点均位于半径为 $R = (d/\epsilon)^{\mathrm{polylog}(d/\epsilon)}$ 的球内，并且我们还假设 $\mathbf{\Sigma} \succeq (\epsilon/d)^{\mathrm{polylog}(d/\epsilon)}\mathbf{I}$。我们的算法忽略球外的所有点，并将剩余点舍入为 x'，使得 $\|x - x'\|_2 \leq \eta$。我们希望 η 足够小，以免影响算法的正确性。选择 $\eta = (\epsilon/d)^{\mathrm{polylog}(d/\epsilon)}$ 意味着舍入后的分布的协方差矩阵与原始分布足够接近，从而满足相同的稳定性条件（允许常数上的差异），因此我们的算法将是正确的。到目前为止，通过上述舍入，每个点都可以存储在 d 个字中，每个字的大小为 $O(\log(Rd/\eta)) = \mathrm{polylog}(d/\epsilon)$ 位。关于算法中间计算的位复杂度，开销最大的是计算 $\mathbf{M}z$，其中 $\mathbf{M}$ 是经验协方差矩阵的 $p = \gamma^{-1}\mathrm{polylog}(d/\epsilon)$ 次幂，而 z 是一个随机高斯向量。由于幂 p 包含一个 $1/\gamma$ 因子，最终结果的位复杂度可能增加 $1/\gamma$ 倍，因此我们需要 d 个大小为 $(1/\gamma)\mathrm{polylog}(d/\epsilon)$ 的寄存器来存储所得向量 $\mathbf{M}z$。我们创建的过滤器数量为 $(1/\gamma)\mathrm{polylog}(d/\epsilon)$，每个过滤器的表示形式都是前述形式的一个向量（以及一个一维阈值）。因此，总体而言，我们需要使用 $(d/\gamma)\mathrm{polylog}(d/\epsilon)$ 个寄存器，每个寄存器的大小为 $(1/\gamma)\mathrm{polylog}(d/\epsilon)$ 位。

5 结论

我们提出了首个近线性时间的鲁棒主成分分析算法，在 Σ 的谱中不存在特征值间隙假设的情况下，仍能达到近乎最优的误差保证。此外，我们还提出了首个用于鲁棒主成分分析的亚二次空间流式算法。

关于未来的改进，一个可能的优化方向是计算$\mathbf{B}^{\tilde{p}}$对于半正定矩阵$\mathbf{B}$ 和 $p = \Theta(1/\epsilon)$。通过使用切比雪夫近似，这些矩阵点积很可能可以更快地近似计算 [SV14，第 10]章。此外，真实协方差矩阵的大特征值之间存在间隔时，我们算法的运行时间是否会有所改善，仍有待确定。最后，正如 [DKPP22], 所指出的，如何设计具有最优样本复杂度且能抵御强污染的流式算法，仍是一个开放问题。

参考文献

[BDJ+22] A. Bakshi、I. Diakonikolas、H. Jia、D. M. Kane、P. K. Kothari 和 S. S. Vempala。鲁棒学习任意 k 个高斯分布的混合。载于第 54 届 ACM 计算理论研讨会论文集（STOC），2022年。[BDLS17] S. Balakrishnan、S. S. Du、J. Li 和 A. Singh。高维计算高效的鲁棒稀疏估计。载于第 30 届学习理论会议论文集（COLT 2017），第 169–212 页，2017年。[BJSY22] D. Bienstock、M. Jeong、A. Shukla 和 S. Yun。鲁棒流式主成分分析。载于《神经信息处理系统进展》第 35 卷（NeurIPS），2022年。[CAT+20] Y. Cherapanamjeri、E. Aras、N. Tripuraneni、M. I. Jordan、N. Flammarion 和 P. L. Bartlett。近线性时间内的最优鲁棒线性回归。abs/2007.08137，2020年。[CDG19] Y. Cheng、I. Diakonikolas 和 R. Ge。近线性时间内的高维鲁棒均值估计。载于第 30 届年度离散算法研讨会论文集（SODA），第 2755–2771 页。SIAM，2019年。[CDGW19] Y. Cheng、I. Diakonikolas、R. Ge 和 D. P. Woodruff。更快的高维鲁棒协方差估计算法。载于第 32 届年度学习理论会议论文集（COLT），2019年。[CDK+22] Y. Cheng、I. Diakonikolas、D. M. Kane、R. Ge、S. Gupta 和 M. Soltanolkotabi。通过非凸优化进行抗离群值鲁棒稀疏估计。载于《神经信息处理系统进展》第 35 卷（NeurIPS），2022年。[CLMW11] E. J. Candès、X. Li、Y. Ma 和 J. Wright。鲁棒主成分分析？《ACM期刊》，58(3)，2011年6月。[CTBJ22] Y. Cherapanamjeri、N. Tripuraneni、P. L. Bartlett 和 M. I. Jordan。无方差条件下的最优均值估计。载于第 35 届年度学习理论会议论文集（COLT），2022年。[DHL19] Y. Dong、S. B. Hopkins 和 J. Li。用于快速鲁棒均值估计和改进离群值检测的量子熵评分。载于《神经信息处理系统进展》第 32 卷（NeurIPS），2019年。[DK19] I. Diakonikolas 和 D. M. Kane。算法高维鲁棒统计的最新进展。CoRR，abs/1911.05911，2019年。[DK23] I. Diakonikolas 和 D. M. Kane。《算法高维鲁棒统计》。剑桥大学出版社，2023年。[DKK+16] I. Diakonikolas、G. Kamath、D. M. Kane、J. Li、A. Moitra 和 A. Stewart。摆脱计算上的难处理性，实现高维鲁棒估计量。载于 FOCS，第 655–664 页，2016年。[DKK+17] I. Diakonikolas、G. Kamath、D. M. Kane、J. Li、A. Moitra 和 A. Stewart。实现鲁棒性（在高维中）可以很实用。载于第 34 届国际机器学习会议论文集（ICML），2017年。

[DKK+19a] I. Diakonikolas、G. Kamath、D. M. Kane、J. Li、J. Steinhardt 和 A. Stewart。《Sever：一种用于随机优化的鲁棒元算法》。载于：第36届国际机器学习会议（ICML）论文集，2019年。[DKK+19b] I. Diakonikolas、D. M. Kane、S. Karmalkar、E. Price 和 A. Stewart。通过迭代过滤实现抗异常值的高维稀疏估计。载于：神经信息处理系统进展第32卷（NeurIPS），2019年。[DKK+22a] I. Diakonikolas、D. M. Kane、S. Karmalkar、A. Pensia 和 T. Pittas。通过平方和实现鲁棒稀疏均值估计。载于：第35届年度学习理论会议（COLT）论文集，2022年。[DKK+22b] I. Diakonikolas、D. M. Kane、D. Kongsgaard、J. Li 和 K. Tian。通过列表可解码均值估计，以近线性时间对混合模型进行聚类。载于：第54届年度 ACM SIGACT计算理论研讨会（STOC）论文集，2022年。[DKLP22] I. Diakonikolas、D. M. Kane、J. C. H. Lee 和 A. Pensia。重尾分布的抗异常值稀疏均值估计。载于：神经信息处理系统进展第35卷（NeurIPS），2022年。[DKP20] I. Diakonikolas、D. M. Kane 和 A. Pensia。通过稳定性实现具有次高斯速率的抗异常值均值估计。载于：神经信息处理系统进展第33卷（NeurIPS），2020年。[DKPP22] I. Diakonikolas、D. M. Kane、A. Pensia 和 T. Pittas。高维鲁棒统计的流式算法。载于：第39届国际机器学习会议（ICML）论文集，2022年。预印本见 https://arxiv.org/abs/2204.12399。[DKS17] I. Diakonikolas、D. M. Kane 和 A. Stewart。鲁棒估计高维高斯分布和高斯混合分布的统计查询下界。载于：第58届 IEEE 计算机科学基础研讨会（FOCS）论文集，第73–84页，2017年。[DKS18] I. Diakonikolas、D. M. Kane 和 A. Stewart。列表可解码鲁棒均值估计与球形高斯混合分布学习。载于：第50届年度 ACM SIGACT计算理论研讨会论文集（STOC 2018），第1047–1060页，2018年。完整版本见 https://arxiv.org/abs/1711.07211。[DKS19] I. Diakonikolas、W. Kong 和 A. Stewart。鲁棒线性回归的高效算法与下界。载于：第30届年度离散算法研讨会（SODA）论文集，第2745–2754页。SIAM，2019年。[DL22] J. Depersin 和 G. Lecué。以近线性时间对均值向量进行鲁棒次高斯估计。《统计学年鉴》，50(1)：511–536，2022年。[HL18] S. B. Hopkins 和 J. Li。混合模型、鲁棒性与平方和证明。载于：第50届年度 ACM 计算理论研讨会（STOC）论文集，2018年。[HLZ20] S. B. Hopkins、J. Li 和 F. Zhang。通过遗憾最小化，轻松实现鲁棒重尾均值估计。载于：神经信息处理系统进展第33卷（NeurIPS），2020年。

[HR21] L. Hu 和 O. Reingold。含任意异常值的高度不完整数据上的鲁棒均值估计。载于：第24届国际人工智能与统计会议（AISTATS）论文集，2021年。[JLT20] A. Jambulapati、J. Li 和 K. Tian。鲁棒次高斯主成分分析与宽度无关的 Schatten 打包。《神经信息处理系统进展》第33卷（NeurIPS），2020年。arXiv 预印本：https://arxiv.org/abs/2006.06980。[KKK19] S. Karmalkar、A. Klivans 和 P. K. Kothari。列表可解码线性回归。载于：神经信息处理系统进展第32卷（NeurIPS），2019年。[KKM18] A. Klivans、P. K. Kothari 和 R. Meka。抗异常值回归的高效算法。载于：第31届年度学习理论会议（COLT）论文集，第75卷，第1420–1430页。PMLR，2018年。[KSKO20] W. Kong、R. Somani、S. Kakade 和 S. Oh。小批次混合线性回归的鲁棒元学习。载于：神经信息处理系统进展第33卷（NeurIPS），2020年。[KSS18] P. K. Kothari、J. Steinhardt 和 D. Steurer。鲁棒矩估计与通过平方和改进聚类。载于：第50届年度 ACM 计算理论研讨会（STOC）论文集，第1035–1046页。ACM出版社，2018年。[LM21] A. Liu 和 A. Moitra。解决高斯混合的鲁棒可学习性问题。载于：第53届年度 ACM 计算理论研讨会（STOC）论文集，2021年。[LRV16] K. A. Lai、A. B. Rao 和 S. Vempala。均值和协方差的无先验估计。载于：第57届 IEEE 计算机科学基础研讨会（FOCS）论文集，第665–674页，2016年。[MMA18] T. V. Marinov、P. Mianjy 和 R. Arora。噪声环境下的流式主成分分析。载于：第35届国际机器学习会议（ICML）论文集，2018年。[PJL20] A. Pensia、V. Jog 和 P. Loh。采用协变量过滤的鲁棒回归：重尾与对抗性污染。CoRR，abs/2009.12976，2020年9月。[PSBR20] A. Prasad、A. S. Suggala、S. Balakrishnan 和 P. Ravikumar。通过稳健梯度估计实现稳健估计。《英国皇家统计学会杂志 B 辑》，82(3)：601–627，2020年7月。[SV14] S. Sachdeva 和 N. K. Vishnoi。通过逼近理论实现更快的算法。《理论计算机科学基础与趋势》，9：125–210，2014年。[Ver10] R. Vershynin。随机矩阵非渐近分析导论。arXiv 预印本 arXiv:1011.3027，2010年。[Ver18] R. Vershynin。《高维概率：数据科学中的应用导论》。剑桥大学出版社，2018年。[XCM13] H. Xu、C. Caramanis 和 S. Mannor。抗异常值主成分分析：高维情形。《IEEE 信息论汇刊》，59(1)：546–572，2013年。

[ZJS21] B. Zhu、J. Jiao 和 J. Steinhardt。通过广义拟梯度进行稳健估计。*Information and Inference: A Journal of the IMA*，11(2):581–636，2021。

附录

组织结构 附录的安排如下： 附录 A收录了正文中省略的其他标准结果（及其证明）。最后，附录 B 重点介绍实现定理 1.5的流式算法。

A 预备知识：省略的事实

A.1 信息论误差

我们简要说明一下，次高斯分布所能达到的最佳近似保证为何是 ˜阶（$1-\Theta(\epsilon)$）。为此，我们将聚焦于高斯分布，并证明一个（$1-\Omega(\epsilon)$）的下界。令 P为 $\mathbb{R}^d$中的标准高斯分布。对于一个单位向量 v 和$\epsilon \in (0,\ 0。1)$，令 Q为分布 $\mathcal{N}(0,\ \mathbf{I}+\epsilon vv^\top)$。根据下文给出的简单计算可知， $d_{\mathrm{TV}}(P,\ Q)\ \le \epsilon/2$：利用 Pinsker 不等式以及两个高斯分布之间 KL 散度的表达式，可得以下一系列不等式

$$\begin{aligned}
d_{\mathrm{TV}}(Q,P) &\le \sqrt{d_{\mathrm{KL}}(Q||P)} && \text{(Using Pinsker's inequality)}\\
&= \sqrt{\frac{1}{2}\left(-\log|\mathbf{\Sigma}_Q| + \mathrm{tr}(\mathbf{\Sigma}_Q) - d\right)} && \text{（两个高斯分布之间的 KL 散度）}\\
&= \sqrt{\frac{1}{2}\left(-\log(1+\epsilon)+\epsilon\right)} && \text{(Using } \mathbf{\Sigma}_Q = \mathbf{I}+\epsilon vv^\top\text{)}\\
&\le \epsilon/2. && \text{(since } \log(1+x)\ge x - x^2/2 \text{ for } x\ge 0\text{)}
\end{aligned}$$

设 $\widehat{v}\in\mathbb{R}^d$ 是任意鲁棒主成分分析算法在以 P 作为输入的独立同分布样本上运行时的输出。分布 Q 沿方向 $\widehat{v}$ 的真实方差为 $1+\epsilon(\widehat{v}^\top v)^2$。特别地，输出 $\widehat{v}$ 的算法所获得的近似因子（针对分布 Q）为

$$\frac{\widehat{v}^\top \mathbf{\Sigma}_Q \widehat{v}}{\|\mathbf{\Sigma}_Q\|_{\mathrm{op}}} = \frac{1+\epsilon(\widehat{v}^\top v)^2}{1+\epsilon} \le (1-0.5\epsilon)(1+\epsilon(\widehat{v}^\top v)^2).$$

由于 P 和 Q 的全变差距离至多为 $\epsilon/2$，因此允许损坏 ϵ 比例样本的对手，可以根据来自 Q 的独立同分布样本，模拟来自 P 的独立同分布样本，反之亦然。现在，假设真实分布是 Q，但对手给出来自 P 的独立同分布样本（如上所述，这是一个有效的对手）。由于 P 不包含任何关于 v 的信息，因此没有算法能够从 Q 的 ϵ 损坏样本集合中识别出 v。由于 v 可以是任意单位向量，因此对于大于某个常数的 d，没有算法能够输出一个 ˆv，使得 $|v^\top \widehat{v}| \le 0.01$。因此，对于高斯分布，不可能获得优于 $(1-0.1\epsilon)$ 的近似，因为 $(1-0.5\epsilon)(1+0.1\epsilon)\ \le (1-0.1\epsilon)$。

A.2 两种停止条件的比较

考虑 第 3.1 节中的情形。设 w： $\mathbb{R}^d \to \{0,1\}$ 为一组权重，使得内点的概率至少为 $1-3\epsilon$。设 $r = \mathbf{E}_{X\sim P}[w_t(X)]$。回顾一下， $\mathbf{M} = \left(\mathbf{E}_{X\sim P}[w_t(X)XX^\top]\right) = (\mathbf{E}_{X\sim P}[w_t(X)]\mathbf{\Sigma}_t)^p = r^p\mathbf{\Sigma}_t^p$。

当 $\langle \mathbf{\Sigma}_t, \mathbf{M}^2\rangle \le (1+C\gamma)\langle \mathbf{\Sigma}, \mathbf{M}^2\rangle$ 时，我们的算法停止。现在我们将证明，对于较大的 p，这一停止条件蕴含仅取决于 Schatten 范数的停止条件。根据 Schatten 范数的定义，我们从以下推导开始：

$$\|\mathbf{\Sigma}_t\|_{2p+1}^{2p+1} = \langle \mathbf{\Sigma}_t, \mathbf{\Sigma}_t^{2p}\rangle = r^{-2p}\langle \mathbf{\Sigma}_t, \mathbf{M}^2\rangle$$

$$
\begin{aligned}
&\le (1+C\gamma) r^{-2p}\langle \mathbf{\Sigma}, \mathbf{M}^2\rangle && \text{(Using stopping condition)}\\
&\le (1+C\gamma) r^{-2p}\|\mathbf{\Sigma}\|_{\mathrm{op}}\|\mathbf{M}\|_{\mathrm{F}}^2\\
&= (1+C\gamma)\|\mathbf{\Sigma}\|_{\mathrm{op}}\|\mathbf{\Sigma}_t\|_{2p}^{2p} && \text{(Using definition of } \mathbf{\Sigma}_t)\\
&\le (1+C\gamma)\|\mathbf{\Sigma}\|_{\mathrm{op}}\left(\|\mathbf{\Sigma}_t\|_{2p+1} d^{\frac{1}{2p(2p+1)}}\right)^{2p} && \text{(Using Fact 2.7)}\\
&\le (1+2C\gamma)\|\mathbf{\Sigma}\|_{\mathrm{op}}\|\mathbf{\Sigma}_t\|_{2p+1}^{2p},
\end{aligned}
$$

其中我们利用了 $p \gg \log(d)/\gamma$，因此 $d^{1/(2p+1)} \le (1+C\gamma)$。重新整理可得

$$\|\mathbf{\Sigma}_t\|_{2p+1} \le (1+2C\gamma)\|\mathbf{\Sigma}\|_{\mathrm{op}},$$

因此 $\|\mathbf{\Sigma}_t\|_{2p+1}^{2p+1} \le (1+2C\gamma)^{2p+1}\|\mathbf{\Sigma}\|_{\mathrm{op}}^{2p+1}$，这是仅取决于 $\mathbf{\Sigma}_t$ 和 $\mathbf{\Sigma}$ 的 Schatten 范数的相应停止条件。因此，引理 3.3 的停止条件更强。

A.3 关于稳定性的省略的证明：引理 引理 2.14

本节重述并证明 引理 2.14。我们没有对常数进行优化（除了常数 C 的选择）。

引理 *2.14*。在设定 *2.13* 中，定义如下：*(i)* $g(x) := \|Ux\|_2^2$ 对于某个 $U \in \mathbb{R}^{m\times d}$,$L$，是 P_w *(*P 由 w 重新加权*)* 下 $g(x)$ 的顶部 3ϵ 分位数，以及 $S := \{x : \ x > L\}$，*(iii)*$\widehat{\sigma} := E_{X\sim P}[w(X)g(X)1(g(X) \le L)]$。那么，

(i) $\left|\mathop{\mathbf{E}}_{X\sim G}[w(X)g(X)] - \left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle\right| \le 2\gamma \left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle,$

(ii) $\mathbf{E}_{X\sim G}[w(X)g(X)\mathbb{1}\{X\in S\}] \le 2.35\gamma \left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle,$

(iii) $L \le 1.65(\gamma/\epsilon)\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\rangle,$

(iv) $\left|\widehat{\sigma} - \left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle\right| \le 4\gamma \left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle,$

(v) $\mathbf{Pr}_{X\sim G}\left[\|X\|_2^2 > 2(d/\epsilon)\|\mathbf{\Sigma}\|_{\mathrm{op}}\right] \le \epsilon.$

我们分别证明引理的各个部分。对于第一部分，有如下结论，它在稍强的条件 $\mathbf{E}_{X\sim G}[w(X)] \ge 1-7\epsilon$ 下成立（这是因为稍后会用到这一版本）。

引理 A.1. 设 $0 < 20\epsilon \le \gamma < 1$。设 G 是关于 $\mathbf{\Sigma}$ 的 $(20\epsilon, \gamma)$ 稳定分布。设矩阵 $\mathbf{U} \in \mathbb{R}^{m\times d}$ 和函数 $w\colon \mathbb{R}^d \to [0,1]$ 满足 $\mathbf{E}_{X\sim G}[w(X)] \ge 1-7\epsilon$。对于函数 $g(x) = \|\mathbf{U}x\|_2^2$，有

$$(1-1.35\gamma)\left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle \le \mathop{\mathbf{E}}_{X\sim G}[w(X)g(X)] \le (1+\gamma)\left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle .$$

证明。 对于上界，

$$\mathop{\mathbf{E}}_{X\sim G}[w(X)g(X)] = \left\langle \mathbf{U}^\top\mathbf{U}, \mathop{\mathbf{E}}_{X\sim G}[w(X)XX^\top]\right\rangle \le (1+\gamma)\mathop{\mathbf{E}}_{X\sim G}[w(X)]\left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle \le (1+\gamma)\left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle,$$

其中，第一步将其改写为迹内积的符号形式，第二步使用稳定性条件和 事实 2.2，其中 $\mathbf{A} = \mathbf{U}^\top\mathbf{U}$、$\mathbf{B} = \mathbf{E}_{X\sim G}[w(X)XX^\top]$、$\mathbf{C} = \mathbf{E}_{X\sim G}[w(X)](1+\gamma)\mathbf{\Sigma}$，以及

最后一步使用 $w(x) \le 1$。另一方向的证明类似，唯一的区别在于我们对$\mathbf{E}_{X\sim G}[w(X)] \ge 1-7\epsilon$取下界：

$$\mathop{\mathbf{E}}_{X\sim G}[w(X)g(X)] \ge (1-\gamma)(1-7\epsilon)\left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle \ge (1-\gamma-7\epsilon)\left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle \ge (1-1.35\gamma)\left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle ,$$

其中我们还用到了 $\epsilon \le \gamma/20$。 □

我们将需要以下中间结果：

引理 A.2。 令 $0 < 20\epsilon \le \gamma < 1$。令 G 是关于 $\mathbf{\Sigma}$ 的 $(20\epsilon、\ \gamma)$ 稳定分布。令矩阵 $\mathbf{U} \in \mathbb{R}^{m\times d}$、函数 $w: \mathbb{R}^d \to [0,1]$，满足 $\mathbf{E}_{X\sim G}[w(X)] \ge 1-3\epsilon$，以及集合 S，使得$\mathbf{E}_{X\sim G}[w(X)\mathbb{1}\{X\in S\}] \le 4\epsilon$。那么，对于函数 $g(x) := \|\mathbf{U}x\|_2^2$，我们有如下结论：

$$\mathop{\mathbf{E}}_{X\sim G}[w(X)g(X)\mathbb{1}\{X\in S\}] \le 2.35\gamma\left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle .$$

证明。 令 $w'(x) := w(x)\mathbb{1}\{X \notin S\} = w(x) - w(x)\mathbb{1}\{X \in S\}$，并注意到根据我们的假设，$\mathbf{E}_{X\sim G}[w'(X)] = \mathbf{E}_{X\sim G}[w(X)] - \mathbf{E}_{X\sim G}[w(X)\mathbb{1}\{X\in S\}] \ge 1-7\epsilon$。因此，我们有

$$\begin{aligned}\mathop{\mathbf{E}}_{X\sim G}[w(X)g(X)\mathbb{1}\{X\in S\}] &= \mathop{\mathbf{E}}_{X\sim G}[w(X)g(X)] - \mathop{\mathbf{E}}_{X\sim G}[w'(X)g(X)] \\ &\le (1+\gamma)\left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle - (1-1.35\gamma)\left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle \\ &= 2.35\gamma\left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle ,\end{aligned}$$

其中第二行应用了 引理 A.1 于 $w'(x)$。 □

现在，引理 2.14 的条目 (ii) 如下所示可直接得到。

推论 A.3。 在引理 *A.2*的设定下，令混合分布 $P = (1-\epsilon)G + \epsilon B$ 和$S = \{x:\ x > L\}$，其中 L 定义为 3ϵ-分位数 $g(x)$ 在 P_w （按 w加权的 P版本）下的最高分位数。于是，

$$\mathop{\mathbf{E}}_{X\sim G}[w(X)g(X)\mathbb{1}\{X\in S\}] \le 2.35\gamma\left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle .$$

证明。 为了应用引理 A.2，我们只需验证 $\mathbf{E}_{X\sim G}[w(X)\mathbb{1}\{X\in S\}] \le 4\epsilon$。由于$P = (1-\epsilon)G + \epsilon B$ 且 L 根据定义是 3ϵ-分位数 $g(x)$ 在 P_w下的，并且 $\epsilon < 1/10$，我们有

$$\begin{aligned}\mathop{\mathbf{E}}_{X\sim G}[w(X)\mathbb{1}\{g(X) > L\}] &\le \frac{1}{1-\epsilon}\mathop{\mathbf{E}}_{X\sim P}[w(X)\mathbb{1}\{g(X) > L\}] \\ &\le \frac{1}{1-\epsilon}\mathop{\mathbf{Pr}}_{X\sim P_w}[g(X) > L] \\ &\le \frac{3\epsilon}{1-\epsilon} < 4\epsilon . \end{aligned} \tag{44}$$

应用引理 A.2 即可完成证明。 □

我们将在下文证明条目 (iii)，即引理 2.14 的内容。

引理 *A.4.* 设 $0 < 20\epsilon \le \gamma < 1$。设 $\mathbb{R}^d$ 上的混合分布为 $P = (1-\epsilon)G + \epsilon B$，其中 G 是相对于矩阵 $\mathbf{\Sigma}$ 的 $(20\epsilon, \gamma)$- 稳定分布。设 $U \in \mathbb{R}^{m\times d}$，且函数 $w: \mathbb{R}^d \to [0,1]$ 满足 $E_{X\sim G}[w(X)] \ge 1-3\epsilon$。定义 $g(x) := \|Ux\|_2^2$，并设 L 是 $g(x)$ 在 P_w 下的最高 3ϵ 分位数。那么， $L \le 1.65(\gamma/\epsilon)\langle U^\top U\ \mathbf{\Sigma}\rangle$。

证明。 设 $w'(x) := w(x)\mathbb{1}(g(x) > L)$。注意到

$$\begin{aligned}\mathop{\mathbf{E}}_{X\sim G}[w'(X)] &= \frac{\mathbf{E}_{X\sim P}[w(X)\mathbb{1}(g(X) > L)] - \epsilon\,\mathbf{E}_{X\sim B}[w(X)\mathbb{1}(g(X) > L)]}{1-\epsilon} \\ &\geq \frac{\mathbf{E}_{X\sim P}[w(X)\mathbb{1}(g(X) > L)] - \epsilon}{1-\epsilon} \\ &= \frac{\mathbf{E}_{X\sim P}[w(X)]\,\mathbf{Pr}_{X\sim P_w}[g(X) > L] - \epsilon}{1-\epsilon} \\ &\geq \frac{(1-\epsilon)(1-3\epsilon)3\epsilon - \epsilon}{1-\epsilon} \geq \frac{1.43\epsilon}{1-\epsilon}\,,\end{aligned}$$

其中，第二步利用了 $w(x) \leq 1$，第四步利用了 $\mathrm{E}_{X\sim P}[w(X)] \geq (1-\epsilon)\,\mathrm{E}_{X\sim G}[w(X)] \geq (1-\epsilon)(1-3\epsilon)$，以及 L 是 g 在 P_w下的 3ϵ分位数，最后一步利用了 $\epsilon < 1/20$。将 $\mathrm{E}_{X\sim G}[w'(X)]$ 的这个下界与推论 A.3结合，我们得到。

$$\mathop{\mathbf{E}}_{X\sim G_{w'}}[g(X)] = \frac{\mathbf{E}_{X\sim G}[w'(X)g(X)]}{\mathbf{E}_{X\sim G}[w'(X)]} \leq \frac{1-\epsilon}{1.43\epsilon}\mathop{\mathbf{E}}_{X\sim G}[w(X)g(X)\mathbb{1}(g(X) > L)] \leq \frac{1.65\gamma}{\epsilon}\left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle\,.$$

由于最小值小于均值，上述结果意味着存在一个点属于 $G_{w'}$ 的支撑集，且该点小于 $1.65(\gamma/\epsilon)\langle \mathbf{U}^\top\mathbf{U}\ \mathbf{\Sigma}\rangle$。由于 $G_{w'}$ 仅支撑在大于 L 的点上，这意味着 $L \leq 1.65(\gamma/\epsilon)\langle \mathbf{U}^\top\mathbf{U}\ \mathbf{\Sigma}\rangle$。

□

现在我们给出下面引理 2.14的第 (iv) 项的证明。

引理 *A.5*（方差估计量）。设 $0 < 20\epsilon \leq \gamma < 1$。令混合分布 $P = (1-\epsilon)G + \epsilon B$定义在 $\mathbb{R}^d$上，其中 G 是关于矩阵 $\mathbf{\Sigma}$的 $(20\epsilon, \gamma)$-稳定分布。令 $U \in \mathbb{R}^{m\times d}$ 以及函数 $w: \mathbb{R}^d \to [0,1]$，满足 $E_{X\sim G}[w(X)] \geq 1 - 3\epsilon$。定义 $g(x) := \|Ux\|_2^2$，并令 L 为在 P_w下 $g(x)$ 的前3ϵ分位数。对于估计量 $\hat{\sigma} := E_{X\sim P}[w(X)g(X)\mathbb{1}(g(X) \leq L)],$，我们有

$$\left|\hat{\sigma} - \left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle\right| \leq 4\gamma\left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle\,.$$

证明。我们分别考虑分布 P 的两个组成部分对 $\hat{\sigma}$ 的贡献。对于组成部分 G，我们首先注意到 $\mathrm{E}_{X\sim G}[w(X)\mathbb{1}(g(X) \leq L)] \geq \mathrm{E}_{X\sim G}[w(X)] - \mathrm{E}_{X\sim G}[w(X)\mathbb{1}(g(X) > L)] \geq 1 - 3\epsilon - 4\epsilon = 1 - 7\epsilon$（最后一步可按照公式 (44)中的方法证明）。因此，根据引理 A.1，我们有

$$\left|\mathop{\mathbf{E}}_{X\sim G}[w(X)g(X)\mathbb{1}(g(X) \leq L)] - \left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle\right| \leq 1.35\gamma\left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle\,. \tag{45}$$

现在我们考虑 $\hat{\sigma}$ 来自 B的贡献。首先注意到 $L \leq 1.65(\gamma/\epsilon)\langle \mathbf{U}^\top\mathbf{U}\ \mathbf{\Sigma}\rangle$ 根据引理 A.4成立。因此，

$$\left|\mathop{\mathbf{E}}_{X\sim B}[w(X)g(X)\mathbb{1}(g(X) \leq L)] - \left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle\right| \leq L + \left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle \leq \left(\frac{1.65\gamma}{\epsilon} + 1\right)\left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle\,. \tag{46}$$

最后，结合公式 (45)和(46) 并利用 $\epsilon \leq \gamma$ 可得:

$$\begin{aligned}\left|\mathop{\mathbf{E}}_{X\sim P}[w(X)g(X)\mathbb{1}(g(X) \leq L)] - \left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle\right| &\leq (1-\epsilon)\left|\mathop{\mathbf{E}}_{X\sim G}[w(X)g(X)\mathbb{1}(g(X) \leq L)] - \left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle\right| \\ &\quad + \epsilon\left|\mathop{\mathbf{E}}_{X\sim B}[w(X)g(X)\mathbb{1}(g(X) \leq L)] - \left\langle \mathbf{U}^\top\mathbf{U}, \mathbf{\Sigma}\right\rangle\right|\end{aligned}$$

$$\leq 4\gamma \left\langle \mathbf{U}^\top \mathbf{U}, \mathbf{\Sigma} \right\rangle .$$

□

我们以.....的最后一部分结束引理 2.14。

断言 A.6. 设 $0 < 20\epsilon \leq \gamma < 1$。设 G 是关于矩阵 $\mathbf{\Sigma}$ 的 $(20\epsilon, \gamma)$-稳定分布。那么，$\mathbf{Pr}_{X\sim G}\left[\|X\|_2^2 > 2(d/\epsilon)\|\mathbf{\Sigma}\|_{\mathrm{op}}\right] \leq \epsilon$。

证明。 利用马尔可夫不等式和 引理 A.1 ，取 $\mathbf{U} = \mathbf{I}$

$$\mathop{\mathbf{Pr}}_{X\sim G}\left[\|X\|_2^2 > 2(d/\epsilon)\|\mathbf{\Sigma}\|_{\mathrm{op}}\right] \leq \frac{\mathbf{E}_{X\sim G}[\|X\|_2^2]}{2(d/\epsilon)\|\mathbf{\Sigma}\|_{\mathrm{op}}} \leq \frac{(1+\gamma)\mathrm{tr}(\mathbf{\Sigma})}{2(d/\epsilon)\|\mathbf{\Sigma}\|_{\mathrm{op}}} \leq \epsilon .$$

□

我们最后重述并证明引理 2.15 ，它来自第 2.2 节。

引理 *A.7.* 令 $0 < 20\epsilon \leq \gamma < \gamma_0$ ，其中 γ_0 是足够小的绝对常数。设 $P = (1-\epsilon)G + \epsilon B$，其中 G 是关于 $\mathbf{\Sigma}$ 的 $(20\epsilon, \gamma)$-稳定分布。设 $w\colon \mathbb{R}^d \to [0,1]$ ，且 $E_{X\sim G}[w(X)] \geq 1-3\epsilon$；并回顾一下，在我们的符号约定中，$\mathbf{\Sigma}_{P_w} := E_{X\sim P_w}[XX^\top] = E_{X\sim P}[w(X)XX^\top]/E_{X\sim P}[w(X)]$。若 u 是一个向量，满足 $u^\top \mathbf{\Sigma}_{P_w} u/\|u\|_2^2 \geq (1-\gamma)\|\mathbf{\Sigma}_{P_w}\|_{\mathrm{op}}$ 且 $u^\top \mathbf{\Sigma} u \geq (1-O(\gamma))u^\top \mathbf{\Sigma}_{P_w} u$，则 $u^\top \mathbf{\Sigma} u/\|u\|_2^2 \geq (1-O(\gamma))\|\mathbf{\Sigma}\|_{\mathrm{op}}$。

证明。 由于 $\mathbf{E}_{X\sim G}[w(X)] \geq 1-3\epsilon$，我们可以如下使用稳定性条件：令 a 为 $\mathbf{\Sigma}_{P_w}$ 对应的归一化最大特征值特征向量，令 b 为 $\mathbf{\Sigma}$对应的归一化最大特征值特征向量。根据算子范数的定义和稳定性，我们首先得到以下关于 $\|\mathbf{\Sigma}\|_{\mathrm{op}}$的 $\|\mathbf{\Sigma}_{P_w}\|$ 下界：

$$\|\mathbf{\Sigma}_{P_w}\|_{\mathrm{op}} = a^\top \mathbf{\Sigma}_{P_w} a \geq b^\top \mathbf{\Sigma}_{P_w} b = b^\top \left(\frac{\mathbf{E}_{X\sim P}[w(X)XX^\top]}{\mathbf{E}_{X\sim P}[w(X)]}\right) b \tag{47}$$

$$\geq (1-\epsilon) b^\top \left(\frac{\mathbf{E}_{X\sim G}[w(X)XX^\top]}{\mathbf{E}_{X\sim P}[w(X)]}\right) b = (1-\epsilon) b^\top \left(\frac{\mathbf{E}_{X\sim G}[w(X)]}{\mathbf{E}_{X\sim P}[w(X)]}\mathbf{\Sigma}_{G_w}\right) b \tag{48}$$

$$\geq (1-\epsilon)(1-3\epsilon) b^\top \mathbf{\Sigma}_{G_w} b \geq (1-4\epsilon)(1-\gamma) b^\top \mathbf{\Sigma} b \tag{49}$$

$$\geq (1-2\gamma)\, b^\top \mathbf{\Sigma} b = (1-2\gamma)\|\mathbf{\Sigma}\|_{\mathrm{op}} , \tag{50}$$

其中倒数第二步用到了 $\epsilon < \gamma/20$。结合 $u^\top \mathbf{\Sigma} u$ 远大于 $u^\top \mathbf{\Sigma}_{P_w} u$这一假设，我们得到：

$$\begin{aligned}\frac{u^\top \mathbf{\Sigma} u}{\|u\|_2^2} &\geq (1-O(\gamma))\frac{u^\top \mathbf{\Sigma}_{P_w} u}{\|u\|_2^2} \geq (1-O(\gamma))(1-\gamma)\|\mathbf{\Sigma}_{P_w}\|_{\mathrm{op}} \\ &\geq (1-O(\gamma))(1-\gamma)(1-2\gamma)\|\mathbf{\Sigma}\|_{\mathrm{op}} = (1-O(\gamma))\|\mathbf{\Sigma}\|_{\mathrm{op}} .\end{aligned}$$

□

A.4 筛选

在本小节中，我们证明第 2.3 节中所描述的筛选保证。我们使用一个略微更一般的滤波器版本，以便稍后开发流式算法时能够采用它。其区别在于引入了一个额外的误差参数 δ 到$\widehat{}$滤波器中。这将在附录 B中发挥作用，其中数量$\mathbf{E}_{X\sim P}[w(x)\tau(x)]$ 以及 T将必须通过对样本取平均来估计，因此 δ 将用于计入该近似中的一个小加性误差。对于非流式算法，我们将使用 $\delta = 0$

算法 8 HardThresholdingFilter

1：输入：分布 $P=(1-\epsilon)G+\epsilon B$、权重 w、分数 τ、参数 T、R、δ。2：$w_0(x) \leftarrow w(x)$，以及 $r_0 \leftarrow R$、$\ell \leftarrow 1$。3：当 $\mathrm{E}_{X\sim P}[w(X)\tau(X)] > \frac{5}{2}(\widehat{T}+\delta)$ 时 4：抽取 $r_\ell \sim \mathcal{U}([0, r_{\ell-1}])$。5：$w_{\ell+1}(x) \leftarrow w_\ell(x) \cdot \mathbb{1}(\tau(x) > r_\ell)$。6：$\ell \leftarrow \ell+1$。7：返回 $w_\ell(x)$。

引理 A.8 (过滤保证)。 考虑调用一次HardThresholdingFilter$(P, w, \tau, T, R, \delta)$。假设存在 T使得$(1-\epsilon)\,\mathbf{E}_{X\sim G}\left[w(X)\tau(X)\right] < T$，并且 T使得 $|T-T| < T/5+\delta$。记 F为滤波器的随机性（即所使用的随机阈值 r_1、$r_2\ldots$的集合）。令 w' 为返回的权重函数。那么，

1. $\mathbf{E}_{X\sim P}[w'(X)\tau(X)] \le 3T+5d$关于滤波器的随机性 $\mathcal{F}$，以概率 1 成立。
2. $\mathbf{E}_F[\epsilon\,\mathbf{E}_{X\sim B}[w(X)-w'(X)]] > (1-\epsilon)\,\mathbf{E}_{X\sim G}[w(X)-w'(X)]]$.

证明。 引理的第一部分由终止条件$\mathbf{E}_{X\sim P}[w'(X)\tau(X)] \le (5/2)(T+\delta)$ 以及假设 $|T-T| < T/5+\delta$ 可得。

我们证明，在给定之前各轮过滤结果的条件下，第二个期望性质对每一轮过滤都成立。考虑第 ℓ轮过滤。一个点被移除的概率（对 r_ℓ 的随机选择取平均）为 $\tau(x)/r_{\ell-1}$。因此，对于任意尚未被过滤的点 x，一个被移除的点为 的概率等于 $\mathbf{E}[(w_\ell(x)-w_{\ell+1}(x))] = \tau(x)/r_{\ell-1}$ 此外，还要注意，由于权重是二值递减函数，$w_\ell(x)-w_{\ell+1}(x) = w_\ell(x)(w_\ell(x)-w_{\ell+1}(x))$. 对于内点，我们有

$$
\begin{aligned}
\mathop{\mathbf{E}}_{r_\ell}\left[(1-\epsilon)\mathop{\mathbf{E}}_{X\sim G}[w_\ell(X)-w_{\ell+1}(X)]\right] &= \mathop{\mathbf{E}}_{r_\ell}\left[(1-\epsilon)\mathop{\mathbf{E}}_{X\sim G}[w_\ell(X)(w_\ell(X)-w_{\ell+1}(X))]\right] \\
&= (1-\epsilon)\mathop{\mathbf{E}}_{X\sim G}\left[w_\ell(X)\mathop{\mathbf{E}}_{r_\ell}[w_\ell(X)-w_{\ell+1}(X)]\right] \\
&= \frac{1-\epsilon}{r_{\ell-1}}\mathop{\mathbf{E}}_{X\sim G}[w_\ell(X)\tau(X)] < \frac{T}{r_{\ell-1}}\,.
\end{aligned}
$$

现在我们来处理异常值。如果过滤尚未停止，这意味着while 循环的停止条件仍未满足，因此 $\mathbf{E}_{X\sim P}\left[w_\ell(X)\tau(X)\right] > (5/2)(T+\delta) > 2T$（因为我们已假设 $|T-T| < T/5+\delta$）。因此，

$$
\begin{aligned}
\mathop{\mathbf{E}}_{r_\ell}\left[\epsilon\mathop{\mathbf{E}}_{X\sim B}[w_\ell(X)-w_{\ell+1}(X)]\right] &= \mathop{\mathbf{E}}_{r_\ell}\left[\epsilon\mathop{\mathbf{E}}_{X\sim B}[w_\ell(X)(w_\ell(X)-w_{\ell+1}(X))]\right] \\
&= \mathop{\mathbf{E}}_{X\sim P}\left[w_\ell(X)\mathop{\mathbf{E}}_{r_\ell}[w_\ell(X)-w_{\ell+1}(X)]\right] \\
&\quad -(1-\epsilon)\mathop{\mathbf{E}}_{X\sim G}\left[w_\ell(X)\mathop{\mathbf{E}}_{r_\ell}[w_\ell(X)-w_{\ell+1}(X)]\right] \\
&= \frac{1}{r_{\ell-1}}\mathop{\mathbf{E}}_{X\sim P}[w_\ell(X)\tau(X)] - \frac{1-\epsilon}{r_{\ell-1}}\mathop{\mathbf{E}}_{X\sim G}[w_\ell(X)\tau(X)] \\
&> \frac{2T}{r_{\ell-1}} - \frac{T}{r_{\ell-1}} = \frac{T}{r_{\ell-1}}\,.
\end{aligned}
$$

□

现在我们给出 RobustPCA 整个执行过程中该过滤算法的保证。

引理 A.9. 设 $0 < 20\epsilon \le \gamma < 1/20$。设 $P = (1-\epsilon)G + \epsilon B$，其中 G 是关于 $\mathbf{\Sigma}$ 的 $(20\epsilon, \gamma)$-稳定分布。考虑一次 RobustPCA 的执行。对于算法的随机性，至少以 0.2 的概率，有 $(1-\epsilon)\,\mathbf{E}_{X\sim G}[1 - w_{k,t}(X)] + \epsilon \mathbf{E}_{X\sim B}[w_{k,t}(X)] \le 2.9\epsilon$。

证明。为简化符号，首先按自然顺序为多重指标 $\{(k,t) : k \in \{1,\dots,k_{\mathrm{end}}\}, t \in \{1,\dots,t_{\mathrm{end}}\}\}$ 定义一个排序：$(k',t') \le (k,t)$ 表示以下任一情况成立：$k' \le k$，或 $k' = k$ 且 $t' \le t$。另定义后继算子 $\mathsf{s}(\cdot)$，它按照 RobustPCA 中两个循环的结构返回下一个多重指标，也就是说，如果 $t < t_{\mathrm{end}}$，则 $\mathsf{s}(k,t) = (k, t+1)$；否则 $\mathsf{s}(k,t) = (k+1, 0)$。

令 (k^*, t^*) 为首次出现 $(1-\epsilon)\mathbf{E}_{X\sim G}[1 - w_{k,t}(X)] + \epsilon\, \mathbf{E}_{X\sim B}[w_{k,t}(X)] \ge 2.9\epsilon$ 的时刻（如果存在；否则令 $(k^*, t^*) = (k_{\mathrm{end}}, t_{\mathrm{end}})$）。定义序列

$$\Delta_{(k,t)} = (1-\epsilon) \mathop{\mathbf{E}}_{X\sim G}\left[1 - w_{\min\{(k,t),(k^*,t^*)\}}(X)\right] + \epsilon \mathop{\mathbf{E}}_{X\sim B}\left[w_{\min\{(k,t),(k^*,t^*)\}}(X)\right] .$$

注意，$\Delta_{(k,t)}$ 关于滤波器的随机性构成一个超鞅：若 $(k,t) < (k^*, t^*)$，则 $\mathrm{E}[\Delta_{\mathsf{s}(k,t)} \mid \mathcal{F}_{(k,t)}] \le \Delta_{(k,t)}$，其中 $\mathcal{F}_{(k,t)}$ 表示迄今为止所应用过滤器的随机性。只要引理 A.8 适用，这一结论便由该引理得出：我们在该引理中取 $T = 2.35\gamma v_k^\top{}_{,t}\mathbf{\Sigma} v_{k,t}$ 和 $\widehat{T} = 2.35\gamma\widehat{\sigma}_{k,t}$。其第一项要求，即 $(1-\epsilon)\,\mathrm{E}_{X\sim G}[w_{k,t}(x)\tau_{k,t}(x)] < T$，由引理 2.14 的条目 (ii)（具体而言，是使用 $\mathrm{U} = v_{k,t}^\top$ 的特殊情形）满足。对于第二项要求 $|\widehat{T} - T| < T/5$，我们有 $|\widehat{T} - T| = 2.35\gamma|\widehat{\sigma}_{k,t} - v_{k,t}^\top\mathbf{\Sigma} v_{k,t}| \le 2.35\gamma \cdot 4\gamma \cdot v_{k,t}^\top\mathbf{\Sigma} v_{k,t} \le (1/5) \cdot 2.35\gamma v_{k,t}^\top\mathbf{\Sigma} v_{k,t} = T/5$；其中第二步使用引理 2.14 的第 (iv) 项（取 $\mathrm{U} = v_{k,t}^\top$），最后一个不等式则利用了 $\gamma < 1/20$。刚才引用的所有这些中间引理，均在 $(1-\epsilon)\,\mathrm{E}_{X\sim G}[1 - w_{k,t}(X)] + \epsilon\,\mathrm{E}_{X\sim B}[w_{k,t}(X)] \le 2.9\epsilon$ 时适用。

否则，如果 $(k,t) \ge (k^*, t^*)$，我们仍有 $\mathbf{E}[\Delta_{\mathsf{s}(k,t)} \mid \mathcal{F}_{(k,t)}] \le \Delta_{(k,t)}$，因为根据定义，序列 $\Delta_{\mathsf{s}(k,t)}$ 保持不变。

一开始，我们有 $\Delta_{(1,1)} \le 2\epsilon$。这是因为我们从含有 ϵ比例异常值的分布开始，而算法第 5行的朴素剪枝最多只能从分布中移除 ϵ 的质量G，如引理 2.14的第 (v) 项所示。

根据鞅的 Doob 不等式，我们有

$$\mathbf{Pr}\left[\max_{(1,1)\le(k,t)\le(k_{\mathrm{end}},t_{\mathrm{end}})} \Delta_{(k,t)} > 2.5\epsilon\right] \le \frac{\Delta_{(1,1)}}{2.5\epsilon} \le 0.8\,. \tag{51}$$

这意味着，以 0.2 的概率，在整个算法迭代过程中，$(1-\epsilon)\,\mathrm{E}_{X\sim G}[1 - w_{k,t}(X)] + \epsilon\,\mathrm{E}_{X\sim B}[w_{k,t}(X)] \le 2.5\epsilon$成立（这已经比我们要证明的 ~ ~引理中的结论稍强一些）。为此，反设结论不成立，并将 (k,t) 定义为首次出现 $(1-\epsilon)\,\mathrm{E}_{X\sim G}[1 - w_{k,t}(X)] + \epsilon\,\mathrm{E}_{X\sim B}[w_{k,t}(X)] > 2.5\epsilon$ 的时刻。结合显然的观察 $(\widetilde{k},\widetilde{t}) \le (k^*, t^*)$，可得 $\Delta_{(\widetilde{k},\widetilde{t})} > 2.5\epsilon$；根据公式 (51)，这种情况发生的概率不可能超过 0.2。

□

B 第 4 节省略的证明第 4 节

现在我们补充第 4 节中省略的细节。具体而言，我们将证明条件 4.2中的确定性条件以高概率成立。本节将重点讨论

成立于 条件 4.2 的第 1 至 4 项 条件 4.2，因为我们已经证明了第 5 和 6 项 第 4.1 节。 第 1、2 和 4 项的证明将紧密遵循 [DKPP22],，而 第 3 项 则是 Chernoff-Hoeffding 界的标准推论。

B.1 条目 1和2

条件 1 和 2 中的 条件 4.2 依赖于这样一个事实：估计量中使用的经验协方差$\prod_{\ell=1}^{p_k}\widehat{\mathbf{B}}_{k,t,\ell}$与 $\mathbf{B}_{k,t}$足够接近，如下文的 引理 B.1 所述。我们先证明该引理，再说明这两个条件如何成立。

引理 B.1。 设 $\delta<1$。考虑 算法 5 并假设 $P_{k,t}$ 支撑于一个 ℓ_2-球内，其半径为 $r\sqrt{d}\|\mathbf{\Sigma}\|_{\mathrm{op}}$。若 $\widehat{W}_{k,t}$ 且 $\widehat{\mathbf{B}}_{k,t,\ell}$ 在 算法 5 中均使用 $\tilde{n}>C\frac{r^2dp^2}{\delta^2}\log\left(\frac{d}{\tau}\right)$个样本计算，其中 C 是一个充分大的常数，则以至少 $1-\tau$的概率，有

$$\left\|\widehat{\mathbf{M}}_{k,t}-\mathbf{M}_{k,t}\right\|_2\le\delta\left\|\mathbf{M}_{k,t}\right\|_{\mathrm{op}}\ ,$$

$\widehat{\ }$ $\widehat{\ }$证明。 令 $\|\mathbf{B}_{k,\,t,\,\ell}-\mathbf{B}_{k,\,t}\|_{\mathrm{op}}\le\delta_1\|\mathbf{B}_{k,\,t}\|_{\mathrm{op}}$和$|W_{k,\,t}-\mathbf{E}_{X\sim P}[w_{k,t}(X)]|\le\delta_2$，其中 δ_1、δ_2待定。$\widehat{\ }$ $\widehat{\ }$估计量 $\mathbf{B}_{k,\,t,\,\ell}$ 与 $\mathbf{M}_{k,\,t}$ 的精度关系如下：

引理 B.2 (引理 *4.14*，见 [DKPP22])。设 $\mathbf{A}$、$\mathbf{B}$、$\mathbf{B}_1$、. . . 、$\mathbf{B}^p$ 为对称 $d\times d$ 矩阵，并定义 $\mathbf{M}=\mathbf{B}^p$、$\mathbf{M}_S=\prod_{i=1}^{p}\mathbf{B}_i$。若 $\|\mathbf{B}_i-\mathbf{B}\|_2\le\delta_1\|\mathbf{B}\|_2$，则 $\|\mathbf{M}_S-\mathbf{B}^p\|_{\mathrm{op}}\le p\delta_1(1+\delta_1)^p\|\mathbf{B}\|_{\mathrm{op}}^p$。

根据引理 B.2，只需确保 $p\delta_1e^{p\delta_1}<\delta$。为此，只需使用 $\delta_1:=\delta/(3p)$。在$\widehat{\ }$的回顾中，我们着重确保：$\|\mathbf{B}_{k,\,t,\,\ell}-\mathbf{B}_{k,\,t}\|_{\mathrm{op}}\le\delta_1\|\mathbf{B}_{k,\,t}\|_{\mathrm{op}}$ 对于该 δ_1 选择。
$\widehat{\ }$令$\|\mathbf{\Sigma}_{k,\,t,\,\ell}-\mathbf{\Sigma}_{k,\,t}\|_{\mathrm{op}}\le\delta_3\|\mathbf{\Sigma}_{k,\,t}\|_{\mathrm{op}}$，其中 δ_3 将稍后指定。我们有

$$\begin{aligned}
\left\|\widehat{\mathbf{B}}_{k,t,\ell}-\mathbf{B}_{k,t}\right\|_{\mathrm{op}}&=\left\|\widehat{W}\widehat{\mathbf{\Sigma}}_{k,t,\ell}-\mathop{\mathbf{E}}_{X\sim P}[w_{k,t}(X)]\mathbf{\Sigma}_{k,t}\right\|_{\mathrm{op}}\\
&\le\mathop{\mathbf{E}}_{X\sim P}[w_{k,t}(X)]\left\|\widehat{\mathbf{\Sigma}}_{k,t,\ell}-\mathbf{\Sigma}_{k,t}\right\|_{\mathrm{op}}+\left|\widehat{W}_{k,t}-\mathop{\mathbf{E}}_{X\sim P}[w_{k,t}(X)]\right|\cdot\left\|\widehat{\mathbf{\Sigma}}_{k,t,\ell}\right\|_{\mathrm{op}}\\
&\le\left\|\widehat{\mathbf{\Sigma}}_{k,t,\ell}-\mathbf{\Sigma}_{k,t}\right\|_{\mathrm{op}}+\delta_2\left\|\widehat{\mathbf{\Sigma}}_{k,t,\ell}\right\|_{\mathrm{op}}\\
&\le(1+\delta_2)\left\|\widehat{\mathbf{\Sigma}}_{k,t,\ell}-\mathbf{\Sigma}_{k,t,\ell}\right\|_{\mathrm{op}}+\delta_2\left\|\mathbf{\Sigma}_{k,t,\ell}\right\|_{\mathrm{op}}\\
&\le(1+\delta_2)\delta_3\left\|\mathbf{\Sigma}_{k,t,\ell}\right\|_{\mathrm{op}}+\delta_2\left\|\mathbf{\Sigma}_{k,t,\ell}\right\|_{\mathrm{op}}\\
&\le2\delta_3\left\|\mathbf{\Sigma}_{k,t,\ell}\right\|_{\mathrm{op}}+\delta_2\left\|\mathbf{\Sigma}_{k,t,\ell}\right\|_{\mathrm{op}}\ .
\end{aligned}$$

为使上述右侧有界于 $\delta_1\|\mathbf{\Sigma}_{k,t,\ell}\|_{\mathrm{op}}$，我们选择 $\delta_3=0.1\delta_1$ 和 $\delta_2=:0.1\delta_1$。通过$\widehat{\ }$霍夫丁不等式，以 $|\widehat{W}_{k,t}-\mathbf{E}_{X\sim P}[w_{k,t}(X)]|\le\delta_2=\delta/(30p)$ 达成目标的样本复杂度为 $O(\frac{p^2}{\delta^2}\log(1/\tau))$。通过事实 2.10，可如下得到达成 $\|\mathbf{\Sigma}_{k,t}-\mathbf{\Sigma}_{k,t}\|_{\mathrm{op}}\le\delta_3\|\mathbf{\Sigma}_{k,t}\|_{\mathrm{op}}$ 的样本复杂度：

$$\tilde{n}\lesssim\frac{r^2d\|\mathbf{\Sigma}\|_{\mathrm{op}}}{\delta_3^2\|\mathbf{\Sigma}_{k,t}\|_{\mathrm{op}}}\log\left(\frac{d}{\tau}\right)\lesssim\frac{r^2dp^2\|\mathbf{\Sigma}\|_{\mathrm{op}}}{\delta^2\|\mathbf{\Sigma}_{k,t}\|_{\mathrm{op}}}\log\left(\frac{d}{\tau}\right)\lesssim\frac{r^2dp^2}{\delta^2}\log\left(\frac{d}{\tau}\right)\ ,$$

其中，最后一个不等式利用了以下事实：根据我们的假设$\mathbf{E}_{X\sim P}[w_{k,t}(X)]\ge1-O(\epsilon)$ 和$\|\mathbf{B}_{k,\,t}\|_{\mathrm{op}}\ge0.5\|\mathbf{\Sigma}\|_{\mathrm{op}}$，我们有 $\|\mathbf{\Sigma}_{k,t}\|_{\mathrm{op}}=\|\mathbf{B}_{k,t}\|_{\mathrm{op}}/\,\mathbf{E}_{X\sim P}[w_{k,t}(X)]\gtrsim\|\mathbf{\Sigma}\|_{\mathrm{op}}$。

□

证明 条件 4.2 的第 1 项和第 2 项。这两个条件可直接由引理 B.1 得到，使用 $\delta = \frac{0.01}{\sqrt{d}}\min\left(\sqrt{\gamma/\epsilon}, 1\right)$。对于第一个条件：由于

$\|\widehat{\mathbf{M}}_{k,t} - \mathbf{M}_{k,t}\|_{\mathrm{op}} \le \frac{0.1}{r\sqrt{d}}\sqrt{\gamma/\epsilon}\|\mathbf{M}_{k,t}\|_{\mathrm{F}}$ 和 $\|x\|_2 \le r\sqrt{d}\|\mathbf{\Sigma}\|_{\mathrm{op}}$，并且我们只关注满足 的点，因此有

$$
\begin{aligned}
g(x) &\le \|\mathbf{M}_{k,t}x\|_2^2 \le 2\left\|\widehat{\mathbf{M}}_{k,t}x\right\|_2^2 + 2\left\|\left(\mathbf{M}_{k,t} - \widehat{\mathbf{M}}_{k,t}\right)x\right\|_2^2 \\
&\le 2\left\|\widehat{\mathbf{M}}_{k,t}x\right\|_2^2 + 2\|x\|_2^2\left\|\mathbf{M}_{k,t} - \widehat{\mathbf{M}}_{k,t}\right\|_{\mathrm{op}}^2 \le 2\widehat{g}(x) + 0.02(\gamma/\epsilon)\|\mathbf{M}_{k,t}\|_{\mathrm{F}}^2\|\mathbf{\Sigma}\|_{\mathrm{op}}\,.
\end{aligned}
$$

等价地，$\widehat{g}(x) \ge 0.5g(x) - 0.01(\gamma/\epsilon)\|\mathbf{M}_{k,t}\|_{\mathrm{F}}^2\|\mathbf{\Sigma}\|_{\mathrm{op}}$。关于条目 2的证明可见[DKPP22]。

□

B.2 项目 4

我们将在下面的引理中证明条目 4。该引理是对 [DKPP22]中引理 4.16 的简单改编。为完整起见，我们附上证明。

条目 4 用于估计 HardThresholdingFilter 第 3 行以及算法 7 第 12 行和算法 6 第 21 行中的 $\mathrm{E}_{X\sim P}[w(x)\tau(x)]$。在给出证明之前，我们简要说明它的具体用法；我们将以算法 7 的第 12 行为例，因为其他情况也可类似验证。为此，我们应用引理 B.3，其中 $w(x) = w_{k,t}(x)\mathbf{1}((u^\top x)^2 \le Q)$。随后，引理给出的保证是（使用第 12 行的符号约定）$|\widehat{\sigma}'_u - \widehat{\sigma}_u| \le 0.01\gamma\widehat{\sigma}_u + \frac{0.01\gamma}{r^2 d}\|u\|_2^2\|\mathbf{\Sigma}_k\|_{\mathrm{op}}$，其中最后一个不等式由以下事实得出：$\|\mathbf{\Sigma}_{k,t}\|_{\mathrm{op}} \le r^2 d\|\mathbf{\Sigma}\|_{\mathrm{op}}$，因为 $P_{k,t}$ 的支撑集位于半径为 $r\sqrt{d}\|\mathbf{\Sigma}\|_{\mathrm{op}}$ 的球内（参见算法 6 第 4 行）。

引理 B.3。在算法 6 的设定中，令 r 为满足 $Pr_{X\sim G}[\|X\|_2 > r\sqrt{d}\|\mathbf{\Sigma}\|_{\mathrm{op}}] \le \epsilon$ 的半径，并假设 $\|\mathbf{B}_{k,t}\|_{\mathrm{op}} \ge 0.5\|\mathbf{\Sigma}\|_{\mathrm{op}}$。对于任意 τ、任意权重函数 $w : \mathbb{R}^d \to [0,1]$ 以及任意向量 u，存在一个针对量 $F_{k,t} := E_{X\sim P}[w(X)(u^\top X)^2]$ 的估计量 $\widehat{F}$，它使用来自 P 的 $n = O\left(\frac{r^4 d^2}{\gamma^2}\log(1/\tau)\right)$ 个样本，运行时间为 $O(nd)$，使用内存 $O(\log(1/\tau))$，并且以概率 $1-\tau$ 满足 $|\widehat{F} - F_{k,t}| \le 0.01\gamma F_{k,t} + \frac{0.01\gamma}{r^2 d}\|u\|_2^2\|\mathbf{\Sigma}_{k,t}\|_{\mathrm{op}}$。

证明。我们证明该引理在成功概率为常数时成立，因为随后可以通过重复执行 $\log(1/\tau)$ 次并取中位数，将其提升至 $1-\tau$。估计量就是样本均值。根据切比雪夫不等式，以常数概率有

$$
\left|\frac{1}{n}\sum_{i=1}^n w(x_i)(u^\top x_i)^2 - \mathop{\mathbf{E}}_{X\sim P}[w(X)(u^\top X)^2]\right| \lesssim \frac{\sqrt{\mathrm{Var}_{X\sim P}[w(X)(u^\top X)^2]}}{n}\,.
$$

我们希望将右端上界为 $0.01\gamma\,\mathbf{E}_{X\sim P}[w(X)(u^\top X)^2] + \frac{0.01\gamma}{r^2 d}\|u\|_2^2\|\mathbf{\Sigma}_{k,t}\|_{\mathrm{op}}$。为此，样本数量只需足够大，即为以下各项的某个足够大的倍数：

$$
\begin{aligned}
\frac{\mathrm{Var}_{X\sim P}[w(X)(u^\top X)^2]}{\left(\gamma\,\mathbf{E}_{X\sim P}[w(X)(u^\top X)^2] + \frac{\gamma}{r^2 d}\|u\|_2^2\|\mathbf{\Sigma}_{k,t}\|_{\mathrm{op}}\right)^2} &\le \frac{r^2 d\,\mathrm{Var}_{X\sim P}[w(X)(u^\top X)^2]}{\gamma^2\,\mathbf{E}_{X\sim P}[w(X)(u^\top X)^2]\|u\|_2^2\|\mathbf{\Sigma}_{k,t}\|_{\mathrm{op}}} \\
&\le \frac{r^2 d\,\mathbf{E}_{X\sim P}[w^2(X)(u^\top X)^4]}{\gamma^2\,\mathbf{E}_{X\sim P}[w(X)(u^\top X)^2]\|u\|_2^2\|\mathbf{\Sigma}_{k,t}\|_{\mathrm{op}}} \\
&\le \frac{r^2 d\|u\|_2^2 r^2 d\|\mathbf{\Sigma}\|_{\mathrm{op}}\,\mathbf{E}_{X\sim P}[w(X)(u^\top X)^2]}{\gamma^2\,\mathbf{E}_{X\sim P}[w(X)(u^\top X)^2]\|u\|_2^2\|\mathbf{\Sigma}_{k,t}\|_{\mathrm{op}}}
\end{aligned}
$$

$$= \frac{r^4 d^2 \|\mathbf{\Sigma}\|_{\mathrm{op}}}{\gamma^2 \|\mathbf{\Sigma}_{k,t}\|_{\mathrm{op}}} \lesssim \frac{r^4 d^2}{\gamma^2} .$$

其中倒数第三行用到了 $w^2(x) \le w(x)$，$(u^\top x)^2 \le \|u\|_2^2 \|x\|_2^2$，以及 $\|x\|_2^2 \le r^2 d \|\mathbf{\Sigma}\|_{\mathrm{op}}$(算法 6 中第 3 行和第 4 行的剪枝）和第四行用到了我们的假设 $\|\mathbf{B}_{k,\,t}\|_{\mathrm{op}} \ge 0.5 \|\mathbf{\Sigma}\|_{\mathrm{op}}$。

□

B.3 项目 3

引理 B.4。对于任意分布 P（定义在 $\mathbb{R}$ 上）以及任意 $\epsilon, \delta \in (0,1)$，存在一个估计量 L，它使用$O\left(\frac{1}{\epsilon} \log\left(\frac{1}{\tau}\right)\right)$ 个来自 D 的样本，并满足

$$\left| \Pr_{X \sim P}[X > \widehat{L}] - \epsilon \right| \le \frac{\epsilon}{100} ,$$

概率至少为 $1-\tau$。该估计量的内存使用量和运行时间也为 $O\left(\frac{1}{\epsilon}\right.$对数 $\left.\left(\frac{1}{\tau}\right)\right)$。

证明。来自 m 个样本 $X_1, \ldots, X_m \sim P$ 的估计量 $L(X_1, \ldots, X_m)$ 定义为第 $(m \cdot \epsilon)$ 大的样本。令 L 表示目标阈值，即满足 $\mathbf{Pr}_{X \sim P}[X > L] = \epsilon$ 的实数。此外，令 $L_{-\epsilon}, L_{+\epsilon}$ 满足 $\mathbf{Pr}_{X \sim P}[L_- \le X \le L] = \mathbf{Pr}_{X \sim P}[L \le X \le L_+] = \epsilon/100$。那么，$L$ 不准确的概率为

$$\Pr_{X_1, \ldots, X_m \sim P}\left[\left| \Pr_{X \sim P}[X > \widehat{L}] - \epsilon \right| > \frac{\epsilon}{100} \right] \le \mathbf{Pr}[\widehat{L} > L_+] + \mathbf{Pr}[\widehat{L} < L_-] .$$

每一项都通过应用乘法形式的 Chernoff 界来进行界定。关于第一个，$L > L_+$ 意味着我们在区间（L_+, $+\infty$）中至少有 $m \cdot \epsilon$ 个样本。然而，这是一个低概率事件，因为每个样本属于该区域的概率仅为 $\epsilon(1 - 1/100)$:

$$\mathbf{Pr}[\widehat{L} > L_+] \le \Pr_{X_1, \ldots, X_m \sim P}\left[\frac{1}{m} \sum_{j=1}^{m} \mathbb{1}(X_j > L_+) - \Pr_{X \sim P}[X > L_+] > \frac{\epsilon}{100} \right] \le e^{-\Omega(\epsilon m)} .$$

选择 m 为 $\epsilon^{-1} \log\left(\frac{1}{\tau}\right)$ 的一个足够大的倍数，使得失败概率小于 $\tau/2$。另一项 $\mathbf{Pr}[L < L_-] < \tau/2$ 也可以用类似的方法证明。

该算法存储 $O(\epsilon^{-1} \log(1/\tau))$ 个一维样本。如果算法先对数据进行排序以选择前$\epsilon \cdot m$ 个元素，其运行时间为 $O(\epsilon^{-1} \log(1/\tau) \log(\epsilon^{-1} \log(1/\tau)))$。然而，使用概率分治算法进行选择也可以达到 $O(\epsilon^{-1} \log(1/\tau))$。

□