arXiv:2603.01951v1 [cs.LG] 2 Mar 2026

广义线性预测的单遍随机梯度下降加速

Qian Chen† Shihong Ding† Cong Fang†

†北京大学

摘要

我们研究流式设定下的广义线性预测，其中每次迭代仅使用一个新数据点进行梯度级更新。尽管动量在确定性优化中已得到充分确立，但一个根本性的开放问题是，它能否加速这种单遍非二次随机优化。我们提出了首个通过一种新颖的数据依赖近端方法成功融入动量、实现双动量加速的算法。我们推导的超额风险界分解为三个部分：得到改善的优化误差、达到极小极大最优的统计误差，以及高阶的模型误设误差。证明通过对内层更新进行细致的驻点分析来处理模型设定错误，并通过两阶段外循环分析将统计误差局部化。因此，我们解决了Jain 等人 [2018 提出的] 开放问题，并证明在流式设定下，对于广义线性预测，动量加速比方差缩减更有效。

Key关键词：动量加速；数据依赖近端方法；单遍随机梯度下降

1 引言

本文研究以下形式的广义线性预测：

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) = \mathbb{E}_{(\mathbf{a},b)\sim\mathcal{D}}\ \ell(\mathbf{a}^\top\mathbf{x}, b), \tag{GLP}$$

其中，目标函数 F 是对线性投影数据上的期望凸损失 ℓ 进行最小化 $\mathbf{a}$，其中（$\mathbf{a}$ b）取自某个底层分布 $\mathcal{D}$。这一问题在机器学习中无处不在，并在各个领域得到广泛应用。形式上，它涵盖了广义线性模型（GLM）的最大似然估计（MLE）；通过对损失 ℓ作特定选择，即可得到线性回归和逻辑回归等基础学习问题。我们在大规模流式设定下研究这一问题，在该设定中，算法每次迭代仅允许对一个新鲜的数据点进行梯度级 $\mathcal{O}(d)$ 计算。我们主要关注降低近似求解（GLP）所需的数据和运行时间复杂度。我们主要研究满足以下条件的损失函数：关于第一个参数 ℓ 强凸且梯度利普希茨连续，并将 α 记为损失函数的条件数（定义见第3.1 节）。

与我们受约束的流式算法进行直接比较的基准，是那些不受计算和内存限制的算法。其中一种常用方法是计算如下形式的经验风险最小化问题：

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}} \frac{1}{N}\sum_{i=1}^{N} \ell(\mathbf{a}_i^\top\mathbf{x}, b). \tag{1}$$

对于 (1)，可以通过著名的局部化 Rademacher 技术 [Bartlett 等人，2005 年，Wainwright，2019]，建立超额风险 $F(\hat{x}) - F(x^*)$ 的非渐近样本复杂度 $\frac{\alpha\sigma_0^2}{\varepsilon}$，其中 x^* 表示 (GLP) 的总体最小化解，σ_0^2 与 ℓ 的 Lipschitz 连续性有关。样本复杂度至少为 $\frac{\alpha\sigma_*^2}{\varepsilon}$，其中 $\sigma_*^2 = \mathrm{tr}(H^{-1}Q)$，Q 是 x^* 处的随机梯度协方差，H 是 $\nabla^2 F$ 的一个上界（定义见第 3.2 节）。我们在附录 G 中证明，当对 ℓ 仅施加强凸性和梯度 Lipschitz 连续性时，该样本复杂度是最优的。

转向在线设定，一个自然的问题是：在我们的流式约束下，还需要多少额外数据？需要指出的是，标准随机梯度下降（SGD）在求解（GLP）时效率低下，因为它未能利用问题结构。对于线性回归（其中 $\alpha = 1$），已知某些随机梯度下降变体确实能够改善样本复杂度。例如，在设定正确的模型下，采用尾部平均的随机梯度下降 [Jain 等人，2018b、Zou 等人，2023] 以及采用指数步长衰减的随机梯度下降 [Ge 等人，2019、Wu 等人，2022]，的样本复杂度分别为 $\kappa\log(1/\varepsilon) + \frac{\sigma_*^2}{\varepsilon}$。有趣的是，该复杂度可分解为最优统计项和一个优化项；后者仅随目标精度呈对数增长，并由 κ 缩放。这里，κ 表示数据分布的条件数（详见第 3.1 节）。只要数据满足 $\|a\|^2 \leqslant L$，它就可以被界定为 $\kappa \leqslant L/\mu$；而 μ 是数据协方差矩阵 $\Sigma \stackrel{\mathrm{def}}{=} \mathbb{E}_{a\sim\mathcal{D}}[aa^\top]$的最小特征值。

近年来，已有多项研究 [Frostig 等人、2015年、Li 等人，2022] 尝试在一般设定下求解（GLP），并得到了复杂度界，该界分解为统计项和由流式约束所导致的优化项。这些算法的核心思想是采用方差缩减技术，将梯度方差分解为最优点处的方差和可通过优化控制的误差分量，从而局部化统计误差。然而，这些结果的一个共同局限是，优化复杂度依赖于问题的条件数 $\alpha^2\kappa$。优化复杂度刻画了从初始误差收敛所需的时间。具体而言，除非样本量超过 $\alpha^2\kappa$，否则超额风险不会降低到非平凡的水平（即变为 $o(1)$）。

在确定性优化中，众所周知，Nesterov 加速等动量技术 [Nesterov、1983] 以及重球法 [Polyak，1964] 能够在不同设定下加速收敛。然而，如何将动量有效地融入我们的流式优化问题，在很大程度上仍是一个开放问题。

对于一般的随机凸问题，普遍观点认为动量可能无法带来有效加速。例如，Agarwal 等人 [2012] 证明了在具有梯度方差的一般强凸目标函数下，随机梯度下降的最优性。相比之下，对于设定正确的线性回归——这是一种结构化估计问题——Jain 等人 [2018a] 表明，在保持最优统计复杂度的同时，确实可以加速优化项。具体而言，Jain 等人 [2018a] 引入了数据的统计条件数，记作 $\tilde{\kappa}$（详见第3.1节），它严格小于 κ，并将优化复杂度提升至$\sqrt{\kappa\tilde{\kappa}}$。他们的工作留下了两个开放问题：（i）将分析扩展至允许模型设定错误的情形；（ii）探索更一般的估计场景，在这些场景中动量仍然有效。本文通过提出首个成功将动量应用于广义线性预测的算法，解决了这两项挑战，且无需依赖固定的海森矩阵结构或模型设定。

1.1 回顾：已有结果

设定正确的线性回归。使随机梯度下降类算法达到最优统计复杂度并附加一个优化项的核心思想，是将随机梯度下降建模为随机过程，并证明变量在分布意义下收敛 [Dieuleveut and Bach，2016，Jain 等人，2018b，Ge 等人，2019，Zou 等人，2023，Wu 等人，2022年，Li 等人，2024]。这种收敛分析自然得出优化复杂度，而对平稳分布的方差计算则刻画了统计误差。Jain 等人[2018a]改进了优化项，并达到了

$$\sqrt{\kappa\tilde{\kappa}}+\frac{\sigma_*^2}{\varepsilon},$$

通过采用动量加速和更精细的分析，忽略对数因子。然而，这一结果仅限于设定正确的线性回归，因为其平稳分布分析严重依赖目标函数的二次形式，以及梯度噪声协方差与问题海森矩阵的可交换性。

广义线性预测的方差缩减。 Frostig 等人 [2015] 和 Li 等人[2022] 研究了一个稍具一般性的随机凸优化问题，该问题不要求各个函数具有线性结构。他们要求目标函数 F 是 μ_h-强凸的，并且每个分量均满足 L_h-梯度利普希茨连续；此外，他们还关键性地施加了一个针对广义线性预测的海森矩阵支配条件 $\nabla^2 F(\mathbf{x}^*) \preceq \alpha_h \nabla^2 F(\mathbf{x})$。将他们的结果调整到我们的设定后，Frostig 等人 [2015] 达到了如下复杂度[1]：

$$\alpha^2\kappa+\frac{\alpha^2\sigma_*^2}{\varepsilon}, \tag{2}$$

其中，$\alpha^2\kappa$ 对应优化成本，而 $\frac{\alpha^2\sigma_*^2}{\varepsilon}$ 是统计样本复杂度。Li 等人 [2022]得到的样本复杂度为[2] 的

$$\alpha^2\kappa\cdot\frac{\lambda_{\max}(\boldsymbol{\Sigma})}{\lambda_{\min}(\boldsymbol{\Sigma})}+\frac{\alpha\sigma_*^2}{\varepsilon}+\mathcal{O}\left(\frac{1}{\varepsilon^{2/3}}\right). \tag{3}$$

这项工作需要额外的海森矩阵光滑性假设，以进一步降低统计复杂度。此外，上面的最后一项隐藏了依赖于问题的因子，包括海森光滑性参数。对于（GLP），方差缩减是否从根本上不可或缺，以及如何改善对优化复杂度的依赖，仍是开放问题。

1.2 我们的结果与启示

本文针对流式设定中的（GLP）提出动量加速。我们算法的主要思想在于一种数据依赖的近端方案，其中近端算子被显式地构-

[1]Frostig 等人 [2015] 在推论4中得到样本复杂度 $\alpha_h\kappa_h+\frac{\alpha_h}{\varepsilon}\mathrm{tr}((\nabla F(\mathbf{x}^*))^{-1}\mathbf{Q})$，其中 $\kappa_h=L_h/\mu_h$。在最坏情况下，有 $\alpha_h=\alpha$、$\kappa_h=\alpha\kappa$ 以及 $\mathrm{tr}((\nabla F(\mathbf{x}^*))^{-1}\mathbf{Q})=\alpha\sigma_*^2$，这就是式（2）中的复杂度。[2]Li 等人 [2022] 在推论3中得到样本复杂度 $\kappa_h+\frac{\ell_\Xi^2}{\mu_h^2}+\mathrm{tr}((\nabla F(\mathbf{x}^*))^{-1}\mathbf{Q})_{\varepsilon}+\mathcal{O}(\varepsilon^{-2/3})$。在最坏情况下，有 $\kappa_h=\alpha\kappa$、$\ell_\Xi^2/\mu_h^2=\alpha^2\kappa$。$\lambda\max(\boldsymbol{\Sigma})\,\lambda\min(\boldsymbol{\Sigma})$以及 $\mathrm{tr}((\nabla F(\mathbf{x}^*))^{-1}\mathbf{Q})=\alpha\sigma_*^2$，这就是式（3）中的复杂度。

造自期望数据协方差 $\boldsymbol{\Sigma}$。这使得双重动量加速可同时应用于内循环和外循环，并且我们得到的复杂度为

$$\left(\sqrt{\alpha\kappa\tilde{\kappa}}+\alpha^2\tilde{\kappa}\right)+\frac{\alpha\sigma_*^2}{\varepsilon}+\left(\frac{\alpha^2\tilde{\kappa}^2\,\mathrm{tr}\,\mathbf{Q}}{L_\ell\mu\varepsilon}\right)^{1/3}.$$

上述超额风险界由三个可解释的部分组成：(i) 优化项对问题条件数和数据条件数的依赖有所改善，其中第一项来自双重加速方案，第二项则体现了噪声的影响；(ii) 估计项与最优统计风险相匹配；(iii) 高阶模型设定错误项，用于刻画受限计算与模型设定错误之间的耦合效应，我们对其进行了更精细的刻画。我们的框架可扩展到弱凸目标，并能自然地适应无标签数据的使用、小批量处理和并行实现。

我们对算法的分析由两部分组成：(i) 在内循环中，子问题类似于线性回归，但存在模型设定错误。我们对模型设定错误的影响进行了细致刻画；(ii) 外循环要求我们将统计误差局部化到极小点。我们提出了一种两阶段分析，表明在我们的问题中并不需要方差缩减。

因此，我们解决了Jain 等人 [2018提出的]开放问题。我们还证明，对于流式广义线性预测，动量加速比方差缩减更高效。这一发现与流式设定下非凸优化的已知结果截然不同。在那种情况下，目标是找到一个近似驻点，而即使采用海森矩阵光滑性假设，动量也无法改善普通随机梯度下降的最坏情况收敛率 [Fang 等人，2019，Jin 等人，2018]。然而，方差缩减方法（如随机方差缩减梯度）改进了 $\varepsilon^{-3.5}$至 $\varepsilon^{-10/3}$ [Allen-Zhu 和 Hazan，2016，Reddi 等人，2016]，采用 SPIDER [Fang 等人，2018] 和 SARAH [Nguyen 等人，2017]达到了最优的 ε^{-3}。我们希望本研究能为设计适用于更广泛凸问题和非凸问题的加速流式方法带来启发。

记号。对于向量 $\mathbf{x}$，令 $\|\mathbf{x}\|$表示二范数；对于对称半正定矩阵 $\mathbf{A}$，$\|\mathbf{x}\|_{\mathbf{A}}=\sqrt{\mathbf{x}^\top\mathbf{A}\mathbf{x}}$ 表示诱导范数。令 $\|\mathbf{A}\|$ 表示 $\mathbf{A}$ 的谱范数。对于半正定矩阵 $\boldsymbol{\Sigma}$，我们分别用 $\lambda_{\max}(\boldsymbol{\Sigma})$ 和 $\lambda_{\min}(\boldsymbol{\Sigma})$ 表示 $\boldsymbol{\Sigma}$ 的最大特征值和最小特征值。将 $\kappa(\boldsymbol{\Sigma})=\lambda_{\max}(\boldsymbol{\Sigma})/\lambda_{\min}(\boldsymbol{\Sigma})$ 定义为 $\boldsymbol{\Sigma}$ 的条件数。我们将均值为 $\boldsymbol{\mu}$、协方差为 $\boldsymbol{\Sigma}$ 的高斯分布记作 $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$。对于两个非负序列 $\{a_n\}$ 和 $\{b_n\}$，若存在普适常数 $C>0$，使得对所有充分大的 n 都有 $a_n\le Cb_n$（相应地，$a_n\ge Cb_n$），则记作 $a_n=\mathcal{O}(b_n)$（相应地，$a_n=\Omega(b_n)$）；若两者同时成立，则我们~~称~~ $a_n=\Theta(b_n)$。我们使用 $\mathcal{O}(\cdot)$、$\Omega(\cdot)$ 和 $\Theta(\cdot)$ 来隐藏对数因子。若 $a_n=\mathcal{O}(b_n)$（相应地，$a_n=\Omega(b_n)$），则记作 $a_n\lesssim b_n$（相应地，$a_n\gtrsim b_n$）。

2 相关工作

随机逼近。 随机逼近方法可追溯至 Robbins 和 Monro [1951]的开创性工作。此后，一系列研究针对凸和强凸设定下的随机梯度下降，推导了非渐近收敛速率，并分析了步长调度和平均方案的影响 [Moulines 和 Bach，2011年，Bach 和 Moulines，2013，D´efossez 和 Bach，2015年，Dieuleveut 和 Bach，2016，Ge 等人，2019，Zhang 等人，2025年，Sheshukova

等人， 2025]。在线性回归设定下，更精细的非渐近分析表明，迭代平均的随机梯度下降可达到最优统计速率 $\sigma^2 d/n$ [Bach 和 Moulines，2013，Jain 等人，2018b，Dieuleveut 等人，2017，Jain 等人，2018a]。然而，这类最优速率通常无法推广到二次目标函数之外。近期研究提出了能够达到此类最优统计速率的算法 [Frostig 等人，2015，Li 等人， 2022]。具体而言，流式随机方差缩减梯度法（Streaming SVRG） [Frostig 等人，2015] 将方差缩减的思路应用于流式设定，并达到与经验风险最小化（ERM）相当的统计性能。ROOT-SGD [Li 等人， 2022] 实现了非渐近速率，且其渐近速率达到 Cram´ér-Rao 下界。这些分析依赖于自协调性或海森光滑性等三阶连续条件，而我们只需要二阶条件。

动量加速与方差缩减。动量方法源于重球法 [Polyak、 1964] 以及 Nesterov 的加速梯度下降 [Nesterov， 1983],这些方法实现了加速的最优收敛速率。在流式设定中，动量的作用更为微妙，因为与 SGD 相比，动量通常并不能改善最坏情况下的统计速率 [Agarwal 等人， 2012]。然而，在线性回归中，一些工作通过利用二次结构，分别实现了优化误差的加速收敛和统计误差的最优速率 [Jain 等人，2018a，Dieuleveut，2017，Pan，2024 年，Li 等人，2024 年，Zhang，2024 年，Liu 等人， 2025]。不过，这些分析高度依赖矩阵计算，而将其推广到一般目标函数并非易事。对于非凸问题，Fang 等人 [2019] 证明，SGD 找到二阶驻点的复杂度为 $\mathcal{O}(\varepsilon^{-3.5})$，与基于动量的方法相匹配 [Jin， 2018]。方差缩减技术通过利用问题结构来减少随机噪声，从而加速随机优化。对于凸问题，在有限和设置中，SDCA [Shalev-Shwartz 和 Zhang， 2013]考虑了与 (GLP) 相同的问题。一系列工作开发了采用动量加速的各种算法 [Shalev-Shwartz 和 Zhang，2014，Lin 等人，2018年，Allen-Zhu， 2018]。在流式设定中，SVRG [Frostig 等人，2015] 和 ROOT-SGD [Li 等人，2022] 改善了统计项，但未能加速优化项。对于非凸问题，方差缩减技术已被证明在有限和设置与流式设定中均有益处。具体而言，在流式设定中，SVRG 达到 $\varepsilon^{-10/3}$ [Allen-Zhu 和 Hazan，2016，Reddi 等人， 2016], 的复杂度，而递归方差缩减方法（如 SARAH） [Nguyen 等人，2017], 达到了 ε^{-3}的最优速率。

3 问题设定

我们将广义线性预测重述如下：

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \stackrel{\text{def}}{=} \mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\, \ell(\mathbf{a}^\top\mathbf{x}, b). \tag{GLP}$$

3.1 假设

我们作出以 下关于目标函数（GLP）和数据分布的假设 $\mathcal{D}$。

假设 1 （损失函数的条件数）。令 $\ell''(\cdot,\cdot)$ 表示关于其第一个参数的二阶导数。假设存在 L_ℓ, $\mu_\ell > 0$，使得 $\mu_\ell \leqslant \ell''(\mathbf{a}^\top\mathbf{x}\ b) \leqslant L_\ell$。令 $\alpha \stackrel{\text{def}}{=} L_\ell/\mu_\ell$ 为损失函数 ℓ 的条件数。

假设 1 意味着 ℓ 是 L_ℓ-光滑且 μ_ℓ-强凸，这在优化文献中被广泛采用 [Nesterov， 2018]。这一假设也常用于广义线性模型的分析，例如借助局部拉德马赫复杂度等工具 [Bartlett 等人，2005 年，Wainwright， 2019]。

假设 2 （数据二阶矩条件）。 我们假设二阶矩矩阵 $\mathbf{\Sigma} \xlongequal{\text{定义为}} \mathbb{E}\mathbf{a}\mathbf{a}^\top$ 存在且有限。此外，我们假设 $\mathbf{\Sigma}$ 为正定矩阵，其最小特征值为 $\mu \xlongequal{\text{定义为}} \lambda_{\min}(\mathbf{\Sigma}) > 0$。

假设 2 是一个温和的正则性条件。结合假设 1，它意味着目标函数 F 是 $L_\ell \lambda_{\max}(\mathbf{\Sigma})$ 光滑且 $\mu_\ell \mu$-强凸。我们将在第7节将结果扩展到弱凸情形。

假设 3 （数据四阶矩条件）。 我们假设四阶矩满足以下有界性条件：

$$\mathbb{E}_{\mathbf{a}\sim\mathcal{D}}\left(\|\mathbf{a}\|^2\mathbf{a}\mathbf{a}^\top\right) \preceq R^2\mathbf{\Sigma}, \quad \mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\|\mathbf{a}\|^2_{\mathbf{\Sigma}^{-1}}\mathbf{a}\mathbf{a}^\top\right) \preceq \tilde{\kappa}\mathbf{\Sigma}.$$

假设 3 常用于线性回归中随机梯度下降的分析 [Jain 等人，2018a、Zou 等人，2023、Wu 等人，2022、Li 等人， 2024]。量 R 可以看作数据的有效半径 $\sim \mathcal{D}$。具体而言，如果 $\|\mathbf{a}\| \leqslant L$ 几乎处处成立，则 $R = L$。量 $\tilde{\kappa}$也可以类似地理解。在我们的设定中， R^2 和 $\tilde{\kappa}$ 分别刻画梯度噪声的二范数和 $\mathbf{\Sigma}^{-1}$-范数。我们将 $\kappa = R^2/\mu$ 称为 $\mathcal{D}$ 的条件数，并沿用 D´efossez 和 Bach [2015], Jain 等人。 [2018a]的术语，将 $\tilde{\kappa}$称为统计条件数。注意，我们始终有 $\tilde{\kappa} \leqslant \kappa$，因为 $\mathbb{E}(\|\mathbf{a}\|^2_{\mathbf{\Sigma}^{-1}}\mathbf{a}\mathbf{a}^\top) \preceq \frac{1}{\mu}\mathbb{E}_{\mathbf{a}\sim\mathcal{D}}(\|\mathbf{a}\|^2\mathbf{a}\mathbf{a}^\top) \preceq \kappa\mathbf{\Sigma}$ [Jain 等人，2018a]。

在线性回归设定中，Jain 等人 [2018a] 将优化复杂度从 κ改进到 $\sqrt{\kappa\tilde{\kappa}}$，其中 $\tilde{\kappa} \leqslant \kappa$ 对所有分布都成立。当 $\mathbf{\Sigma}$ 条件较差时，即 $\tilde{\kappa} \ll \kappa$。统计条件 $\tilde{\kappa}$ 是必要的，因为 $\tilde{\kappa}$ 控制着经验协方差矩阵向 $\mathbf{\Sigma}$ 的集中，并且在某种意义上，它刻画了无噪声标签设定下的统计复杂度 [Jain 等人，2018a]。我们建议读者参阅Jain 等人 [2018a] ，以了解关于 $\tilde{\kappa}$必要性的更多讨论。

注1（加速的影响）。在高斯设计$\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$中，可以取 $\kappa = 3\operatorname{tr}\mathbf{\Sigma}/\mu$和 $\tilde{\kappa} = 3d$。于是 $\kappa \geqslant \tilde{\kappa}$。更一般地，如果 $\mathbf{\Sigma}^{-1/2}\mathbf{a}$ 为 σ_a^2-次高斯，则该条件成立，可取$\kappa = 16\sigma_a^4\operatorname{tr}\mathbf{\Sigma}/\mu$和 $\tilde{\kappa} = 16\sigma_a^4 d$ [Zou 等人， 2023]。在最坏情况下，$\operatorname{tr}\mathbf{\Sigma}/\mu = \kappa(\mathbf{\Sigma})\tilde{\kappa}$。

3.2 问题相关量总结

我们汇总这些量，并给出整体复杂度的目标。记 $\mathbf{H} = L_\ell\mathbf{\Sigma}$ 为目标函数海森矩阵 $\nabla^2 F$ 的一个上界。令 $\mathbf{x}^*$ 表示 F 的极小化点，我们将极小化点$\mathbf{x}^*$ 处梯度噪声的二阶矩阵定义为

$$\mathbf{Q} = \mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\ell'(\mathbf{a}^\top\mathbf{x}^*, b)\right)^2\mathbf{a}\mathbf{a}^\top.$$

- 对于优化复杂度，目标是改善对损失条件数的依赖 α 以及数据条件数的依赖 κ， $\tilde{\kappa}$。特别是，我们希望改善对 κ的依赖。如备注1所示，它可能是 $\kappa(\mathbf{\Sigma}$的 $\tilde{\kappa}$倍。

算法 1 随机加速数据依赖算法（SADA）

输入：初始化 $\tilde{\mathbf{x}}_0$, 正则化参数 $\{h_k\}_{k=1}^K$、步长 η, γ 和动量参数 $\{\beta_k\}_{k=1}^K$, θ, $\tilde{\mathbf{x}}_{-1} = \tilde{\mathbf{x}}_0$, 对于 $k = 1, 2, \ldots, K$, 执行$\tilde{\mathbf{y}}_{k-1} \leftarrow \tilde{\mathbf{x}}_{k-1} + \beta_k(\tilde{\mathbf{x}}_{k-1} - \tilde{\mathbf{x}}_{k-2})$ ▷ 外推$\tilde{\mathbf{x}}_0 \leftarrow \mathbf{y}_{k-1}, \mathbf{z}_0 \leftarrow \mathbf{y}_{k-1}$对于 $t = 1, 2, \ldots, T$, 执行 ▷ 求解子问题 (4) 的内循环 采样新数据 $(\mathbf{a}_t, b_t) \sim \mathcal{D}\mathbf{y}_{t-1} \leftarrow \frac{1}{1+\theta}\mathbf{x}_{t-1} + \frac{\theta}{1+\theta}\mathbf{z}_{t-1}\hat{\mathbf{g}}_t \leftarrow h_k\ell'(\mathbf{a}_t^\top \tilde{\mathbf{y}}_{k-1}, b_t)\mathbf{a}_t + \left[\mathbf{a}_t^\top(\mathbf{y}_{t-1} - \tilde{\mathbf{y}}_{k-1})\right]\mathbf{a}_t$ ▷ 通过流式数据计算 $\hat{\mathbf{g}}_t$ $\hat{\mathbf{x}}_t \leftarrow \mathbf{y}_{t-1} - \eta\mathbf{g}_t\mathbf{z}_t \leftarrow \theta\mathbf{y}_{t-1} + (1-\theta)\mathbf{z}_{t-1} - \gamma\hat{\mathbf{g}}_t$结束循环$\tilde{\mathbf{x}}_k \leftarrow \frac{2}{T}\sum_{t=T/2+1}^{T}\mathbf{x}_t$ ▷ 尾部平均方案 结束循环 返回 $\tilde{\mathbf{x}}_K$

—— ——

–

- 对于统计复杂度，我们的目标是保留 最优速率 $\alpha_{\mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q})\,\varepsilon}$。在正确设定的线性回归模型 $b = \mathbf{a}^\top\mathbf{x}^* + \varepsilon_{\text{噪声}}$中，该表达式可简化为 $\frac{\sigma^2 d}{\varepsilon}$，其中 σ^2 是噪声 $\varepsilon_{\text{噪声}}$的方差。 ——

- 额外的复杂度可能源于优化过程与模型设定错误之间的耦合。我们预计，这种复杂度对 ε^{-1} 和 κ的依赖程度较低（即作为风险率中的高阶项出现）。

4 随机加速数据依赖算法

我们提出 **随机加速数据依赖算法（SADA）**来求解（GLP），如算法 1所示。SADA结合动量与数据依赖近端方法，在流式设定下实现加速。总体而言，外循环通过迭代构造数据依赖近端子问题。近端项由数据协方差诱导，无法显式获取。内循环利用流式数据近似近端项，从而近似求解子问题，并通过对最后一半迭代结果进行尾部平均，返回所得解。

4.1 内循环：带尾部平均的加速求解器

内循环采用动量来加速优化误差的收敛，并采用尾部平均方案来降低方差。在第 k 次外层迭代中，令 $\tilde{\mathbf{y}}_{k-1}$ 表示外推变量。内循环求解以下依赖数据的近端子问题：

$$\min_{\mathbf{x}\in\mathbb{R}^d} \mathbb{E}_{\mathbf{a},b\sim\mathcal{D}} h_k\langle\ell'(\mathbf{a}^\top\tilde{\mathbf{y}}_{k-1}, b)\mathbf{a}, \mathbf{x} - \tilde{\mathbf{y}}_{k-1}\rangle + \frac{1}{2}\|\mathbf{x} - \tilde{\mathbf{y}}_{k-1}\|_{\mathbf{\Sigma}}^2, \tag{4}$$

其中 $\ell'(\cdot,\cdot)$ 表示 ℓ 关于其第一个自变量的一阶偏导数。

由于总体协方差 $\mathbf{\Sigma}$ 无法直接获得，内循环使用新样本（$\mathbf{a}$ b）$\sim \mathcal{D}$ 来近似总体协方差 $\mathbf{\Sigma}$，即通过 $\mathbf{a}\mathbf{a}^\top$。这类似于使用数据对（$\mathbf{a}$ $-h_k\ell'$（$\mathbf{a}^\top \tilde{\mathbf{y}}_{k-1}$， b））的线性回归。然而，将会出现模型设定错误，且其分析并非易事。以往研究仅关注设定正确的情形。为填补这一空白，我们提出逐层剥离分解方法来研究其平稳分布。具体细节见第6.1节。

4.2 O外循环：带加速的数据依赖近端方法 $\mathbf{n}$

外循环迭代构造近端子问题（4），并利用动量加速收敛。由于加速往往会放大误差，我们需要仔细控制内循环的优化精度。SADA采用两阶段步长，以实现加速的优化收敛速率并控制随机噪声。第一阶段使用较大的常数步长以快速降低优化误差，第二阶段使用逐渐衰减的步长以控制随机噪声。

5 收敛结果与算法复杂度

In 本节给出SADA的收敛速率与复杂度分析 .

超参数选择。 我们如下选择内循环超参数 s:

$$\eta \leqslant \frac{1}{16R^2}, \quad \gamma = \frac{1}{4}\sqrt{\frac{\eta}{\tilde{\kappa}\mu}}, \quad \theta = \frac{1}{4}\sqrt{\frac{\mu\eta}{\tilde{\kappa}}}, \quad T \geqslant \tilde{\Omega}\left(\sqrt{\frac{\tilde{\kappa}}{\mu\eta}}\right). \tag{5}$$

对于外循环的超参数，我们设置步长 $h_k = 2\alpha\tilde{\theta}_k^2/L_\ell$，以及动量参数$\tilde{}$ $\tilde{}$ $\beta_k = (1-\theta_k)/(1+\theta_k)$，其中

$$\tilde{\theta}_{\max} \asymp \min\left\{\frac{1}{\sqrt{\alpha}}, \frac{T}{\alpha^2\tilde{\kappa}}\right\}, \quad \tilde{\theta}_k = \begin{cases} \tilde{\theta}_{\max}, & k \leqslant K/2, \\ \dfrac{4}{4/\tilde{\theta}_{\max} + k - K/2}, & k > K/2. \end{cases} \tag{6}$$

备注 2。在上述设定中，正则化参数 h_k 可视为外循环的步长；下文将 h_k 称为步长。为保证收敛，η 必须小于$1/R^2$； γ 和 θ 是加速超参数，最初分别用于随机坐标下降（ [Nesterov， 2012]）和线性回归（ [Jain 等人，2018a；Li 等人，2024]）。我们要求内循环迭代次数超过 $T \geqslant \tilde{\Omega}(\sqrt{\tilde{\kappa}}/(\mu\eta))$，以确保内循环充分收敛。在外循环中，将 (6) 中的 $\tilde{\theta}_k$ 代入 $h_k = 2\alpha\tilde{\theta}_k^2/L_\ell$后，可知步长 h_k 受光滑性参数的逆矩阵 $1/L_\ell$ 以及由额外噪声引入的附加项 $\frac{T^2}{L_\ell\alpha^3\tilde{\kappa}^2}$ 所约束。在第一阶段（$k \leqslant K/2$），我们选择最大步长，以加速分布收敛；在第二阶段（$t > K/2$），我们谨慎地衰减步长，以控制噪声。

备注3。在备注1中的高斯和次高斯情形下，内循环超参数为 $\eta \lesssim 1/\mathrm{tr}\ \mathbf{\Sigma}$, $\gamma \asymp \sqrt{\eta/(\mu d)}$, $\theta \asymp \sqrt{\mu\eta/d}$和 $T \gtrsim \tilde{\Omega}(1/(\mu\gamma))$。对于 $\mathrm{tr}\,\mathbf{\Sigma}/\mu = \kappa(\mathbf{\Sigma})d$ 的最坏情况，步长 γ 比 η 大 $\sqrt{\kappa(\mathbf{\Sigma})}$ 倍，其中 η 也是普通随机梯度下降的步长 [（Jain 等人，2018b，Zou 等人， 2023]）。这表明收敛速率可加速 $\sqrt{\kappa(\mathbf{\Sigma})}$ 倍。

定理 1。 假设 *1*、 *2* 和 *3* 均成立。令算法*1* 从 $\tilde{\mathbf{x}}_0$, 开始，并按照上文 (5)和 (6)中指定的方式选择超参数。令 $\tilde{\mathbf{x}}_K$ 表示算法 *1* 经过 K 次外层迭代后的输出，每次迭代包含 $T \geqslant \tilde{\Omega}(\sqrt{\tilde{\kappa}}/(\mu\eta))$ 次内层更新。于是有

$$\mathbb{E}F(\tilde{\mathbf{x}}_K) - F(\mathbf{x}^*) \lesssim \exp\left(-\frac{c_0 K}{\sqrt{\alpha} + \alpha^2\tilde{\kappa}/T}\right)\left(F(\tilde{\mathbf{x}}_0) - F(\mathbf{x}^*)\right) + \frac{\alpha\,\mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q})}{n} + \frac{\alpha\eta\tilde{\kappa}\,\mathrm{tr}\,\mathbf{Q}}{L_\ell n},$$

其中， $n = KT$ 为样本量， c_0 为普适常数。

推论 1 （样本复杂度）。 在定理 *1*的设定下，要获得一个超额风险至多为 $\tilde{\mathbf{x}}_K$ 的解，所需的样本量 $n = KT$ 为

$$\tilde{\mathcal{O}}\Bigg(\underbrace{\left(\sqrt{\alpha\kappa\tilde{\kappa}} + \alpha^2\tilde{\kappa}\right)}_{\textit{Optimization term}} + \underbrace{\frac{\alpha\,\mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q})}{\varepsilon}}_{\textit{Statistical term}} + \underbrace{\left(\frac{\alpha^2\tilde{\kappa}^2\,\mathrm{tr}\,\mathbf{Q}}{L_\ell\mu\varepsilon}\right)^{1/3}}_{\textit{Mis-specification term}}\Bigg).$$

我们将超参数选为 $\tilde{\Theta}\left(\frac{\alpha\tilde{\kappa}}{\mu n^2}\right) \leqslant \eta \leqslant \tilde{\Theta}\left(\min\left\{\frac{1}{R^2}, \frac{\alpha\tilde{\kappa}}{\mathrm{tr}\,\mathbf{Q}}\right\}\right)$, $T = \tilde{\Theta}(\sqrt{\tilde{\kappa}}/(\mu\eta))$, $K = n/T$ 其他超参数则按照 (5) 和 (6)中的规定进行选择。

备注 4。 样本复杂度要求足以保证存在一个满足上述条件的步长 η （见附录E.3）。请注意，此 η 是针对有限时域设定选择的，因为它取决于预先已知的样本量 n。利用倍增技巧，可以将这一选择扩展至无限时域设定（算法可在任意时刻停止） [Hazan 和 Kale， 2014], 且只会带来对数级开销。此外，选择超参数需要预先知道一些依赖于问题的参数，这在优化方法中很常见 [Nesterov， 2018]。通过重启 [O'donoghue 和 Candes，2015年，Necoara 等人， 2019]或线搜索程序 [Bubeck 等人， 2015]或许可以消除这种依赖。我们将进一步研究留待未来工作。

优化项。 优化复杂度由两项组成。第一项 $\sqrt{\alpha\kappa\tilde{\kappa}}$对应于初始误差的加速衰减。该收敛速率体现了双重加速效应，这是由算法 1的外循环和内循环中的动量机制共同引起的。第二项 $\alpha^2\tilde{\kappa}$ 源于局部化所带来的额外噪声，其大小与距离 $\|\tilde{\mathbf{y}}_k - \mathbf{x}^*\|_{\mathbf{\Sigma}}^2$成正比。我们将方差缩减方法中先前的 $\alpha^2\kappa$ 复杂度改进为 $\alpha^2\tilde{\kappa}$。

作为一个代表性示例，考虑备注1 中的高斯情形，并假设外层条件数为 $\alpha \asymp 1$ （例如，具有归一化特征和常数水平最优解的线性回归或逻辑回归）。在这种情形下，算法 1 达到如下阶的优化复杂度：$\mathcal{O}((d\mathrm{tr}\ \mathbf{\Sigma}/\mu)^{1/2}$，优于未加速速率 $\mathcal{O}(\mathrm{tr}\ \mathbf{\Sigma}/\mu)$（注意 $d \leqslant \mathrm{tr}\,\mathbf{\Sigma}/\mu$）。最坏情况下，$\mathbf{tr}\ \mathbf{\Sigma}/\mu = \kappa(\mathbf{\Sigma})d$，因此我们的算法将优化收敛速率提升了 $\sqrt{\kappa}(\mathbf{\Sigma}$倍。当 $\mathbf{\Sigma}$ 条件较差时，这一改进十分显著。

统计项。 统计项可以解释如下。该项对应于极小化点 $\mathbf{x}^*$ 处的噪声协方差 $\mathbf{Q}$ ，并沿目标函数海森矩阵的下界特征方向受到抑制，即 $\mathbf{H}/\alpha$。具体而言，噪声$\mathbf{Q}$ 会经过局部曲率的滤波：曲率较大的方向会收缩噪声，而较平坦的方向则会放大噪声。

若不对 ℓ 的三阶光滑性作出额外假设，就无法改进这一统计项。我们在附录G 中构造了一个问题类，其极小极大风险的下界为 $\Omega(\alpha\mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q})/n)$，从而确立了我们所给界的最坏情况最优性。

模型设定错误项。 模型设定错误项是近似误差 $\mathbf{a}\mathbf{a}^\top \neq \mathbf{\Sigma}$ （亦称四阶矩效应 [Jain 等人， 2018a, Li 等人， 2024]）与噪声协方差 $\mathbf{Q} \neq \mathbf{H}$ 可能存在的耦合作用。请注意，复杂度不依赖于 κ，也不依赖于 $\alpha^{2/3}$， $\mu^{-1/3}$，以及 $\varepsilon^{-1/3}$。若要达到高精度且满足 ε，则需要较小的步长。定理 1表明，模型设定错误渐近趋于消失。

对于满足 $\mathbf{Q} = \sigma^2\mathbf{\Sigma}$ 的设定正确的模型，经过一些额外分析，可以将模型设定错误项吸收到统计项中。对于备注 1 中的高斯分布和次高斯分布，直接将 $\kappa \asymp \operatorname{tr}\mathbf{\Sigma}/\mu$ 和 $\tilde{\kappa} \asymp d$ 应用于定理 1中的界即可得到这一结论，因此模型设定错误项为 $\frac{\alpha\eta\tilde{\kappa}_{\mathrm{tr}\mathbf{Q}}}{L_\ell n} = \frac{\sigma^2\alpha\eta d\mathrm{tr}\ \mathbf{\Sigma}}{n} \leqslant \frac{\sigma^2\alpha d}{n} = \frac{\alpha\mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q})\,n}{}$，其中我们使用了 $\eta\mathrm{tr}\ \mathbf{\Sigma} \leqslant 1$。

对开放问题的回答。 我们的结果回答了 Jain 等人 [2018a] 在最小假设下的一般线性预测设定中提出的开放问题。具体而言，我们证明了加速方法在以下意义下达到了渐近最优收敛率：

$$\limsup_{n\to\infty} \frac{\mathbb{E}F(\mathbf{x}_n^{\mathrm{SADA}}) - F(\mathbf{x}^*)}{\mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q})/n} = \mathcal{O}(\alpha),$$

其中，如上所述，项 $\mathcal{O}(\alpha)$在我们的设定下是最优的。

6 证明概要

6.1 第一部分：内循环分析

子问题（4）类似于使用数据对（$\mathbf{a}\ -h_k\ell'(\mathbf{a}^\top \tilde{\mathbf{y}}_{k-1}, b)$ ）进行线性回归。我们将（4）的极小化点记为 $\tilde{\mathbf{x}}_k^*$。应用标准的偏差-方差分解 [DieuleveutandBach、2016、Jain 等人、2018a、Zou 等人、2023、Wu 等人、2022年、Li 等人， 2024], 我们可以将 $\boldsymbol{\eta}_t \overset{\mathrm{def}}{=} \begin{pmatrix} \mathbf{x}_t - \tilde{\mathbf{x}}_k^* \\ \mathbf{y}_t - \tilde{\mathbf{x}} \end{pmatrix} \in \mathbb{R}^{2d}$ 的动力学分解为偏差项 $\boldsymbol{\eta}_t^{\mathrm{bias}}$与方差项 $\boldsymbol{\eta}_t^{\mathrm{var}}$之和：

$$\begin{cases} \boldsymbol{\eta}_t^{\mathrm{bias}} = \hat{\mathbf{A}}_t\boldsymbol{\eta}_{t-1}^{\mathrm{bias}}, & \boldsymbol{\eta}_0^{\mathrm{bias}} = \boldsymbol{\eta}_0, \\ \boldsymbol{\eta}_t^{\mathrm{var}} = \hat{\mathbf{A}}_t\boldsymbol{\eta}_{t-1}^{\mathrm{var}} + \boldsymbol{\zeta}_t, & \boldsymbol{\eta}_0^{\mathrm{var}} = \mathbf{0}, \end{cases} \quad \text{where} \quad \hat{\mathbf{A}}_t = \begin{pmatrix} \mathbf{O} & \mathbf{I} - \eta\mathbf{a}_t\mathbf{a}_t^\top \\ -\frac{1-\theta}{1+\theta}\mathbf{I} & \frac{2}{1+\theta}\mathbf{I} - \frac{\eta+\theta\gamma}{1+\theta}\mathbf{a}_t\mathbf{a}_t^\top \end{pmatrix}.$$

内循环输出 $\tilde{\mathbf{x}}_k$ 可以分解为 $\tilde{\mathbf{x}}_k = \tilde{\mathbf{x}}_k^* + h_k\mathbf{r}_k + h_k\mathbf{v}_k$，其中 $\mathbf{r}_k \in \mathbb{R}^d$ 是刻画优化误差的偏差项，而 $\mathbf{v}_k \in \mathbb{R}^d$ 是均值为零的噪声项。内循环中的整体结果见引理 4。 $\mathbf{r}_k$ 的界Jain 等人[2018a] 利用附录 D.3中展示的标准加速技术即可得出。主要关注点是推导 $\|\mathbf{v}_k\|_{\mathbf{H}}^2$的界，这并非易事。

$\mathbf{v}_k$ 的界。 这需要分析协方差 $\mathbf{C}_t = \mathbb{E}\left(\boldsymbol{\eta}_t^{\mathrm{var}}(\boldsymbol{\eta}_t^{\mathrm{var}})^\top\right)$。其动力学 $\mathbf{C}_t$满足

$$\mathbf{C}_t = \mathcal{B} \circ \mathbf{C}_{t-1} + \begin{pmatrix} \eta^2\mathbf{R} & \eta q\mathbf{R} \\ \eta q\mathbf{R} & q^2\mathbf{R} \end{pmatrix}, \quad \mathbf{C}_0 = \mathbf{O}.$$

其中$\mathbf{R} \in \mathbb{R}^{2d\times 2d}$是噪声协方差，而 $\mathcal{B}\colon \mathbf{C} \mapsto \mathbb{E}\,\hat{}\mathbf{A}_t\mathbf{C}\hat{}\mathbf{A}_t^\top$ 是一个作用于$\mathbb{R}^{2d\times 2d}$ 的线性矩阵算子， $\circ$ 表示线性矩阵算子作用于矩阵的运算。请注意，在先前关于设定正确的线性回归的研究中，$\mathbf{R} = \sigma^2\mathbf{\Sigma}$ 与数据协方差 $\mathbf{\Sigma}$可交换，

此时分析$\mathbf{C}_t$的稳态性质相对简单。相比之下，我们的设定需要理解一般$\mathbf{R}$下$\mathbf{C}_t$的行为。

我们提出层剥离分解，将$\mathbf{C}_t$分解为多层动力学$\tilde{\mathbf{C}}_t$和$\tilde{\mathbf{C}}_t^{(\ell)}$。具体而言，我们引入$\mathbf{C}_t = \tilde{\mathbf{C}}_t + \sum_{\ell=0}^{T} \tilde{\mathbf{C}}_t^{(\ell)}$，其中

$$\begin{aligned}
&\tilde{\mathbf{C}}_t = \mathbf{A}\tilde{\mathbf{C}}_{t-1}\mathbf{A} + \mathbf{R}, && \tilde{\mathbf{C}}_0 = \mathbf{O}, && \tilde{\mathbf{C}}_t^{(0)} = \tilde{\mathbf{C}}_t,\\
&\tilde{\mathbf{C}}_t^{(\ell)} = \mathbf{A}\tilde{\mathbf{C}}_{t-1}^{(\ell)}\mathbf{A} + (\mathcal{B} - \mathbf{A}\otimes\mathbf{A})\circ\tilde{\mathbf{C}}_t^{(\ell-1)}, && \tilde{\mathbf{C}}_0^{(\ell)} = \mathbf{O}, && \text{for } \ell = 1, 2, \dots
\end{aligned}$$

以及$\mathbf{A}_t$。直观而言，$\tilde{\mathbf{C}}_t$可视为将$\mathbf{aa}^\top$替换为$\mathbf{\Sigma}$后动力学的协方差，而$\tilde{\mathbf{C}}_t^{(\ell)}$则汇集了这些近似误差。为了研究$\tilde{\mathbf{C}}_t$和$\tilde{\mathbf{C}}_t^{(\ell)}$，我们为$\mathbf{M} \in \mathbb{R}^{d\times d}$引入核心辅助动力学$\tilde{\mathbf{\Pi}}_t(\mathbf{M}) \in \mathbb{R}^{2d\times 2d}$，如下所示：

$$\tilde{\mathbf{\Pi}}_t(\mathbf{M}) = \mathbf{A}\tilde{\mathbf{\Pi}}_{t-1}(\mathbf{M})\mathbf{A} + \begin{pmatrix} \eta^2\mathbf{M} & \eta q\mathbf{M} \\ \eta q\mathbf{M} & q^2\mathbf{M} \end{pmatrix}, \quad \tilde{\mathbf{\Pi}}_0(\mathbf{M}) = \mathbf{O}. \tag{7}$$

由于我们有$\tilde{\mathbf{C}}_t = \tilde{\mathbf{\Pi}}_t(\mathbf{R})$，且$\tilde{\mathbf{C}}_{t-1}^{(\ell)}$可由$\tilde{\mathbf{\Pi}}_t(\mathbf{R}^{(\ell)})$界定，其中某个$\mathbf{R}^{(\ell)}$与$\mathbf{C}_t^{(\ell-1)}$的平稳协方差相关。因此，我们可以递归地将$\tilde{\mathbf{C}}_t^{(\ell)}$的结果推至其极限，从而界定方差。

关于方差上界的详细证明见附录 D.2。在定义 2 中，我们引入线性映射L_{var}，用于计算一般协方差矩阵序列的损失。在附录 D.2.2 中，我们研究式 (7) 的动力学。具体而言，我们在引理16中得到了$L_{\text{var}}(\{\tilde{\mathbf{\Pi}}_t(\mathbf{M})\}_{t\in\mathbb{N}})$的紧上界，并在引理 19 中得到了平稳协方差$\tilde{\mathbf{\Pi}}_\infty(\mathbf{M}) = \lim_{t\to\infty}\tilde{\mathbf{\Pi}}_t(\mathbf{M})$。在附录 D.2.3 中，我们研究噪声协方差矩阵 R，并分解 $\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathrm{R})$ 和 $\mathrm{tr}(\mathrm{R})$。最后，在附录 D.2.5 中，我们引入$\tilde{\mathrm{C}}_t$和$\tilde{\mathrm{C}}_t^{(\ell)}$。我们首先在引理 25 中证明，方差还可以进一步分解如下：$\mathbb{E}\|\mathrm{v}_k\|_{\mathbf{\Sigma}}^2 = L_{\text{var}}(\{\tilde{\mathrm{C}}_t\}_{t\in\mathbb{N}}) + \sum_{\ell=1}^{T} L_{\text{var}}(\{\tilde{\mathrm{C}}_t^{(\ell)}\}_{t\in\mathbb{N}})$。然后，代入引理 30，即可得到方差的最终上界。

6.2 第二部分：外循环分析

外循环的界主要利用内循环返回的解$\tilde{\mathbf{x}}_k$的结构。我们的证明技巧沿用加速梯度下降的分析，但会仔细控制噪声项 (C)。主要结果如下引理所示。

引理 1. 假设我们有$h_k \lesssim$最小化$\left\{\frac{1}{L_\ell}, \frac{T^2}{L_\ell\alpha^3\tilde{\kappa}^2}\right\}$并设$\theta_k = \sqrt{\frac{h_k}{2\alpha}}$。令能量为$L_k = \mathbb{E}F(\tilde{\mathbf{x}}_k) - F(\mathbf{x}^*) + \frac{2\theta_k^2}{3L_\ell h_k}\mathbb{E}_{k-1}\|\tilde{\mathbf{z}}_k - \mathbf{x}^*\|_{\mathbf{H}}^2$，则有

$$L_k \leqslant \left(1 - \frac{\theta_k}{2}\right)L_{k-1} + \frac{3h_k\sigma^2}{2}, \quad \textit{where } \sigma^2 \lesssim \frac{L_\ell\,\mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q})}{T} + \frac{\eta\tilde{\kappa}\,\mathrm{tr}\,\mathbf{Q}}{T}.$$

我们采用与（6）中相同的两阶段步长方案。对于第一阶段（$t \leqslant K/2$），我们选择最大步长，以加速初始误差的收敛。对于第二阶段（$t > K/2$），我们使用$h_{k+K/2} \asymp \frac{1}{\mu_\ell(a+k)^2}$来控制噪声累积。

7 扩展与讨论

7.1 弱凸情形

我们应用约化技术，将结果推广到弱凸情形。假设 $\|\mathbf{x}^*\| \leqslant D$ 和 $\|\mathbf{x}^*\|_{\mathbf{H}} \leqslant M$，那么我们考虑优化以下替代函数：

$$G(\mathbf{x}) = \mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\ell(\mathbf{a}^\top\mathbf{x}, b) + \underbrace{\frac{\varepsilon}{2M^2}\left(\mathbf{a}^\top\mathbf{x}\right)^2}_{\text{Term (A)}}\right) + \underbrace{\frac{\varepsilon}{2D^2}\|\mathbf{x}\|^2}_{\text{Term (B)}},$$

其中，项 (A) 是在 ℓ 为弱凸（$\mu_\ell \leqslant \varepsilon$）时所需的，而项 (B) 是在 $\mathbf{\Sigma}$ 病态（$\mu \leqslant \varepsilon$）时所必需的。算法 SADA-WC（算法 2）详见附录 H。

推论 2 (样本复杂度)。 在定理 *1*的设定下。假设我们运行算法 *2*，得到一个解 $\tilde{\mathbf{x}}_K$，其超额风险至多为 ε，所需样本量 n 为

$$\left(\sqrt{\frac{L_\ell R^2\tilde{\kappa}}{\mu'_\ell\mu'}} + \frac{L_\ell^2\tilde{\kappa}}{(\mu'_\ell)^2}\right) + \frac{L_\ell \operatorname{tr}((\mathbf{H}+\varepsilon\mathbf{I}/D^2)^{-1}\mathbf{Q})}{\mu'_\ell\varepsilon} + \left(\frac{L_\ell\tilde{\kappa}^2\operatorname{tr}\mathbf{Q}}{(\mu'_\ell)^2\mu'\varepsilon}\right)^{1/3},$$

其中 $\mu'_\ell = \text{最大值}\{\mu_\ell, \varepsilon\}$ 和 $\mu' = \text{最大值}\{\mu, \varepsilon\}$。

7.2 未标记数据的使用

在内循环中，我们使用单个数据点 $\mathbf{a}\mathbf{a}^\top$ 来近似 $\mathbf{\Sigma}$。该过程不需要相应的标签 b。我们可以利用可用的无标签数据来改进对$\mathbf{\Sigma}$的估计。具体而言，在某些步骤中，我们使用 m 个无标签样本以获得更好的估计（见附录 H 中的算法 3 ）。然后，我们可以分别将 R^2、κ 和 $\tilde{\kappa}$ 改进为 $R_m^2 = {}^{R^2+m\lambda_{\max}(\mathbf{\Sigma})}{}_{m+1}$、$\kappa_m = {}^{R_m^2}{}_{\mu}$ 和 $\tilde{\kappa}_m = {}^{\tilde{\kappa}+m}{}_{m+1}$。换言之，为了得到超额风险至多为 ε 的解，所需的有标签样本量 n 为

$$n \geqslant \tilde{\Theta}\left(\left(\sqrt{\alpha\kappa_m\tilde{\kappa}_m} + \alpha^2\tilde{\kappa}\right) + \frac{\alpha\operatorname{tr}(\mathbf{H}^{-1}\mathbf{Q})}{\varepsilon} + \left(\frac{\alpha^2\tilde{\kappa}_m^2\operatorname{tr}\mathbf{Q}}{L_\ell\mu\varepsilon}\right)^{1/3}\right).$$

无标签数据的总量至少为 $\tilde{\Omega}(m\sqrt{\alpha\tilde{\kappa}_m/\mu\eta})$，其中我们设定 $\eta = \tilde{\Theta}\left(\min\left\{\frac{1}{R_m^2}, \frac{n\varepsilon L_\ell}{\alpha\tilde{\kappa}_m\operatorname{tr}\mathbf{Q}}\right\}\right)$ 和 $K = \Theta(\sqrt{\alpha})$ 以及 $T = n/K$。有标签数据复杂度中的冗余可归因于优化和模型设定错误，并由 R_m^2、κ_m 和 $\tilde{\kappa}_m$ 直接量化。

7.3 小批量处理与并行化

小批量处理。 该算法默认采用单样本更新，但也可以扩展到批处理设置。使用批量大小为 B 的批次，我们可以分别将 R^2、κ 和 $\tilde{\kappa}$ 重新定义为 $R_B^2 = {}^{R^2+(B-1)\lambda_{\max}(\mathbf{\Sigma})}{}_{B}$、$\kappa_B = \frac{R_B^2}{\mu}$ 和 $\tilde{\kappa}_B = {}^{\tilde{\kappa}+(B-1)}{}_{B}$。要获得超额风险至多为 ε 的解，样本量 n 为

$$n \geqslant \tilde{\Theta}\left(B\left(\sqrt{\alpha\kappa_B\tilde{\kappa}_B} + \alpha^2\tilde{\kappa}_B\right) + \frac{\alpha\operatorname{tr}(\mathbf{H}^{-1}\mathbf{Q})}{\varepsilon} + \left(\frac{\alpha^2\tilde{\kappa}_B^2B^2\operatorname{tr}\mathbf{Q}}{L_\ell\mu\varepsilon}\right)^{1/3}\right),$$

以及最大批量大小 $B_{\max} = n\sqrt{\mu\eta}/(\alpha\kappa_B)$，其中我们设定 $\eta = \tilde{\Theta}\left(\min\left\{\frac{1}{R_B^2}, \frac{n\varepsilon L_\ell}{\alpha\tilde{\kappa}_B\operatorname{tr}\mathbf{Q}}\right\}\right)$。以及 $K = \Theta(\sqrt{\alpha})$ 和 $T = n/K$。

并行化。算法 1 的内循环可以在 $T \geqslant T_0 = \tilde{\Theta}(\sqrt{\tilde{\kappa}/(\mu\eta)})$ 时并行实现。 T 内迭代可以划分为 T/T_0 个相互独立的内层运行，外层迭代 $\tilde{\mathbf{x}}_k$ 定义为这些运行输出的平均值。在这种并行实现下，定理 1中的收敛保证仍然成立。

7.4 高阶光滑性假设

与我们所关联的传统基于梯度的优化和局部拉德马赫复杂度框架类似，我们的分析并不假设 ℓ 具有三阶光滑性。在这一最弱正则性条件下，得到的统计速率是极小极大最优的。若假设海森光滑性，则可以期待更快的速率 $\mathrm{tr}((\nabla^2 F(\mathbf{x}^*))^{-1}\mathrm{Q})/n$，与经验风险最小化（ERM）的渐近效率及 Cram´er-Rao 下界相匹配。一种直接的方法是先运行 SADA，直到迭代点进入 $\mathbf{x}^*$ 的邻域，例如 $\|\tilde{\mathbf{x}} - \mathbf{x}^*\| \leqslant \frac{1}{\mu_\ell \mu}$。然后，在求解子问题 4 时，使用经过仔细选择的步长，将 $\mathbf{a}\mathbf{a}^\top$ 替换为 $\ell''(\mathbf{a}^\top \tilde{\mathbf{x}} - b)\mathbf{a}\mathbf{a}^\top$。然而，分析会复杂得多。此外，当海森矩阵光滑时，三次正则化等高阶方法 [Nesterov 和 Polyak, 2006]，可能更为合适，因为它们能使外循环达到更快的速率。我们将这项研究留待未来开展。

参考文献

Alekh Agarwal、Peter L. Bartlett、Pradeep Ravikumar 和 Martin J. Wainwright。随机凸优化预言机复杂度的信息论下界。*IEEE* 信息论汇刊，58(5):3235–3249，2012。Zeyuan Allen-Zhu。Katyusha：首个直接加速随机梯度方法的算法。机器学习研究杂志，18(221):1–51，2018。Zeyuan Allen-Zhu 和 Elad Hazan。通过方差缩减加速非凸优化。发表于国际机器学习会议，2016。Francis Bach 和 Eric Moulines。收敛率为 $O(1/n)$的非强凸光滑随机逼近。发表于神经信息处理系统，2013。Peter L. Bartlett、Olivier Bousquet 和 Shahar Mendelson。局部拉德马赫复杂度。统计年刊，33(4):1497–1537，2005。S´ebastien Bubeck 等人。凸优化：算法与复杂度。基础与趋势® 机器学习，8(3-4):231–357，2015。Alexandre D´efossez 和 Francis Bach。平均最小均方：偏差-方差权衡与最优抽样分布。发表于人工智能与统计学，2015。Aymeric Dieuleveut 和 Francis Bach。采用大步长的非参数随机逼近。统计年刊，44(4):1363–1399，2016。Aymeric Dieuleveut、Nicolas Flammarion 和 Francis Bach。更难、更好、更快、更强：最小二乘回归的收敛率。机器学习研究杂志，18(101):1–51，2017。

Cong Fang、Chris Junchi Li、Zhouchen Lin 和 Tong Zhang。SPIDER：通过随机路径积分微分估计器实现近乎最优的非凸优化。发表于神经信息处理系统，2018。Cong Fang、Zhouchen Lin 和 Tong Zhang。对逃离鞍点的非凸随机梯度下降进行精细分析。发表于学习理论会议，2019。Roy Frostig、Rong Ge、Sham M Kakade 和 Aaron Sidford。单次遍历中与经验风险最小化器竞争。发表于学习理论会议，2015。Rong Ge、Sham M Kakade、Rahul Kidambi 和 Praneeth Netrapalli。阶梯衰减计划：一种近乎最优、按几何级数衰减的最小二乘学习率方案。发表于神经信息处理系统，2019。Elad Hazan 和 Satyen Kale。超越遗憾最小化障碍：随机强凸优化的最优算法。机器学习研究期刊，15(1)：2489–2512，2014。Prateek Jain、Sham M Kakade、Rahul Kidambi、Praneeth Netrapalli 和 Aaron Sidford。加速用于最小二乘回归的随机梯度下降。发表于学习理论会议，2018a。Prateek Jain、Sham M Kakade、Rahul Kidambi、Praneeth Netrapalli 和 Aaron Sidford。并行化用于最小二乘回归的随机梯度下降：小批量处理、平均与模型设定错误。机器学习研究杂志，18(223):1–42，2018b。Chi Jin、Praneeth Netrapalli 和 Michael I Jordan。加速梯度下降比梯度下降更快地逃离鞍点。发表于学习理论会议，2018。Chris Junchi Li、Wenlong Mou、Martin Wainwright 和 Michael Jordan。ROOT-SGD：在单一算法中实现精细的非渐近分析与渐近效率。发表于学习理论会议，2022。Xuheng Li、Yihe Deng、Jingfeng Wu、Dongruo Zhou 和 Quanquan Gu。过参数化线性回归中加速随机梯度下降的风险界。发表于国际学习表征会议，2024。Hongzhou Lin、Julien Mairal 和 Zaid Harchaoui。一阶凸优化的 Catalyst 加速：从理论到实践。机器学习研究杂志，18(212):1–54，2018。Yuanshi Liu、Haihan Zhang、Qian Chen 和 Cong Fang。协变量偏移下线性回归的最优算法：预条件的重要性。*arXiv* 预印本 *arXiv:2502.09047*，2025。Eric Moulines 和 Francis Bach。机器学习随机逼近算法的非渐近分析。发表于神经信息处理系统，2011。Jaouad Mourtada。线性最小二乘的精确极小极大风险，以及样本协方差矩阵的下尾。统计年刊，50(4):2157–2178，2022。

Ion Necoara、尤里·涅斯捷罗夫和 Francois Glineur。非强凸优化一阶方法的线性收敛。数学规划，175(1):69–107，2019。尤里·涅斯捷罗夫。坐标下降法在超大规模优化问题上的效率。*SIAM*优化期刊，22(2):341–362，2012。尤里·涅斯捷罗夫。一种求解凸规划问题的方法，其收敛速率为$O(1/k^2)$。苏联科学院学报，269:543–547，1983。尤里·涅斯捷罗夫。凸优化讲义，第137卷。施普林格，2018。尤里·涅斯捷罗夫和鲍里斯·波利亚克。牛顿法的三次正则化及其全局性能。数学规划，108(1):177–205，2006。Lam M Nguyen、Jie Liu、Katya Scheinberg 和 Martin Tak´aˇc。SARAH：一种用于机器学习问题的新型随机递归梯度方法。载于国际机器学习会议，2017。Brendan O'donoghue 和 Emmanuel Candes。加速梯度格式的自适应重启。计算数学基础，15(3):715–732，2015。Rui Pan、Yuxing Liu、Xiaoyu Wang 和 Tong Zhang。各向异性梯度噪声下随机重球法的加速收敛。载于国际学习表征会议，2024。鲍里斯·波利亚克。若干加速迭代法收敛的方法。苏联计算数学与数理物理，4(5):1–17，1964。Sashank J Reddi、Ahmed Hefny、Suvrit Sra、Barnabas Poczos 和 Alex Smola。非凸优化的随机方差缩减。载于国际机器学习会议，2016。Herbert Robbins 和 Sutton Monro。一种随机逼近方法。数理统计年刊，第400–407页，1951。Shai Shalev-Shwartz 和 Tong Zhang。正则化损失的随机对偶坐标上升法。机器学习研究期刊，14(1):567–599，2013。Shai Shalev-Shwartz 和 Tong Zhang。用于正则化损失最小化的加速近端随机对偶坐标上升。载于国际机器学习会议，2014。Marina Sheshukova、Denis Belomestny、Alain Oliviero Durmus、Eric Moulines、Alexey Naumov 和 Sergey Samsonov。利用 Richardson–Romberg 外推对随机梯度下降进行非渐近分析。载于第十三届国际学习表征会议，2025。Martin J Wainwright。高维统计：非渐近观点，第48卷。剑桥大学出版社，2019。Jingfeng Wu、Difan Zou、Vladimir Braverman、Quanquan Gu 和 Sham Kakade。衰减步长下过参数化线性回归的随机梯度下降最后迭代风险界。载于国际机器学习会议，2022。

Haihan Zhang、Yuanshi Liu、陈倩雯和 Cong Fang。高维二次优化中（加速）随机梯度下降的最优性。*arXiv*预印本 *arXiv:2409.09745*，2024。Haihan Zhang、林伟成、Yuanshi Liu 和 Cong Fang。核回归中随机梯度下降的学习曲线。载于第四十二届国际机器学习大会，2025年。

Difan Zou、Jingfeng Wu、Vladimir Braverman、Quanquan Gu 和 Sham M Kakade。恒定步长随机梯度下降在线性回归中的良性过拟合。机器学习研究杂志，24(326):1–58，2023。

目录

1 引言 1.1 回顾：先前结果.3 1.2 我们的结果及其影响.3 **2 相关工作** 4 **3 问题设置** 5 3.1 假设.5 3.2 问题相关量汇总.6 **4 随机加速数据依赖算法** 7 4.1 内循环：带尾部平均的加速求解器.7 4.2 外循环：带加速的数据依赖近端方法.8 **5 收敛结果与算法复杂度** 8 **6 证明概要** 10 6.1 第一部分：内循环分析.10 6.2 第二部分：外循环分析.11

7 扩展与讨论 12

7.1 弱凸情形.12 7.2 无标签数据的使用.12 7.3 小批量处理与并行化.12 7.4 高阶光滑性假设.13

A 组织结构 18 **B 广义假设** 19 **C 超参数选择** 20 **D 第一部分：分析**

sis of 内循环 20

D.1 预备知识.21 D.1.1 偏差-方差分解.21 D.1.2 协方差动态.22 D.1.3 动量矩阵的性质.23 D.2 方差上界.25 D.2.1 层剥离分解.27 D.2.2 核心辅助动力学.28 D.2.3 噪声协方差分解.32 D.2.4 第0层动力学的界 $\mathbf{C}_t$ .33 D.2.5 第 ℓ 层动力学 $\tilde{\mathbf{C}}_t^{(\ell)}$ .34 D.2.6 引理 9 的证明.36 D.3 偏差上界.37

D.4 引理 4 的证明.39D.5 辅助引理.39

E 第二部分：分析 外循环分析3 **9**
E.1 外循环的超额风险界.40E.2 定理1的证明.43E.3 推论1 的证明.44

F 经验风险最小化的超额风险 **44**
F.1 证明 定理3的证明.4 5

G 最坏情况最优性p性 y 统计项 **46**
G.1 定理 4 的证明.47

H SADA 的扩展 **48**
H.1 弱凸目标的 SADA.48H.2 使用无标签数据的 SADA.49

A 组织

附录的组织结构如下：

- 在附录B中，我们提出了两个一般性假设（假设4 和5），用于刻画目标函数 F的随机梯度噪声。在正文中的假设1、2 和3 成立的条件下，假设4 和5 在 $B = \alpha R^2$ 和 $L = \alpha\tilde{\kappa}$ 条件下成立（见备注5）。附录中的结果以 B 和 L为条件给出，并通过选择 $B = \alpha R^2$ 和 $L = \alpha\tilde{\kappa}$，立即转化为正文中的结果。

- 在附录C中，我们详细说明超参数选择，并明确给出常数因子和对数因子的依赖关系。

- 在附录D中，我们分析算法1的内循环。

- 在附录E中，我们分析算法1的外循环。

- 在附录F中，我们使用局部化技术给出了经验风险最小化的非渐近超额风险上界的证明 [Bartlett 等人，2005 年，Wainwright， 2019]。

- 在附录G中，我们通过构造一个困难问题实例，证明定理1 中的统计项是极小极大最优的。对于这个问题，超额风险的下界必须为 $\Omega(\alpha\mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q})/n)$。

- 在附录H中，我们给出了适用于弱凸情形的算法（算法2），以及可以使用无标签数据的算法（算法3）。

B 广义假设

本节提出了两个一般性假设，可用于精细刻画梯度噪声；随后我们证明，这些假设在假设1、2和3下均成立。

假设 4（梯度噪声二范数的界）。 我们假设存在 $B > 0$，使得对所有$\mathbf{x}$，以下成立：

$$\mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\ell'(\mathbf{a}^\top\mathbf{x},b)-\ell'(\mathbf{a}^\top\mathbf{x}^*,b)\right)^2\|\mathbf{a}\|^2 \leqslant 2L_\ell B(F(\mathbf{x})-F(\mathbf{x}^*)).$$

假设 5（ $\mathbf{\Sigma}^{-1}$范数下梯度噪声的界）。 我们假设存在 $L > 0$，使得对所有$\mathbf{x}$，以下成立：

$$\mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\ell'(\mathbf{a}^\top\mathbf{x},b)-\ell'(\mathbf{a}^\top\mathbf{x}^*,b)\right)^2\|\mathbf{a}\|_{\mathbf{\Sigma}^{-1}}^2 \leqslant 2L_\ell L(F(\mathbf{x})-F(\mathbf{x}^*)).$$

注记 5。 在假设1、2 和 3下，上述假设以 $B = \alpha R^2$ 和 $L = \alpha\tilde{\kappa}$成立，如引理2 和引理3所示。此外，若能得到更紧的界，则定理1 中的收敛速率可以改进。参见定理2，其中可能更紧的收敛速率显式依赖于 B 和 L。

我们验证假设 4 和 5 在假设 1、2 和3下。

引理 2（假设 4的验证）。 设假设 *1*、*2* 和 *3* 成立。则对任意$\mathbf{x} \in \mathbb{R}^d$，我们有

$$\mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\ell'(\mathbf{a}^\top\mathbf{x},b)-\ell'(\mathbf{a}^\top\mathbf{x}^*,b)\right)^2\|\mathbf{a}\|^2 \leqslant 2L_\ell\alpha R^2(F(\mathbf{x})-F(\mathbf{x}^*)).$$

证明。根据假设 1，我们有 $0 \leqslant \ell''(\cdot,\cdot) \leqslant L_\ell$。因此，$|\ell'(\mathbf{a}^\top\mathbf{x}\ b)-\ell'(\mathbf{a}^\top\mathbf{x}^*,b)| \leqslant L_\ell|\mathbf{a}^\top(\mathbf{x}-\mathbf{x}^*)|$。根据假设 3，左侧可作如下界定：

$$\begin{aligned}\mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\ell'(\mathbf{a}^\top\mathbf{x},b)-\ell'(\mathbf{a}^\top\mathbf{x}^*,b)\right)^2\|\mathbf{a}\|^2 &\leqslant L_\ell^2\mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\mathbf{a}^\top(\mathbf{x}-\mathbf{x}^*)\right)^2\|\mathbf{a}\|^2\\ &\leqslant L_\ell^2R^2\|\mathbf{x}-\mathbf{x}^*\|_{\mathbf{\Sigma}}^2 \leqslant 2L_\ell\alpha R^2(F(\mathbf{x})-F(\mathbf{x}^*)).\end{aligned}$$

证明至此完成。 □

引理 3 (假设 5 的验证)。 设假设 *1*、*2* 和 *3* 均成立。则对于任意$\mathbf{x} \in \mathbb{R}^d$，有

$$\mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\ell'(\mathbf{a}^\top\mathbf{x},b)-\ell'(\mathbf{a}^\top\mathbf{x}^*,b)\right)^2\|\mathbf{a}\|_{\mathbf{\Sigma}^{-1}}^2 \leqslant 2L_\ell\alpha\tilde{\kappa}(F(\mathbf{x})-F(\mathbf{x}^*)).$$

证明。 根据假设 1 和3，左侧可作如下界定：

$$\begin{aligned}\mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\ell'(\mathbf{a}^\top\mathbf{x},b)-\ell'(\mathbf{a}^\top\mathbf{x}^*,b)\right)^2\|\mathbf{a}\|_{\mathbf{\Sigma}^{-1}}^2 &\leqslant L_\ell^2\mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\mathbf{a}^\top(\mathbf{x}-\mathbf{x}^*)\right)^2\|\mathbf{a}\|_{\mathbf{\Sigma}^{-1}}^2\\ &\leqslant L_\ell^2\tilde{\kappa}\|\mathbf{x}-\mathbf{x}^*\|_{\mathbf{\Sigma}}^2 \leqslant 2L_\ell\alpha\tilde{\kappa}(F(\mathbf{x})-F(\mathbf{x}^*)).\end{aligned}$$

证明至此完成。 □

C 超参数选择

本节给出依赖于假设 4~和 5中的常数的超参数设置。我们还给出超参数 T 和 $\theta_{\max}$的显式常数和对数因子。我们选择超参数 η 和 T 使得

$$\eta \leqslant \frac{1}{16R^2}, \quad T \geqslant \sqrt{\frac{\tilde{\kappa}}{\mu\eta}} \ln \kappa(\boldsymbol{\Sigma}) \ln \frac{4}{\tilde{\theta}_K^2}.$$

然后，将内循环的超参数选择如下：

$$\gamma = \frac{1}{4}\sqrt{\frac{\eta}{\tilde{\kappa}\mu}}, \quad \theta = \frac{1}{4}\sqrt{\frac{\mu\eta}{\tilde{\kappa}}}.$$

对于外循环，令 $L_{\text{eff}} = 160（6L + \tilde{\kappa}（7 + 16\eta B）$，其中 B和 L定义于假设 4和5 （另见备注 5）。我们定义

$$\tilde{\theta}_{\max} = \min\left\{\sqrt{\frac{1}{12\alpha}}, \frac{T}{12\sqrt{2}\alpha L_{\text{eff}}}\right\}, \quad \tilde{\theta}_k = \begin{cases} \tilde{\theta}_{\max}, & k \leqslant K/2, \\ \dfrac{4}{4/\tilde{\theta}_{\max} + k - K/2}, & k > K/2. \end{cases}$$

然后，将外循环的超参数选择如下：

$$h_k = \frac{2\alpha\tilde{\theta}_k^2}{L_\ell}, \quad \beta_k = \frac{1 - \tilde{\theta}_k}{1 + \tilde{\theta}_k}.$$

D 第一部分：内循环分析

组织结构。 在本节中，我们研究算法 k 的第 1 个外层迭代所对应的内循环。

- 在附录D.1中，我们介绍分析线性回归所需的预备工具。在附录D.1.1中，我们介绍偏差-方差分解技术 [Dieuleveutand Bach、2016、Jain 等人、2018a、Zou 等人， 2023] 以将动力学分解为偏差项（用于刻画优化误差）和方差项（用于刻画随机噪声）。在附录D.1.2中，我们推导偏差迭代协方差$\mathbf{B}_t$ 和方差迭代协方差$\mathbf{C}_t$的动力学。在附录D.1.3中，我们总结有关动量矩阵$\mathbf{A}$的结果，该矩阵决定动力学$_{\text{bias}t}$ 和 $\boldsymbol{\eta}_{\text{var}t}$ 的收缩。

- 在附录D.2中，我们给出界定$\mathbf{C}_t$的完整证明；该证明在第~节6.1中作了概述。在附录D.2.2中，我们提出核心辅助动力学 $\boldsymbol{\Pi}_t(\cdot)$。随后，我们推导 ~ ~$_{\text{var}}(\{\boldsymbol{\Pi}_t(\mathbf{M})\}_{t\in\mathbb{N}})$的 L上界，并分析平稳协方差 $\boldsymbol{\Pi}_\infty(\mathbf{M})$。借助对 $\boldsymbol{\Pi}_t(\cdot)$的~紧确界，我们得以推导出方差项的界。在附录D.2.4中，~我们应用附录D.2.2 中的工具，证明第0层动力学$\mathbf{C}_t$的界。在附录D.2.5中，我们应用附录D.2.2 中的工具，证明动力学 ~$\mathbf{C}_{(\ell)\ t}$对于 $\ell \geqslant 1$的界。

- 在附录D.3中，我们给出了对偏差迭代进行界定的证明 $\boldsymbol{\eta}_{\text{偏差}t}$ 。该证明主要沿用了Jain 等人 [2018a]。

- 在附录D.4中，我们给出了引理4的证明，该引理是引理4的完整版本。该引理是分析外循环的基础。

为简化记号，我们记 $\tilde{\mathbf{x}} \stackrel{\text{def}}{=} \tilde{\mathbf{x}}_{k-1}$，$\tilde{\mathbf{y}} \stackrel{\text{定义}}{=} \tilde{\mathbf{y}}_{k-1}$ 以及 $\tilde{\mathbf{x}}_+ \stackrel{\text{定义}}{=} \tilde{\mathbf{x}}_k$。令 $\tilde{\mathbf{x}}_+^*$ 表示第 k 次外层迭代目标函数的极小化点

$$\min_{\mathbf{x}\in\mathbb{R}^d} \mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\langle \ell'(\mathbf{a}^\top \tilde{\mathbf{y}}_{k-1}, b)\mathbf{a}, \mathbf{x}-\tilde{\mathbf{y}}_{k-1}\rangle + \frac{1}{2h_k}\|\mathbf{x}-\tilde{\mathbf{y}}_{k-1}\|_{\mathbf{\Sigma}}^2.$$

本节旨在推导以下分解:

$$\tilde{\mathbf{x}}_+ = \tilde{\mathbf{x}}_+^* + h_k\mathbf{r}_k + h_k\mathbf{v}_k,$$

其中$\mathbf{r}_k$ 和 $\mathbf{v}_k$ 是满足以下引理的随机变量。

引理 4. 假设 *1*、*2*、*3*、*4*和 *5* 成立，则

$$\mathbb{E}\|\mathbf{r}\|_{\mathbf{H}}^2 \leqslant \frac{L_\ell h_k}{8\alpha}\|\nabla F(\tilde{\mathbf{y}}_{k-1})\|_{\mathbf{H}^{-1}}^2, \quad \mathbb{E}\mathbf{v} = \mathbf{0},$$
$$\mathbb{E}\|\mathbf{v}\|_{\mathbf{H}}^2 \leqslant \frac{320L_\ell\left(3L_\ell \operatorname{tr}(\mathbf{H}^{-1}\mathbf{Q}) + 8\eta\tilde{\kappa}\operatorname{tr}\mathbf{Q}\right)}{T} + \frac{160L_\ell^2(6L+\tilde{\kappa}(7+16\eta B))(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*))}{T},$$

其中，期望是对第 k 次外层迭代中抽取的样本取的。

不失一般性，在本节以下内容中，我们假设 $h_k = 1$。从现在起，我们选择一个能将 $\mathbf{\Sigma}$ 对角化的基。因此， $\mathbf{\Sigma}$ 和$\mathbf{H}$ 是对角矩阵。

D.1 预备知识

D.1.1 偏差-方差分解

我们首先给出内循环更新的简洁矩阵形式。

引理 5. 内循环的更新形式如下:

$$\begin{pmatrix}\mathbf{x}_t - \tilde{\mathbf{x}}_+^* \\ \mathbf{y}_t - \tilde{\mathbf{x}}_+^*\end{pmatrix} = \begin{pmatrix}\mathbf{O} & \mathbf{I}-\eta\mathbf{a}_t\mathbf{a}_t^\top \\ -c\mathbf{I} & (1+c)\mathbf{I} - q\mathbf{a}_t\mathbf{a}_t^\top\end{pmatrix}\begin{pmatrix}\mathbf{x}_{t-1} - \tilde{\mathbf{x}}_+^* \\ \mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\end{pmatrix} + \begin{pmatrix}\eta\varepsilon_t\mathbf{a}_t \\ q\varepsilon_t\mathbf{a}_t\end{pmatrix}. \tag{8}$$

其中

$$c = \frac{1-\theta}{1+\theta}, \quad q = \frac{\eta+\theta\gamma}{1+\theta}, \quad \varepsilon_t = -\ell'(\mathbf{a}_t^\top\tilde{\mathbf{y}}, b_t) + \mathbf{a}_t^\top(\tilde{\mathbf{y}} - \tilde{\mathbf{x}}_+^*). \tag{9}$$

证明。 回顾一下，更新规则为

$$\begin{aligned}\mathbf{x}_t &= \mathbf{y}_{t-1} - \eta\mathbf{g}_t, \\ \mathbf{z}_t &= \theta\mathbf{y}_{t-1} + (1-\theta)\mathbf{z}_{t-1} - \gamma\mathbf{g}_t, \\ \mathbf{y}_t &= \frac{1}{1+\theta}\mathbf{x}_t + \frac{\theta}{1+\theta}\mathbf{z}_t,\end{aligned}$$

其中梯度 $\mathbf{g}_t$ 为

$$\mathbf{g}_t = \ell'(\mathbf{a}_t^\top\tilde{\mathbf{y}}, b_t)\mathbf{a}_t + \mathbf{a}_t\mathbf{a}_t^\top(\mathbf{y}_{t-1} - \tilde{\mathbf{y}}),$$

通过代数变换消去 $\mathbf{z}_t$ 和 $\mathbf{z}_{t-1}$，我们得到以下矩阵形式：

$$\begin{pmatrix}\mathbf{x}_t\\ \mathbf{y}_t\end{pmatrix}=\begin{pmatrix}\mathbf{O} & \mathbf{I}\\ -c\mathbf{I} & (1+c)\mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{x}_{t-1}\\ \mathbf{y}_{t-1}\end{pmatrix}-\begin{pmatrix}\eta\mathbf{g}_t\\ q\mathbf{g}_t\end{pmatrix}, \tag{10}$$

其中 c 和 q 在（9）中定义。注意，根据（9）中 ε_t 的定义，我们有

$$\mathbf{g}_t=\mathbf{a}_t\mathbf{a}_t^\top(\mathbf{y}_{t-1}-\tilde{\mathbf{x}}_+^*)-\varepsilon_t\mathbf{a}_t. \tag{11}$$

将（11）代入（10），即可完成证明。 □

定义

$$\boldsymbol{\eta}_t=\begin{pmatrix}\mathbf{x}_t-\tilde{\mathbf{x}}_+^*\\ \mathbf{y}_t-\tilde{\mathbf{x}}_+^*\end{pmatrix},\quad \hat{\mathbf{A}}_t=\begin{pmatrix}\mathbf{O} & \mathbf{I}-\eta\mathbf{a}_t\mathbf{a}_t^\top\\ -c\mathbf{I} & (1+c)\mathbf{I}-q\mathbf{a}_t\mathbf{a}_t^\top\end{pmatrix},\quad \boldsymbol{\zeta}_t=\begin{pmatrix}\eta\varepsilon_t\mathbf{a}_t\\ q\varepsilon_t\mathbf{a}_t\end{pmatrix}.$$

那么，$\boldsymbol{\eta}_t$ 的动力学形式如下，

$$\boldsymbol{\eta}_t=\hat{\mathbf{A}}_t\boldsymbol{\eta}_{t-1}+\boldsymbol{\zeta}_t,\quad \boldsymbol{\eta}_0=\begin{pmatrix}\tilde{\mathbf{y}}-\tilde{\mathbf{x}}_+^*\\ \tilde{\mathbf{y}}-\tilde{\mathbf{x}}_+^*\end{pmatrix}.$$

按照偏差-方差分解技术 [DieuleveutandBach，2016，Jain 等人，2018a，Zou 等人，2023]，我们将 $\boldsymbol{\eta}_t$ 分解为偏差动力学 $\boldsymbol{\eta}_t^{\text{bias}}$和方差动力学 $\boldsymbol{\eta}_t^{\text{var}}$如下：

$$\begin{aligned}\boldsymbol{\eta}_t^{\text{bias}}&=\hat{\mathbf{A}}_t\boldsymbol{\eta}_{t-1}^{\text{bias}}, & \boldsymbol{\eta}_0^{\text{bias}}&=\begin{pmatrix}\tilde{\mathbf{y}}-\tilde{\mathbf{x}}_+^*\\ \tilde{\mathbf{y}}-\tilde{\mathbf{x}}_+^*\end{pmatrix},\\ \boldsymbol{\eta}_t^{\text{var}}&=\hat{\mathbf{A}}_t\boldsymbol{\eta}_{t-1}^{\text{var}}+\boldsymbol{\zeta}_t, & \boldsymbol{\eta}_0^{\text{var}}&=\mathbf{0}.\end{aligned} \tag{12}$$

注意，$\boldsymbol{\eta}_t=\boldsymbol{\eta}_t^{\text{bias}}+\boldsymbol{\eta}_t^{\text{var}}$，因为动力学是线性的。根据上述偏差-方差分解，我们构造$\mathbf{r}$和$\mathbf{v}$ 如下。

定义 1. 设

$$\mathbf{r}=\frac{2}{h_kT}\left(\sum_{t=T/2+1}^{T}\boldsymbol{\eta}_t^{\text{bias}}\right)_1,\quad \mathbf{v}=\frac{2}{h_kT}\left(\sum_{t=T/2+1}^{T}\boldsymbol{\eta}_t^{\text{var}}\right)_1,$$

其中，对于 $\boldsymbol{\eta}\in\mathbb{R}^{2d}$，$\boldsymbol{\eta}_1\in\mathbb{R}^d$ 表示 $\boldsymbol{\eta}$ 的前 d 个元素。

D.1.2 协方差动态

定义

$$\mathbf{R}=\mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\varepsilon^2\mathbf{a}\mathbf{a}^\top\right).$$

其中 ε 定义于（9）中。我们在 $\mathbb{R}^{d\times d}$上引入以下线性算子，

$$\mathcal{B}=\mathbb{E}(\hat{\mathbf{A}}_t\otimes\hat{\mathbf{A}}_t),\quad \tilde{\mathcal{B}}=\mathbf{A}\otimes\mathbf{A},$$

其中，$\mathbf{A}\otimes\mathbf{B}$定义了作用于矩阵$\mathbf{A}$ 和$\mathbf{B}$的线性矩阵算子，满足 $(\mathbf{A}\otimes\mathbf{B})\circ\mathbf{C}=\mathbf{A}\mathbf{C}\mathbf{B}^\top$，而 $\circ$ 表示线性矩阵算子对矩阵的作用，并且。令

$$\mathcal{M}=\mathbb{E}(\mathbf{a}\otimes\mathbf{a}\otimes\mathbf{a}\otimes\mathbf{a}),\quad \tilde{\mathcal{M}}=\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma}. \tag{13}$$

那么，对于$\mathbf{M} = \begin{pmatrix} \mathbf{M}_{11} & \overset{12}{\mathbf{M}_{22}} \end{pmatrix} \in \mathbb{R}^{2d\times 2d}$

$\mathbf{M}_{21}$ $\mathbf{M}$,

$$(\mathcal{B} - \tilde{\mathcal{B}}) \circ \mathbf{M} = \begin{pmatrix} \eta^2(\mathcal{M} - \tilde{\mathcal{M}}) \circ \mathbf{M}_{22} & \eta q(\mathcal{M} - \tilde{\mathcal{M}}) \circ \mathbf{M}_{22} \\ \eta q(\mathcal{M} - \tilde{\mathcal{M}}) \circ \mathbf{M}_{22} & q^2(\mathcal{M} - \tilde{\mathcal{M}}) \circ \mathbf{M}_{22} \end{pmatrix}.$$

Let

$$\mathbf{C}_t = \mathbb{E}\left(\boldsymbol{\eta}_t^{\text{var}} (\boldsymbol{\eta}_t^{\text{var}})^\top\right),$$

那么我们可以将$\mathbf{C}_t$ 的更新写为

$$\mathbf{C}_t = \mathcal{B} \circ \mathbf{C}_{t-1} + \begin{pmatrix} \eta^2\mathbf{R} & \eta q\mathbf{R} \\ \eta q\mathbf{R} & q^2\mathbf{R} \end{pmatrix}, \quad \mathbf{C}_0 = \mathbf{O}.$$

D.1.3 动量矩阵的性质

注意到 $\mathbf{A}$ 是一个块对角矩阵，并且 i阶块[3] 为

$$\mathbf{A}_i = \begin{pmatrix} 0 & 1 - \eta\lambda_i \\ c & 1 + c - q\lambda_i \end{pmatrix},$$

其中 λ_i 是 $\mathbf{\Sigma}$的第 i个特征值。我们定义

$$\mathbf{A}(\lambda) = \begin{pmatrix} 0 & 1 - \eta\lambda \\ c & 1 + c - q\lambda \end{pmatrix}.$$

引理 6。我们有

$$(\mathbf{I} - \mathbf{A}(\lambda))^{-1} \begin{pmatrix} \eta \\ q \end{pmatrix} = \frac{1}{\lambda} \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

证明。 $\mathbf{I} - \mathbf{A}(\lambda)$的逆矩阵为

$$(\mathbf{I} - \mathbf{A}(\lambda))^{-1} = \frac{1}{(q - c\eta)\lambda} \begin{pmatrix} q\lambda - c & 1 - \eta\lambda \\ -c & 1 \end{pmatrix}.$$

因此,

$$(\mathbf{I} - \mathbf{A}(\lambda))^{-1} \begin{pmatrix} \eta \\ q \end{pmatrix} = \frac{1}{\lambda(q - c\eta)} \begin{pmatrix} \eta(q\lambda - c) + q(1 - \eta\lambda) \\ c\eta + q \end{pmatrix} = \frac{1}{\lambda} \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

证明至此完成。 □

动量矩阵的谱性质。A 的特征值(λ) 为

$$x_1(\lambda) = \frac{1 + c - q\lambda}{2} - \frac{\sqrt{(1 + c - q\lambda)^2 - 4c(1 - \eta\lambda)}}{2},$$
$$x_1(\lambda) = \frac{1 + c - q\lambda}{2} + \frac{\sqrt{(1 + c - q\lambda)^2 - 4c(1 - \eta\lambda)}}{2}.$$

[3]更准确地说，$\mathbf{A}_i$是由$\mathbf{A}$中条目 (i, i) 、 $(d+i, i)$ 、 $(i, d+i)$ 和 $(d+i, d+i)$ 组成的子矩阵。

设 $(1+c-q\lambda)^2-4c(1-\eta\lambda)<0$，我们得到满足 $x_1(\lambda)\ x_2(\lambda)\in\mathbb{C}$的 λ 区域:

$$\underbrace{\frac{(1-c)^2}{\left(\sqrt{q-c\eta}+\sqrt{c(q-\eta)}\right)^2}}_{\lambda_L}<\lambda<\underbrace{\frac{(1-c)^2}{\left(\sqrt{q-c\eta}-\sqrt{c(q-\eta)}\right)^2}}_{\lambda_U}.$$

我们采用引理 E.2，引自 Li 等人 [2024], ，该引理给出 $x_1(\lambda)$ 和 $x_2(\lambda)$ 的界。

引理 7。设 $\lambda\geqslant 0$ 和 $\eta\lambda<1$。

- 若 $\lambda\leqslant\lambda_L$，则 $x_1(\lambda)$ 与 $x_2(\lambda)$ 为实数，且 $x_1(\lambda)\leqslant x_2(\lambda)\leqslant 1-(q-c\delta)\lambda/(1-c)$。
- 若 $\lambda_L<\lambda<\lambda_U$，则 $x_1(\lambda)$ 与 $x_2(\lambda)$ 为复数，且 $\|x_1(\lambda)\|=\|x_2(\lambda)\|=\sqrt{c\,(1-\delta\lambda)}$。
- 若 $\lambda\geqslant\lambda_U$，则 $x_1(\lambda)$ 与 $x_2(\lambda)$ 为实数，且 $x_1(\lambda)\leqslant x_2(\lambda)\leqslant c\delta/q$。

动量矩阵幂的界。k次幂$\mathbf{A}(\lambda)$为

$$\mathbf{A}^k(\lambda)=\begin{pmatrix}-c(1-\eta\lambda)a_{k-1}(\lambda) & (1-\eta\lambda)a_k(\lambda)\\ -ca_k(\lambda) & a_{k+1}(\lambda)\end{pmatrix},\quad\text{where}\quad a_k(\lambda)=\frac{x_2^k(\lambda)-x_1^k(\lambda)}{x_2(\lambda)-x_1(\lambda)},$$

且 $a_k(\lambda)\in\mathbb{R}$。

下面的引理由 [Liu 等人的 2025], 引理 20 修改而来，该引理给出了$\mathbf{A}^k(\lambda)$的界。为完整起见，我们给出证明。

引理 8。假设 $\eta\lambda<1$ 且 $\eta\leqslant q$，则有

$$\left\|\left(\mathbf{A}^k(\lambda)\begin{pmatrix}1\\1\end{pmatrix}\right)_1\right\|\leqslant 2.$$

*Pr*明。为简化记号，我们将 $a_k=a_k(\lambda)$ 和 $x_{1,2}=x_{1,\ 2}(\lambda)$ 记作。我们有

$$\left(\mathbf{A}^k(\lambda)\begin{pmatrix}1\\1\end{pmatrix}\right)_1=(1-\eta\lambda)(a_k-ca_{k-1}).$$

该结论对 $k=0$成立，因此我们需要证明对于 $k\geqslant 1$，$\|a_k-ca_{k-1}\|\leqslant 2$ 成立。

1. 若 $\lambda\leqslant\lambda_L$，根据引理 7，以及 $\delta\leqslant q$，我们有 $a_k\geqslant 0$，并且

$$x_1\leqslant x_2\leqslant 1-\frac{q-c\delta}{1-c}\lambda\leqslant 1-\delta\lambda.$$

根据韦达公式，$x_1x_2=c(1-\delta\lambda)$，因此 $c\leqslant x_1\leqslant x_2$。由此，我们将 a_k-ca_{k-1} 界定如下:

$$\begin{aligned}a_k-ca_{k-1}&\geqslant a_k-x_1a_{k-1}=x_2^{k-1}>0,\\ a_k-ca_{k-1}&\leqslant a_k-x_1x_2a_{k-1}=x_2^{k-1}+(1-x_2)\sum_{i=1}^{k-1}x_1^ix_2^{k-i-1}\\ &\leqslant x_2^{k-1}\left[1+(k-1)(1-x_2)\right]\overset{a}{\leqslant}1,\end{aligned}$$

^{a}其中 $\leqslant$ 使用引理 35 和 $k\geqslant 1$。

2. 如果 $\lambda_L < \lambda < \lambda_U$，令 $x_{1,\,2} = r(\cos\,\theta \pm \mathrm{i}\sin\,\theta)$，其中 $0 \leqslant \theta \leqslant \pi/2$。则我们有 $r = \sqrt{c(1-\delta\lambda)} \leqslant 1$ 和 $2r\cos\theta = x_1 + x_2 = 1 + c - q\lambda \geqslant 0$。因此，

$$\begin{aligned} a_k - ca_{k-1} &= \frac{r^{k-1}\sin(k\theta)}{\sin\theta} - \frac{r^{k-2}\sin((k-1)\theta)}{\sin\theta} \\ &= r^{k-2}\left(r\cos((k-1)\theta) + \frac{r-c}{\sin\theta}\sin((k-1)\theta) - r\tan\frac{\theta}{2}\sin((k-1)\theta)\right), \end{aligned}$$

应用三角不等式，以及 $\|\sin k\theta\| \leqslant 1$、$\|\cos\,k\theta\| \leqslant 1$、$\|\tan(\theta/2)\| \leqslant 1$

$$\|a_k - ca_{k-1}\| \overset{a}{\leqslant} r^{k-2}\left(r + (k-1)(1-r)\right) + r^{k-1} = r^{k-2}\left(1 + (k-2)(1-r)\right) + r^{k-1} \overset{b}{\leqslant} 2,$$

其中 $\overset{a}{\leqslant}$ 成立，因为 $r^2 \leqslant c \leqslant 1$ 意味着 $\|r - c\| \leqslant 1 - r$ 且 $\|\sin(k\theta)/\sin\,\theta\| \leqslant k$，$\overset{b}{\leqslant}$ 由引理 35可得，$k \geqslant 1$ 且 $0 \leqslant r \leqslant 1$。

3. 若 $\lambda > \lambda_U$，则有 $a_k \geqslant 0$，并且 $x_1 \leqslant x_2 \leqslant c\delta/q \leqslant c$ 根据引理 7，且 $\delta \leqslant q$。于是得到以下界：

$$\begin{aligned} a_k - ca_{k-1} \geqslant & a_k - a_{k-1} = \sum_{i=0}^{k-1} x_1^i x_2^{k-i-1} - \sum_{i=0}^{k-2} x_1^i x_2^{k-i-2} \\ = & x_1^{k-1} - (1-x_2)\sum_{i=0}^{k-2} x_1^i x_2^{k-i-2} \\ \geqslant & x_1^{k-1} - (k-1)(1-x_2)x_2^{k-2} \\ \geqslant & -x_2^{k-2}\left(1 + (k-2)x_2^{k-1}\right) \overset{a}{\geqslant} -1, \\ a_k - ca_{k-1} \leqslant & a_k - x_2 a_{k-1} = x_1^{k-1} \leqslant 1, \end{aligned}$$

$_a$其中 $\geqslant$ 如果由引理 35 和 $k \geqslant 1$。

结合上述各情况，即完成证明。 □

D.2 方差上界

结构安排。下文将介绍 $L_{\mathrm{var}}(\{\mathbf{X}_t\}_{t\in\mathbb{N}})$，它的作用类似于协方差矩阵 $\{\mathbf{X}_t\}_{t\in\mathbb{N}}$的损失函数。在附录 D.2.1 中，我们陈述层剥离分解技术。我们介绍动力学 $\tilde{\mathbf{C}}_t^{(\ell)}$，并在引理 13 中建立关系 $\mathbf{C}_t = \tilde{\mathbf{C}}_t + \sum_{\ell=1}^{\infty}\tilde{\mathbf{C}}_t^{(\ell)}$。在附录 D.2.3 中，我们分解式 (25) 中定义的噪声矩阵 R；这是建立依赖于极小化点 $\mathbf{x}^*$ 处噪声协方差矩阵（即 Q）的统计项的关键步骤。在附录 D.2.2 中，我们构造式 (21) 中的核心辅助迭代 $\mathbf{\Pi}_t(\mathbf{M})$，其中协方差矩阵为一般的半正定矩阵 M。在附录 D.2.4 中，我们令 $\mathbf{M} = \mathbf{R}$，以得到 $\mathbf{C}_t$的界。在附录 D.2.5 中，我们通过递归地获得 $\tilde{\mathbf{C}}_t^{(\ell)}$ 的界，来界定 $\tilde{\mathbf{C}}_t^{(\ell)}$。最后，结合附录 D.2.4 和附录 D.2.5 中的界，我们通过证明附录 D.2.6 中的引理 9，得到方差界。

本节的目标是证明以下引理。

引理 9（方差上界）。 假设 *1*、 *2*、 *3*、 *4* 和 *5* 成立。令 $\mathbf{v}$ 由定义 *1* 定义，则 $\mathbb{E}\mathbf{v} = \mathbf{0}$ 且

$$\mathbb{E}\|\mathbf{v}\|_{\mathbf{\Sigma}}^2 \leqslant \frac{320\left(3\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{Q}) + 8\eta\tilde{\kappa}L_\ell \operatorname{tr}\mathbf{Q}\right)}{T} + \frac{160L_\ell(6L + \tilde{\kappa}(7 + 16\eta B))(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*))}{T},$$

其中 期望是针对第 k次外层迭代中抽取的样本计算的 n.

证明推迟到附录D.2.6。
为简化记号，我们在半正定矩阵序列上引入如下映射 L_{var} 。

定义 2。 对于半正定矩阵序列 $\{\mathbf{X}_t \in \mathbb{R}^{2d\times 2d}\}_{t\in\mathbb{N}}$，我们定义

$$L_{\text{var}}(\{\mathbf{X}_t\}_{t\in\mathbb{N}}) \stackrel{\text{def}}{=} \frac{4}{T^2}\left\langle \begin{pmatrix}\mathbf{\Sigma} & \mathbf{O}\\ \mathbf{O} & \mathbf{O}\end{pmatrix}, \sum_{s,t=T/2+1}^{T} \mathbf{A}^{\max\{s-t,0\}}\mathbf{X}_{\min\{s,t\}}\left(\mathbf{A}^{\max\{t-s,0\}}\right)^\top \right\rangle. \tag{14}$$

映射 L_{var} 具有如下意义下的线性性和单调性。

引理 10. 对于半正定矩阵序列 $\{\mathbf{X}_t\}_{t\in\mathbb{N}}$ 和 $\{\mathbf{T}_t\}_{t\in\mathbb{N}}$，有

$$L_{\text{var}}(\{\alpha\mathbf{X}_t + \beta\mathbf{Y}_t\}_{t\in\mathbb{N}}) = \alpha L_{\text{var}}(\{\mathbf{X}_t\}_{t\in\mathbb{N}}) + \beta L_{\text{var}}(\{\mathbf{Y}_t\}_{t\in\mathbb{N}}). \tag{15}$$

引理 11. 对于半正定矩阵序列 $\{\mathbf{X}_t\}_{t\in\mathbb{N}}$ 和 $\{\mathbf{Y}_t\}_{t\in\mathbb{N}}$，若对所有 $t\in\mathbb{N}$均有 $\mathbf{X}_t \preceq \mathbf{Y}_t$，则有

$$L_{\text{var}}(\{\mathbf{X}_t\}_{t\in\mathbb{N}}) \leqslant L_{\text{var}}(\{\mathbf{Y}_t\}_{t\in\mathbb{N}}). \tag{16}$$

由 L_{var}的定义可直接得到证明。我们可以将 $\mathbb{E}\|\mathbf{v}\|_{\mathbf{\Sigma}}^2$ 表示为 L_{var}，如下所示。

引理 12. 我们有

$$\mathbb{E}\|\mathbf{v}\|_{\mathbf{\Sigma}}^2 = L_{\text{var}}(\{\mathbf{C}_t\}_{t\in\mathbb{N}}).$$

证明。 根据$\mathbf{v}$的定义，有

$$\begin{aligned}
\mathbb{E}\|\mathbf{v}\|_{\mathbf{\Sigma}}^2 &= \left\langle \begin{pmatrix}\mathbf{\Sigma} & \mathbf{O}\\ \mathbf{O} & \mathbf{O}\end{pmatrix}, \mathbb{E}\left[\left(\frac{2}{T}\sum_{t=T/2+1}^{T}\boldsymbol{\eta}_t^{\text{var}}\right)\left(\frac{2}{T}\sum_{t=T/2+1}^{T}\boldsymbol{\eta}_t^{\text{var}}\right)^\top\right]\right\rangle \\
&= \left\langle \begin{pmatrix}\mathbf{\Sigma} & \mathbf{O}\\ \mathbf{O} & \mathbf{O}\end{pmatrix}, \frac{4}{T^2}\sum_{s,t=T/2+1}^{T}\mathbb{E}(\boldsymbol{\eta}_s^{\text{var}}\boldsymbol{\eta}_t^{\text{var}})\right\rangle, \\
&\stackrel{a}{=} \left\langle \begin{pmatrix}\mathbf{\Sigma} & \mathbf{O}\\ \mathbf{O} & \mathbf{O}\end{pmatrix}, \frac{4}{T^2}\sum_{s,t=T/2+1}^{T}\mathbf{A}^{\max\{s-t,0\}}\mathbf{C}_{\min\{s,t\}}\left(\mathbf{A}^{\max\{t-s,0\}}\right)^\top\right\rangle,
\end{aligned}$$

其中 $\stackrel{a}{=}$使用

$$\mathbb{E}\boldsymbol{\eta}_s^{\text{var}}(\boldsymbol{\eta}_t^{\text{var}})^\top = \mathbb{E}\left(\mathbb{E}\left(\boldsymbol{\eta}_s^{\text{var}}|\mathcal{F}_t\right)(\boldsymbol{\eta}_t^{\text{var}})^\top\right) = \mathbf{A}^{s-t}\mathbb{E}\boldsymbol{\eta}_t^{\text{var}}(\boldsymbol{\eta}_t^{\text{var}})^\top = \mathbf{A}^{s-t}\mathbf{C}_t, \quad s\geqslant t$$

并收集各项。 □

量 $L_{\mathrm{var}}(\{\mathbf{X}_t\}_{t\in\mathbb{N}})$ 可解释为协方差为 $\mathbf{X}_t$ 的尾部平均序列所产生的损失。具体而言，我们按如下方式构造动力学 $\boldsymbol{\psi}_t \in \mathbb{R}^{2d\times 2d}$：

$$\boldsymbol{\psi}_t = \mathbf{A}\boldsymbol{\psi}_{t-1} + \begin{pmatrix} \eta\boldsymbol{\xi}_t \\ q\boldsymbol{\xi}_t \end{pmatrix}, \quad \boldsymbol{\psi}_0 = \mathbf{0},$$

其中 $\boldsymbol{\xi}_t \sim$独立同分布 $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}})$，并且 $\mathbb{E}\boldsymbol{\psi}_t\boldsymbol{\psi}_t^\top = \mathbf{X}_t$，那么我们有（证明与引理 12 的证明相同）

$$L_{\mathrm{var}}(\{\mathbf{X}_t\}_{t\in\mathbb{N}}) = \left\langle \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}, \mathbb{E}\left[\left(\frac{2}{T}\sum_{t=T/2+1}^{T} \boldsymbol{\psi}_t \right) \left(\frac{2}{T}\sum_{t=T/2+1}^{T} \boldsymbol{\psi}_t \right)^\top \right] \right\rangle.$$

然而，对于任意序列 $\{\mathbf{X}_t\}_{t\in\mathbb{N}}$，都可以计算 $L_{\mathrm{var}}(\{\mathbf{X}_t\}_{t\in\mathbb{N}})$，且不一定要求$\mathbf{X}_t$ 是某种动力学的协方差 $\boldsymbol{\psi}_t$。

D.2.1 层剥离分解

本节的目标是界定 $\mathbb{E}\|\mathbf{v}\|_{\boldsymbol{\Sigma}}^2 = L_{\mathrm{var}}(\{\mathbf{C}_t\}_{t\in\mathbb{N}})$。为精确刻画 $\mathbf{C}_t$，本文提出了**层剥离分解**，将协方差矩阵$\mathbf{C}_t$ 分解为以下动力学之和。

定义 3（第0层动力学）。设

$$\tilde{\mathbf{C}}_t = \mathbf{A}\tilde{\mathbf{C}}_{t-1}\mathbf{A}^\top + \begin{pmatrix} \eta^2\mathbf{R} & \eta q\mathbf{R} \\ \eta q\mathbf{R} & q^2\mathbf{R} \end{pmatrix}, \quad \tilde{\mathbf{C}}_0 = \mathbf{O}, \tag{17}$$

定义 4（层 ℓ 动力学）。设

$$\tilde{\mathbf{C}}_t^{(\ell)} = \mathbf{A}\tilde{\mathbf{C}}_{t-1}^{(\ell)}\mathbf{A}^\top + (\mathcal{B} - \tilde{\mathcal{B}}) \circ \tilde{\mathbf{C}}_{t-1}^{(\ell-1)}, \quad \tilde{\mathbf{C}}_0^{(\ell)} = \mathbf{O}. \tag{18}$$

我们将 (17) 称为第0层动力学，并将 (18) 称为第 ℓ 层动力学。下面的引理表明 $\mathbf{C}_t = \tilde{\mathbf{C}}_t + \sum_{\ell=1}^{\infty} \tilde{\mathbf{C}}_t^{(\ell)}$。

引理 13. 我们有

$$\mathbf{C}_t = \tilde{\mathbf{C}}_t + \sum_{\ell=1}^{\infty} \tilde{\mathbf{C}}_t^{(\ell)}.$$

我们还有 $\tilde{\mathbf{C}}_t^{(\ell)} = \mathbf{O}$ 对于 $\ell \geqslant t$，因此上述展开式是有限和。

证明。我们构造一系列动力学（由 ℓ索引），如下所示：

$$\mathbf{C}_t^{(\ell)} = \mathcal{B} \circ \mathbf{C}_{t-1}^{(\ell)} + (\mathcal{B} - \tilde{\mathcal{B}}) \circ \tilde{\mathbf{C}}_{t-1}^{(\ell-1)}, \quad \mathbf{C}_0^{(\ell)} = \mathbf{O}. \tag{19}$$

回忆一下，$\tilde{\mathcal{B}} = \mathbf{A} \otimes \mathbf{A}$，因此 $\mathbf{C}_t = \tilde{\mathbf{C}}_t + \mathbf{C}_t^{(1)}$，$\mathbf{C}_t^{(\ell)} = \tilde{\mathbf{C}}_t^{(\ell)} + \mathbf{C}_t^{(\ell+1)}$，并且

$$\mathbf{C}_t^{(\ell)} = \mathbf{O}, \quad \tilde{\mathbf{C}}_t^{(\ell)} = \mathbf{O}, \quad \text{for all } 0 \leqslant t \leqslant \ell,$$

因此我们得到如下分解:

$$\begin{aligned}\mathbf{C}_t^{(1)} &= \tilde{\mathbf{C}}_t^{(1)} + \mathbf{C}_t^{(2)} \\ &= \tilde{\mathbf{C}}_t^{(1)} + \tilde{\mathbf{C}}_t^{(2)} + \mathbf{C}_t^{(3)} \\ &\cdots \\ &= \sum_{\ell=1}^{\infty} \tilde{\mathbf{C}}_t^{(\ell)}. \qquad (\text{note that } \tilde{\mathbf{C}}_t^{(\ell)} = \mathbf{O} \text{ for all } \ell \geqslant t)\end{aligned} \tag{20}$$

证明至此完成。 □

D.2.2 核心辅助动力学

在本节中，我们分析以下动力学 $\tilde{\mathbf{\Pi}}_t(\mathbf{M}) \in \mathbb{R}^{2d\times 2d}$,

$$\tilde{\mathbf{\Pi}}_t(\mathbf{M}) = \mathbf{A}\tilde{\mathbf{\Pi}}_{t-1}(\mathbf{M})\mathbf{A}^\top + \begin{pmatrix}\eta^2\mathbf{M} & \eta q\mathbf{M} \\ \eta q\mathbf{M} & q^2\mathbf{M}\end{pmatrix}, \quad \tilde{\mathbf{\Pi}}_0 = \mathbf{O}, \tag{21}$$

˜其中 $\mathbf{M}$ 是任意半正定矩阵。为简化记号，我们还记 $\mathbf{\Pi}_t$ 。我们重点关注[4] 的对角线 $\tilde{\mathbf{\Pi}}_t$，即

$$\tilde{\mathbf{\Pi}}_{t,i} = \mathbf{A}_i\tilde{\mathbf{\Pi}}_{t-1,i}\mathbf{A}_i^\top + m_i\begin{pmatrix}\eta^2 & \eta q \\ \eta q & q^2\end{pmatrix},$$

其中 m_i 表示$\mathbf{M}$ 的第 i 个对角元素。

˜ ˜**关于** $L_{\text{var}}(\{\mathbf{\Pi}_t(\mathbf{M})\}_{t\in\mathbb{N}})$**的上界**。 我们对 $L_{\text{var}}(\{\mathbf{\Pi}_t(\mathbf{M})\}_{t\in\mathbb{N}})$ 求上界。第一步是推导其矩阵形式。

引理14. 我们有 $L_{\text{var}}(\{\tilde{\mathbf{\Pi}}_t(\mathbf{M})\}_{t\in\mathbb{N}}) = 4(A_{T/2} + B_{T/2})/T^2$，其中

$$\begin{aligned}A_m &= \left\langle \begin{pmatrix}\mathbf{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O}\end{pmatrix}, \left(\sum_{k=0}^{T-m-1}\mathbf{A}^k\right)\tilde{\mathbf{\Pi}}_{m+1}\left(\sum_{k=0}^{T-m-1}\mathbf{A}^k\right)^\top \right\rangle, \\ B_m &= \sum_{s=1}^{T-m-1}\left\langle \begin{pmatrix}\mathbf{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O}\end{pmatrix}, \left(\sum_{k=0}^{s-1}\mathbf{A}^k\right)\begin{pmatrix}\eta^2\mathbf{M} & \eta q\mathbf{M} \\ \eta q\mathbf{M} & q^2\mathbf{M}\end{pmatrix}\left(\sum_{k=0}^{s-1}\mathbf{A}^k\right)^\top \right\rangle.\end{aligned}$$

证明。 我们用归纳法证明

$$L_m \overset{\text{def}}{=} \left\langle \begin{pmatrix}\mathbf{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O}\end{pmatrix}, \sum_{s,t=m+1}^{T}\mathbf{A}^{\max\{s-t,0\}}\tilde{\mathbf{\Pi}}_{\min\{s,t\}}\left(\mathbf{A}^{\max\{t-s,0\}}\right)^\top \right\rangle = A_m + B_m.$$

[4]更准确地说，$\tilde{\mathbf{\Pi}}_{t,\,i}$ 是由 $\tilde{\mathbf{\Pi}}_{t,\,i}$ 中条目 $(i,\ \ i)$、$(d+i,\ \ i)$、$(i,\ \ d+i)$ 和 $(d+i,\ \ d+i)$ 构成的 2×2 子矩阵 $\tilde{\mathbf{\Pi}}_t$。

上述等式对 $m = T-1$ 成立。假设上述等式对 $m \leqslant T-1$ 成立，我们证明它对 $m-1$ 也成立。首先，注意到

$$\begin{aligned} L_{m-1} - L_m = &\left\langle \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}, \tilde{\boldsymbol{\Pi}}_m \right\rangle + \left\langle \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}, \left(\sum_{k=1}^{T-m} \mathbf{A}^k \right) \tilde{\boldsymbol{\Pi}}_m \right\rangle \\ &+ \left\langle \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}, \tilde{\boldsymbol{\Pi}}_m \left(\sum_{k=1}^{T-m} \mathbf{A}^k \right)^\top \right\rangle. \end{aligned} \tag{22}$$

因此，我们有

$$\begin{aligned} & A_{m-1} - A_m \\ = &\left\langle \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}, \left(\mathbf{A} \sum_{k=0}^{T-m-1} \mathbf{A}^k + \mathbf{I} \right) \tilde{\boldsymbol{\Pi}}_m \left(\mathbf{A} \sum_{k=0}^{T-m-1} \mathbf{A}^k + \mathbf{I} \right)^\top \right\rangle \\ &- \left\langle \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}, \left(\sum_{k=0}^{T-m-1} \mathbf{A}^k \right) \tilde{\boldsymbol{\Pi}}_{m+1} \left(\sum_{k=0}^{T-m-1} \mathbf{A}^k \right)^\top \right\rangle \\ \overset{a}{=} &(L_{m-1} - L_m) + \left\langle \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}, \left(\sum_{k=0}^{T-m-1} \mathbf{A}^k \right) \left(\tilde{\boldsymbol{\Pi}}_{m+1} - \mathbf{A} \tilde{\boldsymbol{\Pi}}_m \mathbf{A}^\top \right) \left(\sum_{k=0}^{T-m-1} \mathbf{A}^k \right)^\top \right\rangle \\ \overset{b}{=} &(L_{m-1} - L_m) - \left\langle \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}, \left(\sum_{k=0}^{T-m-1} \mathbf{A}^k \right) \begin{pmatrix} \eta^2 \mathbf{M} & \eta q \mathbf{M} \\ \eta q \mathbf{M} & q^2 \mathbf{M} \end{pmatrix} \left(\sum_{k=0}^{T-m-1} \mathbf{A}^k \right)^\top \right\rangle \\ = &(L_{m-1} - L_m) - (B_{m-1} - B_m), \end{aligned}$$

其中，我们展开乘积，并在22）中应用 $\overset{a}{=}$，而 $\overset{b}{=}$ 则应用了 $\tilde{\boldsymbol{\Pi}}_t$ 在（21）中的更新规则。因此，$A_{m-1} + B_{m-1} = A_m + B_m +$（$L_{m-1} - L_m$） $= L_{m-1}$。 □

˜我们进一步推导出 L_{var}（$\{\boldsymbol{\Pi}_t$（$\mathbf{M}$）$\}_{t\in\mathbb{N}}$）的逐维求和形式。

引理15。设 m_i 表示$\mathbf{M}$对角线上的第 i个元素，则

$$\begin{aligned} L_{\text{var}}(\{\tilde{\boldsymbol{\Pi}}_t(\mathbf{M})\}_{t\in\mathbb{N}}) = &\frac{4}{T^2} \sum_{i=1}^{d} \sum_{k=0}^{T/2} \lambda_i m_i \left(\mathbf{A}_i^k \left(\mathbf{I} - \mathbf{A}_i^{T/2} \right) (\mathbf{I} - \mathbf{A}_i)^{-1} \begin{pmatrix} \eta \\ q \end{pmatrix} \right)_1^2 \\ &+ \frac{4}{T^2} \sum_{i=1}^{d} \sum_{k=1}^{T/2} \lambda_i m_i \left(\left(\mathbf{I} - \mathbf{A}_i^{T/2-k} \right) (\mathbf{I} - \mathbf{A}_i)^{-1} \begin{pmatrix} \eta \\ q \end{pmatrix} \right)_1^2. \end{aligned}$$

证明。 根据引理 14，我们有 $L_{\text{var}}(\{\tilde{\boldsymbol{\Pi}}_t(\mathbf{M})\}_{t\in\mathbb{N}}) = 4(A_{T/2} + B_{T/2})/T^2$，其中

$$\begin{aligned} A_{T/2} &= \left\langle \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}, \left(\sum_{k=0}^{T/2-1} \mathbf{A}^k \right) \tilde{\boldsymbol{\Pi}}_{T/2+1} \left(\sum_{k=0}^{T/2-1} \mathbf{A}^k \right)^\top \right\rangle, \\ B_{T/2} &= \sum_{k=1}^{T/2} \left\langle \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix}, \left(\sum_{k=0}^{T/2-k-1} \mathbf{A}^k \right) \begin{pmatrix} \eta^2 \mathbf{M} & \eta q \mathbf{M} \\ \eta q \mathbf{M} & q^2 \mathbf{M} \end{pmatrix} \left(\sum_{k=0}^{T/2-k-1} \mathbf{A}^k \right)^\top \right\rangle. \end{aligned}$$

注意， $B_{T/2}$ 中的矩阵是块对角的。因此，

$$B_{T/2} = \sum_{i=1}^{d} \sum_{k=1}^{T/2} \lambda_i m_i \left(\left(\mathbf{I} - \mathbf{A}_i^{T/2-k} \right) (\mathbf{I} - \mathbf{A}_i)^{-1} \begin{pmatrix} \eta \\ q \end{pmatrix} \right)_1^2 .$$

根据 $\mathbf{\Pi}$ 在（21）中的定义，我们有

$$\tilde{\mathbf{\Pi}}_{T/2+1} = \sum_{k=0}^{T/2} \mathbf{A}^k \begin{pmatrix} \eta^2 \mathbf{M} & \eta q \mathbf{M} \\ \eta q \mathbf{M} & q^2 \mathbf{M} \end{pmatrix} \left(\mathbf{A}^k \right)^\top .$$

因此， $A_{T/2}$ 具有以下形式：

$$A_{T/2} = \sum_{i=1}^{d} \sum_{k=0}^{T/2} \lambda_i m_i \left(\mathbf{A}_i^k \left(\mathbf{I} - \mathbf{A}_i^{T/2} \right) (\mathbf{I} - \mathbf{A}_i)^{-1} \begin{pmatrix} \eta \\ q \end{pmatrix} \right)_1^2 ,$$

其中 $(\cdot)_1$ 表示二维向量的第一个元素。通过应用$L_{\mathrm{var}}(\{\tilde{\mathbf{\Pi}}_t(\mathbf{M})\}_{t\in\mathbb{N}}) = 4(A_{T/2} + B_{T/2})/T^2$，即可得到结果。 □

下面的引理给出了 $L_{\mathrm{var}}(\{\mathbf{\Pi}_t(\mathbf{M})\}_{t\in\mathbb{N}})$的界。

引理16。假　　设我们按照附录 C中指定的方式选择超参数，则有

$$L_{\mathrm{var}}(\{\tilde{\mathbf{\Pi}}_t(\mathbf{M})\}_{t\in\mathbb{N}}) \leqslant \frac{64 \operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{M})}{T}.$$

证明。 根据引理 6 和引理 8，我们有

$$A_{T/2} = \sum_{i=1}^{d} \sum_{k=0}^{T/2} \frac{m_i}{\lambda_i} \left(\mathbf{A}_i^k \left(\mathbf{I} - \mathbf{A}_i^{T/2} \right) \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right)_1^2 \leqslant \sum_{i=1}^{d} \sum_{k=0}^{T/2} \frac{16 m_i}{\lambda_i} = \frac{8 \operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{M})}{T},$$

$$B_{T/2} = \sum_{i=1}^{d} \sum_{k=1}^{T/2} \lambda_i m_i \left(\left(\mathbf{I} - \mathbf{A}_i^{T/2-k} \right) \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right)_1^2 \leqslant \sum_{i=1}^{d} \sum_{k=0}^{T/2} \frac{16 m_i}{\lambda_i} = \frac{8 \operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{M})}{T}.$$

我们通过 $L_{\mathrm{var}}(\{\tilde{\mathbf{\Pi}}_t(\mathbf{M})\}_{t\in\mathbb{N}}) = 4(A_{T/2} + B_{T/2})/T^2 \leqslant 64 \operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{M})/T$ 完成证明。 □

平稳协方差的推导。 我们推导 $\mathbf{\Pi}_t(\mathbf{M})$的平稳状态。下面的引理表明， $\mathbf{\Pi}_t(\mathbf{M})$是递增的。

引理17。 设 $\mathbf{\Pi}_t(\mathbf{M})$ 定义于 (21)。则

$$\tilde{\mathbf{\Pi}}_0(\mathbf{M}) \preceq \tilde{\mathbf{\Pi}}_1(\mathbf{M}) \preceq \cdots \preceq \tilde{\mathbf{\Pi}}_t(\mathbf{M}) \preceq \tilde{\mathbf{\Pi}}_{t+1}(\mathbf{M}) \preceq \cdots$$

证明。 我们展开 $\mathbf{\Pi}_t$ 的递推，得到

$$\tilde{\mathbf{\Pi}}_t(\mathbf{M}) = \sum_{k=0}^{t-1} \mathbf{A}^k \begin{pmatrix} \eta^2 \mathbf{M} & \eta q \mathbf{M} \\ \eta q \mathbf{M} & q^2 \mathbf{M} \end{pmatrix} \left(\mathbf{A}^k \right)^\top .$$

由于它是半正定矩阵的和，因此结论成立。 □

˜我们将 $\mathbf{\Pi}$的平稳状态定义如下：

$$\tilde{\mathbf{\Pi}}_\infty(\mathbf{M}) = \sum_{k=0}^{\infty} \mathbf{A}^k \begin{pmatrix} \eta^2\mathbf{M} & \eta q\mathbf{M} \\ \eta q\mathbf{M} & q^2\mathbf{M} \end{pmatrix} \left(\mathbf{A}^k\right)^\top, \tag{23}$$

˜ ˜因此，对于所有 $t \in \mathbb{N}$， $\mathbf{\Pi}_t(\mathbf{M}) \preceq \mathbf{\Pi}_\infty(\mathbf{M})$。注意，

$$\tilde{\mathbf{\Pi}}_\infty(\mathbf{M}) = \mathbf{A}\tilde{\mathbf{\Pi}}_\infty(\mathbf{M})\mathbf{A}^\top + \begin{pmatrix} \eta^2\mathbf{M} & \eta q\mathbf{M} \\ \eta q\mathbf{M} & q^2\mathbf{M} \end{pmatrix}. \tag{24}$$

˜下面的引理明确计算了 $\mathbf{\Pi}_\infty(\mathbf{M})$。

引理 18. 设 m_i 表示 $\mathbf{M}$ 的第 i 个对角元素，则

$$\tilde{\mathbf{\Pi}}_{\infty,i}(\mathbf{M}) = m_i \begin{pmatrix} * & * \\ * & \dfrac{(1+c)(q-c\eta) + c\eta(q+c\eta)\lambda_i}{(2+2c-(q+c\eta)\lambda_i)(1-c+c\eta\lambda_i)\lambda_i} \end{pmatrix}$$

证明。 注意，(24) 是关于 $\mathbf{M}$ 的线性方程组。所需结果可通过直接计算得到。 □

以下界限 $\left\langle \begin{pmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{\Sigma} \end{pmatrix}, \tilde{\mathbf{\Pi}}_\infty(\mathbf{M}) \right\rangle$。

引理 19。 假设我们按照附录 C 的规定选择超参数，则我们有 e

$$\left\langle \begin{pmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{\Sigma} \end{pmatrix}, \tilde{\mathbf{\Pi}}_\infty(\mathbf{M}) \right\rangle \leqslant \frac{\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{M})}{4\tilde{\kappa}} + 4\eta \operatorname{tr}\mathbf{M}.$$

证明。我们有

$$\begin{aligned} \left\langle \begin{pmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{\Sigma} \end{pmatrix}, \tilde{\mathbf{\Pi}}_\infty(\mathbf{M}) \right\rangle &\leqslant \sum_{i=1}^{d} \lambda_i \left(\tilde{\mathbf{\Pi}}_\infty(\mathbf{M})\right)_{22} \\ &\leqslant \sum_{i=1}^{d} \lambda_i m_i \cdot \frac{(1+c)(q-c\eta) + c\eta(q+c\eta)\lambda_i}{(2+2c-(q+c\eta)\lambda_i)(1-c+c\eta\lambda_i)\lambda_i}. \end{aligned}$$

由于 $\eta \geqslant \theta\gamma$，且

$$q = \frac{\eta+\theta\gamma}{1+\theta} \leqslant \frac{2\eta}{1+\theta} = 2c\eta,$$

因此我们有

$$2+2c-(q+c\eta)\lambda_i \geqslant 2+2c-(2+c)\eta\lambda_i \geqslant 1.$$

对于第一项，

$$\begin{aligned} \frac{(1+c)(q-c\eta)}{(1-c+c\eta\lambda_i)\lambda_i} &= \frac{1+c}{(1-c+c\eta\lambda_i)\lambda_i}\left(q - \frac{(1+c)\eta}{2} + \frac{(1-c)\eta}{2}\right) \\ &\leqslant \frac{1+c}{c\eta\lambda_i^2}\left(q - \frac{(1+c)\eta}{2}\right) + \frac{(1+c)\eta}{2\lambda_i} \\ &= \frac{2\theta\gamma}{(1-\theta^2)\eta\lambda_i^2} + \frac{\eta}{(1+\theta)\lambda_i} \overset{a}{=} \frac{1}{8(1-\theta^2)\tilde{\kappa}\lambda_i^2} + \frac{\eta}{(1+\theta)\lambda_i} \\ &\overset{b}{\leqslant} \frac{1}{4\tilde{\kappa}\lambda_i} + \frac{\eta}{\lambda_i} \end{aligned}$$

其中 $\overset{a}{=}$ 使用 $\eta = 16\tilde{\kappa}\theta\gamma$，且 $\overset{b}{\leqslant}$ 使用 $1-\theta^2 \geqslant 1/2$。对于第二项，

$$\frac{c\eta(q+c\eta)\lambda_i}{(1-c+c\eta\lambda_i)\lambda_i} \overset{a}{\leqslant} \frac{q+c\eta}{\lambda_i} = \frac{(3+\theta)\eta}{(1+\theta)\lambda_i} \leqslant \frac{3\eta}{\lambda_i},$$

$_a$其中，$\leqslant$ 使用了 $c\eta\lambda_i \leqslant 1-c+c\eta\lambda_i$。结合这些结果，我们得到

$$\left\langle \begin{pmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{\Sigma} \end{pmatrix}, \tilde{\mathbf{\Pi}}_\infty(\mathbf{M}) \right\rangle \leqslant \sum_{i=1}^{d} \frac{m_i}{4\tilde{\kappa}\lambda_i} + 4\eta m_i = \frac{\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{M})}{4\tilde{\kappa}} + 4\eta \operatorname{tr}\mathbf{M}.$$

证明至此完成。 □

D.2.3 噪声协方差的分解

定义

$$\mathbf{R} = \mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\left(\varepsilon^2 \mathbf{a}\mathbf{a}^\top\right). \tag{25}$$

其中，ε 在（9）中定义。在本小节中，我们将随机梯度的 $\mathbf{\Sigma}^{-1}$-范数 $\mathbf{tr(\Sigma^{-1}R)}$ 分解为与 $\mathbf{\Sigma^{-1}Q)}$ 对应的、位于 $\mathbf{x}^*$处的噪声项，以及一个与超额风险成正比的附加项 F（˜$\mathbf{y}$）$-F$（$\mathbf{x}^*$），见引理 21。我们在引理 **22** 中对随机梯度的二范数 $\mathbf{tr(R)}$ 推导出类似的分解。这些结果是推导局部化于 $\mathbf{x}^*$ 的统计复杂度的关键。也就是说，得到仅依赖于 $\mathbf{Q}$ 的统计项，其中 $\mathbf{Q}^*$ 是随机梯度在 $\mathbf{x}$ 处的协方差。

我们需要以下引理来界定 F 的梯度范数。

引理 20。 假设 *1* 成立。则对于任意 $\mathbf{x} \in \mathbb{R}^d$，有

$$\|\nabla F(\mathbf{x})\|^2_{\mathbf{\Sigma}^{-1}} \leqslant 2L_\ell(F(\mathbf{x}) - F(\mathbf{x}^*)),$$

其中 $\mathbf{x}^*$ 是 F 的极小化点。

证明。 令 $\mathbf{u} = \mathbf{\Sigma}^{-1}\nabla F(\mathbf{x})/L_\ell$。由于 $\nabla^2 F \preceq L_\ell \mathbf{\Sigma}$，因此

$$F(\mathbf{x}^*) \leqslant F(\mathbf{x}-\mathbf{u}) \leqslant F(\mathbf{x}) - \langle \nabla F(\mathbf{x}), \mathbf{u}\rangle + \frac{1}{2}\|\mathbf{u}\|^2_{L_\ell \mathbf{\Sigma}} = F(\mathbf{x}) - \frac{1}{2L_\ell}\|\nabla F(\mathbf{x})\|^2_{\mathbf{\Sigma}^{-1}}.$$

通过重新整理各项即可得到该结果。 □

根据引理 20，我们可以如下界定 $\mathbf{tr(\Sigma^{-1}R)}$。

引理21（噪声上界）。 假设 *3* 和 *5* 成立。则

$$\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{R}) \leqslant 5\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{Q}) + 5L_\ell(L+\tilde{\kappa})(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*)).$$

证明。 注意，根据（9），我们有

$$\begin{aligned} \varepsilon^2\mathbf{a}\mathbf{a}^\top \preceq\ & 5\left(\ell'(\mathbf{a}^\top\mathbf{x}^*, b)\right)^2 \mathbf{a}\mathbf{a}^\top + \frac{5}{2}\left(\ell'(\mathbf{a}^\top\tilde{\mathbf{y}}, b) - \ell'(\mathbf{a}^\top\mathbf{x}^*, b)\right)^2 \mathbf{a}\mathbf{a}^\top \\ & + \frac{5}{2}\left(\mathbf{a}^\top(\tilde{\mathbf{y}} - \tilde{\mathbf{x}}^*_+)\right)^2 \mathbf{a}\mathbf{a}^\top, \end{aligned} \tag{26}$$

其中我们使用了 $(a+b+c)^2 \leqslant 5a^2+5b^2/2+5c^2/2$。与 $\mathbf{\Sigma}^{-1}$ 取内积并求期望，第一项由 $\mathbf{tr(\Sigma}^{-1}\mathbf{Q)}$ 界定，第二项根据假设 $2L_\ell L(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*))$ 界定，因为 5，$\ell'(\mathbf{a}^\top \tilde{\mathbf{y}}\ b)\,\mathbf{a}$ 是 $F(\tilde{\mathbf{y}})$ 的随机梯度。第三项界定如下

$$\mathbb{E}\left(\mathbf{a}^\top(\tilde{\mathbf{y}} - \tilde{\mathbf{x}}_+^*)\right)^2 \|\mathbf{a}\|_{\mathbf{\Sigma}^{-1}}^2 \overset{a}{\leqslant} \tilde{\kappa}\|\tilde{\mathbf{y}} - \tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}}^2 = \tilde{\kappa}\|\nabla F(\tilde{\mathbf{y}})\|_{\mathbf{\Sigma}^{-1}}^2 \overset{b}{\leqslant} 2L_\ell\tilde{\kappa}(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*)),$$

$\overset{a}{\leqslant}$ 使用假设 3，而 $\overset{b}{\leqslant}$ 使用引理 20。结合这三个界即可得到结果。 □

我们还需要 $\mathbf{trR}$ 的上界。证明与引理 21 的证明类似 .

引理 22。 假设 *3* 和 *4* 成立，则

$$\operatorname{tr}\mathbf{R} \leqslant 5\operatorname{tr}\mathbf{Q} + 5L_\ell(B+R^2)(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*)).$$

证明。我们对（26）的迹进行界定。第一项由 $\mathbf{trQ}$ 界定。第二项由

$$\mathbb{E}\left(\ell'(\mathbf{a}^\top\tilde{\mathbf{y}}, b) - \ell'(\mathbf{a}^\top\mathbf{x}^*, b)\right)^2 \|\mathbf{a}\|^2 \leqslant 2L_\ell B(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*)),$$

其中使用假设 4。第三项由

$$R^2\|\tilde{\mathbf{y}} - \tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}}^2 = R^2\|\nabla F(\tilde{\mathbf{y}})\|_{\mathbf{\Sigma}^{-1}}^2 \overset{a}{\leqslant} 2L_\ell R^2(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*)),$$

其中我们使用假设3。结合这三个界即可得到结果 . □

˜D.2.4 第0 层动力学 $\mathbf{C}_t$

˜根据 $\mathbf{\Pi}$ 式（21）中的定义，我们有

$$\tilde{\mathbf{C}}_t = \tilde{\mathbf{\Pi}}_t(\mathbf{R}).$$

因此，应用 引理16 和引理19 附录D.2.2，可得以下引理 .

引理 23。 *Su* 假设我们按照附录 *C* 中的指定选择超参数，则有

$$L_{\text{var}}\left(\left\{\tilde{\mathbf{C}}_t\right\}_{t\in\mathbb{N}}\right) \leqslant \frac{320\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{Q})}{T} + \frac{320(L+\tilde{\kappa})(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*))}{T}.$$

引理24。 假设我们按照附录 *C* 中指定的方式选择超参数，则有

$$\left\langle \begin{pmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{\Sigma} \end{pmatrix}, \tilde{\mathbf{C}}_\infty \right\rangle \leqslant \frac{5\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{Q}) + 5(L+\tilde{\kappa})(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*))}{4\tilde{\kappa}}$$
$$+20\eta\operatorname{tr}\mathbf{Q} + 20\eta(B+R^2)(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*))$$

D.2.5 层 ℓ 动力学 $\tilde{\mathbf{C}}_t^{(\ell)}$ 的界

下面的引理推导出 $\mathbb{E}\|\mathbf{v}_k\|_{\mathbf{H}}^2$的分解。

引理 25。我们有

$$\mathbb{E}\|\mathbf{v}_k\|_{\mathbf{H}}^2 = L_{\text{var}}\left(\left\{\tilde{\mathbf{C}}_t\right\}_{t\in\mathbb{N}}\right) + \sum_{\ell=1}^{\infty} L_{\text{var}}\left(\left\{\tilde{\mathbf{C}}_t^{(\ell)}\right\}_{t\in\mathbb{N}}\right),$$

证明。 根据引理 13，我们有

$$\mathbf{C}_t = \tilde{\mathbf{C}}_t + \sum_{\ell=1}^{\infty} \tilde{\mathbf{C}}_t^{(\ell)}.$$

所需结果由引理 L_{var} 中的线性性10得到。 □

定义 ℓ 级平稳状态 $\tilde{\mathbf{C}}_\infty^{(\ell)}$ 如下:

$$\tilde{\mathbf{C}}_\infty^{(\ell)} = \begin{pmatrix} \tilde{\mathbf{C}}_{\infty,11}^{(\ell)} & \tilde{\mathbf{C}}_{\infty,12}^{(\ell)} \\ \tilde{\mathbf{C}}_{\infty,21}^{(\ell)} & \tilde{\mathbf{C}}_{\infty,22}^{(\ell)} \end{pmatrix} \stackrel{\text{def}}{=} \tilde{\mathbf{\Pi}}_\infty(\mathbf{R}^{(\ell)}), \quad \text{where } \mathbf{R}^{(\ell)} = (\mathcal{M} - \tilde{\mathcal{M}}) \circ \tilde{\mathbf{C}}_{\infty,22}^{(\ell-1)}.$$

我们设 $\tilde{\mathbf{C}}_\infty^{(0)} = \tilde{\mathbf{C}}_\infty \stackrel{\text{def}}{=} \tilde{\mathbf{\Pi}}_\infty(\mathbf{R})$ 定义于（23）中。可以验证 $\tilde{\mathbf{C}}_t^{(\ell)} \preceq \tilde{\mathbf{C}}_\infty^{(\ell)}$。因此,

$$L_{\text{var}}\left(\left\{\tilde{\mathbf{C}}_t^{(\ell)}\right\}_{t\in\mathbb{N}}\right) \leqslant L_{\text{var}}\left(\left\{\tilde{\mathbf{\Pi}}_t(\mathbf{R}^{(\ell)})\right\}_{t\in\mathbb{N}}\right)$$

我们应用附录 D.2.2 中关于 $\mathbf{M} = \mathbf{R}^{(\ell)}$ 的结果，得到以下引理 .

引理 26. 我们有

$$L_{\text{var}}\left(\left\{\tilde{\mathbf{C}}_t^{(\ell)}\right\}_{t\in\mathbb{N}}\right) \leqslant L_{\text{var}}\left(\left\{\tilde{\mathbf{\Pi}}_t(\mathbf{R}^{(\ell)})\right\}_{t\in\mathbb{N}}\right) \leqslant \frac{64\,\text{tr}(\mathbf{\Sigma}^{-1}\mathbf{R}^{(\ell)})}{T}.$$

引理 27. 我们有

$$\left\langle \begin{pmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{\Sigma} \end{pmatrix}, \tilde{\mathbf{C}}_t^{(\ell)} \right\rangle \leqslant \left\langle \begin{pmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{\Sigma} \end{pmatrix}, \tilde{\mathbf{C}}_\infty^{(\ell)} \right\rangle \leqslant \frac{\text{tr}(\mathbf{\Sigma}^{-1}\mathbf{R}^{(\ell)})}{4\tilde{\kappa}} + 4\eta\,\text{tr}\,\mathbf{R}^{(\ell)}.$$

我们定义

$$\sigma^{(\ell)} = \left\langle \begin{pmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{\Sigma} \end{pmatrix}, \tilde{\mathbf{C}}_\infty^{(\ell)} \right\rangle.$$

以下引理是对假设 3的推论，这是建立 $\sigma^{(\ell-1)}$ 与$\sigma^{(\ell)}$之间联系。

引理 28。 假设*3* 成立，则有

$$\text{tr}\,\mathbf{R}^{(\ell)} \leqslant R^2\sigma^{(\ell-1)}, \quad \text{tr}(\mathbf{\Sigma}^{-1}\mathbf{R}^{(\ell)}) \leqslant \tilde{\kappa}\sigma^{(\ell-1)}.$$

证明。 根据$\mathbf{R}^{(\ell)}$的定义，

$$\mathbf{R}^{(\ell)} = (\mathcal{M} - \tilde{\mathcal{M}}) \circ \tilde{\mathbf{C}}_{\infty,22}^{(\ell-1)} = \mathbb{E}\left(\|\mathbf{a}\|^2_{\tilde{\mathbf{C}}_{\infty,22}^{(\ell-1)}} \mathbf{a}\mathbf{a}^\top\right) - \mathbf{\Sigma}\tilde{\mathbf{C}}_{\infty,22}^{(\ell-1)}\mathbf{\Sigma}.$$

我们舍去最后一项，它是一个半正定矩阵，因此

$$\begin{aligned}
\operatorname{tr}\mathbf{R}^{(\ell)} &\leqslant \mathbb{E}\left(\|\mathbf{a}\|^2\|\mathbf{a}\|^2_{\tilde{\mathbf{C}}_{\infty,22}^{(\ell-1)}}\right) = \left\langle \mathbb{E}\left(\|\mathbf{a}\|^2\mathbf{a}\mathbf{a}^\top\right), \tilde{\mathbf{C}}_{\infty,22}^{(\ell-1)}\right\rangle \\
&\overset{a}{\leqslant} \left\langle R^2\mathbf{\Sigma}, \tilde{\mathbf{C}}_{\infty,22}^{(\ell-1)}\right\rangle = R^2\left\langle \begin{pmatrix}\mathbf{O} & \mathbf{O}\\ \mathbf{O} & \mathbf{\Sigma}\end{pmatrix}, \tilde{\mathbf{C}}_{\infty}^{(\ell-1)}\right\rangle \overset{b}{\leqslant} R^2\sigma^{(\ell-1)}.
\end{aligned}$$

$$\begin{aligned}
\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{R}^{(\ell)}) &\leqslant \mathbb{E}\left(\|\mathbf{a}\|^2_{\mathbf{\Sigma}^{-1}}\|\mathbf{a}\|^2_{\tilde{\mathbf{C}}_{\infty,22}^{(\ell-1)}}\right) = \left\langle \mathbb{E}\left(\|\mathbf{a}\|^2_{\mathbf{\Sigma}^{-1}}\mathbf{a}\mathbf{a}^\top\right), \tilde{\mathbf{C}}_{\infty,22}^{(\ell-1)}\right\rangle \\
&\overset{a}{\leqslant} \left\langle \tilde{\kappa}\mathbf{\Sigma}, \tilde{\mathbf{C}}_{\infty,22}^{(\ell-1)}\right\rangle = \tilde{\kappa}\left\langle \begin{pmatrix}\mathbf{O} & \mathbf{O}\\ \mathbf{O} & \mathbf{\Sigma}\end{pmatrix}, \tilde{\mathbf{C}}_{\infty}^{(\ell-1)}\right\rangle \overset{b}{\leqslant} \tilde{\kappa}\sigma^{(\ell-1)}.
\end{aligned}$$

^{a} ^{b}其中 $\leqslant$ 由假设3得出，而 $\leqslant$使用引理27。 □

结合引理 27和引理28，我们得到 $\sigma^{(\ell-1)}$ 与 $\sigma^{(\ell)}$ 之间的关系如下。

引理 29。 对于所有 $\ell \geqslant 1$，我们有 $\sigma^{(\ell)} \leqslant \sigma^{(\ell-1)}/2$ 。

证明。 注意到

$$\sigma^{(\ell)} \overset{a}{\leqslant} \frac{\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{R}^{(\ell)})}{4\tilde{\kappa}} + 4\eta\operatorname{tr}\mathbf{R}^{(\ell)} \overset{a}{\leqslant} \frac{1}{4\tilde{\kappa}}\cdot\tilde{\kappa}\sigma^{(\ell-1)} + 4\eta\cdot R^2\sigma^{(\ell-1)} \overset{b}{\leqslant} \frac{1}{2}\sigma^{(\ell-1)},$$

其中 $\overset{a}{\leqslant}$ 使用引理 27， $\overset{b}{\leqslant}$ 应用引理 28，而 $\overset{c}{\leqslant}$ 使用 $\eta \leqslant 1/(16R^2)$。 □

The follow lemma bounds $L_{\text{var}}(\{\tilde{\mathbf{C}}_t^{(\ell)}\}_{t\in\mathbb{N}})$ for $\ell \geqslant 1$

引理 30. 我们有

$$\begin{aligned}
L_{\text{var}}\left(\left\{\tilde{\mathbf{C}}_t^{(\ell)}\right\}_{t\in\mathbb{N}}\right) \leqslant \frac{64\tilde{\kappa}}{2^{\ell-1}T}\Bigg(&\frac{5\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{Q}) + 5L_\ell(L+\tilde{\kappa})(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*))}{4\tilde{\kappa}} \\
&+20\eta\operatorname{tr}\mathbf{Q} + 20\eta L_\ell(B+R^2)(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*))\Bigg).
\end{aligned}$$

证明。 根据引理 24，我们有

$$\begin{aligned}
\sigma^{(0)} \overset{\text{def}}{=} \left\langle \begin{pmatrix}\mathbf{O} & \mathbf{O}\\ \mathbf{O} & \mathbf{\Sigma}\end{pmatrix}, \tilde{\mathbf{C}}_\infty\right\rangle \leqslant &\frac{5\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{Q}) + 5L_\ell(L+\tilde{\kappa})(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*))}{4\tilde{\kappa}} \\
&+20\eta L_\ell\operatorname{tr}\mathbf{Q} + 20\eta(B+R^2)(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*)).
\end{aligned}$$

根据引理 29，我们有

$$\begin{aligned}
L_{\text{var}}\left(\left\{\tilde{\mathbf{C}}_t^{(\ell)}\right\}_{t\in\mathbb{N}}\right) \leqslant L_{\text{var}}\left(\left\{\tilde{\mathbf{\Pi}}_t(\mathbf{R}^{(\ell)})\right\}_{t\in\mathbb{N}}\right) &\leqslant \frac{64\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{R}^{(\ell)})}{T} \leqslant \frac{64\tilde{\kappa}\sigma^{(\ell)}}{T} \leqslant \frac{64\tilde{\kappa}\sigma^{(0)}}{2^{\ell-1}T} \\
\leqslant \frac{64\tilde{\kappa}}{2^{\ell-1}T}\Bigg(&\frac{5\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{Q}) + 5L_\ell(L+\tilde{\kappa})(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*))}{4\tilde{\kappa}} \\
&+20\eta\operatorname{tr}\mathbf{Q} + 20\eta L_\ell(B+R^2)(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*))\Bigg),
\end{aligned}$$

证明至此完成。 □

最后，我们可以对 $\sum_{\ell=1}^{\infty} L_{\text{var}}(\{\tilde{\mathbf{C}}_t^{(\ell)}\}_{t\in\mathbb{N}})$ 进行界定。

引理 31。 假设 *3*、*4* 和 *5*成立，则有

$$\sum_{\ell=1}^{\infty} L_{\text{var}}\left(\left\{\tilde{\mathbf{C}}_t^{(\ell)}\right\}_{t\in\mathbb{N}}\right) \leqslant \frac{640\,\text{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{Q}) + 640L_\ell(L+\tilde{\kappa})(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*))}{T} + \frac{2560\eta\tilde{\kappa}\,\text{tr}\,\mathbf{Q} + 2560\eta\tilde{\kappa}L_\ell(B+R^2)(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*))}{T}.$$

证明。 应用引理29,

$$\begin{aligned}\sum_{\ell=1}^{\infty} L_{\text{var}}\left(\left\{\tilde{\mathbf{C}}_t^{(\ell)}\right\}_{t\in\mathbb{N}}\right) &\leqslant \sum_{\ell=1}^{\infty}\frac{64\tilde{\kappa}}{2^{\ell-1}T}\Bigg(\frac{5\,\text{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{Q}) + 5L_\ell(L+\tilde{\kappa})(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*))}{4\tilde{\kappa}} \\ &\quad +20\eta\,\text{tr}\,\mathbf{Q} + 20\eta L_\ell(B+R^2)(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*))\Bigg) \\ &\leqslant \frac{640\,\text{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{Q}) + 640L_\ell(L+\tilde{\kappa})(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*))}{T} \\ &\quad + \frac{2560\eta\tilde{\kappa}\,\text{tr}\,\mathbf{Q} + 2560\eta\tilde{\kappa}L_\ell(B+R^2)(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*))}{T}.\end{aligned}$$

证明完毕。 □

D.2.6 引理 9 的证明

引理 *9* 的证明。 根据 $\boldsymbol{\eta}_t^{\text{var}}$ 的定义，有 $\mathbb{E}\boldsymbol{\eta}_t = \boldsymbol{\eta}_0 = \mathbf{0}$。因此，根据定义 $\mathbf{v}$ 中的定义 1，有 $\mathbb{E}\mathbf{v} = \mathbf{0}$。为了界定 $\mathbb{E}\|\mathbf{v}\|_{\boldsymbol{\Sigma}}^2$，应用引理 25,

$$\mathbb{E}\|\mathbf{v}_k\|_{\mathbf{H}}^2 = L_{\text{var}}\left(\left\{\tilde{\mathbf{C}}_t\right\}_{t\in\mathbb{N}}\right) + \sum_{\ell=1}^{\infty} L_{\text{var}}\left(\left\{\tilde{\mathbf{C}}_t^{(\ell)}\right\}_{t\in\mathbb{N}}\right).$$

对于$\mathbf{v}_k$的界，有

$$\begin{aligned}\mathbb{E}\|\mathbf{v}_k\|_{\boldsymbol{\Sigma}}^2 &= L_{\text{var}}\left(\left\{\tilde{\mathbf{C}}_t\right\}_{t\in\mathbb{N}}\right) + \sum_{\ell=1}^{\infty} L_{\text{var}}\left(\left\{\tilde{\mathbf{C}}_t^{(\ell)}\right\}_{t\in\mathbb{N}}\right) \\ &\overset{a}{\leqslant} \frac{320\,\text{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{Q})}{T} + \frac{320L_\ell(L+\tilde{\kappa})(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*))}{T} \\ &\quad + \frac{640\,\text{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{Q}) + 640L_\ell(L+\tilde{\kappa})(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*))}{T} \\ &\quad + \frac{2560\eta\tilde{\kappa}\,\text{tr}\,\mathbf{Q} + 2560\eta\tilde{\kappa}L_\ell(B+R^2)(F(\tilde{\mathbf{y}})-F(\mathbf{x}^*))}{T}.\end{aligned}$$

^{a}其中 $\leqslant$ 应用引理 23和引理31。证明至此完成。 □

D.3 偏差上界

本节的目标是证明以下引理。

引理 32 (偏差上界)。 假设 *1*、*2*、*3*、*4* 和 *5* 均成立。令 $\mathbf{r}$ 为定义 *1*中定义的量。则有

$$\mathbb{E}\|\mathbf{r}\|_{\mathbf{\Sigma}}^2 \leqslant \frac{L_\ell}{8\alpha}\|\nabla F(\tilde{\mathbf{y}}_{k-1})\|_{\mathbf{\Sigma}^{-1}}^2,$$

其中 期望是针对第 k次外层迭代中抽取的样本计算的 n.

证明推迟到本小节末尾给出。

回顾式 $\boldsymbol{\eta}_t^{\text{bias}} \in \mathbb{R}^{2d}$ （12）中的定义：

$$\boldsymbol{\eta}_t^{\text{bias}} = \hat{\mathbf{A}}_t \boldsymbol{\eta}_{t-1}^{\text{bias}}, \quad \boldsymbol{\eta}_0^{\text{bias}} = \begin{pmatrix} \tilde{\mathbf{y}} - \tilde{\mathbf{x}}_+^* \\ \tilde{\mathbf{y}} - \tilde{\mathbf{x}}_+^* \end{pmatrix}.$$

Let

$$\boldsymbol{\eta}_t^{\text{bias}} = \begin{pmatrix} \mathbf{x}_t^{\text{bias}} \\ \mathbf{y}_t^{\text{bias}} \end{pmatrix}, \quad \text{where } \mathbf{x}_t^{\text{bias}}, \mathbf{y}_t^{\text{bias}} \in \mathbb{R}^d$$

为简单起见，在本小节中省略$\mathbf{x}_t^{\text{bias}}$、$\mathbf{y}_t^{\text{bias}}$、$\mathbf{z}_t^{\text{bias}}$的上标。偏差迭代可写成以下形式：

$$\begin{aligned}
\mathbf{y}_{t-1} &= \frac{1}{1+\theta}\mathbf{x}_{t-1} + \frac{\theta}{1+\theta}\mathbf{z}_{t-1}, \\
\mathbf{x}_t &= \mathbf{y}_{t-1} - \eta \mathbf{a}_t \mathbf{a}_t^\top (\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*), \\
\mathbf{z}_t &= \theta \mathbf{y}_{t-1} + (1-\theta)\mathbf{z}_{t-1} - \gamma \mathbf{a}_t \mathbf{a}_t^\top (\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*),
\end{aligned}$$

其中$\mathbf{x}_0 = \mathbf{z}_0 = \tilde{\mathbf{y}} - \tilde{\mathbf{x}}_+^*$。以下引理表明， $\|\mathbf{x}_t - \tilde{\mathbf{x}}_+^*\|$ 以指数速度收敛。

引理 33。 假设 *3* 成立，并按照附录 *C*中的规定选择超参数，则有

$$\mathbb{E}\|\mathbf{x}_T - \tilde{\mathbf{x}}_+^*\|^2 + \frac{\theta}{2\gamma}\mathbb{E}\|\mathbf{z}_T - \tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2 \leqslant (1-\theta)^T \left(\|\mathbf{x}_0 - \tilde{\mathbf{x}}_+^*\|^2 + \frac{\theta}{2\gamma}\|\mathbf{x}_0 - \tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2 \right). \tag{27}$$

证明。 令 $\mathbb{E}_t$ 表示在 $\{(\mathbf{a}_s,\ b_s)\}_{s=1}^t$ 条件下的期望。我们考虑能量函数$\|\mathbf{x}_t - \tilde{\mathbf{x}}_+^*\|^2/2$。根据更新规则，

$$\begin{aligned}
\frac{1}{2}\mathbb{E}_{t-1}\|\mathbf{x}_t - \tilde{\mathbf{x}}_+^*\|^2 &= \frac{1}{2}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 + \mathbb{E}_{t-1}\langle \mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*, \mathbf{x}_t - \mathbf{y}_{t-1}\rangle + \frac{1}{2}\mathbb{E}_{t-1}\|\mathbf{x}_t - \mathbf{y}_{t-1}\|^2 \\
&= \frac{1}{2}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 - \left\langle \mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*, \mathbf{\Sigma}(\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*) \right\rangle \\
&\quad + \frac{\eta^2}{2}\mathbb{E}_{t-1}\left\| \mathbf{a}_t \mathbf{a}_t^\top (\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*) \right\|^2 \\
&\overset{a}{\leqslant} \frac{1}{2}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 - \left(\eta - \frac{\eta^2 R^2}{2}\right)\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}}^2 \\
&\overset{b}{\leqslant} \frac{1}{2}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 - \frac{\eta}{2}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}}^2,
\end{aligned} \tag{28}$$

其中 $\overset{a}{\leqslant}$ 使用 $\mathbb{E}\left(\|\mathbf{a}\|^2\mathbf{a}\mathbf{a}^\top\right) \preceq R^2\mathbf{\Sigma}$，而 $\overset{b}{\leqslant}$ 使用 $\eta \leqslant 1/R^2$。接下来，我们需要对 $\|\mathbf{y}_{t-1} - \tilde{}\mathbf{x}_+^*\|^2/2$ 进行界定。根据 $\|\cdot\|$ 的凸性，我们有

$$\frac{1}{2}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 \leqslant \langle \mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*, \mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\rangle - \frac{1}{2}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2, \tag{29}$$

$$\frac{1}{2}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 \leqslant \frac{1}{2}\|\mathbf{x}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 + \langle \mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*, \mathbf{y}_{t-1} - \mathbf{x}_{t-1}\rangle. \tag{30}$$

设 $\mathbf{x} = (1-\theta)\mathbf{x}_{t-1} + \theta\mathbf{x}_+^*$，则 $\theta \times$(29) $+ (1-\theta) \times$ (30) 得到

$$\frac{1}{2}\|\mathbf{y}_{t-1} - \mathbf{x}_{t-1}\|^2 \leqslant \frac{1}{2}(1-\theta)\|\mathbf{x}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 + \langle \mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*, \mathbf{y}_{t-1} - \mathbf{x}\rangle - \frac{\theta}{2}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2.$$

设 $\mathbf{w}_{t-1} = \theta\mathbf{y}_{t-1} + (1-\theta)\mathbf{z}_{t-1}$，并注意到

$$\mathbb{E}(\mathbf{x}_t - \mathbf{y}_{t-1}) = -\eta\mathbf{\Sigma}(\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*),$$

$$\mathbf{y}_{t-1} - \mathbf{x} = \theta(\mathbf{w}_k - \tilde{\mathbf{x}}_+^*), \quad \mathbf{x}_t - \mathbf{y}_{t-1} = \frac{\eta}{\gamma}(\mathbf{z}_t - \mathbf{w}_{t-1}).$$

因此，我们有

$$\begin{aligned}
&\frac{1}{2}\|\mathbf{y}_{t-1} - \mathbf{x}_{t-1}\|^2 \\
\leqslant\ &\frac{1}{2}(1-\theta)\|\mathbf{x}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 + \frac{\theta}{\gamma}\mathbb{E}\langle \mathbf{w}_{t-1} - \mathbf{z}_{t-1}, \mathbf{w}_{t-1} - \tilde{\mathbf{x}}_+^*\rangle_{\mathbf{\Sigma}^{-1}} - \frac{\theta}{2}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 \\
=\ &\frac{1}{2}(1-\theta)\|\mathbf{x}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 - \frac{\theta}{2}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 \\
&+\frac{\theta}{2\gamma}\left(\|\mathbf{w}_{t-1} - \tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2 + \mathbb{E}_{t-1}\|\mathbf{z}_t - \mathbf{w}_{t-1}\|_{\mathbf{\Sigma}^{-1}}^2 - \mathbb{E}_{t-1}\|\mathbf{z}_t - \tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2\right) \\
\overset{a}{\leqslant}\ &\frac{1}{2}(1-\theta)\|\mathbf{x}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 - \frac{\theta}{2}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 + \frac{\theta^2}{2\gamma}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2 \\
&+\frac{\theta}{2\gamma}\left((1-\theta)\|\mathbf{z}_{t-1} - \tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2 + \mathbb{E}_{t-1}\|\mathbf{z}_t - \mathbf{w}_{t-1}\|_{\mathbf{\Sigma}^{-1}}^2 - \mathbb{E}_{t-1}\|\mathbf{z}_t - \tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2\right).
\end{aligned}$$

应用假设 3，我们有

$$\begin{aligned}
\frac{1}{2}\|\mathbf{y}_{t-1} - \mathbf{x}_{t-1}\|^2 \leqslant\ &\frac{1}{2}(1-\theta)\|\mathbf{x}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 + \frac{\tilde{\kappa}\theta\gamma}{2}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}}^2 - \frac{\theta(\mu\gamma - \theta)}{2\mu\gamma}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 \\
&+\frac{\theta}{2\gamma}\left((1-\theta)\|\mathbf{z}_{t-1} - \tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2 - \mathbb{E}_{t-1}\|\mathbf{z}_t - \tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2\right).
\end{aligned}$$

结合上述不等式与（28），我们有

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}_{t-1}\|\mathbf{x}_t - \tilde{\mathbf{x}}_+^*\|^2 \\
\leqslant\ &\frac{1}{2}(1-\theta)\|\mathbf{x}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 + \frac{\theta}{2\gamma}\left((1-\theta)\|\mathbf{z}_{t-1} - \tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2 - \mathbb{E}_{t-1}\|\mathbf{z}_t - \tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2\right) \\
&-\frac{\eta - \tilde{\kappa}\theta\gamma}{2}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}}^2 - \frac{\theta(\gamma\mu - \theta)}{2\mu\gamma}\|\mathbf{y}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2.
\end{aligned}$$

根据超参数选择，我们有 $\eta - \tilde{\kappa}\theta\gamma \geqslant 0$ 和 $\gamma\mu - \theta \geqslant 0$，因此

$$\mathbb{E}_{t-1}\|\mathbf{x}_t - \tilde{\mathbf{x}}_+^*\|^2 + \frac{\theta}{2\gamma}\mathbb{E}_{t-1}\|\mathbf{z}_t - \tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2 \leqslant (1-\theta)\left(\|\mathbf{x}_{t-1} - \tilde{\mathbf{x}}_+^*\|^2 + \frac{\theta}{2\gamma}\|\mathbf{z}_{t-1} - \tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2\right).$$

因此，我们取期望并得到

$$\mathbb{E}\|\mathbf{x}_T - \tilde{\mathbf{x}}_+^*\|^2 + \frac{\theta}{2\gamma}\mathbb{E}\|\mathbf{z}_T - \tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2 \leqslant (1-\theta)^T\left(\|\mathbf{x}_0 - \tilde{\mathbf{x}}_+^*\|^2 + \frac{\theta}{2\gamma}\|\mathbf{x}_0 - \tilde{\mathbf{x}}_+^*\|_{\mathbf{\Sigma}^{-1}}^2\right).$$

证明至此完成。 □

我们现在可以证明引理32。

引理 *32* 的证明。 根据引理 33，我们有

$$\begin{aligned}\mathbb{E}\|\mathbf{r}_k\|_{\mathbf{\Sigma}}^2 &\leqslant \lambda_{\max}(\mathbf{\Sigma})(1-\theta)^T\left(\|\mathbf{x}_0 - \mathbf{x}_+^*\|^2 + \frac{\theta}{2\gamma}\|\mathbf{x}_0 - \mathbf{x}_+^*\|_{\mathbf{\Sigma}^{-1}}^2\right)\\ &\overset{a}{\leqslant} \frac{L_\ell}{8\alpha}\|\nabla F(\tilde{\mathbf{y}}_{k-1})\|_{\mathbf{\Sigma}^{-1}}^2,\end{aligned}$$

$_a$其中 $\leqslant$ 由 T 和 θ 在附录C中的选择可得。 □

D.4 引理 4 的证明

引理 *4* 的证明。回顾一下，$\mathbf{H} = L_\ell\mathbf{\Sigma}$。所需结果可由结合引理 32 和引理 9 得出。

□

D.5 辅助引理

引理 34。 假设 *3* 成立，则 $R^2 \geqslant$ 迹 $\mathbf{\Sigma}$ 且 $\tilde{\kappa} \geqslant d$。

证明。 对于第一个结论，注意到

$$(\operatorname{tr}\mathbf{\Sigma})^2 = \left(\mathbb{E}\|\mathbf{a}\|^2\right)^2 \leqslant \mathbb{E}\|\mathbf{a}\|^4 \leqslant R^2\operatorname{tr}\mathbf{\Sigma}.$$

对于第二个结论，注意到

$$d^2 = \left(\mathbb{E}\|\mathbf{a}\|_{\mathbf{\Sigma}^{-1}}^2\right)^2 \leqslant \mathbb{E}\|\mathbf{a}\|_{\mathbf{\Sigma}^{-1}}^4 \leqslant \tilde{\kappa}d.$$

通过消去 $\operatorname{tr}\mathbf{\Sigma}$ 或 d，即可得到这些结果。 □

引理 35. 对于 $k \geqslant 0$，我们有 $x^k\left[1 + k(1-x)\right] \leqslant 1$，对于 $x \in [0,1]$。

证明。 令 $f(x) = x^k\left[1 + k(1-x)\right]$。 $f(x)$ 的导数为 $f'(x) = k(k+1)x^{k-1}(1-x) \geqslant 0$，对于 $x \in [0,1]$。因此，我们有 $f(x) \leqslant f(1) = 1$。 □

E 第二部分：OuterLoop 的分析

在本节中，我们分析算法1的外循环。在第E.1节中，我们重述引理4，并在该引理下分析外循环。

E.1 外层循环的超额风险界

令 $\mathcal{F}_k$ 表示由前 k 个内层循环的样本所生成的滤过。根据引理4，我们将外循环的更新写为如下：

$$\tilde{\mathbf{y}}_{k-1} = \tilde{\mathbf{x}}_{k-1} + \beta(\tilde{\mathbf{x}}_{k-1} - \tilde{\mathbf{x}}_{k-2}), \tag{31}$$

$$\tilde{\mathbf{x}}_k = \tilde{\mathbf{y}}_{k-1} - L_\ell h_k \mathbf{H}^{-1}\nabla F(\tilde{\mathbf{y}}_{k-1}) + h_k \mathbf{r}_k + h_k \mathbf{v}_k. \tag{32}$$

其中$\mathbf{r}_k$ 和$\mathbf{v}_k$ 满足

$$\begin{aligned}
\mathbb{E}(\|\mathbf{r}_k\|_{\mathbf{H}}^2|\mathcal{F}_{k-1}) &\leqslant \frac{L_\ell h_k}{8\alpha}\|\nabla F(\tilde{\mathbf{y}}_{k-1})\|_{\mathbf{H}^{-1}}^2, \quad \mathbb{E}(\mathbf{v}_k|\mathcal{F}_{k-1}) = 0,\\
\mathbb{E}(\|\mathbf{v}_k\|_{\mathbf{H}}^2|\mathcal{F}_{k-1}) &\leqslant \frac{320L_\ell\left(3L_\ell \operatorname{tr}(\mathbf{H}^{-1}\mathbf{Q}) + 8\eta\tilde{\kappa}\operatorname{tr}\mathbf{Q}\right)}{T}\\
&\quad + \frac{160L_\ell^2(6L + \tilde{\kappa}(7+16\eta B))(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*))}{T}\\
&\overset{\text{def}}{=} L_\ell\sigma^2 + L_{\text{eff}}L_\ell^2(F(\tilde{\mathbf{y}}_{k-1}) - F(\mathbf{x}^*)),
\end{aligned}$$

其中

$$\sigma^2 = \frac{320\left(3L_\ell \operatorname{tr}(\mathbf{H}^{-1}\mathbf{Q}) + 8\eta\tilde{\kappa}\operatorname{tr}\mathbf{Q}\right)}{T}, \quad L_{\text{eff}} = \frac{160(6L + \tilde{\kappa}(7+16\eta B))(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*))}{T}.$$

引理 36. 假设我们有

$$h_k \leqslant \min\left\{\frac{1}{6L_\ell}, \frac{T^2}{288\alpha L_\ell L_{\text{eff}}^2}\right\}, \quad \theta_k = \sqrt{\frac{L_\ell h_k}{2\alpha}}, \quad \beta_k = \frac{1-\theta_k}{1+\theta_k}.$$

令 $\tilde{}\mathbf{z}_k = \tilde{}\mathbf{y}_k + (\tilde{}\mathbf{y}_k - \tilde{}\mathbf{x}_k)/\theta_k$，则有

$$\begin{aligned}
&\mathbb{E}_{k-1}F(\tilde{\mathbf{x}}_k) - F(\mathbf{x}^*) + \frac{2\theta_k^2}{3L_\ell h_k}\mathbb{E}_{k-1}\|\tilde{\mathbf{z}}_k - \mathbf{x}^*\|_{\mathbf{H}}^2\\
&\leqslant \left(1 - \frac{\theta_k}{2}\right)\left(F(\tilde{\mathbf{x}}_{k-1}) - F(\mathbf{x}^*) + \frac{2\theta_k^2}{3L_\ell h_k}\|\tilde{\mathbf{z}}_{k-1} - \mathbf{x}^*\|_{\mathbf{H}}^2\right) + \frac{3}{2}h_k\sigma^2.
\end{aligned}$$

证明。 由于 $\nabla^2 F \preceq \mathbf{H}$，我们有

$$F(\tilde{\mathbf{x}}_k) \leqslant F(\tilde{\mathbf{y}}_{k-1}) + \langle\nabla F(\tilde{\mathbf{y}}_{k-1}), \tilde{\mathbf{x}}_k - \tilde{\mathbf{y}}_{k-1}\rangle + \frac{1}{2}\|\tilde{\mathbf{x}}_k - \tilde{\mathbf{y}}_{k-1}\|_{\mathbf{H}}^2.$$

取条件期望，条件为 $\mathcal{F}_{k-1}$，并使用（32）两次，我们有

$$\begin{aligned}
&\mathbb{E}_{k-1}\langle\nabla F(\tilde{\mathbf{y}}_{k-1}), \tilde{\mathbf{x}}_k - \tilde{\mathbf{y}}_{k-1}\rangle\\
&= -\frac{1}{L_\ell h_k}\mathbb{E}_{k-1}\langle\mathbf{H}(\tilde{\mathbf{x}}_k - \tilde{\mathbf{y}}_{k-1}), \tilde{\mathbf{x}}_k - \tilde{\mathbf{y}}_{k-1}\rangle + \frac{1}{L_\ell}\mathbb{E}_{k-1}\langle\mathbf{H}(\mathbf{r}_k + \mathbf{v}_k), \tilde{\mathbf{x}}_k - \tilde{\mathbf{y}}_{k-1}\rangle\\
&= -\frac{1}{L_\ell h_k}\mathbb{E}_{k-1}\|\tilde{\mathbf{x}}_k - \tilde{\mathbf{y}}_{k-1}\|_{\mathbf{H}}^2 - h_k\mathbb{E}_{k-1}\langle\nabla F(\tilde{\mathbf{y}}_{k-1}), \mathbf{r}_k\rangle + \frac{h_k}{L_\ell}\mathbb{E}_{k-1}\|\mathbf{r}_k + \mathbf{v}_k\|_{\mathbf{H}}^2.
\end{aligned} \tag{33}$$

注意，我们还有

$$\mathbb{E}_{k-1}\langle\nabla F(\tilde{\mathbf{y}}_{k-1}), \tilde{\mathbf{x}}_k - \tilde{\mathbf{y}}_{k-1}\rangle = -L_\ell h_k\|\nabla F(\tilde{\mathbf{y}}_{k-1})\|_{\mathbf{H}^{-1}}^2 + h_k\langle\nabla F(\tilde{\mathbf{y}}_{k-1}), \mathbf{r}_k\rangle \tag{34}$$

因此，将 $\langle \nabla F(\tilde{\mathbf{y}}_{k-1})\ \tilde{\mathbf{x}}_k - \tilde{\mathbf{y}}_{k-1}\rangle$ 替换为 $(3/4)\times(33)+(1/4)\times(34)$，从而得到

$$\begin{aligned}\mathbb{E}_{k-1}F(\tilde{\mathbf{x}}_k) &\leqslant F(\tilde{\mathbf{y}}_{k-1}) - \frac{3-2L_\ell h_k}{4L_\ell h_k}\|\tilde{\mathbf{x}}_k-\tilde{\mathbf{y}}_{k-1}\|_{\mathbf{H}}^2 + \frac{3h_k}{2L_\ell}\left(\mathbb{E}_{k-1}\|\mathbf{r}_k\|_{\mathbf{H}}^2 + \mathbb{E}_{k-1}\|\mathbf{v}_k\|_{\mathbf{H}}^2\right)\\ &\quad -\frac{h_k}{2}\mathbb{E}_{k-1}\langle\nabla F(\tilde{\mathbf{y}}_{k-1}),\mathbf{r}_k\rangle - \frac{L_\ell h_k}{4}\|\nabla F(\tilde{\mathbf{y}}_{k-1})\|_{\mathbf{H}^{-1}}^2.\\ &\overset{a}{\leqslant} F(\tilde{\mathbf{y}}_{k-1}) - \frac{3-2L_\ell h_k}{4L_\ell h_k}\|\tilde{\mathbf{x}}_k-\tilde{\mathbf{y}}_{k-1}\|_{\mathbf{H}}^2 - \frac{L_\ell h_k}{8}\|\nabla F(\tilde{\mathbf{y}}_{k-1})\|_{\mathbf{H}^{-1}}^2\\ &\quad +\frac{3h_k\sigma^2}{2} + \frac{3L_{\text{eff}}L_\ell h_k(F(\tilde{\mathbf{y}}_{k-1})-F(\mathbf{x}^*))}{2},\end{aligned}$$

其中 $\overset{a}{\leqslant}$ 使用 $\mathbb{E}_{k-1}\langle\nabla F(\tilde{\mathbf{y}}_{k-1}),\ \mathbf{r}_k\rangle \leqslant L_\ell\|\nabla F(\tilde{\mathbf{y}}_{k-1})\|^2{}_{\mathbf{H}^{-1}}/4$。因此，

$$\begin{aligned}\mathbb{E}_{k-1}F(\tilde{\mathbf{x}}_k) - F(\mathbf{x}^*) &\leqslant \left(1+\frac{3L_{\text{eff}}L_\ell h_k}{2}\right)(F(\tilde{\mathbf{y}}_{k-1})-F(\mathbf{x}^*)) + \frac{3h_k\sigma^2}{2}\\ &\quad -\frac{3-2L_\ell h_k}{4L_\ell h_k}\|\tilde{\mathbf{x}}_k-\tilde{\mathbf{y}}_{k-1}\|_{\mathbf{H}}^2 - \frac{L_\ell h_k}{8}\|\nabla F(\tilde{\mathbf{y}}_{k-1})\|_{\mathbf{H}^{-1}}^2.\end{aligned} \tag{35}$$

令 $\beta_k=(1-\theta_k)/(1+\theta_k)$，$\tilde{\mathbf{z}}_k=\tilde{\mathbf{y}}_k+(\tilde{\mathbf{y}}_k-\tilde{\mathbf{x}}_k)/\theta_k$，$\tilde{\mathbf{w}}_k=\tilde{\mathbf{y}}_k+(1-\theta_k)(\tilde{\mathbf{y}}_k-\tilde{\mathbf{x}}_k)$，其中 θ_k 将在稍后选定。根据假设 1，我们有 $\nabla^2 F\geqslant \mathbf{H}/\alpha$，因此

$$F(\tilde{\mathbf{y}}_{k-1}) \leqslant F(\mathbf{x}^*) + \langle\nabla F(\tilde{\mathbf{y}}_{k-1}),\tilde{\mathbf{y}}_{k-1}-\mathbf{x}^*\rangle - \frac{1}{2\alpha}\|\tilde{\mathbf{y}}_{k-1}-\mathbf{x}^*\|_{\mathbf{H}}^2, \tag{36}$$

$$F(\tilde{\mathbf{y}}_{k-1}) \leqslant F(\tilde{\mathbf{x}}_{k-1}) + \langle\nabla F(\tilde{\mathbf{y}}_{k-1}),\tilde{\mathbf{y}}_{k-1}-\tilde{\mathbf{x}}_{k-1}\rangle - \frac{1}{2\alpha}\|\tilde{\mathbf{y}}_{k-1}-\tilde{\mathbf{x}}_{k-1}\|_{\mathbf{H}}^2. \tag{37}$$

令 $\tilde{\mathbf{x}}=(1-\theta_k)\tilde{\mathbf{x}}_{k-1}+\theta_k\mathbf{x}^*$，则 $\theta_k\times(36)+(1-\theta_k)\times(37)$ 得到

$$\begin{aligned}F(\tilde{\mathbf{y}}_{k-1})-F(\mathbf{x}^*) &\leqslant (1-\theta_k)(F(\tilde{\mathbf{x}}_{k-1})-F(\mathbf{x}^*)) + \langle\nabla F(\tilde{\mathbf{y}}_{k-1}),\tilde{\mathbf{y}}_{k-1}-\tilde{\mathbf{x}}\rangle\\ &\quad -\frac{\theta_k}{2\alpha}\|\tilde{\mathbf{y}}_{k-1}-\mathbf{x}^*\|_{\mathbf{H}}^2 - \frac{1-\theta_k}{2\alpha}\|\tilde{\mathbf{y}}_{k-1}-\tilde{\mathbf{x}}_{k-1}\|_{\mathbf{H}}^2\\ &\overset{a}{\leqslant} (1-\theta_k)(F(\tilde{\mathbf{x}}_{k-1})-F(\mathbf{x}^*)) + \langle\nabla F(\tilde{\mathbf{y}}_{k-1}),\tilde{\mathbf{y}}_{k-1}-\tilde{\mathbf{x}}\rangle\\ &\quad -\frac{\theta_k}{4\alpha}\|\tilde{\mathbf{y}}_{k-1}-\mathbf{x}^*\|_{\mathbf{H}}^2 - \frac{1}{4\alpha}\|\tilde{\mathbf{y}}_{k-1}-\tilde{\mathbf{x}}\|_{\mathbf{H}}^2,\end{aligned}$$

其中 $\overset{a}{\leqslant}$ 使用了 $\|\cdot\|_{\mathbf{H}}^2$ 的凸性，如下所示

$$\theta_k\|\tilde{\mathbf{y}}_{k-1}-\mathbf{x}^*\|_{\mathbf{H}}^2 + (1-\theta_k)\|\tilde{\mathbf{y}}_{k-1}-\tilde{\mathbf{x}}_{k-1}\|_{\mathbf{H}}^2 \geqslant \|\tilde{\mathbf{y}}_{k-1}-\tilde{\mathbf{x}}\|_{\mathbf{H}}^2.$$

注意，$\nabla F(\tilde{\mathbf{y}}_{k-1}) = \mathbf{H}\mathbb{E}_{k-1}(\tilde{\mathbf{y}}_{k-1}-\tilde{\mathbf{x}}_k)/(L_\ell h_k) + \mathbf{H}\mathbb{E}_{k-1}\mathbf{r}_k/L_\ell$。因此

$$\begin{aligned}&\langle\nabla F(\tilde{\mathbf{y}}_{k-1}),\tilde{\mathbf{y}}_{k-1}-\tilde{\mathbf{x}}\rangle = \frac{1}{L_\ell h_k}\mathbb{E}_{k-1}\langle\mathbf{H}(\tilde{\mathbf{y}}_{k-1}-\tilde{\mathbf{x}}_k),\tilde{\mathbf{y}}_{k-1}-\tilde{\mathbf{x}}\rangle + \frac{1}{L_\ell}\mathbb{E}_{k-1}\langle\mathbf{H}\mathbf{r}_k,\tilde{\mathbf{y}}_{k-1}-\tilde{\mathbf{x}}\rangle\\ &= \frac{\theta_k^2}{L_\ell h_k}\mathbb{E}_{k-1}\langle\mathbf{H}(\tilde{\mathbf{w}}_{k-1}-\tilde{\mathbf{z}}_k),\tilde{\mathbf{w}}_{k-1}-\mathbf{x}^*\rangle + \frac{1}{L_\ell}\mathbb{E}_{k-1}\langle\mathbf{H}\mathbf{r}_k,\tilde{\mathbf{y}}_{k-1}-\tilde{\mathbf{x}}\rangle\\ &\leqslant \frac{\theta_k^2}{2L_\ell h_k}\left(\|\tilde{\mathbf{w}}_{k-1}-\mathbf{x}^*\|_{\mathbf{H}}^2 + \mathbb{E}_{k-1}\|\tilde{\mathbf{z}}_k-\tilde{\mathbf{w}}_{k-1}\|_{\mathbf{H}}^2 - \mathbb{E}_{k-1}\|\tilde{\mathbf{z}}_k-\mathbf{x}^*\|_{\mathbf{H}}^2\right)\\ &\quad +\frac{\alpha}{L_\ell^2}\mathbb{E}_{k-1}\|\mathbf{r}_k\|_{\mathbf{H}}^2 + \frac{1}{4\alpha}\|\tilde{\mathbf{y}}_{k-1}-\tilde{\mathbf{x}}\|_{\mathbf{H}}^2.\end{aligned}$$

因此，通过应用 $\alpha\mathbb{E}_{k-1}\|\mathbf{r}_k\|_{\mathbf{H}}^2 \leqslant L_\ell^3 h_k \|\nabla F(\tilde{\mathbf{y}}^{k-1})\|^2{}_{\mathbf{H}^{-1}}/4$，我们有

$$\begin{aligned}
&F(\tilde{\mathbf{y}}_{k-1}) - F(\mathbf{x}^*) \\
&\leqslant (1-\theta_k)(F(\tilde{\mathbf{x}}_{k-1}) - F(\mathbf{x}^*)) + \frac{\alpha}{L_\ell^2}\mathbb{E}_{k-1}\|\mathbf{r}_k\|_{\mathbf{H}}^2 - \frac{\theta_k}{4\alpha}\|\tilde{\mathbf{y}}_{k-1} - \mathbf{x}^*\|_{\mathbf{H}}^2 \\
&\quad + \frac{\theta_k^2}{2L_\ell h_k}\left(\|\tilde{\mathbf{w}}_{k-1} - \mathbf{x}^*\|_{\mathbf{H}}^2 + \mathbb{E}_{k-1}\|\tilde{\mathbf{z}}_k - \tilde{\mathbf{w}}_{k-1}\|_{\mathbf{H}}^2 - \mathbb{E}_{k-1}\|\tilde{\mathbf{z}}_k - \mathbf{x}^*\|_{\mathbf{H}}^2\right) \\
&\leqslant (1-\theta_k)(F(\tilde{\mathbf{x}}_{k-1}) - F(\mathbf{x}^*)) + \frac{\alpha}{L_\ell^2}\mathbb{E}_{k-1}\|\mathbf{r}_k\|_{\mathbf{H}}^2 - \left(\frac{\theta_k}{4\alpha} - \frac{\theta_k^3}{2h_k}\right)\|\tilde{\mathbf{y}}_{k-1} - \mathbf{x}^*\|_{\mathbf{H}}^2 \\
&\quad + \frac{\theta_k^2}{2L_\ell h_k}\left((1-\theta_k)\|\tilde{\mathbf{z}}_{k-1} - \mathbf{x}^*\|_{\mathbf{H}}^2 + \mathbb{E}_{k-1}\|\tilde{\mathbf{z}}_k - \tilde{\mathbf{w}}_{k-1}\|_{\mathbf{H}}^2 - \mathbb{E}_{k-1}\|\tilde{\mathbf{z}}_k - \mathbf{x}^*\|_{\mathbf{H}}^2\right).
\end{aligned} \tag{38}$$

最后，组合 上述界和（35），并使用 $3L_{\text{eff}}L_\ell h_k \leqslant \theta_k/4 \leqslant 1/3$ 和引理 4，我们得 e

$$\begin{aligned}
\mathbb{E}_{k-1}F(\tilde{\mathbf{x}}_k) - F(\mathbf{x}^*) \leqslant (1-\theta_k/2)(F(\tilde{\mathbf{x}}_k) - F(\mathbf{x}^*)) + \frac{3}{2}h_k\sigma^2 + \frac{4\alpha}{3}\mathbb{E}_{k-1}\|\mathbf{r}_k\|_{\mathbf{H}}^2 \\
+ \frac{2\theta_k^2}{3L_\ell h_k}\left((1-\theta_k)\|\tilde{\mathbf{z}}_{k-1} - \mathbf{x}^*\|_{\mathbf{H}}^2 - \mathbb{E}_{k-1}\|\tilde{\mathbf{z}}_k - \mathbf{x}^*\|_{\mathbf{H}}^2\right) \\
- \frac{1-6L_\ell h_k}{12L_\ell h_k}\mathbb{E}_{k-1}\|\tilde{\mathbf{x}}_k - \tilde{\mathbf{y}}_{k-1}\|_{\mathbf{H}^{-1}}^2 - \frac{h_k}{8}\|\nabla F(\tilde{\mathbf{y}}_{k-1})\|_{\mathbf{H}^{-1}}^2.
\end{aligned}$$

通过整理各项，我们完成证明。 □

为方便起见，我们分析以下递推，

$$L_k \leqslant (1-a\theta_k)L_{k-1} + c\theta_k^2,$$

其中 $\theta_k \leqslant 1/d$。下面的引理给出了该递推的解。

引理 37。 假设我们选择

$$\theta_k = \begin{cases} \theta_{\max}, & k \leqslant K/2, \\ \dfrac{2}{a(2/(a\theta_{\max}) + k - K/2)}, & k > K/2. \end{cases}$$

于是我们有

$$L_K \leqslant (1-a\theta_{\max})^{K/2} L_0 + \frac{12c}{a^2 K}.$$

证明。 我们首先考虑 $K \leqslant 4d/a$ 的情况。注意到我们有

$$L_k \leqslant (1-a\theta_k)L_{k-1} + c\theta_k\theta_{\max}.$$

重新排列各项后，我们有

$$L_k - \frac{c\theta_{\max}}{a} \leqslant (1-a\theta_{\max})\left(L_{k-1} - \frac{c}{a}\right).$$

因此，我们求解该递推，

$$L_K \leqslant (1-a\theta_{\max})^K L_0 + \frac{c\theta_{\max}}{a} \leqslant (1-a\theta_{\max})^K L_0 + \frac{4c}{a^2 K}.$$

对于情况 $K > 4/(a\theta_{\max})$，我们有以下初始条件：

$$L_{K/2} \leqslant (1 - a\theta_{\max})^{K/2} L_0 + \frac{c\theta_{\max}}{a}.$$

根据 θ 的选择，对于 $t > K/2$，我们有

$$L_k \leqslant \frac{2/(a\theta_{\max}) + k - K/2 - 2}{2/(a\theta_{\max}) + k - K/2} L_{k-1} + \frac{4c}{a^2(2/(a\theta_{\max}) + k - K/2)^2}.$$

将两边乘以 $(2/(a\theta_{\max}) + k - K/2)^2$,，我们得到

$$(2/(a\theta_{\max}) + k - K/2)^2 L_k \leqslant (2/(a\theta_{\max}) + k - K/2 - 1)^2 L_{k-1} + \frac{4c}{a^2}.$$

我们求解该递推，

$$\begin{aligned} L_K &\leqslant \frac{16}{a^2\theta_{\max}^2 K^2} L_{K/2} + \frac{8c}{a^2 K} \leqslant \frac{16}{a^2\theta_{\max} K^2} \left[(1 - a\theta_{\max})^{K/2} L_0 + \frac{c\theta_{\max}}{a} \right] + \frac{8c}{a^2 K}, \\ &\leqslant (1 - a\theta_{\max})^{K/2} L_0 + \frac{16c}{a^3\theta_{\max} K^2} + \frac{8c}{a^2 K} \leqslant (1 - a\theta_{\max})^{K/2} L_0 + \frac{12c}{a^2 K}. \end{aligned}$$

结合这两种情况，即完成证明。 □

E.2 定理 1 的证明

定理 2 （定理 1 的完整版本）。 假设我们按照附录 *C* 中的说明选择超参数。设 $\tilde{\mathbf{x}}_K$ 表示算法 *1* 经过 K 次外层迭代后的输出，每次外层迭代包含$T \geqslant \tilde{\Omega}(\sqrt{\tilde{\kappa}}/(\mu\eta))$ 次内层更新。那么我们有

$$\begin{aligned} \mathbb{E}F(\tilde{\mathbf{x}}_K) - F(\mathbf{x}^*) \leqslant \frac{5}{3} \exp\left(\frac{K}{8\sqrt{3\alpha} + 24\sqrt{2}\alpha L_{\text{eff}}/T} \right) \left(F(\tilde{\mathbf{x}}_0) - F(\mathbf{x}^*) \right) \\ + \frac{46080\alpha \operatorname{tr}(\mathbf{H}^{-1}\mathbf{Q})}{KT} + \frac{122880\alpha\eta\tilde{\kappa} \operatorname{tr} \mathbf{Q}}{L_\ell KT}. \end{aligned}$$

注 6. 设假设 1、2 和 3 成立，则我们有 $L_{\text{eff}} \lesssim \alpha\tilde{\kappa}$。 因此，定理2蕴含定理 1。

证明。 我们首先应用引理 37。设

$$a = \frac{1}{2}, \quad c = \frac{\alpha\sigma^2}{L_\ell}, \quad \frac{1}{\theta_{\max}} = \min\left\{ \sqrt{\frac{1}{12\alpha}}, \frac{L_\ell T}{12\sqrt{2}\alpha L_{\text{eff}}} \right\},$$

则我们有

$$\mathbb{E}F(\tilde{\mathbf{x}}_K) - F(\mathbf{x}^*) \leqslant \frac{5}{3} \exp\left(\frac{K}{8\sqrt{3\alpha} + 24\sqrt{2}\alpha L_{\text{eff}}/T} \right) \left(\mathbb{E}F(\tilde{\mathbf{x}}_K) - F(\mathbf{x}^*) \right) + \frac{48\alpha\sigma^2}{L_\ell K}.$$

注意，由引理 4，我们有

$$\sigma^2 = \frac{960 L_\ell \operatorname{tr}(\mathbf{H}^{-1}\mathbf{Q}) + 2560\eta\tilde{\kappa} \operatorname{tr} \mathbf{Q}}{T}.$$

证明至此完成。 □

E.3 推论1的证明

推论的证明 *1*。 注意，当

$$n \geqslant \tilde{\mathcal{O}}\left(\left(\sqrt{\alpha\kappa\tilde{\kappa}}+\alpha^2\tilde{\kappa}\right)+\frac{\alpha\,\mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q})}{\varepsilon}+\left(\frac{\alpha^2\tilde{\kappa}^2\,\mathrm{tr}\,\mathbf{Q}}{L_\ell\mu\varepsilon}\right)^{1/3}\right).$$

根据定理 1，优化误差和统计误差小于 $\mathcal{O}(\varepsilon)$,

$$\mathbb{E}F(\tilde{\mathbf{x}}_K)-F(\mathbf{x}^*)\leqslant\mathcal{O}(\varepsilon)+\frac{\alpha\eta\tilde{\kappa}\,\mathrm{tr}\,\mathbf{Q}}{L_\ell n}.$$

由于我们有

$$n\geqslant\tilde{\Omega}\left(\left(\frac{\alpha^2\tilde{\kappa}^2\,\mathrm{tr}\,\mathbf{Q}}{L_\ell\mu\varepsilon}\right)^{1/3}\right)\implies\tilde{\Theta}\left(\frac{\alpha\tilde{\kappa}}{\mu n^2}\right)\leqslant\tilde{\Theta}\left(\frac{n\varepsilon L_\ell}{\alpha\tilde{\kappa}\mathbf{Q}}\right),$$

$$n\geqslant\tilde{\Omega}(\sqrt{\alpha\kappa\tilde{\kappa}})\implies\tilde{\Theta}\left(\frac{\alpha\tilde{\kappa}}{\mu n^2}\right)\leqslant\frac{1}{R^2}.$$

因此，选择 η是可行的。最后，请注意，选择 $\eta\leqslant\tilde{\Theta}\left(\frac{n\varepsilon L_\ell}{\alpha\tilde{\kappa}\,\mathrm{tr}\,\mathbf{Q}}\right)$意味着$\frac{\alpha\eta\tilde{\kappa}\,\mathrm{tr}\mathbf{Q}}{L_\ell n}\lesssim\varepsilon$。至此证明完毕。

□

F 经验风险最小化的超额风险

在本节中，我们应用局部 Rademacher 技术，给出广义线性预测中经验风险最小化的超额风险界。我们考虑如下形式的经验风险最小化：

$$\hat{\mathbf{x}}=\arg\min_{\mathbf{x}\in\mathbb{R}^d}\frac{1}{N}\sum_{i=1}^{N}\ell(\mathbf{a}_i^\top\mathbf{x},b_i),$$

其中，$\{(\mathbf{a}_i,\ b_i)\}_{i=1}^N$ 是来自分布 $\mathcal{D}$的独立同分布随机变量。我们作出以下假设。

假设 6。 我们对损失函数 ℓ 和数据分布$\mathcal{D}$作出以下假设：

1. 损失函数 $\ell(z,y)$ 关于第一个参数是 L_0-利普希茨连续的，即对于所有 $z_1,z_2\in[-1,1]$,

$$\|\ell(z_1,y)-\ell(z_2,y)\|\leqslant L_0\|z_1-z_2\|\quad\mathcal{D}\text{-}a.s.$$

2. 损失函数 $\ell(\cdot,\cdot)$ 关于第一个参数是 μ_ℓ-强凸的 .

3. 数据分布 $\mathcal{D}$ 满足： $\mathbf{\Sigma}^{-1/2}\mathbf{a}$ 是 σ_a^2-次高斯的。

令函数类

$$\mathcal{F}_\delta=\left\{\mathbf{a}\mapsto\frac{\mathbf{a}^\top\mathbf{x}}{AX}\mathbb{I}_{\{\|\mathbf{a}^\top\mathbf{x}\|\leqslant AX\}}:\frac{1}{AX}\sqrt{\mathbb{E}_{a\sim\mathcal{D}}(\mathbf{a}^\top\mathbf{x})^2\mathbb{I}_{\{\|\mathbf{a}^\top\mathbf{x}\|\leqslant AX\}}}\leqslant\delta\right\},$$

其中我们选择

$$A = C\sigma_a d\sqrt{\ln\frac{1}{\delta}}, \quad X = 2\|\mathbf{x}^*\|_{\mathbf{\Sigma}},$$

并且 C 是一个充分大的常数，使得 $\mathbb{E}_{\mathbf{a}\sim\mathcal{D}}\mathbf{a}\mathbf{a}^\top\mathbb{I}_{\|\mathbf{\Sigma}^{-1/2}\mathbf{a}\|\leqslant A} \succeq \mathbf{\Sigma}/2$，且对所有 $i\in[N]$ 以至少 $1-N\delta$的概率有 $\|\mathbf{\Sigma}^{-1/2}\mathbf{a}_i\| \leqslant A$。

下面的定理给出了经验风险最小化的超额风险界。

定理 3 （经验风险最小化的超额风险界）。 假设 *6* 成立。令 $N \geqslant 441L_0^2 d\ln(1/\delta)/\mu_\ell^2$且 $\sigma_0^2 = L_0^2 d/L_\ell$ 则以至少 $1-(N+1)\delta$的概率，有

$$L(\hat{\mathbf{x}}) - L(\mathbf{x}^*) \leqslant \frac{\alpha\sigma_0^2}{\mu_\ell N}$$

Furt此外，我们证明 σ_0 的下界为 $\|\mathbf{H}^{-1/2}\mathbf{Q}\mathbf{H}^{-1/2}\|d \geqslant \mathbf{tr(H^{-1}Q)}$。

引理 38。 假设 *6* 成立，则有 $\sigma_0^2 \geqslant \|\mathbf{H}^{-1/2}\mathbf{Q}\mathbf{H}^{-1/2}\|d \geqslant \mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q}$).

证明。由于 $\|\ell(\mathbf{a}^\top\mathbf{x}^*)\| \leqslant L_0$，我们有

$$\mathbf{Q} = \mathbb{E}_{\mathbf{a},\ell\sim\mathcal{D}}\left(\ell'(\mathbf{a}^\top\mathbf{x}^*)\right)^2\mathbf{a}\mathbf{a}^\top \preceq L_0\mathbf{\Sigma} = \frac{L_0\mathbf{H}}{L_\ell}.$$

因此，我们有 $L_0 \geqslant L_\ell\|\mathbf{H}^{-1/2}\mathbf{Q}\mathbf{H}^{-1/2}\|$，这等价于 $\sigma_0^2 \geqslant \|\mathbf{H}^{-1/2}\mathbf{Q}\mathbf{H}^{-1/2}\|d \geqslant \mathbf{tr(H^{-1}Q)}$。

□

F.1 定理的证明 3

以下引理给出了 $\mathcal{F}_\delta$ 的拉德马赫复杂度上界。

引理 39. 假设条件*6* 成立，则有

$$\mathcal{R}_N(\mathcal{F}_\delta) \leqslant 2\delta\sqrt{\frac{d}{N}}.$$

¯证明。 注意， $\mathcal{F}_\delta \subset \mathcal{F}_\delta$，其中

$$\bar{\mathcal{F}}_\delta = \left\{\mathbf{a}\mapsto\frac{\mathbf{a}^\top\mathbf{x}}{AX}\mathbb{I}_{\{\|\mathbf{a}^\top\mathbf{x}\|\leqslant AX\}} : \frac{\|\mathbf{x}-\mathbf{x}^*\|_{\mathbf{\Sigma}}}{2AX} \leqslant \delta\right\}$$

根据定义，我们有

$$\begin{aligned}
&\mathcal{R}_N(\mathcal{F}_\delta) \leqslant \mathcal{R}_N(\bar{\mathcal{F}}_\delta)\\
&= \mathbb{E}_{\mathbf{a},\boldsymbol{\varepsilon}}\sup_{\|\mathbf{x}-\mathbf{x}^*\|_{\mathbf{\Sigma}}\leqslant 2AX\delta}\frac{1}{N}\sum_{i=1}^N\frac{\varepsilon_i\mathbf{a}_i^\top\mathbf{x}}{AX}\mathbb{I}_{\{|\mathbf{a}_i^\top\mathbf{x}|\leqslant AX\}} \leqslant \frac{1}{AX}\mathbb{E}_{\mathbf{a},\boldsymbol{\varepsilon}}\sup_{\|\mathbf{h}\|_{\mathbf{\Sigma}}\leqslant 2AX\delta}\frac{1}{N}\sum_{i=1}^N\varepsilon_i\mathbf{a}_i^\top\mathbf{h}\mathbb{I}_{\{\|\mathbf{h}\|\leqslant AX\}}\\
&\leqslant \frac{\delta}{AX}\sqrt{\mathbb{E}_{\mathbf{a},\boldsymbol{\varepsilon}}\left\|\frac{1}{N}\sum_{i=1}^N\varepsilon_i\mathbf{a}_i\right\|^2} \leqslant 2\delta\sqrt{\frac{d}{N}}.
\end{aligned}$$

至此完成证明。 □

定理的证明 3。我们考虑如下约束经验风险最小化：

$$\tilde{\mathbf{x}} = \operatorname*{arg\,min}_{\|\mathbf{x}-\mathbf{x}^*\|_{\boldsymbol{\Sigma}} \leqslant 1} \frac{1}{N} \sum_{i=1}^{N} \ell_i(\mathbf{a}_i^\top \mathbf{x}).$$

令 $\tilde{\ell}(z) = \ell(AXz)$，因此 $\tilde{\ell}$ 是 L_0AX-利普希茨且为 $\mu_\ell A^2X^2$-强凸，目标函数可写为：

$$\min_{\|\mathbf{x}-\mathbf{x}^*\|_{\boldsymbol{\Sigma}} \leqslant X} \frac{1}{N} \sum_{i=1}^{N} \tilde{\ell}_i(\mathbf{a}_i^\top \mathbf{x}/(AX))$$

令 $\delta_n = 2\sqrt{\frac{d}{N} \ln \frac{1}{\delta}}$，则 $\delta_N^2 \geqslant \mathcal{R}_N(\mathcal{F}_{\delta_N})$。我们应用局部化技术（Wainwright著作中的定理 14.20 [2019]）可得

$$\mathbb{E}\left(\frac{(\mathbf{a}^\top(\tilde{\mathbf{x}} - \mathbf{x}^*))^2 \mathbb{I}_{\{\|\mathbf{a}^\top \mathbf{x}\| \leqslant AX\}}}{A^2X^2} \middle| \mathbf{a}_1, \ldots, \mathbf{a}_N \right) \leqslant \frac{441 L_0^2 d}{\mu_\ell^2 A^2 X^2 N} \ln \frac{1}{\delta},$$

$$L(\tilde{\mathbf{x}}) - L(\mathbf{x}^*) \leqslant \frac{220 L_0^2 d}{\mu_\ell N} \ln \frac{1}{\delta},$$

概率至少为 $1-\delta$。因此，根据 $N \geqslant 441 L_0^2 d \ln(1/\delta)/\mu_\ell^2$，有 $\|\tilde{\mathbf{x}} - \mathbf{x}^*\|_{\boldsymbol{\Sigma}} \leqslant 1$。因此，由强凸性可得 $\hat{\mathbf{x}} = \tilde{\mathbf{x}}$。 □

G 统计项的最坏情况最优性

本节中，我们证明，若不对 $\ell \sim \mathcal{D}$ 的光滑性作出额外假设，则定理 1 中的统计项无法进一步改进。

数据生成过程。 我们考虑以下统计模型：

$$\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \quad b \sim \mathcal{N}(\mathbf{a}^\top \mathbf{x}^*, \sigma^2), \tag{39}$$

其中$\mathbf{x}^*$ 是真实值。我们考虑以下问题类。

定义 5 （问题类）。 问题类 $\mathcal{P}_n$ 定义为 $\mathcal{P}_n = \{\mu_{\mathbf{x}^*}^{\otimes n} \colon \mathbf{x}^* \in \mathbb{R}^d\}$，其中 $\mu_{\mathbf{x}^*}$ 表示（39）中的分布，$\mu_{\mathbf{x}^*}^{\otimes n}$表示 $\{(\mathbf{a}_i, b_i)\}_{i=1}^n$ 独立同分布$\sim$ $\mu_{\mathbf{x}^*}$的分布。

损失函数。 令 $\delta > 0$，我们定义以下函数：

$$\ell_\delta(z) = \begin{cases} \mu_\ell z^2/2 + (L_\ell - \mu_\ell)\delta z - (L_\ell - \mu_\ell)\delta^2/2, & z < -\delta. \\ L_\ell z^2/2, & \|z\| \leqslant \delta, \\ \mu_\ell z^2/2 + (L_\ell - \mu_\ell)\delta z - (L_\ell - \mu_\ell)\delta^2/2, & z > \delta. \end{cases}$$

总体损失函数定义为

$$F_\delta(\mathbf{x}) = \mathbb{E}_{\mathbf{a},b\sim\mathcal{D}}\, \ell_\delta(\mathbf{a}^\top \mathbf{x} - b).$$

令 $F_\delta^* = F_\delta(\mathbf{x}^*)$，我们用以下符号表示极小极大风险：

$$\mathcal{R}_n(\mathcal{P}_n) = \inf_{\hat{\mathbf{x}}} \sup_{\mathcal{D}_n \in \mathcal{P}_n} \mathbb{E}_{\{(\mathbf{a}_i, b_i)\}_{i=1}^n \sim \mathcal{D}_n} F_\delta\left(\hat{\mathbf{x}}(\{(\mathbf{a}_i, b_i)\}_{i=1}^n)\right) - F_\delta(\mathbf{x}^*(\mathcal{D})), \tag{40}$$

其中，inf 取遍所有估计量 $\hat{\mathbf{x}} : (\mathbb{R}^d \times \mathbb{R})^n \to \mathbb{R}^d$。

The 下面的定理确立了统计项的最坏情况最优性。

定理 4. 考虑定义于 $\mathcal{P}_n$ 的问题类。对于充分小的 $\delta > 0$，我们有以下极小极大风险下界 $\mathcal{R}_n(\mathcal{P}_n)$ 定义于 (40)：

$$\mathcal{R}_n(\mathcal{P}_n) \geqslant \frac{\alpha \operatorname{tr}(\mathbf{H}^{-1}\mathbf{Q}_\delta)}{2n}.$$

G.1 定理 4 的证明

下面的引理给出了 $\mathcal{P}_n$ 的极小极大风险上界。

引理 40. 考虑问题类 $\mathcal{P}_n$，我们有以下极小极大风险下界 $\mathcal{R}_n(\mathcal{P}_n)$ 定义于 (40)：

$$\mathcal{R}_n(\mathcal{P}_n) \geqslant \frac{\mu_\ell d}{n} + \frac{\mu_\ell \sigma^2}{n} - F_\delta^*$$

证明。 应用Mourtada [2022]中的定理 1，可得：

$$\inf_{\hat{\mathbf{x}}} \sup_{\mathcal{D}_n \in \mathcal{P}_n} \mathbb{E}_{\{(\mathbf{a}_i, b_i)\}_{i=1}^n \sim \mathcal{D}_n} \|\hat{\mathbf{x}}(\{(\mathbf{a}_i, b_i)\}_{i=1}^n) - \mathbf{x}^*\|_{\mathbf{\Sigma}}^2 = \frac{1}{n} \mathbb{E} \operatorname{tr}\left(\mathbf{\Sigma}\hat{\mathbf{\Sigma}}_n^{-1}\right),$$

其中，$\hat{\mathbf{\Sigma}}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{a}_i \mathbf{a}_i^\top$是样本协方差矩阵。根据矩阵逆的算子凸性，我们有 $\mathbb{E} \operatorname{tr}(\mathbf{\Sigma}\hat{\mathbf{\Sigma}}_n^{-1}) \geqslant d$。最后，注意到 ℓ_δ是 μ_ℓ-强凸的，因此我们有

$$F_\delta(\mathbf{x}) \geqslant \mathbb{E}_{\mathbf{a},b\sim\mathcal{D}} \frac{\mu_\ell}{2}(\mathbf{a}^\top \mathbf{x} - \mathbf{a}^\top \mathbf{x}^* + \varepsilon)^2 = \frac{\mu_\ell}{2}\|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{\Sigma}}^2 + \frac{\mu_\ell \sigma^2}{2}.$$

因此，我们有

$$\begin{aligned}\mathcal{R}_n(\mathcal{P}_n) &\geqslant \inf_{\hat{\mathbf{x}}} \sup_{\mathcal{D}_n \in \mathcal{P}} \mathbb{E}_{\{(\mathbf{a}_i, b_i)\}_{i=1}^n \sim \mathcal{D}_n} \frac{\mu_\ell}{2} \|\hat{\mathbf{x}}(\{(\mathbf{a}_i, b_i)\}_{i=1}^n) - \mathbf{x}^*(\mathcal{D})\|_{\mathbf{\Sigma}}^2 + \frac{\mu_\ell \sigma^2}{2} - F_\delta^* \\ &\geqslant \frac{\mu_\ell d}{n} + \frac{\mu_\ell \sigma^2}{n} - F_\delta^*,\end{aligned}$$

证明至此完成。 □

令 $\mathbf{H} = L_\ell \mathbf{\Sigma}$，且 $\mathbf{Q}_\delta = \mathbb{E}((\ell'(\mathbf{a}^\top \mathbf{x}^*))^2 \mathbf{a}\mathbf{a}^\top)$ 表示梯度协方差矩阵的二阶矩，则定理 1 中的统计项为

$$\frac{\alpha \operatorname{tr}(\mathbf{H}^{-1}\mathbf{Q}_\delta)}{n}.$$

下面我们证明定理 4。

算法 2 随机加速的数据依赖算法，用于弱凸目标函数
(SADA-WC)

Input: Initialization $\tilde{\mathbf{x}}_0 = \tilde{\mathbf{x}}_{-1}$, regularization parameters $\{h_k\}_{k=1}^K$, step sizes η, γ, and momentum parameters $\{\beta_k\}_{k=1}^K$, θ, target accuracy ε, radius D, M

for $k = 1, 2, \ldots, K$ **do**

$\tilde{\mathbf{y}}_{k-1} \leftarrow \tilde{\mathbf{x}}_{k-1} + \beta_k(\tilde{\mathbf{x}}_{k-1} - \tilde{\mathbf{x}}_{k-2})$ ▷ Extraplotation

$\mathbf{x}_0 \leftarrow \tilde{\mathbf{y}}_{k-1}$, $\mathbf{z}_0 \leftarrow \tilde{\mathbf{y}}_{k-1}$

for $t = 1, 2, \ldots, T$ **do** ▷ Inner loop for solving subproblem (41)

Sample fresh data $(\mathbf{a}_t, b_t) \sim \mathcal{D}$

$\mathbf{y}_{t-1} \leftarrow \frac{1}{1+\theta}\mathbf{x}_{t-1} + \frac{\theta}{1+\theta}\mathbf{z}_{t-1}$

$\hat{\mathbf{g}}_t \leftarrow h_k \ell'(\mathbf{a}_t^\top \tilde{\mathbf{y}}_{k-1}, b_t)\mathbf{a}_t + \frac{h_k \varepsilon}{2M^2}\mathbf{a} + \left[\mathbf{a}_t^\top(\mathbf{y}_{t-1} - \tilde{\mathbf{y}}_{k-1})\right]\mathbf{a}_t + \frac{\varepsilon}{L_\ell D^2}(\mathbf{y}_{t-1} - \tilde{\mathbf{y}}_{k-1})$

$\mathbf{x}_t \leftarrow \mathbf{y}_{t-1} - \eta\hat{\mathbf{g}}_t$

$\mathbf{z}_t \leftarrow \theta\mathbf{y}_{t-1} + (1-\theta)\mathbf{z}_{t-1} - \gamma\hat{\mathbf{g}}_t$

end for

$\tilde{\mathbf{x}}_k \leftarrow \frac{2}{T}\sum_{t=T/2+1}^{T}\mathbf{x}_t$ ▷ Tail-averaging scheme

end for

return $\tilde{\mathbf{x}}_K$

定理的证明 4。 令 $\mathbf{Q} = \lim_{\delta\to 0}\mathbf{Q}_\delta = \mu_\ell^2\mathbf{\Sigma}$。注意，我们有

$$\lim_{\delta\to 0} F_\delta^* = \lim_{\delta\to 0}\mathbb{E}\ell_\delta(\varepsilon) = \frac{\mu_\ell\sigma^2}{2},$$

$$\lim_{\delta\to 0}\mathbf{Q}_\delta = \lim_{\delta\to 0}\mathbb{E}(\ell_\delta'(\varepsilon))^2\mathbf{a}\mathbf{a}^\top = \mathbb{E}(\mu_\ell\varepsilon)^2\mathbf{a}\mathbf{a}^\top = \mu_\ell^2\mathbf{\Sigma}.$$

结合以上结果，我们有

$$\lim_{\delta\to 0}\left\{\mathcal{R}_n(\mathcal{P}_n) - \frac{\alpha\,\mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q}_\delta)}{2n}\right\} = \frac{\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{Q})}{\mu_\ell n} - \frac{\alpha\,\mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q})}{2n} = \frac{\alpha\,\mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q})}{2n} > 0.$$

因此，存在 $\delta > 0$ 使得

$$\inf_{\hat{\mathbf{x}}}\sup_{\mathcal{D}\in\mathcal{P}} F_\delta(\hat{\mathbf{x}}) - F_\delta(\mathbf{x}^*) - \frac{\alpha\,\mathrm{tr}(\mathbf{H}^{-1}\mathbf{Q}_\delta)}{2n} > 0.$$

重新排列各项即可得到所需结果。 □

H SADA 的扩展

H.1 用于弱凸目标的 SADA

算法 SADA-WC 如算法2所示。不同之处在于，内循环求解以下子问题：

$$\min_{\mathbf{x}\in\mathbb{R}^d}\ \mathbb{E}_{\mathbf{a},b\sim\mathcal{D}} h_k\left\langle \ell'(\mathbf{a}^\top\tilde{\mathbf{y}}_{k-1}, b)\mathbf{a} + \frac{\varepsilon}{2M^2}\mathbf{a}, \mathbf{x} - \tilde{\mathbf{y}}_{k-1}\right\rangle + \frac{1}{2}\|\mathbf{x} - \tilde{\mathbf{y}}_{k-1}\|^2_{\mathbf{\Sigma}+\frac{\varepsilon}{L_\ell D^2}\mathbf{I}}. \tag{41}$$

算法 3 Stochastic Accelerated Data-Dependent Algorithm with Unlabeled Data (SADA-UD)

Input: Initialization $\tilde{\mathbf{x}}_0$, regularization parameters $\{h_k\}_{k=1}^K$, step sizes η, γ, and momentum parameters $\{\beta_k\}_{k=1}^K$, θ, $\tilde{\mathbf{x}}_{-1} = \tilde{\mathbf{x}}_0$, $T_0 = \tilde{\Theta}(\sqrt{\tilde{\kappa}/(\mu\eta)}$

for $k = 1, 2, \ldots, K$ **do**

  $\tilde{\mathbf{y}}_{k-1} \leftarrow \tilde{\mathbf{x}}_{k-1} + \beta_k(\tilde{\mathbf{x}}_{k-1} - \tilde{\mathbf{x}}_{k-2})$ ▷ Extraplotation

  $\mathbf{x}_0 \leftarrow \tilde{\mathbf{y}}_{k-1}$, $\mathbf{z}_0 \leftarrow \tilde{\mathbf{y}}_{k-1}$

  for $t = 1, 2, \ldots, T$ **do** ▷ Inner loop for solving subproblem (4)

    $\mathbf{y}_{t-1} \leftarrow \frac{1}{1+\theta}\mathbf{x}_{t-1} + \frac{\theta}{1+\theta}\mathbf{z}_{t-1}$

    Sample fresh labeled data $(\mathbf{a}_t, b_t) \sim \mathcal{D}$

    if $t \leqslant T_0$ **then**

      Sample fresh m unlabeled data $\{\mathbf{a}_{t,i}^{\mathrm{ul}}\}_{i=1}^m \sim$ i.i.d.$\mathcal{D}$ ▷ Sample m unlabeled data

      $$\hat{\mathbf{g}}_t \leftarrow h_k \ell'(\mathbf{a}_t^\top \tilde{\mathbf{y}}_{k-1}, b_t)\mathbf{a}_t + \frac{1}{m+1}\left(\mathbf{a}_t\mathbf{a}_t^\top + \sum_{i=1}^m \mathbf{a}_{t,i}^{\mathrm{ul}}(\mathbf{a}_{t,i}^{\mathrm{ul}})^\top\right)(\mathbf{y}_{t-1} - \tilde{\mathbf{y}}_{k-1})$$

    else

      $\hat{\mathbf{g}}_t \leftarrow h_k \ell'(\mathbf{a}_t^\top \tilde{\mathbf{y}}_{k-1}, b_t)\mathbf{a}_t + \left[\mathbf{a}_t^\top(\mathbf{y}_{t-1} - \tilde{\mathbf{y}}_{k-1})\right]\mathbf{a}_t$

    end if

    $\mathbf{x}_t \leftarrow \mathbf{y}_{t-1} - \eta\hat{\mathbf{g}}_t$

    $\mathbf{z}_t \leftarrow \theta\mathbf{y}_{t-1} + (1-\theta)\mathbf{z}_{t-1} - \gamma\hat{\mathbf{g}}_t$

  end for

  $\tilde{\mathbf{x}}_k \leftarrow \frac{2}{T}\sum_{t=T/2+1}^T \mathbf{x}_t$ ▷ Tail-averaging scheme

end for

return $\tilde{\mathbf{x}}_K$

附录中给出的证明D.2 和附录D.3 稍作修改噪声界的证明后，即可推广到这种情况。对于内循环，我们将定义修改为˜的 $\mathcal{M}$ 和 $\mathcal{M}$ 在（13）中修改为

$$\mathcal{M}_\varepsilon = \mathbb{E}\left(\mathbf{a}\mathbf{a}^\top + \frac{\varepsilon}{L_\ell D^2}\mathbf{I}\right) \otimes \left(\mathbf{a}\mathbf{a}^\top + \frac{\varepsilon}{L_\ell D^2}\mathbf{I}\right), \quad \tilde{\mathcal{M}}_\varepsilon = \left(\mathbf{\Sigma} + \frac{\varepsilon}{L_\ell D^2}\mathbf{I}\right) \otimes \left(\mathbf{\Sigma} + \frac{\varepsilon}{L_\ell D^2}\mathbf{I}\right).$$

˜ ˜注意到 $\mathcal{M}_\varepsilon - \mathcal{M}_\varepsilon = \mathcal{M} - \mathcal{M}$，因此将特征值 λ_i替换为 $\lambda_i + \frac{\varepsilon}{L_\ell D^2}$后，附录D 中的界仍然成~~立，~~ 从而得到所声称的界。

H.2 使用无标签数据的 SADA

算法 SADA-UD 如算法3所示。区别在于，每次内迭代中，我们从$\mathbf{a}$ b 中采样数据对 $\mathcal{D}$ ，并从同一分布 m 中额外采样 $\{\mathbf{a}_i^{\mathrm{ul}}\}_{i=1}^m$ 个独立的无标签数据 $\underline{\mathcal{D}}$。然后，我们使用 $\frac{1}{m+1}\left(\mathbf{a}\mathbf{a}^\top + \sum_{i=1}^m \mathbf{a}_i^{\mathrm{ul}}\ (\mathbf{a}_i^{\mathrm{ul}})^\top\right)$ 作为$\mathbf{\Sigma}$的估计量。

我们简要介绍算法的分析3。设

$$\hat{\mathbf{\Sigma}} = \frac{1}{m+1}\left(\mathbf{a}\mathbf{a}^\top + \sum_{i=1}^m \mathbf{a}_i^{\mathrm{ul}}(\mathbf{a}_i^{\mathrm{ul}})^\top\right).$$

我们将假设3推广至以下引理。

引理41。 假设3 成立，则有

$$\mathbb{E}\hat{\mathbf{\Sigma}}^2 \preceq R_m^2\mathbf{\Sigma}, \quad \mathbb{E}\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}} \preceq \tilde{\kappa}_m\mathbf{\Sigma},$$

其中 $R_m^2 = \frac{R^2+m\lambda_{\max}(\mathbf{\Sigma})}{m+1}$ 且 $\tilde{\kappa}_m = \frac{\tilde{\kappa}+m}{m+1}$。

备注 7. 如果我们令 $\hat{\mathbf{\Sigma}} = \mathbf{a}\mathbf{a}^\top$，那么引理 41 与假设 3完全相同。

证明。 根据定义，我们有

$$\begin{aligned}
\mathbb{E}\hat{\mathbf{\Sigma}}^2 &= \frac{1}{(m+1)^2}\mathbb{E}\left(\mathbf{a}\mathbf{a}^\top + \sum_{i=1}^m \mathbf{a}_i^{\mathrm{ul}}(\mathbf{a}_i^{\mathrm{ul}})^\top\right)\left(\mathbf{a}\mathbf{a}^\top + \sum_{i=1}^m \mathbf{a}_i^{\mathrm{ul}}(\mathbf{a}_i^{\mathrm{ul}})^\top\right) \\
&\overset{a}{=} \frac{1}{m+1}\mathbb{E}\|\mathbf{a}\|^2\mathbf{a}\mathbf{a}^\top + \frac{m}{m+1}\mathbf{\Sigma}^2 \\
&\overset{b}{\preceq} \frac{R^2+m\lambda_{\max}(\mathbf{\Sigma})}{m+1}\mathbf{\Sigma} = R_m^2\mathbf{\Sigma},
\end{aligned}$$

其中，对于 $\overset{a}{=}$，我们使用 $\mathbf{a}$和 $\mathbf{a}_i^{\mathrm{ul}}$是独立同分布随机变量，而 $\overset{b}{\preceq}$ 使用了假设3。对于第二个不等式，我们有

$$\begin{aligned}
\mathbb{E}\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}\hat{\mathbf{\Sigma}} &= \frac{1}{(m+1)^2}\mathbb{E}\left(\mathbf{a}\mathbf{a}^\top + \sum_{i=1}^m \mathbf{a}_i^{\mathrm{ul}}(\mathbf{a}_i^{\mathrm{ul}})^\top\right)\mathbf{\Sigma}^{-1}\left(\mathbf{a}\mathbf{a}^\top + \sum_{i=1}^m \mathbf{a}_i^{\mathrm{ul}}(\mathbf{a}_i^{\mathrm{ul}})^\top\right) \\
&\overset{a}{=} \frac{1}{m+1}\mathbb{E}\|\mathbf{a}\|^2_{\mathbf{\Sigma}^{-1}}\mathbf{a}\mathbf{a}^\top + \frac{m}{m+1}\mathbf{\Sigma} \\
&\overset{b}{\preceq} \frac{\tilde{\kappa}+m}{m+1}\mathbf{\Sigma} = \tilde{\kappa}_m^2\mathbf{\Sigma},
\end{aligned}$$

其中，对于 $\overset{a}{=}$，我们使用 $\mathbf{a}$ 和 $\mathbf{a}_i^{\mathrm{ul}}$是独立同分布随机变量，而 $\overset{b}{\preceq}$ 使用了假设 3。 □

对于内循环，$\mathbf{x}$处的随机梯度为

$$\ell'(\mathbf{a}^\top\tilde{\mathbf{y}}, b)\mathbf{a} + \hat{\mathbf{\Sigma}}(\mathbf{x}-\tilde{\mathbf{y}}),$$

因此，噪声协方差$\mathbf{Q}$定义为

$$\mathbf{R} = \mathbb{E}\left(\ell'(\mathbf{a}^\top\tilde{\mathbf{y}}, b)\mathbf{a} - \hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}\nabla F(\tilde{\mathbf{y}})\right)\left(\ell'(\mathbf{a}^\top\tilde{\mathbf{y}}, b)\mathbf{a} + \hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}\nabla F(\tilde{\mathbf{y}})\right)^\top.$$

按照引理的证明21 和引理22，可得

$$\begin{aligned}
\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{R}) &\leqslant 5\operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{Q}) + 5L_\ell(L+\tilde{\kappa}_m)(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*)), \\
\operatorname{tr}\mathbf{R} &\leqslant 5\operatorname{tr}\mathbf{Q} + 5L_\ell(B+R_m^2)(F(\tilde{\mathbf{y}}) - F(\mathbf{x}^*)).
\end{aligned}$$

通过将假设3 替换为引理41，并相应地将量R^2、 κ和 $\tilde{\kappa}$ 替换为 R_m^2、 κ_m 和 $\tilde{\kappa}_m$，即可将证明直接推广到此设置，从而得到所声称的结果。最后，注意到 $L \leqslant \alpha\tilde{\kappa}$，我们推导出第7.2节中的结果。