arXiv:2602.05460v1 [math.ST] 5 Feb 2026

在线随机牛顿方法的复杂度降低：总成本有望达到 $\mathcal{O}(Nd)$

Antoine Godichon-Baggioni[(1)]、Bruno Portier[(2)] 和 Guillaume Sallé[(2)](1) 概率、统计与建模实验室，索邦大学，法国巴黎 75005，antoine.godichon_baggioni@upmc.fr (2) 法国国家应用科学学院数学实验室，诺曼底鲁昂国立应用科学学院，法国鲁夫赖圣艾蒂安 76800

摘要

在随机情境下优化光滑凸函数，而此时只能获得梯度和海森矩阵的含噪估计，是优化领域的一项基础问题。尽管一阶方法的单次迭代成本较低，但在病态问题上收敛缓慢。随机牛顿方法利用二阶信息校正局部曲率，但计算并求逆完整海森矩阵的 $\mathcal{O}(d^3)$ 单次迭代成本，在高维情况下难以承受，其中 d 为问题维度。本文提出一种在线小批量随机牛顿算法。该方法采用随机掩码策略，在每次迭代时选取海森矩阵列的一个子集，从而大幅降低单步计算成本。在小批量设置下，该方法使算法对 N 个数据点进行单次遍历的总计算成本达到 $\mathcal{O}(Nd)$，与一阶方法相当，同时保留了二阶信息的优势。我们证明了所得估计量几乎必然收敛且具有渐近效率。该性质无需迭代平均即可获得，这使本研究有别于以往的分析。

关键词： 凸优化；在线算法；随机牛顿方法；小批量；计算复杂度；渐近效率

1 引言

估计光滑凸函数 F 的极小化点 $\theta^* \in \mathbb{R}^d$ ： $\mathbb{R}^d \to \mathbb{R}$ 是应用数学和机器学习中的一个基本问题。本文在随机在线框架下研究这一问题，在该框架中，函数 F 无法直接访问。相反，我们假设可以访问随机预言机，它们能够提供梯度 $\nabla F(\theta)$ 和海森矩阵 $\nabla^2 F(\theta)$ 的含噪估计。在大规模学习中，这种设定十分常见，此时目标函数 F 定义为已知函数 $f, F(\theta) = \mathbb{E}_\xi[f(\xi, \theta)]$, 关于随机变量 ξ 的期望。随后，根据 ξ 的序贯样本估计极小化点 θ^*。

对于大规模问题，尤其是数据集超出内存容量或按序到达时，在线算法必不可少。一阶方法，特别是随机梯度下降 （随机梯度下降），因其简单且每次迭代的计算成本较低而得到广泛应用，通常为 $\mathcal{O}(d)$。其理论性质已得到充分理解，基础性的渐近结果见 Pelletier（2000）；Duflo（1996），非渐近分析见 Bach 和 Moulines（2011）。全面综述可参见 Bottou 等人（2018）。然而，随机梯度下降采用单一标量步长，这可能导致病态问题的收敛速度缓慢。为解决这一局限，一大类称为条件化*SGD*的算法将预条件矩阵 C_{n-1} 纳入更新：

$$\theta_n = \theta_{n-1} - \alpha_n C_{n-1} g_n(\theta_{n-1}),$$

其中， $\alpha_n > 0$ 是步长， $g_n(\theta_{n-1})$是梯度$\nabla F(\theta_{n-1})$的随机估计。 C_{n-1} 的选择决定了收敛性能与计算成本之间的权衡。AdaGrad Duchi 等人提出的著名自适应方法

(2011)、RMSProp Hinton（2012）和 Adam Kingma 与 Ba（2017）使用对角矩阵作为 C_{n-1}。这种策略为每个坐标调整步长，但无法捕捉非对角曲率信息。

受牛顿法启发，理想的条件矩阵将 C_{n-1} 设为海森矩阵的逆的近似值，即$(\nabla^2 F(\theta_{n-1}))^{-1}$。显式估计该矩阵的方法称为随机牛顿方法。这种预条件方式校正了目标函数的局部曲率，使算法能够稳健地处理病态问题。随机牛顿方法能够实现渐近效率，这是 Pelletier （ 2000 ）和 Duflo （ 1996 ）提出的渐近最优性基本基准。该性质意味着，在相关统计设定下，估计量的渐近协方差矩阵达到 Cramér-Rao下界；例如参见 Leluc 和 Portier（2023）。如 Leluc 和 Portier（2023）所示，如果预条件矩阵 C_n 几乎必然收敛到最优点处海森矩阵的逆，即 $H^{-1} := (\nabla^2 F(\theta^*))^{-1}$，当 $n \to \infty$ 时，就能实现渐近效率。这一结果放宽了以往分析所要求的更强收敛速率条件（Boyer 和 Godichon-Baggioni（2023））。

然而，阻碍其实际应用的主要障碍仍然是计算成本。直接实现（例如 Leluc 和 Portier（2023）中的实现）需要在每次迭代时计算并求逆一个经验 Hessian 矩阵。这是一项 $\mathcal{O}(d^3)$ 运算，在高维情形（$d \gg 1$）下成本高得难以承受。

相关工作。 大量研究致力于降低随机牛顿方法的 $\mathcal{O}(d^3)$ 计算负担。随机拟牛顿方法（例如适用于在线设置的 (L)-BFGS 变体，Schraudolph et al.（2007）；Moritz et al.（2016））通过利用梯度差分构建海森矩阵的逆的近似值，避免直接计算海森矩阵，通常只需 $\mathcal{O}(md)$ 的计算成本。其他方法则通过使用仅需海森向量积的迭代线性求解器，避免构造或求逆完整的海森矩阵，例如随机 Newton-CG 方法（Byrd et al.（2011））。

尽管这些方法降低了单次迭代成本，但通常缺乏渐近效率保证，因为它们无法确保预条件矩阵收敛

到真实的海森矩阵的逆。

在海森矩阵估计值为秩一矩阵的特殊情况下（例如广义线性模型），已有算法能够将 $\mathcal{O}(d^2)$ 计算成本降至更低。

例如，Bercu 等人（2020）提出了一种 $\mathcal{O}(d^2)$ 算法，利用 Sherman-Morrison 公式高效地执行秩一更新（Dennis 和 Moré（1977））。Boyer 和 Godichon-Baggioni（2023）将这一方法扩展到更广泛的模型类别，使算法在每次迭代中以 $\mathcal{O}(d^2)$ 复杂度实现渐近效率。

最近，Godichon-Baggioni 等人（2025）提出了一种在线算法，该算法以每次迭代 $\mathcal{O}(d^2)$ 的计算成本递归估计海森矩阵的逆，适用于一般的 Hessian 结构，并具有已确立的收敛保证。然而，该算法还需要额外的平均步骤才能实现渐近效率。

为降低计算成本，研究者提出了用于矩阵求逆的随机迭代算法。这些方法包括拟牛顿更新的随机变体，它们通过反复求解线性矩阵方程的“草图化”或投影版本，来近似求得固定矩阵的逆（Gower 和 Richtárik（2017））。此类算法必然是在离线设定下运行，即目标矩阵在每次迭代时均已知且可用。这一框架不同于本文所考虑的在线估计设置；在该设置中，目标海森矩阵未知，必须根据一系列噪声数据依次进行估计。因此，为离线迭代设计的方法无法直接应用于我们的问题。

除二阶预条件外，实现渐近效率的另一种途径是平均方法。该方法由 Polyak（1990）和 Ruppert（1988）独立提出，并由 Polyak 和 Juditsky（1992）进一步发展，通常称为 Polyak-Ruppert 平均。它使一阶随机梯度下降变体能够达到最优渐近协方差。尽管这种方法效果显著，但与采用 Hessian 信息的方法不同，平均方法本身并不能以同样的方式适应局部几何结构。

对于大规模应用，小批量处理也是一项重要考量，因为它能够稳定估计值、利用并行架构并降低更新开销（Bottou 等人（2018））。针对海森矩阵估计值为秩一矩阵这一特殊情形，Godichon-Baggioni 和 Werge（2025）提出了一种渐近有效的小批量随机牛顿算法，通过降低预条件矩阵的更新频率来减少计算成本。然而，该方法无法直接应用于一般的海森矩阵结构，而且利用的海森矩阵估计信息较少，因为它仅使用小批量中的单个样本来更新预条件矩阵，而非汇总整个批量的信息。对于一般情形，Godichon-Baggioni 等人（2025）的算法在纯在线设置（批量大小为 $b = 1$）下具有渐近效率，但将其扩展至小批量设置（$b > 1$）仍然面临挑战。

贡献。 本文提出了一种渐近有效的在线小批量（或流式处理）算法，即掩码随机牛顿算法 （mSNA），该算法扩展并改进了 Godichon-Baggioni 等人（2025）提出的方法。我们的工作带来了两项主要进展。首先，我们证明了

mSNA 估计量几乎必然收敛且具有渐近效率，无需 进行迭代平均；此前的工作并未提供这些理论保证。其次，我们开发了一种实用方法，可降低处理小批量 Hessian 信息的计算成本。我们的算法完全在在线设置下运行（不同于离线迭代求解器），并适用于一般的海森矩阵结构（而非仅限于秩一结构）。其计算效率的提升源于：通过对特定矩阵泛函执行随机梯度下降（SGD），直接构造逆海森矩阵估计量。随后，我们对这一底层 SGD 采用随机掩码策略，在每次迭代中选取一部分海森矩阵列；这一技术可视为对带坐标采样的 SGD 的改造，参见 Leluc 和 Portier（2022）。这种方法使算法能够处理大小为 b 的小批量，且每次迭代的计算成本为 $\mathcal{O}(\ell bd + \ell d^2)$，其中 $\ell \ll d$为秩参数。对于大小为 N的数据集，该算法执行 N/b 次迭代。通过设定批量大小 $b = d$和 $\ell = 1$，对数据进行一遍处理的总复杂度为 $\mathcal{O}(Nd)$，其复杂度增长阶与一阶方法相当，同时保留了二阶信息的优势。我们还研究了该算法的平均版本；它只需极少的额外计算成本，便可提升实际性能。

因此，我们的工作旨在提出：（i）在线随机牛顿估计量；（ii）在尽可能一般的设定下构建；（iii）降低计算复杂度并支持小批量处理（对数据进行一次遍历时，操作次数可能为 $O(Nd)$）；（iv）提供强有力的理论保证，尤其是渐近效率。据我们所知，此前尚无研究能够同时成功兼顾这四个方面。

论文结构。 第 2 节介绍记号、优化框架以及所依据的假设。第 3 节介绍递归估计海森矩阵的逆矩阵的算法。随后，第 4 节将该估计量纳入所提出的小批量随机牛顿算法，并对其进行理论收敛分析。第 5 节详细说明该算法在流式处理（小批量）设定下的实现，并分析其计算复杂度。最后，第 6 节通过数值实验展示该算法的性能。证明见附录。

2 框架

记号。 梯度和海森算子分别记为 ∇ 和 ∇^2。我们用 $\mathcal{M}_d(\mathbb{R})$表示实系数 d 阶方阵的集合，用$\mathcal{S}_d^{++}(\mathbb{R})$表示 d阶对称正定矩阵的集合。令 I_d 为 d阶单位矩阵。在 $\mathcal{M}_d(\mathbb{R})$上，我们用 $\|\cdot\|_F$（分别用 $\|\cdot\|_{op}$）表示 Frobenius 范数（分别表示算子范数），用 $\langle \cdot, \cdot \rangle_F$ 表示 Frobenius 内积。对于向量 $v \in \mathbb{R}^d$，用 $\|v\|$ 表示其欧几里得范数。对于矩阵 $A \in \mathcal{M}_d(\mathbb{R})$， A^T 表示其转置，$\mathrm{Tr}(A)$表示其迹。我们用 $\xrightarrow{\mathcal{L}}$表示依分布收敛。

我们不研究满足 Robbins-Monro条件$\sum_n \alpha_n = +\infty$ 和 $\sum_n \alpha_n^2 < +\infty$的一般步长 $(\alpha_n)_{n\geq 1}$，为简化起见，而是将步长限制为

如下形式 $\alpha_n \coloneqq c_\alpha/(n^\alpha + c'_\alpha)$，其中 $c_\alpha > 0$，$c'_\alpha \geq 0$、$1/2 < \alpha \leq 1$；按照 Carpentier等人（2015）的说法，我们将这样的序列称为 $\sigma(\alpha,\ c_\alpha,\ c'_\alpha)$序列。

问题设定 设 $F\colon \mathbb{R}^d \longrightarrow \mathbb{R}$ 是一个凸的二阶可微函数。假设极小化点存在且唯一，我们考虑估计 F 的极小化点 θ^* 这一问题，即:

$$\theta^* \coloneqq \arg\min_{\theta \in \mathbb{R}^d} F(\theta),$$

利用估计量序列 $(\theta_n)_{n\geq 0}$。在每次迭代 $n \geq 1$ 中，我们假设可以访问梯度和海森矩阵的随机预言机。更准确地说，给定点 $\theta \in \mathbb{R}^d$，随机预言机返回梯度 $\nabla F(\theta)$（分别为海森矩阵 $\nabla^2 F(\theta)$）的随机估计值 $g_n(\theta)$（分别为 $h_n(\theta)$）。此外，我们用 $H \coloneqq \nabla^2 F(\theta^*)$ 表示极小化点处的海森矩阵。设 $(\mathcal{F}_n)_{n\geq 0}$ 为由迭代值 $(\theta_n)_{n\geq 0}$ 和预言机 $(g_n、\ h_n)_{n\geq 1}$ 生成的自然滤子。

（A1）（无偏估计）。对于所有 $n \geq 1$ 和 $\theta \in \mathbb{R}^d$:

a) $\mathbb{E}\left[g_n(\theta) \mid \mathcal{F}_{n-1}\right] = \nabla F(\theta) \quad$ a.s.
b) $\mathbb{E}\left[h_n(\theta) \mid \mathcal{F}_{n-1}\right] = \nabla^2 F(\theta) \quad$ a.s.

（A2）（增长条件）。存在 $\mathcal{L}_g,\ \mathcal{L}_h,\ \sigma^2 \geq 0$，使得对所有 $n \geq 1$ 和 $\theta \in \mathbb{R}^d$:

a) $\mathbb{E}\left[\|g_n(\theta)\|^2 \mid \mathcal{F}_{n-1}\right] \leq \mathcal{L}_g\left(F(\theta) - F(\theta^*)\right) + \sigma^2 \quad$ a.s.
b) $\mathbb{E}\left[\|h_n(\theta)\|_{op}^2 \mid \mathcal{F}_{n-1}\right] \leq \mathcal{L}_h \quad$ a.s.

（A3）（极小化点处的海森矩阵）。极小化点处的海森矩阵 $H \coloneqq \nabla^2 F(\theta^*)$ 是正定的，且映射 $\theta \mapsto \nabla^2 F(\theta)$ 在 θ^* 处连续。

（A4）（李雅普诺夫条件）。存在 $q,\ q' > 2$ 和 $M,\ M' > 0$，使得:

a) $\sup_{\substack{n\geq 1 \\ \|\theta-\theta^*\|<M}} \mathbb{E}\left[\|g_n(\theta) - \nabla F(\theta)\|^q \mid \mathcal{F}_{n-1}\right] < +\infty \quad$ a.s.

b) $\sup_{\substack{n\geq 1 \\ \|\theta-\theta^*\|<M'}} \mathbb{E}\left[\left\|h_n(\theta) - \nabla^2 F(\theta)\right\|_F^{q'} \mid \mathcal{F}_{n-1}\right] < +\infty \quad$ a.s.

（A5）（协方差极限）。如果 $(\theta_n)_{n\geq 0}$ 几乎必然收敛到 θ^*，则存在 $\Gamma \in \mathcal{S}_d^{++}$，使得 $\mathrm{Cov}\left(g_n(\theta_{n-1}) \mid \mathcal{F}_{n-1}\right) \xrightarrow[n\to\infty]{\text{几乎必然}} \Gamma$。

这些假设是随机逼近中的标准假设，遵循 Pelletier（1998b）和 Pelletier（1998a）中的设定（另见 Leluc 和 Portier（2023））。假设(**A1**) 可以放宽以容纳消失偏差，但为简便起见，我们不考虑这种情况；相关分析请参见 Surendran等人（2024）。假设(**A2**) a) 对于获得随机梯度估计量的强相合性至关重要。它与 Gower等人（2019）提出的期望光滑性 有关，并且与

Leluc 和 Portier（2023）一样，允许梯度预言机的分布发生变化。假设 **(A2) b)** 的限制更强，尤其意味着 F 的海森矩阵一致有界。假设 **(A3)** 意味着函数 F 在 θ^* 附近局部强凸，并为使用牛顿方法提供依据。对于任何基于梯度的方法而言，该假设都是实现渐近效率的关键（例如，平均随机梯度算法可参见 Pelletier（2000）和 Godichon-Baggioni（2019），条件梯度算法可参见 Leluc 和 Portier（2023））。假设 **(A4)** 是一个标准且已知的李雅普诺夫条件，只需在极小化点 θ^* 附近局部成立。最后，假设 **(A5)** 是确立基于梯度的估计量渐近正态性所需的标准条件（Pelletier（1998b））。

在实践中 这一框架对通用预言机作出假设，比标准设定更为一般。在标准设定中，$F(\theta) = \mathbb{E}_\xi[f(\xi, \theta)]$ 被定义为已知函数 f 的期望，即 $F(\theta) = \mathbb{E}_\xi[f(\xi, \theta)]$。这种一般性允许预言机分布随时间变化（Leluc 和 Portier（2023）也考虑了这一设定）。此外，这种基于预言机的框架便于研究小批量策略，包括批量大小可变或采用非均匀采样的策略。

在标准设定中，$F(\theta) = \mathbb{E}_\xi[f(\xi, \theta)]$，并且可以获得独立同分布样本 $(\xi_n)_{n\geq 1}$，此时自然的预言机为

$$g_n(\theta) = \nabla f(\xi_n, \theta) \quad \text{and} \quad h_n(\theta) = \nabla^2 f(\xi_n, \theta).$$

在线小批量（流式处理）设定会按大小为 $b > 1$ 的数据块处理数据。令 $\{\xi_{n,i}\}_{i=1}^b$ 表示第 n 次迭代时数据块中的样本，则预言机为经验平均值：

$$g_n(\theta) = \frac{1}{b}\sum_{i=1}^{b} \nabla f(\xi_{n,i}, \theta) \quad \text{and} \quad h_n(\theta) = \frac{1}{b}\sum_{i=1}^{b} \nabla^2 f(\xi_{n,i}, \theta).$$

第 5 节详细分析了这一小批量设定。

如果函数 f 满足相应的假设，则这些预言机假设也成立。这些假设见附录 C。对于逻辑回归和 p 均值等经典问题，Godichon-Baggioni 等人（2025）的附录 A 中给出了对这些假设的验证。

3 正定矩阵逆矩阵的在线估计

在本节中，我们旨在借助一系列适应于滤过 $(\mathcal{F}_n)_{n\geq 0}$ 的 H 估计值 $(H_n)_{n\geq 1}$，递归地估计假设为正定的矩阵 $H := \nabla^2 F(\theta^*)$ 的逆矩阵。在我们的框架中，这些估计值通过在一个收敛到 θ^* 的估计量 θ_{n-1} 处评估海森矩阵预言机 h_n 获得。该序列 $\hat{}(\theta_n)_{n\geq 0}$ 可以是主迭代序列 $(\theta_n)_{n\geq 0}$ 本身，也可以是由其导出的平均版本 $\bar{}(\theta_n)_{n\geq 0}$。

3.1 逆矩阵的随机梯度估计

我们首先注意到， H^{-1} 是二次泛函 $J:\mathcal{M}_d(\mathbb{R}) \longrightarrow \mathbb{R}_+$ 的唯一极小化点，该泛函对所有 $A \in \mathcal{M}_d(\mathbb{R})$ 定义如下:

$$J(A) = \left\| H^{1/2}\left(A - H^{-1}\right)\right\|_F^2,$$

其中， $H^{1/2}$ 表示矩阵 H唯一的对称正定平方根。函数 J 二阶可微，利用标准的矩阵微积分（参见附录D）可得其梯度:

$$\nabla J(A) = 2(H^{1/2})^T H^{1/2}(A - H^{-1}) = 2\left(HA - I_d\right).$$

因此，泛函 J 是 $2\lambda_{\min}(H)$-强凸且 $2\lambda_{\max}(H)$-光滑的。

泛函 J 的主要特点是，其梯度仅通过乘积 HA依赖于 H 。因此，可以在函数J上使用随机梯度下降，递归地估计 H^{-1} ，并以其估计值 H_n替代 H 。我们将在下一节介绍的算法源自一种对称正定分解。为便于分析，我们还定义了相关的对称化泛函:

$$J_{sym}(A) \coloneqq \frac{1}{2}\left(J(A) + J(A^T)\right),$$

其梯度为

$$\nabla J_{sym}(A) = \frac{1}{2}\left(\nabla J(A) + \left(\nabla J(A^T)\right)^T\right) = HA + AH - 2I_d,$$

当 A 为对称矩阵时，该梯度也是对称的。基于 J 或 J_{sym} 的朴素随机梯度下降估计量 $(A_n)_{n\geq 0}$ 存在两个显著局限。首先，这类估计量的主要缺点是计算成本高。事实上，在每一步中，使用标准算法进行稠密矩阵乘法需要 $\mathcal{O}(d^3)$次运算。为了降低这一成本，我们将考虑梯度的草图化版本。第二个局限是，这些更新无法保证与 A_n 相关的二次型保持正定，即对所有 $v \in \mathbb{R}^d \setminus \{0\}$， $v^T A_n v > 0$都成立；而当 A_n 用作预条件器时，这一点对于确保下降方向至关重要。因此，我们将提出一种随机梯度下降更新的修改方法，以确保估计量的正定性。

3.2 逆矩阵的低成本正估计量

令 $\ell \in \{1, \ldots, d\}$ 为一个整数。设 $(M_{\ell,\, n})_{n\geq 1}$ 为适应于滤过 $(\mathcal{F}_n)_{n\geq 0}$ 的一列随机对角投影矩阵。在每一步 $n \geq 1$， $M_{\ell,\, n}$ 通过均匀随机地选取一个大小为 ℓ 的子集 $I_n \subset \{1, \ldots, d\}$ 构造而成；当 $i \in I_n$ 时，令 $(M_{\ell,\, n})_{ii} = 1$ ，否则令其为 0。我们假设第 n 步的随机选择独立于 $\mathcal{F}_{n-1}$ 和随机海森矩阵 H_n。由此可直接得到 $\mathbb{E}[M_{\ell,n} \mid \mathcal{F}_{n-1}] = \frac{\ell}{d} I_d$。令 $\tilde{H}_n \coloneqq M_{\ell,\, n} H_n$ 为草图化随机海森矩阵。从

一个正定初始估计量 $A_0 := I_d$ 出发，我们提出通过估计量序列 $(A_n)_{n\geq 0}$ 来估计 H^{-1}，该序列对所有 $n \geq 1$ 递归定义如下：

$$A_n := A_{n-1} - \mathbf{1}_{\gamma_n \|\tilde{H}_n\|_{op} \leq \frac{1}{2}} \left(\gamma_n \left(\tilde{H}_n A_{n-1} + A_{n-1} \tilde{H}_n^T - 2M_{\ell,n} \right) - \gamma_n^2 \tilde{H}_n A_{n-1} \tilde{H}_n^T \right), \quad (1)$$

其中，$(\gamma_n)_{n\geq 1}$是一个 $\sigma(\gamma, c_\gamma, c'_\gamma)$序列，满足 $\gamma \in \left(\frac{1}{2}, 1\right)$。

更新 (1) 值得作几点说明。首先，估计量 A_n 按构造是对称且正定的。由于 A_0 是正定的，根据归纳法，序列 $(A_n)_{n\geq 0}$ 始终保持正定，如下列分解所示：

$$A_n = \left(I_d - \gamma_n \mathbf{1}_{\gamma_n \|\tilde{H}_n\|_{op} \leq \frac{1}{2}} \tilde{H}_n \right) A_{n-1} \left(I_d - \gamma_n \mathbf{1}_{\gamma_n \|\tilde{H}_n\|_{op} \leq \frac{1}{2}} \tilde{H}_n \right)^T + 2\gamma_n \mathbf{1}_{\gamma_n \|\tilde{H}_n\|_{op} \leq \frac{1}{2}} M_{\ell,n}. \quad (2)$$

事实上，$\left(I_d - \gamma_n \mathbf{1}_{\gamma_n \|\tilde{H}_n\|_{op} \leq \frac{1}{2}} \tilde{H}_n \right)$ 可逆，因此根据 Sylvester惯性定律，若 A_{n-1} 正定，则第一项正定；而第二项为半正定，因为$M_{\ell,\, n}$ 是一个正交投影矩阵。其次，更新中的 $\mathcal{O}(\gamma_n)$项 $\left(\tilde{H}_n A_{n-1} + A_{n-1} \tilde{H}_n^T - 2M_{\ell,n} \right)$是投影随机梯度$2M_{\ell,\, n}(H_n A_{n-1} - I_d)$在点 A_{n-1}处对 J的对称化，这里利用了 A_{n-1} 是对称矩阵这一事实。给定 $\mathcal{F}_{n-1}$ 时，其条件期望为 $\frac{\ell}{d} \nabla J_{sym}(A_{n-1})$。草图化引入了预期缩放因子 ℓ/d，从而减缓估计量向 H^{-1}邻域收敛。第三，更新中的 $\mathcal{O}(\gamma_n^2)$项 $-\gamma_n^2 \tilde{H}_n A_{n-1} \tilde{H}_n^T$恰好是补全上述分解中二次型所需的修正项，这保证了正定性。最后，示性函数 $\mathbf{1}_{\gamma_n \|\tilde{H}_n\|_{op} \leq \frac{1}{2}}$是一个标准的

通过截断来确保递推的稳定性，限制大型随机样本的影响。

3.3 收敛结果

现在我们陈述估计量 $(A_n)_{n\geq 0}$ 的收敛性质。

命题3.1。 设$(\theta_n)_{n\geq 0}$ 是 θ^* 的一个适应于滤过$(\mathcal{F}_n)_{n\geq 0}$的估计量序列，并设$(A_n)_{n\geq 0}$ 是使用 $H_n := h_n(\theta_{n-1})$通过(1) 定义的序列。假设在 $q' > \frac{2}{\gamma}$ 下，假设***(A1)***、***(A2)***、***b)***、***(A3)***和***(A4)b)***成立。

若 $\theta_n \to \theta^*$ 几乎必然成立，则 $A_n \to H^{-1}$ 几乎必然成立。

在这种情况下，令 $\delta_n := \text{上确界}_{k\geq n} \left\| \theta_{k-1} - \theta^* \right\|$。令 $\mu: \mathbb{R}_+ \to [0; +\infty]$ 为 $\nabla^2 F$ 在 θ^* 处的局部连续模，定义为 $\mu(r) := \text{上确界}_{\|\theta-\theta^*\|\leq r} \left\| \nabla^2 F(\theta) - \nabla^2 F(\theta^*) \right\|_F$. 根据假设 ***(A3)***，映射 μ 在0 的某个邻域内有限，并满足极限$\lim_{r\to 0} \mu(r) = 0$。若 $\gamma < 1$，则对任意 $\eta > 0$:

$$\left\| A_n - H^{-1} \right\|_F^2 = \mathcal{O}\left(\gamma_n (\ln n)^{1+\eta} + \mu\left(\delta_{\lceil \frac{n}{2} \rceil} \right)^2 \right) \quad a.s.$$

命题3.1证明，估计量序列 $(A_n)_{n\geq 0}$ 强相合，条件是 $\theta_n \to \theta^*$ 几乎必然成立。此外，该命题通过将渐近误差分解为两个部分来加以量化：方差项 $\mathcal{O}(\gamma_n(\ln n)^{1+\eta})$，这是随机逼近固有的；以及偏差项 $\mathcal{O}(\mu(\delta_{\lceil \frac{n}{2} \rceil})^2)$，它受估计量 θ_n 的收敛速率控制。下一节所提出的随机牛顿算法将使用这一序列 $(A_n)_{n\geq 0}$ 作为条件矩阵。

3.4 与以往工作的关系

在 Godichon-Baggioni 等人（2025）中，作者提出了一种类似的在线算法，基于 Robbins-Monro 方法寻找映射$A \mapsto HA + AH - 2I_d$的零点，以估计 H^{-1} 。该映射是上文定义的泛函 J_{sym} 的梯度，并且算法结合了草图化。通过采用比我们更严格的截断，并将结果投影到半径缓慢增长的矩阵球上，他们的算法确保估计量在每一步都保持正定。

在离线设定中，Agarwal等人（2017）提出了一种算法，通过围绕已知矩阵对逆矩阵进行泰勒展开，来估计正定矩阵的逆。他们提出的递推实际上就是对上文定义的二次泛函J 进行随机梯度下降，使用固定步长 $\gamma_n = 1$，且不进行草图化，也不包含正定校正项。经过几次递推后，他们只计算该逆海森矩阵估计与随机梯度的乘积，因此每次迭代的计算成本为 $\mathcal{O}(Sd^2)$，其中 S 为递推次数。每次主算法迭代时，逆海森矩阵估计都从头开始计算。

3.5 加权平均版本的说明

我们以关于加权平均的一点说明结束本节。为了获得更快的收敛速率，也可以考虑平均估计量（例如参见 Polyak 和 Juditsky（1992）、Pelletier（2000））。由于标准平均可能对初始化较为敏感，我们考虑一种加权平均版本（参见 Mokkadem和Pelletier（2011）或 Boyer 和 Godichon-Baggioni（2023）），其递归定义为 $A_0 := A_0$ 并且，对所有 $n \geq 1$:

$$\bar{A}_n = \left(1 - \frac{\omega_n}{\sum_{k=0}^{n} \omega_k}\right) \bar{A}_{n-1} + \frac{\omega_n}{\sum_{k=0}^{n} \omega_k} A_n, \quad \text{where} \quad \omega_n = (\ln(n+1))^\tau, \tag{3}$$

其中 $\tau \geq 0$ 是一个加权参数； $\tau = 2$ 是一种常见选择。由于 A_n 是正定估计量（A_k）$_{0\leq k\leq n}$的凸组合，因此仍然是正定的。

尽管众所周知，平均化有助于优化渐近方差，但后续将不采用该估计量（A_n）$_{n\geq 0}$ 。该估计量（A_n）$_{n\geq 0}$ 面临的主要挑战并非最优点附近的随机噪声，而是瞬态阶段收敛缓慢，因为算法每次迭代仅修改 ℓ 行和列。平均化会引入惯性，进一步阻碍初期收敛。此外，平均更新（3）是一种稠密运算，需要$\mathcal{O}(d^2)$ 内存访问来读写估计量的所有分量。这种稠密

访问需求与非平均更新（1）形成对比；尽管计算该更新项需要 $\mathcal{O}(\ell d^2)$ 次运算，但更新本身仅修改 A_{n-1} 中与所选 ℓ 行和列对应的 $\mathcal{O}(\ell d)$ 个分量。鉴于平均步骤会带来不利的惯性，并要求进行稠密内存访问，下一节的随机牛顿算法将保留非平均估计量 $(A_n)_{n\geq 0}$。关于用于估计正定矩阵逆矩阵的相关加权平均方案的收敛分析，可参阅 Godichon-Baggioni 等人（2025）。

4 算法

4.1 一种新的随机牛顿算法

我们在此提出一种名为“掩码随机牛顿算法”（mSNA）的算法，其递归定义为：$n \geq 1$

$$\theta_n \coloneqq \theta_{n-1} - \alpha_n \left(A_{n-1} + \nu_n I_d\right) g_n(\theta_{n-1}) \tag{4}$$

$$\tilde{H}_n \coloneqq M_{\ell,n} h_n(\theta_{n-1}) \tag{5}$$

$$A_n \coloneqq A_{n-1} - \mathbf{1}_{\gamma_n \left\|\tilde{H}_n\right\|_{op} \leq \frac{1}{2}} \left(\gamma_n \left(\tilde{H}_n A_{n-1} + A_{n-1}\tilde{H}_n^T - 2M_{\ell,n}\right) - \gamma_n^2 \tilde{H}_n A_{n-1} \tilde{H}_n^T\right), \tag{6}$$

其中，$\theta_0 \in \mathbb{R}^d$ 和 $A_0 \in \mathcal{M}_d(\mathbb{R})$ 可任意选取，$(M_{\ell,\ n})_{n\geq 1}$ 是第 3 节定义的随机投影矩阵序列，$(\alpha_n)_{n\geq 1}$ 是一个 $\sigma(\alpha,\ c_\alpha,\ c'_\alpha)$-序列，$(\gamma_n)_{n\geq 1}$ 是一个 $\sigma(\gamma,\ c_\gamma,\ c'_\gamma)$-序列，且 $\gamma \in \left(\frac{1}{2},\ 1\right)$，而 $(\nu_n)_{n\geq 1}$ 递减至 0 并满足 $\sum_{n\geq 1} \alpha_n \nu_n = +\infty$。为此，对于任意 $\nu > 0$，若 $\alpha < 1$，则取 $\nu_n = \frac{\nu}{n^{1-\alpha}}$；若 $\alpha = 1$，则取 $\nu_n = \frac{\nu}{\ln n}$。项 ν_n 在理论上对于给条件矩阵的最小特征值设定下界至关重要，但在实践中，ν 可以任意取小。以下定理给出了该随机牛顿算法的渐近性质。

定理 4.1。在假设 ***(A1)*** 和 ***(A2)*** 下，若 $\alpha + \gamma > \frac{3}{2}$，由 (4) 定义的估计量 $(\theta_n)_{n\geq 0}$ 几乎必然收敛到 θ^*。进一步假设 ***(A3)***、在 $q' > \frac{2}{\gamma}$ 条件下的 ***(A4)*** 以及 ***(A5)*** 成立。记 $\zeta\colon \quad - = \mathbf{1}_{\alpha=1}/2c_\alpha$，若 $1 - \zeta > 0$，则有

$$A_n \xrightarrow[n\to+\infty]{a.s.} H^{-1} \quad and \quad \frac{1}{\sqrt{\alpha_n}} \left(\theta_n - \theta^*\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{2(1-\zeta)} H^{-1} \Gamma H^{-1}\right).$$

证明见附录，其依据是 Leluc 和 Portier（2023）的一个近期结果，该结果表明，只要条件矩阵强一致，条件化SGD 就具有渐近正态性。定理 4.1 的结果与 Godichon-Baggioni 等人（2025）的结果有一个主要区别：我们无需对 θ_n 执行平均步骤，就能获得算法的收敛性和渐近效率。由以下推论给出：

推论 4.2. 在定理 *4.1* 的假设下，如果 $\alpha_n = 1/(n + c'_\alpha)$，那么

$$\sqrt{n} \left(\theta_n - \theta^*\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, H^{-1} \Gamma H^{-1}\right).$$

为了利用 Boyer 和 Godichon-Baggioni（2023）以及 Godichon-Baggioni 和 Werge（2025）的结果建立几乎必然的渐近收敛速率，我们必须采用比 **(A3)** 更强的假设。

(A6) （在 θ^*处利普希茨连续）。映射 $\theta \mapsto \nabla^2 F(\theta)$ 在 θ^*处是利普希茨连续的，即存在 $r > 0$ 和 L，使得对于所有 $\theta \in \mathcal{B}(\theta^*, r)$，$\left\|\nabla^2 F(\theta) - H\right\|_F \leq L \left\|\theta - \theta^*\right\|$。

据我们所知，文献中对平均条件算法的研究都额外采用了假设**(A6)**，即海森矩阵在极小化点处是利普希茨连续的；而非平均 mSNA 在步长为 $\alpha_n = 1/(n + c'_\alpha)$ 时则不需要这一假设。

定理 4.3. 在满足定理 *4.1* 的假设和假设*(A6)*的条件下，如果假设 *(A4)* 中的指数 q 满足 $q > \frac{2}{\alpha}$，那么对于任意 $\eta > 0$：

$$\left\|\theta_n - \theta^*\right\|^2 = \mathcal{O}\left(\frac{\ln n}{n^\alpha}\right) \ a.s. \quad and \quad \left\|A_n - H^{-1}\right\|_F^2 = \mathcal{O}\left(\frac{(\ln n)^{1+\eta}}{n^{\min\{\alpha,\gamma\}}}\right) \ a.s.$$

4.2 加权平均版本

我们已经看到，只有选择 $\alpha = 1$和 $c_\alpha = 1$，才能直接实现渐近效率，这相当于采用大小为 $1/n$的步长。然而，这会导致更新幅度很小，在初始化不佳时可能会造成问题（参见 Cénac 等（2025）；Boyer 和 Godichon-Baggioni（2023））。为了解决这一问题，可以选择 $\alpha < 1$ 以采用更大的步长，然后通过平均步骤恢复渐近效率。这一平均步骤可应用于各种条件化随机梯度下降算法，包括 mSNA 算法。对于任意估计量 $(\theta_n)_{n\geq 0}$，我们将平均估计量 $(\theta_n)_{n\geq 0}$ 定义为迭代值 θ_k 在 $k \leq n$下的加权平均，其权重形式为 $\omega_k := (\ln(k+1))^\tau$，其中 $\tau \geq 0$ 由 Boyer 和 Godichon-Baggioni（2023）选定：

$$\bar{\theta}_0 := \theta_0, \quad \text{and for } n \geq 1, \quad \bar{\theta}_n := \frac{\sum_{k=0}^n \omega_k \theta_k}{\sum_{k=0}^n \omega_k}. \tag{7}$$

按照约定 $0^0 = 1$，取 $\tau = 0$ 对应于对迭代值进行简单平均，而取 $\tau > 0$ 则可以为最后的估计值赋予更大的权重 θ_n。这个平均步骤可以通过递归方式计算：

$$\bar{\theta}_n = \left(1 - \frac{\omega_n}{\sum_{k=0}^n \omega_k}\right) \bar{\theta}_{n-1} + \frac{\omega_n}{\sum_{k=0}^n \omega_k} \theta_n.$$

备注4.4。注意到，由于平均估计量 θ_n 预计会比 θ_n迭代值收敛得更快，因此可以在 θ_n 而不是 θ_n处查询海森矩阵预言机，从而将方程 (5)修改为

$$\tilde{H}_n := M_{\ell,n} h_n(\bar{\theta}_{n-1}). \tag{8}$$

然而，计算海森向量积通常需要将梯度作为中间步骤，而为了更新 θ_n ，必须在 θ_n处计算¯随机梯度。因此，在 θ_n 而不是θ_n 处查询海森矩阵的随机预言机甚至可能效率更低。

以下定理给出了平均估计量（采用更新 (5) 或 (8)）的强相合性、其几乎必然收敛速率以及渐近效率。

定理 4.5。 在假设 ***(A1)*** 和 ***(A2)***下，若 $\alpha+\gamma>\frac{3}{2}$，由 (4)、(5)_（或 (8)）、(6) 和 (7) 定义的估计值 θ_n 和_θ_n 满足

$$\theta_n \xrightarrow[n\to+\infty]{a.s.} \theta^* \quad and \quad \bar{\theta}_n \xrightarrow[n\to+\infty]{a.s.} \theta^*.$$

另外，假设 ***(A3)***、***(A4)*** 满足 $q>\frac{2}{\alpha}$ 和 $q'>\frac{2}{\gamma}$，且满足 ***(A6)***和 $\alpha<1$。那么，对任意 $\eta>0$:

$$\|\theta_n-\theta^*\|^2=\mathcal{O}\left(\frac{\ln n}{n^\alpha}\right) \quad and \quad \|\bar{\theta}_n-\theta^*\|^2=\mathcal{O}\left(\frac{\ln n}{n}\right) \quad a.s.,$$

$$\left\|A_n-H^{-1}\right\|_F^2=\mathcal{O}\left(\frac{(\ln n)^{1+\eta}}{n^{\min\{\gamma,\alpha\}}}\right) \ a.s. \quad and \quad \left\|\bar{A}_n-H^{-1}\right\|_F^2=\mathcal{O}\left(\frac{(\ln n)^{1+\eta}}{n}\right) \quad a.s.$$

此外，在假设***(A5)***下，我们有

$$\sqrt{n}\left(\bar{\theta}_n-\theta^*\right)\xrightarrow[n\to+\infty]{\mathcal{L}}\mathcal{N}\left(0,H^{-1}\Gamma H^{-1}\right).$$

值得注意的是，若 $\alpha<1$，则对于任意$c_\alpha>0$，平均估计量都能达到渐近效率。因此，可以通过采用较大的 c_α ，并以较大的 c'_α 作为平衡，来减缓步长 α_n 的递减速度，同时仍能得到渐近有效的算法。

5 流式牛顿算法

5.1 算法

本节重点介绍 mSNA 算法的流式版本，该算法对数据进行一次遍历所需的总运算量可能为$\mathcal{O}(Nd)$。为此，我们考虑目标函数为某个已知函数的期望，且该函数关于其第二个参数 θ 二阶可微： f

$$F(\theta)=\mathbb{E}\left[f(\xi,\theta)\right],$$

其中， ξ 是一个取值于测度空间 $\mathcal{X}$的随机变量。沿用（Godichon-Baggioni 和 Werge，2025）中提出的思路，我们从现在起假设，在每次迭代 n时，都可以访问 b 个新的、独立同分布的 ξ副本，这些副本以块 $\{\xi_{n,1},\ldots,\xi_{n,b}\}$的形式到达。流式 mSNA 算法（及其加权平均版本）随后通过在更新（4）和（6）中使用以下随机预言机来定义：

$$g_n(\theta)=\frac{1}{b}\sum_{i=1}^{b}\nabla f(\xi_{n,i},\theta) \quad \text{and} \quad h_n(\theta)=\frac{1}{b}\sum_{i=1}^{b}\nabla^2 f(\xi_{n,i},\theta). \tag{9}$$

将 $N = nb$ 视为经过 n 次迭代后处理的样本总数，处理 N 个样本所需的迭代次数为 N/b。这比纯在线（$b = 1$）设置少 b 倍，从而降低了总计算复杂度，具体讨论如下。尽管迭代次数减少了 b倍，小批量预言机（9）的协方差也减少了b倍。因此，假设 **(A5)** 中的极限协方差矩阵 Γ 变为Σ/b，其中 $\Sigma := \mathrm{Cov}\left[\nabla f(\xi, \theta^*)\right]$ 是由假设**(B5)**定义的单个样本的协方差矩阵。这些估计值仍然是渐近有效的，以下推论由定理4.1和4.5给出。

推论 5.1. 在假设 ***(B1)*** 至 ***(B5)***（见附录 ***C***）对由 (9) 定义的预言机成立，且假设 ***(B4)b)*** 在 $q' > \frac{2}{\gamma}$ 下成立时，采用步长 $\alpha_n = 1/(n + c'_\alpha)$ 和小批量大小 b 的流式 ***mSNA*** 算法的估计量 θ_n 渐近有效：

$$\sqrt{N}\left(\theta_n - \theta^*\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, H^{-1}\Sigma H^{-1}\right),$$

其中 $\Sigma := \mathrm{Cov}\left[\nabla f(\xi, \theta^*)\right]$。此外，在假设 ***(A6)*** 下，采用步长指数 $\alpha < 1$、小批量大小 b，且假设 ***(B4)*** 在 $q > \frac{2}{\alpha}$ 和 $q' > \frac{2}{\gamma}$ 下成立时，流式平均 ***mSNA*** 算法的估计量 θ_n 渐近有效：

$$\sqrt{N}\left(\bar{\theta}_n - \theta^*\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, H^{-1}\Sigma H^{-1}\right).$$

值得注意的是，采用 $\alpha_n := 1/(n + c'_\alpha)$ 的流式 mSNA 算法，以及采用任意 $(\alpha, c_\alpha, c'_\alpha)$ 序列步长且满足 $\alpha < 1$ 的流式平均 mSNA 算法，均能达到渐近效率，并具有与小批量大小为 $b = 1$ 的纯在线版本相同的渐近分布。请注意，对于流式平均 mSNA，可以令 $c_\alpha = b^a$，其中 $a \in [0, 1]$（通常为 $a = 1 - \alpha$）。这样，在迭代次数减少时便可采用更大的步长（更多详情见 Godichon-Baggioni 等人，2023）。

5.2 计算复杂度讨论

我们的算法每次迭代的工作量可总结如下：

计算 H_n。计算海森矩阵预言机所选列的复杂度取决于函数 f。我们假设，无需构造完整矩阵即可高效计算海森矩阵$\nabla^2 f(\xi_{n,\,i},\ \theta_{n-1})$与单个向量的乘积（即海森向量积）（例如，参见 Pearlmutter，1994）。因此，对于单个样本$\xi_{n,\,i}$，计算由$M_{\ell,\,n}$ 选出的 ℓ 列，在许多标准的情形下需要 $\mathcal{O}(\ell d)$次运算。计算 H_n 需要针对每个 b 样本计算这些 ℓ 列并求其平均值，复杂度为 $\mathcal{O}(\ell b d)$量级。

计算 $\left\|\tilde{H}_n\right\|_{op}$。计算 $\tilde{H}_n$ 的算子范数需要 $\mathcal{O}(\ell^2 d)$次运算，因为 $\tilde{H}_n$ 中只有 ℓ 列非零。只要 $\ell \leq \sqrt{d}$，其量级便为$\mathcal{O}(d^2)$。

更新 A_n。计算乘积 $A_{n-1}\tilde{H}_n^T$ 和 $\tilde{H}_n A_{n-1}\tilde{H}_n^T$ 需要 $\mathcal{O}(\ell d^2)$ 次运算。因此，A_n 的整体更新主要由这些矩阵乘法决定，需要 $\mathcal{O}(\ell d^2)$ 次运算。

更新 θ_n。首先需要 $\mathcal{O}(bd)$ 次运算来计算梯度预言机$g_n(\theta_{n-1})$，即小批量上的平均值。接着，计算乘积 $(A_{n-1} + \nu_n I_d)\, g_n(\theta_{n-1})$ 需要 $\mathcal{O}(d^2)$ 次运算。总成本为 $\mathcal{O}(bd + d^2)$。

Up **更新** θ_n。此操作需要 $\mathcal{O}(d)$ 次运算（向量缩放与加法）。

算法每次迭代的总成本 将这些成本相加，每次迭代的总成本为：

$$\underbrace{\mathcal{O}\left(\ell b d + \ell d^2\right)}_{\text{Update } A_n} + \underbrace{\mathcal{O}\left(bd + d^2\right)}_{\text{Update } \theta_n} + \underbrace{\mathcal{O}\left(d\right)}_{\text{Update } \bar{\theta}_n \text{ (if used)}} = \mathcal{O}\left(\ell\left(bd + d^2\right)\right)$$

因此，只要批量大小 b 不大于维度 d，一次迭代的成本阶为 $\mathcal{O}(\ell d^2)$。在只需对 N 个数据点进行一次遍历的流式设置中，该算法执行 N/b 次迭代，总复杂度为 $\mathcal{O}(\frac{N}{b}\ell d^2)$。因此，通过选择 $b = d$ 和 $\ell = 1$，该算法在一次遍历中的总复杂度可达到 $\mathcal{O}(Nd)$ 次运算。

6 数值实验

本节通过实证评估所提出的掩码随机牛顿算法（mSNA）及其加权平均变体的性能，尤其关注流式设置。我们在合成数据集和真实世界数据集上，将所提方法与随机梯度下降（SGD）及其平均变体进行比较。我们的实现采用 PyTorch 框架，代码可在 https://github.com/guillaume-salle/SNA 获取。所有实验均在一台普通笔记本电脑上进行，计算由多核 CPU 执行，并使用 6 个线程进行并行处理；未使用 GPU 加速。虽然我们的模拟主要关注线性回归和逻辑回归的经典示例，但该方法还有许多其他应用，例如 Godichon-Baggioni 等人（2025）的研究。

6.1 实验设置

我们比较以下算法的性能：

- **随机梯度下降：** 由下式定义的标准随机梯度下降更新

$$\theta_n = \theta_{n-1} - \alpha_n g_n(\theta_{n-1}). \tag{10}$$

- **平均随机梯度下降：**采用估计量 θ_n 的加权平均变体，该估计量由（7）定义 .
- **mSNA：**我们提出的掩码随机牛顿算法，由（4）和（6）定义。_ (5),

- **平均掩码随机牛顿算法：**我们方法的加权平均变体，采用估计量 θ_n，由（7）定义。

对于所有优化器，批量大小均设为 b，等于 θ的维度 d。

对于非平均方法随机梯度下降和 mSNA，我们采用标准步长 α_n :=1/（$n + n_0$）。对于平均随机梯度下降和平均掩码随机牛顿算法，我们采用较大的步长 α_n := $d^{0.25}$/（$n^{0.75} + d^{0.25} * n_0$），并将 $\tau = 2$ 用于平均步骤（7）。超参数 n_0 决定初始步长 α_1，并与目标函数的期望光滑常数 $\mathcal{L}$ 相关。合成数据的取值设为 $n_0 = d$，真实数据则通过线搜索进行调优。

此外，对于 mSNA 和平均掩码随机牛顿算法，参数ℓ决定每次迭代中条件矩阵 A_n 被更新的列数和行数。考虑了 ℓ的三种选择： $\ell = 1$, $d^{0.25}$, $\sqrt{d}$。我们对条件矩阵更新（6）使用步长 $\gamma_n = 1/$（$n^{0.75}+n_0$），并对 θ_n 更新（4）使用正则化参数 $\nu = 0$。请注意，对于平均掩码随机牛顿算法，$\gamma = 0.75$ 的这一选择是定理 4.5 中关于 θ_n收敛的条件 $\alpha+\gamma > 3/2$ 的边界情形。

6.2 合成数据实验

为评估不同方法的性能，我们分析二次误差 $\|\theta_n - \theta^*\|^2$ 随样本量的变化。我们还考察逆 Hessian 估计的平方 Frobenius 范数误差 $\|A_n - H^{-1}\|_F^2$ 的变化，以及不同算法的总计算时间。请注意，对于这两种方法，初始化 θ_0 是在以 θ^* 为中心的单位球面上随机选取的，并且A_0 被设为 I_d。

6.2.1 线性回归

我们考虑一个线性回归问题，其中 $F(\theta) = \frac{1}{2}\mathbb{E}[(Y - X^T\theta)^2]$，且 $\theta \in \mathbb{R}^d$。这是一个有趣的情形，因为海森矩阵已知且为常数，由 $\nabla^2 F(\theta) = \mathbb{E}[XX^T] =$给出： Σ_X 对所有 $\theta \in \mathbb{R}^d$ 成立，因此我们可以显式计算 H^{-1}。为了构造一个具有挑战性的场景，我们考虑如下模型：

- X服从多元高斯分布 $\mathcal{N}(0, \Sigma_X)$。协方差矩阵Σ_X被构造成病态矩阵。具体而言，$\Sigma_X = U\Lambda U^T$，其中 Λ是一个对角矩阵，其特征值在 10^{-2} 和 1 之间均匀分布，而 U 是

一个随机正交矩阵，使 Σ_X的条件数为 10^2 。这种构造确保协方差矩阵是病态的，且主成分不与坐标轴对齐，因此即使对于使用对角条件矩阵的条件化随机梯度下降方法，这一优化问题仍然具有挑战性。

- $Y = X^T\theta^* + \epsilon$，其中参数 θ^* 从标准化高斯分布中随机选取，且 $\epsilon \sim \mathcal{N}(0, 1)$。

我们取 $d = 1000$ 和 $N = 10^7$，以在维度较大的同时，仍能以批量大小为 d 进行合理次数的算法迭代。

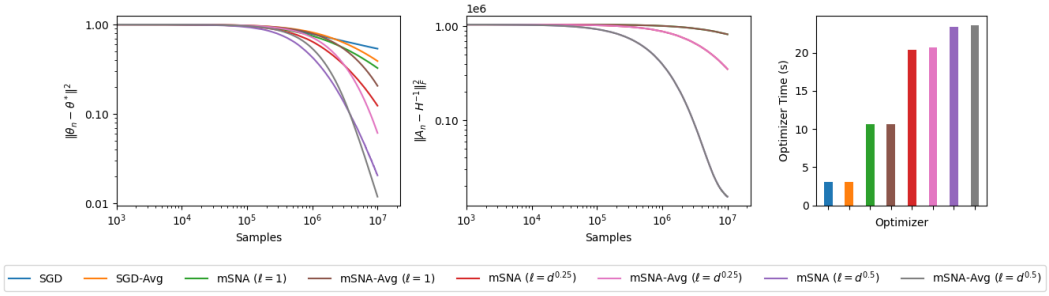


图1： $d = 1000$ 的线性回归。从左至右依次比较：二次误差 $\|\theta_n - \theta^*\|^2$ 的演变、二次误差 $\left\|A_n - H^{-1}\right\|_F^2$、不同方法的计算时间。

从图1可以看出，正如病态条件下所预期的那样，标准的基于梯度的方法无法有效收敛。相比之下，牛顿型方法的表现显著更好，而计算成本仅略有增加。对于所考虑的所有算法变体，平均版本的表现始终优于非平均版本。正如预期，增大掩码大小 ℓ 可以提升性能，但计算时间也会随之增加。

6.2.2 逻辑回归

我们在此考虑一个逻辑回归问题：

$$F(\theta) = \mathbb{E}[\log(1 + \exp(X^T\theta)) - X^T\theta Y],$$

其中 $X \sim \mathcal{N}(0, \Sigma_X)$ 与第 6.2.1 节中的相同，且 $Y|X \sim \mathcal{B}(\sigma(X^T\theta^*))$，其中 $\sigma(x) = \frac{e^x}{1+e^x}$ 是 sigmoid函数。

对于这种情况，我们没有极小化点处海森矩阵的闭式表达式：$H = \nabla^2 F(\theta^*) = \mathbb{E}[\sigma(X^T\theta^*)(1 - \sigma(X^T\theta^*))XX^T]$。通过蒙特卡洛方法，我们在一个大样本上计算 H 的经验估计，并利用它计算逆 H^{-1}的估计值。

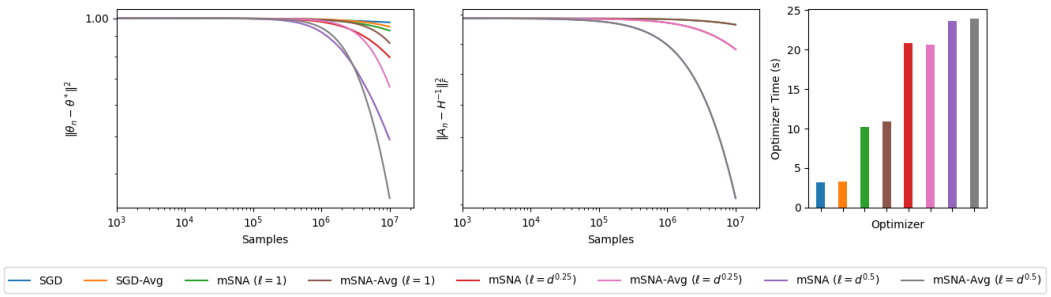


图2：使用 $d = 1000$ 的逻辑回归。从左到右依次比较：二次误差 $\|\theta_n - \theta^*\|^2$ 的变化、二次误差 $\|A_n - H^{-1}\|_F^2$、不同方法的计算时间。

在图2中，我们观察到与线性回归情形相似的趋势。在逻辑回归的情况下，海森矩阵 H 表现出非常小的特征值和较大的条件数，这进一步降低了随机梯度下降优化器的性能。

6.3 真实数据上的岭逻辑回归

我们使用岭逻辑回归，在真实世界数据集上评估这些算法的二分类性能。所研究数据集的特征汇总于表1。

表1：本研究所用数据集的主要特征。

	维度	训练集大小	初始集大小	测试集大小
Adult	98	38,682	390	9,769
四子棋	85	53,504	540	13,512
COVTYPE	55	460,160	4,648	116,203
MNIST	785	54,430	1,570	14,000
蘑菇	95	6,309	190	1,625
网络钓鱼	39	8,755	88	2,211

每个数据集都分为训练集和测试集。为了评估不同方法的性能，我们同时考察预测准确率以及在训练集和测试集上计算的经验损失函数。

Adult（Becker 和 Kohavi，1996）包含用于收入预测的人口普查数据，其中多个分类变量被转换为二元指示变量。四子棋（Tromp，1995）是一个关于四子棋游戏的数据集，目标是根据当前棋盘状态预测下一步棋。COVTYPE（Blackard，1998）最初包含多种森林覆盖类型；在本研究中，我们专注于区分“云杉/冷杉”（标记为 1）与所有其他类别（标记为 0 ）。MNIST （ Lecun 等人，1998 ）是一个用于手写数字识别的数据集，我们专注于区分数字“0”（标记为 1）与所有其他数字（标记为 0）。蘑菇（UCI 机器学习资料库，

1981年）包含用于判定蘑菇是否可食用的形态学属性。网络钓鱼（Mohammad 和 McCluskey，2012）旨在检测恶意网站，其特征全部为类别特征，我们将其编码为二进制形式。这些数据集常用于对二元分类器进行基准测试（Toulis 和 Airoldi，2017）。

我们的目标是应用岭回归模型来预测每个数据集的二元响应变量。岭回归的目标函数定义为

$$F(\theta) = \mathbb{E}\left[\log(1+\exp(X^T\theta)) - X^T\theta Y + \frac{\lambda}{2}\|\theta\|^2\right]$$

其中，$Y \in \{0,1\}$、$X \in \mathbb{R}^d$ 和 $\lambda > 0$ 是 L_2正则化参数。对于所有优化器，我们都会对初始估计量 θ_0 进行初始化：从 0 开始，使用通过线搜索找到的固定学习率，在最多包含 $\max\{N/100, 2d\}$ 个数据的初始批次上执行 100 步梯度下降。由于所有优化器共用此初始化过程，因此我们不将 θ_0 的初始化计入算法的计算时间。

对于 mSNA 和平均掩码随机牛顿算法优化器，我们还会初始化条件矩阵 A_0：在同一批次上计算先前得到的 F 在 θ_0 处的海森矩阵估计量，并将其求逆。此初始化的计算成本为 $O(d^3)$，并计入计算时间。

在表 2 中，我们观察到，mSNA 算法及其平均版本表现相近，或优于基于梯度的方法，计算时间虽较长但仍具有可比性。这些结果不仅凸显了我们算法的卓越性能，也表明它们已具备实际部署的条件。

结论

本文介绍了一种高效的在线小批量随机牛顿算法（mSNA），专为随机情境下的光滑凸优化问题而设计。我们的工作推进了 Godichon-Baggioni 等人（2025）提出的通用随机牛顿算法（USNA）：在极小化点附近具有局部强凸性的情况下，我们证明了该算法具有渐近效率，且无需进行迭代平均，这是一项重要的理论改进。为此，我们证明了一个关于线性函数的 Robbins-Monro 算法收敛性的一般定理，该定理本身也具有独立的研究价值。

所提出的 mSNA 算法以流式方式运行，按小批量处理数据，因此尤其适用于大规模数据集。对于单次遍历 N个数据点，该算法的总计算复杂度为 $O(Nd)$。在合成数据集和真实世界数据集上，对带或不带岭惩罚的线性回归和逻辑回归进行的数值实验表明，与随机梯度下降和平均随机梯度下降相比，mSNA 及其平均变体始终具有竞争力。这些结果不仅凸显了我们方法强有力的理论保证，也表明它已具备解决真实世界问题的实际应用条件。

表 2：流式优化器在各种数据集上的性能。

数据集	优化器	训练准确率	测试准确率	训练损失	测试损失	Time
adult	SGD-Avg	**76.86**	**76.87**	5.72e-01	5.75e-01	31.8 毫秒
	SGD	76.75	76.72	**5.67e-01**	**5.70e-01**	**27.0 毫秒**
	mSNA-平均 ($\ell = 0.25$)	76.17	76.17	5.74e-01	5.77e-01	151.4 毫秒
	mSNA-平均 ($\ell = 0.5$)	76.17	76.17	5.74e-01	5.77e-01	149.8 毫秒
	mSNA-平均	76.17	76.17	5.74e-01	5.77e-01	78.6 毫秒
	mSNA ($\ell = 0.25$)	76.17	76.17	5.74e-01	5.77e-01	131.3 毫秒
	mSNA ($\ell = 0.5$)	76.17	76.17	5.74e-01	5.77e-01	146.7 毫秒
	mSNA	76.17	76.17	5.74e-01	5.77e-01	78.1 毫秒
四子棋	SGD-Avg	77.04	76.72	4.68e-01	4.69e-01	**42.2 毫秒**
	SGD	77.00	76.61	4.69e-01	4.70e-01	44.8 毫秒
	mSNA-平均 ($\ell = 0.25$)	**79.26**	**79.18**	4.38e-01	4.39e-01	205.0 毫秒
	mSNA-平均 ($\ell = 0.5$)	79.26	79.18	**4.38e-01**	**4.39e-01**	217.4 毫秒
	mSNA-平均	79.25	79.17	4.38e-01	4.39e-01	123.5 毫秒
	mSNA ($\ell = 0.25$)	79.16	79.08	4.39e-01	4.41e-01	201.5 毫秒
	mSNA ($\ell = 0.5$)	79.14	79.10	4.39e-01	4.41e-01	217.8 毫秒
	mSNA	79.15	79.05	4.39e-01	4.41e-01	119.8 毫秒
covtype	SGD-Avg	**58.36**	**58.23**	**6.66e-01**	**6.66e-01**	527.6 毫秒
	SGD	57.56	57.35	6.70e-01	6.71e-01	**494.9 毫秒**
	mSNA-平均 ($\ell = 0.25$)	55.64	55.68	6.79e-01	6.79e-01	2.35 秒
	mSNA-平均 ($\ell = 0.5$)	55.64	55.68	6.79e-01	6.79e-01	2.22 秒
	mSNA-平均	55.64	55.68	6.79e-01	6.79e-01	1.51 秒
	mSNA ($\ell = 0.25$)	55.62	55.65	6.79e-01	6.79e-01	2.31 秒
	mSNA ($\ell = 0.5$)	55.62	55.65	6.79e-01	6.79e-01	2.19 秒
	mSNA	55.62	55.65	6.79e-01	6.79e-01	1.47 秒
MNIST	SGD-Avg	86.76	86.66	5.05e-01	5.21e-01	23.8 毫秒
	SGD	**86.95**	**87.11**	**3.09e-01**	**3.14e-01**	**19.0 毫秒**
	mSNA-平均 ($\ell = 0.25$)	86.08	85.76	3.28e-01	3.35e-01	102.3 毫秒
	mSNA-平均 ($\ell = 0.5$)	86.08	85.76	3.28e-01	3.35e-01	98.8 毫秒
	mSNA-平均	86.08	85.80	3.28e-01	3.35e-01	65.0 毫秒
	mSNA ($\ell = 0.25$)	85.75	85.49	3.34e-01	3.41e-01	243.2 毫秒
	mSNA ($\ell = 0.5$)	85.75	85.49	3.34e-01	3.41e-01	133.7 毫秒
	mSNA	85.76	85.49	3.35e-01	3.41e-01	50.2 毫秒
蘑菇	SGD-Avg	94.37	95.08	2.43e-01	2.19e-01	9.3 毫秒
	SGD	94.37	95.08	2.41e-01	2.17e-01	**4.6 毫秒**
	mSNA-平均 ($\ell = 0.25$)	99.18	98.95	3.50e-02	4.38e-02	47.1 毫秒
	mSNA-平均 ($\ell = 0.5$)	99.18	98.95	3.50e-02	4.37e-02	48.6 毫秒
	mSNA-平均	99.18	98.95	3.50e-02	4.38e-02	27.8 毫秒
	mSNA ($\ell = 0.25$)	99.69	99.08	**2.98e-02**	3.83e-02	47.4 毫秒
	mSNA ($\ell = 0.5$)	**99.70**	**99.20**	2.98e-02	**3.81e-02**	49.3 毫秒
	mSNA	99.66	99.08	2.98e-02	3.84e-02	30.3 毫秒
网络钓鱼	SGD-Avg	89.45	89.69	2.56e-01	2.46e-01	20.2 毫秒
	SGD	89.26	89.51	2.59e-01	2.49e-01	**13.3 毫秒**
	mSNA-平均 ($\ell = 0.25$)	93.79	94.08	1.61e-01	1.63e-01	95.0 毫秒
	mSNA-平均 ($\ell = 0.5$)	93.77	94.08	**1.61e-01**	**1.62e-01**	124.2 毫秒
	mSNA-平均	93.75	94.08	1.62e-01	1.63e-01	52.5 毫秒
	mSNA ($\ell = 0.25$)	93.77	94.12	1.66e-01	1.66e-01	78.8 毫秒
	mSNA ($\ell = 0.5$)	**93.80**	**94.17**	1.65e-01	1.65e-01	95.4 毫秒
	mSNA	93.77	94.08	1.67e-01	1.66e-01	62.7 毫秒

局限性与未来工作：当前工作聚焦于凸情形，在这一范围内仍有若干改进和拓展方向。

首先，将矩阵逆估计表述为优化问题，为在结构约束下估计逆矩阵（例如采用岭回归或 Lasso 正则化）提供了可能。当矩阵稀疏或其最小特征值过于接近零时，这种做法可能很有价值。

其次，探究算法在较弱条件下的鲁棒性，例如放宽光滑性假设（如 (**A2**) b）），将是未来值得研究的方向。在不依赖 Hessian Lipschitz 连续性假设 (**A6**) 的情况下，确立平均掩码随机牛顿算法的渐近效率，仍是一项开放且值得关注的挑战。

最后，尽管用于海森矩阵近似的随机掩码策略简单而有效，但研究其他草图技术或用于选择海森矩阵信息的自适应机制，或许能带来进一步改进。

参考文献

Agarwal, N., Bullins, B., and Hazan, E. (2017). 面向机器学习的线性时间二阶随机优化。机器学习研究期刊, 18(116):1–40. Bach, F. and Moulines, E. (2011). 面向机器学习的随机逼近算法的非渐近分析。载于神经信息处理系统进展。 Becker, B. and Kohavi, R. (1996). Adult. UCI 机器学习资料库。 Bercu, B., Godichon, A., and Portier, B. (2020). 一种用于逻辑回归参数估计的高效随机牛顿算法。*SIAM* 控制与优化杂志, 58(1):348–367. Blackard, J. (1998). Covertype. UCI 机器学习资料库。 Bottou, L., Curtis, F. E., and Nocedal, J. (2018). 大规模机器学习的优化方法。*SIAM* 评论, 60(2):223–311. Boyer, C. and Godichon-Baggioni, A. (2023). 关于随机牛顿算法及其加权平均版本的渐近收敛速率。计算优化与应用, 84(3):921–972. Byrd, R. H., Chin, G. M., Neveitt, W., and Nocedal, J. (2011). 关于在机器学习优化方法中使用随机 Hessian 信息。*SIAM* 优化杂志, 21(3):977–995. Carpentier, P., Chancelier, J.-P., Cohen, G., and De Lara, M. (2015). 随机多阶段优化：处于离散时间随机控制与随机规划的交汇点，第 75 卷，收录于概率论与随机建模。Springer 国际出版公司。

Cénac, P., Godichon-Baggioni, A., and Portier, B. (2025). 一种用于估计非线性回归模型参数的高效平均随机高斯-牛顿算法。伯努利, 31(1):1 – 29. Dennis, Jr., J. E. and Moré, J. J. (1977). 拟牛顿方法：动机与理论。*SIAM* 评论, 19(1):46–89. Duchi, J., Hazan, E., and Singer, Y. (2011). 用于在线学习和随机优化的自适应次梯度方法。机器学习研究期刊, 12(61):2121–2159. Duflo, M. (1996). 随机算法。数学与应用。施普林格柏林·海德堡。 Godichon-Baggioni, A. (2019). 平均随机梯度算法渐近方差的在线估计。统计规划与推断期刊, 203:1–19. Godichon-Baggioni, A., Lu, W., and Portier, B. (2025). 随机优化中海森矩阵逆矩阵的在线估计及其在通用随机牛顿算法中的应用。预印本，参见 https://hal.science/hal-04391570。 Godichon-Baggioni, A. and Werge, N. (2025). 关于流式数据的自适应随机优化：一种运算量为 o(dN) 的牛顿法。机器学习研究期刊, 26(59):1–49. Godichon-Baggioni, A., Werge, N., and Wintenberger, O. (2023). 流式数据随机逼近算法的非渐近分析。*ESAIM*：概率与统计, 27:482–514. Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. (2019). 随机梯度下降：一般分析与改进的收敛速率。载于 Chaudhuri, K. and Salakhutdinov, R. 编，第 *36* 届国际机器学习会议论文集，第 97 卷，收录于机器学习研究论文集，第 5200–5209 页。PMLR。 Gower, R. M. and Richtárik, P. (2017). 随机拟牛顿更新是线性收敛的矩阵求逆算法。*SIAM* 矩阵分析与应用杂志, 38(4):1380–1409. 出版方：工业与应用数学学会。 Hinton, G. (2012). 面向机器学习的神经网络，第 6 讲。在线课程，Coursera。 Kingma, D. P. and Ba, J. (2017). Adam：一种随机优化方法。 Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). 应用于文档识别的基于梯度的学习。*IEEE*论文集, 86(11):2278–2324. Leluc, R. and Portier, F. (2022). 采用坐标采样的随机梯度下降：理论与实践。机器学习研究期刊, 23(342):1–47。

Leluc, R. 和 Portier, F. (2023). 条件随机梯度下降的渐近分析。机器学习研究汇刊。Mohammad, R. 和 McCluskey, L. (2012). 钓鱼网站。UCI 机器学习资料库。Mokkadem, A. 和 Pelletier, M. (2011). 平均过程的一种推广：双时间尺度算法的应用。*SIAM* 控制与优化杂志，49(4):1523–1543。Moritz, P., Nishihara, R. 和 Jordan, M. (2016). 一种线性收敛随机 L-BFGS 算法。载于第*19*届人工智能与统计国际会议论文集，第249–258页。PMLR。Pearlmutter, B. A. (1994). 海森矩阵的快速精确乘法。神经计算，6(1):147–160。Pelletier, M. (1998a). 随机算法的几乎必然渐近行为。随机过程及其应用，78(2):217–244。Pelletier, M. (1998b). 随机逼近的弱收敛速率及其在多个目标和模拟退火中的应用。应用概率年刊，8(1):10–44。Pelletier, M. (2000). 平均随机算法的渐近几乎必然效率。*SIAM* 控制与优化杂志，39(1):49–72。Polyak, B. T. (1990). 一种新的随机逼近型方法。自动化与远程控制，(7):98–107。俄文。英文译文见：自动化与远程控制，51(7):937–946，1990。Polyak, B. T. 和 Juditsky, A. B. (1992). 通过平均加速随机逼近。*SIAM* 控制与优化杂志，30(4):838–855。Ruppert, D. (1988). 从缓慢收敛的 Robbins–Monro 过程中高效估计。技术报告，康奈尔大学运筹学与工业工程系。Schraudolph, N. N., Yu, J. 和 Günter, S. (2007). 一种用于在线凸优化的随机拟牛顿方法。载于第十一届人工智能与统计国际会议论文集，第436–443页。PMLR。Surendran, S., Fermanian, A., Godichon-Baggioni, A. 和 Le Corff, S. (2024). 有偏自适应随机逼近的非渐近分析。神经信息处理系统进展，37:12897–12943。Toulis, P. 和 Airoldi, E. M. (2017). 基于随机梯度的估计量的渐近性质和有限样本性质。统计年刊，45(4):1694–1727。Tromp, J. (1995). 四子棋。UCI 机器学习资料库。

UCI 机器学习资料库（1981年）。Mushroom。UCI 机器学习资料库 .

A 证明

A.1 命题3.1的证明

命题*3.1*的证明。 下面的引理给出了 A_n 的算子范数几乎必然且渐近的界，而无需依赖关于 θ_n 行为的信息。

引理 A.1。 在假设 *(A1)* 和 *(A2)b)* 下，对于任意 $\eta > 0$，我们有：

$$\lambda_{\max}(A_n) = \mathcal{O}(n^{1-\gamma}(\ln n)^{1+\eta}) \quad a.s.$$

该引理的证明较为复杂，故推迟到下一节 A.3。我们有

$$A_n = A_{n-1} - \gamma_n \left(\nabla J_{sym}(A_{n-1}) + \epsilon_n + r_n\right),$$

带有噪声

$$\begin{aligned}\epsilon_n \coloneqq & \left(\tilde{H}_n A_{n-1} + A_{n-1}\tilde{H}_n^T - 2M_{n,\ell} - \gamma_n \tilde{H}_n A_{n-1}\tilde{H}_n^T\right)\mathbf{1}_{\gamma_n\left\|\tilde{H}_n\right\|_{op}\le\frac{1}{2}} \\ & - \mathbb{E}\left[\left(\tilde{H}_n A_{n-1} + A_{n-1}\tilde{H}_n^T - 2M_{n,\ell} - \gamma_n \tilde{H}_n A_{n-1}\tilde{H}_n^T\right)\mathbf{1}_{\gamma_n\left\|\tilde{H}_n\right\|_{op}\le\frac{1}{2}} \mid \mathcal{F}_{n-1}\right],\end{aligned}$$

和残差项

$$\begin{aligned} r_n &\coloneqq r_n^{(1)} + r_n^{(2)} + r_n^{(3)} \\ r_n^{(1)} &\coloneqq A_{n-1}\left(\nabla^2 F(\theta_{n-1}) - H\right) + \left(\nabla^2 F(\theta_{n-1}) - H\right)A_{n-1} \\ r_n^{(2)} &\coloneqq \mathbb{E}\left[\gamma_n \tilde{H}_n A_{n-1}\tilde{H}_n^T \mathbf{1}_{\gamma_n\left\|\tilde{H}_n\right\|_{op}\le\frac{1}{2}} \mid \mathcal{F}_{n-1}\right] \\ r_n^{(3)} &\coloneqq -\mathbb{E}\left[\left(\tilde{H}_n A_{n-1} + A_{n-1}\tilde{H}_n^T + 2M_{n,\ell}\right)\mathbf{1}_{\gamma_n\left\|\tilde{H}_n\right\|_{op}>\frac{1}{2}} \mid \mathcal{F}_{n-1}\right]. \end{aligned}$$

项 $r^{(1)}$ 计入了估计量 θ_n 的误差，从而导致对 H 进行估计时产生偏差。项 $r^{(2)}$ 计入了为确保 A_n 的正定性而添加的项，项 $r^{(3)}$ 则计入了截断。根据构造，噪声 ϵ_n 满足 $\mathbb{E}\left[\epsilon_n \mid \mathcal{F}_{n-1}\right] = 0$。此外，利用针对 $p \ge 2$ 的不等式 $(a+b)^p \le 2^{p-1}(a^p + b^p)$，我们有

$$\begin{aligned}\mathbb{E}\left[\|\epsilon_n\|_F^2 \mid \mathcal{F}_{n-1}\right] &\le \mathbb{E}\left[\left\|\tilde{H}_n A_{n-1} + A_{n-1}\tilde{H}_n^T + 2M_{n,\ell} + \gamma_n \tilde{H}_n A_{n-1}\tilde{H}_n^T\right\|_F^2 \mathbf{1}_{\gamma_n\left\|\tilde{H}_n\right\|_{op}\le\frac{1}{2}} \mid \mathcal{F}_{n-1}\right] \\ &\le \mathbb{E}\left[\left(2\left\|A_{n-1}\right\|_F\left\|\tilde{H}_n\right\|_F + 2\left\|M_{n,\ell}\right\|_F + \frac{1}{2}\left\|A_{n-1}\right\|_F\left\|\tilde{H}_n\right\|_F\right)^2 \mid \mathcal{F}_{n-1}\right] \\ &\le \mathbb{E}\left[2 \times \left(\frac{5}{2}\left\|A_{n-1}\right\|_F\left\|\tilde{H}_n\right\|_F\right)^2 + 2 \times \left(2\left\|M_{n,\ell}\right\|_F\right)^2 \mid \mathcal{F}_{n-1}\right] \\ &= \mathcal{O}\left(1 + \|A_{n-1}\|_F^2\right).\end{aligned}$$

由于 θ_n 几乎处处$\xrightarrow[n\to\infty]{}\theta^*$，并且利用假设 **(A4)b)**，当 n 足够大时，$\|\epsilon_n\|_F$ 具有 q' 阶矩；按照相同的推理，我们对二阶矩也有：

$$\mathbb{E}\left[\|\epsilon_n\|_F^{q'} \mid \mathcal{F}_{n-1}\right] = \mathcal{O}\left(1 + \|A_{n-1}\|_F^{q'}\right).$$

For 余项 r_n，取 $\nabla^2 F$ 关于 θ^* 的非递增连续模 μ ，
we have

$$\left\|\nabla^2 F(\theta_{n-1}) - H\right\| \le \mu(\|\theta_{n-1} - \theta^*\|) \le \mu(\delta_n).$$

我们得到

$$\begin{aligned}
\left\|r_n^{(1)}\right\|_F &\le 2\|A_{n-1}\|_F \left\|\nabla^2 F(\theta_{n-1}) - H\right\|_F = \mathcal{O}\left(\|A_{n-1}\|_F \mu(\delta_n)\right), \\
\left\|r_n^{(2)}\right\|_F &\le \gamma_n \|A_{n-1}\|_F \mathbb{E}\left[\left\|\tilde{H}_n\right\|_F^2 \mid \mathcal{F}_{n-1}\right] = \mathcal{O}\big(\gamma_n \|A_{n-1}\|_F\big) \text{ a.s.}, \\
\left\|r_n^{(3)}\right\|_F &\le \mathbb{E}\left[\left(2\|A_{n-1}\|_F \left\|\tilde{H}_n\right\|_F + 2\|M_{n,\ell}\|_F\right) \times 2\gamma_n \left\|\tilde{H}_n\right\|_F \mid \mathcal{F}_{n-1}\right] \\
&= \mathcal{O}\left(\gamma_n\left(1 + \|A_{n-1}\|_F\right)\right) \text{ a.s.}
\end{aligned}$$

因此，

$$\|r_n\|_F = \mathcal{O}\left(\left(1 + \|A_{n-1}\|_F\right)\left(\mu(\delta_n) + \gamma_n\right)\right) \text{ a.s.}$$

将命题 B.1 应用于 $a_n := \mu(\delta_n) + \gamma_n$，它是非递增的，可得对任意 $\eta > 0$:

$$\begin{aligned}
\left\|A_n - H^{-1}\right\|_F^2 &= \mathcal{O}\left(\gamma_n(\ln n)^{1+\eta} + \mu\left(\delta_{\lceil \frac{n}{2} \rceil}\right)^2 + \gamma_{\lceil \frac{n}{2} \rceil}^2\right) \quad \text{a.s.} \\
&= \mathcal{O}\left(\gamma_n(\ln n)^{1+\eta} + \mu\left(\delta_{\lceil \frac{n}{2} \rceil}\right)^2\right) \quad \text{a.s.}
\end{aligned}$$

因为 $\gamma_{\lceil \frac{n}{2} \rceil}^2 = o(\gamma_n)$，由此完成证明。 □

A.2 定理 4.1 的证明

定理 *4.1* 的证明。我们首先证明 (θ_n) 的强相合性。回顾一下，根据分解 (2) 进行归纳可知，对于任意 $n \ge 0$，A_n 都是正定的。利用引理 A.1，我们对任意 $\eta > 0$ 有：

$$\lambda_{\min}(A_n + \nu_n I_d) \ge \nu_n. \quad \text{and} \quad \lambda_{\max}(A_n + \nu_n I_d) = \mathcal{O}(n^{1-\gamma}(\ln n)^{1+\eta}) \quad \text{a.s.}$$

根据 (ν_n) 的定义，若 $\alpha + \gamma > 3/2$，则有

$$\sum_{n\ge 1} \alpha_n \nu_n = +\infty \quad \text{and} \quad \sum_{n\ge 1} \alpha_n n^{1-\gamma}(\ln n)^{1+\eta} < +\infty.$$

因此，我们可以应用 Godichon-Baggioni 和 Werge（2025）的定理 1，得到 (θ_n) 收敛到 θ^* 的强相合性。然后，通过应用定理 B.1，我们得到

(A_n) 几乎必然收敛到 H^{-1}的收敛速率，特别是 (A_n) 的强相合性。因此，如果 $1-\zeta>0$，我们可以应用 Leluc 和 Portier（2023）的定理 2，得到 θ_n 的渐近正态性：

$$\frac{1}{\sqrt{\alpha_n}}(\theta_n-\theta^*)\xrightarrow[n\to+\infty]{\mathcal{L}}\mathcal{N}\left(0,\frac{1}{2(1-\zeta)}H^{-1}\Gamma H^{-1}\right).$$

为了在假设 **(A6)**下获得几乎必然收敛速率，依据 Boyer 和 Godichon-Baggioni（2023），若 $\alpha=1$，则应用定理 3.2；若 $\alpha\in(2/b,1)$，则应用定理 4.2。

□

定理 4.5 的证明。 平均矩阵 A_n 仍满足引理 A.1 的上界，因为 A_n 是 A_j 的凸组合（$j\leq n$），且该上界是递增的。因此，我们可以沿用定理 4.1 证明中的相同推理，得到关于 θ_n 和 A_n的结果。

应用 Toeplitz 引理，θ_n（分别为 A_n）的强相合性意味着 θ_n（$resp.A_n$）的强相合性。为了得到 θ_n的收敛速率和渐近正态性，我们可以应用 Godichon-Baggioni 和 Werge（2025）中的定理 3。他们的假设 5 可由假设 **(A6)** 和 **(A2)**b) 推出。接着，结合假设 **(A6)** 应用定理 B.1，即可得到 A_n 的收敛速率。 □

A.3 对 A_n 的最大特征值的控制

引理 A.1 的证明与 Godichon-Baggioni 等人（2025）中命题 6.1 的证明非常接近。

引理 A.1 的证明。 令 $n\geq 1$。我们有：

$$\mathbb{E}\left[\|A_n\|_F^2\mid\mathcal{F}_{n-1}\right]=\|A_{n-1}\|_F^2+2\langle A_{n-1},\mathbb{E}\left[A_n-A_{n-1}\mid\mathcal{F}_{n-1}\right]\rangle_F+\mathbb{E}\left[\|A_n-A_{n-1}\|_F^2\mid\mathcal{F}_{n-1}\right]. \tag{11}$$

对于 (11) 右侧的最后一项，利用截断和不等式 $(a+b)^2\leq 2a^2+2b^2$，我们有：

$$\begin{aligned}
&\mathbb{E}\left[\|A_n-A_{n-1}\|_F^2\mid\mathcal{F}_{n-1}\right]\\
&=\mathbb{E}\left[\left\|\gamma_n\left(\tilde{H}_nA_{n-1}+A_{n-1}\tilde{H}_n^T-2M_{n,\ell}\right)-\gamma_n^2\tilde{H}_nA_{n-1}\tilde{H}_n^T\right\|_F^2\mathbf{1}_{\gamma_n\|\tilde{H}_n\|_{op}\leq\frac{1}{2}}\mid\mathcal{F}_{n-1}\right]\\
&\leq\gamma_n^2\mathbb{E}\left[\left(2\|\tilde{H}_n\|_F\|A_{n-1}\|_F+2\|M_{n,\ell}\|_F+\gamma_n\|\tilde{H}_n\|_F^2\|A_{n-1}\|_F\right)^2\mathbf{1}_{\|\gamma_n\tilde{H}_n\|_{op}\leq\frac{1}{2}}\mid\mathcal{F}_{n-1}\right]\\
&\leq\gamma_n^2\mathbb{E}\left[\left(2\|\tilde{H}_n\|_F\|A_{n-1}\|_F+2\|M_{n,\ell}\|_F+\frac{1}{2}\|\tilde{H}_n\|_F\|A_{n-1}\|_F\right)^2\mid\mathcal{F}_{n-1}\right]\\
&\leq\gamma_n^2\left(2\times\frac{25}{4}\,\mathbb{E}\left[\|\tilde{H}_n\|_F^2\mid\mathcal{F}_{n-1}\right]\|A_{n-1}\|_F^2+2\times 4\mathbb{E}\left[\|M_{n,\ell}\|_F^2\mid\mathcal{F}_{n-1}\right]\right)\\
&=\mathcal{O}\left(\gamma_n^2\left(\|A_{n-1}\|_F^2+1\right)\right)\ \text{a.s.}
\end{aligned}$$

˜关于 (11) 中的内积项，为了构造 H_{n-1}的期望，我们可以写成

$$
\begin{aligned}
&\mathbb{E}\left[A_n - A_{n-1} \mid \mathcal{F}_{n-1}\right] \\
&= \mathbb{E}\left[\left(-\gamma_n\left(\tilde{H}_n A_{n-1} + A_{n-1}\tilde{H}_n^T - 2M_{n,\ell}\right) + \gamma_n^2 \tilde{H}_n A_{n-1}\tilde{H}_n^T\right)\mathbf{1}_{\gamma_n\|\tilde{H}_n\|_{op}\le\frac{1}{2}} \mid \mathcal{F}_{n-1}\right] \\
&= -\gamma_n\mathbb{E}\left[\tilde{H}_n A_{n-1} + A_{n-1}\tilde{H}_n^T - 2M_{n,\ell} \mid \mathcal{F}_{n-1}\right] + \gamma_n^2\mathbb{E}\left[\tilde{H}_n A_{n-1}\tilde{H}_n^T\mathbf{1}_{\gamma_n\|\tilde{H}_n\|_{op}\le\frac{1}{2}} \mid \mathcal{F}_{n-1}\right] \\
&\quad + \gamma_n\mathbb{E}\left[\left(\tilde{H}_n A_{n-1} + A_{n-1}\tilde{H}_n^T - 2M_{n,\ell}\right)\mathbf{1}_{\gamma_n\|\tilde{H}_n\|_{op}>\frac{1}{2}} \mid \mathcal{F}_{n-1}\right]. \qquad (12)
\end{aligned}
$$

现在，我们对 A_{n-1} 与式（12）中的三个项分别取内积并进行界定。根据假设 **(A1)**，$\tilde{H}_n$ 的条件期望为海森矩阵 $\nabla^2 F(\theta_{n-1})$；由于 F 是凸的，该矩阵为正。于是得到：

$$
\begin{aligned}
&\langle A_{n-1}, \nabla^2 F(\theta_{n-1})A_{n-1}\rangle_F = \mathrm{Tr}\left(A_{n-1}^T\nabla^2 F(\theta_{n-1})A_{n-1}\right) \ge 0 \\
&\langle A_{n-1}, A_{n-1}\nabla^2 F(\theta_{n-1})\rangle_F = \mathrm{Tr}\left(A_{n-1}^T A_{n-1}\nabla^2 F(\theta_{n-1})\right) = \mathrm{Tr}\left(A_{n-1}\nabla^2 F(\theta_{n-1})A_{n-1}^T\right) \ge 0.
\end{aligned}
$$

因此，对于任意 $\zeta_n > 0$ （我们稍后将对其进行选择，以平衡各个分量），A_{n-1} 与式（12）第一项的内积满足以下界：

$$
\begin{aligned}
\langle A_{n-1}, -\gamma_n\mathbb{E}\left[\tilde{H}_n A_{n-1} + A_{n-1}\tilde{H}_n^T - 2M_{n,\ell} \mid \mathcal{F}_{n-1}\right]\rangle_F &\le 2\gamma_n\langle A_{n-1}, \mathbb{E}\left[M_{n,\ell} \mid \mathcal{F}_{n-1}\right]\rangle_F \\
&\le \frac{\gamma_n}{\zeta_n}\|A_{n-1}\|_F^2 + \frac{\ell^2}{d^2}\gamma_n\zeta_n.
\end{aligned}
$$

对于 A_{n-1} 与式（12）第二项的内积，我们得到

$$
\langle A_{n-1}, \gamma_n^2\mathbb{E}\left[\tilde{H}_n A_{n-1}\tilde{H}_n^T\mathbf{1}_{\gamma_n\|\tilde{H}_n\|_{op}\le\frac{1}{2}} \mid \mathcal{F}_{n-1}\right]\rangle_F \le \gamma_n^2\|A_{n-1}\|^2\mathbb{E}\left[\|\tilde{H}_n\|_F^2 \mid \mathcal{F}_{n-1}\right].
$$

对于 A_{n-1} 与式（12）第三项的内积，我们得到：

$$
\begin{aligned}
&\langle A_{n-1}, \gamma_n\mathbb{E}\left[\left(\tilde{H}_n A_{n-1} + A_{n-1}\tilde{H}_n^T - 2M_{n,\ell}\right)\mathbf{1}_{\gamma_n\|\tilde{H}_n\|_{op}>\frac{1}{2}} \mid \mathcal{F}_{n-1}\right]\rangle_F \\
&\le 2\gamma_n\|A_{n-1}\|_F^2\mathbb{E}\left[\|\tilde{H}_n\|_F\mathbf{1}_{\gamma_n\|\tilde{H}_n\|_{op}>\frac{1}{2}} \mid \mathcal{F}_{n-1}\right] + 2\gamma_n|\langle A_{n-1}, \mathbb{E}\left[M_{n,\ell} \mid \mathcal{F}_{n-1}\right]\rangle_F| \\
&\le 2\gamma_n\|A_{n-1}\|_F^2\mathbb{E}\left[2\gamma_n\|\tilde{H}_n\|_F^2 \mid \mathcal{F}_{n-1}\right] + \frac{\gamma_n}{\zeta_n}\|A_{n-1}\|_F^2 + \gamma_n\zeta_n d \\
&= \mathcal{O}\left(\left(\gamma_n^2 + \frac{\gamma_n}{\zeta_n}\right)\|A_{n-1}\|_F^2 + \gamma_n\zeta_n\right) \text{ a.s.}
\end{aligned}
$$

最后，我们得到

$$
\mathbb{E}\left[\|A_n\|_F^2 \mid \mathcal{F}_{n-1}\right] \le (1 + E_n)\|A_{n-1}\|_F^2 + B_n,
$$

其中 $E_n = \mathcal{O}(\gamma_n^2 + \gamma_n/\zeta_n)$ 和 $B_n = \mathcal{O}(\gamma_n^2 + \gamma_n\zeta_n)$。现在我们可以设定 $\zeta_n := n^{1-\gamma}(\ln n)^{1+\eta}$，其中 $\eta > 0$，从而使 $(\frac{\gamma_n}{\zeta_n})$ 可求和。将引理 B.4 应用于 $a_n := \zeta_n^2$ 和 $D_n := 0$，可得：

$$
\|A_{n-1}\|_F^2 = o(\zeta_{n-1}^2) \quad \text{a.s.}
$$

□

B 辅助结果

定理 B.1。 设 $H \in \mathcal{M}_d(\mathbb{R})$ 为一个正定矩阵，且 $z^* \in \mathbb{R}^d$。考虑线性 *Robbins-Monro* 算法

$$z_n = z_{n-1} - \gamma_n \left(h(z_{n-1}) + \epsilon_n + r_n \right),$$

其中 $h(z) = H(z - z^*)$、(ϵ_n) 和(r_n) 是适应于滤过$(\mathcal{F}_n)$的随机序列，z_0是一个 $\mathcal{F}_0$-可测的 $\mathbb{R}^d$值随机变量，而 $\gamma_n = c/(n^\gamma + c')$满足 $\gamma \in \left(\frac{1}{2}, 1\right)$、$c > 0$ 和 $c' \geq 0$。假设存在一个指数 $p \geq 0$，使得 $\|z_n\| = \mathcal{O}(n^p)$ 几乎必然成立。进一步假设对所有 $n \geq 1$,

$$\begin{aligned} \mathbb{E}\left[\epsilon_n \mid \mathcal{F}_{n-1}\right] &= 0 \ \ a.s., \\ \mathbb{E}\left[\|\epsilon_n\|^b \mid \mathcal{F}_{n-1}\right] &= \mathcal{O}\left((1 + \|z_{n-1}\|)^b \right) \ \ a.s., \\ \|r_n\| &= \mathcal{O}\left(a_n \left(1 + \|z_{n-1}\|\right) \right) \ \ a.s., \end{aligned}$$

对于某个 $b > -\frac{2}{\gamma}$，以及一个几乎必然单调不增并收敛至0的随机序列(a_n)。那么，对于任意 $\eta > 0$,

$$\|z_n - z^*\|^2 = \mathcal{O}\left(\gamma_n (\ln n)^{1+\eta} + a_{\lceil n/2 \rceil}^2 \right) \ \ a.s.$$

证明。 对所有 $k \leq n$，定义 $\beta_{k,n} := \prod_{j=k}^{n} (I_d - \gamma_j H)$，并对 $k > n$ 定义 $\beta_{k,n} := I_d$。通过归纳法，我们得到

$$z_n - z^* = \underbrace{\beta_{1,n}(z_0 - z^*)}_{Z_n} + \underbrace{\sum_{k=1}^{n} \beta_{k+1,n} \gamma_k \epsilon_k}_{E_n} + \underbrace{\sum_{k=1}^{n} \beta_{k+1,n} \gamma_k r_k}_{R_n}. \tag{13}$$

选取 $n_0 \in \mathbb{N}$，使得对所有 $n \geq n_0$， $\lambda_{\max}(H)\gamma_n \leq 1$。由于 H 为正，对于 $n \geq n_0$,

$$\|I_d - \gamma_n H\|_{op} \leq \max\{1 - \lambda_{\min}(H)\gamma_n, 1 - \lambda_{\max}(H)\gamma_n\} = 1 - \lambda_{\min}(H)\gamma_n.$$

界定 Z_n。 项 Z_n 趋于 0 的速度呈指数级。事实上，对于 $n \geq n_0$,

$$\|\beta_{1,n}\|_{op} \leq \|\beta_{1,n_0-1}\|_{op} \prod_{j=n_0}^{n} (1 - \lambda_{\min}(H)\gamma_j) \leq \|\beta_{1,n_0-1}\|_{op} \exp\left(-\lambda_{\min}(H) \sum_{j=n_0}^{n} \gamma_j \right),$$

并且 $\sum_{j=n_0}^{n} \gamma_j \sim cn^{1-\gamma}/(1-\gamma)$。

界定 E_n。 令 $p \geq 0$ 满足 $\|z_{n-1}\| = \mathcal{O}(n^p)$ 几乎必然成立。由于 $b > \frac{2}{\gamma}$，应用 Cénac 等 (2025) 的定理 6.1 和备注 6.1，取 $T_n := (1 + \|z_{n-1}\|) I_d$ 和$\xi_n := \epsilon_n/(1 + \|z_{n-1}\|)$。于是，对于任意 $\eta > 0$, $\|T_n\|_{op} = o\left(n^p (\ln n)^{\eta/2}\right)$ 几乎必然成立。尽管该定理假设指数 a 非负，b 且满足 $\|T_n\|_{op} = o\left((\ln n)^b / n^a\right)$ 几乎必然成立，但用相同的证明可将其推广到负值，因此我们取 $a := -p$。由此得到

$$\|E_n\|^2 = \mathcal{O}\left(n^{2p} \gamma_n (\ln n)^{1+\eta} \right) \ \text{a.s.}$$

界定 R_n。 对于 $n \geq n_0$，我们有：

$$\|R_n\| \leq \|I_d - \gamma_n H\|_{op} \|R_{n-1}\| + \gamma_n \|r_n\| \leq (1 - \lambda_{\min}(H)\gamma_n) \|R_{n-1}\| + \gamma_n \|r_n\|.$$

利用方程 (13) 以及关于 (r_n) 的假设，我们得到

$$\|r_n\| = \mathcal{O}\left(a_n \left(1 + \|Z_{n-1}\| + \|E_{n-1}\| + \|R_{n-1}\|\right)\right) \text{ a.s.}$$

我们已证明 $1 + \|Z_n\| + \|E_n\| = \mathcal{O}\left(\max\left\{1, n^{p-\gamma/2} (\ln n)^{(1+\eta)/2}\right\}\right)$ 几乎必然成立。利用关于 (a_n) 的假设，我们可以几乎处处应用引理 B.2 的确定性结果。因此，存在某个 $\kappa > 0$，使得：

$$\|R_n\| = \mathcal{O}\left(\left(\exp(-\kappa n^{1-\gamma}) + a_{\lceil n/2 \rceil}\right) \max\left\{1, n^{p-\gamma/2} (\ln n)^{(1+\eta)/2}\right\}\right) \text{ a.s.}$$

利用 (13)，忽略指数项并利用 $a_n = o$，我们得到 (1) 几乎必然成立 .:

$$\|z_n - z^*\| = \mathcal{O}\left(a_{\lceil n/2 \rceil} + n^{p-\gamma/2} (\ln n)^{(1+\eta)/2}\right) \text{ a.s.} \tag{14}$$

令 $k \in \mathbb{N}$ 为满足 $\|z_{n-1}\| = \mathcal{O}(n^{k\gamma/4})$ 几乎必然成立的最小整数。将 $p := k\gamma/4$ 代入方程 (14)，再利用 $(\ln n)^{(1+\eta)/2} = o(n^{k\gamma/4})$ 以及 $a_n = o(1)$ 几乎必然成立，我们得到

$$\|z_n\| \leq \|z_n - z^*\| + \|z^*\| = \mathcal{O}\left(a_{\lceil n/2 \rceil} + n^{(k-2)\gamma/4} (\ln n)^{(1+\eta)/2} + 1\right) \subset \mathcal{O}\left(n^{(k-1)\gamma/4} + 1\right) \text{ a.s.}$$

若 $k > 0$，则有 $\|z_n\| = \mathcal{O}(n^{(k-1)\gamma/4})$ 几乎必然成立，这与k的定义矛盾。因此 $k = 0$，将 $p = 0$ 代入 (14) 即可得到所需的收敛速率。 □

引理 B.2。 设 $(r_n)_{n \geq 0}$、$(s_n)_{n \geq 1}$ 和$(a_n)_{n \geq 1}$ 是满足以下条件的三个非负数列$n \geq n_0$

$$r_n \leq (1 - \gamma_n) r_{n-1} + \gamma_n a_n (s_n + r_n),$$

其中 $\gamma_n := c_\gamma n^{-\gamma}$，且 $\gamma \in (1/2, 1]$ 和 $c_\gamma > 0$。假设(a_n) 单调不增并收敛于0，且 $s_n = \mathcal{O}(v_n)$ 满足(v_n) 为正且单调不减。那么，对于任意 $0 < \kappa < c_\gamma(1 - 2^{\gamma-1})/(1 - \gamma)$，有

$$r_n = \mathcal{O}\left(\left(\exp(-\kappa n^{1-\gamma}) + a_{\lceil n/2 \rceil}\right) v_n\right).$$

证明。 取 $n_1 \geq n_0$ 使得 $a_{n_1} \leq 1/2$，并取 $M \geq 0$ 使得对任意 $n \in \mathbb{N}$，$s_n \leq M v_n$。对于任意 $n \geq n_1$：

$$r_n \leq (1 - \gamma_n) r_{n-1} + M\gamma_n a_n v_n + \frac{1}{2}\gamma_n r_n \leq \left(1 - \frac{1}{2}\gamma_n\right) r_{n-1} + M\gamma_n a_n v_n.$$

由于 (v_n) 为正且单调不减，因此对于 $n \geq n_1$：

$$\frac{r_n}{v_n} \leq \left(1 - \frac{1}{2}\gamma_n\right) \frac{r_{n-1}}{v_n} + M\gamma_n a_n \leq \left(1 - \frac{1}{2}\gamma_n\right) \frac{r_{n-1}}{v_{n-1}} + \frac{1}{2}\gamma_n (2M a_n).$$

我们可以将命题 B.3 应用于 (r_n/v_n)，从而对任意 $\kappa < c_\gamma(1 - 2^{\gamma-1})(1 - \gamma)^{-1}$，得到：

$$\frac{r_n}{v_n} = \mathcal{O}\left(\exp\left(-\kappa n^{1-\gamma}\right) + a_{\lceil n/2 \rceil}\right).$$

□

命题 B.3. 设$(r_n)^{n\geq 0}$、$(\gamma_n)^{n\geq 1}$和$(a_n)^{n\geq 1}$ 是三个非负数列，并满足 $n \geq n_0$:

$$r_n \leq (1-\gamma_n)\, r_{n-1} + \gamma_n a_n. \tag{15}$$

假设(a_n) 是非递增的。那么，对于所有的 n、m，若有 $n_0 < m \leq n$，我们得到上界:

$$r_n \leq \exp\left(-\sum_{k=m}^{n} \gamma_k\right)\left(r_{n_0} + \sum_{k=n_0+1}^{m-1} \gamma_k a_k\right) + a_m.$$

特别地，如果 $\gamma_n = c/(n^\gamma + c')$ 满足 $c > 0$、$\gamma \in (0,\ 1)$、$c' \geq 0$，并且 $\lim_{n\to\infty} a_n = 0$，那么对于任意满足 $0 < \kappa < c(1-2^{\gamma-1})(1-\gamma)^{-1}$ 的 κ，我们有:

$$r_n = \mathcal{O}\left(\exp(-\kappa n^{1-\gamma}) + a_{\lceil \frac{n}{2} \rceil}\right)$$

证明。 对 n 使用归纳法，并在任意秩 $n_0 < m \leq n$ 处分拆总和，可得 s

$$r_n \leq \prod_{j=n_0+1}^{n} (1-\gamma_j)\, r_{n_0} + \sum_{k=n_0+1}^{m-1} \left(\prod_{j=k+1}^{n} (1-\gamma_j)\right) \gamma_k a_k + \sum_{k=m}^{n} \left(\prod_{j=k+1}^{n} (1-\gamma_j)\right) \gamma_k a_k$$

利用 $1 - x \leq e^{-x}$、$0 \leq 1-\gamma_j \leq 1$ 以及 (a_n) 是非递增的这一事实，我们得到

$$r_n \leq \exp\left(-\sum_{j=n_0+1}^{n} \gamma_j\right) r_{n_0} + \exp\left(-\sum_{j=m}^{n} \gamma_j\right) \sum_{k=n_0+1}^{m-1} \gamma_k a_k + \underbrace{a_m \sum_{k=m}^{n} \left(\prod_{j=k+1}^{n} (1-\gamma_j)\right) \gamma_k}_{=:S}$$

通过识别一个伸缩和，我们得到:

$$S = a_m \sum_{k=m}^{n} \left(\prod_{j=k+1}^{n} (1-\gamma_j) - \prod_{j=k}^{n} (1-\gamma_j)\right) = a_m \left(1 - \prod_{j=m}^{n} (1-\gamma_j)\right) \leq a_m,$$

这就证明了第一个结论。

对于这一特例，假设 $\gamma_n = c/(n^\gamma + c')$ 且 $c > 0$、$c' \geq 0$、$\gamma \in (0,\ 1)$，并且 $\lim_{n\to\infty} a_n = 0$。那么，$\sum_{k=m}^{n} \gamma_k \sim c(1-\gamma)^{-1}(n^{1-\gamma} - m^{1-\gamma})$，取 $m = \lceil n/2 \rceil$ 后，该和以 $\kappa n^{1-\gamma}$ 增长，其中 $\kappa := c(1-\gamma)^{-1}(1-2^{\gamma-1})$ 为正。指数项 $\exp\left(-\sum_{k=m}^{n} \gamma_k\right)$ 以 $\exp(-\kappa n^{1-\gamma} + \text{const})$ 的形式衰减。于是界变为

$$r_n = \mathcal{O}\left(\exp\left(-\kappa n^{1-\gamma}\right)\left(1 + n^{1-\gamma}\right) + a_{\lceil n/2 \rceil}\right) \subset \mathcal{O}\left(\exp(-\kappa' n^{1-\gamma}) + a_{\lceil n/2 \rceil}\right)$$

对于 $\kappa' < \kappa$，证明完成。 □

下面的引理 B.4 是 Robbins-Siegmund 定理的一个简单推论，已在 Godichon-Baggioni 等人（2025）中陈述并证明。

引理 B.4. 设 (V_n)、(B_n)、(E_n)、(D_n) 和 (a_n) 为适应于$(\mathcal{F}_n)_{n\in\mathbb{N}}$ 的五个正序列，使得

$$\mathbb{E}\left[V_n \mid \mathcal{F}_{n-1}\right] \leq (1+E_n)V_{n-1} + B_n - D_n \quad a.s.$$

还假设 $\sum_{n=0}^{\infty} \frac{B_n}{a_n} < +\infty$ 几乎处成立。若 $a_n \to \infty$，则 $V_n = o\left(a_n\right)$ 几乎处处成立。

C 流式处理情形的假设

本节旨在给出推论 5.1 的假设，这些假设直接针对函数 f 陈述。更准确地说，在 $F(\theta) = \mathbb{E}_\xi[f(\xi, \theta)]$ 的情况下，并考虑独立同分布或小批量设置时，只要函数 f 满足相应假设，预言机的假设便成立：

(B1) （f的可微性）。对于所有 $\theta \in \mathbb{R}^d$，随机函数 $\theta \mapsto f(\xi, \theta)$在 θ处几乎必然二阶可微。

(B2) （增长条件）。存在 $\mathcal{L}_g$、$\mathcal{L}_h$、$\sigma^2 \geq 0$，使得对于所有 $\theta \in \mathbb{R}^d$：

a) $\mathbb{E}\left[\|\nabla f(\xi, \theta)\|^2\right] \leq 2\mathcal{L}_g\left(F(\theta) - F(\theta^*)\right) + \sigma^2$

b) $\mathbb{E}\left[\left\|\nabla^2 f(\xi, \theta)\right\|_{op}^2\right] \leq \mathcal{L}_h.$

(B3) （极小化点处的海森矩阵）。极小化点处的海森矩阵 $H := \nabla^2 F(\theta^*)$是正定的，且映射 $\theta \mapsto \nabla^2 F(\theta)$在 θ^*处连续。

(B4) （李雅普诺夫条件）。存在 q、$q' > 2$ 以及 M、$M' > 0$，使得：

a) $\sup_{\|\theta - \theta^*\| < M} \mathbb{E}\left[\|\nabla f(\xi, \theta) - \nabla F(\theta)\|^q\right] < +\infty$

b) $\sup_{\|\theta - \theta^*\| < M'} \mathbb{E}\left[\left\|\nabla^2 f(\xi, \theta) - \nabla^2 F(\theta)\right\|_F^{q'}\right] < +\infty.$

(B5) （协方差的连续性）。协方差函数 $\theta \mapsto \mathsf{Cov}\left(\nabla f(\xi, \theta)\right)$在 θ^*处连续，并且我们定义 $\Sigma := \mathsf{Cov}\left(\nabla f(\xi, \theta^*)\right)$；当 $\nabla F(\theta^*) = 0$时，它等价于 $\Sigma = \mathbb{E}\left[\nabla f(\xi, \theta^*) \nabla f(\xi, \theta^*)^T\right]$。

D 矩阵梯度

希尔伯特空间中的梯度与伴随

设 $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ 为一个希尔伯特空间，$\|\cdot\|$ 为由内积诱导的范数，$U \subset \mathcal{H}$ 为一个开集，且 $f\colon U \to \mathbb{R}$ 为在点 $a \in U$ 处可微的函数。微分 $df(a)$ 按定义是从 $\mathcal{H}$ 到 $\mathbb{R}$ 的连续线性算子，里斯表示定理保证存在唯一元素 $\nabla f(a) \in \mathcal{H}$，使得

$$\forall h \in \mathcal{H}, \quad df(a)(h) = \langle \nabla f(a), h \rangle.$$

元素 $\nabla f(a)$ 称为 f 在 a 处的梯度。

设 $L : \mathcal{H} \to \mathcal{H}$ 为连续线性算子。对任意 $y \in \mathcal{H}$，映射$h \mapsto \langle y, Lh \rangle$ 是 $\mathcal{H}$ 上的有界线性泛函，因此由里斯表示定理，存在唯一元素 $L^* y \in \mathcal{H}$，使得

$$\forall h \in \mathcal{H}, \quad \langle y, Lh \rangle = \langle L^* y, h \rangle.$$

算子 $L^* : \mathcal{H} \to \mathcal{H}$ 称为 L 的伴随算子。

二次泛函的梯度

设 $L:\ \mathcal{H}\to\mathcal{H}$ 为连续线性算子，且 $b\in\mathcal{H}$。考虑二次泛函 $f\colon\mathcal{H}\to\mathbb{R}$，其定义为

$$f(x)\coloneqq\|Lx+b\|^2.$$

令 $x,\ h\in\mathcal{H}$。我们有

$$\begin{aligned}f(x+h)&=\|L(x+h)+b\|^2\\&=\|Lx+b\|^2+2\langle Lx+b,Lh\rangle+\|Lh\|^2\\&=f(x)+2\langle L^*(Lx+b),h\rangle+\|Lh\|^2.\end{aligned}$$

f 在 x 处的梯度因此为

$$\nabla f(x)=2L^*(Lx+b).$$

应用于矩阵空间

现在将前面的结果应用于希尔伯特空间 $\mathcal{H}=\mathcal{M}_d(\mathbb{R})$，即赋予 Frobenius 内积的实$d\times d$ 矩阵空间：

$$\langle M,N\rangle_F\coloneqq\operatorname{tr}(M^TN).$$

空间 $\mathcal{H}$ 是有限维的，因此 $\mathcal{H}$ 上的所有线性算子都是连续的。令$A,\ B\in\mathcal{H}$。线性算子 $M\mapsto AM$的伴随为 $N\mapsto A^TN$。事实上，对于任意$M,\ N\in\mathcal{H}$,

$$\langle N,AM\rangle_F=\operatorname{tr}(N^TAM)=\operatorname{tr}((A^TN)^TM)=\langle A^TN,M\rangle_F.$$

考虑二次泛函 $f\colon\mathcal{H}\to\mathbb{R}$，其定义为

$$f(M)\coloneqq\|AM+B\|_F^2.$$

f 在点 $M\in\mathcal{H}$ 处的梯度为

$$\nabla f(M)=2A^T(AM+B).$$

复合的梯度与转置

设 $f\colon\mathcal{H}\to\mathbb{R}$ 是一个可微函数，并定义 $g\colon\mathcal{H}\to\mathbb{R}$ 为

$$g(M)\coloneqq f(M^T).$$

记 $t:\ \mathcal{H}\to\mathcal{H}$ 为矩阵的转置算子，则有 $g=f\circ t$。转置算子是线性且自伴的，因为对于任意 $M,\ N\in\mathcal{H}$,

$$\langle N,t(M)\rangle_F=\operatorname{tr}(N^TM^T)=\operatorname{tr}((N^T)^TM)=\langle t(N),M\rangle_F.$$

利用微分的链式法则，我们得到 g 在点 $M \in \mathcal{H}$处的微分为

$$dg(M) = df(M^T) \circ t.$$

因此， g 在点 $M \in \mathcal{H}$ 处的梯度为

$$\nabla g(M) = t^*(\nabla f(M^T)) = t(\nabla f(M^T)) = (\nabla f(M^T))^T.$$