arXiv:2303.01370v1 [math.OC] 1 Mar 2023

自适应随机梯度算法的非渐近分析及其应用

Antoine Godichon-Baggioni * Pierre Tarrago *

摘要

在随机优化中，处理大规模样本的一种常见方法是依次使用广为人知的随机梯度算法。然而，由于每个方向采用相同的步长序列，当问题条件不佳时，实际应用中可能会导致较差的结果。为解决这一问题，应优先采用 Adagrad 或随机牛顿算法等自适应梯度算法。本文旨在对用于强凸目标函数的这类自适应梯度算法进行非渐近分析。所有理论结果都将分别应用于 Adagrad 和随机牛顿算法下的线性回归与正则化广义线性模型。

关键词：非渐近分析；在线估计；自适应梯度算法；Adagrad；随机牛顿算法。

1 引言

随机优化中的一个常见问题是估计凸泛函G 的极小化点 θ： $\mathbb{R}^d \longrightarrow \mathbb{R}$ 其形式为

$$G(h) = \mathbb{E}\left[g(X,h)\right]$$

其中 g： $\mathcal{X} \times \mathbb{R}^d \longrightarrow \mathbb{R}$，且 X 是一个取值于 $\mathcal{X}$ 的随机变量。事实上，常见的回归问题（如线性回归和逻辑回归）（Bach，2014），以及几何中位数和分位数的估计（Cardot 等人，2013，2015；Godichon-Baggioni，2016）都符合这一情况。已有多种技术用于估计该问题的解，可分为两大类：迭代方法和递归方法。迭代方法是利用样本生成经验函数，并借助常见的凸优化方法逼近其极小化点（Boyd 和 Vandenberghe，2004），或采用小批量算法等改进方法（Koneˇcnˇy 等人，2015）。虽然这些方法以性能优异著称，但在处理大样本时可能会遇到计算问题。此外，它们不适用于处理按顺序到达的数据，因此可以转而关注递归方法。

*索邦大学概率、统计与建模实验室，法国 antoine.godichon_baggioni@sorbonne-universite.fr, pierre.tarrago@sorbonne-universite.fr 本项目由法国国家研究署资助，项目名称为 CORTIPOM，编号为 ANR-21-CE40-0019。

最著名且研究最广泛的递归方法之一，无疑是随机梯度算法（Robbins 和 Monro，1951）及其平均版本（Ruppert，1988；Polyak 和 Juditsky，1992）。对于按顺序到达的数据 $X_1, \ldots, X_n$，X_{n+1}，...，对所有 $n \geq 0$，其递归定义如下：

$$\theta_{n+1} = \theta_n - \gamma_{n+1} \nabla_h g\left(X_{n+1}, \theta_n\right), \qquad \overline{\theta}_{n+1} = \overline{\theta}_n + \frac{1}{n+2}\left(\theta_{n+1} - \overline{\theta}_n\right)$$

其中 (γ_n) 是一个趋于 0 的正步长序列。这些估计值已被研究一段时间：关于一些渐近结果，可参见 (Pelletier，1998，2000)；关于非渐近结果的较新文献则包括估计值的均方收敛 (Bach and Moulines，2013；Gadat 和 Panloup，2017；Gower 等人，2019；Godichon-Baggioni，2021)。众所周知，在某些正则性假设下，平均估计值具有渐近有效性，并达到 Cramér–Rao 界（忽略余项）。

然而，步长序列 (γ_n) 无法针对梯度的各个方向进行调整，这在实际应用于病态问题时可能导致不理想的结果。为缓解这一问题，可以更多地关注以下形式的自适应随机梯度算法：

$$\theta_{n+1} = \theta_n - \gamma_{n+1} A_n \nabla_h g\left(X_{n+1}, \theta_n\right)$$

其中 (A_n) 是一个（随机）矩阵序列，能够针对各个坐标进行调整。最著名的自适应算法之一是 Adagrad (Duchi 等人，2011)，它可以被视为一种对梯度 $\nabla_h g\left(X_{n+1}, \theta\right)$进行标准化的方法。在近期的研究中，Bercu 等人(2020) 和 Boyer 和 Godichon-Baggioni (2020) 将 (A_n) 视为海森矩阵逆矩阵的估计值序列，从而得到随机牛顿算法。当我们希望最小化的函数的海森矩阵具有不同量级的特征值时，例如，这种方法尤其值得关注。

需要指出的是，关于自适应方法已有若干渐近结果，其中值得关注的近期研究包括Leluc 和 Portier(2020)以及Gadat 和 Gavra (2020)等，而非渐近结果则较为少见。尽管如此，在一项近期研究中，De Vilmarest 和 Wintenberger(2021)针对逻辑回归中特殊情形下的卡尔曼递推，给出了高概率界；而Défossez 等人 (2020)则研究了 Adagrad 和 Adam 的 L^2 收敛速率。此外，Bercu 等人 (2021)得到了最优传输中随机高斯–牛顿算法的均方收敛速率。不过请注意，在上述所有情形中，都假设g 的梯度一致有界。

本文关注强凸函数（因此梯度无界）的非渐近收敛速率。更具体地说，我们针对序列 A_n 可能发散、但其发散程度受到控制的情形，提出了自适应估计的首个收敛速率。此外，假设 A_n 具有一致有界的四阶矩，我们证明 $\mathbb{E}\left[G\left(\theta_n\right) - G(\theta)\right]$ 以通常的收敛速率收敛。最后，我们建立了一个不受约束的一般框架，用于获得随机牛顿法和 Adagrad 算法的收敛速率。这些结果将

应用于线性回归和岭广义线性模型。

本文安排如下：第 2 节介绍一般框架。第 3 节给出算法及收敛的理论结果，而第 4 节和第 5 节分别介绍线性回归和广义线性模型的应用。证明将在第 6 节和附录中给出。

2 框架

下面，我们考虑一个取值于可测空间$\mathcal{X}$ 的随机变量X ，并固定$d \geq 2$。我们重点研究凸函数G 的极小化点 θ 的估计： $\mathbb{R}^d \longrightarrow \mathbb{R}$ ，该函数对所有$h \in \mathbb{R}^d$ 定义为

$$G(h) := \mathbb{E}\left[g\left(X, h\right)\right]$$

其中g ： $\mathcal{X} \times \mathbb{R}^d \longrightarrow \mathbb{R}$。从现在起，我们假设以下条件成立：

（A1）对于几乎所有$x \in \mathcal{X}$，泛函$g(x, .)$ 在 $\mathbb{R}^d$ 上可微，并且存在$p \geq 2$ 以及非负常数$C_1^{(p)}$ 、$C_2^{(p)}$ ，使得对于所有$h \in \mathbb{R}^d$，

$$\mathbb{E}\left[\left\|\nabla_h g\left(X, h\right)\right\|^{2p}\right] \leq C_1^{(p)} + C_2^{(p)} \left\|h - \theta\right\|^{2p}.$$

（A2）泛函G 二阶连续可微。

（A3）G 的海森矩阵在 $\mathbb{R}^d$ 上一致有界，即存在正常数 $L_{\nabla G}$，使得对所有 $h \in \mathbb{R}^d$，

$$\left\|\nabla^2 G(h)\right\|_{op} \leq L_{\nabla G}$$

其中 $\|.\|_{op}$ 是矩阵通常使用的谱范数。

（A4）泛函 G 是 μ 拟强凸的：对所有 $h \in \mathbb{R}^d$，

$$\left\langle \nabla G\left(h\right), h - \theta \right\rangle \geq \mu \left\|h - \theta\right\|^2.$$

请注意，在一个特例中，假设（**A3**） 可确保 G 的梯度是 $L_{\nabla G}$- Lipschitz。请注意，这些假设通常用于获得随机梯度算法及其平均版本的 L^2 收敛速率（Bach and Moulines，2013；Gower 等人，2019）。

3 自适应随机梯度算法

3.1 算法

设 X_1、...、X_n、X_{n+1}、... 为 X 的独立同分布副本。则自适应随机梯度算法递归定义如下，对所有 $n \geq 0$

$$\theta_{n+1} = \theta_n - \gamma_{n+1} A_n \nabla_h g\left(X_{n+1}, \theta_n\right),$$

其中，θ_0 为任意选取的，$\gamma_n = c\gamma n^{-\gamma}$，且满足 $c_\gamma > 0$，$\gamma \in (0,1)$；A_n 是一个对称正定矩阵序列，使得存在一个滤过$(\mathcal{F}_n)_{n\geq 0}$ 满足：

- 对所有$n \geq 0$，A_n 均为 $\mathcal{F}_n$-可测的。
- X_{n+1} 与 $\mathcal{F}_n$相互独立。

通常，可以考虑 A_n 仅依赖于 $\mathrm{X}_1, \ldots, \mathrm{X}_n, \theta_0, \ldots, \theta_n$，并考虑由样本生成的滤过，即 $\mathcal{F}_n = \sigma(\mathrm{X}_1, \ldots, \mathrm{X}_n)$。若考虑对角矩阵 A_n，其中 $(\mathrm{A}_n)_{k,k} = \left(\mathrm{n}+1\left(\mathrm{a}_k + \sum_{i=1}^n \nabla_h \mathrm{g}\left(\mathrm{X}_i, \theta_{i-1}\right)^2 \mathrm{i}, \mathrm{i}\right)\right)^{-1/2}$，则得到 Adagrad 算法（Duchi 等人，2011）。此外，当 A_n 是海森矩阵逆矩阵的递归估计时，对应于随机牛顿算法（Bercu 等人，2020；Boyer 和 Godichon-Baggioni，2020）；而当 $\mathrm{A}_n = \mathrm{n}+1\left(\mathrm{A}_0 + \sum_{i=1}^n \nabla_h \mathrm{g}\left(\mathrm{X}_i, \theta_{i-1}\right) \nabla_h \mathrm{g}\left(\mathrm{X}_i, \theta_{i-1}\right)^T\right)$ 时，对应于随机高斯-牛顿算法（Cénac 等人，2020；Bercu 等人，2021）。

3.2 收敛结果

3.2.1 首个收敛结果

为了获得估计值的初步收敛速率，现在我们对随机矩阵序列$(A_n)_{n\geq 0}$引入一些假设：

(H1) 可以对A_n的最小和最大特征值进行控制：

(H1a) 存在$(v_n)_{n\geq 0}, \lambda_0 > 0$ 和 $\delta, q \geq 0$，使得

$$\mathbb{P}\left[\lambda_{\min}\left(A_n\right) \leq \lambda_0 t\right] \leq v_{n+1} t^q (n+1)^{-\delta},$$

对于 $0 < t \leq 1$，其中$(v_{n+1}(n+1)^{-\delta})_{n\geq 0}$ 递减。
如果 $\gamma \leq 1/2$，还需假设存在更强的假设，即存在 $\lambda'_n = \lambda'_0 (n+1)^{-\lambda'}$，其中 $\lambda'_0 > 0$、$\lambda' < \gamma$ 满足对于所有 $n \geq 0$,

$$\lambda_{\min}(A_n) \geq \lambda'_n.$$

(H1b) 存在 $\beta_n = c_\beta n^\beta$，满足 $c_\beta \geq 0$，并且如果 $\gamma \leq 1/2$ 或 $0 < \beta < \gamma - 1/2$（如果 $\gamma > 1/2$），则 $0 < \beta < \frac{\gamma}{2}$；对于所有 $n \geq 0$,

$$\|A_n\|_{op} \leq \beta_{n+1}.$$

请注意，情形 $\delta = 0$ 在 **(H1a)** 中是允许的，并且始终可以在 β 的允许范围内选择 **(H1b)**。在大多数情况下，尤其是对于 Adagrad 和随机牛顿算法，**(H1a)** 很容易得到验证。项 v_n 在 **(H1a)** 中的存在，考虑了一个普遍现象（通常由 Rosenthal 不等式所蕴含）：X的高阶矩误差贡献虽然在 n较小时占主导地位，但会随着 n 趋于无穷而逐渐消失。关于 **(H1b)**，附录 F中给出了一些反例，表明当这一假设不满足时，估计值可能发散，这意味着该假设很遗憾是至关重要的。无论如何，一种容易验证它的方法是将随机矩阵 A_n 替换为

$$\tilde{A}_n = \frac{\min\left\{\|A_n\|_{op}, \beta_{n+1}\right\}}{\|A_n\|_{op}} A_n$$

并且可以直接验证 $\left\| \tilde{A}_n \right\|_{op} \leq \beta_{n+1}$。在 $\gamma \leq 1/2$ 的情况下，也可以通过类似调整确保 **(H1a)** 。

我们来考虑牛顿法的情形，尤其是海森矩阵的估计值具有以下形式的情形：$H_n = {}^{1}{}_{n+1}\left(H_0 + \sum_{k=1}^{n} a_k \Phi_k \Phi_k^T\right)$ 并且可以借助 Riccati/Shermann-Morrisson 公式递归求逆（参见 Bercu 等人(2020）；Boyer 和 Godichon-Baggioni (2020）；Godichon-Baggioni 等人 (2022））。为验证**(H1b)**，可以考虑以下海森矩阵估计值的形式

$$\tilde{H}_n = H_n + \frac{1}{n+1} \sum_{k=1}^{n} \frac{\tilde{c}_\beta}{k^\beta} e_k e_k^T$$

其中e_k是第k个（模d意义下）标准基向量（参见Bercu 等人(2021)；Godichon-Baggioni 等人(2022)）。现在我们可以得到估计值的一个初步收敛速率。为简便起见，现将风险误差记为$V_n := G(\theta_n) - G(\theta)$。注意，由于$G$ 是 μ拟强凸的，因此有$\|\theta_n - \theta\|^2 \leq \frac{2}{\mu} V_n$。

定理 3.1。假设 （***A1***） 至 （***A3***） 以及 （***H1***） 成立。那么，对所有 $n \geq 1$以及任意$\lambda < \min\{\gamma - 2\beta, 1 - \gamma\}$,

$$\mathbb{E}[V_n] \leq \exp\left(-c_\gamma \mu \lambda_0 n^{1-(\lambda+\gamma)}(1-\varepsilon(n))\right)\left(K_1^{(1)} + K_{1'}^{(1)} \max_{1\leq k\leq n+1} k^{\gamma-2\beta-\delta/2-(q/2+1)\lambda}\sqrt{v_k}\right) + K_2^{(1)} n^{-(\gamma-2\beta-\lambda)} + K_3^{(1)}\sqrt{v_{\lfloor n/2\rfloor}}\, n^{-(\delta+q\lambda)/2},$$

其中 $\varepsilon(n) = o(1)$ 见 （20） 中给出，且 $K_1^{(1)}$、$K_{1'}^{(1)}$、$K_2^{(1)}$、$K_3^{(1)}$ 分别见（21） 和 （22）。

在 $\delta/2 \geq \gamma - 2\beta$ 这一特殊情形下（只要 $\delta \geq 1$就会出现），只需在上述公式中令 $\lambda = 0$ 即可：我们将看到，对于采用随机牛顿算法的广义线性模型，情况正是如此。然而，对于 Adagrad 算法，则无法避免首先使用 $\lambda > 0$，因为A_n 取决于 $\nabla g(X, \cdot)$ 而不是 $\nabla^2 g(X, \cdot)$（后者的期望在 $\mathbb{R}^d$ 上有界，而前者的期望通常无界）。为摆脱这一较弱的结论，我们需要以下关于高阶矩的定理3.1 的等价形式：

高阶矩。

命题 3.1。 假设假设（***A1***）满足 $p > 2$，（***A2***）、（***A3***）和（***H1***）成立。那么对任意 $p' < p$ 及任意 $\lambda <$ 最小值$\{\gamma - 2\beta, 1-\gamma\}$,

$$\mathbb{E}\left[V_n^{p'}\right] \leq \exp\left(-c_\gamma \mu\lambda_0 n^{1-(\lambda+\gamma)}(1-\varepsilon'(n))\right)\left(K_1^{(1')} + K_{1'}^{(1')} \max_{1\leq k\leq n+1} k^{\gamma-2\beta-\lambda-\frac{p-p'}{p}(\delta+q\lambda)} v_k^{\frac{p-p'}{p}}\right)$$
$$+ K_2^{(1')} n^{-p'(\gamma-2\beta-\lambda)} + K_3^{(1')} v_{\lfloor n/2\rfloor}^{\frac{p-p'}{p}} (n+1)^{-\frac{p-p'}{p}(\delta+q\lambda)},$$

其中 $\epsilon'(n)$, $K^{(1')}{}_1$, $K^{(1')}{}_{1'}$, $K^{(1')}{}_2$ *and* $K^{(1')}{}_3$ 分别由(57)、(58) 和(60)给出。

3.2.2 当A_n 具有有界矩时的收敛

为了获得更好的收敛速率，现在我们针对随机矩阵序列(A_n)引入一些新的假设:

（H2a）随机矩阵A_n 的二阶矩一致有界：存在C_S，使得对所有$n \geq 0$:

$$\mathbb{E}\left[\|A_n\|^2\right] \leq C_S^2.$$

(H2b) 随机矩阵A_n 的四阶矩一致有界：存在C_S，使得对所有$n \geq 0$:

$$\mathbb{E}\left[\|A_n\|^4\right] \leq C_S^4.$$

为简化表述，此处及下一段中我们假设$q > 0$ 属于**(H1a)**，尽管在一般情况下也能得到类似的界。

定理 3.2. 对于某个 $p > 2$，若假设（***A1***）至（***A3***）、***(H1)*** 和***(H2a)*** 在 $\delta > 0$ 下成立，则对所有 $n \geq 0$,

$$\mathbb{E}[V_n] \leq \exp\left(-c_\gamma \mu\lambda_0 n^{1-\gamma}(1-\varepsilon(n))\right)\cdot\left(K_1^{(2)} + K_{1'}^{(2)} \max_{1\leq k\leq n+1} v_k^{\frac{p-1}{p}} k^{\gamma-2\beta-\frac{p-1}{p}\delta}\right)$$
$$+ K_2^{(2)} v_{\lfloor n/2\rfloor}^{\frac{p-1}{p}} n^{-\frac{(p-1)}{p}\delta} + K_3^{(2)} n^{-\gamma},$$

其中 $\varepsilon(n) = o(1)$ 由式 (25) 给出，且 $K_1^{(2)}$、$K_{1'}^{(2)}$、$K_2^{(2)}$、$K_3^{(2)}$ 分别由式 (26)、(27) 和(28)给出。

最后，为了得到随机牛顿估计的均方收敛速率，我们现在给出L^2 阶的$G(\theta_n)$ 在 $\gamma > 1/2$ 时的收敛速率。

命题 3.2。 假设对于某个 $p > 2$，（***A1***）至（***A3***）、（***H1***）和***(H2b)*** 成立，且$\gamma > 1/2$, $\delta > 0$ 以及 $\beta < \gamma - 1/2$。那么

$$\mathbb{E}\left[V_n^2\right] \leq \exp\left(-\frac{3}{2}c_\gamma \lambda_0 \mu n^{1-\gamma}\right)\left(K_1^{(2')} + K_{1'}^{(2')} \max_{1\leq k\leq n+1} v_k^{\frac{p-2}{p}} k^{\gamma-\frac{p-2}{p}\delta}\right)$$
$$+ K_2^{(2')} n^{-2\gamma} + K_3^{(2')} v_{\lfloor n/2\rfloor}^{(p-2)/p} n^{-\delta(p-2)/p} =: M_n.$$

其中 $K^{(2')}{}_1$，$K^{(2')}{}_{1'}$，$K^{(2')}{}_2$，$K^{(2')}{}_3$ 分别由 (63)、(64) 和(65)给出。

注意到 $M_n = O\left(n^{-\min\left\{2\gamma, \delta(p-2)/p\right\}}\right)$。因此，当 δ 足够大时（即 $\delta > 2p/(p-2)\gamma$），后一个界中的第二项是主要贡献。于是，对于任意仅依赖于 v_n 和 γ 的 $0 \le \gamma' \le \min\left\{2\gamma, \delta(p-2)/p\right\}$，我们有

$$w_\infty(\gamma') := \sup_{n\ge 1} M_n n^{\gamma'} < +\infty. \tag{1}$$

函数 $w_\infty: \left[0, \min\left\{2\gamma, \delta(p-2)/p\right\}\right] \to \mathbb{R}$ 可以通过数值方法计算，但无论如何请注意，$w_\infty(\gamma') \le K^{(2')}{}_1 \sup_{t\ge 1}\left\{t^{\gamma'} \exp\left(-\frac{1}{2}\lambda_0 \mu t^{1-\gamma}\right)\right\} + K^{(2')}{}_2 + K^{(2')}{}_3$，因此通过函数分析可得，对于 $\gamma' \in \left[0, \min\left\{2\gamma, \delta(p-2)/p\right\}\right]$，

$$w_\infty(\gamma') \le K_1^{(2')}\left(\frac{2\gamma'}{\lambda_0 \mu e(1-\gamma)}\right)^{\frac{\gamma'}{1-\gamma}} + K_2^{(2')} + K_3^{(2')}. \tag{2}$$

我们将在大多数应用中看到，在适当的假设下，γ' 可以等于 2γ（即当 $\delta \ge 2p/(p-2)\gamma$时）。

3.2.3 随机牛顿算法的收敛结果

现在我们来关注随机牛顿算法的收敛速率。为此，记 $H := \nabla^2 G(\theta)$，并从现在起假设以下条件也成立：

（A1'） 存在 $L_{\nabla g}$，使得对所有 $h \in \mathbb{R}^d$，

$$\mathbb{E}\left[\left\|\nabla_h g(X,h) - \nabla_h g(X,\theta)\right\|^2\right] \le L_{\nabla g}\|h-\theta\|^2 \tag{3}$$

（A5） 存在一个非负常数 L_δ，使得对所有 $h \in \mathbb{R}^d$，

$$\left\|\nabla G(h) - \nabla^2 G(\theta)(\theta - h)\right\| \le L_\delta \|h-\theta\|^2$$

(H3) 估计值 A_n 收敛到 H^{-1}：存在一个递减的正数列 $(v_{A,n})$，使得对所有 $n \ge 0$，

$$\mathbb{E}\left[\left\|A_n - H^{-1}\right\|^2\right] \le v_{A,n}.$$

请注意，假设 **（A1'）** 在文献中通常称为期望光滑性（Bach and Moulines，2013），并且在线性回归和逻辑回归（Bach and Moulines，2013；Bach，2014）或几何分位数和中位数的估计（Cardot 等人，2013）等大多数例子中都得到满足。关于 **（A5）**，在 **（A3）** 下，只要海森矩阵在 θ的某个邻域内满足 Lipschitz 条件，该条件便得到满足。例如，在线性回归的情形下，$L_\delta = 0$。最后，假设 **(H3)** 若估计值具有如下的一阶收敛速率，便得到满足：

θ（例如，根据定理3.2 或命题 3.2，这将导致 A_n具有一阶收敛速率，而这在实践中通常可以得到验证（例如参见 Boyer 和 Godichon-Baggioni （2020））。

定理 3.3。假设（***A1***）至（***A5***）以及***(H1)*** 至***(H3)*** 成立，且满足 $\gamma > 1/2$， $\delta > 0$和 $\beta < \gamma - 1/2$。那么，

$$\mathbb{E}\left[\|\theta_n - \theta\|^2\right] \leq e^{-\frac{1}{2}c_\gamma n^{1-\gamma}}\left(K_1^{(3)} + K_{1'}^{(3)} \max_{0\leq k\leq n}(k+1)^\gamma d_k\right) + n^{-\gamma}\left(2^{3+\gamma}c_\gamma Tr\left(H^{-1}\Sigma H^{-1}\right) + \frac{K_2^{(3)}}{n^\gamma} + K_{2'}^{(3)} v_{A,n/2}\right) + d_{\lfloor n/2\rfloor}.$$

其中 $K^{(3)}{}_i$， $i = 1$， $1'$， 2， $2'$ 定义见(29)、(30) 和(31)，而 d_k 仅依赖于 M_k 和 $v_{A,\,k}$见(30)。

由（30）可知，对于某个常数$C > 0$，$d_k \leq C(v_{A,k} + M_k)$。若在(H1a)中还假设指数 δ 足够大，则上述结果可以进一步简化。

推论 3.1。假设（***A1***）至（***A4***）以及***(H1)*** 至***(H3)*** 成立，且满足 $\gamma > 1/2$， $\delta > {}^{2\gamma}p\, p-2$和 $\beta < \gamma - 1/2$。那么，

$$\mathbb{E}\left[\|\theta_n - \theta\|^2\right] \leq n^{-\gamma}\left(2^{3+\gamma}c_\gamma Tr\left(H^{-1}\Sigma H^{-1}\right) + \frac{K_2^{(3')}}{n^\gamma} + K_{2'}^{(3')} v_{A,n/2} + K_{2''}^{(3')}\sqrt{v_{A,n/2}}\right) + K_1^{(3')} e^{-\frac{1}{2}c_\gamma n^{1-\gamma}},$$

其中 $K^{(3')}{}_i$， $i = 1...2''$ 见 （32）和（33）。

那么，如果$v_{A,n}$ 收敛到 0，我们就得到通常的收敛速率 $^1{}_{n^\gamma}$。

3.2.4 自适应梯度（Adagrad）的收敛结果

回顾一下，Adagrad 算法要求指定d 个初始参数$a_1, \ldots, a_d \in \mathbb{R}_+$，并选择对角矩阵$A_n$，其中

$$(\overline{A_n})_{kk'} = \delta_{kk'}\frac{1}{\sqrt{\frac{1}{n+1}\left(a_k + \sum_{i=0}^{n-1} n\left(\nabla_h g(X_{i+1}, \theta_i)_k\right)^2\right)}}. \tag{4}$$

那么，原始的 Adagrad 算法相当于取 $\gamma = 1/2$。为保证矩阵（A_n）$_{n\geq 0}$非退化，我们假设极小化点处的梯度存在一定的最小波动 θ。

（A6） 存在 $\alpha > 0$，使得对于所有 $1 \leq i \leq d$,

$$\mathbb{E}\left[(\nabla_h g(X, \theta))_i^2\right] > \alpha. \tag{5}$$

(A6′) 存在 $\alpha > 0$，使得对于所有 $h \in \mathbb{R}^d$ 以及 $1 \le i \le d$,

$$\mathbb{E}\left[(\nabla_h g(X,h))_i^2\right] > \alpha. \tag{6}$$

注意，（**A6'**）比（**A6**）强得多。不过，前者通常是成立的，例如带噪声的线性回归情形。无论如何，可以考虑对A_n作如下变换：

$$(A_n)_{kk'} = \begin{cases} \min\left\{c_\beta n^\beta, (\overline{A_n})_{kk'}\right\}, & \text{if } \gamma > 1/2 \\ \max\left\{\min\left\{c_\beta n^\beta, (\overline{A_n})_{kk'}\right\}, \lambda_0' n^{-\lambda'}\right\}, & \text{if } \gamma \le 1/2 \end{cases} \tag{7}$$

其中 $\beta_n = c^\beta n^\beta$，取 $\beta < \min\{\gamma/2, 1/4\}$ (λ', λ_0'，并任意选取$c_\beta > 0$)

定理 3.4. 假设（***A1***）至（***A4***）以及（***A6***）成立，且 $\beta < \min\left\{\frac{(1-\gamma)\gamma(\gamma-2\beta)p}{4\ 2-\gamma}(), 1/4\right\}$。那么，

$$\mathbb{E}\left[\|\theta_n - \theta\|^2\right] \le \tilde{K}_1^{(4)} \exp\left(-c_\gamma \mu \tilde{\lambda}_0 n^{1-\gamma}(1-\tilde{\varepsilon}(n))\right) + \tilde{K}_2^{(4)} \log(n+1)^{\frac{p-1}{p}} n^{-\frac{(p-1)}{p}\min\left\{\frac{2(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma}, 1\right\}} + \tilde{K}_3^{(4)} n^{-\gamma},$$

其中 *(n)* 由 *(35)* 给出，$v_n = v_0 \log(n+1)$，其中 v_0, C_S^4，而 $\tilde{\lambda}_0$ 由 *(73)*、*(74)* 和 *(72)* 给出，且 $p' = \frac{2(1-\gamma)}{2-\gamma} p$。此外，$K_1^{(4)}$、$K_2^{(4)}$ 和 $K_3^{(4)}$ 由 *(36)* 给出。若 *(A6')* 成立，则对 $\beta < 1/4$ 也有相同结论，其中 C_S 由 *(75)* 给出，取 $p' = \frac{2(1-\gamma)}{2-\gamma} p$。

在 $\gamma = 1/2$ 这一特殊情况下，即对应于通常的 Adagrad 算法，我们得到

$$\mathbb{E}\left[\|\theta_n - \theta\|^2\right] \le K_1^{(4)} \exp\left(-c_\gamma \mu \lambda_0 \sqrt{n}(1-\varepsilon(n))\right) + \frac{1}{\sqrt{n}}\left(K_2^{(4)} \log(n+1) n^{1/2 - \frac{(1-4\beta)(p-1)}{6}} + K_3^{(4)}\right),$$

因此，只要 $1/2 - \frac{(1-4\beta)(p-1)}{6} < 0$，即可达到通常的收敛速率 $\frac{1}{\sqrt{n}}$；也就是说，只要 $p > 4\frac{1-\beta}{1-4\beta}$。

4 线性模型的应用

现在考虑线性模型 $Y = X^T\theta + \epsilon$，其中 $X \in \mathbb{R}^d$，且 ϵ 是一个与 X 独立的中心化随机实变量。下面假设 $\mathbb{E}\left[XX^T\right]$ 为正。于是，θ 是泛函 $G\colon \mathbb{R}^d \longrightarrow \mathbb{R}$ 的唯一极小化点，该泛函对所有 $h \in \mathbb{R}^d$ 定义为

$$G(h) = \frac{1}{2}\mathbb{E}\left[\left(Y - X^T h\right)^2\right].$$

如果X具有二阶矩，则函数G是二阶连续可微的，其中$\nabla G(h) = -\mathbb{E}\left[\left(Y - X^T h\right) X\right]$，并且 $\nabla^2 G(h) = \mathbb{E}\left[XX^T\right]$。

4.1 随机牛顿法

随机牛顿算法对所有n ≥ 0均递归定义为（Boyer 和 Godichon-Baggioni，2020）

$$\theta_{n+1} = \theta_n + \gamma_{n+1}\overline{S}_n^{-1}\left(Y_{n+1} - X_{n+1}^T\theta_n\right)X_{n+1}$$

其中 $\tilde{S}_n = \frac{1}{n+1}\left(S_0 + \sum_{i=1}^n X_i X_i^T\right)$，其中 S_0 为正常数，且

$$\overline{S}_n^{-1} = \frac{\min\left(\left\|\tilde{S}_n^{-1}\right\|_{op}, \beta_{n+1}\right)}{\left\|\tilde{S}_n^{-1}\right\|_{op}}\tilde{S}_n^{-1}$$

其中 $\beta_n = c_\beta n^{-\beta}$。注意到，$\tilde{S}^{-1}_{n+1}$ 可以利用 Sherman-Morrison（或 Riccati）公式，仅通过 $O\left(d^2\right)$ 次运算轻松更新。更具体地，考虑 $S_n = (n+1)\tilde{S}_n$，可得

$$S_{n+1}^{-1} = S_n^{-1} - \left(1 + X_{n+1}^T S_n^{-1} X_n\right)^{-1} S_n^{-1} X_{n+1} X_{n+1}^T S_n^{-1}.$$

接下来，可以轻松更新 S_n 和 $\overline{S_n}$。现在，我们可以将定理3.3 改写如下：

定理 4.1. 假设存在 $p > 4$，使得 X, ϵ 的 $2p$ 阶矩和 p 阶矩存在。另假设存在正常数 L_{MK}，使得对于任意 $h \in \mathbb{S}^{d-1}$，$\sqrt{\mathbb{E}\left[hXX^Th\right]} \leq L_{MK}\mathbb{E}\left[\left|X^Th\right|\right]$。那么，对于 $\gamma > 1/2$，有

$$\mathbb{E}\left[\|\theta_n - \theta\|^2\right] \leq e^{-\frac{1}{2}c_\gamma n^{1-\gamma}}\left(K_{1,lin}^{(3)} + K_{1',lin}^{(3)}\max_{0\leq k\leq n} d_k(k+1)^\gamma\right)$$
$$+ n^{-\gamma}\left(2^{3+\gamma}c_\gamma\mathbb{E}\left[\epsilon^2\right]Tr\left(H^{-1}\right) + \frac{K_{2,lin}^{(3)}}{n^\gamma} + K_{2',lin}^{(3)}v_{H,n/2}\right) + d_{\lfloor n/2\rfloor}.$$

其中 $K_{2,lin}, K_{2'}^{(3)}$，线性，$K_1^{(3)}$，线性，$K_{1'}^{(3)}$，线性，d_n 由下式给出 (40) 而 $v_{H,n}$ 在 (39) 中定义。

注意，$d_n = O\left(\frac{1}{n^{\max\left\{\frac{p-2}{2}, 2\gamma\right\}}}\right)$ 且 $v_{H,n} = O\left(n^{-1}\right)$，并且由于 $p > 4$，这些项都可以忽略不计。

4.2 Adagrad 算法

对于线性模型，Adagrad 算法对所有 $n \geq 0$ 定义如下

$$\theta_{n+1} = \theta_n + \gamma_{n+1}\bar{D}_n^{-1}\left(Y_{n+1} - X_{n+1}^T\theta_n\right)X_{n+1},$$

其中D_n 是对角矩阵，且对于 $\gamma \leq 1/2$,

$$(\bar{D}_n)_{kk} = \min\left\{\max\left\{\frac{n^{-\beta}}{c_\beta}, \sqrt{\frac{1}{n+1}\left(a_k + \sum_{i=0}^{n-1}\left(\left(Y_{i+1} - X_{i+1}^T\theta_i\right)(X_{i+1})_k\right)^2\right)}\right\}, \frac{n^{\lambda'}}{\lambda_0'}\right\}.$$

其中 $0 < \beta < (\gamma - \lambda')/2$ 对某个 $a_k > 0$ 成立，并且若 $\gamma > 1/2$,

$$(\bar{D}_n)_{kk} = \max\left\{\frac{n^{-\beta}}{c_\beta}, \sqrt{\frac{1}{n+1}\left(a_k + \sum_{i=0}^{n-1}\left(\left(Y_{i+1} - X_{i+1}^T\theta_i\right)(X_{i+1})_k\right)^2\right)}\right\},$$

对于某个 $0 < \beta < \gamma - 1/2$。通常的 Adagrad 算法取 $\gamma = 1/2$，由此我们得到

$$(\theta_{n+1})_k = (\theta_n)_k + \frac{\left(Y_{n+1} - X_{n+1}^T\theta_n\right)(X_{n+1})_k}{\min\left\{\max\left\{\frac{n^{-\beta+1/2}}{c_\beta}, \sqrt{a_k + \sum_{i=0}^{n-1}\left(\left(Y_{i+1} - X_{i+1}^T\theta_i\right)(X_{i+1})_k\right)^2}\right\}, \frac{n^{\lambda'+1/2}}{\lambda_0'}\right\}},$$

并且几乎必然存在 $n_0 \geq n$，使得对于 $n \geq n_0$,

$$(\theta_{n+1})_k = (\theta_n)_k + \frac{\left(Y_{n+1} - X_{n+1}^T\theta_n\right)(X_{n+1})_k}{\sqrt{a_k + \sum_{i=0}^{n-1}\left(\left(Y_{i+1} - X_{i+1}^T\theta_i\right)(X_{i+1})_k\right)^2}},$$

这就是通常的 Adagrad 算法。于是，我们可以将定理3.4 改写如下（这里仅给出 $\gamma = 1/2$ 的情形，读者很容易将其推广到 $\gamma > 1/2$ 的情形）。

定理 4.2. 假设存在 $p > 2$，使得 X, ϵ 具有 $2p$ 阶矩。那么，对于$\gamma \leq 1/2$，我们有

$$\begin{aligned}\mathbb{E}\left[\|\theta_n - \theta\|^2\right] \leq\ & K_{1,lin}^{ada}\exp\left(-c_\gamma\lambda_{\min}\lambda_{0,lin}^{ada}n^{1-\gamma}\left(1 - \varepsilon_{n,lin}^{ada}\right)\right) \\ & + K_{2,lin}^{ada}\log(n+1)^{\frac{p-1}{p}}n^{-\frac{(p-1)}{p}\min\left\{\frac{2(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma},1\right\}} + K_{3,lin}^{ada}n^{-\gamma},\end{aligned}$$

其中 $\varepsilon_{n,lin}^{ada} = o(1)$ 由 (41) 给出，$K_{1,lin}^{ada}, K_{2,lin}^{ada}K_{3,lin}^{ada}$ 由 (42)、(43) 和 (44)给出。

请注意，类似的结论对 $\gamma > 1/2$ 也成立。注意，在 $\gamma = 1/2$ 的情况下，只要 $(p-1)(1-4\beta)/3 \geq 1/2$，即 $p > \frac{5-4\beta}{2(1-4\beta)}$，就能达到 $\frac{1}{\sqrt{n}}$ 收敛速率。

5 应用于广义线性模型

线性回归的框架可以轻松推广到更一般的有限维线性模型。设 $\ell: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ 为某个定义域$\mathcal{Y} \subset \mathbb{R}$上的代价函数。一般的学习问题是求解以下最小化问题

$$\operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}\left[\ell(Y, f(X))\right],$$

其中 $(X, Y) \sim \mathbb{P}$ 且 $\mathcal{F}$是从 $\mathcal{X}$到 $\mathcal{Y}$的可测函数类，其中 $\mathcal{X}$是一个可测空间。对于有限维线性模型，$\mathcal{Y} = \mathbb{R}$ 和 $\mathcal{F} = \{h^T\Phi(\cdot), h \in \mathbb{R}^m\}$，其中 $\Phi: \mathcal{X} \to \mathbb{R}^m$ 是一个已知的设计函数（请注意，该设定可以轻松推广至 $\mathcal{Y} = \mathbb{R}^p$ 和 $\Phi: \mathcal{X} \to \mathbb{R}^m$ 及$h \in M_{m,p}(\mathbb{R})$的情形）。于是，

假设 ℓ 是凸的，并在 θ上添加一个正则化项后，该最小化问题便转化为本文的框架，其中

$$G(h) = \mathbb{E}\left[g\left(\tilde{Z}, h\right)\right],$$

其中 $\tilde{Z} = (Y, \Phi(X)) := (Y, \tilde{X})$，且对于所有$h \in M_m$，$p(\mathbb{R})$，$g(\tilde{Z}, h) = \ell(Y, h^T \tilde{X})$。下文中，我们从现在起假设代价函数$l$ 关于第二个变量二阶可微，并且存在正常数 $L_{\nabla l}$，使得对于所有h

$$\left|\nabla_h^2 \ell\left(Y, h^T \tilde{X}\right)\right| \leq L_{\nabla l}. \tag{8}$$

其中 $\nabla_h^2 \ell(.,.)$ 是关于第二个变量的二阶导数。备注：通常在要求对于所有 h，$\|\nabla^2 G(h)\|_{op} \leq L_{\nabla G} < \infty$时，都会假设这样的界。例如，当 $\ell(y, y') = f(y - y')$ 且 $\sup_y |f''(y)| < +\infty$ 时，这一条件就成立。举例来说，在最简单的逻辑回归情形下，我们考虑一对取值于 $\mathbb{R}^d \times \{-1, 1\}$，$\Phi = I_d$的随机变量 (X, Y)，并令 $\ell(y, y') = \log(1 + \exp(-yy'))$，此时对于所有 h 和 Y，确实有 $\in \{-1, 1\}$

$$\nabla_h^2 \ell(Y, h^T X) = \frac{1}{1 + \exp(h^T X)} \cdot \frac{1}{1 + \exp(-h^T X)} \leq 1.$$

处理最小化问题的凸性主要有两种方法：要么假设其具有强凸性，要么使用正则化。第一种方法是假设泛函 $h \longmapsto \mathbb{E}\left[\ell\left(Y, h^T X\right)\right]$ 是强凸的；特别地，当存在 $\alpha > 0$ 使得

$$\inf_{y' \in \mathbb{R}} \nabla_h^2 \ell(y, y') > \alpha. \tag{9}$$

且 $\mathbb{E}\left[XX^T\right]$ 为正时，这一情形在下文称为椭圆情形，其结果与线性回归的结果非常相似，因此不再赘述。接下来我们将重点讨论正则化情形。若没有 $\nabla_h^2 \ell(y,y')$ 的一致下界，就需要加入正则化项，从而得到以下正则化最小化问题

$$\operatorname{argmin}_{\theta \in \mathbb{R}^m} \mathbb{E}\left[\ell(Y, \langle \theta, \theta^T X \rangle)\right] + \frac{\sigma}{2}\|\theta\|^2 \tag{10}$$

其中 $\sigma > 0$ 为某个值。下文假设极小化点存在，并记为θ_σ。

5.1 随机牛顿法

随机牛顿算法对所有$n \geq 0$ 递归定义如下

$$\theta_{n+1} = \theta_n - \gamma_{n+1} \overline{S}_n^{-1} \left(\nabla_h l\left(Y_{n+1}, \theta_n^T X_{n+1}\right) X_{n+1} + \sigma \theta_n\right),$$

其中，利用Bercu 等人(2021) 中引入并在Godichon-Baggioni 等人(2022) 中发展的技巧，S_n 是海森矩阵的自然递归估计量，其定义如下

$$\overline{S}_n = \frac{1}{n+1}\sum_{i=0}^{n-1}\nabla_h^2\ell(Y_{i+1},\langle\theta_i,X_{i+1}\rangle)X_{i+1}X_{i+1}^T + \frac{\sigma d}{n+1}\sum_{i=1}^{n}e_{i[d]+1}e_{i[d]+1}^T, \tag{11}$$

其中，$i[d]$ 表示i 模d。注意，可以通过两次使用黎卡提公式轻松更新逆矩阵，即考虑$S_n = (n+1)\overline{S}_n$ 和

$$S_{n+\frac{1}{2}}^{-1} = S_n^{-1} - \nabla_h^2\ell(Y_{n+1},\langle\theta_n,X_{n+1}\rangle)\left(1+\nabla_h^2\ell(Y_{n+1},\langle\theta_n,X_{n+1}\rangle)X_{n+1}^TS_n^{-1}X_{n+1}\right)^{-1}S_n^{-1}X_{n+1}X_{n+1}^TS_n^{-1}$$
$$S_{n+1} = S_{n+\frac{1}{2}}^{-1} - \sigma d\left(1+\sigma d e_{(n+1)[d]+1}^TS_{n+\frac{1}{2}}^{-1}e_{(n+1)[d]+1}\right)^{-1}S_{n+\frac{1}{2}}^{-1}e_{(n+1)[d]+1}e_{(n+1)[d]+1}^TS_{n+\frac{1}{2}}^{-1},$$

可得$S^{-1}{}_{n+1} = (n+2)S^{-1}{}_{n+1}$。接下来，假设以下条件成立：

（GLM1）存在$L_{\nabla^2 L}\geq 0$，使得函数$h\longmapsto\mathbb{E}\left[\nabla_h^2\ell\left(Y,\ h^TX\right)XX^T\right]$关于谱范数是$L_{\nabla^2 L}$-Lipschitz。

（GLM2）存在$p>2$，使得X 具有$2p$ 阶矩，且存在正常数L_σ，满足对所有 $0\leq a\leq 2p$

$$\mathbb{E}\left[\left\|\nabla_h\ell\left(Y,X^T\theta_\sigma\right)X+\sigma\theta_\sigma\right\|^a\right]\leq L_\sigma^a.$$

注意，当对所有y，$\nabla_h^2\ell(y,.)$ 均为Lipschitz函数，且X具有三阶矩时，假设**（GLM1）**成立；例如，逻辑回归很容易满足这些条件。随机变量 $\nabla_h\ell\left(Y,\ X^T\theta_\sigma\right)X$ 具有$2p$阶矩时，假设**（GLM2）**成立。

定理 5.1。假设 ***（GLM1）*** 和 ***（GLM2）*** 成立。那么，

$$\mathbb{E}\left[\|\theta_n-\theta_\sigma\|^2\right]\leq e^{-\frac{1}{2}c_\gamma n^{1-\gamma}}\left(K_{1,GLM}^{(3)}+K_{1',GLM}^{(3)}\max_{0\leq k\leq n}(k+1)^\gamma d_{k,GLM}\right)$$
$$+n^{-\gamma}\left(2^{3+\gamma}c_\gamma Tr\left(H_\sigma^{-1}\Sigma_\sigma H_\sigma^{-1}\right)+\frac{K_{2,GLM}^{(3)}}{n^\gamma}+K_{2',GLM}^{(3)}v_{l,n/2}\right)+d_{\lfloor n/2\rfloor,GLM},$$

其中 $H_\sigma=\mathbb{E}\left[\nabla_h^2\ell\left(Y,X^T\theta_\sigma\right)XX^T\right]+\sigma I_d$, $\Sigma_\sigma=\mathbb{E}\left[\left(\nabla_h\ell\left(Y,X^T\theta_\sigma\right)X+\sigma\theta_\sigma\right)\left(\nabla_h\ell\left(Y,X^T\theta_\sigma\right)X+\sigma\theta_\sigma\right)^T\right]$, $K_1^{(3)},GLM$, $K_{1'}^{(3)},GLM$, $K_2^{(3)},GLM$, $K_{2'}^{(3)},GLM$, d^n,GLM 定义见公式（*50*）、（*51*）和（*52*），而 $v_{l,n}$ 定义见命题 *6.5*。

5.2 Adagrad 算法

对于广义线性模型，Adagrad 算法对所有 $n\geq 0$ 均定义为

$$\theta_{n+1}=\theta_n-\gamma_{n+1}\bar{D}_n^{-1}\nabla_h l\left(Y_{n+1},\theta_n^TX_{n+1}\right)X_{n+1},$$

其中 $\bar{D}_n$ 为对角矩阵，且对于 $\gamma > 1/2$,

$$(\bar{D}_n)_{kk} = \max\left\{\frac{n^{-\beta}}{c_\beta}, \sqrt{\frac{1}{n+1}\left(a_k + \sum_{i=0}^{n-1}\left(\nabla_h l\left(Y_{i+1}, \theta_i^T X_{i+1}\right)(X_{i+1})_k + \sigma(\theta_i)_k\right)^2\right)}\right\},$$

对于某个 $0 < \beta < \gamma - 1/2$，以及对于 $\gamma \leq 1/2$,

$$(\bar{D}_n)_{kk} = \min\left\{\max\left\{\frac{n^{-\beta}}{c_\beta}, \sqrt{\frac{1}{n+1}\left(a_k + \sum_{i=0}^{n-1}\left(\nabla_h l\left(Y_{i+1}, \theta_i^T X_{i+1}\right)(X_{i+1})_k + \sigma(\theta_i)_k\right)^2\right)}\right\}, \frac{n^{\lambda'}}{\lambda_0'}\right\}.$$

其中 $0 < \beta < (\gamma - \lambda')/2$ 且 $a_k > 0$。通常的 Adagrad 算法取 $\gamma = 1/2$，这对我们得到

$$(\theta_{n+1})_k = (\theta_n)_k + \frac{\nabla_h l\left(Y_{n+1}, \theta_n^T X_{n+1}\right)(X_{n+1})_k + \sigma(\theta_n)_k}{\min\left\{\max\left\{\frac{n^{-\beta+1/2}}{c_\beta}, \sqrt{a_k + \sum_{i=0}^{n-1}\left(\nabla_h l\left(Y_{i+1}, \theta_i^T X_{i+1}\right)(X_{i+1})_k + \sigma(\theta_i)_k\right)^2}\right\}, \frac{n^{\lambda'+1/2}}{\lambda_0'}\right\}}.$$

与线性回归类似，广义线性模型只需要极少的随机性，就能保证 Adagrad 的预期收敛速率。事实上，在确定性序列 $(X_n, Y_n)_{n\geq 0}$这一极端情况下，如果不巧出现 $\nabla_h \ell\left(Y_{i+1}, \theta_i^T X_{i+1}\right)(X_{i+1})_k$ 消失或保持极小，Adagrad 算法可能会发散。要求在极小化点处，$\nabla_h \ell\left(Y\right.$ 具有最小方差，即对所有 $1 \leq k \leq d$，$\left.\theta_\sigma^T X\right)(X)_k$，即可避免这种行为。

(GLM3) 存在一个正常数 $\alpha_\sigma > 0$，使得对于所有 $1 \leq k \leq d$

$$Var\left[\nabla_h l\left(Y, X^T\theta_\sigma\right)(X)_k\right] > \alpha_\sigma.$$

请注意

$$Var\left[\nabla_h l\left(Y, X^T\theta_\sigma\right)(X)_k\right] = \mathbb{E}\left[\left|\nabla_h l\left(Y, X^T\theta_\sigma\right)(X)_k + \sigma(\theta_\sigma)_k\right|^2\right], \tag{12}$$

因此 （**GLM3**）可以看作是 （**GLM2**）的镜像假设。我们应当强调，当假设 α_σ 时，只要 X 与 Y 之间存在最低限度的随机性，这种 $\nabla_h l\left(Y, \theta_\sigma^T X\right) X_k$ 为确定性的，就会在 Y 与 X之间形成解析关系。主要的计算问题是估计 α_σ的一个具体值。第 E 节给出了一个关于逻辑回归的例子。

当 （**GLM3**）成立时，利用定理 5.2 可以证明，几乎必然存在 $n_0 \geq n$ 使得对于 $n \geq n_0$,

$$(\theta_{n+1})_k = (\theta_n)_k + \frac{\nabla_h l\left(Y_{n+1}, \theta_n^T X_{n+1}\right)(X_{n+1})_k + \sigma(\theta_n)_k}{\sqrt{a_k + \sum_{i=0}^{n-1}\left(\nabla_h l\left(Y_{i+1}, \theta_i^T X_{i+1}\right)(X_{i+1})_k + \sigma(\theta_i)_k\right)^2}},$$

因此，当 n 较大时，我们就恢复了通常的 Adagrad 算法。接下来可以将定理

3.4 改写为 $\gamma \leq 1/2$ 如下（注意，类似的结论也适用于 $\gamma > 1/2$）。

定理 5.2. 假设 （***GLM1***）、（***GLM2***）和（***GLM3***）成立。那么，对于 $\gamma = 1/2$，我们有

$$\mathbb{E}\left[\|\theta_n - \theta_\sigma\|^2\right] \leq K_{1,GLM}^{ada} \exp\left(-c_\gamma \sigma \tilde{\lambda}_{0,GLM} n^{\frac{p}{2(1+p)}}(1-\varepsilon(n)\right)$$
$$K_{2,GLM}^{ada} \log(n+1)^{\frac{p-1}{p}} n^{-\frac{(p-1)}{p}\min\left\{\frac{2(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma},1\right\}} + K_{3,GLM}^{ada} n^{-\gamma},$$

其中 $\varepsilon(n) = o(1)$，$K_1^{ada}{}_{,GLM}$、$K_2^{ada}{}_{,GLM}$ 和 $K_3^{ada}{}_{,GLM}$ 都有显式公式，且取决于模型的参数。

此处我们不列出这些常数的具体值，因为按照先前结果的思路，可以很容易地推导出来。

6 证明

在整个证明过程中，为简化记号，我们将用相同的符号表示 $\|\circ\|$ 的欧几里得范数 $\mathbb{R}^d$ 以及方阵的谱范数。此外，我们将经常使用以下技术性结果，见（Godichon-Baggioni 等人，2021，命题 A.5）。

命题 6.1. 设 $(\gamma_t)_{t\geq 1}$、$(\eta_t)_{t\geq 1}$和$(\nu_t)_{t\geq 1}$ 为正递减序列，并设 $(\delta_t)_{t\geq 0}$，满足以下条件：

- 序列 δt 遵循以下递归关系：

$$\delta_t \leq (1 - 2\omega\gamma_t + \eta_t\gamma_t)\,\delta_{t-1} + \nu_t\gamma_t, \tag{13}$$

  其中 $\delta_0 \geq 0$且 $\omega > 0$。

- 设 γt 和ηt 收敛至0。

- 设 $t_0 = \inf\{t \geq 1 : \quad \eta_t \leq \omega\}$，并假设对所有 $t \geq t_0 + 1$，均有 $\omega\gamma_t \leq 1$.

那么，对所有 $t \in \mathbb{N}$，有以下上界：

$$\delta_t \leq \exp\left(-\omega \sum_{j=t/2}^{t} \gamma_j\right) \exp\left(2\sum_{i=1}^{t} \eta_i\gamma_i\right)\left(\delta_0 + 2\max_{1\leq i\leq t}\frac{\nu_i}{\eta_i}\right) + \frac{1}{\omega}\max_{t/2\leq i\leq t}\nu_i.$$

其中约定：若 $t/2 < t_0$，则 $\sum_{t_0}^{t/2} = 0$。

此外，我们用 C_1, C_1'、C_2, C_2' 表示常数，使得对于所有 $h \in \mathbb{R}^d$，

$$\mathbb{E}\left[\|\nabla_h g(X,h)\|^2\right] \leq C_1 + C_2\|h-\theta\|^2, \qquad \mathbb{E}\left[\|\nabla_h g(X,h)\|^4\right] \leq C_1' + C_2'\|h-\theta\|^4. \tag{14}$$

6.1 定理3.1 的证明

注意，得益于梯度的泰勒展开，记 $V_n = G(\theta_n) - G(\theta)$ 和 $g'_{n+1} = \nabla_h g\left(X_{n+1}, \theta_n\right)$，

$$\begin{aligned} V_{n+1} &\leq V_n - \gamma_{n+1} \nabla G\left(\theta_n\right)^T A_n g'_{n+1} + \frac{L_{\nabla G}}{2} \gamma_{n+1}^2 \left\|A_n\right\|^2 \left\|g'_{n+1}\right\|^2 \\ &\leq V_n - \gamma_{n+1} \nabla G\left(\theta_n\right)^T A_n g'_{n+1} + \frac{L_{\nabla G}}{2} \gamma_{n+1}^2 \beta_{n+1}^2 \left\|g'_{n+1}\right\|^2, \end{aligned} \tag{15}$$

其中，最后一行用到了假设**(H1b)**。接下来，取条件期望，利用假设**(A1)**，并且由于 $\left\|\theta_n - \theta\right\|^2 \leq \frac{2}{\mu} V_n$，

$$\mathbb{E}\left[V_{n+1} | \mathcal{F}_n\right] \leq \left(1 + \frac{C_2 L_{\nabla G}}{\mu} \beta_{n+1}^2 \gamma_{n+1}^2\right) V_n - \gamma_{n+1} \nabla G\left(\theta_n\right)^T A_n \nabla G\left(\theta_n\right) + \frac{C_1 L_{\nabla G}}{2} \gamma_{n+1}^2 \beta_{n+1}^2$$

此外，由于 G 是 μ 强凸的，因此有

$$\begin{aligned} \nabla G\left(\theta_n\right)^T A_n \nabla G\left(\theta_n\right) &\geq \lambda_{\min}\left(A_n\right) \left\|\nabla G\left(\theta_n\right)\right\|^2 \\ &\geq 2\lambda_n \mu V_n \mathbf{1}_{\lambda_{\min}(A_n) \geq \lambda_n} \\ &= 2\lambda_n \mu V_n - \mathbf{1}_{\lambda_{\min}(A_n) < \lambda_n} 2\lambda_n \mu V_n, \end{aligned} \tag{16}$$

其中 $\lambda_n = \lambda_0 (n+1)^{\lambda}$，且 $0 \leq \lambda < 1 - \gamma$。应用柯西-施瓦茨不等式可得

$$\begin{aligned} \mathbb{E}\left[\nabla G\left(\theta_n\right)^T A_n \nabla G\left(\theta_n\right)\right] \geq & 2\lambda_n \mu \mathbb{E}\left[V_n\right] - 2\lambda_n \mu \sqrt{\mathbb{E}[V_n^2]} \sqrt{\mathbb{P}\left[\lambda_{\min}\left(A_n\right) < \lambda_n\right]} \\ \geq & 2\lambda_n \mu \mathbb{E}\left[V_n\right] - 2\lambda_n \mu V \sqrt{\mathbb{P}\left[\lambda_{\min}\left(A_n\right) < \lambda_n\right]}, \end{aligned}$$

其中 $V^2 \geq \sup_{n \geq 0} \mathbb{E}[V_n^2]$，后文将计算其值。然后，假设**(H1a)** 给出 $\mathbb{P}\left[\lambda_{\min}\left(A_n\right) < \lambda_n\right] \leq v_{n+1}(n+1)^{-\delta - q\lambda} := \bar{v}_n$，因此

$$\begin{aligned} \mathbb{E}\left[V_{n+1}\right] \leq & \left(1 - 2\mu\lambda_0 (n+1)^{-\lambda} \gamma_{n+1} + \frac{C_2 L_{\nabla G}}{\mu} \beta_{n+1}^2 \gamma_{n+1}^2\right) \mathbb{E}\left[V_n\right] \\ & + 2\lambda_0 (n+1)^{-\lambda} \mu V \gamma_{n+1} \sqrt{\bar{v}_n} + \frac{C_1 L_{\nabla G}}{2} \gamma_{n+1}^2 \beta_{n+1}^2. \end{aligned}$$

为了应用命题6.1，我们记

$$C_M = \max\left\{\frac{C_2 L_{\nabla G} c_\beta^2 c_\gamma}{\mu}, (\mu\lambda_0)^{\frac{2\gamma - 2\beta}{\gamma + \lambda}} c_\gamma^{\frac{\gamma - 2\beta - \lambda}{\gamma + \lambda}}\right\}, \tag{17}$$

添加最后一个上界，是为了使下文（18）中的各项满足命题6.1的第三个条件。令 $\tilde{\gamma}_n = c\gamma n^{-(\lambda+\gamma)}$，并注意到

$$\mathbb{E}\left[V_{n+1}\right] \leq \left(1-2\mu\lambda_0\tilde{\gamma}_{n+1}+C_M(n+1)^{2\beta+\lambda-\gamma}\tilde{\gamma}_{n+1}\right)\mathbb{E}\left[V_n\right]+2\lambda_0\mu V\sqrt{v_n}\tilde{\gamma}_{n+1} + \frac{C_1L_{\nabla G}}{2}(n+1)^{\lambda}\gamma_{n+1}\beta_{n+1}^2\tilde{\gamma}_{n+1}. \tag{18}$$

于是，由于 $2\gamma-2\beta-1\neq 1$，借助命题 6.1 以及一个用于判定收敛性的积分判别法可得 $\sum_{k=1}^{n}k^{2\beta-2\gamma}\leq 1+\frac{n^{(1+2\beta-2\gamma)^+}}{|2\gamma-2\beta-1|}$ 和 $\sum^{n}{}_{t=\lfloor n/2\rfloor}t^{-\gamma}\geq\frac{1-2^{\gamma-1}}{1-\gamma}n^{1-\gamma}\geq n^{1-\gamma}$ 当$\gamma\in(0,1)$ 时，

$$\mathbb{E}\left[V_n\right]\leq\exp\left(-c_\gamma\mu\lambda_0n^{1-(\lambda+\gamma)}\right)\exp\left(2C_Mc_\gamma\left(1+\frac{n^{(1+2\beta-2\gamma)^+}}{|2\gamma-2\beta-1|}\right)\right)\cdot \left(\mathbb{E}\left[V_0\right]+4\frac{\lambda_0\mu V}{C_M}\max_{1\leq k\leq n}k^{\gamma-2\beta-\lambda}\sqrt{\bar{v}_k}+\frac{C_1L_{\nabla G}c_\gamma c_\beta^2}{C_M}\right)+2V\sqrt{\bar{v}_{n/2}}+\frac{C_1L_{\nabla G}}{2^{1+\lambda}\mu\lambda_0}n^{\lambda}\beta_{n/2}^2\gamma_{n/2}, \tag{19}$$

其中我们回顾一下 $\bar{}v_n=v_{n+1}(n+1)^{-\delta-q\lambda}\geq\mathbb{P}\left[\lambda_{\min}\left(A_n\right)<\lambda_n\right]$。备注：

$$k^{\gamma-2\beta-\lambda}\sqrt{\bar{v}_k}=\sqrt{v_{k+1}}(k+1)^{\gamma-2\beta-\lambda}(k+1)^{-(\delta+q\lambda)/2}=\sqrt{v_{k+1}}(k+1)^{\gamma-2\beta-\delta/2-(q/2+1)\lambda},$$

因此 $\max_{0\leq k\leq n}(k+1)^{\gamma-2\beta-\lambda}\sqrt{\bar{}v_k}=\max_{1\leq k\leq n+1}k^{\gamma-2\beta-\delta/2-(q/2+1)\lambda}\sqrt{v_k}$。由此，我们得到

$$\mathbb{E}\left[V_n\right]\leq\exp\left(-c_\gamma\mu\lambda_0n^{1-(\lambda+\gamma)}\right)\exp\left(2C_Mc_\gamma\left(1+\frac{n^{(1+2\beta-2\gamma)^+}}{|2\gamma-2\beta-1|}\right)\right) \cdot\left(\mathbb{E}\left[V_0\right]+4\frac{\lambda_0\mu V}{C_M}\max_{1\leq k\leq n+1}k^{\gamma-2\beta-\delta/2-(q/2+1)\lambda}\sqrt{v_k}+\frac{C_1L_{\nabla G}c_\gamma c_\beta^2}{C_M}\right) +2^{1+(\delta+q\lambda)/2}V\sqrt{v_{\lfloor n/2\rfloor}}n^{-(\delta+q\lambda)/2}+2^{\gamma-2\beta-\lambda-1}\frac{C_1L_{\nabla G}c_\gamma c_\beta^2}{\mu\lambda_0}n^{2\beta+\lambda-\gamma}$$

其中，V 在引理6.1中定义。因此，只要满足 $\gamma+\lambda+(1+2\beta-2\gamma)^+<1$，由于 $\lambda<\min\{\gamma-2\beta,1-\gamma\}$，该条件成立，我们便有

$$\mathbb{E}\left[V_n\right]\leq\exp\left(-c_\gamma\mu\lambda_0n^{1-(\lambda+\gamma)}(1-\varepsilon'(n)\right)\left(K_1^{(1)}+K_{1'}^{(1)}\max_{1\leq k\leq n+1}k^{\gamma-2\beta-\delta/2-(q/2+1)\lambda}\sqrt{v_k}\right) +K_2^{(1)}n^{-(\gamma-2\beta-\lambda)}+K_3^{(1)}\sqrt{v_{\lfloor n/2\rfloor}}n^{-(\delta+q\lambda)/2},$$

with

$$\varepsilon'(n)=\frac{2C_Mn^{-1+\lambda+\gamma}}{\mu\lambda_0}\left(1+\frac{n^{(1+2\beta-2\gamma)^+}}{|2\gamma-2\beta-1|}\right), \tag{20}$$

$$K_1^{(1)}=\left(\mathbb{E}\left[V_0\right]+\frac{C_1L_{\nabla G}c_\gamma c_\beta^2}{C_M}\right),\quad K_1^{(1')}=4\frac{\lambda_0\mu V}{C_M}, \tag{21}$$

其中 C_M 在（17）中给出，且 V 见引理 6.1 并且

$$K_2^{(1)} = 2^{\gamma-2\beta-\lambda-1}\frac{C_1 L_{\nabla G} c_\gamma c_\beta^2}{\mu\lambda_0},\ K_3^{(1)} = 2^{1+(\delta+q\lambda)/2}V. \tag{22}$$

引理 6.1。 假设对于 $p \geq 2$，假设 ***(A1)*** 和***(H1b)*** 成立。那么，对于所有 $n \geq 0$，如果 $\gamma > 1/2$，则

$$\mathbb{E}\left[V_n^p\right] \leq e^{a_p c_\gamma^2 c_\beta^2 \frac{2\gamma-2\beta}{2\gamma-2\beta-1}} \max\left\{1, \mathbb{E}\left[V_0^2\right]\right\} := V_n^p$$

如果 $\gamma \leq 1/2$，则

$$\mathbb{E}\left[V_n^p\right] \leq \exp\left(-p\mu\lambda_0' c_\gamma \left(1 + \frac{1 + \left(\frac{c_\gamma c_\beta^2 a_p}{p\mu\lambda_0'}\right)^{\frac{1-\gamma-\lambda'}{\gamma-2\beta-\lambda'}}}{1-\gamma-\lambda'}\right) + c_\gamma^2 c_\beta^2 a_p \left(1 + \frac{1 + \left(\frac{c_\gamma c_\beta^2 a_p}{p\mu\lambda_0'}\right)^{\frac{1-2\gamma+2\beta}{\gamma-2\beta-\lambda'}}}{1-2\gamma+2\beta}\right)\right) =: V_p^p$$

其中 a_2 见(67)，且 a^p 由下式给出：(66)，对于 $p > 2$。

该引理的证明见第 B 节。

6.2 定理3.2的证明

注意，由于假设**(H1b)**，有

$$\mathbb{E}\left[\|A_n\|^2 \left\|g'_{n+1}\right\|^2 |\mathcal{F}_n\right] \leq C_1 \|A_n\|^2 + \frac{C_2 L_{\nabla G}}{\mu}\|A_n\|^2 V_n \leq C_1 \|A_n\|^2 + \beta_{n+1}^2 \frac{C_2 L_{\nabla G}}{\mu} V_n.$$

此外，借助假设**(H2a)**，

$$\mathbb{E}\left[\|A_n\|^2 \left\|g'_{n+1}\right\|^2\right] \leq C_1 C_S^2 + \beta_{n+1}^2 \frac{C_2 L_{\nabla G}}{\mu} V_n$$

与定理3.1 的证明中一样，得到

$$\mathbb{E}\left[V_{n+1}\right] \leq \left(1 - 2\mu\lambda_0\gamma_{n+1} + \frac{C_2 L_{\nabla G}}{\mu}\beta_{n+1}^2\gamma_{n+1}^2\right)\mathbb{E}\left[V_n\right] + 2\lambda_0\gamma_{n+1}\mu\mathbb{E}\left[\mathbf{1}_{\lambda_{\min}(A_n)<\lambda_n} V_n\right] + \frac{C_1 L_{\nabla G} C_S^2}{2}\gamma_{n+1}^2.$$

使用赫尔德不等式和 p，可得

$$\mathbb{E}\left[\mathbf{1}_{\lambda_{\min}(A_n)<\lambda_n} V_n\right] \leq \left(\mathbb{P}\left[\mathbf{1}_{\lambda_{\min}(A_n)<\lambda_n}\right]\right)^{\frac{p-1}{p}} \mathbb{E}\left[V_n^p\right]^{1/p} \leq \bar{v}_n^{\frac{p-1}{p}} V_p$$

其中 $\bar{v}_n = v_{n+1}(n+1)^{-\delta}$ 以及 V_p 见引理 6.1。考虑由以下公式定义的 C_M：

$$C_M = \max\left\{\frac{C_2 L_{\nabla G} c_\beta^2 c_\gamma}{\mu}, (\mu\lambda_0)^{\frac{2\gamma-2\beta}{\gamma}} c_\gamma^{\frac{\gamma-2\beta}{\gamma}}\right\}, \tag{23}$$

有

$$\mathbb{E}[V_{n+1}] \leq \left(1 - 2\mu\lambda_0\gamma_{n+1} + C_M(n+1)^{2\beta-\gamma}\gamma_{n+1}\right)\mathbb{E}[V_n] + 2\lambda_0\mu V_p \bar{v}_n^{\frac{p-1}{p}}\gamma_{n+1} + \frac{C_1 L_{\nabla G} C_S^2}{2}\gamma_{n+1}^2.$$

然后，应用命题6.1 并借助收敛积分判别法，可得

$$\begin{aligned}\mathbb{E}[V_n] \leq \exp\left(-c_\gamma\mu\lambda_0 n^{1-\gamma}\right)\exp\left(2C_M c_\gamma\left(1 + \frac{n^{(1+2\beta-2\gamma)^+}}{|2\gamma-2\beta-1|}\right)\right)\cdot \\ \left(\mathbb{E}[V_0] + 4\frac{\lambda_0\mu V_p \max_{1\leq k\leq n} k^{\gamma-2\beta}\bar{v}_k^{\frac{p-1}{p}}}{C_M} + \frac{C_1 L_{\nabla G} c_\gamma C_S^2}{C_M}\right) + 2V_p\bar{v}_{n/2}^{\frac{p-1}{p}} \\ + 2^{\gamma-1}\frac{C_1 L_{\nabla G} c_\gamma C_S^2}{\mu\lambda_0} n^{-\gamma}.\end{aligned} \tag{24}$$

按照定理3.1的证明进行推导，我们得到

$$\begin{aligned}\mathbb{E}[V_n] \leq \exp\left(-c_\gamma\mu\lambda_0 n^{1-\gamma}(1-\varepsilon(n))\right)\cdot\left(K_1^{(2)} + K_{1'}^{(2)}\max_{1\leq j\leq n+1} v_k^{\frac{p-1}{p}} k^{\gamma-2\beta-\frac{p-1}{p}\delta}\right) \\ + K_2^{(2)} v_{\lfloor n/2\rfloor}^{\frac{p-1}{p}} n^{-\frac{(p-1)}{p}\delta} + K_3^{(2)} n^{-\gamma},\end{aligned}$$

with

$$\varepsilon(n) = \frac{2C_M n^{-1+\gamma}}{\mu\lambda_0}\left(1 + \frac{n^{(1+2\beta-2\gamma)^+}}{|2\gamma-2\beta-1|}\right), \tag{25}$$

其中 C_M 由（23）定义，且

$$K_1^{(2)} = \left(\mathbb{E}[V_0] + \frac{C_1 L_{\nabla G} c_\gamma C_S^2}{C_M}\right), \quad K_{1'}^{(2)} = 4\frac{\lambda_0\mu V_p}{C_M}, \tag{26}$$

$$K_2^{(2)} = 2^{1+\delta\frac{p-1}{p}} V_p, \tag{27}$$

$$K_3^{(2)} = 2^{\gamma-1}\frac{C_1 L_{\nabla G} c_\gamma C_S^2}{\mu\lambda_0}. \tag{28}$$

6.3 定理 3.3 和推论 3.1 的证明

定理的证明 3.3。 注意，可以将其改写为

$$\theta_{n+1} - \theta = \theta_n - \theta - \gamma_{n+1}H^{-1}g'_{n+1} - \gamma_{n+1}\left(A_n - H^{-1}\right)g'_{n+1}$$

由于 H 是对称的，因此得到

$$\begin{aligned}\|\theta_{n+1}-\theta\|^2 \leq \|\theta_n-\theta\|^2 - 2\gamma_{n+1}\left\langle g'_{n+1}, H^{-1}(\theta_n-\theta)\right\rangle - 2\gamma_{n+1}\left\langle \left(A_n - H^{-1}\right) g'_{n+1}, \theta_n-\theta\right\rangle \\ + 2\gamma_{n+1}^2\left\|H^{-1}g'_{n+1}\right\|^2 + 2\gamma_{n+1}^2\left\|A_n-H^{-1}\right\|^2\|g'_{n+1}\|^2\end{aligned}$$

首先，根据假设（**A3**）并由柯西-施瓦茨不等式，

$$\begin{aligned}(*) := \left|\mathbb{E}\left[2\gamma_{n+1}\left\langle\left(A_n-H^{-1}\right)g'_{n+1}, \theta_n-\theta\right\rangle|\mathcal{F}_n\right]\right| &= 2\gamma_{n+1}\left|\left\langle\left(A_n-H^{-1}\right)\nabla G(\theta_n), \theta_n-\theta\right\rangle\right| \\ &\leq 2L_{\nabla G}\gamma_{n+1}\left\|A_n-H^{-1}\right\|\|\theta_n-\theta\|^2.\end{aligned}$$

然后，根据假设（**A1'**），可得

$$(**) := \mathbb{E}\left[2\gamma_{n+1}^2\left\|H^{-1}g'_{n+1}\right\|^2|\mathcal{F}_n\right] \leq 4\gamma_{n+1}^2\mathrm{Tr}\left(H^{-1}\Sigma H^{-1}\right) + 4\gamma_{n+1}^2\left\|H^{-1}\right\|^2 L_{\nabla g}\|\theta_n-\theta\|^2$$

最后，可得

$$(***) = \mathbb{E}\left[-2\gamma_{n+1}\left\langle g'_{n+1}, H^{-1}(\theta_n-\theta)\right\rangle|\mathcal{F}_n\right] \leq -2\gamma_{n+1}\|\theta_n-\theta\|^2 + 2\gamma_{n+1}\left\|H^{-1}\right\|\|\delta_n\|\|\theta_n-\theta\|$$

其中，根据假设 （**A4**），$\|\delta_n\| := \|\nabla G(\theta_n) - H(\theta_n-\theta)\| \leq L_\delta\|\theta_n-\theta\|^2$。因此，

$$(***) \leq -2\gamma_{n+1}\|\theta_n-\theta\|^2 + 2\gamma_{n+1}\left\|H^{-1}\right\|L_\delta\|\theta_n-\theta\|^3,$$

由此可得，利用 $\|\theta_n-\theta\|^3 \leq \frac{1}{2a}\|\theta_n-\theta\|^2 + \frac{a}{2}\|\theta_n-\theta\|^4$ 满足 $a = \left\|H^{-1}\right\|L_\delta$这一事实 ，

$$(***) \leq -\gamma_{n+1}\|\theta_n-\theta\|^2 + \gamma_{n+1}\left\|H^{-1}\right\|^2 L_\delta^2\|\theta_n-\theta\|^4.$$

此外，

$$\begin{aligned}(****) &:= \mathbb{E}\left[2\gamma_{n+1}^2\left\|A_n-H^{-1}\right\|^2\|g'_{n+1}\|^2|\mathcal{F}_n\right] \\ &\leq 2\gamma_{n+1}^2\left\|A_n-H^{-1}\right\|^2 C_1 + 2\gamma_{n+1}^2 C_2\left\|A_n-H^{-1}\right\|^2\|\theta_n-\theta\|^2 \\ &\leq 2\gamma_{n+1}^2\left\|A_n-H^{-1}\right\|^2 C_1 + C_2\gamma_{n+1}\|\theta_n-\theta\|^4 + C_2\gamma_{n+1}^3\left\|A_n-H^{-1}\right\|^4.\end{aligned}$$

综上，应用 $(*)$ 上的柯西-施瓦茨不等式，可得:

$$\begin{aligned}\mathbb{E}\left[\|\theta_{n+1}-\theta\|^2\right] \leq \left(1-\gamma_{n+1}+4\left\|H^{-1}\right\|^2\gamma_{n+1}^2 L_{\nabla g}\right)\mathbb{E}\left[\|\theta_n-\theta\|^2\right] + 4\gamma_{n+1}^2\mathrm{Tr}\left(H^{-1}\Sigma H^{-1}\right) \\ + \gamma_{n+1}\left(\left\|H^{-1}\right\|^2 L_\delta^2 + C_2\right)\mathbb{E}\left[\|\theta_n-\theta\|^4\right] + C_2\gamma_{n+1}^3\mathbb{E}\left[\left\|A_n-H^{-1}\right\|^4\right] \\ + 2C_1\gamma_{n+1}^2\mathbb{E}\left[\left\|A_n-H^{-1}\right\|^2\right] + 2\gamma_{n+1}L_{\nabla G}\sqrt{\mathbb{E}\left[\|\theta_n-\theta\|^4\right]\mathbb{E}\left[\|A_n-H^{-1}\|^2\right]},\end{aligned}$$

利用命题 3.2 以及 $\mathbb{E}\left[\|\theta_n-\theta\|^4\right] \leq \frac{4}{\mu^2}\mathbb{E}\left[\text{均匀分布}_n^2\right]$，根据 （**A2**）和**(H2b)**及 **(H3)**，得到

$$
\begin{aligned}
\mathbb{E}\left[\|\theta_{n+1}-\theta\|^2\right] \leq & \left(1-\gamma_{n+1}+4\left\|H^{-1}\right\|^2\gamma_{n+1}^2 L_{\nabla g}\right)\mathbb{E}\left[\|\theta_n-\theta\|^2\right]+4\gamma_{n+1}^2\mathrm{Tr}\left(H^{-1}\Sigma H^{-1}\right) \\
& +\gamma_{n+1}\left(\left\|H^{-1}\right\|^2 L_\delta^2+C_2\right)\frac{M_n}{\mu^2}+C_2\gamma_{n+1}^3 2^3\left(C_S^4+\frac{1}{\mu^4}\right) \\
& +2C_1\gamma_{n+1}^2 v_{A,n}+2\gamma_{n+1}\frac{L_{\nabla G}}{\mu}\sqrt{M_n v_{A,n}} \\
\leq & \left(1-\gamma_{n+1}+4\left\|H^{-1}\right\|^2\gamma_{n+1}^2 L_{\nabla g}\right)\mathbb{E}\left[\|\theta_n-\theta\|^2\right] \\
& +\gamma_{n+1}\cdot\left[4\gamma_{n+1}\mathrm{Tr}\left(H^{-1}\Sigma H^{-1}\right)+\left(\frac{L_\delta^2}{\mu^2}+C_2\right)\frac{4M_n}{\mu^2}\right. \\
& \left.+C_2\gamma_{n+1}^2 2^3\left(C_S^4+\frac{1}{\mu^4}\right)+2C_1\gamma_{n+1}v_{A,n}+4\frac{L_{\nabla G}}{\mu}\sqrt{M_n v_{A,n}}\right].
\end{aligned}
$$

最后，记$C_A=c_\gamma\max\left\{4\left\|H^{-1}\right\|^2 L^{\nabla g},\frac{1}{4}\right\}$。那么，借助命题6.1，可得

$$
\begin{aligned}
&\mathbb{E}\left[\|\theta_n-\theta\|^2\right] \leq e^{-\frac{1}{2}c_\gamma n^{1-\gamma}}e^{2C_A c_\gamma\frac{2\gamma}{2\gamma-1}}\left(\mathbb{E}\left[\|\theta_0-\theta\|^2\right]+\frac{8\mathrm{Tr}\left(H^{-1}\Sigma H^{-1}\right)}{C_A}+c_\gamma\frac{16C_2\left(\mu^{-4}+C_S^4\right)}{C_A}+\frac{4C_1 v_{A,0}}{C_A}\right) \\
&+e^{-\frac{1}{2}c_\gamma n^{1-\gamma}}e^{2C_A c_\gamma\frac{2\gamma}{2\gamma-1}}\max_{1\leq k\leq n}(k+1)^\gamma\cdot\left(8\frac{L_\delta^2\mu^{-2}+C_2}{\mu^2 C_A}M_{k-1}+8\frac{L_{\nabla G}}{C_A\mu}\sqrt{M_{k-1}v_{A,k-1}}\right) \\
&+\frac{2^{3+\gamma}c_\gamma\mathrm{Tr}\left(H^{-1}\Sigma H^{-1}\right)}{n^\gamma}+\frac{8\left(\frac{L_\delta^2}{\mu^2}+C_2\right)}{\mu^2}M_{n/2}+\frac{8L_{\nabla G}}{\mu}\sqrt{M_{n/2}v_{A,n/2}} \\
&+\frac{2^{4+2\gamma}C_2 c_\gamma\left(\mu^{-4}+C_S^4\right)c_\gamma^2}{n^{2\gamma}}+\frac{2^{2+\gamma}C_1 c_\gamma v_{A,n/2}}{n^\gamma}.
\end{aligned}
$$

最后，

$$
\begin{aligned}
\mathbb{E}\left[\|\theta_n-\theta\|^2\right] \leq & e^{-\frac{1}{2}c_\gamma n^{1-\gamma}}\left(K_1^{(3)}+K_{1'}^{(3)}\max_{0\leq k\leq n}d_k(k+1)^\gamma\right) \\
& +n^{-\gamma}\left(2^{3+\gamma}c_\gamma\mathrm{Tr}\left(H^{-1}\Sigma H^{-1}\right)+\frac{K_2^{(3)}}{n^\gamma}+K_{2'}^{(3)}v_{A,n/2}\right)+d_{\lfloor n/2\rfloor}.
\end{aligned}
$$

with

$$
K_1^{(3)}=e^{2C_A c_\gamma^2\frac{2\gamma}{2\gamma-1}}\left(\mathbb{E}\left[\|\theta_0-\theta\|^2\right]+\frac{8\mathrm{Tr}\left(H^{-1}\Sigma H^{-1}\right)}{C_A}+c_\gamma\frac{16C_2\left(\mu^{-4}+C_S^4\right)}{C_A}+\frac{4C_1 v_{A,0}}{C_A}\right), \tag{29}
$$

$$
K_{1'}^{(3)}=\frac{1}{C_A}e^{2C_A c_\gamma^2\frac{2\gamma}{2\gamma-1}},\quad d_n=8L_{\nabla G}\sqrt{M_n v_{A,n}}+8\frac{L_\delta^2\mu^{-2}+C_2}{\mu^2}M_n, \tag{30}
$$

其中我们回顾一下，$C_A = c_\gamma \max\left\{4\left\|H^{-1}\right\|^2 L_{\nabla s}$，以及

$$K_2^{(3)} = 2^{4+2\gamma} C_2 c_\gamma \left(\mu^{-4} + C_S^4\right) c_\gamma^2, \quad K_{2'}^{(3)} = 2^{2+\gamma} C_1 c_\gamma. \tag{31}$$

□

推论 3.1 的证明。只要 $\delta^{\frac{p-2}{p}} \geq 2\gamma$，根据命题 3.2 及后续讨论，

$$\begin{aligned}\max_{0\leq k\leq n} d_k (k+1)^\gamma &= \max_{0\leq k\leq n}\left((k+1)^\gamma 8 L_{\nabla G}\sqrt{M_k v_{A,k}} + 8\frac{L_\delta^2\mu^{-2}+C_2}{\mu^2} M_k\right)\\ &\leq \frac{8 L_{\nabla G}\sqrt{v_{A,0}}}{c_\gamma}\sqrt{w_\infty(2\gamma)} + 8\frac{L_\delta^2\mu^{-2}+C_2}{\mu^2} w_\infty(\gamma).\end{aligned}$$

同理，

$$M_{n/2} \leq \frac{2^{2\gamma} w_\infty(2\gamma)}{n^{2\gamma}}.$$

因此，将这些不等式代入定理3.3 可得

$$\begin{aligned}\mathbb{E}\left[\|\theta_n - \theta\|^2\right] \leq & n^{-\gamma}\left(2^{3+\gamma} c_\gamma \operatorname{Tr}\left(H^{-1}\Sigma H^{-1}\right) + \frac{K_2^{(3')}}{n^\gamma} + K_{2'}^{(3')} v_{A,n/2} + K_{2''}^{(3')}\sqrt{v_{A,n/2}}\right)\\ &+ K_1^{(3')} e^{-\frac{1}{2}c_\gamma n^{1-\gamma}},\end{aligned}$$

with

$$K_1^{(3')} = K_1^{(3)} + K_1^{(3')}\left(\frac{8 L_{\nabla G}\sqrt{v_{A,0}}}{c_\gamma}\sqrt{w_\infty(2\gamma)} + 8\frac{L_\delta^2\mu^{-2}+C_2}{\mu^2} w_\infty(\gamma)\right), \tag{32}$$

$$K_2^{(3')} = K_2^{(3)} + 2\frac{L_\delta^2\mu^{-2}+C_2}{\mu^2} 2^{2\gamma} w_\infty(2\gamma), \quad K_{2'}^{(3')} = K_{2'}^{(3)},\ K_{2''}^{(3')} = 2^{2+\gamma} L_{\nabla G}\sqrt{w_\infty(2\gamma)}. \tag{33}$$

□

6.4 定理3.4 的证明

为证明该定理，我们将应用定理3.2。首先需要验证 $(A_n)_{n\geq 0}$满足假设**(H1a), (H1b)** 和 **(H2)**。假设 **(H1b)** 由构造可知（见（7））， 而 **(H1a)** 由以下引理给出：

引理 6.2。 假设*(A1)* 对某个 $p > 2$ 成立。那么，对所有 $0 < t < 1$,

$$\mathbb{P}\left[\lambda_{\min}(A_n) < c_\beta t\right] \leq v_n t^{2p},$$

with

$$v_n = c_\beta^p\left(\left(\frac{1}{n}\sum_{i=1}^d a_k\right)^p + C_1'' + \frac{2^p C_2'' V_p^p}{\mu^p}\right).$$

证明见附录 B。注意，根据引理 $\mathbb{E}\left[V_n^p\right] < +\infty$，结合 **(A1)**。从现在起，假设 $p > 2$，并令 $p' = \frac{2}{(1-\gamma)2-\gamma} p$ 以及 $\lambda = (1-\gamma)(\gamma - 2\beta)$。注意，

$\lambda < 1-\gamma$，$\lambda < \gamma - 2\beta$ 以及 $p' < p$。因此，应用命题 3.1，其中 $\lambda_0 = c\beta$、$\delta = 0$、$q = 2$ p，

$$\mathbb{E}\left[V_n^{p'}\right] \le \exp\left(-c_\gamma \mu \lambda_0 n^{1-(\lambda+\gamma)}(1-\varepsilon'(n))\right)\left(K_1^{(1')} + K_{1'}^{(1')} \max_{1\le k\le n+1} k^{\gamma-2\beta-\lambda-2(p-p')\lambda} v_0^{\frac{p-p'}{p}}\right)$$
$$+ K_2^{(1')} n^{-p'(\gamma-2\beta-\lambda)} + K_3^{(1')} v_0^{\frac{p-p'}{p}} (n+1)^{-2(p-p')\lambda},$$

其中，ϵ'(n)、K$^{(1')}$ 1、K$^{(1')}$ $_{1'}$、K$^{(1')}$ 2 和 K$^{(1')}$ 3 分别由式 (57)、(58) 和 (60) 给出，且 $\lambda_0 = \mathrm{c}\beta$。根据 λ 的选择，p′ 有

$$p'(\gamma - 2\beta - \lambda) = p\frac{2(1-\gamma)}{2-\gamma}\gamma(\gamma-2\beta) = 2(p-p')\lambda,$$

从而

$$\mathbb{E}\left[V_n^{p'}\right] \le \tilde{K}_1 \exp\left(-c_\gamma \mu c_\beta n^{1-((1-\gamma)(\gamma-2\beta)+\gamma)}(1-\varepsilon'(n)\right) + \tilde{K}_2 (n+1)^{-\frac{2(1-\gamma)\gamma(\gamma-2\beta)}{2-\gamma}p} := c_n \tag{34}$$

with

$$\tilde{K}_1 = K_1^{(1')} + K_{1'}^{(1')} v_0^{\frac{\gamma}{2-\gamma}}, \quad \tilde{K}_2 = K_2^{(1')} + K_3^{(1')} v_0^{\frac{\gamma}{2-\gamma}}.$$

由强凸性，可据此得到估计值的第一个收敛速率。以下引理可确保 **(H1a)** 成立，但其速率可能优于引理6.2中的速率。

引理 6.3. 假设对某个 $p > 2$，***(A1)*** 成立。则，

$$\mathbb{P}[\lambda_{\min}(A_n) < \tilde{\lambda}_0] \le \frac{v_0 \log(n+1)}{(n+1)^{\frac{2(1-\gamma)\gamma(\gamma-2\beta)}{2-\gamma}p\wedge 1}},$$

其中，$\tilde{\lambda}_0 = \left[{}^{2}_{(1-\gamma)}{}^{2-\gamma}p\left(C_{\left(\frac{2(1-\gamma)}{2-\gamma}\right)}+1\right)\right]^{-\frac{2-\gamma}{2\,1-\gamma}}{}_{()p}$，且 v_0 由 (73) 给出，其中 $p' = {}^{2}_{(1-\gamma)}{}^{2-\gamma}p$。

证明见附录B。由（34）还可以推导出关于 $\mathbb{E}\left[\|A_n\|^4\right]$在仅满足（**A6**）时的界。

引理 6.4。假设（***A1***）-（***A6***）和（***A1'***）对某个 p 成立 > 2。那么，对于 β <最小 $\left\{\frac{(1-\gamma)\gamma(\gamma-2\beta)p}{4\ 2-\gamma}\ (\ ),\ 1/4\right\}$，由（$A_n$）定义的随机矩阵序列（4）满足

$$\mathbb{E}\left[\|A_n\|^4\right] \le C_S^4,$$

其中 C_S^4 由（74）给出。

证明见附录 B。若更强假设 （**A6'**）成立，则可得到关于 $\mathbb{E}\left[\|A_n\|^4\right]$ 的更优且更简洁的界，如下一个引理所示。

引理 6.5。 假设 （***A1***）-（***A6'***） 和 （***A1'***） 对某个 $p > 2$成立。那么，对于 $\beta <$最小值 $\{\gamma/2 \wedge 1/4\}$，由 （A_n） 定义的随机矩阵序列 （4） 满足

$$\mathbb{E}\left[\|A_n\|^4\right] \leq C_S^4,$$

其中 C_S^4 由(75)给出。

证明见附录 B。定理 3.4 因此是定理 3.2的推论，其假设由引理 6.2、6.3 和6.4 （或6.5） 满足。于是我们有

$$\mathbb{E}\left[V_n\right] \leq \exp\left(-c_\gamma \mu \tilde{\lambda}_0 n^{1-\gamma}\left(1-\varepsilon(n)\right)\right) \cdot \left(K_1^{(2)} + K_{1'}^{(2)} \max_{1\leq j\leq n+1} v_k^{\frac{p-1}{p}} k^{\gamma-2\beta-\frac{p-1}{p}\delta}\right)$$
$$+ K_2^{(2)} v_{\lfloor n/2\rfloor}^{\frac{p-1}{p}} n^{-\frac{(p-1)}{p}\min\left\{\frac{2(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma},1\right\}} + K_3^{(2)} n^{-\gamma}$$

其中 $\mathrm{K}_1^{(2)}$、$\mathrm{K}_{1'}^{(2)}$、$\mathrm{K}_2^{(2)}$ 和 $\mathrm{K}_3^{(2)}$ 分别由 (26)、(27) 和 (28) 给出，$\delta = \min\left\{\frac{2(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma}、1\right\}$、$\lambda_0$ 由 (72) 给出，$\mathrm{v}_n = \mathrm{v}_0 \log(\mathrm{n}+1)$，其中 v_0 由 (73) 给出，C_S 由 (74) 或 (75) 给出，具体取决于 (A6) 或 (A6') 是否成立。根据强凸性

$$\mathbb{E}\left[\|\theta_n-\theta\|^2\right] \leq \tilde{K}_1^{(4)} \exp\left(-c_\gamma \mu \tilde{\lambda}_0 n^{1-\gamma}\left(1-\tilde{\varepsilon}(n)\right)\right) + \tilde{K}_2^{(4)}\left(v_0 \log(n+1)\right)^{\frac{p-1}{p}} n^{-\frac{(p-1)}{p}\min\left\{\frac{2(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma},1\right\}}$$
$$+ \tilde{K}_3^{(4)} n^{-\gamma},$$

˜其中 λ_0 由 (72)定义

$$\tilde{\varepsilon}(n) = \frac{2C_M n^{-1+(1-\gamma)(2\gamma-\beta)+\gamma}}{\mu\tilde{\lambda}_0}\left(1+\frac{n^{(1+2\beta-2\gamma)^+}}{|2\gamma-2\beta-1|}\right), \tag{35}$$

with

$$\tilde{K}_1^{(4)} = \frac{2}{\mu}\left(K_1^{(2)} + K_{1'}^{(2)} v_0\right), \quad \tilde{K}_2^{(4)} = \frac{2K_2^{(4)}}{\mu}, \quad \tilde{K}_3^{(4)} = \frac{2K_3^{(4)}}{\mu}. \tag{36}$$

其中 v_0 由 （73） 给出。

6.5 定理 4.1 和定理 4.2 的证明

该证明依赖于验证定理 3.3所需的每个假设。

验证假设 （A1）、（A1'） 至 （A6）。首先，注意到

$$\|\nabla_h g(X,Y,h)\| \leq \left\|\left(X^T h - X^T\theta - \epsilon\right)X\right\| \leq |\epsilon|\,\|X\| + \|X\|^2\,\|h-\theta\|.$$

那么，如果X 和 ϵ 分别具有 $4p$ 阶和 $2p$ 阶矩，由于 ϵ 和X 相互独立，

$$\mathbb{E}\left[\|\nabla_h g(X,Y,h)\|^{2p}\right] \leq \sigma_{(2p)} + C_{(2p)}\|h-\theta\|^{2p}$$

其中 $\sigma_{(t)} = 2^{t-1}\mathbb{E}\left[|\epsilon|^t\right]\mathbb{E}\left[\|X\|^t\right]$ 和 $C_{(t)} = 2^{t-1}\mathbb{E}\left[\|X\|^{2t}\right]$。在一种特定情况下，如果 $p \geq 2$，

假设 （**A1**） 成立。此外，由于对所有 h， $\nabla^2 G(h) = \mathbb{E}\left[XX^T\right]$ 均为正， （**A2**） 至（**A4**） 成立，其中 $\mu = \lambda_{\min}\left(\mathbb{E}\left[XX^T\right]\right) =: \lambda_{\min}$，$L_{\nabla G} = \lambda_{\max}\left(\mathbb{E}\left[XX^T\right]\right) =: \lambda_{\max}$，且（**A5**） 成立，其中 $L_\delta = 0$。最后，假设（**A1'**） 成立，因为

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla_h g(X,Y,h) - \nabla_h g(X,Y,\theta)\right\|^2\right] &= \mathbb{E}\left[\left\|X^T(h-\theta)X\right\|^2\right] \\ &\leq \underbrace{\mathbb{E}\left[\|X\|^4\right]}_{=:L_{\nabla g}} \|h-\theta\|^2 .\end{aligned}$$

现在我们可以证明定理 4.1

定理 4.1 的证明。 **验证随机牛顿算法的假设（H1）**。我们先检查假设 **(H1)** 对 $\widetilde{S}_n = \frac{1}{n+1}\left[S_0 + \sum_{i=1}^n X_i X_i^T\right]$。

引理 6.6。 假设 X 的$4p$ 阶矩存在，其中 $p > 2$。那么，对于 $\lambda_0 = \frac{1}{2\mathbb{E}\|X\|^2}$，有

$$\mathbb{P}\left[\lambda_{\min}\left(\widetilde{S}_n^{-1}\right) < \lambda_0\right] \leq \tilde{v}_n$$

with

$$\tilde{v}_n = \frac{2^{p-1}}{\left(\mathbb{E}\left[\|X\|^2\right]\right)^p}\left(C_1(p) n^{1-p}\mathbb{E}\left[|Z|^p\right] + C_2(p) n^{-p/2}\left(\mathbb{E}\left[|Z|^2\right]\right)^{p/2} + \|S_0\|^p n^{-p}\right),$$

其中 $Z = \|X\|^2 - \mathbb{E}\left[\|X\|^2\right]$ 且 $C_1(p)$，$C_2(p)$ 是 *Rosenthal* 不等式中的数值常数，参见 *Pinelis(1994)*。

证明见第C节。为了处理$S_n = \frac{\|\widetilde{S}_n^{-1}\|}{\min(n, \|\widetilde{S}_n^{-1}\|)}\widetilde{S}_n$，首先需要控制 $\lambda_{\min}(\widetilde{S}_n)$ 的行为。设$H = \mathbb{E}\left[XX^T\right]$。

命题 6.2 (参见 Koltchinskii 和 Mendelson (2015)，定理 1.5 和定理 3.3)。假设 $0 < \lambda_{\min} I_d \leq H := \mathbb{E}\left[XX^T\right] \leq \lambda_{\max} I_d$并且存在 $L_{MK} > 0$使得$\mathbb{E}\left[\langle X, t\rangle^2\right] \leq L_{MK}\mathbb{E}\left[|\langle X, t\rangle|\right]$ 对所有 $t \in \mathbb{S}^{d-1}$。那么，对于 $n \geq c_1 d$，

$$\mathbb{P}\left[\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^n X_i X_i^T\right) \leq c_2\right] \leq 2\exp\left(-c_3 n\right),$$

其中 $c_1 = \frac{\lambda_{\max}^2(16 L_{MK}^4)}{\lambda_{\min}^2}$，$c_2 = \frac{\lambda_{\min}}{8\sqrt{2}L_{MK}^2}$ 且 $c_3 = \frac{1}{128 L_{MK}^4}$。

注意，常数 c_1, c_2 和 c_3 关于 L_{MK} 和 $\lambda_{\min}$。根据后一结果、引理 6.6 和命题 6.2，我们可以推出 S_n 满足假设（H1）。我们还会多次用到阈值

$$n_0 = \max\left\{c_1 d, \left(\frac{1}{c_\beta c_2}\left(1 + \frac{1}{c_1 d}\right)\right)^{-1/\beta}\right\}. \tag{37}$$

引理 6.7。 设 X 满足命题 6.2 的假设，且具有$4p$ 阶矩，其中 $p > 2$。那么，对于 $\lambda_0 = \frac{1}{2\mathbb{E}[\|X\|^2]}$，有

$$\mathbb{P}\left[\lambda_{\min}\left(\overline{S}_n^{-1}\right) < \lambda_0\right] \leq v_{n+1}(n+1)^{-p/2}$$

其中 $\delta = p/2$，$v_{n+1} = (n+1)^{\delta}$ 对于 $n \leq n_0$，并且对于 $n > n_0$，

$$v_n = 2\exp(-c_3 n)n^{p/2} + \frac{2^{p-1}\left(C_2(p)\mathbb{E}\left[|Z|^2\right]^{p/2} + C_1(p)n^{1-p/2}\mathbb{E}\left[|Z|^p\right] + \|S_0\|^p n^{-p/2}\right)}{\mathbb{E}\left[\|X\|^2\right]^p},$$

其中 c_1, c_2, c_3 见命题 6.2，$C_1(p)$ 和 $C_2(p)$ 是依赖于 p 和 $Z = \|X\|^2 - \mathbb{E}\left[\|X\|^2\right]$的数值常数。

证明见第 C 节。作为一个特例，假设**(H1a)** 以速率 $\delta = p/2$ 得到验证，当 $\gamma > 1/2$。

验证随机牛顿算法的假设 (H2)。上述引理可直接推得如下结论。

引理 6.8。 假设命题 6.2 的假设成立，且 X 存在$4p$ 阶矩，其中 $p > 2$。那么，对所有 $\kappa > 0$，有

$$\mathbb{E}\left[\|\bar{S}_n^{-1}\|^{\kappa}\right] \leq 2\beta_{n+1}^{\kappa}\exp(-c_3 n) + c_2^{-\kappa}$$

对于 $n \geq c_1 d$，且

$$\mathbb{E}\left[\|\bar{S}_n^{-1}\|^{\kappa}\right] \leq \left[(c_1 d + 1)\|S_0^{-1}\|\right]^{\kappa}$$

对于 $n \leq c_1 d$，其中 c_1, c_2, c_3 如命题 6.2所给出。

证明见第 C 节。最后，以下命题给出了假设**(H2)**的精确界。

命题 6.3。 设命题 6.2 的假设成立，且 X 存在阶数为 $4p$ 的矩，其中 $p > 2$。则

$$\mathbb{E}\left[\|\bar{S}_n^{-1}\|^2\right] \leq \max\left\{2c_\beta^2\left(\frac{2\beta}{ec_3}\right)^{2\beta} + c_2^{-2}, \left[(c_1 d + 1)\left\|S_0^{-1}\right\|\right]^2\right\} \leq C_S^2$$

and

$$\mathbb{E}\left[\|\bar{S}_n^{-1}\|^4\right] \leq \max\left\{2c_\beta^4\left(\frac{4\beta}{ec_3}\right)^{4\beta} + c_2^{-4}, \left[(c_1 d + 1)\left\|S_0^{-1}\right\|\right]^4\right\} \leq C_S^4$$

for all $n \geq 0$, *with* $C_S^4 := \max\left\{\left(2c_\beta^2\left(\frac{4\beta}{ec_3}\right)^{2\beta} + c_2^{-2}\right)^2, \left[(c_1 d + 1)\left\|S_0^{-1}\right\|\right]^4\right\}$

证明见第 C 节。注意 $C_S = O(d)$。

首个收敛结果。 由于在线性模型的情形下，有 $C^1 = \sigma_{(2)}, C_1' = \sigma_{(4)}, C_2 = C_{(2)}, C_2' = C_{(4)}, L_{\nabla G} = \lambda_{\max}, \mu = \lambda_{\min}, \lambda_0 = \frac{1}{2\mathbb{E}\left[\|X\|^2\right]}$ $\delta = p/2$，命题 3.2 现可表述如下：

命题 6.4。 假设存在 $p > 2$，使得 X, ϵ 分别具有阶数为$4p$ 和$2p$ 的矩。再假设存在正常数 L_{MK}，使得对任意 $h \in \mathbb{S}^{d-1}, \sqrt{\mathbb{E}\left[hXX^Th\right]} \leq L_{MK}\mathbb{E}\left[\left|X^Th\right|\right]$。那么，记 $\lambda_{\min}$ 和 $\lambda_{\max}$ 为 $\mathbb{E}\left[XX^T\right]$的最小和最大特征值，

$$\mathbb{E}\left[V_n^2\right] \leq \exp\left(-\frac{3c_\gamma\lambda_{\min}}{4\mathbb{E}\left[\|X\|^4\right]}n^{1-\gamma}\right)\left(K_{1,lin}^{(2')} + K_{1',lin}^{(2')}\max_{1\leq k\leq n+1} v_k^{\frac{p-2}{p}}k^{\gamma-\frac{p-2}{p}}\right)$$
$$+ K_{2,lin}^{(2')}n^{-2\gamma} + K_{3,lin}^{(2')}v_{\lfloor n/2\rfloor}^{(p-2)/p}n^{-(p-2)/2} := c_{n,lin}.$$

其中 v_n 由引理 6.7给出

$$K_{1,lin}^{(2')} = e^{2a_{M,lin}\frac{2\gamma-2\beta}{2\gamma-2\beta-1}}\left(\mathbb{E}\left[V_0^2\right] + \frac{2a_{1,lin}c_\gamma^2}{a_{M,lin}}\right), \quad K_{1',lin}^{(2')} = e^{2a_{M,lin}\frac{2\gamma-2\beta}{2\gamma-2\beta-1}}\frac{4\lambda_{\min}V_{p,lin}^2}{a_{M,lin}\mathbb{E}\left[\|X\|^2\right]},$$

$$K_{2,lin}^{(2')} = \frac{2^{1+2\gamma}a_{1,lin}c_\gamma^2\mathbb{E}\left[\|X\|^2\right]}{3\lambda_{\min}}, \quad K_{3,lin}^{(2')} = \frac{2^{p/2+1}}{3}V_{p,lin}^2,$$

其中，回顾记号 $\sigma_{(t)} = 2^{t-1}\mathbb{E}\left[|\epsilon|^t\right]\mathbb{E}\left[\|X\|^t\right]$ 和 $C_{(t)} = 2^{t-1}\mathbb{E}\left[\|X\|^{2t}\right]$，

$$a_{M,lin} := \max\left\{\left(\frac{2\lambda_{\max}C_{(2)}}{\lambda_{\min}} + \frac{2\lambda_{\max}^2}{\lambda_{\min}^2}\left(4C_{(2)} + C_{(4)}c_\gamma^2c_\beta^2\right)\right)c_\gamma c_\beta^2, \left(\frac{3\lambda_{\min}}{4\mathbb{E}\left[\|X\|^2\right]}\right)^{\frac{2\gamma-2\beta}{\gamma}}c_\gamma^{\frac{\gamma-2\beta}{\gamma}}\right\},$$

其中 C_S 由命题 6.3给出，$a_{1,lin} := C_S^4\lambda_{\max}^2\left(\frac{16\lambda_{\max}^2\sigma_{(2)}^2\mathbb{E}\left[\|X\|^2\right]}{\lambda_{\min}^3} + \sigma_{(4)}c_\gamma + 2C_{(2)}^2\frac{\mathbb{E}\left[\|X\|^2\right]}{\lambda_{\min}}\right)$ 并且

$$\mathbb{E}\left[V_{n,}^p\right] \leq e^{a_{p,lin}c_\gamma^2c_\beta^2\frac{2\gamma-2\beta}{2\gamma-2\beta-1}}\max\left\{1, \mathbb{E}\left[V_0^2\right]\right\} := V_{p,lin}^p$$

其中

$$a_{p,lin} := p\left(\frac{C_{(2)}}{\lambda_{\min}} + \frac{\sigma_{(2)}}{2}\right) + 2^{p-2}(p-1)p\lambda_{\max}^2\left(c_\gamma^2c_\beta^2\left(\sigma_{(4)} + \frac{4C_{(4)}}{\lambda_{\min}^2}\right) + \frac{2\sigma_{(2)}}{\lambda_{\min}} + \frac{4C_{(2)}}{\lambda_{\min}^2}\right)$$
$$+ 2^{p-2}(p-1)p\lambda_{\max}^p\left(c_\gamma^{2p-2}c_\beta^{2p-2}\left(\sigma_{(2p)} + \frac{2^pC_{(2p)}}{\lambda_{\min}^2}\right) + c_\gamma^{p-2}c_\beta^{p-2}\left(\frac{1}{2}\sigma_{(2p)} + \frac{2p}{\lambda_{\min}^2}\left(\frac{1}{2} + \sqrt{C_{(2p)}}\right)\right)\right). \tag{38}$$

验证随机牛顿算法的假设 (H3)。假设**(H3)** 随后可由S_n 向H的收敛以及假设**(H2)**直接得到。

引理 6.9。 假设 X 存在阶为 $2p$ 的矩，其中 $p > 4$，并且进一步假设 X 的分布满足命题 6.2 的假设。那么，对于 $n \geq n_0$ （其中 n_0 定义于(37)），

$$\mathbb{E}\left[\left\|\overline{S}_n^{-1} - H^{-1}\right\|^2\right] \leq \frac{4\left(\mathbb{E}\left[\|X\|^{2p}\right]\right)^{2/p}}{\left(\lambda_{\min}\beta_n\right)^2} e^{-c_3(p-2)n/p} + \frac{2\mathbb{E}\left[\|X\|^4\right]}{n\left(\lambda_{\min}c_2\right)^2} + \frac{2\left\|S_0 - H\right\|_F^2}{n^2\left(\lambda_{\min}c_2\right)^2} =: v_{H,n}. \tag{39}$$

对于$n < n_0$,，我们只需进行界定

$$\mathbb{E}\left[\left\|\overline{S}_n^{-1} - H^{-1}\right\|^2\right] \leq \max\left\{\frac{2}{\lambda_{\min}^2} + 2C_S^2, v_{H,n_0}\right\} := v_{H,n}.$$

根据引理 6.7，**(H1a)** 满足条件 $\delta = p/2$。应用定理 3.3，结合前述引理和命题中计算出的常数，最终得到

$$\begin{aligned}\mathbb{E}\left[\|\theta_n - \theta\|^2\right] &\leq e^{-\frac{1}{2}c_\gamma n^{1-\gamma}}\left(K_{1,\text{lin}}^{(3)} + K_{1',\text{lin}}^{(3)} \max_{0\leq k\leq n} d_k(k+1)^\gamma\right) \\ &+ n^{-\gamma}\left(2^{3+\gamma}c_\gamma\mathbb{E}\left[\epsilon^2\right]\operatorname{Tr}\left(H^{-1}\right) + \frac{K_{2,\text{lin}}^{(3)}}{n^\gamma} + K_{2',\text{lin}}^{(3)} v_{H,n/2}\right) + d_{\lfloor n/2\rfloor}.\end{aligned}$$

其中 $v_{H,\ n}$ 由（39）定义。回顾 $\lambda_{\min}$ 和 $\lambda_{\max}$ 分别是 $\mathbb{E}\left[XX^T\right]$的最小和最大特征值，并且在线性情形下有 $C_A = 4c_\gamma\frac{\mathbb{E}\left[\|X\|^4\right]}{\lambda_{\min}^2} \geq 4c\gamma$,

$$\begin{aligned}K_{1,\text{lin}}^{(3)} &= e^{8\frac{\mathbb{E}\left[\|X\|^4\right]}{\lambda_{\min}^2}c_\gamma^3\frac{2\gamma}{2\gamma-1}}\left(\mathbb{E}\left[\|\theta_0 - \theta\|^2\right] + \frac{2\mathbb{E}\left[\epsilon^2\right]\operatorname{Tr}\left(H^{-1}\right)}{c_\gamma} + 4C_{(2)}\left(\lambda_{\min}^4 + C_S^4\right) + \frac{\sigma_{(2)}v_{H,0}}{c_\gamma}\right), \\ K_{1',\text{lin}}^{(3)} &= \frac{1}{4c_\gamma}e^{8\frac{\mathbb{E}\left[\|X\|^4\right]}{\lambda_{\min}^2}c_\gamma^3\frac{2\gamma}{2\gamma-1}}, \quad d_n = 8\lambda_{\max}\sqrt{c_{n,\text{lin}}v_{H,n}} + 8\frac{C_{(2)}}{\lambda_{\min}^2}c_{n,\text{lin}}, \\ K_{2,\text{lin}}^{(3)} &= 2^{4+2\gamma}C_{(2)}c_\gamma\left(\lambda_{\min}^{-4} + C_S^4\right)c_\gamma^2, \quad K_{2',\text{lin}}^{(3)} = 2^{2+\gamma}\sigma_{(2)}c_\gamma,\end{aligned} \tag{40}$$

以及 $c_{n,\ \text{lin}}$ 和 C_S^4 分别定义于命题6.4 和 6.3中。 □

定理 4.2的证明。 我们首先证明假设 **(A6′)** 得到满足。对于所有 h,

$$\mathbb{E}\left[\nabla_h g\left(X,Y,h\right)\nabla_h g\left(X,Y,h\right)^T\right] = \mathbb{E}\left[\left(Y - X^Th\right)^2 XX^T\right] = \mathbb{E}\left[\epsilon^2\right]\mathbb{E}\left[XX^T\right] + \mathbb{E}\left[\left(X^Th - X^T\theta\right)^2 XX^T\right]$$

并且 **(A6′)** 满足 $\alpha = \mathbb{E}\left[\epsilon^2\right]\lambda_{\min}$，根据（75），有

$$\mathbb{E}\left[\|A_n\|^4\right] \leq \frac{4d\left(1 + \sigma_{(4)} + C_{(4)}\frac{4V_{2,ada}^2}{\lambda_{\min}^2}\right)}{\mathbb{E}\left[\epsilon^2\right]^2\lambda_{\min}^2} := C_{S,ada}^4,$$

其中 V_2 由引理 6.1 针对 $p = 2$给出。然后，应用定理3.4,

$$\mathbb{E}\left[\|\theta_n - \theta\|^2\right] \leq K_{1,lin}^{ada} \exp\left(-c_\gamma \lambda_{\min} \lambda_{0,lin}^{ada} n^{1-\gamma}\left(1 - \varepsilon_{n,lin}^{ada}\right)\right) + K_{2,lin}^{ada}\left(v_{0,lin}^{ada}\log(n+1)\right)^{\frac{p-1}{p}} n^{-\frac{(p-1)}{p}\min\left\{\frac{2(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma},1\right\}} + K_{3,lin}^{ada} n^{-\gamma},$$

其中 λ_0^{ada}、$lin = \underline{\left[\begin{matrix}4\\(1-\gamma\\2-\gamma\end{matrix}\right.}\left(\right) pC\left(\frac{4p(1-\gamma)}{2-\gamma}\right)+1\Bigg)\Bigg]^{-\frac{2-\gamma}{4\ 1-\gamma}}{}_{p\ ()}$，并回顾一下，$\lambda_{\min}$和 $\lambda_{\max}$是 $\mathbb{E}\left[XX^T\right]$的最小特征值和最大特征值,

$$\varepsilon_{n,lin}^{ada} = \frac{2C_{M,lin}^{ada} n^{-1+(1-\gamma)(2\gamma-\beta)+\gamma}}{\lambda_{\min}\lambda_{0,lin}^{ada}}\left(1 + \frac{n^{(1+2\beta-2\gamma)^+}}{|2\gamma - 2\beta - 1|}\right), \tag{41}$$

$$K_{1,\text{lin}}^{ada} = \frac{2}{\lambda_{\min}}\left(\mathbb{E}\left[V_0\right] + \frac{c_\gamma \lambda_{\max}\sigma_{(2)}C_{S,ada}^2}{C_{M,lin}^{ada}} + \frac{4\lambda_{\min}\lambda_{0,lin}^{ada}V_{p,lin}^{ada}}{C_{M,lin}^{ada}}\right), \tag{42}$$

$$K_{2,\text{lin}}^{ada} = \frac{1}{\lambda_{\min}} 2^{p/2+3/2} V_{p,lin}^{ada} \tag{43}$$

$$K_{3,\text{lin}}^{ada} = \frac{2^\gamma c_\gamma \lambda_{\max}\sigma_{(2)}C_{S,ada}^2}{\lambda_{\min}^2 \lambda_{0,lin}^{ada}}. \tag{44}$$

where $v_0 = dM(\beta) + \frac{d2^{\frac{2(1-\gamma)}{2-\gamma}p}\left(\sigma_{\left(\frac{4(1-\gamma)}{2-\gamma}p\right)} + 2^{\frac{2(1-\gamma)}{2-\gamma}p} C_{\left(\frac{4(1-\gamma)}{2-\gamma}p\right)} \frac{V_{p,ada}^{\frac{2(1-\gamma)}{2-\gamma}p}}{\lambda_{min}^{\frac{2(1-\gamma)}{2-\gamma}p}}\right)}{\sigma_{\left(\frac{4(1-\gamma)}{2-\gamma}p\right)}+1}$.

$$C_{M,lin}^{ada} = \max\left\{\frac{C_{(2)}\lambda_{\max}c_\beta^2 c_\gamma}{\lambda_{\min}}, (\lambda_{\min}\lambda_{0,lin}^{ada})^{\frac{2\gamma-2\beta}{\gamma}} c_\gamma^{\frac{\gamma-2\beta}{\gamma}}\right\}$$

and

$$V_{p,ada}^p = e^{-p\lambda_{\min}\lambda_0' c_\gamma\left(1+\frac{1+\left(\frac{c_\gamma c_\beta^2 a_{p,lin}^{ada}}{p\lambda_{\min}\lambda_0'}\right)^{\frac{1-\gamma-\lambda'}{\gamma-2\beta-\lambda'}}}{1-\gamma-\lambda'}\right) + c_\gamma^2 c_\beta^2 a_p\left(1+\frac{1+\left(\frac{c_\gamma c_\beta^2 a_{p,lin}^{ada}}{p\lambda_{\min}\lambda_0'}\right)^{\frac{1-2\gamma+2\beta}{\gamma-2\beta-\lambda'}}}{1-2\gamma+2\beta}\right)}$$

其中

$$a_{p,lin}^{ada} = p\left(\frac{C_{(2)}}{\lambda_{\min}} + \frac{\sigma_{(2)}}{2}\right) + 2^{p-2}(p-1)p\lambda_{\max}^2\left(c_\gamma^2 c_\beta^2\left(\sigma_{(4)} + \frac{4C_{(4)}}{\lambda_{\min}^2}\right) + \frac{2\sigma_{(2)}}{\mu} + \frac{4C_{(2)}}{\lambda_{\min}^2}\right)$$
$$+ 2^{p-2}(p-1)p\lambda_{\max}^p\left(c_\gamma^{2p-2}c_\beta^{2p-2}\left(\sigma_{(2p)} + \frac{2^p C_{(2p)}}{\lambda_{\min}^2}\right) + c_\gamma^{p-2}c_\beta^{p-2}\left(\frac{1}{2}\sigma_{(2p)} + \frac{2p}{\lambda_{\min}^2}\left(\frac{1}{2} + \sqrt{C_{(2p)}}\right)\right)\right), \tag{45}$$

and

$$a_{2,lin}^{ada} = \sigma_{(2)} + \frac{2C_{(2)}}{\lambda_{\min}} + \frac{4\lambda_{\max}^2}{\lambda_{\min}}\sigma_{(2)} + \frac{8\lambda_{\max}^2 C_{(2)}}{\lambda_{\min}^2} + 2\lambda_{\max}^2\sigma_{(4)}c_\gamma^2 c_\beta^2 + \frac{8\lambda_{\max}^2 C_{(4)}}{\lambda_{\min}^2}c_\gamma^2 c_\beta^2 \tag{46}$$

□

6.6 定理5.1的证明

证明依赖于对定理中每个假设的验证3.3。
验证假设（A1）、（A1'）至（A6）。首先，注意到对所有 $0 \leq a \leq 2p$，有

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla_h l\left(Y, X^T h\right) X+\sigma h\right\|^a\right] & \leq 2^{a-1} \mathbb{E}\left[\left\|\nabla_h l\left(Y, X^T \theta_\sigma\right) X+\sigma \theta_\sigma\right\|^a\right] \\ & +2^{a-1} \mathbb{E}\left[\left\|\nabla_h l\left(Y, X^T h\right) X-\nabla_h \ell\left(Y, X^T \theta_\sigma\right) X+\sigma\left(h-\theta_\sigma\right)\right\|^a\right] \\ & \leq 2^{a-1} L_\sigma^a+2^{a-1} \underbrace{\mathbb{E}\left[\left(L_{\nabla l}\|X\|+\sigma\right)^a\right]}_{=: C_{\mathrm{GLM}}^{(a)}}\left\|h-\theta_\sigma\right\|^a \end{aligned} \tag{47}$$

因此假设（**A1**）得以验证。类似地，

$$\mathbb{E}\left[\left\|\left(\nabla_h g\left(X, h\right)-\nabla_h g\left(X, \theta_\sigma\right)\right)\right\|^2\right] \leq \mathbb{E}\left[\left(L_{\nabla l}\|X\|+\sigma\right)^2\right]\left\|h-\theta_\sigma\right\|^2 \leq C_{\mathrm{GLM}}^{(2)}\left\|h-\theta_\sigma\right\|^2$$

因此（**A1'**）得以验证。注意，（**A2**）和（**A4**）根据假设并结合 $\mu=\sigma$ 得以验证；对于（**A3**），有

$$\left\|\mathbb{E}\left[\nabla_h^2 \ell\left(Y, X^T h\right) X X^T+\sigma I_d\right]\right\|_{op} \leq L_{\nabla l} \mathbb{E}\left[\|X\|^2\right]+\sigma =: C_{\mathrm{GLM}}. \tag{48}$$

注意，假设（**A5**）由（**GLM1**）给出，而对于假设（**A6**），（**GLM3**）与（12）结合可得

$$\mathbb{E}\left[\left(\nabla_h g(X, \theta_v)\right)_k^2\right]=\mathbb{E}\left[\left|\nabla_h l\left(Y, X^T \theta_\sigma\right) X_k+\sigma(\theta_\sigma)_k\right|^2\right]>\alpha_\sigma$$

对所有 $1 \leq k \leq d$。

验证假设（H1）。以下引理保证假设（**H1**）成立。

引理6.10。首先假设(8) 成立，且 X 的$2p$ 阶矩存在，其中某个 $p>0$。在(10)定义的正则化情形下，记 $\lambda_0=\frac{1}{2 L_{\nabla l} \mathbb{E}\left[\|X\|^2\right]+2 \lambda}$，我们有

$$\mathbb{P}\left[\lambda_{\min }\left(\overline{S}_n^{-1}\right)<\lambda_0\right] \leq v_n$$

with

$$v_n=\frac{2^{p-1}}{\left(L_{\nabla l} \mathbb{E}\left[\|X\|^2\right]+\sigma\right)^p}\left(n^{-p}\left\|S_0\right\|^p+C_1(p) n^{1-p} \mathbb{E}\left[|T|^p\right]+C_2(p) n^{-p / 2}\left(\mathbb{E}\left[|T|^2\right]\right)^{p / 2}\right),$$

其中 $T=L_{\nabla l}\left(\|X\|^2-\mathbb{E}\left[\|X\|^2\right]\right)+\sigma\left(\|Z\|^2-1\right)$ 且 Z 是一个与 X 独立的标准 d 维随机变量。此外，$C_1(p)$ 和 $C_2(p)$ 见 *Pinelis (1994)*。

证明见附录 D。注意，若 $p > 4\gamma$，则有 $v_n = o(\gamma_n)$。**验证假设 (H2)**。 以下命题保证 **(H2)** 成立。

Prop命题 6.5。 考虑由 （10）给出的正则化问题，对于所有 $n \geq 0$ ，

$$\left\|\bar{S}_n^{-1}\right\| \leq 2d \max\left\{\frac{1}{\sigma}, \left\|S_0^{-1}\right\|\right\} =: C_{S,\sigma}$$

注6.1。 注意，如果 （9）对某个常数成立 $\alpha > 0$ ，并且 $\mathbb{E}\left[XX^T\right]$ 为正，在命题 6.2的假设下，对于所有 $n \geq 0$ 以及 $\sigma = 0$，有

$$\mathbb{E}\left[\|\bar{S}_n^{-1}\|^2\right] \leq \frac{1}{\alpha^2} \max\left\{2c_\beta^2\left(\frac{2\beta}{ec_3}\right)^{2\beta} + c_2^{-2}, \left((c_1 d+1)\left\|S_0^{-1}\right\|\right)^2\right\} \leq C_{S,0}^2,$$
$$\mathbb{E}\left[\|\bar{S}_n^{-1}\|^4\right] \leq \frac{1}{\alpha^4} \max\left\{2c_\beta^4\left(\frac{2\beta}{ec_3}\right)^{4\beta} + c_2^{-4}, \left((c_1 d+1)\left\|S_0^{-1}\right\|\right)^4\right\} \leq C_{S,0}^4$$

with $C_{S,0}^4 = \frac{1}{\alpha^4} \max\left\{\left(2c_\beta^2\left(\frac{2\beta}{ec_3}\right)^{2\beta} + c_2^{-2}\right)^2, \left((c_1 d+1)\left\|S_0^{-1}\right\|\right)^4\right\}$.

第一个结果

注意，可以将命题3.2 改写如下:

命题 6.6。 假设存在 $p > 2$ 使得 X 具有 $2p$ 阶矩，并且存在满足以下条件的 L_σ

$$\mathbb{E}\left[\left|\nabla_h l\left(Y, X^T\theta_\sigma\right)\right|^p \|X\|^p\right] + \sigma\theta_\sigma \leq L_\sigma^p. \tag{49}$$

那么，

$$\mathbb{E}\left[V_n^2\right] \leq \exp\left(-\frac{3c_\gamma\sigma}{4C_{GLM}}n^{1-\gamma}\right)\left(K_{1,GLM}^{(2')} + K_{1',GLM}^{(2')} \max_{1\leq k\leq n+1} v_k^{\frac{p-2}{p}} k^{\gamma-\frac{p-2}{p}\delta}\right)$$
$$+ K_{2,GLM}^{(2')} n^{-2\gamma} + K_{3,GLM}^{(2')} v_{\lfloor n/2\rfloor}^{(p-2)/p} n^{-\delta(p-2)/p} =: v_{n,GLM},$$

其中 v_n 在引理6.10中定义，$C_{S,\ \sigma}$ 在引理6.5中定义，C_{GLM} 和 $C^{(a)}$ *GLM* 在方程中定义

(48) 和(47),

$$a_{1,GLM} = C_{S,\sigma}^4 C_{GLM}^2 \left(\frac{64 L_\sigma^4 C_{GLM}^5}{\sigma^3} + 4 c_\gamma L_\sigma^4 + \frac{4 L_\sigma^4 C_{GLM}}{\sigma} \right)$$

$$a_{M,GLM} = \max \left\{ \left(\frac{4 C_{GLM} C_{GLM}^{(2)}}{\sigma} + \frac{2 C_{GLM}^2}{\sigma^2} \left(8 C_{GLM}^{(2)} + 8 C_{GLM}^{(4)} c_\gamma^2 C_{S,\sigma}^2 \right) \right) c_\gamma C_{S,\sigma}^2, \left(\frac{3\sigma}{4 C_{GLM}} \right)^2 c_\gamma \right\}$$

$$K_{1,GLM}^{(2')} = \exp \left(2 a_{M,GLM} \frac{2\gamma}{2\gamma - 1} \right) \left(\mathbb{E}\left[V_0^2 \right] + \frac{2 a_{1,GLM} c_\gamma^2}{a_{M,GLM}} \right)$$

$$K_{1',GLM}^{(2')} = \exp \left(2 a_{M,GLM} \frac{2\gamma}{2\gamma - 1} \right) \cdot \frac{4\sigma V_{p,GLM}^2}{a_{M,GLM} C_{GLM}}$$

$$K_{2,GLM}^{(2')} = \frac{2^{2\gamma+1} a_{1,GLM} C_{GLM} c_\gamma^2}{3\sigma}$$

$$K_{3,GLM}^{(2')} = \frac{2^{2+(p-2)\delta/p}}{3} V_{p,GLM}^2,$$

其中 $V^p p,\ GLM\ = e^{a_{p,\ GLM} c_\gamma^2 C_{S,\ \sigma}^2 \frac{2\gamma}{2\gamma-1}}$最大$\{1, \mathbb{E}\left[V p_{\ 0} \right]\}$
其中

$$a_{p,GLM} := p \left(\frac{2 C_{GLM}^{(2)}}{\sigma} + L_\sigma^2 \right) + 2^{p-2}(p-1) p C_{GLM}^2 \left(c_\gamma^2 C_{S,\sigma}^2 \left(8 L_\sigma^4 + \frac{32 C_{GLM}^{(4)}}{\sigma^2} \right) + \frac{4 L_\sigma^2}{\sigma} + \frac{8 C_{GLM}^{(2)}}{\sigma^2} \right)$$
$$+ 2^{p-2}(p-1) p C_{GLM}^p \left(c_\gamma^{2p-2} C_{S,\sigma}^{2p-2} \left(2^{2p-1} L_\sigma^{2p} + \frac{2^{3p-1} C_{GLM}^{(2p)}}{\sigma^2} \right) + c_\gamma^{p-2} C_{S,\sigma}^{p-2} \left(2^{2p-2} L_\sigma^{2p} + \frac{2p}{\sigma^2} \left(\frac{1}{2} + 2^{p-1/2} \sqrt{C_{GLM}^{(2p)}} \right) \right) \right).$$

验证假设 (H3)。我们在此证明**(H3)** 对一般线性模型成立。现在记

$$H_\sigma =: \mathbb{E}\left[\nabla_h^2 \ell \left(Y, \theta_\sigma^T X \right) X X^T \right] + \sigma I_d.$$

命题 6.7。假设***(GLM1)*** 和***(GLM2)*** 成立，则对所有 $n \geq 0$,

$$\mathbb{E}\left[\left\| \bar{S}_n^{-1} - H_\sigma^{-1} \right\|^2 \right] \leq \frac{4 C_{S,\sigma}^2}{\sigma^2 n} \left(L_{\nabla l}^2 \mathbb{E}\left[\|X\|^4 \right] + \frac{L_{\nabla^2 L}^2}{\sigma} \sum_{i=0}^{n-1} v_{i,GLM} + \frac{1}{n} \|S_0 - H(\theta_\sigma)\|^2 \right) + \frac{16 d^4 C_{S,\sigma}^2}{n^2} =: v_{\ell,n}$$

其中 $v_{i,\ GLM}$ 定义见命题 6.6。

现在可以完成定理5.1的证明。为此，先注意到对所有h，h',

$$\mathbb{E}\left[\left\| \nabla_h \ell \left(y, X^T h \right) X + \sigma h - \nabla_h \ell \left(y, X^T h' \right) X - \sigma h' \right\|^2 \right] \leq 2 \left(L_{\nabla l}^2 \mathbb{E}\left[\|X\|^2 \right] + \sigma^2 \right) \|h - h'\|^2.$$

然后，借助定理 3.3，可得

$$\mathbb{E}\left[\|\theta_n - \theta_\sigma\|^2 \right] \leq e^{-\frac{1}{2} c_\gamma n^{1-\gamma}} \left(K_{1,\mathrm{GLM}}^{(3)} + K_{1',\mathrm{GLM}}^{(3)} \max_{0 \leq k \leq n} (k+1)^\gamma d_{k,\mathrm{GLM}} \right)$$
$$+ n^{-\gamma} \left(2^{3+\gamma} c_\gamma \mathrm{Tr}\left(H_\sigma^{-1} \Sigma_\sigma H_\sigma^{-1} \right) + \frac{K_{2,\mathrm{GLM}}^{(3)}}{n^\gamma} + K_{2',\mathrm{GLM}}^{(3)} v_{l,n/2} \right) + d_{\lfloor n/2 \rfloor,\mathrm{GLM}},$$

with $\Sigma_\sigma := \mathbb{E}\left[\left(\nabla_h \ell\left(y, X^T\theta_\sigma\right) X + \sigma\theta_\sigma\right)\left(\nabla_h \ell\left(y, X^T\theta_\sigma\right) X + \sigma\theta_\sigma\right)^T\right]$ and since $c_\gamma 4\frac{C^{(2)}_{\mathrm{GLM}}}{\sigma^2} \geq C_{A,\mathrm{GLM}} =\geq 4c_\gamma$,

$$K^{(3)}_{1,\mathrm{GLM}} = e^{8\frac{C^{(2)}_{\mathrm{GLM}}}{\sigma^2} c_\gamma^3 \frac{2\gamma}{2\gamma-1}} \left(\mathbb{E}\left[\|\theta_0 - \theta_\sigma\|^2\right] + \frac{2\mathrm{Tr}\left(H_\sigma^{-1}\Sigma_\sigma H_\sigma^{-1}\right)}{c_\gamma} + 8\sigma^2\left(\sigma^{-4} + C^4_{S,\sigma}\right) + \frac{2L^2_\sigma v_{l,0}}{c_\gamma}\right), \tag{50}$$

$$K^{(3)}_{1',\mathrm{GLM}} = \frac{1}{4c_\gamma} e^{8\frac{C^{(2)}_{\mathrm{GLM}}}{\sigma^2} c_\gamma^3 \frac{2\gamma}{2\gamma-1}}, \quad d_{n,\mathrm{GLM}} = 8C_{\mathrm{GLM}}\sqrt{v_{n,\mathrm{GLM}} v_{l,n}} + 8\frac{L^2_{\nabla^2 L}\sigma^{-2} + 2C^{(2)}_{\mathrm{GLM}}}{\sigma^2} v_{n,\mathrm{GLM}}, \tag{51}$$

$$K^{(3)}_{2,\mathrm{GLM}} = 2^{5+2\gamma} C^{(2)}_{\mathrm{GLM}} c_\gamma \left(\sigma^{-4} + C^4_{S,\sigma}\right) c_\gamma^2, \quad K^{(3)}_{2',\mathrm{GLM}} = 2^{3+\gamma} L^2_\sigma c_\gamma. \tag{52}$$

定理 5.2 的证明。 证明完全遵循定理 4.2 的证明模式，利用假设 （**A6**） 以及引理 6.4 来计算常数 C_S，使得 （**H2**）得到满足。

□

A 技术命题的证明

A.1 命题3.1 的证明

回顾一下，

$$V_{n+1} = V_n \underbrace{-\gamma_{n+1}\left(g'_{n+1}\right)^T A_n \int_0^1 \nabla G\left(\theta_n + t\left(\theta_{n+1} - \theta_n\right)\right) dt}_{=:U_{n+1}}$$

注意，对于$a \geq 2$和x，$h \in \mathbb{R}$满足$x \geq 0$且$x + h \geq 0$时，根据泰勒展开可得

$$(x+h)^a \leq x^a + ax^{a-1}h + 2^{p-2}a(a-1)(x^{a-2}|h|^2 + |h|^a). \tag{53}$$

这意味着，对于$a = p'$、$x = V_n$和$h =$均匀分布$_{n+1}$，并在对 $\mathcal{F}_n$ 进行条件化后

$$\begin{aligned}\mathbb{E}\left[V^{p'}_{n+1}|\mathcal{F}_n\right] \leq\ & V^{p'}_n + p'V^{p'-1}_n \mathbb{E}\left[U_{n+1}|\mathcal{F}_n\right] \\ & + 2^{p'-2}p'(p'-1)\left(\mathbb{E}\left[|U_{n+1}|^2|\mathcal{F}_n\right] V^{p'-2}_n + \mathbb{E}\left[|U_{n+1}|^{p'}|\mathcal{F}_n\right]\right).\end{aligned} \tag{54}$$

由于 G 是凸的，且 ∇G 是 Lipschitz 的，

$$\begin{aligned}\mathbb{E}\left[U_{n+1}V^{p'-1}_n|\mathcal{F}_n\right] &\leq -\mathbb{E}\left[\gamma_{n+1}\left(g'_{n+1}\right)^T A_n \int_0^1 \nabla G\left(\theta_n\right) dt|\mathcal{F}_n\right] V^{p'-1}_n \\ &\quad + \mathbb{E}\left[\gamma_{n+1}\left(g'_{n+1}\right)^T A_n \int_0^1 \left(\nabla G\left(\theta_n\right) - \nabla G\left(\theta_n + t\left(\theta_{n+1} - \theta_n\right)\right)\right) dt|\mathcal{F}_n\right] V^{p'-1}_n \\ &\leq -\gamma_{n+1}\nabla G\left(\theta_n\right)^T A_n \nabla G\left(\theta_n\right) V^{p'-1}_n + \frac{L_{\nabla G}}{2}\gamma^2_{n+1}\mathbb{E}\left[\left\|g'_{n+1}\right\|^2|\mathcal{F}_n\right] \|A_n\|^2 V^{p'-1}_n. \\ &\leq -\gamma_{n+1}\nabla G\left(\theta_n\right)^T A_n \nabla G\left(\theta_n\right) V^{p'-1}_n + \gamma^2_{n+1}\beta^2_{n+1}\frac{L_{\nabla G}}{2}\left(C_1 V^{p'-1}_n + \frac{2C_2}{\mu}V^{p'}_n\right).\end{aligned}$$

根据强凸性，我们有

$$\begin{aligned}\nabla G\left(\theta_n\right)^T A_n \nabla G\left(\theta_n\right) V_n^{p'-1} &\geq \lambda_{\min}\left(A_n\right)\left\|\nabla G\left(\theta_n\right)\right\|^2 V_n^{p'-1} \\ &\geq 2\lambda_n \mu V_n^{p'} \mathbf{1}_{\lambda_{\min}(A_n)\geq\lambda_n} \\ &= 2\lambda_n \mu V_n^{p'} - 2\mathbf{1}_{\lambda_{\min}(A_n)<\lambda_n}\lambda_n \mu V_n^{p'},\end{aligned}$$

其中 $\lambda_n = \lambda_0(n+1)^\lambda$ 且 $0 \leq \lambda < \min\{\gamma - 2\beta, 1-\gamma\}$。然后应用赫尔德不等式可得

$$\begin{aligned}\mathbb{E}\left[\nabla G\left(\theta_n\right)^T A_n \nabla G\left(\theta_n\right)\right] \geq & 2\lambda_n \mu \mathbb{E}\left[V_n^{p'}\right] - 2\lambda_n \mu \mathbb{E}[V_n^p]^{p'/p}\left(\mathbb{P}\left[\lambda_{\min}(A_n) < \lambda_n\right]\right)^{\frac{p-p'}{p}} \\ \geq & 2\lambda_n \mu \mathbb{E}\left[V_n^{p'}\right] - 2\lambda_n \mu V_p^{p'}\left(\mathbb{P}\left[\lambda_{\min}(A_n) < \lambda_n\right]\right)^{\frac{p-p'}{p}},\end{aligned}$$

其中 $V_p^p \geq \sup_{n\geq 0}\mathbb{E}[V_n^p]$ 由引理 6.1给出。然后，假设 **(H1a)** 给出 $\mathbb{P}\left[\lambda_{\min}\left(A_n\right) < \lambda_n\right] \leq \bar{v}_{n+1}(n+1)^{-\delta-q\lambda} := \bar{v}_n$，因此最终得到

$$\begin{aligned}&\mathbb{E}\left[U_{n+1}V_n^{p'-1}\right] \\ &\leq -2\gamma_{n+1}\lambda_n \mathbb{E}\left[\mu V_n^{p'}\right] + 2\lambda_n \gamma_{n+1}\mu V_p^{p'}\bar{v}_n^{\frac{p-p'}{p}} + \gamma_{n+1}^2\beta_{n+1}^2\frac{L_{\nabla G}}{2}\left(C_1\mathbb{E}\left[V_n^{p'-1}\right] + \frac{2C_2}{\mu}\mathbb{E}\left[V_n^{p'}\right]\right).\end{aligned} \tag{55}$$

此外，由于 ∇G 是 $L_{\nabla G}$-Lipschitz 的，因此有

$$\begin{aligned}\left\|\int_0^1 \nabla G\left(\theta_n + t\left(\theta_{n+1} - \theta_n\right)\right) dt\right\| &\leq L_{\nabla G}\int_0^1 \left(\left\|\theta_n - \theta\right\| + t\left\|\theta_{n+1} - \theta_n\right\|\right) dt \\ &\leq L_{\nabla G}\left(\left\|\theta_n - \theta\right\| + \frac{1}{2}\gamma_{n+1}\left\|A_n\right\|\left\|g'_{n+1}\right\|\right).\end{aligned} \tag{56}$$

因此，利用 **(H1b)** 以及 G 的强凸性可得

$$\begin{aligned}\mathbb{E}\left[\left|U_{n+1}\right|^{p'}|\mathcal{F}_n\right] &\leq L_{\nabla G}^{p'}\left\|A_n\right\|^{p'}\gamma_{n+1}^{p'}\mathbb{E}\left[\left\|g'_{n+1}\right\|^{p'}\left(2^{p'-1}\left\|\theta_n - \theta\right\|^{p'} + 2^{-1}\gamma_{n+1}^{p'}\left\|A_n\right\|^{p'}\left\|g'_{n+1}\right\|^{p'}\right)|\mathcal{F}_n\right] \\ &\leq \frac{L_{\nabla G}^{p'}}{2}\gamma_{n+1}^{p'}\beta_{n+1}^{p'}\left(2^{p'}\left(C_1^{(p'/2)}\frac{2^{p'/2}V_n^{p'/2}}{\mu^{p'/2}} + C_2^{(p'/2)}\frac{2^{p'}V_n^{p'}}{\mu^{p'}}\right)\right. \\ &\qquad \left. + \gamma_{n+1}^{p'}\beta_{n+1}^{p'}\left(C_1^{(p')} + C_2^{(p')}\frac{2^{p'}V_n^{p'}}{\mu^{p'}}\right)\right)\end{aligned}$$

将后一个不等式特化为 $p' = 2$，于是得到（回顾不等式（14））

$$\begin{aligned}&\mathbb{E}\left[\left|U_{n+1}\right|^2|\mathcal{F}_n\right]V_n^{p'-2} \\ &\leq \frac{L_{\nabla G}^2}{2}\gamma_{n+1}^2\beta_{n+1}^2\left(2^{p'}\left(C_1\frac{2V_n}{\mu} + C_2\frac{2^2V_n^2}{\mu^2}\right) + \gamma_{n+1}^2\beta_{n+1}^2\left(C_1' + C_2'\frac{4V_n^2}{\mu^2}\right)\right)V_n^{p'-2},\end{aligned}$$

因此

$$
\begin{aligned}
&\mathbb{E}\left[|U_{n+1}|^{p'}|\mathcal{F}_n\right]+\mathbb{E}\left[|U_{n+1}|^2|\mathcal{F}_n\right]V_n^{p'-2}\\
\leq&\frac{L_{\nabla G}^{p'}C_1^{(p')}}{2}\gamma_{n+1}^{2p'}\beta_{n+1}^{2p'}+\frac{2^{3p'/2-1}L_{\nabla G}^{p'}C_1^{(p'/2)}}{\mu^{p'/2}}\gamma_{n+1}^{p'}\beta_{n+1}^{p'}V_n^{p'/2}+\frac{2^{p'}L_{\nabla G}^2C_1}{\mu}\gamma_{n+1}^2\beta_{n+1}^2V_n^{p'-1}\\
&+\frac{L_{\nabla G}^2C_1'}{2}\gamma_{n+1}^4\beta_{n+1}^4V_n^{p'-2}+V_n^{p'}\left(\frac{2^{2p'-1}L_{\nabla G}^{p'}C_2^{(p'/2)}}{\mu^{p'}}\gamma_{n+1}^{p'}\beta_{n+1}^{p'}+\frac{2^{p'-1}L_{\nabla G}^{p'}C_2^{(p')}}{\mu^{p'}}\gamma_{n+1}^{2p'}\beta_{n+1}^{2p'}\right.\\
&\qquad\left.+\frac{2^{p'+1}C_2L_{\nabla G}^2}{\mu^2}\gamma_{n+1}^2\beta_{n+1}^2+\frac{2C_2'L_{\nabla G}^2}{\mu^2}\gamma_{n+1}^4\beta_{n+1}^4\right).
\end{aligned}
$$

将后一个不等式与式（55）代入式（54），可得

$$
\mathbb{E}\left[V_{n+1}^{p'}\right]\leq\mathbb{E}\left[V_n^{p'}\right]-2p'\mu\gamma_{n+1}\lambda_n\mathbb{E}\left[V_n^{p'}\right]+2p'\lambda_n\gamma_{n+1}\mu V_p^{p'}\bar{v}_n^{\frac{p-p'}{p}}+\mathbb{E}\left[P\left(\gamma_{n+1}^2\beta_{n+1}^2,V_n\right)\right]
$$

其中 $P(x,y)=A_0x^{p'}+A_{p'}/2x^{p'}/^2y^{p'}/^2+A^{p'-1}xy^{p'-1}+A^{p'-2}x^2y^{p'-2}+A^{p'}xy^{p'}$，其中

$$
A_0=2^{p'-3}p'(p'-1)L_{\nabla G}^{p'}C_1^{(p')},\ A_{p'/2}=\frac{2^{5p'/2-3}p'(p'-1)L_{\nabla G}^{p'}C_1^{(p'/2)}}{\mu^{p'/2}},
$$

$$
A_{p'-1}=p'\frac{L_{\nabla G}}{2}+\frac{2^{2p'-2}p'(p'-1)L_{\nabla G}^2C_1}{\mu},\ A_{p'-2}=2^{p'-3}p'(p'-1)L_{\nabla G}^2C_1',
$$

and

$$
\begin{aligned}
A_{p'}=\frac{p'L_{\nabla G}C_2}{\mu}+p'(p'-1)\Bigg(&\frac{2^{3p'-3}L_{\nabla G}^{p'}C_2^{(p'/2)}}{\mu^{p'}}c_\gamma^{p'-2}c_\beta^{p'-2}+\frac{2^{2p'-3}L_{\nabla G}^{p'}C_2^{(p')}}{\mu^{p'}}c_\gamma^{2p'-2}c_\beta^{2p'-2}\\
&+\frac{2^{2p'-1}C_2L_{\nabla G}^2}{\mu^2}+\frac{2^{p'-1}C_2'L_{\nabla G}^2}{\mu^2}c_\gamma^2c_\beta^2\Bigg).
\end{aligned}
$$

现在应用杨氏不等式，可得 $a^ib^{p'-i}\leq ia^{p'}p'+(p'-i)b^{p'}p'$，对于 $0<i<p'$ 和 a、$b\geq 0$，对任意 $t>0$ 和 $i\in\{1,2,p'/2\}$ 均成立

$$
A_{p-i}x^iy^{p'-i}=\left(\frac{A_i^{1/i}x}{(t\lambda_n\gamma_n)^{\frac{p'-i}{p'}}}\right)^i\left((t\lambda_n\gamma_n)^{\frac{i}{p'}}y\right)^{p'-i}\leq\frac{iA_i^{\frac{p'}{i}}x^{p'}}{(t\lambda_n\gamma_n)^{\frac{p'-i}{i}}}+\frac{(p'-i)t\lambda_n\gamma_ny^{p'}}{p'},
$$

因此，利用后一个不等式，并令 $t=p3\ '^2\mu\ '-(pi)\mu$，其中 $i\in\{1,2,p'/2\}$，并利用

$$
\begin{aligned}
\frac{\gamma_{n+1}^{2p'}\beta_n^{2p'}}{(\gamma_{n+1}\lambda)^{\frac{p'}{i}-1}}&=(\gamma_{n+1}\lambda_n)c_\gamma^{\frac{2i-1}{i}p'}c_\beta^{2p'}\lambda_0^{-\frac{p'}{i}}(n+1)^{-\frac{(2i-1)p'}{i}\gamma+2p'\beta+\frac{p'}{i}\lambda}\\
&\leq(\gamma_{n+1}\lambda_n)c_\gamma^{\frac{2i-1}{i}p'}c_\beta^{2p'}\lambda_0^{-\frac{p'}{i}}(n+1)^{-p'(\gamma-2\beta-\lambda)}
\end{aligned}
$$

可得

$$\mathbb{E}\left[P\left(\gamma_{n+1}^2\beta_{n+1}^2, V_n\right)\right] \leq L\left(\frac{p'\mu}{2}\lambda_n\gamma_{n+1}\right)(n+1)^{-p'(\gamma-2\beta-\lambda)} + \left(p'\mu(\lambda_n\gamma_{n+1}) + A_{p'}(\gamma_{n+1}\beta_{n+1})^2\right)\mathbb{E}\left[V_n^{p'}\right]$$

with

$$L = \frac{c_\gamma^{2p'-1}c_\beta^{2p'}}{\lambda_0}A_0 + \frac{3c_\gamma^{2(p'-1)}c_\beta^{2p'}\lambda_0^{-2}\sqrt{A_{p'}}}{4\mu}A_{p'/2} + \frac{2c_\gamma^{\frac{3}{2}p'}c_\beta^{2p'}\lambda_0^{-\frac{p'}{2}}}{\left(\frac{p'^2\mu}{3(p'-2)}\right)^{\frac{p'-2}{2}}}A_2^{p'/2} + \frac{c_\gamma^{p'}c_\beta^{2p'}\lambda_0^{-p'}}{\left(\frac{p'^2\mu}{3(p'-1)}\right)^{p'-1}}A_1.$$

综合前面的不等式并取期望，可得

$$\begin{aligned}\mathbb{E}\left[V_{n+1}^{p'}\right] \leq &\left(1 - p'\mu\gamma_{n+1}\lambda_n + \frac{A_{p'}c_\gamma c_\beta^2}{\lambda_0}(n+1)^{-\gamma+2\beta+\lambda}\gamma_{n+1}\lambda_n\right)\mathbb{E}\left[V_n^{p'}\right] \\ &+ \lambda_n\gamma_{n+1}\left(2p'\mu V_p^{p'}\bar{v}_n^{\frac{p-p'}{p}} + \frac{Lp'\mu}{2}(n+1)^{-p'(\gamma-2\beta-\lambda)}\right).\end{aligned}$$

然后，回想 $\bar{v}_n = v_{n+1}(n+1)^{-\delta-q\lambda}$ 并利用命题 6.1 可得

$$\begin{aligned}\mathbb{E}\left[V_n^{p'}\right] \leq &\exp\left(-\frac{c_\gamma p'\mu\lambda_0}{2}n^{1-(\lambda+\gamma)}(1-\varepsilon(n))\right)\left(K_1^{(1')} + K_{1'}^{(1')}\max_{1\leq k\leq n+1}k^{\gamma-2\beta-\lambda-\frac{p-p'}{p}(\delta+q\lambda)}v_k^{\frac{p-p'}{p}}\right) \\ &+ K_2^{(1')}n^{-p'(\gamma-2\beta-\lambda)} + K_3^{(1')}v_{\lfloor n/2\rfloor}^{\frac{p-p'}{p}}(n+1)^{-\frac{p-p'}{p}(\delta+q\lambda)},\end{aligned}$$

with

$$\varepsilon(n) = \frac{4C'_M n^{-1+\lambda+\gamma}}{\mu p'\lambda_0}\left(1 + \frac{n^{(1+2\beta-2\gamma)^+}}{|2\gamma-2\beta-1|}\right), \tag{57}$$

and

$$K_1^{(1')} = \left(\mathbb{E}\left[V_0\right] + \frac{p'\mu L}{C'_M}\right), \quad K_{1'}^{(1')} = \frac{4p'\mu V_p^{p'}}{C'_M}, \tag{58}$$

其中

$$C'_M = \max\left\{\frac{A_{p'}c_\gamma c_\beta^2}{\lambda_0}, \left(\frac{\mu p'\lambda_0}{8}\right)^{\frac{2\gamma-2\beta}{\gamma+\lambda}}c_\gamma^{\frac{\gamma-2\beta-\lambda}{\gamma+\lambda}}\right\}, \tag{59}$$

and

$$K_2^{(1')} = 2^{p'(\gamma-2\beta-\lambda)}L,\ K_3^{(1')} = 2^{2+\frac{p-p'}{p}(\delta+q\lambda)}V_p^{p'}. \tag{60}$$

其中 V^p 由引理6.1给出。

A.2 命题3.2的证明

注意，借助 G 的泰勒展开，可得

$$\begin{aligned} V_{n+1} &= V_n + (\theta_{n+1}-\theta_n)^T \int_0^1 \nabla G\left(\theta_n + t\left(\theta_{n+1}-\theta_n\right)\right) dt \\ &= V_n - \gamma_{n+1}\left(g'_{n+1}\right)^T A_n \int_0^1 \nabla G\left(\theta_n + t\left(\theta_{n+1}-\theta_n\right)\right) dt. \end{aligned}$$

然后，利用（56）可得

$$\begin{aligned} V_{n+1}^2 \leq V_n^2 &\overbrace{- 2\gamma_{n+1} V_n \left(g'_{n+1}\right)^T A_n \int_0^1 \nabla G\left(\theta_n + t\left(\theta_{n+1}-\theta_n\right)\right) dt}^{:=(\star)} \\ &\underbrace{+ L_{\nabla G}^2 \|A_n\|^2 \left\|g'_{n+1}\right\|^2 \gamma_{n+1}^2 \left(2\|\theta_n-\theta\|^2 + \frac{1}{2}\gamma_{n+1}^2 \|A_n\|^2 \left\|g'_{n+1}\right\|^2\right)}_{:=(\star\star)} \end{aligned}$$

现在我们对 $(\star)$ 和 $(\star\star)$进行界定。首先，由于假设 **(H1)** 且 $\|\theta_n-\theta\|^2 \leq \frac{2}{\mu}V_n$，可得

$$\begin{aligned} \mathbb{E}\left[(\star\star)|\mathcal{F}_n\right] &\leq \frac{4L_{\nabla G}^2 C_1}{\mu}\gamma_{n+1}^2 \|A_n\|^2 V_n + \frac{8L_{\nabla G}^2 C_2}{\mu^2}\|A_n\|^2 \gamma_{n+1}^2 V_n^2 \\ &\quad + \frac{1}{2}L_{\nabla G}^2 C_1' \gamma_{n+1}^4 \|A_n\|^4 + \frac{2L_{\nabla G}^2 C_2'}{\mu^2}\gamma_{n+1}^4 \|A_n\|^4 V_n^2 \\ &\leq \frac{8L_{\nabla G}^4 C_1^2}{\mu^3\lambda_0}\gamma_{n+1}^3 \|A_n\|^4 + \frac{1}{2}\mu\lambda_0\gamma_{n+1}V_n^2 + \frac{L_{\nabla G}^2 C_1'}{2}\gamma_{n+1}^4\|A_n\|^4 \\ &\quad + \frac{2L_{\nabla G}^2}{\mu^2}\left(4C_2 + C_2' c_\gamma^2 c_\beta^2\right)\gamma_{n+1}^2\beta_{n+1}^2 V_n^2 \end{aligned}$$

然后，在假设**(H2b)**下取期望，

$$\begin{aligned} \mathbb{E}\left[(\star\star)\right] &\leq \frac{8L_{\nabla G}^4 C_1^2}{\mu^3\lambda_0}\gamma_{n+1}^3 C_S^4 + \frac{\mu\lambda_0}{2}\gamma_{n+1}\mathbb{E}\left[V_n^2\right] + \frac{L_{\nabla G}^2 C_1'}{2}\gamma_{n+1}^4 C_S^4 \\ &\quad + \frac{2L_{\nabla G}^2}{\mu^2}\left(4C_2 + C_2' c_\gamma^2 c_\beta^2\right)\gamma_{n+1}^2\beta_{n+1}^2\mathbb{E}\left[V_n^2\right]. \end{aligned}$$

此外，由于 ∇G 是 $L_{\nabla G}$-Lipschitz 的，可以验证

$$\begin{aligned} \left\|\int_0^1 \nabla G\left(\theta_n + t\left(\theta_{n+1}-\theta_n\right)\right) - \nabla G\left(\theta_n\right) dt\right\| &\leq L_{\nabla G}\int_0^1 t dt \gamma_{n+1}\|A_n\|\left\|g'_{n+1}\right\| \\ &\leq \frac{L_{\nabla G}}{2}\gamma_{n+1}\|A_n\|\left\|g'_{n+1}\right\|. \end{aligned}$$

于是，有

$$\begin{aligned}\mathbb{E}\left[(\star)|\mathcal{F}_n\right] &\geq 2\gamma_{n+1}\nabla G\left(\theta_n\right)^T A_n \nabla G\left(\theta_n\right) V_n - L_{\nabla G}\gamma_{n+1}^2 \left\|A_n\right\|^2 \mathbb{E}\left[\left\|g'_{n+1}\right\|^2 |\mathcal{F}_n\right] V_n \\ &\geq 2\gamma_{n+1}\nabla G\left(\theta_n\right)^T A_n \nabla G\left(\theta_n\right) V_n - L_{\nabla G}\gamma_{n+1}^2 \left\|A_n\right\|^2 C_1 V_n - \frac{2L_{\nabla G}C_2}{\mu}\gamma_{n+1}^2 \left\|A_n\right\|^2 V_n^2 \\ &\geq 2\gamma_{n+1}\nabla G\left(\theta_n\right)^T A_n \nabla G\left(\theta_n\right) V_n - \frac{C_1^2 L_{\nabla G}^2}{2\mu\lambda_0}\gamma_{n+1}^3 \left\|A_n\right\|^4 - \frac{\mu\lambda_0\gamma_{n+1}}{2}V_n^2 - \frac{2L_{\nabla G}C_2}{\mu}\gamma_{n+1}^2\beta_{n+1}^2 V_n^2.\end{aligned}$$

此外，借助不等式（16），可得

$$\gamma_{n+1}\nabla G\left(\theta_n\right)^T A_n \nabla G\left(\theta_n\right) V_n \geq 2\lambda_0\mu\gamma_{n+1}V_n^2 - 2\lambda_0\mu\gamma_{n+1}\mathbf{1}_{A_n<\lambda_0}V_n^2.$$

那么，结合霍尔德不等式以及当 **(H1a)** 对$t = 1$成立时，可得

$$\mathbb{E}\left[(\star)\right] \geq \frac{7}{2}\lambda_0\mu\gamma_{n+1}V_n^2 - 4\lambda_0\mu\gamma_{n+1}\bar{v}_n^{(p-2)/p}V_p^2 - \frac{C_1^2L_{\nabla G}^2}{2\mu\lambda_0}\gamma_{n+1}^3C_S^4 - \frac{2L_{\nabla G}C_2}{\mu}\gamma_{n+1}^2\beta_{n+1}^2\mathbb{E}\left[V_n^2\right]$$

其中，V^p 由引理6.1 定义，且 $\bar{v}_n := v_n(n+1)^{-\delta}$ 是由**(H1a)** 给出的关于$\mathbb{P}\left[\lambda_{\min}\left(A_n\right) \leq \lambda_0\right]$的上界。令

$$a_M := \max\left\{\left(\frac{2L_{\nabla G}C_2}{\mu} + \frac{2L_{\nabla G}^2}{\mu^2}\left(4C_2 + C_2'c_\gamma^2c_\beta^2\right)\right)c_\gamma c_\beta^2, \left(\frac{3\lambda_0\mu}{2}\right)^{\frac{2\gamma-2\beta}{\gamma}} c_\gamma^{\frac{\gamma-2\beta}{\gamma}}\right\}, \tag{61}$$

可得

$$\begin{aligned}\mathbb{E}\left[V_{n+1}^2\right] \leq &\left(1 - 3\lambda_0\mu\gamma_{n+1} + a_M n^{2\beta-\gamma}\gamma_{n+1}\right)\mathbb{E}\left[V_n\right] + 4\lambda_0\mu\gamma_{n+1}\bar{v}_n^{(p-2)/p}V_p^2 \\ &+ \underbrace{C_S^4L_{\nabla G}^2\left(\frac{8L_{\nabla G}^4C_1^2}{\mu^3\lambda_0} + \frac{C_1'c_\gamma}{2} + \frac{C_1^2}{2\mu\lambda_0}\right)}_{=:a_1}\gamma_{n+1}^3\end{aligned} \tag{62}$$

应用命题 6.1，可得（采用与定理证明中类似的推导过程 3.1）

$$\begin{aligned}&\mathbb{E}\left[V_n^2\right] \leq \exp\left(-\frac{3}{2}c_\gamma\lambda_0\mu n^{1-\gamma}\right)\exp\left(2a_M\frac{2\gamma-2\beta}{2\gamma-2\beta-1}\right) \\ &\cdot\left(\mathbb{E}\left[V_0^2\right] + \frac{2a_1c_\gamma^2}{a_M} + \frac{8\lambda_0\mu c_\gamma V_p^{2/p}}{a_M}\max_{1\leq k\leq n+1} v_k^{\frac{p-2}{p}}k^{\gamma-\frac{(p-2)}{p}\delta}\right) + \frac{2^{2\gamma}a_1c_\gamma^2}{3\lambda_0\mu}n^{-2\gamma} + \frac{4}{3}V_p^2\bar{v}_{\lfloor n/2\rfloor}^{(p-2)/p}.\end{aligned}$$

其中 V^p 由引理 6.1 给出，且 $\bar{v}_{\lfloor n/2\rfloor} \leq v_{n/2}2^\delta(n+1)^{-\delta}$。令

$$K_1^{(2')} = \exp\left(2a_M\frac{2\gamma-2\beta}{2\gamma-2\beta-1}\right)\left(\mathbb{E}\left[V_0^2\right] + \frac{2a_1c_\gamma^2}{a_M}\right), \tag{63}$$

$$K_{1'}^{(2')} = \exp\left(2a_M\frac{2\gamma-2\beta}{2\gamma-2\beta-1}\right)\cdot\frac{8\lambda_0\mu V_p^2}{a_M}, \tag{64}$$

其中 a_M 由（61）给出， a_1 由（62）给出，且 V^p 由引理 6.1给出，并且

$$K_2^{(2')} = \frac{2^{2\gamma} a_1 c_\gamma^2}{3\lambda_0 \mu}, \quad K_3^{(2')} = \frac{2^{2+(p-2)\delta/p}}{3} V_p^2, \tag{65}$$

最终得到

$$\mathbb{E}\left[V_n^2\right] \leq \exp\left(-\frac{3}{2} c_\gamma \lambda_0 \mu n^{1-\gamma}\right) \left(K_1^{(2')} + K_{1'}^{(2')} \max_{1\leq k\leq n+1} v_k^{\frac{p-2}{p}} k^{\gamma - \delta\frac{p-2}{p}}\right) + K_2^{(2')} n^{-2\gamma} + K_3^{(2')} v_{\lfloor n/2 \rfloor}^{(p-2)/p} n^{-\delta(p-2)/p}.$$

B 技术引理的证明

B.1 引理6.1 (新版本)

注意到，由于证明过程类似，我们只对$p > 2$的情形进行证明；对于$p = 2$的情形，如果证明中存在差异，将通过备注予以说明。

借助泛函G的泰勒展开，可得

$$V_{n+1} = V_n - \gamma_{n+1} \left(g'_{n+1}\right)^T A_n \int_0^1 \nabla G\left(\theta_n + t\left(\theta_{n+1} - \theta_n\right)\right) dt.$$

然后，应用不等式

$$\begin{aligned}(a+h)^p &\leq a^p + pa^{p-1}h + \frac{p(p-1)h^2}{2} \max(1, 2^{p-3})(a^{p-2} + |h|^{p-2}) \\ &\leq a^p + pa^{p-1}h + p(p-1)2^{p-3}h^2(a^{p-2} + |h|^{p-2})\end{aligned}$$

对于 a、$a + h \geq 0$ 到 $a = V_n$ 以及 $h = -\gamma_{n+1}\left(g'_{n+1}\right)^T A_n \int_0^1 (1-t)\nabla G\left(\theta_n + t(\theta_{n+1} - \theta_n)\right) dt$，有

$$\begin{aligned}V_{n+1}^p \leq V_n^p &- p\gamma_{n+1}\left(g'_{n+1}\right)^T A_n \int_0^1 \nabla G\left(\theta_n + t\left(\theta_{n+1} - \theta_n\right)\right) dt V_n^{p-1} \\ &+ 2^{p-3}p(p-1)\left\|\gamma_{n+1}\left(g'_{n+1}\right)^T A_n \int_0^1 \nabla G\left(\theta_n + t\left(\theta_{n+1} - \theta_n\right)\right) dt\right\|^2 V_n^{p-2} \\ &+ 2^{p-3}p(p-1)\left\|\gamma_{n+1}\left(g'_{n+1}\right)^T A_n \int_0^1 \nabla G\left(\theta_n + t\left(\theta_{n+1} - \theta_n\right)\right) dt\right\|^p\end{aligned}$$

注 B.1. 注意，在 $p = 2$ 的情况下，有

$$(a+h)^2 = a^2 + 2ah + h^2 = a^p + 2a^{p-1}h + p(p-1)2^{p-3}h^2|h|^{p-2}$$

前一个不等式右侧的最后一项可视为等于0。

回想一下，由于 ∇G 是$L_{\nabla G}$-Lipschitz 连续的，因此有

$$\left\| \int_0^1 (1-t)\nabla G\left(\theta_n + t\left(\theta_{n+1}-\theta_n\right)\right) dt \right\| \leq L_{\nabla G}\left(\|\theta_n-\theta\| + \gamma_{n+1}\|A_n\|\left\|g'_{n+1}\right\|\right),$$

这意味着

$$\begin{aligned}
V_{n+1}^p \leq V_n^p &- p\gamma_{n+1}\left(g'_{n+1}\right)^T A_n \int_0^1 \nabla G\left(\theta_n + t\left(\theta_{n+1}-\theta_n\right)\right) dt V_n^{p-1} \Big\} =: (*) \\
&+ 2^{p-2}p(p-1)L_{\nabla G}^2\gamma_{n+1}^2\left\|g'_{n+1}\right\|^2\|A_n\|^2\left(\|\theta_n-\theta\|^2 + \gamma_{n+1}^2\|A_n\|^2\left\|g'_{n+1}\right\|^2\right)V_n^{p-2}\Big\} =: (**) \\
&+ 2^{p-2}p(p-1)L_{\nabla G}^p\gamma_{n+1}^p\left\|g'_{n+1}\right\|^p\|A_n\|^p\left(\|\theta_n-\theta\|^p + \gamma_{n+1}^p\|A_n\|^p\left\|g'_{n+1}\right\|^p\right)\Big\} =: (***)
\end{aligned}$$

此外，有

$$\begin{aligned}
(*) &= -p\gamma_{n+1}\left(g'_{n+1}\right)^T A_n \int_0^1 \nabla G\left(\theta_n + t\left(\theta_{n+1}-\theta_n\right)\right) dt V_n^{p-1} \\
&= -p\gamma_{n+1}\left(g'_{n+1}\right)^T A_n \nabla G\left(\theta_n\right) V_n^{p-1} \\
&- p\gamma_{n+1}\left(g'_{n+1}\right)^T A_n \int_0^1 \left(\nabla G\left(\theta_n + t\left(\theta_{n+1}-\theta_n\right)\right) - \nabla G\left(\theta_n\right)\right) dt V_n^{p-1}
\end{aligned}$$

由于 A_n 为正，且由于 ∇G 是 $L_{\nabla G}$-Lipschitz，取条件期望可得：对所有 a、$b \geq 0$，$ab \leq \frac{1}{p}a^p + \frac{p-1}{p}b^{p/(p-1)}$，并借助假设**(H1a)**，——

$$\begin{aligned}
\mathbb{E}\left[(*)|\mathcal{F}_n\right] &\leq -p\gamma_{n+1}\nabla G\left(\theta_n\right)^T A_n \nabla G\left(\theta_n\right) V_n^{p-1} + \frac{p}{2}\gamma_{n+1}^2\|A_n\|^2\,\mathbb{E}\left[\left\|g'_{n+1}\right\|^2|\mathcal{F}_n\right]V_n^{p-1} \\
&\leq -p\gamma_{n+1}\lambda_{\min}\left(A_n\right)\|\nabla G\left(\theta_n\right)\|^2 V_n^{p-1} + \frac{p}{2}\beta_{n+1}^2\gamma_{n+1}^2\left(C_1 + C_2\|\theta_n-\theta\|^2\right)V_n^2 \\
&\leq -p\mu\gamma_{n+1}\lambda_{\min}\left(A_n\right)V_n^p + \frac{pC_2}{\mu}\beta_{n+1}^2\gamma_{n+1}^2 V_n^p + \frac{pC_1}{2}\beta_{n+1}^2\gamma_{n+1}^2 V_n^{p-1} \\
&\leq -p\mu\gamma_{n+1}\lambda'_{n+1}\mathbf{1}_{\gamma\leq 1/2}V_n^p + \left(\frac{pC_2}{\mu} + \frac{C_1(p-1)}{2}\right)\beta_{n+1}^2\gamma_{n+1}^2 V_n^p + \frac{C_1}{2}\beta_{n+1}^2\gamma_{n+1}^2,
\end{aligned}$$

其中 $\lambda'_n = \lambda'_0 n^{-\lambda'}$。我们还在第一个不等式中使用了假设**(A1)**，并在第三个不等式中利用了$\|\theta_n-\theta\|^2 \leq \frac{2}{\mu}V_n \leq \frac{2}{\mu^2}\|\nabla G\left(\theta_n\right)\|^2$ 这一事实。基于相同的理由，可得

$$\begin{aligned}
\mathbb{E}\left[(**)|\mathcal{F}_n\right] &\leq 2^{p-2}p(p-1)L_{\nabla G}^2\left(\gamma_{n+1}^4\beta_{n+1}^4\left(C'_1 + \frac{4C'_2}{\mu^2}V_n^2\right) + \gamma_{n+1}^2\beta_{n+1}^2\left(\frac{2C_1}{\mu}V_n + \frac{4C_2}{\mu^2}V_n^2\right)\right)V_n^{p-2} \\
&\leq 2^{p-2}(p-1)L_{\nabla G}^2\gamma_{n+1}^4\beta_{n+1}^4\left(2C'_1 + \left((p-2)C'_1 + \frac{4pC'_2}{\mu^2}\right)V_n^p\right) \\
&+ 2^{p-2}(p-1)L_{\nabla G}^2\gamma_{n+1}^2\beta_{n+1}^2\left(\frac{2C_1}{\mu} + \left(\frac{2(p-1)C_1}{\mu} + \frac{4pC_2}{\mu^2}\right)V_n^p\right)
\end{aligned}$$

同理，由于假设**(A1″)** 和**(H1)**，有

$$\begin{aligned}\mathbb{E}\left[(***)|\mathcal{F}_n\right] \leq\ & 2^{p-2}p(p-1)L_{\nabla G}^p\gamma_{n+1}^{2p}\beta_{n+1}^{2p}\left(C_1^{(p)}+\frac{2^pC_2^{(p)}}{\mu^p}V_n^p\right)\\ &+2^{p-2}p(p-1)L_{\nabla G}^p\gamma_{n+1}^{p}\beta_{n+1}^{p}\left(\frac{1}{2}C_1^{(p)}+\frac{2^p}{\mu^p}\left(\frac{1}{2}+\sqrt{C_2^{(p)}}\right)V_n^p\right)\end{aligned}$$

对 $\mathbb{E}\left[(*)|\mathcal{F}_n\right]+\mathbb{E}\left[(**)|\mathcal{F}_n\right]+\mathbb{E}\left[(***)|\mathcal{F}_n\right]$取期望，并应用后面的不等式，可得

$$\mathbb{E}\left[V_{n+1}^p\right]\leq\max\left\{\mathbb{E}\left[V_n^p\right],1\right\}\left(1-p\mu\lambda'_{n+1}\gamma_{n+1}\mathbf{1}_{\gamma\leq1/2}+a_p\gamma_{n+1}^2\beta_{n+1}^2\right)$$

with

$$\begin{aligned}a_p:=\ & p\left(\frac{C_2}{\mu}+\frac{C_1}{2}\right)+2^{p-2}(p-1)pL_{\nabla G}^2\left(c_\gamma^2c_\beta^2\left(C_1'+\frac{4C_2'}{\mu^2}\right)+\frac{2C_1}{\mu}+\frac{4C_2}{\mu^2}\right)\\ &+2^{p-2}(p-1)pL_{\nabla G}^p\left(c_\gamma^{2p-2}c_\beta^{2p-2}\left(C_1^{(p)}+\frac{2^pC_2^{(p)}}{\mu^2}\right)+c_\gamma^{p-2}c_\beta^{p-2}\left(\frac{1}{2}C_1^{(p)}+\frac{2p}{\mu^2}\left(\frac{1}{2}+\sqrt{C_2^{(p)}}\right)\right)\right).\end{aligned}\tag{66}$$

注 B.2。 注意，在 $p=2$ 的情况下，有

$$a_2=C_1+\frac{2C_2}{\mu}+\frac{4L_{\nabla G}^2}{\mu}C_1+\frac{8L_{\nabla G}^2C_2}{\mu^2}+2L_{\nabla G}^2C_1'c_\gamma^2c_\beta^2+\frac{8L_{\nabla G}^2C_2'}{\mu^2}c_\gamma^2c_\beta^2\tag{67}$$

若 $\gamma>1/2$，则通过求和，

$$\mathbb{E}\left[V_n^p\right]\leq e^{a_pc_\gamma^2c_\beta^2\frac{2\gamma-2\beta}{2\gamma-2\beta-1}}\max\left\{1,\mathbb{E}\left[V_0^p\right]\right\}=:V_p^p.$$

若 $\gamma\leq1/2$，则令n_0为满足 $\gamma_{n+1}^2\beta_{n+1}^2a_p>p\mu\lambda'_n\gamma_{n+1}$的最小整数。记$\lambda'_n=\lambda'_0(n+1)^{-\lambda'}$，我们有$n_0=\left\lfloor\left(\frac{c_\gamma c_\beta^2 p}{\mu\lambda'_0}\right)^{\frac{1}{\gamma-2\beta-\lambda'}}\right\rfloor ap$。于是，

$$\begin{aligned}\mathbb{E}\left[V_n^p\right]\leq&\exp\left(\sum_{n=0}^{n_0}-p\mu\lambda'_n\gamma_{n+1}+a_p\gamma_{n+1}^2\beta_{n+1}^2\right)\max\left\{1,\mathbb{E}\left[V_0^p\right]\right\}\\ \leq&\exp\left(-p\mu\lambda'_0c_\gamma\left(1+\frac{1+\left(\frac{c_\gamma c_\beta^2a_p}{p\mu\lambda'_0}\right)^{\frac{1-\gamma-\lambda'}{\gamma-2\beta-\lambda'}}}{1-\gamma-\lambda'}\right)+c_\gamma^2c_\beta^2a_p\left(1+\frac{1+\left(\frac{c_\gamma c_\beta^2a_p}{p\mu\lambda'_0}\right)^{\frac{1-2\gamma+2\beta}{\gamma-2\beta-\lambda'}}}{1-2\gamma+2\beta}\right)\right)=:V_p^p.\end{aligned}$$

B.2 引理证明6.2

Recall that $(A_n)_{kk'} = \max\left\{\min\left\{c_\beta n^\beta, \left(\overline{A_n}\right)_{kk'}\right\}, \lambda_0' n^{-\lambda'} \mathbf{1}_{\gamma \leq 1/2}\right\}$ with $\left(\overline{A_n}\right)_{kk'} = \frac{\delta_{kk'}}{\sqrt{\frac{1}{n+1}\left(a_k + \sum_{i=0}^{n-1} \left(\nabla_h g\left(X_{i+1}, \theta_i\right)_k\right)^2\right)}}$.
Since $\lambda_{\min}(A_n) \geq \lambda_{\min}\left(\overline{A_n}\right)$ on the event $\left\{\lambda_{\min}\left(\overline{A_n}\right) < c_\beta\right\}$, we have for $0 < t < 1$

$$\begin{aligned}
\mathbb{P}\left[\lambda_{\min}(A_n) < t c_\beta\right] \leq & \mathbb{P}\left[\lambda_{\min}\left(\overline{A_n}\right) < t c_\beta\right] \\
\leq & \mathbb{P}\left[\max_{1 \leq k \leq d} \frac{1}{n+1}\left(a_k + \sum_{i=0}^{n-1} \left(\nabla_h g\left(X_{i+1}, \theta_i\right)\right)_k\right)^2 > \frac{1}{c_\beta^2 t^2}\right].
\end{aligned}$$

于是，由$p > 2$的马尔可夫不等式和詹森不等式可得

$$\begin{aligned}
& \mathbb{P}\left[\max_{1 \leq k \leq d} \sqrt{\frac{1}{n+1}\left(a_k + \sum_{i=0}^{n-1} \left(\nabla_h g\left(X_{i+1}, \theta_i\right)_k\right)^2\right)} > \frac{1}{c_\beta t}\right] \\
& \qquad \leq c_\beta^{2p} t^{2p} \mathbb{E}\left[\left(\max_{1 \leq k \leq d} \frac{1}{n+1}\left(a_k + \sum_{i=0}^{n-1} \left(\nabla_h g\left(X_{i+1}, \theta_i\right)_k\right)^2\right)\right)^p\right] \\
& \qquad \leq c_\beta^{2p} t^{2p} \mathbb{E}\left[\left(\frac{1}{n+1}\left(\sum_{i=1}^{d} a_k + \sum_{i=0}^{n-1} \left\|\nabla_h g\left(X_{i+1}, \theta_i\right)\right\|^2\right)\right)^p\right] \\
& \qquad \leq c_\beta^{2p} t^{2p} \frac{1}{n+1}\left(\left(\sum_{i=1}^{d} a_k\right)^p + \sum_{i=0}^{n-1} \mathbb{E}\left[\left\|\nabla_h g\left(X_{i+1}, \theta_i\right)\right\|^{2p}\right]\right).
\end{aligned}$$

然后，利用假设（**A1**），再利用（**A2**），我们得到

$$\begin{aligned}
& \mathbb{P}\left[\max_{1 \leq k \leq d} \sqrt{\frac{1}{n+1}\left(a_k + \sum_{i=0}^{n-1} \left(\nabla_h g(X_{i+1}, \theta_i)_k\right)^2\right)} > \frac{1}{c_\beta t}\right] \\
& \qquad \leq c_\beta^{2p} t^{2p} \frac{1}{n+1}\left(\left(\sum_{i=1}^{d} a_k\right)^p + n C_1'' + C_2'' \sum_{i=0}^{n-1} \mathbb{E}\left[\left\|\theta_i - \theta\right\|^{2p}\right]\right) \\
& \qquad \leq c_\beta^{2p} t^{2p} \frac{1}{n+1}\left(\left(\sum_{i=1}^{d} a_k\right)^p + n C_1'' + \frac{2^p C_2''}{\mu^p} \sum_{i=0}^{n-1} \mathbb{E}\left[V_n^p\right]\right).
\end{aligned}$$

根据引理 $\mathbb{E}\left[V_n^p\right] \leq V_p^p$给出的界6.1，最终可得

$$\mathbb{P}\left[\max_{1 \leq k \leq d} \sqrt{\frac{1}{n+1}\left(a_k + \sum_{i=0}^{n-1} \left(\nabla_h g\left(X_{i+1}, \theta_i\right)_k\right)^2\right)} > \frac{1}{c_\beta t}\right] \leq v_n t^{2p}$$

with

$$v_n = c_\beta^{2p}\left(\left(\frac{1}{n}\sum_{i=1}^{d} a_k\right)^p + C_1'' + \frac{2^p C_2'' V_p^p}{\mu^p}\right). \tag{68}$$

B.3 引理6.3 的证明

令 $E_k = \mathbb{E}\left[\nabla_h g(X,\theta)^2{}_k\right]$，并令 $\partial_k^2 g(h) = \mathbb{E}\left[\nabla_h g(X,h)_k^2\right]$。于是，由$p' \geq 2$的詹森不等式，

$$\left|\left(\overline{A_n}\right)_{kk}\right|^{-2p'} \leq 2^{p'-1}\left|\frac{1}{n+1}\sum_{i=0}^{n-1}\nabla_h g\left(X_{i+1},\theta_i\right)_k^2 - \partial_k^2 g\left(\theta_i\right)\right|^{p'} + 2^{p'-1}\left|\frac{a_k}{n+1} + \frac{1}{n+1}\sum_{i=0}^{n-1}\partial_k^2 g(\theta_i)\right|^{p'}.$$

因此，对于任意 $x > 0$,

$$\begin{aligned}\mathbb{P}\left[\left|\left(\overline{A_n}\right)_{kk}\right| < \frac{1}{x}\right] = \mathbb{P}\left[\left|\left(\overline{A_n}\right)_{kk}\right|^{-2p'} > x^{2p'}\right] \leq & \mathbb{P}\left[\left|\frac{1}{n+1}\sum_{i=0}^{n-1}\nabla_h g\left(X_{i+1},\theta_i\right)_k^2 - \partial_k^2 g\left(\theta_i\right)\right|^{p'} > \frac{x^{2p'}}{2^{p'}}\right] \\ & + \mathbb{P}\left[\left|\frac{a_k}{n+1} + \frac{1}{n+1}\sum_{i=0}^{n-1}\partial_k^2 g(\theta_i)\right|^{p'} > \frac{x^{2^{p'}}}{2^{p'}}\right].\end{aligned} \tag{69}$$

令 $M_0 = 0$，对于$n \geq 1$,

$$M_n = \sum_{i=0}^{n-1}\nabla_h g(X_{i+1},\theta_i)_k^2 - \partial_k^2 g(\theta_i).$$

那么，$(M_n)_{n\geq 0}$ 是一个鞅，因此根据伯克霍尔德不等式，参见（Hall 和 Heyde，2014，定理 2.10），存在一个显式常数 $C_{p'}$ 使得

$$\begin{aligned}\mathbb{E}\left[|M_n|^{p'}\right] \leq C_{p'}\mathbb{E}\left[\left|\sum_{i=1}^{n}\left(M_i - M_{i-1}\right)^2\right|^{p'/2}\right] \leq & C_{p'} n^{p'/2-1}\sum_{i=1}^{n}\mathbb{E}\left[\left|M_i - M_{i-1}\right|^{p'}\right] \\ \leq & C_{p'} n^{p'/2-1}\sum_{i=0}^{n-1}\mathbb{E}\left[\left|\nabla_h g\left(X_{i+1},\theta_i\right)_k^2 - \partial_k^2 g\left(\theta_i\right)\right|^{p'}\right],\end{aligned}$$

这里我们在第二个不等式中使用了詹森不等式。根据假设 **(A1)**，G 的强凸性以及引理 6.1,

$$\begin{aligned}\mathbb{E}\left[\left(\nabla_h g\left(X_{i+1},\theta_i\right)_k^2 - \partial_k^2 g\left(\theta_i\right)\right)^{p'}\right] \leq 2^{p'}\mathbb{E}\left[\left(\nabla_h g\left(X_{i+1},\theta_i\right)_k\right)^{2p}\right] \leq & 2^{p'}\mathbb{E}\left[\left\|\nabla_h g\left(X_{i+1},\theta_i\right)\right\|^{2p'}\right] \\ \leq & 2^{p'}C_1^{(p')} + 2^{p'}C_2^{(p')}\mathbb{E}\left[\left\|\theta_i - \theta\right\|^{2p'}\right] \\ \leq & 2^{p'}C_1^{(p')} + 2^{2p'}C_2^{(p')}\frac{V_p^{p'}}{\mu^{p}}.\end{aligned}$$

因此，

$$\mathbb{E}\left[\left|\frac{1}{n+1}\sum_{i=0}^{n-1}\nabla_h g\left(X_{i+1},\theta_i\right)_k^2 - \partial_k^2 g\left(\theta_i\right)\right|^{p'}\right] = \mathbb{E}\left[\left|\frac{1}{n+1}M_n\right|^{p'}\right] \leq 2^{p'}\frac{C_1^{(p')} + 2^{p'}C_2^{(p')}\frac{V_p^{p'}}{\mu^{p'}}}{(n+1)^{p'/2}}, \tag{70}$$

从而得到 $x > 0$

$$\mathbb{P}\left[\left|\frac{1}{n+1}\sum_{i=0}^{n-1}\nabla_h g\left(X_{i+1},\theta_i\right)_k^2-\partial_k^2 g\left(\theta_i\right)\right|^{p'}>\frac{x^{2p'}}{2^{p'}}\right]\leq\frac{2^{2p'}}{x^{2p'}}\frac{C_1^{(p')}+2^{p'}C_2^{(p')}\frac{V_p^{p'}}{\mu^{p'}}}{(n+1)^{p'/2}}. \tag{71}$$

接下来，根据詹森不等式，

$$\left|\frac{a_k}{n+1}+\frac{1}{n+1}\sum_{i=0}^{n-1}\left(\partial_k^2 g\left(\theta_i\right)\right)\right|^{p'}\leq\frac{1}{n+1}\left(|a_k|^{p'}+\sum_{i=0}^{n-1}\left|\partial_k^2 g(\theta_i)\right|^{p'}\right).$$

利用假设**(A1)**，再利用强凸性可得

$$\left|\partial_k^2 g\left(\theta_i\right)\right|^{p'}\leq C_1^{(p')}+2^{p'}C_2^{(p')}\frac{V_p^{p'}}{\mu^{p'}},$$

因此

$$\left|\frac{a_k}{n+1}+\frac{1}{n+1}\sum_{i=0}^{n-1}\left(\partial_k^2 g\left(\theta_i\right)\right)\right|^{p'}\leq C_1^{(p')}+\frac{|a_k|^{p'}}{n+1}+\frac{2^{p'}C_2^{(p')}}{\mu^{p'}}\left(\frac{1}{n+1}\sum_{i=0}^{n-1}V_i^{p'}\right).$$

因此，对于 $\frac{x^{2p'}}{2}{}_{p,} > C^{(p')}\,1,$

$$\begin{aligned}\mathbb{P}\left[\left|\frac{a_k}{n+1}+\frac{1}{n+1}\sum_{i=0}^{n-1}\partial_k^2 g(\theta_i)\right|^{p'}>\frac{x^{2p'}}{2^{p'}}\right]\leq&\mathbb{P}\left[\frac{1}{n+1}\left(|a_k|^{p'}+\frac{2^{p'}C_2^{(p')}}{\mu^{p'}}\sum_{i=0}^{n-1}V_i^{p'}\right)>\frac{x^{2p'}}{2^{p'}}-C_1^{(p')}\right]\\ \leq&\frac{1}{n+1}\frac{\mathbb{E}\left[|a_k|^{p'}+\frac{2^{p'}C_2^{(p')}}{\mu^{p'}}\sum_{i=0}^{n-1}V_i^{p'}\right]}{\frac{x^{2p'}}{2^{p'}}-C_1^{(p')}}.\end{aligned}$$

由（34）以及 $\frac{1}{+1}\sum_{=0}^{n-1}\mathrm{n}\,\mathrm{i}\,(\mathrm{i}+1)^{-}{}^{2}(_{1-\gamma})\gamma(_{\gamma-2\beta})\mathrm{p}^{\,2-\gamma}\leq\ \frac{^{1}}{\mathrm{n}+1}+\frac{}{\left|1-\frac{2\,1\ \gamma\ \gamma\ \ \gamma\ \ 2\beta}{}\right.}{}^{1}(\,)\,(\,)\mathrm{p}_{\,2-\gamma}\ \ \mathbf{1}_{\frac{2(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma}\neq1}\Big|}$

$\log(\mathrm{n}+1)\,(\mathrm{n}+1)\,2(_{1-\gamma})\gamma(_{\gamma-2\beta})\mathrm{p}^{\,2-\gamma}\quad{}^{\wedge1}$ 这一事实，并记 $\tilde{1}=1+\left|1-\frac{2\,1-\gamma\ \gamma\ \gamma-2\beta}{}{}^{1}(\,)\,(\,)\mathrm{p}_{\,2-\gamma}\quad\mathbf{1}_{\frac{2(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma}\neq1}\right|$，可得

$$\begin{aligned}&\frac{1}{n+1}\mathbb{E}\left[|a_k|^{p'}+\frac{2^{p'}C_2^{(p')}}{\mu^{p'}}\sum_{i=0}^{n-1}V_i^{p'}\right]=\frac{|a_k|^{p'}+\frac{2^{p'}C_2^{(p')}}{\mu^{p'}}\sum_{i=0}^{n-1}\mathbb{E}\left[V_i^{p'}\right]}{n+1}\\ &\quad\leq\frac{2^{p'}C_2^{(p')}}{\mu^{p'}}\tilde{K}_2\tilde{1}\frac{\log(n+1)}{(n+1)^{\frac{2(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma}\wedge1}}+\frac{2^{p'}C_2^{(p')}}{\mu^{p'}(n+1)}\left[1+|a_k|^{p'}+\tilde{K}_1\sum_{i=0}^{\infty}\exp\left(-c_\gamma\mu\lambda_0 i^{1-(\lambda+\gamma)}(1-\varepsilon'(i)\right)\right]\\ &\quad\leq M(\beta)\frac{\log(n+1)}{(n+1)^{\frac{2(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma}\wedge1}}\end{aligned}$$

其中 $n \geq 2$

$$M(\beta) = \frac{2^{p'} C_2^{(p')}}{\mu^{p'}} \left[\tilde{K}_2 \tilde{1} + 1 + |a_k|^{p'} + \tilde{K}_1 \sum_{n=0}^{+\infty} \exp\left(-c_\gamma \mu \lambda_0 n^{1-(\lambda+\gamma)} (1 - \varepsilon'(n)) \right) \right]$$

选择

$$\lambda_0 = \left[2^{p'} (C_1^{(p')} + 1) \right]^{-\frac{1}{2p'}} \tag{72}$$

于是得到

$$\mathbb{P}\left[\left| \frac{a_k}{n+1} + \frac{1}{n+1} \sum_{i=0}^{n-1} \partial_k^2 g(\theta_i) \right|^{p'} > \frac{\lambda_0^{-2p'}}{2^{p'}} \right] \leq \frac{M(\beta) \log(n+1)}{(n+1)^{\frac{2(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma} \wedge 1}}.$$

将后一个不等式与（69）和（71）结合，得到

$$\begin{aligned}
\mathbb{P}\left[\lambda_{\min}\left(\overline{A_n} \right) < \lambda_0 \right] &\leq \sum_{k=1}^{d} \mathbb{P}\left[\left| \left(\overline{A_n} \right)_{kk} \right| < \lambda_0 \right] \\
&\leq \frac{d M(\beta) \log(n+1)}{(n+1)^{\frac{2(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma} \wedge 1}} + \frac{d 2^{p'} \left(C_1^{(p')} + 2^{p'} C_2^{(p')} \frac{V_p^{p'}}{\mu^{p'}} \right)}{(C_1^{(p')} + 1) n^{p'/2}} \\
&\leq \frac{v_0 \log(n+1)}{(n+1)^{\frac{2(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma} \wedge 1}}
\end{aligned}$$

with

$$v_0 = d M(\beta) + \frac{d 2^{p'} \left(C_1^{(p')} + 2^{p'} C_2^{(p')} \frac{V_p^{p'}}{\mu^{p'}} \right)}{C_1^{(p')} + 1}. \tag{73}$$

由于 $\mathbb{P}\left[\lambda_{\min}(A_n) < \lambda_0\right] \leq \mathbb{P}\left[\lambda_{\min}\left(\overline{A_n}\right) < \lambda_0\right]$，可得所需结果。

B.4 引理6.4 的证明

令 $E_k = \mathbb{E}\left[\nabla_h (X, \theta)^2 {}_k \right]$ 且 $\partial_k^2 g(h) = \mathbb{E}\left[\nabla_h (X, h)_k^2 \right]$。于是

$$\begin{aligned}
\mathbb{E}\left[\left| \left(\overline{A_n} \right)_{kk}^{-2} - E_k \right|^{p'} \right] \leq & 2^{p'-1} \mathbb{E}\left[\left| \frac{1}{n+1} \sum_{i=0}^{n-1} \nabla_h g\left(X_{i+1}, \theta_i \right)_k^2 - \partial_k^2 g\left(\theta_i \right) \right|^{p'} \right] \\
& + 2^{p'-1} \mathbb{E}\left[\left| \frac{a_k - E_k}{n+1} + \frac{1}{n+1} \sum_{i=0}^{n-1} (\partial_k^2 g(\theta_i) - E_k) \right|^{p'} \right].
\end{aligned}$$

根据（70），

$$\mathbb{E}\left[\left| \frac{1}{n+1} \sum_{i=0}^{n-1} \nabla_h g\left(X_{i+1}, \theta_i \right)_k^2 - \partial_k^2 g\left(\theta_i \right) \right|^{p'} \right] \leq 2^{p'} \frac{C_1^{(p')} + 2^{p'} C_2^{(p')} \frac{V_p^{p'}}{\mu^{p'}}}{(n+1)^{p'/2}}.$$

接下来，由詹森不等式，

$$\mathbb{E}\left[\left|\frac{a_k-E_k}{n+1}+\frac{1}{n+1}\sum_{i=0}^{n-1}\left(\partial_k^2 g\left(\theta_i\right)-E_k\right)\right|^{p'}\right] \leq \frac{1}{n+1}\left(\left(a_k-E_k\right)^{p'}+\sum_{i=0}^{n-1}\mathbb{E}\left[\left|\partial_k^2 g(\theta_i)-E_k\right|^{p'}\right]\right).$$

利用柯西-施瓦茨不等式、假设（**A1'**），再利用假设（**A1**）可得

$$\begin{aligned}
&\mathbb{E}\left[\left|\partial_k^2 g\left(\theta_i\right)-E_k\right|^{p'}\right]=\mathbb{E}\left[\left|\mathbb{E}\left[\nabla_h g\left(\theta_i, X\right)_k^2-\nabla_h g\left(\theta, X\right)_k^2\left|\theta_i\right.\right]\right|^{p'}\right] \\
\leq & \mathbb{E}\left[\left|\mathbb{E}\left[\left(\nabla_h g\left(\theta_i, X\right)_k-\nabla_h g\left(\theta, X\right)_k\right)\left(\nabla_h g\left(\theta_i, X\right)_k+\nabla_h g\left(\theta, X\right)_k\right)\left|\theta_i\right.\right]\right|^{p'}\right] \\
\leq & \mathbb{E}\left[\mathbb{E}\left[\left(\nabla_h g\left(\theta_i, X\right)_k-\nabla_h g\left(\theta, X\right)_k\right)^2\left|\theta_i\right.\right]^{p'/2}\mathbb{E}\left[\left(\nabla_h g\left(\theta_i, X\right)_k+\nabla_h g\left(\theta, X\right)_k\right)^2\left|\theta_i\right.\right]^{p'/2}\right] \\
\leq & 2^{p'/2-1}L_{\nabla g}^{p'/2}\mathbb{E}\left[\|\theta_i-\theta\|^{p'}\left(2C_1^{p'/2}+C_2^{p'/2}\|\theta_i-\theta\|^{p'}\right)\right] \\
\leq & \frac{2^{p'}L_{\nabla g}^{p'/2}C_1^{p'/2}}{\mu^{p'/2}}\mathbb{E}\left[V_i^{p'/2}\right]+\frac{2^{3p'/2-1}C_2^{p'/2}L_{\nabla g}^{p'/2}}{\mu^{p'}}\mathbb{E}\left[V_i^{p'}\right] \\
\leq & \frac{2^{p'}L_{\nabla g}^{p'/2}C_1^{p'/2}}{\mu^{p'/2}}\sqrt{c_i}+\frac{2^{3p'/2-1}C_2^{p'/2}L_{\nabla g}^{p'/2}}{\mu^{p'}}c_i,
\end{aligned}$$

其中，c_i 由式（34）给出。将上述所有界合并，并利用 $E_k \leq C_1$，

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{a_k-E_k}{n+1}+\frac{1}{n+1}\sum_{i=0}^{n-1}\left(\partial_k^2 g\left(\theta_i\right)-E_k\right)\right|^{p'}\right] \\
\leq & \frac{1}{n+1}\left[2^{p'-1}(a_k^{p'}+C_1^{p'})+\sum_{i=0}^{n-1}\left(\frac{2^{p'}L_{\nabla g}^{p'/2}C_1^{p'/2}}{\mu^{p'/2}}\sqrt{c_i}+\frac{2^{3p'/2-1}C_2^{p'/2}L_{\nabla g}^{p'/2}}{\mu^{p'}}c_i\right)\right].
\end{aligned}$$

因此，注意到根据假设（**A1'**）和引理 6.1，$V_p<\infty$

$$\begin{aligned}
&\mathbb{E}\left[\left|(\overline{A_n})_{kk}^{-2}-E_k\right|^{p'}\right] \\
\leq & \underbrace{2^{p'-1}\frac{C_1^{(p')}+2^{p'}C_2^{(p')}\frac{V_p^{p'}}{\mu^{p'}}}{n}+\frac{2^{p'-1}}{n+1}\left[2^{p'-1}(a_k^{p'}+C_1^{p'})+\sum_{i=0}^{n-1}\left(\frac{2^{p'}L_{\nabla g}^{p'/2}C_1^{p'/2}}{\mu^{p'/2}}\sqrt{c_i}+\frac{2^{3p'/2-1}C_2^{p'/2}L_{\nabla g}^{p'/2}}{\mu^{p'}}c_i\right)\right]}_{:=\bar{c}_n},
\end{aligned}$$

其中，根据（34），$\bar{c}_n=O\left(\log(n)n^{-\left[\frac{(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma}\wedge 1\right]}\right)$。由于根据（**A6**）我们有 $E_k \geq \alpha$，

由马尔可夫不等式可得

$$\mathbb{P}\left[(\overline{A_n})_{kk}^{-1}\leq\sqrt{\alpha/2}\right]=\mathbb{P}\left[(\overline{A_n})_{kk}^{-2}\leq\alpha/2\right]\leq\frac{2^{p'}}{\alpha^{p'}}\mathbb{E}\left[\left|(\overline{A_n})_{kk}^{-2}-E_k\right|^{p'}\right]\leq\frac{2^{p'}\bar{c}_n}{\alpha^{p'}}.$$

因此，我们有

$$\begin{aligned}\mathbb{E}\left[(A_n)_{kk}^4\right] =&\mathbb{E}\left[\mathbf{1}_{\left(\overline{A_n}\right)_{kk}\geq\sqrt{\frac{2}{\alpha}}}(A_n)_{kk}^4\right]+\mathbb{E}\left[\mathbf{1}_{\left(\overline{A_n}\right)_{kk}<\sqrt{\frac{2}{\alpha}}}(A_n)_{kk}^4\right]\\ \leq&\mathbb{E}\left[\mathbf{1}_{\left(\overline{A_n}\right)_{kk}\geq\sqrt{\frac{2}{\alpha}}}c_\beta^4 n^{4\beta}\right]+\mathbb{E}\left[\mathbf{1}_{\left(\overline{A_n}\right)_{kk}<\sqrt{\frac{2}{\alpha}}}\left(\overline{A_n}\right)_{kk}^4\right]\\ \leq&c_\beta^4 n^{4\beta}\mathbb{P}\left[\left(\overline{A_n}\right)_{kk}^{-1}\leq\sqrt{\alpha/2}\right]+\frac{4}{\alpha^2}\leq\frac{2^{p'}c_\beta^4 n^{4\beta}\bar{c}_n}{\alpha^{p'}}+\frac{4}{\alpha^2}.\end{aligned}$$

由于 $\bar{c}_n = O\left(\log(n)n^{-\left[\frac{(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma}\wedge 1\right]}\right)$，对于 $\beta<\frac{(1-\gamma)\gamma(\gamma-2\beta)p}{4(2-\gamma)}\wedge\frac{1}{4}$，我们有 $\left[\frac{(1-\gamma)\gamma}{2-\gamma}\right.$ $\left.\frac{\gamma-2\beta}{}\wedge 1\right]-()p4\beta>0$，因此

$$w(\beta)=\sup_{n\geq 1}\bar{c}_n n^{4\beta}<+\infty,$$

最后

$$\mathbb{E}\left[\|A_n\|^4\right]\leq\sum_{k=1}^{d}\mathbb{E}\left[(A_n)_{kk}^4\right]\leq C_S^4$$

with

$$C_S^4=d\left[\frac{2^{p'}c_\beta^4 w(\beta)}{\alpha^{p'}}+\frac{4}{\alpha^2}\right]. \tag{74}$$

B.5 引理的证明6.5

首先，根据**(A6′)**，我们有

$$\mathbb{E}\left[\left(\left(\overline{A}_n\right)_{kk}\right)^{-2}\right]=\mathbb{E}\left[\frac{1}{n+1}\sum_{i=0}^{n-1}\nabla_h g\left(X_{i+1},\theta_i\right)_k^2\right]=\frac{1}{n+1}\sum_{i=0}^{n-1}\mathbb{E}\left[\nabla_h g\left(X_{i+1},\theta_i\right)_k^2\right]\geq\alpha.$$

接着，如前一个引理的证明中所示，

$$\mathbb{E}\left[\left|\frac{1}{n+1}\sum_{i=0}^{n-1}\nabla_h g\left(X_{i+1},\theta_i\right)_k^2-\frac{1}{n+1}\sum_{i=0}^{n-1}\mathbb{E}\left[\nabla_h g\left(X_{i+1},\theta_i\right)_k^2\right]\right|^2\right]\leq\frac{C_1'+C_2'\frac{4V_2^2}{\mu^2}}{n}.$$

因此，根据马尔可夫不等式，

$$\begin{aligned}\mathbb{P}\left[\left(\left(\overline{A}_n\right)_{kk}\right)^{-2}\leq\alpha/2\right]\leq&\mathbb{P}\left[\left|\frac{1}{n+1}\sum_{i=0}^{n-1}\nabla_h g\left(X_{i+1},\theta_i\right)_k^2-\frac{1}{n+1}\sum_{i=0}^{n-1}\mathbb{E}\left[\nabla_h g\left(X_{i+1},\theta_i\right)_k^2\right]\right|^2>\frac{\alpha^2}{4}\right]\\ \leq&\frac{4\left(C_1'+C_2'\frac{4V_2^2}{\mu^2}\right)}{n\alpha^2}.\end{aligned}$$

我们像在前一个引理中那样推导可得

$$\mathbb{E}\left[(A_n)_{kk}^4\right]\leq c_\beta^4 n^{4\beta}\mathbb{P}\left[\left(\overline{A_n}\right)_{kk}^{-1}\leq\sqrt{\alpha/2}\right]+\frac{4}{\alpha^2}\leq\frac{4\left(C_1'+C_2'\frac{4V_2^2}{\mu^2}\right)}{n^{1-4\beta}\alpha^2}+\frac{4}{\alpha^2}.$$

当 $\beta < 1/4$ 时，我们最终得到

$$\mathbb{E}\left[\|A_n\|^4\right] \leq \sum_{k=1}^{d} \mathbb{E}\left[(A_n)_{kk}^4\right] \leq C_S^4$$

with

$$C_S^4 = \frac{4d\left(1 + C_1' + C_2' \frac{4V_2^2}{\mu^2}\right)}{\alpha^2}. \tag{75}$$

C 线性回归技术引理和命题的证明

C.1 引理6.6 的证明

注意到

$$\left\|\widetilde{S}_n\right\| \leq \frac{1}{n+1}\left(\|S_0\| + \sum_{i=1}^{n}\left\|X_i X_i^T\right\|\right) \leq \frac{1}{n}\left(\|S_0\| + \sum_{i=1}^{n}\|X_i\|^2\right).$$

因此，对于 $\lambda > 0$,

$$\mathbb{P}\left[\lambda_{\min}\left(\widetilde{S}_n^{-1}\right) < \lambda\right] = \mathbb{P}\left[\|\widetilde{S}_n\| > 1/\lambda\right] \leq \mathbb{P}\left[\frac{1}{n}\left(\|S_0\| + \sum_{i=1}^{n}\|X_i\|^2\right) > \lambda^{-1}\right].$$

取 $\lambda_0 = \left(2\mathbb{E}\left[\|X\|^2\right]\right)^{-1}$，则得到

$$\mathbb{P}\left[\lambda_{\min}\left(\widetilde{S}_n^{-1}\right) < \lambda_0\right] \leq \mathbb{P}\left[\frac{1}{n}\left(\|S_0\| + \sum_{i=1}^{n}\left(\|X_i\|^2 - \mathbb{E}\left[\|X\|^2\right]\right)\right) > \mathbb{E}\left[\|X\|^2\right]\right].$$

取 p 次方，应用马尔可夫不等式，然后应用 Rosenthal 不等式，可得

$$\begin{aligned}
&\mathbb{P}\left[\frac{1}{n}\left(\|S_0\| + \sum_{i=1}^{n}\left(\|X_i\|^2 - \mathbb{E}\left[\|X\|^2\right]\right)\right) > \mathbb{E}\left[\|X\|^2\right]\right] \\
&\leq \mathbb{P}\left[\left(\frac{1}{n}\left(\|S_0\| + \left|\sum_{i=1}^{n}\left(\|X_i\|^2 - \mathbb{E}\left[\|X\|^2\right]\right)\right|\right)\right)^p > \left(\mathbb{E}\left[\|X\|^2\right]\right)^p\right] \\
&\leq \frac{1}{\left(\mathbb{E}\left[\|X\|^2\right]\right)^p}\mathbb{E}\left[\frac{1}{n^p}\left(\|S_0\| + \left|\sum_{i=1}^{n}\left(\|X_i\|^2 - \mathbb{E}\left[\|X\|^2\right]\right)\right|\right)^p\right] \\
&\leq \frac{2^{p-1}}{\left(\mathbb{E}\left[\|X\|^2\right]\right)^p}\left(C_1(p)n^{1-p}\mathbb{E}\left[|Z|^p\right] + C_2(p)n^{-p/2}\left(\mathbb{E}\left[|Z|^2\right]\right)^{p/2} + \|S_0\|^p n^{-p}\right),
\end{aligned}$$

与 $Z = \|X\|^2 - \mathbb{E}\left[\|X\|^2\right]$。

C.2 引理6.7的证明

根据 S_n，$S_n = \widetilde{S}_n$ 在事件 $T_n = \{\lambda_{\min}\left(\widetilde{S}_n\right) \geq \frac{1}{\beta} cn^{\beta}\}$。因此，对于与引理 λ_0相同的6.6，

$$\begin{aligned}\mathbb{P}\left[\lambda_{\min}\left(\bar{S}_n^{-1}\right) < \lambda_0\right] &= \mathbb{P}\left[T_n \cap \left\{\lambda_{\min}\left(\widetilde{S}_n^{-1}\right) < \lambda_0\right\}\right] + \mathbb{P}\left[T_n^c\right] \\ &\leq \mathbb{P}\left[\lambda_{\min}\left(\widetilde{S}_n^{-1}\right) < \lambda_0\right] + \mathbb{P}\left[T_n^c\right]. \end{aligned} \tag{76}$$

根据引理 6.6，

$$\mathbb{P}\left[\lambda_{\min}\left(\widetilde{S}_n^{-1}\right) < \lambda_0\right] \leq \tilde{v}_n, \tag{77}$$

其中 $\tilde{v}_n$ 由引理 6.6给出。那么，对于 $n \geq n_0$，，其中 n_0 在（37）中定义，我们有 $n \geq \left(\frac{1}{\beta} cc_2 \left(n+1\right)\right)^{-1/\beta}$，因此 $nn+1c_2 \geq \frac{1}{c\beta n^{\beta}}$。特别地，在事件 $\{\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^n X_i X_i^T\right) > c_2\}$ 上，我们有

$$\begin{aligned}\lambda_{\min}\left(\widetilde{S}_n\right) =& \lambda_{\min}\left(\frac{1}{n+1}\left(S_0 + \sum_{i=1}^n X_i X_i^T\right)\right) \\ \geq& \frac{n}{n+1}\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^n X_i X_i^T\right) > \frac{n}{n+1}c_2 \geq \frac{1}{c_\beta n^\beta}.\end{aligned}$$

因此，对于 $n \geq n_0$，$\{\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^n X_i X_i^T\right) > c_2\} \subset T_n$ 根据命题 6.2 以及这一事实$n \geq c_1 d$，

$$\mathbb{P}\left[T_n^c\right] \leq \mathbb{P}\left[\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^n X_i X_i^T\right) < c_2\right] \leq \exp(-c_3 n). \tag{78}$$

将（77）和（78）代入（76），即可得到

$$\mathbb{P}\left[\lambda_{\min}\left(\bar{S}_n^{-1}\right) < \lambda_0\right] \leq \tilde{v}_n + 2\exp(-c_3 n)$$

当 $n \geq n_0$ 时，引理的陈述便是对后一个不等式的改写。

C.3 引理6.8的证明

由于我们有

$$\|\bar{S}_n^{-1}\| = \min\left\{\|\tilde{S}_n^{-1}\|, \beta_{n+1}\right\} = \min\left\{\frac{1}{\lambda_{\min}\left\{\tilde{S}_n\right)}, \beta_{n+1}\right\},$$

对于 c_1,c_2,c_3 由命题 6.2，$n \geq c_1 d$ 和 $\kappa > 0$,

$$\mathbb{E}\left[\|\bar{S}_n^{-1}\|^\kappa\right] \leq \beta_{n+1}^\kappa \mathbb{P}\left[\lambda_{\min}\left(\tilde{S}_n\right) \leq c_2\right] + c_2^{-\kappa} \leq 2\beta_{n+1}^\kappa \exp\left(-c_3 n\right) + c_2^{-\kappa}.$$

由于 $\tilde{S}_n = \frac{1}{n+1}\left(S_0 + \sum_{i=1}^n X_i X_i^T\right)$ 且 $\sum_{i=1}^n X_i X_i^T \geq 0$，我们有 $\tilde{S}_n \geq \frac{1}{n+1} S_0$，因此$\|\bar{S}_n^{-1}\| \leq \|\tilde{S}_n^{-1}\| \leq (n+1)\|S_0^{-1}\|$，当 $n \geq 1$ 时成立。因此，对于 $n \leq c_1 d$，$\|\tilde{S}_n^{-1}\| \leq (c_1 d + 1)\|S_0^{-1}\|$，并且

最终得到该结果。

C.4 命题6.3的证明

回顾一下， $\beta_n = c^\beta n^\beta$。由于，对于 $\kappa > 0$，映射 $g : t \mapsto (c^\beta t^\beta)^\kappa \exp(-c_3 t)$ 有上界 $c_\beta^\kappa \left(\frac{\beta\kappa}{ec_3}\right)^{\beta\kappa}$，我们得到

$$\sup_{n \geq c_1 d} \mathbb{E}\left[\|\bar{S}_n^{-1}\|^\kappa\right] \leq 2c_\beta^\kappa \left(\frac{\beta\kappa}{ec_3}\right)^{\beta\kappa} + c_2^{-\kappa}.$$

考虑到情形 $n \leq c_1 d$，于是得到

$$\sup_{n \geq 1} \mathbb{E}\left[\|\bar{S}_n^{-1}\|^2\right] \leq \max\left\{2c_\beta^2 \left(\frac{2\beta}{ec_3}\right)^{2\beta} + c_2^{-2}, \left[(c_1 d + 1)\|S_0^{-1}\|\right]^2\right\},$$

and

$$\sup_{n \geq 1} \mathbb{E}\left[\|\bar{S}_n^{-1}\|^4\right] \leq \max\left\{2c_\beta^4 \left(\frac{4\beta}{ec_3}\right)^{4\beta} + c_2^{-4}, \left[(c_1 d + 1)\|S_0^{-1}\|\right]^4\right\}.$$

C.5 引理6.9的证明

首先注意到

$$\left\|\overline{S}_n^{-1} - H^{-1}\right\| = \left\|\overline{S}_n^{-1}(H - \overline{S}_n)H^{-1}\right\| \leq \left\|\overline{S}_n^{-1}\right\| \|H - \overline{S}_n\| \left\|H^{-1}\right\|.$$

在命题6.2的假设下，

$$\mathbb{P}\left[\lambda_{\min}\left(\tilde{S}_n\right) \leq c_2\right] \leq \exp\left(-c_3 n\right)$$

对于$n \geq c_1 d$。由于$\|\bar{S}_n^{-1}\| \leq \|\tilde{S}_n^{-1}\|$，$\lambda_{\min}(\bar{S}_n) \geq \lambda_{\min}\left(\tilde{S}_n\right)$，因此我们也有

$$\mathbb{P}\left[\lambda_{\min}\left(\bar{S}_n\right) \leq c_2\right] \leq \exp\left(-c_3 n\right)$$

对于$n \geq c_1 d$。因此，对于$n \geq n_0$，

$$\begin{aligned}\mathbb{E}\left[\left\|\overline{S}_n^{-1} - H^{-1}\right\|^2\right] =& \mathbb{E}\left[\mathbf{1}_{\lambda_{\min}(\overline{S}_n) \leq c_2}\left\|\overline{S}_n^{-1} - H^{-1}\right\|^2\right] + \mathbb{E}\left[\mathbf{1}_{\lambda_{\min}(\overline{S}_n) > c_2}\left\|\overline{S}_n^{-1} - H^{-1}\right\|^2\right] \\ \leq& \frac{1}{(\lambda_{\min}\beta_n)^2}\mathbb{E}\left[\mathbf{1}_{\lambda_{\min}(\overline{S}_n) \leq c_2}\left\|\overline{S}_n - H\right\|^2\right] + \frac{1}{(\lambda_{\min}c_2)^2}\mathbb{E}\left[\left\|\tilde{S}_n - H\right\|^2\right],\end{aligned}$$

其中，最后一个等号用到了$n \geq n_0$时，$\bar{S}_n = \tilde{S}_n$ 在事件 $\{\lambda_{\min}(\bar{S}_n > c_2)\}$，如引理6.7的证明中所示。第一项可利用赫尔德不等式进行估计

其中 $\frac{1}{q} + \frac{1}{q'} = 1$且 $q' = p/2$ 时

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}_{\lambda_{\min}\left(\overline{S}_n\right)\leq c_2}\left\|\overline{S}_n - H\right\|^2\right] \leq& \mathbb{P}\left[\lambda_{\min}\left(\overline{S}_n\right) \leq c_2\right]^{1/q}\mathbb{E}\left[\left\|\overline{S}_n - H\right\|^{2q'}\right]^{1/q'} \\ \leq& \exp(-c_3(p-2)n/p)\mathbb{E}\left[\left\|\overline{S}_n - H\right\|^p\right]^{2/p}.\end{aligned}$$

利用H 的上界和凸性不等式$(a+b)^p \leq 2^{p-1}(a^p + b^p)$ 可得粗略上界

$$\begin{aligned}\mathbb{E}\left[\left\|\overline{S}_n - H\right\|^p\right]^{2/p} \leq \mathbb{E}\left[\left(\left\|\overline{S}_n\right\| + \|H\|\right)^p\right]^{2/p} \leq& 2^{2-2/p}\left(\mathbb{E}\left[\left\|\overline{S}_n\right\|^p\right] + \lambda_{\max}^p\right)^{2/p} \\ \leq& 4\max\left\{\lambda_{\max}^2, \mathbb{E}\left[\left\|\overline{S}_n\right\|^p\right]^{\frac{2}{p}}\right\}\end{aligned}$$

由于X 存在 $2p$ 阶矩，我们得到

$$\mathbb{E}\left[\left\|\overline{S}_n\right\|^p\right] \leq \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n \|X_i\|^2\right)^p\right] \leq \left(\mathbb{E}\left[\|X\|^{2p}\right]\right)^{1/2}.$$

因此我们得到

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}_{\lambda_{\min}\left(\overline{S}_n\right)\leq c_2}\left\|\overline{S}_n - H\right\|^2\right] &\leq 4\exp\left(-c_3(p-2)n/p\right)\max\left\{\lambda_{\max}^2, \left(\mathbb{E}\left[\|X\|^{2p}\right]\right)^{2/p}\right\} \\ &= 4\exp\left(-c_3(p-2)n/p\right)\left(\mathbb{E}\left[\|X\|^{2p}\right]\right)^{2/p}\end{aligned}$$

对于第二项，利用弗罗贝尼乌斯范数与算子范数之间的关系可得

$$\begin{aligned}\mathbb{E}\left[\left\|\overline{S}_n - H\right\|^2\right] &\leq \mathbb{E}\left[\left\|\overline{S}_n - H\right\|_F^2\right] \\ &\leq \frac{2}{(n+1)^2}\|S_0 - H\|_F^2 + \frac{2}{(n+1)^2}\mathbb{E}\left[\left\|\sum_{k=1}^n\left(X_kX_k^T - \mathbb{E}\left[XX^T\right]\right)\right\|_F^2\right] \\ &= \frac{2}{(n+1)^2}\|S_0 - H\|_F^2 + \frac{2}{n+1}\mathbb{E}\left[\left\|XX^T - \mathbb{E}\left[XX^T\right]\right\|_F^2\right] \\ &\leq \frac{2}{(n+1)^2}\|S_0 - H\|_F^2 + \frac{2}{n+1}\mathbb{E}\left[\|X\|^4\right].\end{aligned}$$

综合以上所有界限即可得到该命题中的界限。

D 广义线性模型的技术引理与命题证明

D.1 引理证明6.10

借助不等式（8），可得

$$\|\bar{S}_n\| \leq \frac{1}{n+1}\|S_0\| + \frac{L_{\nabla l}}{n+1}\sum_{i=1}^{n}\|X_i\|^2 + \frac{\sigma d}{n+1}\sum_{i=1}^{n}\|Z_i\|^2.$$

其中 $Z_i = e^{i[d]+1}$。然后，类似于引理 6.7 的证明可得，对于 $\lambda_0 = \left(2L_{\nabla l}\mathbb{E}\left[\|X\|^2\right] + 2\sigma\right)^{-1}$，

$$\begin{aligned}
&\mathbb{P}\left[\lambda_{\min}\left(\overline{S}_n^{-1}\right) < \lambda_0\right] \\
&\quad \leq \mathbb{P}\left[\frac{\|S_0\|}{n} + \frac{L_{\nabla l}}{n}\sum_{i=1}^{n}\left(\|X_i\|^2 - \mathbb{E}\left[\|X\|^2\right]\right) + \frac{\sigma}{n}\sum_{i=1}^{n}\left(\|Z_i\|^2 - 1\right) > L_{\nabla l}\mathbb{E}\left[\|X\|^2\right] + \sigma\right].
\end{aligned}$$

然后，根据 $p \geq 1$ 的马尔可夫不等式，我们得到

$$\begin{aligned}
&\mathbb{P}\left[\lambda_{\min}\left(\overline{S}_n^{-1}\right) < \lambda_0\right] \leq \frac{\mathbb{E}\left[\left(\frac{1}{n}\|S_0\| + \frac{1}{n}\sum_{i=1}^{n} L_{\nabla l}\left(\|X_i\|^2 - \mathbb{E}\left[\|X\|^2\right]\right) + \sigma\left(\|Z_i\|^2 - 1\right)\right)^p\right]}{\left(L_{\nabla l}\mathbb{E}\left[\|X\|_2^2\right] + \sigma\right)^p} \\
&\leq \frac{2^{p-1}}{\left(L_{\nabla l}\mathbb{E}\left[\|X\|^2\right] + \sigma\right)^p}\left(n^{-p}\|S_0\|^p + C_1(p)n^{1-p}\mathbb{E}\left[|T|^p\right] + C_2(p)n^{-p/2}\left(\mathbb{E}\left[\|T\|^2\right]\right)^{p/2}\right),
\end{aligned}$$

其中 $T = L_{\nabla l}\left(\|X\|^2 - \mathbb{E}\left[\|X\|^2\right]\right) + \sigma\left(\|Z\|^2 - 1\right)$。

D.2 命题证明 6.5

对于所有 $n \geq 2d$，显然有

$$\lambda_{\min}\left(\overline{S}_n\right) \geq \frac{\lfloor n/d \rfloor \sigma}{(n+1)} \geq \frac{n+1-d}{d(n+1)}\sigma \geq \frac{1}{2d}\sigma,$$

且 $S_n \geq \frac{1}{2}dS_0$ 对于 $n \leq 2d-1$，因此

$$\sup_{n\geq 1}\left\|\bar{S}_n^{-1}\right\| \leq 2d\max\left\{\frac{1}{\sigma}, \left\|S_0^{-1}\right\|\right\}.$$

D.3 命题 6.7 的证明

记作

$$H(\theta_\sigma) = \mathbb{E}\left[\nabla_h^2\ell\left(Y, \theta_\sigma^T X\right)XX^T\right] \qquad \text{and} \qquad \overline{H}_n = \frac{1}{n+1}\left(S_0 + \sum_{i=0}^{n-1}\nabla_h^2\ell\left(Y_{i+1}, \langle\theta_i, X_{i+1}\rangle\right)X_{i+1}X_{i+1}^T\right).$$

可以将 $H_n - H(\theta_\sigma)$ 分解为

$$\begin{aligned}\overline{H}_n - H\left(\theta_\sigma\right) =& \frac{1}{n+1}\sum_{i=0}^{n-1}\nabla_h^2\ell\left(Y_{i+1},\left\langle\theta_i, X_{i+1}\right\rangle\right)X_{i+1}X_{i+1}^T + \frac{1}{n+1}S_0 - H\left(\theta_\sigma\right)\\ =& \frac{1}{n+1}\sum_{i=0}^{n-1}\nabla_h^2\ell\left(Y_{i+1},\left\langle\theta_i, X_{i+1}\right\rangle\right)X_{i+1}X_{i+1}^T - H\left(\theta_i\right)\\ &+\frac{1}{n+1}\sum_{i=0}^{n-1}\left(H\left(\theta_i\right) - H\left(\theta_\sigma\right)\right) + \frac{1}{n+1}\left(S_0 - H\left(\theta_\sigma\right)\right).\end{aligned}$$

现在给出前一个等式右侧各项的收敛速率。令 $\mathrm{M}_n := \sum^{n-1}\mathrm{i}=0\ \left(\nabla_h^2\ell\left(\mathrm{Y}i+1,\theta_i^T\mathrm{X}_{i+1}\right)\mathrm{X}i+1\mathrm{X}_{i+1}^T - \mathrm{H}(\theta_i)\right)$。由于 $\mathbb{E}\left[\nabla_h^2\ell\left(\mathrm{Y}i+1,\theta_i^T\mathrm{X}_{i+1}\right)\mathrm{X}i+1\mathrm{X}_{i+1}^T|\mathcal{F}_i\right] = \mathrm{H}\left(\theta_i\right)$，其中 $(\mathcal{F}_i)$ 是由样本生成的 σ-代数，即 $\mathcal{F}_i := \sigma\left((\mathrm{X}_1,\mathrm{Y}_1),\ldots,(\mathrm{X}_i,\mathrm{Y}_i)\right)$。因此，$(\mathrm{M}n)n\geq 1$ 是一个鞅，故

$$\frac{1}{(n+1)^2}\mathbb{E}\left[\|M_n\|^2\right] \leq \frac{1}{(n+1)^2}\sum_{i=0}^{n-1}\mathbb{E}\left[\left\|\left(\nabla_h^2\ell\left(Y_{i+1},\theta_i^T X_{i+1}\right)X_{i+1}X_{i+1}^T - H(\theta_i)\right)\right\|^2\right] \leq \frac{L_{\nabla l}^2\mathbb{E}\left[\|X\|^4\right]}{n}$$

接下来只需处理 $\frac{1}{+1}\sum^{n-1}{=0}\ n\not{i}(\not{H}(\theta_i)\ - H\left(\theta_\sigma\right))$。根据假设（**GLM1**），可得

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{n+1}\sum_{i=0}^{n-1}\left(H(\theta_i) - H\left(\theta_\sigma\right)\right)\right\|^2\right] \leq& \frac{1}{n}\sum_{i=0}^{n-1}\mathbb{E}\left[\|H(\theta_i) - H\left(\theta_\sigma\right)\|^2\right]\\ \leq& \frac{L_{\nabla^2 L}^2}{n}\sum_{i=0}^{n-1}\mathbb{E}\left[\|\theta_i - \theta_\sigma\|^2\right] \leq \frac{L_{\nabla^2 L}^2}{\sigma n}\sum_{i=0}^{n-1}v_{i,\mathrm{GLM}},\end{aligned}$$

其中 $v_{i,\ \mathrm{GLM}}$ 在命题 6.6中定义。因此，由于

$$\left\|\frac{d\sigma}{n}\sum_{i=1}^{n}e_{i[d]+1}e_{i[d]+1}^T - \sigma I_d\right\|^2 = \left\|\frac{d\sigma}{n}\sum_{i=d\lfloor\frac{n}{d}\rfloor}^{n}e_{i[d]+1}e_{i[d]+1}^T + \left(\frac{d\sigma}{n}\left\lfloor\frac{n}{d}\right\rfloor - \sigma\right)I_p\right\|^2$$

and

$$\frac{d^2\sigma^2}{n^2}\left\|\sum_{k=d\lfloor\frac{n}{d}\rfloor}^{n}e_{i[d]+1}e_{i[d]+1}^T\right\|^2 \leq \frac{d^2\sigma^2}{n^2}\left(n - d\left\lfloor\frac{n}{d}\right\rfloor\right)\sum_{k=d\lfloor\frac{n}{d}\rfloor}^{n}\left\|e_{i[d]+1}e_{i[d]+1}^T\right\|^2 \leq \frac{d^4\sigma^2}{n^2},$$

可得

$$\left\|\overline{S}_n - H_\sigma\right\|^2 \leq \frac{4}{n}\left(L_{\nabla l}^2\mathbb{E}\left[\|X\|^4\right] + \frac{L_{\nabla^2 L}^2}{\sigma}\sum_{i=0}^{n-1}v_{i,\mathrm{GLM}} + \frac{1}{n}\|S_0 - H\left(\theta_\sigma\right)\|^2\right) + \frac{16d^4\sigma^2}{n^2}$$

现在，注意到根据引理6.9，

$$\left\|\overline{S}_n^{-1} - H_\sigma^{-1}\right\| = \left\|\overline{S}_n^{-1}(H_\sigma - \overline{S}_n)H_\sigma^{-1}\right\| \leq \left\|\overline{S}_n^{-1}\right\|\|H_\sigma - \overline{S}_n\|\left\|H_\sigma^{-1}\right\|,$$

由此可得，这得益于命题6.5

$$\mathbb{E}\left[\left\|\overline{S}_n^{-1}-H_\sigma^{-1}\right\|^2\right]\leq\frac{C_{S,\sigma}^2}{\sigma^2}\mathbb{E}\left[\left\|\overline{S}_n-H_\sigma\right\|^2\right],$$

即有

$$\mathbb{E}\left[\left\|\overline{S}_n^{-1}-H_\sigma^{-1}\right\|^2\right]\leq\frac{4C_{S,\sigma}^2}{\sigma^2 n}\left(L_{\nabla l}^2\mathbb{E}\left[\|X\|^4\right]+\frac{L_{\nabla^2 L}^2}{\sigma}\sum_{i=0}^{n-1}v_{i,\mathrm{GLM}}+\frac{1}{n}\|S_0-H\left(\theta_\sigma\right)\|^2\right)+\frac{16d^4C_{S,\sigma}^2}{n^2}. \tag{79}$$

E 如何验证逻辑回归的假设（GLM3）

R 注意到 θ_σ 是 $\mathbb{E}\left[\nabla_h\ell\left(Y、X^T\theta_\sigma\right)X+\sigma\theta_\sigma\right]=0$的唯一解，因此

$$\mathbb{E}\left[\left|\nabla_h l\left(Y,X^T\theta_\sigma\right)X_k+\sigma(\theta_\sigma)_k\right|^2\right]=Var\left[\nabla_h l\left(Y,X^T\theta_\sigma\right)X_k\right].$$

对于逻辑回归，我们有 $\mathrm{Y}\in\{-1,1\}$和 $\nabla_h\mathrm{l}\left(\mathrm{Y},\mathrm{X}^T\theta_\sigma\right)=\frac{-\mathrm{Y}}{1+\exp(-\mathrm{Y}\theta_\sigma^T\mathrm{X})}$，因此我们需要对所有 $1\leq\mathrm{k}\leq\mathrm{d}$，求出 $\frac{-}{1+}{}^{k}\mathrm{YX}\exp(\mathrm{Y}\theta_\sigma^T\mathrm{X})$ 的方差下界。为保证假设（GLM3），我们对 (X, Y) 施加一个最小随机性条件，即对于所有 $1\leq\mathrm{k}'\leq\mathrm{d}$，存在一个 $\mathrm{x}^{k'}$ $\sigma(\mathrm{Y},\mathrm{X}_i,\mathrm{i}\neq\mathrm{k}')$，其有界于 M，并且存在一个事件 $\mathrm{A}\in\sigma(\mathrm{Y},\mathrm{X}_i,\mathrm{i}\neq\mathrm{k}')$，满足 $\mathbb{P}[\mathrm{A}\cap\{|\mathrm{X}_i|\leq\mathrm{M},1\leq\mathrm{i}\leq\mathrm{d},\mathrm{i}\neq\mathrm{k}'\}]>\eta$ 以及 $\mathrm{c},\epsilon>0$，使得在 A 上有

$$\mathbb{P}\left[X_{k'}>x^{k'}+c|Y,X_i,i\neq k'\right]>\epsilon\quad\text{and}\quad\mathbb{P}\left[X_{k'}<x^{k'}-c|Y,X_i,i\neq k'\right]>\epsilon.$$

特别地，由于 $\mathrm{u}\mapsto\frac{\mathrm{u}}{1+\exp(-\alpha\mathrm{u})}$对所有 $\alpha\in\mathbb{R}$ 和 $\mathbb{C}^1$ 都是单调的，因此存在一个关于 $(\mathrm{Y},\mathrm{X}_i,\mathrm{i}\neq\mathrm{k})$ 可测的 y_k σ，以及一个显式依赖于 $\mathrm{M}\|\theta_\sigma\|$ 的 $\mathrm{c}(\mathrm{M}\|\theta_\sigma\|)$，使得在 $\mathrm{B}:=\mathrm{A}\cap\{|\mathrm{X}_i|\leq\mathrm{M},1\leq\mathrm{i}\leq\mathrm{d},\mathrm{i}\neq\mathrm{k}\}$ 上，

$$\mathbb{P}\left[\frac{-YX_k}{1+\exp(-Y\theta_\sigma^TX)}>y_k+c(M\|\theta_\sigma\|)|Y,X_i,i\neq k\right]>\epsilon,$$

and

$$\mathbb{P}\left[\frac{-YX_k}{1+\exp(-Y\theta_\sigma^TX)}<y_k-c(M\|\theta_\sigma\|)|Y,X_i,i\neq k\right]>\epsilon.$$

由此可得，在事件 B 上有

$$Var\left[\frac{-YX_k}{1+\exp(-Y\theta_\sigma^TX)}|Y,X_i,i\neq k\right]\geq 2\epsilon c(M\|\theta_\sigma\|)^2.$$

因此，

$$Var\left[\frac{-YX_k}{1+\exp(-Y\theta_\sigma^TX)}\right]\geq\mathbb{E}\left[\mathbf{1}_B Var\left[\frac{-YX_k}{1+\exp(-Y\theta_\sigma^TX)}|Y,X_i,i\neq k\right]\right]\geq 2\eta\epsilon c(M\|\theta_\sigma\|)^2,$$

并且我们可以选择

$$\alpha_\sigma = 2\eta\epsilon c(M\|\theta_\sigma\|)^2.$$

F 无正则化随机牛顿算法二次收敛的反例

我们在此证明，即使在最简单的情形 $d = 1$ 下，随机牛顿算法也可能无法在二次均值意义下收敛。假设我们在此定义朴素牛顿自适应矩阵 A_n

$$A_n = \left[\frac{1}{n+1}\left(Id + \sum_{i=0}^{n-1} \nabla_h^2 g(X_{i+1}, \theta_i)\right)\right]^{-1}.$$

回顾一下，众所周知（Boyer 和 Godichon-Baggioni，2020），θ_n 几乎必然收敛到极小化点 θ_0，收敛速度为 $n^{-\gamma}$，其中 $\gamma \in (1/2,1)$。

关于 ∇g 几乎处处有定义的反例

令 $g((x,y),\theta) = (x\theta)^2 + y\lfloor\theta\rfloor\theta$，并设 (X,Y) 为坐标相互独立的随机向量，满足 $X \simeq$ 伯努利分布$(1/2)$，且 $\mathbb{P}[Y=1] = \mathbb{P}[Y=-1] = 1/2$。于是，$G(\theta) = \mathbb{E}\left[X^2\right]\theta^2 + \mathbb{E}[Y]\lfloor\theta\rfloor\theta = \theta^2/2$，并且勒贝格几乎处处有 $\nabla_h g((x,y),h) = 2x^2 h + y\lfloor h\rfloor$ 且 $\nabla_h^2 g((x,y),h) = 2x^2$。

令 $n \geq 1$。于是，$\mathbb{P}\left[X_1 = 0, \ldots, X_n = 0, Y_1 = -1, \ldots, Y_n = -1\right] = 2^{-2n}$，并且在事件 $\{X_1 = 0, \ldots, X_n = 0, Y_1 = -1, \ldots, Y_n = -1\}$ 上，只要对所有 $k \geq 0$ 都有 $\theta_k \notin \mathbb{N}$（暂时假设这一点），

$$A_k^{-1} = \frac{1}{k}\left(1 + \sum_{i=0}^{k-1} 2X_{i+1}^2\right) = \frac{1}{k}.$$

因此，$A_k = k$ 以及 $(\theta_k)_{1\leq k\leq n}$ 由以下递归定义：

$$\theta_k = \theta_{k-1} - \gamma_k A_k \lfloor\theta_{k-1}\rfloor Y_k = \theta_{k-1} + k\gamma_k\lfloor\theta_{k-1}\rfloor.$$

若对某个 $\alpha < 1$ 有 $\gamma_k = k^{-\alpha}$，则有 $k\gamma_k = k^{1-\alpha}$，因此对于 $\theta_0 > 1$

$$\theta_k \geq (1 + k^{1-\alpha}/2)\theta_{k-1}.$$

我们推得 $\theta_n \geq \prod_{k=1}^n (1 + k^{1-\alpha}/2) \geq (n!)^{1-\alpha} 2^{-n}$。特别地，

$$\mathbb{E}\left[\|\theta_n - \theta_0\|^2\right] \geq 2^{-3n}(n!)^{1-\alpha} \xrightarrow{n\to\infty} \infty$$

当 $\theta_k \notin \mathbb{N}$ 对所有 $k \geq 0$ 均成立时。由于对于每个 $k \geq 1$，$\theta_k \notin \mathbb{N}$ 对于几乎处处的 $\theta_0 \in (1,2],$，后一个假设对 $\theta_0 \in]1,2]$ 的勒贝格几乎处处取值均成立。

关于 ∇g 连续性的反例

设 f 满足 $f''(\theta) = \mathbf{1}_{\mathbb{Z}+]-1/3,1/3[}$，并令 $g((x, y), \theta) = (x\theta)^2 + y f(\theta)$。设 (X, Y) 为一个坐标相互独立的随机向量，满足 $X \simeq$ 伯努利分布(1/2)，且 $Y \sim \mathcal{U}([-2, 2])$。则，$G(\theta) = \mathbb{E}\left[X^2\right]\theta^2 + \mathbb{E}[Y] f(\theta) = \theta^2/2$，且 $\nabla_h g((x, y), \theta) = 2x^2\theta + y f'(\theta)$。于是，$A_0 = 1$,

$$A_k^{-1} = \frac{k}{k+1}(A_{k-1}^{-1} + 2X_k^2 + f''(\theta_{k-1})Y_k)$$

对于 $k \geq 1$且

$$\theta_n = \theta_{n-1} - A_{n-1}\gamma_n \nabla_h g((X_n, Y_n), \theta_{n-1}) = \theta_{n-1} - A_{n-1}\gamma_n(2X_n^2\theta_{n-1} + Y_n f'(\theta_{n-1})).$$

令 $\theta_0 = 3/2$ 且 $\gamma_k = k^{-\gamma}$ 对于 $k \geq 1$，并考虑满足以下条件的 $(X_i, Y_i)_{0 \leq i \leq n}$：

- $X_i = 0$ 对所有 $1 \leq i \leq n$</style}，由此得到 $X_1 = 0, \ldots, X_n = 0] = 2^{-}{}_n$，且对所有 $k \geq 1$,

$$\theta_k = \theta_{k-1} - A_{k-1}\gamma_k Y_k f'(\theta_{k-1}).$$

- θ_{k-1} 已知时，$Y_k \in {}_{\gamma}{}_{-1}{}^{1}{}_{\cancel{kAkf'(\theta_{k-1})}}((\mathbb{Z}+]1/3, 2/3[) - \theta_{k-1}) \cap [-2, -1] := T_k$（请注意，$T_k$ 将证明为非空）。

引理 F.1. 当 $k \geq 1$ 时，以下事实成立。

1. $\theta_k \geq k + 1$,
2. $A_k = k + 1$,
3. 其中 ℓ 表示勒贝格测度，

$$\ell\left(\frac{1}{\gamma_k A_{k-1} f'(\theta_{k-1})}((\mathbb{Z}+]1/3, 2/3[) - \theta_{k-1}) \cap [-2, -1])\right) \geq 1/6.$$

证明。 我们将对$k \geq 1$进行归纳，证明这三个事实。对于$k = 1$，我们有$A_0 = \gamma_1 = 1$ 且$f'(3/2) = 1$，因此 $\theta_1 = 3/2 - Y_1$。由于

$$\begin{aligned} T_1 = \frac{1}{A_0\gamma_1 f'(\theta_0)}((\mathbb{Z}+]1/3, 2/3[) - \theta_{k-1}) \cap [-2, -1] &= (-3/2 + \mathbb{Z}+]1/3, 2/3[) \cap [-2, -1] \\ &=]-7/6, -1] \cup [-2, -11/6[, \end{aligned}$$

$\ell(T_1) \geq 1/3$。另一方面，对于$Y_1 \in T1$，$\theta_1 \geq 3/2 + 1 \geq 2$。

下面证明归纳步骤。令$k \geq 2$，并假设结论对$l \leq k - 1$成立。于是$\theta_l \in \mathbb{Z} + [1/3, 2/3]$对所有$l \leq k$都成立，这意味着$A_{k-1} = k - 1$。因此，

$$\theta_k = \theta_{k-1} + k^{1-\gamma} Y_k f'(\theta_{k-1}).$$

根据归纳法， $\theta_{k-1} \geq k$，并且由于$f'(\theta) \geq \theta/2$ 对于 $\theta \geq 0$,

$$T_k = ((b + a\mathbb{Z} + a]1/3, 2/3[) \cap [-2, -1]$$

由鸽巢原理，我们推得 $a = 1/(A_{k-1}\gamma_k f'(\theta_{k-1})) \leq 2k^{2-\gamma} \leq 1$ 且 $b = -\theta_{k-1}/a$。最后，对于 $Y_k \in A_k$，我们有 $Y_k \leq -1$，因此

$$\theta_k \geq k + \frac{1}{2}k^{2-\gamma} \geq k + 1.$$

□

根据前面的结果,

$$\mathbb{P}\left[X_1 = \cdots = X_n = 0, Y_1 \in T_1, \ldots, Y_n \in T_n\right] \geq 2^{-n} \cdot 6^{-n} = 12^{-n}.$$

此外，根据我们之前的证明，在此事件上，对于 $1 \leq k \leq n$，我们有

$$\theta_k = \theta_{k-1} - Y_k A_{k-1}\gamma_k f'(\theta_{k-1}) \geq \theta_{k-1} + k^{2-\gamma}/2 \geq \theta_{k-1}k^{2-\gamma}/2.$$

我们推得 $\theta_n \geq \theta_{k-1}(k-1)^{1-\gamma}/3 \geq (n!)^{2-\gamma}/2^n$。特别地,

$$\mathbb{E}\left[\|\theta_n - \theta_0\|^2\right] \geq (n!)^{2-\gamma}/24^n \xrightarrow{n\to\infty} \infty.$$

注意，后一个结果可以很容易地调整，从而得到一个反例，其中 g 的光滑程度可以任意高。

参考文献

Bach, F.（2014）。平均随机梯度下降对逻辑回归局部强凸性的自适应性。《机器学习研究杂志》，15(1)：595–627。

Bach, F. 和 Moulines, E.（2013）。具有收敛速率 o (1/n) 的非强凸光滑随机逼近。载于神经信息处理系统进展，第 773–781 页。

Bercu, B.、Bigot, J.、Gadat, S. 和 Siviero, E.（2021）。一种用于正则化半离散最优传输的随机高斯-牛顿算法。*arXiv* 预印本 *arXiv:2107.05291*。

Bercu, B.、Godichon, A. 和 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*《控制与优化学报》，58(1)：348–367。

Boyd，S. 和 Vandenberghe, L.（2004）。凸优化。剑桥大学出版社。

Boyer，C. and Godichon-Baggioni, A. (2020). 关于随机牛顿算法及其加权平均版本的渐近收敛速率

。*arXiv* 预印本 *arXiv:2011.09706*。Cardot, H., Cénac, P., and Godichon-Baggioni, A. (2015). 希尔伯特空间中几何中位数的在线估计：非渐近置信球。技术报告，arXiv:1501.06930。Cardot, H., Cénac, P., and Zitt, P.-A. (2013). 使用平均随机梯度算法在希尔伯特空间中高效、快速地估计几何中位数。*Bernoulli*，19(1):18–43。Cénac, P., Godichon-Baggioni, A., and Portier, B. (2020). 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。*arXiv* 预印本 *arXiv:2006.12920*。De Vilmarest, J. and Wintenberger, O. (2021). 使用卡尔曼递推进行随机在线优化。机器学习研究杂志，22:223–1。Défossez, A., Bottou, L., Bach, F., and Usunier, N. (2020). Adam 和 Adagrad 的一个简单收敛证明。*arXiv* 预印本 *arXiv:2003.02395*。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。机器学习研究杂志，12(7)。Gadat, S. and Gavra, I. (2020). 非凸景观中随机自适应算法的渐近研究。*arXiv* 预印本 *arXiv:2012.05640*。Gadat, S. and Panloup, F. (2017). 无强凸性条件下 Ruppert-Polyak 平均的最优非渐近界。*arXiv* 预印本 *arXiv:1709.03342*。Godichon-Baggioni, A. (2016). 使用随机梯度算法在希尔伯特空间中估计几何中位数：Lp 和几乎必然收敛速率。多元分析杂志，146:209–222。Godichon-Baggioni, A. (2021). 无需强凸性或有界梯度的平均随机梯度算法的均方收敛。*arXiv* 预印本 *arXiv:2107.12058*。Godichon-Baggioni, A., Portier, B., and Lu, W. (2022). 使用二阶随机算法的递归岭回归。Godichon-Baggioni, A., Werge, N., and Wintenberger, O. (2021). 流式数据随机逼近算法的非渐近分析。工作论文或预印本。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。载于国际机器学习会议，第5200–5209页。PMLR。

Hall, P. 和 Heyde, C. C.（2014）。鞅极限定理及其应用。学术出版社。Koltchinskii, V. 和 Mendelson, S.（2015）。无需集中不等式即可界定随机矩阵的最小奇异值。《国际数学研究通知》，2015(23):12991– 13008。Koneˇcnˋy, J.、Liu, J.、Richtárik, P. 和 Takáˇc, M.（2015）。近端设置下的迷你批次半随机梯度下降。*IEEE*《信号处理精选主题期刊》，10(2):242–255。Leluc, R. 和 Portier, F.（2020）。条件随机梯度下降的渐近最优性。*arXiv* 预印本 *arXiv:2006.02745*。Pelletier, M.（1998）。随机算法的几乎处处渐近行为。随机过程及其应用，78(2):217–244。Pelletier, M.（2000）。平均随机算法的渐近几乎处处有效性。*SIAM*《控制与优化杂志》，39(1):49–72。Pinelis, I.（1994）。巴拿赫空间中鞅分布的最优界。《概率年刊》，22:1679–1706。Polyak, B. 和 Juditsky, A.（1992）。随机逼近的加速。*SIAM*《控制与优化杂志》，30:838–855。Robbins, H. 和 Monro, S.（1951）。一种随机逼近方法。《数学统计年刊》，第400–407页。Ruppert, D.（1988）。从缓慢收敛的 Robbins-Monro 过程中进行有效估计。技术报告，康奈尔大学运筹学与工业工程系。