# Non asymptotic analysis of Adaptive stochastic gradient algorithms and applications

Antoine Godichon-Baggioni, Pierre Tarrago

# Non asymptotic analysis of Adaptive stochastic gradient algorithms and applications

Antoine Godichon-Baggioni * Pierre Tarrago *

**Abstract**

In stochastic optimization, a common tool to deal sequentially with large sample is to consider the well-known stochastic gradient algorithm. Nevertheless, since the stepsequence is the same for each direction, this can lead to bad results in practice in case of ill-conditionned problem. To overcome this, adaptive gradient algorithms such that Adagrad or Stochastic Newton algorithms should be prefered. This paper is devoted to the non asymptotic analyis of these adaptive gradient algorithms for strongly convex objective. All the theoretical results will be adapted to linear regression and regularized generalized linear model for both Adagrad and Stochastic Newton algorithms.

**Keywords:** Non asymptotic analysis; Online estimation; Adaptive gradient algorithm; Adagrad; Stochastic Newton algorithm.

## 1 Introduction

A usual problem in stochastic optimization is to estimate the minimizer $\theta$ of a convex functional $G : \mathbb{R}^d \longrightarrow \mathbb{R}$ of the form

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

where $g : \mathcal{X} \times \mathbb{R}^d \longrightarrow \mathbb{R}$, and $X$ is a random variable lying in $\mathcal{X}$. Indeed, this is the case for usual regressions such that the linear and logistic ones (Bach, 2014) or the estimation of the geometric median and quantiles (Cardot et al., 2013, 2015; Godichon-Baggioni, 2016) to name a few. Several techniques have been developed to estimate the solution of the problem, which can be split into two main branches: iterative and recursive methods. Iterative methods consist in considering the empirical function generated by the sample and to approximate its minimizer with the help of usual convex optimization methods (Boyd and Vandenberghe, 2004) or considering some refinements such that mini-batch algorithms (Konečnỳ et al., 2015). Although these methods are known to be very competitive, they can encounter computational problems to deal with large samples. In addition, they are not suitable for dealing with data arriving sequentially, and one can so focus on recursive methods.

*Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, France
antoine.godichon_baggioni@sorbonne-universite.fr, pierre.tarrago@sorbonne-universite.fr
This project is supported by the Agence Nationale de la Recherche funding CORTIPOM ANR-21-CE40-0019.

One of the most famous and studied recursive method is unquestionably the stochastic gradient algorithm (Robbins and Monro, 1951) and its averaged version (Ruppert, 1988; Polyak and Juditsky, 1992). Considering data $X_1, \ldots, X_n, X_{n+1}, \ldots$ arriving sequentially, it is defined recursively for all $n \geq 0$ by

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

where $(\gamma_n)$ is a positive step sequence converging to 0. These estimates are studied for a while: one can refer to (Pelletier, 1998, 2000) for some asymptotic results while one can refer to more recent literature for non asymptotic results such that convergence in quadratic mean of the estimates (Bach and Moulines, 2013; Gadat and Panloup, 2017; Gower et al., 2019; Godichon-Baggioni, 2021). The averaged estimates are known to be asymptotically efficient and achieve the Cramer-Rao bound (up to rest terms) under some regularity assumptions.

Nevertheless, the step sequence $(\gamma_n)$ cannot be adapted to each direction of the gradient which can lead to bad results in practice for ill-conditioned problems. In order to alleviate this, one can more focus on adaptive stochastic gradient algorithms of the form

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

where $(A_n)$ is a sequence of (random) matrices which enables to be adapted to each coordinate. One of the most famous adaptive algorithm is Adagrad (Duchi et al., 2011), which can be seen as a way to standardize the gradient $\nabla_h g\left(X_{n+1}, \theta\right)$. In recent works, Bercu et al. (2020) and Boyer and Godichon-Baggioni (2020) consider $(A_n)$ as a sequence of estimates of the inverse of the Hessian, leading to a Stochastic Newton algorithm. This last method is of particular interest in the case where the Hessian of the function we would like to minimize has eigenvalues at different scales for instance.

Remark that several asymptotic results exist on adaptive method and one can focus on the recent works of Leluc and Portier (2020) or Gadat and Gavra (2020) among others, while non asymptotic results are less usual. Nevertheless, in a recent work, De Vilmarest and Wintenberger (2021) give bounds with high probabilities in the special case of Kalman recursion for logistic regression, while Défossez et al. (2020) focus on the $L^2$ rates of convergence for Adagrad and Adam. Furthermore, Bercu et al. (2021) obtain the rate of convergence in quadratic mean of Stochastic Gauss-Newton algorithms for optimal transport. Note however that in all these cases, the gradient of $g$ is supposed to be uniformly bounded.

In this paper, we focus on non asymptotic rates of convergence for strongly convex functions (and so, with unbounded gradient). More precisely, we propose a first rate of convergence of Adaptive estimates in the case where the sequence $A_n$ possibly diverges, but with a control on this possible divergence. Supposing in addition that $A_n$ admits an uniform fourth order moment, we establish that $\mathbb{E}\left[G\left(\theta_n\right) - G(\theta)\right]$ converges at the usual rate of convergence. Finally, we establish a non constraining general framework for obtaining the rate of convergence of Stochastic Newton and Adagrad algorithms. These results will be

applied for linear regression and ridge generalized linear model.

The paper is organized as follows: Section 2, the general framework is introduced. The algorithms and theoretical results of convergence are given in Section 3 while applications consisting in the linear regression and the generalized linear model are respectively given in Sections 4 and 5. The proofs are postponed in Section 6 and in Appendix.

## 2 Framework

In what follows, we consider a random variable $X$ taking values in a measurable space $\mathcal{X}$ and fix $d \geq 2$. We focus on the estimation of the minimizer $\theta$ of the convex function $G : \mathbb{R}^d \longrightarrow \mathbb{R}$ defined for all $h \in \mathbb{R}^d$ by

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

with $g : \mathcal{X} \times \mathbb{R}^d \longrightarrow \mathbb{R}$. Let us suppose from now that the following assumptions are fulfilled:

**(A1)** For almost every $x \in \mathcal{X}$, the functional $g(x, .)$ is differentiable on $\mathbb{R}^d$ and there exists $p \geq 2$ and non-negative constants $C_1^{(p)}$, $C_2^{(p)}$ such that for all $h \in \mathbb{R}^d$,

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

**(A2)** The functional $G$ is twice continuously differentiable.

**(A3)** The Hessian of $G$ is uniformly bounded on $\mathbb{R}^d$, i.e there is a positive constant $L_{\nabla G}$ such that for all $h \in \mathbb{R}^d$,

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

where $\|.\|_{op}$ is the usual spectral norm for matrices.

**(A4)** The functional $G$ is $\mu$ quasi-strongly convex: for all $h \in \mathbb{R}^d$,

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

Remark that in a particular case, Assumption **(A3)** ensures that the gradient of $G$ is $L_{\nabla G}$-Lipschitz. Note that these assumptions are usual for obtaining the $L^2$ rates of convergence of the stochastic gradient algorithms and their averaged versions (Bach and Moulines, 2013; Gower et al., 2019).

## 3 Adaptive stochastic gradient algorithms

### 3.1 The algorithms

Let $X_1, \ldots, X_n, X_{n+1}, \ldots$ be i.i.d copies of $X$. Then, an adaptive stochastic gradient algorithm is defined recursively for all $n \geq 0$ by

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

where $\theta_0$ is arbitrarily chosen, $\gamma_n = c_\gamma n^{-\gamma}$ with $c_\gamma > 0$, $\gamma \in (0,1)$ and $A_n$ is a sequence of symmetric and positive matrices such that there is a filtration $(\mathcal{F}_n)_{n\geq 0}$ satisfying:

- For all $n \geq 0$, $A_n$ is $\mathcal{F}_n$-measurable.
- $X_{n+1}$ is independent of $\mathcal{F}_n$.

Typically, one can consider $A_n$ only depending on $X_1, \ldots, X_n, \theta_0, \ldots, \theta_n$ and consider the filtration generated by the sample, i.e $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. Considering $A_n$ diagonal with $(A_n)_{k,k} = \left(\frac{1}{n+1}\left(a_k + \sum_{i=1}^n \nabla_h g\left(X_i, \theta_{i-1}\right)_{i,i}^2\right)\right)^{-1/2}$ leads to Adagrad algorithm (Duchi et al., 2011). Furthermore, the case where $A_n$ is a recursive estimate of the inverse of the Hessian corresponds to the Stochastic Newton algorithm (Bercu et al., 2020; Boyer and Godichon-Baggioni, 2020) while the case where $A_n = \frac{1}{n+1}\left(A_0 + \sum_{i=1}^n \nabla_h g\left(X_i, \theta_{i-1}\right) \nabla_h g\left(X_i, \theta_{i-1}\right)^T\right)$ corresponds to the stochastic Gauss-newton algorithm (Cénac et al., 2020; Bercu et al., 2021).

### 3.2 Convergence results

#### 3.2.1 A first convergence result

In order to obtain a first rate of convergence of the estimates, let us now introduce some assumptions on the sequence of random matrices $(A_n)_{n\geq 0}$:

**(H1 )** One can control the smallest and largest eigenvalues of $A_n$:

**(H1a)** There exists $(v_n)_{n\geq 0}, \lambda_0 > 0$ and $\delta, q \geq 0$ such that

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

for $0 < t \leq 1$, with $(v_{n+1}(n+1)^{-\delta})_{n\geq 0}$ decreasing.
If $\gamma \leq 1/2$, one also assumes the stronger hypothesis of the existence of $\lambda'_n = \lambda'_0 (n+1)^{-\lambda'}$ with $\lambda'_0 > 0$, $\lambda' < \gamma$ such that for all $n \geq 0$,

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

**(H1b)** There exists $\beta_n = c_\beta n^\beta$ with $c_\beta \geq 0$ and $0 < \beta < \frac{\gamma}{2}$ if $\gamma \leq 1/2$ or $0 < \beta < \gamma - 1/2$ if $\gamma > 1/2$ such that for all $n \geq 0$,

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

Remark that the case $\delta = 0$ is allowed in **(H1a)** and that one can always choose $\beta$ in the allowed range of **(H1b)**. In most cases and especially for Adagrad and Stochastic Newton algorithm, **(H1a)** is easily verified. The presence of the decreasing term $v_n$ in **(H1a)** takes into account a general phenomenon (usually implied by Rosenthal inequality) that error contributions from higher moments of $X$, albeit dominant for small $n$, fade as $n$ goes to infinity. Concerning **(H1b)**, some counter-examples showing that the estimates possibly diverge in the case where this last assumption is not fulfilled are given in Appendix F, meaning that this assumption is unfortunately crucial. Anyway, an easy way to corroborate it is to replace the random matrices $A_n$ by

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

and one can directly check that $\left\|\tilde{A}_n\right\|_{op} \leq \beta_{n+1}$. Similar adjustment can be used to ensure **(H1a)** in the case $\gamma \leq 1/2$.

Let us consider the case of Newton's method, and especially the case where the estimates of the Hessian are of the form $H_n = \frac{1}{n+1}\left(H_0 + \sum_{k=1}^{n} a_k \Phi_k \Phi_k^T\right)$ and which can be so recursively invert with the help of Riccati/Shermann-Morrisson's formula (see Bercu et al. (2020); Boyer and Godichon-Baggioni (2020); Godichon-Baggioni et al. (2022)). In order to verify **(H1b)**, one can consider the following version of the estimate of the Hessian

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

where $e_k$ is the $k$-th (modulo $d$) canonical vector (see Bercu et al. (2021); Godichon-Baggioni et al. (2022)). We can now obtain a first rate of convergence of the estimates. For the sake of simplicity, let us now denote the risk error by $V_n := G(\theta_n) - G(\theta)$. Note that since $G$ is $\mu$ quasi-strongly convex, one has $\|\theta_n - \theta\|^2 \leq \frac{2}{\mu} V_n$.

**Theorem 3.1.** *Suppose Assumptions **(A1)** to **(A3)** and **(H1)** hold. Then, for all $n \geq 1$ and for any $\lambda < \min\{\gamma - 2\beta, 1 - \gamma\}$,*

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

*with $\varepsilon(n) = o(1)$ given in (20) and $K_1^{(1)}$, $K_{1'}^{(1)}$, $K_2^{(1)}$, $K_3^{(1)}$ respectively given in (21) and (22).*

In the particular case where $\delta/2 \geq \gamma - 2\beta$ (which happens as soon as $\delta \geq 1$), one can simply set $\lambda = 0$ in the above formula : we will see that it is the case for the generalized linear model with the stochastic Newton algorithm. However, for Adagrad algorithms, one can not avoid using first $\lambda > 0$, since $A_n$ depends on $\nabla g(X, \cdot)$ rather than $\nabla^2 g(X, \cdot)$ (while the expectation of latter is bounded on $\mathbb{R}^d$, the one of the former is generally unbounded). To get rid of this weaker statement, we need the following equivalent of Theorem 3.1 for

higher moments.

**Proposition 3.1.** *Suppose Assumptions **(A1)** with $p > 2$, **(A2)**, **(A3)** and **(H1)** hold. Then for any $p' < p$ and any $\lambda < \min\{\gamma - 2\beta, 1 - \gamma\}$,*

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

*with $\epsilon'(n)$, $K_1^{(1')}$, $K_{1'}^{(1')}$, $K_2^{(1')}$ and $K_3^{(1')}$ respectively given in (57), (58) and (60).*

### 3.2.2 Convergence when $A_n$ has bounded moments

In order to get a better rate of convergence, let us now introduce some new assumptions on the sequence of random matrices $(A_n)$:

**(H2a)** The random matrices $A_n$ admit uniformly bounded second order moments: there is $C_S$ such that for all $n \geq 0$:

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

**(H2b)** The random matrices $A_n$ admit uniformly bounded fourth order moments: there is $C_S$ such that for all $n \geq 0$:

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

For a simpler statement, we assume here and in the next paragraph that $q > 0$ in **(H1a)**, although similar bound would hold in full generality.

**Theorem 3.2.** *Suppose Assumptions **(A1)** to **(A3)** for some $p > 2$, **(H1)** and **(H2a)** hold with $\delta > 0$. Then, for all $n \geq 0$,*

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

*where $\varepsilon(n) = o(1)$ is given in (25) and $K_1^{(2)}, K_{1'}^{(2)}, K_2^{(2)}, K_3^{(2)}$ are respectively given in (26), (27) and (28).*

Finally, in order to get the rate of convergence in quadratic mean of Stochastic Newton estimates, we now give the $L^2$ rate of convergence of $G\left(\theta_n\right)$ when $\gamma > 1/2$.

**Proposition 3.2.** *Suppose Assumptions **(A1)** to **(A3)** for some $p > 2$, **(H1)** and **(H2b)** hold with $\gamma > 1/2, \delta > 0$ and $\beta < \gamma - 1/2$. Then*

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

*with* $K_1^{(2')}$, $K_{1'}^{(2')}$, $K_2^{(2')}$, $K_3^{(2')}$ *respectively given in* (63), (64) *and* (65).

Remark that one has $M_n = O\left(n^{-\min\left\{2\gamma, \frac{\delta(p-2)}{p}\right\}}\right)$. Hence, for $\delta$ large enough (namely $\delta > \frac{2p}{p-2}\gamma$), the main contribution comes from the second term of the latter bound. Then, for any $0 \leq \gamma' \leq \min\left\{2\gamma, \frac{\delta(p-2)}{p}\right\}$, only depending on $v_n$ and $\gamma$, we have

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

The function $w_\infty : \left[0, \min\left\{2\gamma, \frac{\delta(p-2)}{p}\right\}\right] \to \mathbb{R}$ can be computed numerically, but in any case note that $w_\infty(\gamma') \leq K_1^{(2')} \sup_{t\geq 1}\left\{t^{\gamma'} \exp\left(-\frac{1}{2}\lambda_0 \mu t^{1-\gamma}\right)\right\} + K_2^{(2')} + K_3^{(2')}$, so that a function analysis yields, for $\gamma' \in \left[0, \min\left\{2\gamma, \frac{\delta(p-2)}{p}\right\}\right]$,

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

We will see in most applications that under suitable assumptions, $\gamma'$ can be equal to $2\gamma$ (namely when $\delta \geq \frac{2p}{p-2}\gamma$).

### 3.2.3 Convergence results for stochastic Newton algorithms

Let us now focus on the rate of convergence of Stochastic Newton algorithm. In this aim, let us denote $H := \nabla^2 G(\theta)$ and let us suppose from now that the following assumptions are fulfilled too:

**(A1')** There is $L_{\nabla g}$ such that for all $h \in \mathbb{R}^d$,

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

**(A5)** There is a non negative constant $L_\delta$ such that for all $h \in \mathbb{R}^d$,

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

**(H3)** The estimate $A_n$ converges to $H^{-1}$: there is a decreasing positive sequence $(v_{A,n})$ such that for al $n \geq 0$,

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

Observe that assumption **(A1')** is often called expected smoothness in the literature (Bach and Moulines, 2013) and is satisfied in most of examples such that linear and logistic regression (Bach and Moulines, 2013; Bach, 2014) or the estimation of geometric quantiles and medians (Cardot et al., 2013) among others. Concerning **(A5)**, under **(A3)**, it is satisfied as soon as the Hessian is Lipschitz on a neighborhood of $\theta$. For instance, in the case of the linear regression, $L_\delta = 0$. Finally, Assumption **(H3)** is satisfied if having a first rate of convergence of the estimates of

$\theta$ (thanks to Theorem 3.2 or Proposition 3.2 for instance) leads to have a first rate of convergence of $A_n$, which is often verified in practice (see Boyer and Godichon-Baggioni (2020) for instance).

**Theorem 3.3.** *Suppose Assumptions* ***(A1)*** *to* ***(A5)****, and* ***(H1)*** *to* ***(H3)*** *hold with* $\gamma > 1/2$, $\delta > 0$ *and* $\beta < \gamma - 1/2$. *Then,*

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

*where* $K_i^{(3)}$, $i = 1, 1', 2, 2'$ *are defined in* (29), (30) *and* (31), *and* $d_k$ *only depending on* $M_k$ *and* $v_{A,k}$ *is given in* (30).

Remark from (30) that $d_k \leq C(v_{A,k} + M_k)$ for some constant $C > 0$. The latter results can be further simplified if we also assume a sufficiently large exponent $\delta$ in **(H1a)**.

**Corollary 3.1.** *Suppose Assumptions* ***(A1)*** *to* ***(A4)****, and* ***(H1)*** *to* ***(H3)*** *hold with* $\gamma > 1/2$, $\delta > \frac{2\gamma p}{p-2}$ *and* $\beta < \gamma - 1/2$. *Then,*

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

*with* $K_i^{(3')}$, $i = 1...2''$ *given in* (32) *and* (33).

Then, if $v_{A,n}$ converges to 0, we obtain the usual rate of convergence $\frac{1}{n^\gamma}$.

### 3.2.4 Convergence results for adaptive gradient (Adagrad)

Recall that the Adagrad algorithm amounts to specify $d$ initial parameters $a_1, \ldots, a_d \in \mathbb{R}_+$ choose $\overline{A_n}$ diagonal with

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

The original Adagrad algorithm would then amount to take $\gamma = 1/2$. To guarantee non-degeneracy of the matrices $(\overline{A_n})_{n\geq 0}$, we assume some minimal fluctuation of the gradient at the minimizer $\theta$.

**(A6)** There is $\alpha > 0$ such that for all $1 \leq i \leq d$,

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

**(A6')** There is $\alpha > 0$ such that for all $h \in \mathbb{R}^d$ and $1 \leq i \leq d$,

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

Remark that **(A6')** is much stronger as **(A6)**. However, the former is often satisfied, as it is the case for the linear regression with noise. Anyway, one can consider the following transformation of $\overline{A_n}$:

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

where $\beta_n = c_\beta n^\beta$ with $\beta < \min\{\gamma/2, 1/4\}$ ($\lambda', \lambda_0'$ and $c_\beta > 0$ are chosen arbitrarily).

**Theorem 3.4.** *Suppose Assumptions **(A1)** to **(A4)** and **(A6)** hold with* $\beta < \min\left\{\frac{(1-\gamma)\gamma(\gamma-2\beta)p}{4(2-\gamma)}, 1/4\right\}$. *Then,*

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

*with* $\tilde{\varepsilon}(n)$ *given in* (35), $v_n = v_0 \log(n+1)$, *with* $v_0$, $C_S^4$ *and* $\tilde{\lambda}_0$ *given in* (73), (74) *and* (72) *with* $p' = \frac{2(1-\gamma)}{2-\gamma}p$. *In addition,* $K_1^{(4)}$, $K_2^{(4)}$ *and* $K_3^{(4)}$ *given in* (36). *If **(A6')** is satisfied, same conclusion holds for* $\beta < 1/4$ *with* $C_S$ *given in* (75) *taking* $p' = \frac{2(1-\gamma)}{2-\gamma}p$.

In the special case where $\gamma = 1/2$, which corresponds to the usual Adagrad algorithm, we get

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

and we so achieve the usual rate of convergence $\frac{1}{\sqrt{n}}$ as soon as $1/2 - \frac{(1-4\beta)(p-1)}{6} < 0$, i.e as soon as $p > 4\frac{1-\beta}{1-4\beta}$.

# 4 Application to linear model

Let us now consider the linear model $Y = X^T\theta + \epsilon$ where $X \in \mathbb{R}^d$ and $\epsilon$ is a centered random real variable independent from $X$. Let us suppose from now that $\mathbb{E}\left[XX^T\right]$ is positive. Then, $\theta$ is the unique minimizer of the functional $G : \mathbb{R}^d \longrightarrow \mathbb{R}$ defined for all $h \in \mathbb{R}^d$ by

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

If $X$ admits a second order moment, the function $G$ is twice continuously differentiable with $\nabla G(h) = -\mathbb{E}\left[\left(Y - X^T h\right)X\right]$ and $\nabla^2 G(h) = \mathbb{E}\left[XX^T\right]$.

### 4.1 Stochastic Newton algorithm

The Stochastic Newton algorithm is defined recursively for all $n \geq 0$ by (Boyer and Godichon-Baggioni, 2020)

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

with $\tilde{S}_n = \frac{1}{n+1} \left( S_0 + \sum_{i=1}^n X_i X_i^T \right)$, with $S_0$ positive, and

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

with $\beta_n = c_\beta n^{-\beta}$. Remark that $\tilde{S}_{n+1}^{-1}$ can be easily updated with only $O\left(d^2\right)$ operations using Sherman Morrison (or Ricatti's) formula. More precisely, considering $S_n = (n+1)\tilde{S}_n$, one has

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

Then, one can easily update $\tilde{S}_n$ and $\overline{S}_n$. We can now rewrite Theorem 3.3 as follows:

**Theorem 4.1.** *Suppose that there is $p > 4$ such that $X, \epsilon$ admits a moment of orders $2p$ and $p$. Suppose also that there is a positive constant $L_{MK}$ such that for any $h \in \mathbb{S}^{d-1}$, $\sqrt{\mathbb{E}\left[hXX^Th\right]} \leq L_{MK}\mathbb{E}\left[\left|X^Th\right|\right]$. Then, for $\gamma > 1/2$, we have*

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

*where $K_{2,lin}, K_{2',lin}^{(3)}, K_{1,lin}^{(3)}, K_{1',lin}^{(3)}, d_n$ are given by (40) while $v_{H,n}$ is defined in (39).*

Observe that $d_n = O\left( \frac{1}{n^{\max\left\{\frac{p-2}{2}, 2\gamma\right\}}} \right)$ and $v_{H,n} = O\left(n^{-1}\right)$, and since $p > 4$, these terms are both negligible.

### 4.2 Adagrad algorithm

For linear model, Adagrad algorithm is defined for all $n \geq 0$ by

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

with $\bar{D}_n$ diagonal with, for $\gamma \leq 1/2$,

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

where $0 < \beta < (\gamma - \lambda')/2$ for some $a_k > 0$ and if $\gamma > 1/2$,

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

for some $0 < \beta < \gamma - 1/2$. The usual Adagrad algorithm is done with $\gamma = 1/2$, which yields for us

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

and almost surely there exists $n_0 \geq n$ such that for $n \geq n_0$,

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

which is the usual Adagrad algorithm. We can then rewrite Theorem 3.4 as follows (we simply give it for $\gamma = 1/2$, the reader can easily adapt it to the case $\gamma > 1/2$).

**Theorem 4.2.** *Suppose that there is $p > 2$ such that $X, \epsilon$ admits a moment of orders $2p$. Then, for $\gamma \leq 1/2$, we have*

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

*where $\varepsilon_{n,lin}^{ada} = o(1)$ is given in (41) and $K_{1,lin}^{ada}$, $K_{2,lin}^{ada}$ $K_{3,lin}^{ada}$ are given by (42), (43) and (44).*

Remark that similar statements hold for $\gamma > 1/2$. Observe that in the case where $\gamma = 1/2$, the $\frac{1}{\sqrt{n}}$ rate of convergence is achieved as soon as $(p-1)(1-4\beta)/3 \geq 1/2$, i.e as soone as $p > \frac{5-4\beta}{2(1-4\beta)}$.

## 5 Application to generalized linear models

The framework of the linear regression can be easily generalized to the more general setting of finite dimensional linear models. Let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ a cost function for some domain $\mathcal{Y} \subset \mathbb{R}$. The general learning problem is to solve the minimization problem

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

with $(X, Y) \sim \mathbb{P}$ and $\mathcal{F}$ is a given class of measurable function from $\mathcal{X}$ to $\mathcal{Y}$, where $\mathcal{X}$ is a measurable space. In the case of finite dimensional linear models, $\mathcal{Y} = \mathbb{R}$ and $\mathcal{F} = \left\{h^T\Phi(\cdot), h \in \mathbb{R}^m\right\}$, with $\Phi : \mathcal{X} \to \mathbb{R}^m$ a *known* design function (remark that the setting can be easily generalized to the case $\mathcal{Y} = \mathbb{R}^p$ and $\Phi : \mathcal{X} \to \mathbb{R}^m$ and $h \in M_{m,p}(\mathbb{R})$). Then,

assuming that $\ell$ is convex and adding a regularization term on $\theta$, the minimization problem turns into the framework of this paper with

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

with $\tilde{Z} = (Y, \Phi(X)) := (Y, \tilde{X})$ and for all $h \in M_{m,p}(\mathbb{R})$, $g(\tilde{Z}, h) = \ell(Y, h^T\tilde{X})$. In what follows, let us suppose from now that the cost function $l$ is twice differentiable for the second variable and that there is a positive constant $L_{\nabla l}$ such that for all $h$

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

where $\nabla_h^2 \ell(.,.)$ is the second order derivative with respect to the second variable. Remark that such a bound is generally assumed if we require that for all $h$, $\|\nabla^2 G(h)\|_{op} \leq L_{\nabla G} < \infty$ This is for example satisfied when $\ell(y, y') = f(y - y')$ with $\sup_y |f''(y)| < +\infty$. For example, in the simplest case of the logistic regression, we consider a couple of random variables $(X, Y)$ lying in $\mathbb{R}^d \times \{-1, 1\}$, $\Phi = I_d$ and $\ell(y, y') = \log(1 + \exp(-yy'))$, and we indeed have for all $h$ and $Y \in \{-1, 1\}$

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

There are then two main cases to deal with the convexity of the minimization problem : either assume strong convexity or use a regularization. The first consists in assuming that the functional $h \longmapsto \mathbb{E}\left[\ell\left(Y, h^T X\right)\right]$ is strongly convex, which is in particular verified when there exists $\alpha > 0$ such that

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

and $\mathbb{E}\left[XX^T\right]$ is positive. This case is called the elliptic case in the sequel and the results are very analogous to the ones for the linear regression and are thus not repeated. We will then focus on the regularized case. Without uniform lower bound on $\nabla_h^2 \ell(y, y')$, one needs a regularization term, yielding the following regularized minimization problem

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

for some $\sigma > 0$. In what follows, we suppose that the minimizer exists and we denote it by $\theta_\sigma$.

### 5.1 Stochastic Newton algorithm

The Stochastic Newton algorithm is defined recursively for all $n \geq 0$ by

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

where, using the tricked introduced in Bercu et al. (2021) and developed in Godichon-Baggioni et al. (2022), $\overline{S}_n$ is the natural recursive estimate of the Hessian given by

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

with $i[d]$ denoting $i$ modulo $d$. Remark that one can easily update the inverse using the Riccati's formula used twice, i.e considering $S_n = (n+1)\overline{S}_n$ and

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

one has $\overline{S}_{n+1}^{-1} = (n+2)S_{n+1}^{-1}$. In what follows, let us suppose that the following assumptions hold:

**(GLM1)** There is $L_{\nabla^2 L} \geq 0$ such that the function $h \longmapsto \mathbb{E}\left[\nabla_h^2 \ell\left(Y, h^T X\right) XX^T\right]$ is $L_{\nabla^2 L}$-Lispchitz with respect to the spectral norm.

**(GLM2)** There is $p > 2$ such that $X$ admits a moment of order $2p$ and such that there is a positive constant $L_\sigma$ satisfying for all $0 \leq a \leq 2p$

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

Remark that Assumption **(GLM1)** is verified when for all $y$, $\nabla_h^2 \ell(y, .)$ is Lipschitz and $X$ admits a third order moment, which can be easily verified for the logistic regression for instance. Assumption **(GLM2)** is verified when the random variable $\nabla_h \ell\left(Y, X^T\theta_\sigma\right) X$ admits a moment of order $2p$.

**Theorem 5.1.** *Suppose Assumptions **(GLM1)** and **(GLM2)** hold. Then,*

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

*where* $H_\sigma = \mathbb{E}\left[\nabla_h^2 \ell\left(Y, X^T\theta_\sigma\right) XX^T\right] + \sigma I_d, \Sigma_\sigma = \mathbb{E}\left[\left(\nabla_h \ell\left(Y, X^T\theta_\sigma\right) X + \sigma\theta_\sigma\right)\left(\nabla_h \ell\left(Y, X^T\theta_\sigma\right) X + \sigma\theta_\sigma\right)^T\right]$, $K_{1,GLM}^{(3)}, K_{1',GLM}^{(3)}, K_{2,GLM}^{(3)}, K_{2',GLM}^{(3)}, d_{n,GLM}$ *are defined in equations* (50), (51) *and* (52), *and* $v_{l,n}$ *is defined in Proposition 6.5.*

## 5.2 Adagrad algorithm

For generalized linear model, Adagrad algorithm is defined for all $n \geq 0$ by

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

where $\bar{D}_n$ is diagonal and for $\gamma > 1/2$,

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

for some $0 < \beta < \gamma - 1/2$, and for $\gamma \leq 1/2$,

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

where $0 < \beta < (\gamma - \lambda')/2$ and $a_k > 0$. The usual Adagrad algorithm is done with $\gamma = 1/2$, which yields for us

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

Like the linear regression, the general linear model needs minimal randomness to ensure the expected rate of convergence of Adagrad. Indeed, in the extreme case of a deterministic sequence $(X_n, Y_n)_{n\geq 0}$, Adagrad algorithm may diverge in the unfortunate situation where $\nabla_h \ell\left(Y_{i+1}, \theta_i^T X_{i+1}\right)(X_{i+1})_k$ vanishes or remains very small. Such behavior can be averted by requiring at the minimizer $\theta_\sigma$ a minimal variance for $\nabla_h \ell\left(Y, \theta_\sigma^T X\right)(X)_k$ for all $1 \leq k \leq d$.

**(GLM3)** There is a positive constant $\alpha_\sigma > 0$ such that for all $1 \leq k \leq d$

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

Remark that

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

so that **(GLM3)** can be seen as a mirror assumption to **(GLM2)**. We should stress that the existence of such $\alpha_\sigma$ is almost automatic when a minimal randomness between $X$ and $Y$ is assumed. Indeed, having $\nabla_h l\left(Y, \theta_\sigma^T X\right) X_k$ deterministic would imply an analytic relation between $Y$ and $X$. The main computational issue is to estimate a concrete value of $\alpha_\sigma$. An example dealing with the logistic regression is given in Section E.

When **(GLM3)** is assumed, one can show using Theorem 5.2 that there exists almost surely $n_0 \geq n$ such that for $n \geq n_0$,

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

so that we recover the usual Adagrad algorithm for large $n$. We can then rewrite Theorem

3.4 for $\gamma \leq 1/2$ as follows (remark that similar statements hold for $\gamma > 1/2$).

**Theorem 5.2.** *Suppose Assumptions **(GLM1)**, **(GLM2)** and **(GLM3)** hold. Then, for $\gamma = 1/2$, we have*

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

*where $\varepsilon(n) = o(1)$, $K^{ada}_{1,GLM}$, $K^{ada}_{2,GLM}$ and $K^{ada}_{3,GLM}$ have explicit formulas depending on the parameters of the model.*

We do not specify the exact value of the constants here, since they can easily be obtained along the lines of previous results.

# 6 Proofs

Throughout our proofs, to alleviate notations, we will denote by the same way $\|.\|$ the euclidean norm of $\mathbb{R}^d$ and the spectral norm for square matrices. In addition, we will regularly use the following technical result from (Godichon-Baggioni et al., 2021, Proposition A.5).

**Proposition 6.1.** *Let $(\gamma_t)_{t\geq 1}$, $(\eta_t)_{t\geq 1}$, and $(\nu_t)_{t\geq 1}$ be some positive and decreasing sequences and let $(\delta_t)_{t\geq 0}$, satisfying the following:*

- *The sequence $\delta_t$ follows the recursive relation:*

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

  *with $\delta_0 \geq 0$ and $\omega > 0$.*

- *Let $\gamma_t$ and $\eta_t$ converge to 0.*

- *Let $t_0 = \inf\{t \geq 1 : \eta_t \leq \omega\}$, and let us suppose that for all $t \geq t_0 + 1$, one has $\omega\gamma_t \leq 1$.*

*Then, for all $t \in \mathbb{N}$, we have the upper bound:*

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

*with the convention that $\sum_{t_0}^{t/2} = 0$ if $t/2 < t_0$.*

Moreover, we denote by $C_1, C_1', C_2, C_2'$ constants such that for all $h \in \mathbb{R}^d$,

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

### 6.1 Proof of Theorem 3.1

Remark that thanks to a Taylor's expansion of the gradient, denoting $V_n = G(\theta_n) - G(\theta)$ and $g'_{n+1} = \nabla_h g(X_{n+1}, \theta_n)$,

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

where we used Hypothesis **(H1b)** on the last line. Then, taking the conditional expectation, thanks to assumption **(A1)**, and since $\|\theta_n - \theta\|^2 \leq \frac{2}{\mu} V_n$,

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

Furthermore, since $G$ is $\mu$ strongly convex, it comes

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

with $\lambda_n = \lambda_0 (n+1)^\lambda$ with $0 \leq \lambda < 1 - \gamma$. Applying Cauchy-Schwarz yields

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

with $V^2 \geq \sup_{n \geq 0} \mathbb{E}[V_n^2]$ calculated later. Then, Assumption **(H1a)** gives $\mathbb{P}[\lambda_{\min}(A_n) < \lambda_n] \leq v_{n+1}(n+1)^{-\delta - q\lambda} := \bar{v}_n$, so that

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

In order to apply Proposition 6.1, let us denote

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

the last upper bound being added so that the terms of (18) below satisfy the third condition of Proposition 6.1. Set $\tilde{\gamma}_n = c_\gamma n^{-(\lambda+\gamma)}$, and remark that

$$\mathbb{E}\left[V_{n+1}\right] \leq \left(1 - 2\mu\lambda_0\tilde{\gamma}_{n+1} + C_M(n+1)^{2\beta+\lambda-\gamma}\tilde{\gamma}_{n+1}\right)\mathbb{E}\left[V_n\right] + 2\lambda_0\mu V\sqrt{v_n}\tilde{\gamma}_{n+1} + \frac{C_1 L_{\nabla G}}{2}(n+1)^{\lambda}\gamma_{n+1}\beta_{n+1}^2\tilde{\gamma}_{n+1}. \tag{18}$$

Then, since $2\gamma - 2\beta - 1 \neq 1$, with the help of Proposition 6.1 and an integral test for convergence to get $\sum_{k=1}^{n} k^{2\beta-2\gamma} \leq 1 + \frac{n^{(1+2\beta-2\gamma)^+}}{|2\gamma-2\beta-1|}$ and $\sum_{t=\lfloor n/2\rfloor}^{n} t^{-\gamma} \geq \frac{1-2^{\gamma-1}}{1-\gamma} n^{1-\gamma} \geq n^{1-\gamma}$ for $\gamma \in (0,1)$ ,

$$\mathbb{E}\left[V_n\right] \leq \exp\left(-c_\gamma\mu\lambda_0 n^{1-(\lambda+\gamma)}\right)\exp\left(2C_M c_\gamma\left(1 + \frac{n^{(1+2\beta-2\gamma)^+}}{|2\gamma-2\beta-1|}\right)\right)\cdot \left(\mathbb{E}\left[V_0\right] + 4\frac{\lambda_0\mu V}{C_M}\max_{1\leq k\leq n} k^{\gamma-2\beta-\lambda}\sqrt{\bar{v}_k} + \frac{C_1 L_{\nabla G} c_\gamma c_\beta^2}{C_M}\right) + 2V\sqrt{\bar{v}_{n/2}} + \frac{C_1 L_{\nabla G}}{2^{1+\lambda}\mu\lambda_0} n^{\lambda}\beta_{n/2}^2\gamma_{n/2}, \tag{19}$$

where we recall that $\bar{v}_n = v_{n+1}(n+1)^{-\delta-q\lambda} \geq \mathbb{P}\left[\lambda_{\min}\left(A_n\right) < \lambda_n\right]$. Remark that

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

so that $\max_{0\leq k\leq n}(k+1)^{\gamma-2\beta-\lambda}\sqrt{\bar{v}_k} = \max_{1\leq k\leq n+1} k^{\gamma-2\beta-\delta/2-(q/2+1)\lambda}\sqrt{v_k}$. Hence, we get

$$\mathbb{E}\left[V_n\right] \leq \exp\left(-c_\gamma\mu\lambda_0 n^{1-(\lambda+\gamma)}\right)\exp\left(2C_M c_\gamma\left(1 + \frac{n^{(1+2\beta-2\gamma)^+}}{|2\gamma-2\beta-1|}\right)\right) \cdot \left(\mathbb{E}\left[V_0\right] + 4\frac{\lambda_0\mu V}{C_M}\max_{1\leq k\leq n+1} k^{\gamma-2\beta-\delta/2-(q/2+1)\lambda}\sqrt{v_k} + \frac{C_1 L_{\nabla G} c_\gamma c_\beta^2}{C_M}\right) + 2^{1+(\delta+q\lambda)/2}V\sqrt{v_{\lfloor n/2\rfloor}}n^{-(\delta+q\lambda)/2} + 2^{\gamma-2\beta-\lambda-1}\frac{C_1 L_{\nabla G} c_\gamma c_\beta^2}{\mu\lambda_0} n^{2\beta+\lambda-\gamma}$$

where $V$ is defined in Lemma 6.1. Hence, as long as $\gamma + \lambda + (1+2\beta-2\gamma)^+ < 1$ ,which is satisfied since $\lambda < \min\{\gamma - 2\beta, 1-\gamma\}$, we have

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

with

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

$$K_1^{(1)} = \left(\mathbb{E}\left[V_0\right] + \frac{C_1 L_{\nabla G} c_\gamma c_\beta^2}{C_M}\right), \quad K_1^{(1')} = 4\frac{\lambda_0\mu V}{C_M}, \tag{21}$$

where $C_M$ is given in (17) and $V$ in Lemma 6.1 and

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

**Lemma 6.1.** *Suppose Assumption **(A1)** for $p \geq 2$ and **(H1b)** hold. Then, for all $n \geq 0$, if $\gamma > 1/2$ then*

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

*and if $\gamma \leq 1/2$ then*

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

*with $a_2$ given in (67) and $a_p$ is given by (66) for $p > 2$.*

The proof of this Lemma is given in Section B.

### 6.2 Proof of Theorem 3.2

Remark that thanks to Assumption **(H1b)**, one has

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

Moreover, with the help of Assumption **(H2a)**,

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

leading as in the proof of Theorem 3.1 to

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

Using Hölder inequality with $p$ yields then

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

with $\bar{v}_n = v_{n+1}(n+1)^{-\delta}$ and $V_p$ given in Lemma 6.1. Considering $C_M$ defined by

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

one has

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

Then, applying Proposition 6.1 and with the help of integral tests for convergence, it comes

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

Concluding as in the proof of Theorem 3.1, we get

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

with

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

where $C_M$ is defined by (23) and

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

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

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

### 6.3 Proofs of Theorem 3.3 and Corollary 3.1

*Proof of Theorem 3.3.* Remark that one can rewrite

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

leading, since $H$ is symmetric, to

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

First, thanks to Assumption **(A3)** and by Cauchy-Schwarz inequality,

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

Then, using Assumption **(A1')**, one has

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

Finally, one has

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

with, using Assumption **(A4)**, $\|\delta_n\| := \|\nabla G(\theta_n) - H(\theta_n-\theta)\| \leq L_\delta\|\theta_n-\theta\|^2$. Hence,

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

which yields, using that $\|\theta_n-\theta\|^3 \leq \frac{1}{2a}\|\theta_n-\theta\|^2 + \frac{a}{2}\|\theta_n-\theta\|^4$ with $a = \left\|H^{-1}\right\|L_\delta$,

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

Furthermore,

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

As a conclusion, one has (after using Cauchy-Schwartz inequality on $(*)$),

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

leading, using Proposition 3.2 with the fact that $\mathbb{E}\left[\|\theta_n-\theta\|^4\right] \leq \frac{4}{\mu^2}\mathbb{E}\left[V_n^2\right]$ by **(A2)**, and **(H2b)** and **(H3)**, to

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

Finally, let us denote $C_A = c_\gamma \max\left\{4\left\|H^{-1}\right\|^2 L_{\nabla g}, \frac{1}{4}\right\}$. Then, with the help of Proposition 6.1, one has

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

Finally,

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

with

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

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

where we recall that $C_A = c_\gamma \max\left\{4\left\|H^{-1}\right\|^2 L_{\nabla g}, \frac{1}{4}\right\}$, and

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

□

*Proof of Corollary 3.1.* Remark that as long as $\delta^{\frac{p-2}{p}} \geq 2\gamma$, by Proposition 3.2 and the following discussion,

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

Likewise,

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

Hence, plugging these inequalities into Theorem 3.3 yields

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

with

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

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

□

### 6.4 Proof of Theorem 3.4

To prove this theorem, we will apply Theorem 3.2. We first need to check that $(A_n)_{n\geq 0}$ satisfies Assumptions **(H1a), (H1b)** and **(H2)**. Assumption **(H1b)** is given by construction (see (7)) while **(H1a)** is given by the following lemma:

**Lemma 6.2.** *Assume* ***(A1)*** *is satisfied for some $p > 2$. Then, for all $0 < t < 1$,*

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

*with*

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

The proof is given in Appendix B. Remark that $\mathbb{E}\left[V_n^p\right] < +\infty$ by Lemma 6.1 with **(A1)**. Assume from now that $p > 2$ and let $p' = \frac{2(1-\gamma)}{2-\gamma}p$ and $\lambda = (1-\gamma)(\gamma - 2\beta)$. Remark that

$\lambda < 1-\gamma$, $\lambda < \gamma - 2\beta$ and $p' < p$. Hence, applying Proposition 3.1 with $\lambda_0 = c_\beta$, $\delta = 0$, $q = 2p$,

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

with $\epsilon'(n)$, $K_1^{(1')}$, $K_{1'}^{(1')}$, $K_2^{(1')}$ and $K_3^{(1')}$ respectively given in (57), (58) and (60) with $\lambda_0 = c_\beta$. By the choice of $\lambda, p'$ one has

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

so that

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

with

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

By strong convexity, one can so obtain a first rate of convergence of the estimates. The following lemma enables to ensure that **(H1a)** is satisfied, but with a possibly better rate than with Lemma 6.2.

**Lemma 6.3.** *Assume **(A1)** is satisfied for some $p > 2$. Then,*

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

*with $\tilde{\lambda}_0 = \left[\frac{2(1-\gamma)}{2-\gamma}p\left(C_{\left(\frac{2(1-\gamma)}{2-\gamma}\right)}+1\right)\right]^{-\frac{2-\gamma}{2(1-\gamma)p}}$ and $v_0$ is given in (73) with $p' = \frac{2(1-\gamma)}{2-\gamma}p$.*

The proof is given in Appendix B. We can also deduce from (34) a bound on $\mathbb{E}\left[\|A_n\|^4\right]$ in case only **(A6)** holds.

**Lemma 6.4.** *Assume Assumptions **(A1)-(A6)** and **(A1')** hold for some $p > 2$. Then, for $\beta < \min\left\{\frac{(1-\gamma)\gamma(\gamma-2\beta)p}{4(2-\gamma)}, 1/4\right\}$, the sequence of random matrices $(A_n)$ defined by (4) verifies*

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

*with $C_S^4$ given in (74).*

The proof is given in Appendix B. If the stronger hypothesis **(A6')** holds, an improved and simpler bound on $\mathbb{E}\left[\|A_n\|^4\right]$ can be reached, as next lemma shows.

**Lemma 6.5.** *Assume Assumptions* ***(A1)-(A6')*** *and* ***(A1')*** *hold for some* $p > 2$. *Then, for* $\beta < \min\{\gamma/2 \wedge 1/4\}$, *the sequence of random matrices* $(A_n)$ *defined by* (4) *verifies*

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

*with* $C_S^4$ *given in* (75).

The proof is given in Appendix B. Theorem 3.4 is then a consequence of Theorem 3.2 whose hypotheses are satisfied thanks to Lemma 6.2, 6.3 and 6.4 (or 6.5). We then have

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

with $K_1^{(2)}, K_{1'}^{(2)}, K_2^{(2)}$ and $K_3^{(2)}$ respectively given in (26), (27) and (28) with $\delta = \min\left\{\frac{2(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma}, 1\right\}$, $\lambda_0$ given in (72), $v_n = v_0 \log(n+1)$ with $v_0$ given in (73) and $C_S$ given in (74) or (75) depending on whether **(A6)** or **(A6')** holds. By strong convexity

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

with $\tilde{\lambda}_0$ defined in (72)

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

with

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

where $v_0$ is given in (73).

### 6.5 Proofs of Theorem 4.1 and Theorem 4.2

The proof relies on the verification of each assumption needed in Theorem 3.3.

**Verifying Assumptions (A1), (A1') to (A6).** First, remark that

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

Then, if $X$ and $\epsilon$ respectively admit moments of order $4p$ and $2p$, since $\epsilon$ and $X$ are independent,

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

with $\sigma_{(t)} = 2^{t-1}\mathbb{E}\left[|\epsilon|^t\right]\mathbb{E}\left[\|X\|^t\right]$ and $C_{(t)} = 2^{t-1}\mathbb{E}\left[\|X\|^{2t}\right]$. In a particular case, if $p \geq 2$,

Assumption **(A1)** is verified. Furthermore, since for all $h$, $\nabla^2 G(h) = \mathbb{E}\left[XX^T\right]$ is positive, **(A2)** to **(A4)** hold with $\mu = \lambda_{\min}\left(\mathbb{E}\left[XX^T\right]\right) =: \lambda_{\min}$, $L_{\nabla G} = \lambda_{\max}\left(\mathbb{E}\left[XX^T\right]\right) =: \lambda_{\max}$ and **(A5)** holds with $L_\delta = 0$. Finally Assumption **(A1')** is verified since

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

We can now prove Theorem 4.1

*Proof of Theorem 4.1.* **Verifying Assumption (H1) for Stochastic Newton algorithm.** Let us first check Assumption **(H1)** for $\widetilde{S}_n = \frac{1}{n+1}\left[S_0 + \sum_{i=1}^n X_i X_i^T\right]$.

**Lemma 6.6.** *Suppose that $X$ admits $4p$-moments, with $p > 2$. Then, for $\lambda_0 = \frac{1}{2\mathbb{E}\|X\|^2}$, we have*

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

*with*

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

*where $Z = \|X\|^2 - \mathbb{E}\left[\|X\|^2\right]$ and $C_1(p), C_2(p)$ are numerical constants given in Rosenthal inequality, see Pinelis (1994).*

The proof is given in Section C. To deal with $\overline{S}_n = \frac{\|\widetilde{S}_n^{-1}\|}{\min(n^\beta, \|\widetilde{S}_n^{-1}\|)}\widetilde{S}_n$, one first needs the following control on the behavior of $\lambda_{\min}(\widetilde{S}_n)$. Set $H = \mathbb{E}\left[XX^T\right]$.

**Proposition 6.2** (See Koltchinskii and Mendelson (2015), Theorem 1.5 and Theorem 3.3)**.** *Suppose that $0 < \lambda_{\min} I_d \leq H := \mathbb{E}\left[XX^T\right] \leq \lambda_{\max} I_d$ and that there exists $L_{MK} > 0$ such that $\mathbb{E}\left[\langle X, t\rangle^2\right] \leq L_{MK}\mathbb{E}\left[|\langle X, t\rangle|\right]$ for all $t \in \mathbb{S}^{d-1}$. Then, for $n \geq c_1 d$,*

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

*with $c_1 = \frac{\lambda_{\max}^2 (16 L_{MK})^4}{\lambda_{\min}^2}$, $c_2 = \frac{\lambda_{\min}}{8\sqrt{2}L_{MK}^2}$ and $c_3 = \frac{1}{128 L_{MK}^4}$.*

Remark that the constant $c_1$, $c_2$ and $c_3$ are fairly explicit in terms of $L_{MK}$ and $\lambda_{\min}$. For the latter result and Lemma 6.6 and Proposition 6.2 we deduce Hypothesis (H1) for $\overline{S}_n$. We will need several times the threshold

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

**Lemma 6.7.** *Suppose that $X$ satisfies hypothesis of Proposition 6.2 and admits $4p$-moments, with $p > 2$. Then, for $\lambda_0 = \frac{1}{2\mathbb{E}[\|X\|^2]}$, we have*

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

*with $\delta = p/2$, $v_{n+1} = (n+1)^{\delta}$ for $n \leq n_0$ and, for $n > n_0$,*

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

*where $c_1$, $c_2$, $c_3$ are given in Proposition 6.2, $C_1(p)$ and $C_2(p)$ are numerical constants depending on $p$ and $Z = \|X\|^2 - \mathbb{E}\left[\|X\|^2\right]$.*

The proof is given in Section C. As a particular case, Assumption **(H1a)** is verified with a rate $\delta = p/2$ when $\gamma > 1/2$.

**Verifying Assumption (H2) for Stochastic Newton algorithm.** A straightforward deduction of the above lemma is the following.

**Lemma 6.8.** *Suppose that hypothesis of Proposition 6.2 hold and that $X$ admits a moment of order $4p$ with $p > 2$. Then, for all $\kappa > 0$, we have*

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

*for $n \geq c_1 d$ and*

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

*for $n \leq c_1 d$, with $c_1$, $c_2$, $c_3$ given in Proposition 6.2.*

The proof is given in Section C. Finally, the following proposition gives a precise bound for Assumption **(H2)**.

**Proposition 6.3.** *Suppose that hypothesis of Proposition 6.2 hold and that $X$ admits a moment of order $4p$ with $p > 2$. Then*

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

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

The proof is given in Section C. Remark that $C_S = O(d)$.

**A first convergence result.** Since in the case of the linear model, one as $C_1 = \sigma_{(2)}, C_1' = \sigma_{(4)}, C_2 = C_{(2)}, C_2' = C_{(4)}, L_{\nabla G} = \lambda_{\max}, \mu = \lambda_{\min}, \lambda_0 = \frac{1}{2\mathbb{E}\left[\|X\|^2\right]}$ $\delta = p/2$, Proposition 3.2 can now be written as follows:

**Proposition 6.4.** *Suppose that there is $p > 2$ such that $X, \epsilon$ respectively admit moments of orders $4p$ and $2p$. Suppose also that there is a positive constant $L_{MK}$ such that for any $h \in \mathbb{S}^{d-1}$, $\sqrt{\mathbb{E}\left[hXX^Th\right]} \leq L_{MK}\mathbb{E}\left[\left|X^Th\right|\right]$. Then, denoting $\lambda_{\min}$ and $\lambda_{\max}$ the smallest and largest eigenvalues of $\mathbb{E}\left[XX^T\right]$,*

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

*with $v_n$ given by Lemma 6.7 and*

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

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

*where, recalling the notations $\sigma_{(t)} = 2^{t-1}\mathbb{E}\left[|\epsilon|^t\right]\mathbb{E}\left[\|X\|^t\right]$ and $C_{(t)} = 2^{t-1}\mathbb{E}\left[\|X\|^{2t}\right]$,*

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

*with $C_S$ given by Proposition 6.3, $a_{1,lin} := C_S^4\lambda_{\max}^2\left(\frac{16\lambda_{\max}^2\sigma_{(2)}^2\mathbb{E}\left[\|X\|^2\right]}{\lambda_{\min}^3} + \frac{\sigma_{(4)}c_\gamma}{2} + \frac{2C_{(2)}^2\mathbb{E}\left[\|X\|^2\right]}{\lambda_{\min}}\right)$ and*

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

*where*

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

**Verifying Assumption (H3) for Stochastic Newton algorithm.** Hypothesis **(H3)** is then a straightforward combination of the convergence of $\overline{S}_n$ towards $H$, together with Hypothesis **(H2)**.

**Lemma 6.9.** *Suppose that $X$ admits moments of order $2p$ with $p > 4$, and let suppose as well that the distribution of $X$ satisfies hypothesis of Proposition 6.2. Then, for $n \geq n_0$ (with $n_0$ defined in (37)),*

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

For $n < n_0$, we simply bound

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

By Lemma 6.7, **(H1a)** is satisfied with $\delta = p/2$. Applying Theorem 3.3 with the constants computed in the previous lemmas and proposition, we get finally,

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

with $v_{H,n}$ defined by (39), recalling that $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and largest eigenvalues of $\mathbb{E}\left[XX^T\right]$, and since for the linear case one has $C_A = 4c_\gamma\frac{\mathbb{E}\left[\|X\|^4\right]}{\lambda_{\min}^2} \geq 4c_\gamma$,

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

and $c_{n,\text{lin}}$ and $C_S^4$ are respectively defined in Propositions 6.4 and 6.3. □

*Proof of Theorem 4.2.* Let us first prove that Assumption **(A6')** is fulfilled. For all $h$,

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

and **(A6')** is satisfied with $\alpha = \mathbb{E}\left[\epsilon^2\right]\lambda_{\min}$, we have by (75),

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

with $V_2$ given by Lemma 6.1 for $p = 2$. Then, applying Theorem 3.4,

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

with $\lambda_{0,lin}^{ada} = \left[\frac{4(1-\gamma)p}{2-\gamma}\left(C_{\left(\frac{4p(1-\gamma)}{2-\gamma}\right)}+1\right)\right]^{-\frac{2-\gamma}{4p(1-\gamma)}}$, and recalling that $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and largest eigenvalues of $\mathbb{E}\left[XX^T\right]$,

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

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

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

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

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

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

where

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

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 Proof of Theorem 5.1

The proof relies on the verification of each Assumption in Theorem 3.3.

**Verifying Assumptions (A1), (A1') to (A6).** First, remark that taking for all $0 \leq a \leq 2p$, one has

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

and Assumption **(A1)** is so verified. In a same way,

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

and **(A1')** is so verified. Remark that **(A2)** and **(A4)** are verified by hypothesis with $\mu = \sigma$, while for **(A3)**, one has

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

Observe that Assumption **(A5)** is given by **(GLM1)** while for Assumption **(A6)**, **(GLM3)** together (12), which yields

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

for all $1 \leq k \leq d$.

**Verifying Assumption (H1).** The following lemma ensures that Assumption **(H1)** is fulfilled.

**Lemma 6.10.** *Assume first (8) and that $X$ admits a moment of order $2p$ for some $p < 0$. In the regularized case defined by (10), denoting $\lambda_0 = \frac{1}{2L_{\nabla l}\mathbb{E}\left[\|X\|^2\right] + 2\lambda}$, we have*

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

*with*

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

*where $T = L_{\nabla l}\left(\|X\|^2 - \mathbb{E}\left[\|X\|^2\right]\right) + \sigma\left(\|Z\|^2 - 1\right)$ and $Z$ being a standard $d$-dimensional random variable independent of $X$. In addition, $C_1(p)$ and $C_2(p)$ are given in Pinelis (1994).*

The proof is given in Appendix D. Observe that if $p > 4\gamma$, one has $v_n = o(\gamma_n)$.

**Verifying Assumption (H2).** The following proposition ensures that **(H2)** is fulfilled.

**Proposition 6.5.** *Considering from the regularized problem given by (10), one has for all $n \geq 0$,*

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

**Remark 6.1.** *Remark that if (9) holds for some constant $\alpha > 0$ and if $\mathbb{E}\left[XX^T\right]$ is positive, under hypothesis of Proposition 6.2, for all $n \geq 0$ and for $\sigma = 0$, one has*

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

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

**A first result**

Remark that one can rewrite Proposition 3.2 as follows:

**Proposition 6.6.** *Suppose there exist $p > 2$ such that $X$ admits a $2p$-th order moment and that there is $L_\sigma$ verifying*

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

*Then,*

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

*with $v_n$ defined in Lemma 6.10, $C_{S,\sigma}$ defined in Lemma 6.5, $C_{GLM}$ and $C_{GLM}^{(a)}$ defined in equations*

*(48) and (47),*

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

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

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

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

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

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

*with* $V_{p,GLM}^p = e^{a_{p,GLM} c_\gamma^2 C_{S,\sigma}^2 \frac{2\gamma}{2\gamma-1}} \max\left\{1, \mathbb{E}\left[V_0^p\right]\right\}$ *where*

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

**Verifying Assumption (H3).** We prove here that **(H3)** holds for general linear models. We now denote

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

**Proposition 6.7.** *Suppose Assumptions **(GLM1)** and **(GLM2)** hold, then for all* $n \geq 0$,

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

*with* $v_{i,GLM}$ *defined in Proposition 6.6.*

We can now finish the proof of Theorem 5.1. In this aim, let us first remark that for all $h, h'$,

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

Then, with the help of Theorem 3.3, one has

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

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 $c_\gamma 4\frac{C^{(2)}_{\mathrm{GLM}}}{\sigma^2} \geq C_{A,\mathrm{GLM}} = \geq 4c_\gamma$,

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

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

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

*Proof of Theorem 5.2.* The proof follows exactly the same pattern as the proof of Theorem 4.2, using Assumption **(A6)** together with Lemma 6.4 to compute the constant $C_S$ such that **(H2)** is satisfied. □

# A Proofs of technical proposition

## A.1 Proof of Proposition 3.1

Let us recall that

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

Remark that for $a \geq 2$ and $x, h \in \mathbb{R}$ such that $x \geq 0$ and $x + h \geq 0$, we have by Taylor's expansion

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

This yields for $a = p'$, $x = V_n$ and $h = U_{n+1}$ and after conditioning on $\mathcal{F}_n$

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

Since $G$ is convex and $\nabla G$ is Lipschitz,

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

By strong convexity, we have

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

where $\lambda_n = \lambda_0(n+1)^\lambda$ with $0 \leq \lambda < \min\{\gamma - 2\beta, 1-\gamma\}$. Applying Hölder inequality yields then

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

with $V_p^p \geq \sup_{n\geq 0}\mathbb{E}[V_n^p]$ given by Lemma 6.1. Then, Assumption **(H1a)** gives $\mathbb{P}\left[\lambda_{\min}(A_n) < \lambda_n\right] \leq v_{n+1}(n+1)^{-\delta-q\lambda} := \bar{v}_n$, so that finally

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

Furthermore, since $\nabla G$ is $L_{\nabla G}$-Lipschitz, one has

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

Hence, using **(H1b)** and the strong convexity of $G$ yields

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

Specializing the latter inequality with $p' = 2$ yields then (recalling inequalities (14))

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

so that

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

Using the latter inequality with (55) in (54) yields then

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

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

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

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

and

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

Applying now Young's inequality, which implies $a^ib^{p'-i}\leq\frac{ia^{p'}}{p'}+\frac{(p'-i)b^{p'}}{p'}$ for $0<i<p'$ and $a,b\geq0$, yields for any $t>0$ and $i\in\{1,2,p'/2\}$

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

so that using the latter inequality with $t=\frac{p'^2\mu}{3(p'-i)\mu}$ for $i\in\{1,2,p'/2\}$ and using that

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

gives

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

with

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

Putting together the previous inequalities and taking the expectation yield then

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

Then, recalling that $\bar{v}_n = v_{n+1}(n+1)^{-\delta-q\lambda}$ and using Proposition 6.1 yields

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

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

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

where

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

and

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

where $V_p$ is given in Lemma 6.1.

## A.2 Proof of Proposition 3.2

Remark that with the help of a Taylor's expansion of $G$, one has

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

Then, using (56) one has

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

We now bound $(\star)$ and $(\star\star)$. First, thanks to Assumption **(H1)** and since $\left\|\theta_n - \theta\right\|^2 \leq \frac{2}{\mu} V_n$, one has

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

Then, taking the expectation with Assumption **(H2b)**,

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

Moreover, since $\nabla G$ is $L_{\nabla G}$-Lipschitz, one can check that

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

Then, one has

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

Furtermore, with the help of inequality (16) it comes

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

Then, with the help of Holder's inequality, coupled with **(H1a)** for $t = 1$, one has

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

with $V_p$ defined in Lemma 6.1 and $\bar{v}_n := v_n(n+1)^{-\delta}$ is the upper bound from **(H1a)** on $\mathbb{P}\left[\lambda_{\min}\left(A_n\right) \leq \lambda_0\right]$. Let

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

one has

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

Applying Proposition 6.1, it comes (with analogous calculus to the ones in the proof of Theorem 3.1)

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

where $V_p$ is given by Lemma 6.1 and $\bar{v}_{\lfloor n/2\rfloor} \leq v_{n/2}2^\delta(n+1)^{-\delta}$. Setting

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

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

with $a_M$ given in (61), $a_1$ given in (62) and $V_p$ given in Lemma 6.1, and

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

we finally get

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

# B Proofs of tehcnical lemmas

## B.1 Proof of Lemma 6.1 (Nouvelle version)

Observe that since the proofs are analogous, we only make the proof for $p > 2$, and for the case where $p = 2$, if there are some differences in the proof, it will be indicated with the help of remarks.

With the help of a Taylor expansion of the functional $G$, one has

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

Then, applying the inequality

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

for $a, a+h \geq 0$ to $a = V_n$ and $h = -\gamma_{n+1}\left(g'_{n+1}\right)^T A_n \int_0^1 (1-t)\nabla G\left(\theta_n + t\left(\theta_{n+1} - \theta_n\right)\right) dt$, one has

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

**Remark B.1.** *Observe that in the case where $p = 2$, one has*

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

*the last term on the right hand-side of previous inequality can be considered equal to* 0.

Recalling that since $\nabla G$ is $L_{\nabla G}$-Lipschitz, one has

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

which implies

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

Furthermore, one has

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

Since $A_n$ is positive and since $\nabla G$ is $L_{\nabla G}$-lipschitz, taking the conditional expectation, it comes, since for all $a, b \geq 0$, $ab \leq \frac{1}{p}a^p + \frac{p-1}{p}b^{p/(p-1)}$ and with the help of Assumption **(H1a)**,

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

with $\lambda'_n = \lambda'_0 n^{-\lambda'}$. We also used Assumptions **(A1)** on the first inequality and the fact that $\|\theta_n - \theta\|^2 \leq \frac{2}{\mu} V_n \leq \frac{2}{\mu^2} \|\nabla G\left(\theta_n\right)\|^2$ on the third inequality. For the same reasons, one has

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

In a same way, thanks to Assumptions **(A1")** and **(H1)**, one has

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

Taking the expectation on $\mathbb{E}\left[(*)|\mathcal{F}_n\right]+\mathbb{E}\left[(**)|\mathcal{F}_n\right]+\mathbb{E}\left[(***)|\mathcal{F}_n\right]$, applying the latter inequalities, it comes

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

with

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

**Remark B.2.** *Observe that in the case where $p=2$, one has*

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

If $\gamma > 1/2$, by summation,

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

If $\gamma \leq 1/2$, let $n_0$ be the smallest integer such that $\gamma_{n+1}^2\beta_{n+1}^2a_p > p\mu\lambda'_n\gamma_{n+1}$. Recording that $\lambda'_n = \lambda'_0(n+1)^{-\lambda'}$, we have $n_0 = \left\lfloor\left(\frac{c_\gamma c_\beta^2a_p}{p\mu\lambda'_0}\right)^{\frac{1}{\gamma-2\beta-\lambda'}}\right\rfloor$. Then,

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

## B.2 Proof of Lemma 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 $\left(\overline{A_n}\right)_{kk'} = \frac{\delta_{kk'}}{\sqrt{\frac{1}{n+1}\left(a_k + \sum_{i=0}^{n-1} (\nabla_h g(X_{i+1}, \theta_i)_k)^2\right)}}$.
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 $0 < t < 1$

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

Then, Markov inequality for $p > 2$ and Jensen inequality yields

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

Then, using Assumption **(A1)** and then **(A2)** we get

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

By the bound $\mathbb{E}\left[V_n^p\right] \leq V_p^p$ from Lemma 6.1, we finally get

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

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 Proof of Lemma 6.3

Set $E_k = \mathbb{E}\left[\nabla_h g\left(X,\theta\right)_k^2\right]$ and $\partial_{kg}^2 g(h) = \mathbb{E}\left[\nabla_h g(X,h)_k^2\right]$. Then, by Jensen's inequality for $p' \geq 2$,

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

Hence, for any $x > 0$,

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

Set $M_0 = 0$ and for $n \geq 1$,

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

Then, $(M_n)_{n\geq 0}$ is a martingale, and thus by Burkholder's inequality, see (Hall and Heyde, 2014, Theorem 2.10) there exists an explicit constant $C_{p'}$ such that

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

where we used Jensen's inequality on the second inequality. By Assumption **(A1)**, the strong convexity of $G$ and Lemma 6.1,

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

Hence,

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

which yields for $x > 0$

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

Next, by Jensen inequality,

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

Using Assumption **(A1)** and then strong convexity yields

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

so that

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

Hence, for $\frac{x^{2p'}}{2p'} > C_1^{(p')}$,

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

By (34) and the fact that $\frac{1}{n+1}\sum_{i=0}^{n-1}(i+1)^{-\frac{2(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma}} \leq \frac{1}{n+1} + \frac{1}{\left|1-\frac{2(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma}\mathbf{1}_{\frac{2(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma}\neq 1}\right|}\frac{\log(n+1)}{(n+1)^{\frac{2(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma}\wedge 1}}$,
and denoting $\tilde{1} = 1 + \frac{1}{\left|1-\frac{2(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma}\mathbf{1}_{\frac{2(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma}\neq 1}\right|}$, it comes

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

with for $n \geq 2$

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

Choosing

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

yields then

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

Putting the latter inequality with (69) and (71) gives then

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

with

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

Since $\mathbb{P}\left[ \lambda_{\min}(A_n) < \lambda_0 \right] \leq \mathbb{P}\left[ \lambda_{\min}\left( \overline{A_n} \right) < \lambda_0 \right]$, the result is deduced.

### B.4 Proof of Lemma 6.4

Set $E_k = \mathbb{E}\left[ \nabla_h (X, \theta)_k^2 \right]$ and $\partial_k^2 g(h) = \mathbb{E}\left[ \nabla_h (X, h)_k^2 \right]$. Then

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

By (70),

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

Next, by Jensen inequality,

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

Using Cauchy-Schwarz inequality, Assumption **(A1')** and then Assumption **(A1)** yields

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

where $c_i$ is given in (34). Putting all the latter bounds together yields, using that $E_k \leq C_1$,

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

Hence, noting that $V_p < \infty$ by Assumption **(A1')** and Lemma 6.1,

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

with, by (34), $\bar{c}_n = O\left(\log(n)n^{-\left[\frac{(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma}\wedge 1\right]}\right)$. Since by **(A6)** we have $E_k \geq \alpha$, we deduce by Markov's inequality that

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

Hence, we have

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

Since $\bar{c}_n = O\left(\log(n)n^{-\left[\frac{(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma}\wedge 1\right]}\right)$, for $\beta<\frac{(1-\gamma)\gamma(\gamma-2\beta)p}{4(2-\gamma)}\wedge\frac{1}{4}$ we have $\left[\frac{(1-\gamma)\gamma(\gamma-2\beta)p}{2-\gamma}\wedge 1\right]-4\beta>0$ and thus

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

and finally

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

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 Proof of Lemma 6.5

First, we have by **(A6')**

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

Then, as in the proof of the previous lemma,

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

Hence, by Markov inequality,

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

We deduce as in the previous lemma that

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

When $\beta < 1/4$, we finally get

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

with

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

# C Proof of technical Lemma and Propositions for linear regression

## C.1 Proof of Lemma 6.6

Remark that

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

Hence, for $\lambda > 0$,

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

Taking $\lambda_0 = \left(2\mathbb{E}\left[\|X\|^2\right]\right)^{-1}$ yields then

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

Taking the $p$-power, applying Markov inequality and then Rosenthal inequality yields that

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

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

## C.2 Proof of Lemma 6.7

By definition of $\overline{S}_n$, $\overline{S}_n = \widetilde{S}_n$ on the event $T_n = \{\lambda_{\min}\left(\widetilde{S}_n\right) \geq \frac{1}{c_\beta n^\beta}\}$. Hence, for the same $\lambda_0$ as in Lemma 6.6,

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

By Lemma 6.6,

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

with $\tilde{v}_n$ given in Lemma 6.6. Then, for $n \geq n_0$, where $n_0$ is defined in (37), we have $n \geq \left(\frac{1}{c_\beta c_2}\left(\frac{n+1}{n}\right)\right)^{-1/\beta}$, and thus $\frac{n}{n+1}c_2 \geq \frac{1}{c_\beta n^\beta}$. In particular, on the event $\left\{\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^n X_i X_i^T\right) > c_2\right\}$, we have

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

Hence, for $n \geq n_0$, $\left\{\lambda_{\min}\left(\frac{1}{n}\sum_{i=1}^n X_i X_i^T\right) > c_2\right\} \subset T_n$ and thus by Proposition 6.2 and the fact that $n \geq c_1 d$,

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

Using (77) and (78) in (76) yields then

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

for $n \geq n_0$. The statement of the lemma is then a rewriting of the latter inequality.

## C.3 Proof of Lemma 6.8

Since we have

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

for $c_1, c_2, c_3$ given in Proposition 6.2, $n \geq c_1 d$ and $\kappa > 0$,

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

Since $\tilde{S}_n = \frac{1}{n+1}\left(S_0 + \sum_{i=1}^n X_i X_i^T\right)$ and $\sum_{i=1}^n X_i X_i^T \geq 0$, we have $\tilde{S}_n \geq \frac{1}{n+1}S_0$ and thus $\|\bar{S}_n^{-1}\| \leq \|\tilde{S}_n^{-1}\| \leq (n+1)\|S_0^{-1}\|$ for $n \geq 1$. Hence, for $n \leq c_1 d$, $\|\tilde{S}_n^{-1}\| \leq (c_1 d + 1)\|S_0^{-1}\|$ and

we finally get the result.

### C.4 Proof of Proposition 6.3

Recall that $\beta_n = c_\beta n^\beta$. Since,for $\kappa > 0$, the map $g : t \mapsto (c_\beta t^\beta)^\kappa \exp(-c_3 t)$ is bounded from above by $c_\beta^\kappa \left(\frac{\beta\kappa}{ec_3}\right)^{\beta\kappa}$, we get

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

Taking into account the case $n \leq c_1 d$ yields then

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

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 Proof of Lemma 6.9

First notice that

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

Under hypothesis of Proposition 6.2,

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

for $n \geq c_1 d$. Since $\left\|\bar{S}_n^{-1}\right\| \leq \left\|\tilde{S}_n^{-1}\right\|$, $\lambda_{\min}\left(\bar{S}_n\right) \geq \lambda_{\min}\left(\tilde{S}_n\right)$ and thus we also have

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

for $n \geq c_1 d$. Hence, for $n \geq n_0$,

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

where we used on the last equality that for $n \geq n_0$, $\bar{S}_n = \tilde{S}_n$ on the event $\{\lambda_{\min}(\bar{S}_n > c_2)\}$, as in the proof of Lemma 6.7. The first summand can be bounded using Hölder inequality

with $\frac{1}{q} + \frac{1}{q'} = 1$ and $q' = p/2$ as

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

Using the upper bound on $H$ and the convexity inequality $(a+b)^p \leq 2^{p-1}(a^p + b^p)$ yields the rough bound

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

Since $X$ admits moments of order $2p$, we get

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

We hence get

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

For the second summand, using the relation between Frobenius norm and operator norm yields

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

Putting all the above bounds together yields the bound of the statement.

# D Proof of technical Lemma and Propositions for generalized linear model

### D.1 Proof of Lemma 6.10

With the help of inequality (8), it comes

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

with $Z_i = e_{i[d]+1}$. Then, a similar proof as the one of Lemma 6.7 yields that for $\lambda_0 = \left(2L_{\nabla l}\mathbb{E}\left[\|X\|^2\right] + 2\sigma\right)^{-1}$,

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

Then, by Markov inequality for $p \geq 1$, we then get

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

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

### D.2 Proof of Proposition 6.5

One directly has for all $n \geq 2d$

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

and $\overline{S}_n \geq \frac{1}{2d}S_0$ for $n \leq 2d-1$, so that

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

### D.3 Proof of Proposition 6.7

Let us denote

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

One can decompose $\overline{H}_n - H(\theta_\sigma)$ as

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

Let us now give a rate of convergence of each term on the right-hand side of previous equality. Set $M_n := \sum_{i=0}^{n-1}\left(\nabla_h^2 \ell\left(Y_{i+1}, \theta_i^T X_{i+1}\right) X_{i+1}X_{i+1}^T - H(\theta_i)\right)$. Since $\mathbb{E}\left[\nabla_h^2 \ell\left(Y_{i+1}, \theta_i^T X_{i+1}\right) X_{i+1}X_{i+1}^T | \mathcal{F}_i\right] = H(\theta_i)$, where $(\mathcal{F}_i)$ is the $\sigma$-algebra generated by the sample, i.e $\mathcal{F}_i := \sigma\left((X_1, Y_1), \ldots, (X_i, Y_i)\right)$. Then, $(M_n)_{n\geq 1}$ is a martingale and thus

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

It then remains to handle $\frac{1}{n+1}\sum_{i=0}^{n-1}\left(H(\theta_i) - H(\theta_\sigma)\right)$. With the help of Assumption **(GLM1)**, one has

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

with $v_{i,\mathrm{GLM}}$ defined in Proposition 6.6. Then, since

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

and

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

it comes

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

Now, notice as in Lemma 6.9 that

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

which yields, thanks to Proposition 6.5

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

i.e one has

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

# E How to verify (GLM3) for the logistic regression

Remark that $\theta_\sigma$ is the unique solution to $\mathbb{E}\left[\nabla_h \ell\left(Y, X^T\theta_\sigma\right)X + \sigma\theta_\sigma\right] = 0$, so that

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

For the logistic regression, we have $Y \in \{-1, 1\}$ and $\nabla_h l\left(Y, X^T\theta_\sigma\right) = \frac{-Y}{1+\exp(-Y\theta_\sigma^T X)}$, and thus we need to get a lower bound on the variance of $\frac{-YX_k}{1+\exp(-Y\theta_\sigma^T X)}$ for all $1 \leq k \leq d$. To guarantee Assumptions **(GLM3)**, we impose a minimal randomness on $(X, Y)$ given by the existence for all $1 \leq k' \leq d$ of $x^{k'}$ $\sigma(Y, X_i, i \neq k')$ measurable bounded by $M$ and an event $A \in \sigma(Y, X_i, i \neq k')$ with $\mathbb{P}\left[A \cap \{|X_i| \leq M, 1 \leq i \leq d, i \neq k'\}\right] > \eta$ and $c, \epsilon > 0$ such that on $A$ we have

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

In particular, since $u \mapsto \frac{u}{1+\exp(-\alpha u)}$ is monotonic for all $\alpha \in \mathbb{R}$ and $\mathbb{C}^1$, there is $y_k$ $\sigma(Y, X_i, i \neq k)$ measurable and $c(M\|\theta_\sigma\|)$ explicitly depending on $M\|\theta_\sigma\|$ such that on $B := A \cap \{|X_i| \leq M, 1 \leq i \leq d, i \neq k\}$,

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

and

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

We deduce that on the event $B$ we have

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

Hence,

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

and we can choose

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

# F Counter-example for the quadratic convergence of the stochastic Newton algorithm without regularization

We show here that even in the simplest case $d = 1$, stochastic Newton algorithm may not converge in quadratic mean. Suppose that we define here the naive Newton adaptive matrix $A_n$

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

Recall that is known (Boyer and Godichon-Baggioni, 2020) that $\theta_n$ converges almost-surely to the minimizer $\theta_0$ at speed $n^{-\gamma}$ for $\gamma \in (1/2, 1)$.

### Counter-example with $\nabla g$ almost everywhere defined

Set $g((x,y),\theta) = (x\theta)^2 + y\lfloor\theta\rfloor\theta$ and let $(X,Y)$ be a random vector with independent coordinates such that $X \simeq Ber(1/2)$ and $\mathbb{P}[Y=1] = \mathbb{P}[Y=-1] = 1/2$. Then, $G(\theta) = \mathbb{E}\left[X^2\right]\theta^2 + \mathbb{E}[Y]\lfloor\theta\rfloor\theta = \theta^2/2$ and we have Lebesgue almost surely $\nabla_h g((x,y),h) = 2x^2h + y\lfloor h\rfloor$ and $\nabla_h^2 g((x,y),h) = 2x^2$.

Let $n \geq 1$. Then, $\mathbb{P}\left[X_1 = 0, \ldots, X_n = 0, Y_1 = -1, \ldots, Y_n = -1\right] = 2^{-2n}$ and on the event $\{X_1 = 0, \ldots, X_n = 0, Y_1 = -1, \ldots, Y_n = -1\}$, as long as $\theta_k \notin \mathbb{N}$ for all $k \geq 0$ (which will be temporarily assumed),

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

Hence, $A_k = k$ and $(\theta_k)_{1\leq k\leq n}$ is defined recursively by

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

If $\gamma_k = k^{-\alpha}$ for some $\alpha < 1$, we then have $k\gamma_k = k^{1-\alpha}$, and thus for $\theta_0 > 1$

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

We deduce that $\theta_n \geq \prod_{k=1}^n (1 + k^{1-\alpha}/2) \geq (n!)^{1-\alpha}2^{-n}$. In particular,

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

when $\theta_k \notin \mathbb{N}$ for all $k \geq 0$. Since for each $k \geq 1$, $\theta_k \notin \mathbb{N}$ for almost every $\theta_0 \in (1,2]$, the latter hypothesis holds for Lebesgue almost every choice of $\theta_0 \in ]1,2]$.

### Counter-example with $\nabla g$ continuous

Let $f$ be such that $f''(\theta) = \mathbf{1}_{\mathbb{Z}+]-1/3,1/3[}$, and set $g((x,y),\theta) = (x\theta)^2 + yf(\theta)$. Let $(X,Y)$ be a random vector with independent coordinates satisfying $X \simeq Ber(1/2)$ and $Y \sim \mathcal{U}([-2,2])$. Then, $G(\theta) = \mathbb{E}\left[X^2\right]\theta^2 + \mathbb{E}[Y]f(\theta) = \theta^2/2$ and $\nabla_h g((x,y),\theta) = 2x^2\theta + yf'(\theta)$. Then, $A_0 = 1$,

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

for $k \geq 1$ and

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

Set $\theta_0 = 3/2$ and $\gamma_k = k^{-\gamma}$ for $k \geq 1$, and consider $(X_i,Y_i)_{0\leq i\leq n}$ satisfying the following conditions:

- $X_i = 0$ for all $1 \leq i \leq n$, which yields $\mathbb{P}\left[X_1 = 0,\dots,X_n = 0\right] = 2^{-n}$ and for all $k \geq 1$,

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

- $\theta_{k-1}$ being known, $Y_k \in \frac{1}{\gamma_k A_{k-1} f'(\theta_{k-1})}\left((\mathbb{Z}+]1/3,2/3[) - \theta_{k-1}\right) \cap [-2,-1] := T_k$ (remark that $T_k$ will be shown to be non-empty).

**Lemma F.1.** *The following facts hold for $k \geq 1$.*

1. $\theta_k \geq k+1$,
2. $A_k = k+1$,
3. *with $\ell$ denoting the Lebesgue measure,*

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

*Proof.* We will prove those three facts by induction on $k \geq 1$. For $k = 1$, we have $A_0 = \gamma_1 = 1$ and $f'(3/2) = 1$ so that $\theta_1 = 3/2 - Y_1$. Since

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

$\ell(T_1) \geq 1/3$. On the other hand, for $Y_1 \in T_1$, $\theta_1 \geq 3/2 + 1 \geq 2$.

Let us show the induction. Set $k \geq 2$ and suppose the result is true for $l \leq k-1$. Then $\theta_l \in \mathbb{Z} + [1/3,2/3]$ for all $l \leq k$, which implies that $A_{k-1} = k-1$. Hence,

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

By induction, $\theta_{k-1} \geq k$, and since $f'(\theta) \geq \theta/2$ for $\theta \geq 0$,

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

with $a = 1/(A_{k-1}\gamma_k f'(\theta_{k-1})) \leq \frac{2}{k^{2-\gamma}} \leq 1$ and $b = -\theta_{k-1}/a$. We deduce by pigeonhole principle that $\ell(T_k) \geq 1/3 - \frac{1}{2} \cdot \frac{a}{3} \geq 1/6$. Finally, for $Y_k \in A_k$ we have $Y_k \leq -1$ so that

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

□

By the previous result,

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

Moreover, from what we showed previously, on this event we have for $1 \leq k \leq n$

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

We deduce that $\theta_n \geq \theta_{k-1}(k-1)^{1-\gamma}/3 \geq (n!)^{2-\gamma}/2^n$. In particular,

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

Remark that the latter result can be easily adapted to get a counter-example with $g$ as smooth as desired.

## References


Bach, F. (2014). Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *The Journal of Machine Learning Research*, 15(1):595–627.

Bach, F. and Moulines, E. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). In *Advances in neural information processing systems*, pages 773–781.

Bercu, B., Bigot, J., Gadat, S., and Siviero, E. (2021). A stochastic gauss-newton algorithm for regularized semi-discrete optimal transport. *arXiv preprint arXiv:2107.05291*.

Bercu, B., Godichon, A., and Portier, B. (2020). An efficient stochastic newton algorithm for parameter estimation in logistic regressions. *SIAM Journal on Control and Optimization*, 58(1):348–367.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Boyer, C. and Godichon-Baggioni, A. (2020). On the asymptotic rate of convergence

of stochastic newton algorithms and their weighted averaged versions. *arXiv preprint arXiv:2011.09706*.

Cardot, H., Cénac, P., and Godichon-Baggioni, A. (2015). Online estimation of the geometric median in Hilbert spaces: non asymptotic confidence balls. Technical report, arXiv:1501.06930.

Cardot, H., Cénac, P., and Zitt, P.-A. (2013). Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19(1):18–43.

Cénac, P., Godichon-Baggioni, A., and Portier, B. (2020). An efficient averaged stochastic Gauss-Newton algorithm for estimating parameters of non linear regressions models. *arXiv preprint arXiv:2006.12920*.

De Vilmarest, J. and Wintenberger, O. (2021). Stochastic online optimization using kalman recursion. *J. Mach. Learn. Res.*, 22:223–1.

Défossez, A., Bottou, L., Bach, F., and Usunier, N. (2020). A simple convergence proof of adam and adagrad. *arXiv preprint arXiv:2003.02395*.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

Gadat, S. and Gavra, I. (2020). Asymptotic study of stochastic adaptive algorithm in non-convex landscape. *arXiv preprint arXiv:2012.05640*.

Gadat, S. and Panloup, F. (2017). Optimal non-asymptotic bound of the ruppert-polyak averaging without strong convexity. *arXiv preprint arXiv:1709.03342*.

Godichon-Baggioni, A. (2016). Estimating the geometric median in hilbert spaces with stochastic gradient algorithms: Lp and almost sure rates of convergence. *Journal of Multivariate Analysis*, 146:209–222.

Godichon-Baggioni, A. (2021). Convergence in quadratic mean of averaged stochastic gradient algorithms without strong convexity nor bounded gradient. *arXiv preprint arXiv:2107.12058*.

Godichon-Baggioni, A., Portier, B., and Lu, W. (2022). Recursive ridge regression using second-order stochastic algorithms.

Godichon-Baggioni, A., Werge, N., and Wintenberger, O. (2021). Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Streaming Data. working paper or preprint.

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. (2019). Sgd: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209. PMLR.

Hall, P. and Heyde, C. C. (2014). *Martingale limit theory and its application*. Academic press.

Koltchinskii, V. and Mendelson, S. (2015). Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, 2015(23):12991–13008.

Konečnỳ, J., Liu, J., Richtárik, P., and Takáč, M. (2015). Mini-batch semi-stochastic gradient descent in the proximal setting. *IEEE Journal of Selected Topics in Signal Processing*, 10(2):242–255.

Leluc, R. and Portier, F. (2020). Asymptotic optimality of conditioned stochastic gradient descent. *arXiv preprint arXiv:2006.02745*.

Pelletier, M. (1998). On the almost sure asymptotic behaviour of stochastic algorithms. *Stochastic processes and their applications*, 78(2):217–244.

Pelletier, M. (2000). Asymptotic almost sure efficiency of averaged stochastic algorithms. *SIAM J. Control Optim.*, 39(1):49–72.

Pinelis, I. (1994). Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, 22:1679–1706.

Polyak, B. and Juditsky, A. (1992). Acceleration of stochastic approximation. *SIAM J. Control and Optimization*, 30:838–855.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Ruppert, D. (1988). Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering.