Technical Report 160

Project Title:

Large-scale Optimization with Small-scale Data

Research Supervisor: Constantine Caramanis
Wireless Networking and Communications Group

August 2020

Data-Supported Transportation Operations & Planning Center (D-STOP)

A Tier 1 USDOT University Transportation Center at The University of Texas at Austin

D-STOP is a collaborative initiative by researchers at the Center for Transportation Research and the Wireless Networking and Communications Group at The University of Texas at Austin.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
D-STOP/2020/160		

4. Title and Subtitle	5. Report Date
Statistical Inference without Excess Data Using Only Stochastic Gradients: Volume 2	August 2020
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Tianyang Li, Liu Liu, Anastasios Kyrillidis, and Constantine Caramanis.	Report 160

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Data-Supported Transportation Operations & Planning Center (D-STOP) The University of Texas at Austin 3925 W. Braker Lane, 4th Floor Austin, TX 78759	11. Contract or Grant No. DTRT13-G-UTC58

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
United States Department of Transportation University Transportation Centers 1200 New Jersey Avenue, SE Washington, DC 20590	14. Sponsoring Agency Code

15. Supplementary Notes

Supported by a grant from the U.S. Department of Transportation, University Transportation Centers Program.

This is Volume 2; see Volume 1.

16. Abstract

We present a novel statistical inference framework for convex empirical risk minimization, using approximate stochastic Newton steps. The proposed algorithm is based on the notion of finite differences and allows the approximation of a Hessian-vector product from first-order information. In theory, our method efficiently computes the statistical error covariance in M-estimation, both for unregularized convex learning problems and high-dimensional LASSO regression, without using exact second order information, or resampling the entire data set. We also present a stochastic gradient sampling scheme for statistical inference in non-i.i.d. time series analysis, where we sample contiguous blocks of indices. In practice, we demonstrate the effectiveness of our framework on large-scale machine learning problems, that go even beyond convexity: as a highlight, our work can be used to detect certain adversarial attacks on neural networks.

17. Key Words	18. Distribution Statement
Statistical Inference; Frequentist Inference; M-estimation; High Dimensional Statistics; Time Series; Convex Optimization	No restrictions. This document is available to the public through NTIS (http://www.ntis.gov): National Technical Information Service 5285 Port Royal Road Springfield, Virginia 22161

19. Security Classif.(of this report)	20. Security Classif.(of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified		

Form DOT F 1700.7 (8-72) **Reproduction of completed page authorized**

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Acknowledgements

The authors recognize that support for this research was provided by a grant from the U.S. Department of Transportation, University Transportation Centers.

Appendix A. High dimensional linear regression statistical inference using stochastic gradients (Section 3)

A.1. Statistical inference using approximate proximal Newton steps with stochastic gradients

Here, we present a statistical inference procedure for high dimensional linear regression via approximate proximal Newton steps using stochastic gradients. It uses the plug-in estimator:

$$\widehat{S}^{-1} \quad \frac{1}{n}\sum_{i=1}^{n}\left((x_i^\top\widehat{\theta}-y_i)^2 x_i x_i^\top\right)\left(\widehat{S}^{-1},\right.$$

which is related to the empirical sandwich estimator (Huber, 1967; White, 1980). Lemma A.1 shows this is a good estimate of the covariance when $n \gg \frac{1}{D_\Sigma{}^4}\max\{1,\sigma^2\}s^2(\sigma+\|\theta^\star\|_1)^2$.

Algorithm 2 performs statistical inference in high dimensional linear regression (13), by computing the statistical error covariance in Theorem 3.2, based on the plug-in estimate in Lemma A.1. We denote the soft thresholding of A by ω as an element-wise procedure $(\mathbf{S}_\omega(A))_e = \mathrm{sign}(A_e)(|A_e|-\omega)_+$. For a vector v, we write v's i^{th} coordinate as $v(i)$. The optimization objective (13) is denoted as:

$$\tfrac{1}{2}\theta^\top\left(\widehat{S}-\tfrac{1}{n}\textstyle\sum_{i=1}^{n}x_i x_i^\top\right)\theta+\tfrac{1}{n}\textstyle\sum_{i=1}^{n}f_i,$$

where $f_i = \frac{1}{2}\left(x_i^\top - y_i\right)^2$. Further,

$$\boldsymbol{g}_{\widehat{S}}(v) = \nabla_v\left[\tfrac{1}{2}v^\top\widehat{S}v\right] = \widehat{S}v = \sum_{j=1}^{p}v(j)\cdot\mathbf{S}_\omega \quad \frac{1}{n}\sum_{i=1}^{n}\left[\nabla f_i(\theta+\boldsymbol{e}_j)-\nabla f_i(\theta)\right]$$

where $\boldsymbol{e}_i \in \mathbb{R}^p$ is the basis vector where the i^{th} coordinate is 1 and others are 0, and $\widehat{S}v$ is computed in a column-wise manner.

For point estimate optimization, the proximal Newton step (Lee et al., 2014) at θ solves the optimization problem

$$\min_{\Delta}\tfrac{1}{2\rho}\Delta^\top\widehat{S}\Delta+\left\langle(\widehat{S}-\tfrac{1}{n}\textstyle\sum_{i=1}^{n}x_i x_i^\top)\theta+\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\theta),\Delta\right\rangle+\lambda\|\theta+\Delta\|_1,$$

to determine a descent direction. For statistical inference, we solve a Newton step:

$$\min_{\Delta}\tfrac{1}{2\rho}\Delta^\top\widehat{S}\Delta+\left\langle\frac{1}{S_o}\sum_{k\in I_o}\nabla f_k(\theta_t),\Delta\right\rangle$$

to compute $-\widehat{S}^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\theta)$, whose covariance is the statistical error covariance.

To control variance, we solve Newton steps using SVRG and proximal SVRG (Xiao and Zhang, 2014), because in the high dimensional setting, the variance using SGD (Moulines and Bach, 2011) and proximal SGD (Atchadé et al., 2017) for solving Newton steps is too large. However because $p \gg n$, instead of sampling *by sample*, we sample *by feature*. We start from θ_0 sufficiently close to $\widehat{\theta}$ (see Theorem A.1 for details), which can be effectively achieved using proximal SVRG (Appendix A.3). Line 7 corresponds to SVRG's outer loop part that computes the full gradient, and line 12 corresponds to SVRG's inner loop

update. Line 8 corresponds to proximal SVRG's outer loop part that computes the full gradient, and line 13 corresponds to proximal SVRG's inner loop update.

The covariance estimate bound, asymptotic normality result, and choice of hyper-parameters are described in Appendix A.4. When $L_o^t = \Theta(\log(p) \cdot \log(t))$, we can estimate the covariance with element-wise error less than $O\left(\frac{\max\{1,\sigma\}\text{polylog}(n,p)}{\sqrt{T}}\right)$ with high probability, using $O\left(T \cdot n \cdot p^2 \cdot \log(p) \cdot \log(T)\right)$ numerical operations. Calculation of the de-biased estimator $\widehat{\theta}^{\text{d}}$ (16) via SVRG is described in Appendix A.2.

Algorithm 2 High dimensional linear regression statistical inference

1: **Parameters:** $S_o, S_i \in \mathbb{Z}_+$; $\eta, \tau \in \mathbb{R}_+$; **Initial state:** $\theta_0 \in \mathbb{R}^p$
2: **for** $t = 0$ **to** $T-1$ **do**
3: $I_o \leftarrow$ uniformly sample S_o indices with replacement from $[n]$
4: $g_t^0 \leftarrow -\frac{1}{S_o}\sum_{k\in I_o} \nabla f_k(\theta_t)$
5: $d_t^0 \leftarrow -\left(\boldsymbol{g}_{\widehat{S}}(\theta_t) - \frac{1}{n}\sum_{i=1}^n [\nabla f_i(\theta_t + \theta_t) - \nabla f_i(\theta_t)] + \frac{1}{n}\sum_{i=1}^n \nabla f_i(\theta_t)\right)$
6: **for** $j = 1$ **to** L_o^t **do** // solving Newton steps using SVRG
7: $u_t^j \leftarrow \boldsymbol{g}_{\widehat{S}}(g_t^{j-1}) - g_t^0$
8: $v_t^j \leftarrow \boldsymbol{g}_{\widehat{S}}(d_t^{j-1}) - d_t^0$
9: $g_t^j \leftarrow g_t^{j-1}$, $d_t^j \leftarrow d_t^{j-1}$
10: **for** $l = 1$ **to** L_i **do**
11: $I_i \leftarrow$ uniformly sample S_i indices without replacement from $[p]$
12: $g_t^j \leftarrow g_t^j - \tau\left[u_t^j + \frac{p}{S_i}\sum_{k\in S_i}\left[g_t^j(k) - g_t^{j-1}(k)\right]\left(\mathbf{S}_\omega\left(\nabla f_k(\theta_t + \boldsymbol{e}_k) - \nabla f_k(\theta_t)\right)\right]\right.$
13: $d_t^j \leftarrow \mathbf{S}_{\eta\lambda}\left(d_t^j - \eta\left[v_t^j + \frac{p}{S_i}\sum_{k\in S_i}\left[d_t^j(k) - d_t^{j-1}(k)\right]\left(\mathbf{S}_\omega\left(\nabla f_k(\theta_t + \boldsymbol{e}_k) - \nabla f_k(\theta_t)\right)\right]\right)\right.$
14: **end for**
15: **end for**
16: Use $\sqrt{S_o}\cdot\frac{\bar{g}_t}{\rho_t}$ for statistical inference, where $\bar{g}_t = \frac{1}{L_o^t+1}\sum_{j=0}^{L_o^t} g_t^j$
17: $\theta_{t+1} = \theta_t + \bar{d}_t$, where $\bar{d}_t = \frac{1}{L_o+1}\sum_{j=0}^{L_o^t} d_t^j$ // point estimation (optimization)
18: **end for**

A.2. Computing the de-biased estimator (16) via SVRG

To control variance, we solve each proximal Newton step using SVRG, in stead of SGD as in Algorithm 1. Because However because the number of features is much larger than the number of samples, instead of sampling *by sample*, we sample *by feature*.

The de-biased estimator is

$$\begin{aligned}\widehat{\theta}^{\text{d}} =& \widehat{\theta} + \widehat{S}^{-1}\left[\frac{1}{n}\sum_{i=1}^n y_i x_i - \frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)\widehat{\theta}\right]\\ =& \widehat{\theta} + \widehat{S}^{-1}\left(-\frac{1}{n}\sum_{i=1}^n \nabla f_i(\widehat{\theta})\right)\end{aligned}$$

And we compute $\widehat{S}^{-1}\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\widehat{\theta})$ using SVRG (Johnson and Zhang, 2013) by solving the following optimization problem using SVRG and sampling by feature

$$\min_u \frac{1}{2}u^\top \widehat{S}u + \left\langle \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\widehat{\theta}), u \right\rangle$$

Algorithm 3 Computing the de-biased estimator (16) via SVRG

for $i = 0$ **to** $L_o - 1$ **do**
  $d_i^0 \leftarrow -\eta[\boldsymbol{g}_{\widehat{S}}(u_i) + \frac{1}{n}\sum_{k=1}^{n}\nabla f_k(\widehat{\theta})]$
  for $j = 0$ **to** $L_i - 1$ **do**
    $I \leftarrow$ sample S indices uniformly from $[p]$ without replacement
    $d_i^{j+1} \leftarrow d_i^j + d_t^0 - \eta\left(\frac{1}{S}\sum_{k\in I} d_i^j(k)\cdot \mathbf{S}_\omega(\nabla f_k(\widehat{\theta}+\boldsymbol{e}_k) - f_k(\widehat{\theta}))\right)$
  end for
  $u_{i+1} \leftarrow u_i + \bar{d}_i$, where $\bar{d}_i = \frac{1}{L_i+1}\sum_{j=0}^{L_i} d_i^j$
end for

Similar to Algorithm 2, we choose $\eta = \Theta\left(\frac{1}{p}\right)$ and $L_i = \Theta(p)$.

A.3. Solving the high dimensional linear regression optimization objective (13) using proximal SVRG

We solve our high dimensional linear regression optimization problem using proximal SVRG (Xiao and Zhang, 2014)

$$\widehat{\theta} = \arg\min_\theta \frac{1}{2}\theta^\top \left(\widehat{S} - \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)\theta + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left(x_i^\top\theta - y_i\right)^2 + \lambda\|\theta\|_1. \tag{20}$$

Algorithm 4 Solving the high dimensional linear regression optimization objective (13) using proximal SVRG

for $i = 0$ **to** $L_o - 1$ **do**
  $u_i^0 \leftarrow \theta_i$
  $d_t \leftarrow \boldsymbol{g}_{\widehat{S}}(\theta_i) - \frac{1}{n}\sum_{k=1}^{n}[\nabla f_k(\theta_i+\theta_i) - \nabla f_k(\theta_i)] + \frac{1}{n}\sum_{k=1}^{n}\nabla f_k(\theta_i)$
  for $j = 0$ **to** $L_i - 1$ **do**
    $u_i^{j+1} \leftarrow \mathbf{S}_{\eta\lambda}(u_i^j - \eta[d_t + \frac{1}{S}\sum_{k\in I}\left(u_i^j(k) - \theta_i(k)\right)\left(\mathbf{S}_\omega(\nabla f_k(\theta_t+\boldsymbol{e}_k) - \nabla f_k(\theta_t))\right)])$
  end for
  $\theta_{t+1} \leftarrow \frac{1}{L_i+1}\sum_{j=0}^{L_i} u_i^j$
end for

Similar to Algorithm 2, we choose $\eta = \Theta\left(\frac{1}{p}\right)$ and $L_i = \Theta(p)$.

A.4. Non-asymptotic covariance estimate bound and asymptotic normality in Algorithm 2

We have a non-asymptotic covariance estimate bound and an asymptotic normality result.

Theorem A.1 *Under our assumptions, when* $n \gg \max\{b^2, \frac{1}{{D_\Sigma}^2}\}\log p$, $S_o = O(1)$, $S_i = O(1)$, *and conditioned on* $\{x_i\}_{i=1}^n$ *and following events which simultaneously with probability at least* $1 - p^{-\Theta(1)} - n^{-\Theta(1)}$

- ❑ *[A]:* $\max_{1\le i\le n} |\epsilon_i| \lesssim \sigma\sqrt{\log n}$,
- ❑ *[B]:* $\max_{1\le i\le n} \|x_i\|_\infty \lesssim \sqrt{\log p + \log n}$,
- ❑ *[C]:* $\|\widehat{S}^{-1}\|_\infty \lesssim \frac{1}{D_\Sigma}$,

we choose $L_i = \Theta(p)$, $\tau = \Theta(\frac{1}{p})$, $\eta = \Theta(\frac{1}{p})$ *in Algorithm 2.*

Here, we denote the objective function as

$$P(\theta) = \frac{1}{2}\theta^\top \quad \widehat{S} - \frac{1}{n}\sum_{i=1}^n x_i x_i^\top \Big)\theta + \frac{1}{n}\sum_{i=1}^n \frac{1}{2}\left(x_i^\top\theta - y_i\right)^2 + \lambda\|\theta\|_1.$$

Then, we have a non-asymptotic covariance estimate bound

$$\left\|\tfrac{S_o}{T}\textstyle\sum_{t=1}^T \bar{g}_t\bar{g}_t^\top - \widehat{S}^{-1}\left(\frac{1}{n}\sum_{i=1}^n (x_i^\top\widehat{\theta} - y_i)^2 x_i x_i^\top\right)\widehat{S}^{-1}\right\|_{\max}$$
$$\lesssim \sqrt{\left((\log p + \log n)\|\widehat{\theta} - \theta^\star\|_1 + \sigma\sqrt{(\log p + \log n)\log n}\right)\frac{\log p}{T}}$$
$$+ \frac{1}{u}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^T 0.95^{L_o^t}(1 + \sqrt{P(\theta_0) - P(\widehat{\theta})}\,0.95^{\sum_{i=0}^{t-1} L_o^t}) + \sqrt{p}(\log p + \log n)\sqrt{P(\theta_0) - P(\widehat{\theta})}\,0.95^{\sum_{i=0}^{t-1} L_o^t}\right]$$

where $\|A\|_{\max} = \max\{1 \le j,k \le p\}|A_{jk}|$ *is the matrix max norm, with probability at least* $1 - p^{-\Theta(-1)} - u$.

And we have asymptotic normality

$$\frac{1}{\sqrt{t}}\left(\sum_{t=1}^T \sqrt{S_o}\bar{g}_t + \frac{1}{n}\sum_{i=1}^n x_i(x_i^\top\widehat{\theta} - y_i)\right) = W + R,$$

where W *weakly converges to* $\mathcal{N}\Big(0, \widehat{S}^{-1}\Big[\frac{1}{n}\sum_{i=1}^n (x_i^\top\widehat{\theta} - y_i)^2 x_i x_i^\top - \big(\frac{1}{n}\sum_{i=1}^n x_i(x_i^\top\widehat{\theta} - y_i)\big)\big(\frac{1}{n}\sum_{i=1}^n x_i(x_i^\top\widehat{\theta} - y_i)\big)^\top\Big]\widehat{S}^{-1}\Big)$, *and* $\mathbb{E}[\|R\|_\infty \mid \{x_i\}_{i=1}^n, [A], [B], [C]] \lesssim \frac{1}{\sqrt{T}}\sum_{t=1}^T 0.95^{L_o^t}(1 + \sqrt{P(\theta_0) - P(\widehat{\theta})}\,0.95^{\sum_{i=0}^{t-1} L_o^t}) + \sqrt{p}(\log p + \log n)\sqrt{P(\theta_0) - P(\widehat{\theta})}\,0.95^{\sum_{i=0}^{t-1} L_o^t}$.

Note that when we choose $L_o^t = \Theta(\log(p)\cdot\log(t))$, and start from θ_0 satisfying $P(\theta_0) - P(\widehat{\theta}) \lesssim \frac{1}{p(\log p + \log n)^2}$ which can be effectively achieved using proximal SVRG (Appendix A.3), we can estimate the statistical error covariance with element-wise error less than $O\left(\frac{\max\{1,\sigma\}\mathrm{polylog}(n,p)}{\sqrt{T}}\right)$ with high probability, using $O\left(T\cdot n\cdot p^2\cdot\log(p)\cdot\log(T)\right)$ numerical operations.

A.5. Plug-in statistical error covariance estimate

Algorithm 2 is similar to using plug-in estimator $\frac{1}{n}\sum_{i=1}^n (x_i^\top\widehat{\theta} - y_i)^2 x_i x_i^\top$ for $\sigma^2\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)$ in Theorem 3.2, similar to the sandwich estimator (Huber, 1967; White, 1980). Lemma A.1 gives a bound on using this plug-in estimator in the statistical error covariance (Theorem 3.2) for coordinate-wise confidence intervals.

Lemma A.1 *Under our assumptions, when* $n \gg \max\{b^2, \frac{1}{D_\Sigma{}^2}\} \log p$*, we have*

$$\left\| \widehat{S}^{-1} \left(\frac{1}{n} \sum_{i=1}^n (x_i^\top \widehat{\theta} - y_i)^2 x_i x_i^\top \right) \widehat{S}^{-1} - \sigma^2 \widehat{S}^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i x_i^\top \right) \widehat{S}^{-1} \right\|_{\max}$$
$$\lesssim \frac{1}{D_\Sigma{}^2} \left(\sigma \sqrt{\log n} + s \left(\sigma + \|\theta^\star\|_1 \right) \sqrt{\log p + \log n} \sqrt{\frac{\log p}{n}} \right) s \left(\sigma + \|\theta^\star\|_1 \right) (\log p + \log n)^{\frac{3}{2}} \sqrt{\frac{\log p}{n}},$$

where $\|A\|_{\max} = \max_{1 \le j,k \le p} |A_{jk}|$ *is the matrix max norm, with probability at least* $1 - p^{-\Theta(1)} - n^{-\Theta(1)}$.

Appendix B. Time series statistical inference with approximate Newton steps using only stochastic gradients (Section 4)

Here, we give the complete approximate Newton-based time series statistical inference algorithm using only stochastic gradients.

Algorithm 5 Unregularized M-estimation statistical inference

Parameters: $\mathbf{l}, S_i \in \mathbb{Z}_+$; $\rho_0, \tau_0 \in \mathbb{R}_+$; $d_o, d_i \in \left(\frac{1}{2}, 1\right)$ **Initial state:** $\theta_0 \in \mathbb{R}^p$

for $t = 0$ to $T - 1$ **do** // approximate stochastic Newton descent

  $\rho_t \leftarrow \rho_0 (t+1)^{-d_o}$

  Uniformly select some $i_o \in [n]$, then set I_o to the random contiguous block $\{i_o, i_o+1, \ldots, i_o + \mathbf{l} - 1\} \mod n$, which circularly wraps around

  $g_t^0 \leftarrow -\rho_t \left(\frac{1}{\mathbf{l}} \sum_{i \in I_o} \nabla f_i(\theta_t) \right)$

  for $j = 0$ to $L - 1$ **do** // solving (1) approximately using SGD

    $\tau_j \leftarrow \tau_0 (j+1)^{-d_i}$ and $\delta_t^j \leftarrow O(\rho_t^4 \tau_j^4)$

    $I_i \leftarrow$ uniformly sample S_i indices without replacement from $[n]$

    $g_t^{j+1} \leftarrow g_t^j - \tau_j \left(\frac{1}{S_i} \sum_{k \in I_i} \frac{\nabla f_k(\theta_t + \delta_t^j g_t^j) - \nabla f_k(\theta_t)}{\delta_t^j} \right) + \tau_j g_t^0$

  end for

  Use $\sqrt{\mathbf{l}} \cdot \frac{\bar{g}_t}{\rho_t}$ for statistical inference, where $\bar{g}_t = \frac{1}{L+1} \sum_{j=0}^L g_t^j$

  $\theta_{t+1} \leftarrow \theta_t + g_t^L$

end for

Corollary B.1 gives guarantees for Algorithm 5, and is similar to the i.i.d. case (Theorem 2.1).

Corollary B.1

Under the same assumptions as Theorem 2.1, in Algorithm 5, for the outer iterate θ_t *we have*

$$\mathbb{E}\left[\|\theta_t - \widehat{\theta}\|_2^2 \right] \lesssim t^{-d_o}, \tag{21}$$

$$\mathbb{E}\left[\|\theta_t - \widehat{\theta}\|_2^4 \right] \lesssim t^{-2d_o}. \tag{22}$$

In each outer loop, after L *steps of the inner loop, we have:*

$$\mathbb{E}\left[\left\| \frac{\bar{g}_t}{\rho_t} - [\nabla^2 f(\theta_t)]^{-1} g_t^0 \right\|_2^2 \middle| \theta_t \right] \lesssim \frac{1}{L} \|g_t^0\|_2^2, \tag{23}$$

and at each step of the inner loop, we have:

$$\mathbb{E}\left[\left\|g_t^{j+1} - [\nabla^2 f(\theta_t)]^{-1} g_t^0\right\|_2^4 \mid \theta_t\right] \lesssim (j+1)^{-2d_i} \left\|g_t^0\right\|_2^4. \tag{24}$$

After T steps of the outer loop, we have a non-asymptotic bound on the "covariance":

$$\mathbb{E}\left[\left\|H^{-1}GH^{-1} - \frac{S_o}{T}\sum_{t=1}^{T}\frac{\bar{g}_t\bar{g}_t^\top}{\rho_t^2}\right\|_2\right] \lesssim T^{-\frac{d_o}{2}} + L^{-\frac{1}{2}}, \tag{25}$$

where $H = \nabla^2 f(\widehat{\theta})$, and

$$G = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\widehat{\theta})\, f_i(\widehat{\theta})^\top + \sum_{j=1}^{\mathbf{l}} w(j,\mathbf{l}) \sum_{i=j+1}^{n}\left(\nabla f_i(\widehat{\theta})\,\nabla f_{i-j}(\widehat{\theta})^\top + \nabla f_{i-j}(\widehat{\theta})\,\nabla f_i(\widehat{\theta})^\top\right) \tag{26}$$

with $w(j,\mathbf{l}) = 1 - \frac{j}{\mathbf{l}}$.

Also, in Algorithm 5's outer loop, the average of consecutive iterates satisfies

$$\mathbb{E}\left[\left\|\frac{\sum_{t=1}^{T}\theta_t}{T} - \widehat{\theta}\right\|_2^2\right] \lesssim \frac{1}{T}, \tag{27}$$

$$\frac{1}{\sqrt{T}}\left(\frac{\sum_{t=1}^{T}\theta_t}{T} - \widehat{\theta}\right) \Rightarrow W + \Delta, \tag{28}$$

where W weakly converges to $\mathcal{N}(0, \frac{1}{S_o}H^{-1}GH^{-1})$, and $\Delta = o_P(1)$ when $T \to \infty$ and $L \to \infty$ ($\mathbb{E}[\|\Delta\|_2^2] \lesssim T^{1-2d_o} + T^{d_o-1} + \frac{1}{L}$).

Our approximate Newton time series statistical inference procedure estimates $H^{-1}GH^{-1}$, where G is the Newey-West covariance estimator (19) with weight

$$w(j,\mathbf{l}) = 1 - \frac{j}{\mathbf{l}}, \tag{29}$$

which is because when we estimate the variance in Algorithm 5, for $j > 0$, terms $\nabla f_i\,\nabla f_{i+j}^\top$ and $\nabla f_{i+j}\,\nabla f_i^\top$ appear $\mathbf{l} - j$ times, and the term $\nabla f_i\,\nabla f_i^\top$ appears $\mathbf{l}$ times. Note that the connection between sampling scheme and Newey-West estimator was also observed in (Kunsch, 1989). Thus, our stochastic approximate Newton statistical inference procedure for time series analysis has similar statistical properties compared circular bootstrap (Politis and Romano, 1992, 1994).

Because expectation of the stochastic gradient in line 5 of Algorithm 5 is the full gradient $\frac{1}{n}\sum_{i=1}^{n} f_i(\widehat{\theta})$, we have the same optimization guarantees as the i.i.d. case (Corollary 2.1).

Appendix C. Statistical inference via approximate stochastic Newton steps using first order information with increasing inner loop counts

Here, we present corollaries when the number of inner loops increases in the outer loops (i.e., $(L)_t$ is an increasing series). This guarantees convergence of the covariance estimate to $H^{-1}GH^{-1}$, although it is less efficient than using a constant number of inner loops.

C.1. Unregularized M-estimation

Similar to Theorem 2.1's proof, we have the following result when the number of inner loop increases in the outer loops.

Corollary C.1

In Algorithm 1, if the number of inner loop in each outer loop $(L)_t$ increases in the outer loops, then we have

$$\mathbb{E}\left[\left\|H^{-1}GH^{-1}-\frac{S_o}{T}\sum_{t=1}^{T}\frac{\bar{g}_t\bar{g}_t^\top}{\rho_t^2}\right\|_2\right]\lesssim T^{-\frac{d_o}{2}}+\sqrt{\frac{1}{T}\sum_{i=1}^{T}\frac{1}{(L)_t}}.$$

For example, when we choose choose $(L)_t=L(t+1)^{d_L}$ for some $d_L>0$, then $\sqrt{\frac{1}{T}\sum_{i=1}^{T}\frac{1}{(L)_t}}=O(\frac{1}{\sqrt{L}}T^{-\frac{d_L}{2}})$.

Appendix D. SVRG based statistical inference algorithm in unregularized M-estimation

Here we present a SVRG based statistical inference algorithm in unregularized M-estimation, which has asymptotic normality and improved bounds for the "covariance". Although Algorithm 6 has stronger guarantees than Algorithm 1, Algorithm 6 requires a full gradient evaluation in each outer loop.

Algorithm 6 SVRG based statistical inference algorithm in unregularized M-estimation

for $t\leftarrow 0; t<T; ++t$ **do**
  $d_t^0\leftarrow -\eta\nabla f(\theta_t)=-\eta\left(\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\theta)\right)$ // point estimation via SVRG
  $I_o\leftarrow$ uniformly sample S_o indices with replacement from $[n]$
  $g_t^0\leftarrow -\rho_t\left(\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\theta_t)\right)$ // statistical inference
  for $j\leftarrow 0; j<L; ++j$ **do** // solving (1) approximately using SGD
    $I_i\leftarrow$ uniformly sample S_i indices without replacement from $[n]$
    $d_t^{j+1}\leftarrow d_t^j-\eta\left(\frac{1}{S_i}\sum_{k\in I_i}(\nabla f_k(\theta_t+d_t^j)-\nabla f_k(\theta_t)\right)+d_t^0$ // point estimation via SVRG
    $g_t^{j+1}\leftarrow g_t^j-\tau_j\left(\frac{1}{S_i}\sum_{k\in I_i}\frac{1}{\delta_t^j}[\nabla f_k(\theta_t+\delta_t^j g_t^j)-\nabla f_k(\theta_t)]\right)+\tau_j g_t^0$ // statistical inference
  end for
  Use $\sqrt{S_o}\cdot\frac{\bar{g}_t}{\rho_t}$ for statistical inference // $\bar{g}_t=\frac{1}{L+1}\sum_{j=0}^{L}g_t^j$
  $\theta_{t+1}\leftarrow\theta_t+\bar{d}_t$ // $\bar{d}_t=\frac{1}{L+1}\sum_{j=0}^{L}d_t^j$
end for

Corollary D.1

In Algorithm 6, when $L \geq 20\frac{\max_{1\leq i\leq n}\beta_i}{\alpha}$ *and* $\eta = \frac{1}{10\max_{1\leq i\leq n}\beta_i}$*, after* T *steps of the outer loop, we have a non-asymptotic bound on the "covariance"*

$$\mathbb{E}\left[\left\|H^{-1}GH^{-1} - \frac{S_o}{T}\sum_{t=1}^{T}\frac{\bar{g}_t\bar{g}_t^\top}{\rho_t^2}\right\|_2\right] \lesssim L^{-\frac{1}{2}}, \tag{30}$$

and asymptotic normality

$$\frac{1}{\sqrt{T}}\left(\sum_{t=1}^{T}\frac{\bar{g}_t}{\rho_t}\right) = W + \Delta,$$

where W *weakly converges to* $\mathcal{N}(0, \frac{1}{S_o}H^{-1}GH^{-1})$ *and* $\Delta = o_P(1)$ *when* $T \to \infty$ *and* $L \to \infty$ *(*$\mathbb{E}[\|\Delta\|_2] \lesssim \frac{1}{\sqrt{T}} + \frac{1}{L}$*).*

When the number of inner loops increases in the outer loops (i.e., $(L)_t$ is an increasing series), we have a result similar to Corollary C.1.

A better understanding of concentration, and Edgeworth expansion of the average consecutive iterates averaged (beyond (Dippon, 2008a,b)) in stochastic approximation, would give stronger guarantees for our algorithms, and better compare and understand different algorithms.

D.1. Lack of asymptotic normality in Algorithm 1 for mean estimation

In mean estimation, we solve the following optimization problem

$$\widehat{\theta} = \arg\min_{\theta}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\|\theta - X^{(i)}\|_2^2,$$

where we assume that $\{X^{(i)}\}_{i=1}^n$ are constants.

For ease of explanation we use $S_o = 1$, $\rho_t = \rho$, and $\theta_0 = 0$,and we have

$$\frac{\bar{g}_t}{\rho_t} = -\theta_t + X_t,$$

where X_t is uniformly sampled from $\{X^{(i)}\}_{i=1}^n$.

And for $t \geq 1$ we have

$$\theta_t = \sum_{i=0}^{t-1}\rho(1-\rho)^{t-1-i}X_i.$$

Then, we have

$$\frac{1}{\sqrt{T}}\left(\sum_{i=1}^{T}\frac{\bar{g}_t}{\rho_t}\right)$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{T}}\Big(\sum_{t=1}^{T} X_t - \sum_{t=1}^{T}\sum_{i=0}^{t-1} \rho(1-\rho)^{t-1-i} X_i\Big) \\
&= \frac{1}{\sqrt{T}}\Big(\sum_{t=1}^{T} X_t - \sum_{i=0}^{T-1}\Big(\sum_{t=i+1}^{T} \rho(1-\rho)^{t-1-i}\Big) X_i\Big) \\
&= \frac{1}{\sqrt{T}}\Big(\sum_{t=1}^{T} X_t - \sum_{i=0}^{T-1} \big(1-(1-\rho)^{T-i}\big) X_i\Big) \\
&= \frac{1}{\sqrt{T}}\Big(X_T - X_0 + \sum_{i=1}^{T-1} (1-\rho)^{T-i} X_i\Big),
\end{aligned}
$$

whose ℓ_2 norm's expectation converges to 0 when $T \to \infty$, which implies that it converges to 0 with probability 1. Thus, in this setting $\frac{1}{\sqrt{T}}\left(\sum_{t=1}^{T} \frac{\bar{g}_t}{\rho_t}\right)$ does not weakly converge to $\mathcal{N}\left(0, \frac{1}{S_o} H^{-1} G H^{-1}\right)$.

Appendix E. An intuitive view of SVRG as approximate stochastic Newton descent

Here we present an intuitive view of SVRG as approximate stochastic Newton descent, which is the inspiration behind our work.

Gradient descent solves the optimization problem $\widehat{\theta} = \arg\min_\theta f(\theta)$, where the function is a sum of n functions $f(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta)$, using

$$\theta_{t+1} = \theta_t - \eta \nabla f(\theta_t),$$

and stochastic gradient descent uniformly samples a random index at each step

$$\theta_{t+1} = \theta_t - \eta_t \nabla f_i(\theta_t).$$

- ❏ **Outer loop:**
- ❏ $g \leftarrow \nabla f(\theta_t) = \sum_{i=1}^{n} \nabla f_i(\theta_t)$
- ❏ Let d be the descent direction
- ❏ – **Inner loop:**
  - Choose a random index k
  - $d \leftarrow d - \eta(\nabla f_k(\theta_t + d) - \nabla f_k(\theta_t) + g)$
- ❏ $\theta_{t+1} = \theta_t + d$

SVRG (Johnson and Zhang, 2013) improves gradient descent and SGD by having an outer loop and an inner loop. Here, we give an intuitive explanation of SVRG as stochastic proximal Newton descent, by arguing that

- ❏ each outer loop approximately computes the Newton direction $-(\nabla^2 f)^{-1}\nabla f$

❑ the inner loops can be viewed as SGD steps solving a proximal Newton step $\min_d \langle \nabla f, d\rangle + \frac{1}{2} d^\top (\nabla^2 f) d$

First, it is well known (Bubeck, 2015) that the Newton direction is exactly the solution of

$$\min_d \langle \nabla f(\theta), d\rangle + \frac{1}{2} d^\top [\nabla^2 f(\theta)] d. \tag{31}$$

Next, let's consider solving (31) using gradient descent on a function of d, and notice that its gradient with respect to d is

$$\nabla f(\theta) + [\nabla^2 f(\theta)] d,$$

which can be approximated through f's Taylor expansion ($[\nabla^2 f(\theta)] d \approx \nabla f(\theta + d) - \nabla f(\theta)$) as

$$\nabla f(\theta) + [\nabla f(\theta + d) - \nabla f(\theta)].$$

Thus, SVRG's inner loops can be viewed as using SGD to solve proximal Newton steps in outer loops. And it can be viewed as the power series identity for matrix inverse $H^{-1} = \sum_{i=0}^{\infty} (I - \eta H)$, which corresponds to unrolling the gradient descent recursion for the optimization problem $H^{-1} = \arg\min_\Omega \operatorname{Tr}\left(\frac{1}{2}\Omega^\top H \Omega - \Omega\right)$.

Appendix F. Proofs

F.1. Proof of Theorem 2.1

Given assumptions about strong convexity, Lipschitz gradient continuity and Hessian Lipschitz continuity in Theorem 2.1, we denote:

$$\bar{\beta} = \frac{\beta_i}{n}, \quad \bar{h} = \frac{h_i}{n}.$$

Then, $\forall \theta_1, \theta_2$ we have:

$$\|\nabla f(\theta_2) - \nabla f(\theta_1)\|_2 \le \bar{\beta} \|\theta_2 - \theta_1\|_2, \quad \text{and} \quad \|\nabla^2 f(\theta_2) - \nabla^2 f(\theta_1)\|_2 \le \bar{h} \|\theta_2 - \theta_1\|_2.$$

and $\forall \theta$:

$$\|\nabla^2 f(\theta)\|_2 \le \bar{\beta}.$$

In our proof, we also use the following:

$$\bar{h}_2 = \frac{1}{n} \sum_{i=1}^n h_i^2, \quad \bar{\beta}_2 = \frac{1}{n} \sum_{i=1}^n \beta_i^2, \quad \text{and} \quad \beta = \sup_\theta \|\nabla^2 f(\theta)\|_2.$$

Observe that:

$$\bar{h} \le \sqrt{\bar{h}_2}, \text{ and } \alpha \le \beta \le \bar{\beta} \le \sqrt{\bar{\beta}_2}.$$

F.1.1. PROOF OF (8)

We first prove (8); the proof is similar to standard SGD convergence proofs (e.g. (Li et al., 2018; Chen et al., 2016; Polyak and Juditsky, 1992)). For the rest of our discussion, we assume that

$$\delta_t^j \cdot \bar{h} \le \delta_t^j \cdot \sqrt{\bar{h}_2} \ll 1, \quad \forall t, j.$$

Using $\nabla f(\theta)$'s Taylor series expansion with a Lagrange remainder, we have the following lemma, which bounds the Hessian vector product approximation error.

Lemma F.1 *$\forall, \theta, g, \delta \in \mathbb{R}^p$, we have:*

$$\left\| \frac{\nabla f_i(\theta+\delta g) - \nabla f_i(\theta)}{\delta} - \nabla^2 f_i(\theta) g \right\|_2 \le h_i \cdot |\delta| \cdot \|g\|_2,$$
$$\left\| \frac{\nabla f(\theta+\delta g) - \nabla f(\theta)}{\delta} - \nabla^2 f(\theta) g \right\|_2 \le \bar{h} \cdot |\delta| \cdot \|g\|_2.$$

Denote $H_t = \nabla^2 f(\theta_t)$ and

$$e_t^j = \left(\frac{1}{S_i} \cdot \sum_{k \in I_i} \frac{\nabla f_k(\theta_t + \delta_t^j g_t^j) - \nabla f_k(\theta_t)}{\delta_t^j} \right) - \frac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j},$$

then we have

$$g_t^{j+1} - H_t^{-1} g_t^0 = g_t^j - H_t^{-1} g_t^0 - \tau_j \cdot \frac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} + \tau_j g_t^0 - \tau_j e_t^j. \tag{32}$$

Because $\mathbb{E}[e_j^t \mid g_j^t, \theta_t] = 0$, we have

$$\mathbb{E}\left[\left\| g_t^{j+1} - H_t^{-1} g_t^0 \right\|_2^2 \mid \theta_t \right] = \mathbb{E}\Bigg[\Bigg(\left\| g_t^j - H_t^{-1} g_t^0 \right\|_2^2 - \tau_j \underbrace{\left\langle g_t^j - H_t^{-1} g_t^0, \frac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} - g_t^0 \right\rangle}_{[1]}$$
$$+ \tau_j^2 \underbrace{\left\| \frac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} - g_t^0 \right\|_2^2}_{[2]} + \tau_j^2 \underbrace{\left\| e_t^j \right\|_2^2}_{[3]} \mid \theta_t \Bigg] \Bigg). \tag{33}$$

For term [1], we have

$$\left\langle g_t^j - H_t^{-1} g_t^0, \frac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} - g_t^0 \right\rangle$$

$$
\begin{aligned}
&= \left(g_t^j - H_t^{-1} g_t^0\right)^\top H_t \left(g_t^j - H_t^{-1} g_t^0\right) + \left\langle g_t^j - H_t^{-1} g_t^0, \tfrac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} - H_t \right\rangle \\
&\geq \left(g_t^j - H_t^{-1} g_t^0\right)^\top H_t \left(g_t^j - H_t^{-1} g_t^0\right) - \left| \langle g_t^j - H_t^{-1} g_t^0, \tfrac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} - H_t \rangle \right| \\
&\quad \text{by Hessian approximation} \\
&\geq \left(g_t^j - H_t^{-1} g_t^0\right)^\top H_t \left(g_t^j - H_t^{-1} g_t^0\right) - \delta_t^j \cdot \bar{h} \cdot \left\| g_t^j - H_t^{-1} g_t^0 \right\|_2 \cdot \left\| g_t^j \right\|_2 \\
&\quad \text{by AM-GM inequality} \\
&\geq \left(g_t^j - H_t^{-1} g_t^0\right)^\top H_t \left(g_t^j - H_t^{-1} g_t^0\right) - \tfrac{\delta_t^j \cdot \bar{h}}{2} \cdot \left\| g_t^j - H_t^{-1} g_t^0 \right\|_2^2 - \tfrac{\delta_t^j \cdot \bar{h}}{2} \cdot \left\| g_t^j \right\|_2^2 \\
&= \left(g_t^j - H_t^{-1} g_t^0\right)^\top H_t \left(g_t^j - H_t^{-1} g_t^0\right) - \tfrac{\delta_t^j \cdot \bar{h}}{2} \cdot \left\| g_t^j - H_t^{-1} g_t^0 \right\|_2^2 - \tfrac{\delta_t^j \cdot \bar{h}}{2} \cdot \left\| g_t^j - H_t^{-1} g_t^0 + H_t^{-1} g_t^0 \right\|_2^2 \\
&\quad \|x + a\|_2^2 \leq 2\|x\|_2^2 + 2\|y\|_2^2 \\
&\geq \left(g_t^j - H_t^{-1} g_t^0\right)^\top H_t \left(g_t^j - H_t^{-1} g_t^0\right) - \tfrac{3\delta_t^j \cdot \bar{h}}{2} \cdot \left\| g_t^j - H_t^{-1} g_t^0 \right\|_2^2 - \delta_t^j \bar{h} \cdot \left\| H_t^{-1} g_t^0 \right\|_2^2 \\
&\quad \text{by strong convexity} \\
&\geq \left(g_t^j - H_t^{-1} g_t^0\right)^\top H_t \left(g_t^j - H_t^{-1} g_t^0\right) - \tfrac{3\delta_t^j \cdot \bar{h}}{2} \cdot \left\| g_t^j - H_t^{-1} g_t^0 \right\|_2^2 - \tfrac{\delta_t^j \bar{h}}{\alpha^2} \cdot \left\| g_t^0 \right\|_2^2 . \qquad (34)
\end{aligned}
$$

For term [2], by repeatedly applying AM-GM inequality, using f's smoothness and strong convexity, and assuming $\delta_t^j \bar{h} \ll 1$, we have:

$$
\begin{aligned}
\left\| \tfrac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} - g_t^0 \right\|_2^2 &= \left\| \tfrac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} - H_t g_t^j + H_t g_t^j - g_t^0 \right\|_2^2 \\
&\leq \left\| \tfrac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} - H_t g_t^j \right\|_2^2 \\
&\quad + 2 \left\| \tfrac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} - H_t g_t^j \right\|_2 \cdot \left\| H_t g_t^j - g_t^0 \right\|_2 + \left\| H_t g_t^j - g_t^0 \right\|_2^2 \\
&\leq \left(\delta_t^j \bar{h}\right)^2 \|g_t^j\|_2^2 + 2\delta_t^j \bar{h} \left\| g_t^j \right\|_2 \cdot \left\| H_t g_t^j - g_t^0 \right\|_2 + \left\| H_t g_t^j - g_t^0 \right\|_2^2 \\
&\leq \left(\delta_t^j \bar{h} + \left(\delta_t^j \bar{h}\right)^2\right) \left\| g_t^j \right\|_2^2 + \left(1 + \delta_t^j \bar{h}\right) \|H_t g_t^j - g_t^0\|_2^2 \\
&\leq 2\left(\delta_t^j \bar{h} + \left(\delta_t^j \bar{h}\right)^2\right) \left(\left\| g_k^j - H_t^{-1} g_t^0 \right\|_2^2 + \|H_t^{-1} g_t^0\|_2^2 \right) + \left(1 + \delta_t^j \bar{h}\right) \left\| H_t g_t^j - g_t^0 \right\|_2^2 \\
&\leq \tfrac{2\left(\delta_t^j \bar{h} + \left(\delta_t^j \bar{h}\right)^2\right)}{\alpha^2} \cdot \|g_t^0\|_2^2 + \left(1 + 3\delta_t^j \bar{h} + 2\left(\delta_t^j \bar{h}\right)^2\right) \left\| H_t g_t^j - g_t^0 \right\|_2^2 \\
&\leq \tfrac{4\delta_t^j \bar{h}}{\alpha^2} \cdot \|g_t^0\|_2^2 + \left(1 + 5\delta_t^j \bar{h}\right) \|H_t g_t^j - g_t^0\|_2^2 .
\end{aligned}
$$

For term [3], because we sample uniformly without replacement, we obtain:

$$
\mathbb{E}_{I_i} \left[\left\| e_t^j \right\|_2^2 \middle| g_t^j, \theta_t \right] = \tfrac{1}{S_i} \left(1 - \tfrac{S_i - 1}{n - 1}\right) \mathbb{E}_k \left[\left\| \tfrac{\nabla f_k(\theta_t + \delta_t^j g_t^j) - \nabla f_k(\theta_t)}{\delta_t^j} - \tfrac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} \right\|_2^2 \right]
$$

where k is uniformly sampled from $[n]$. Denote $H_t^k = \nabla^2 f_k(\theta_t)$, and by Lipschitz gradient we have $\|H_t^k\|_2 \leq \beta_k$. We can bound the above

$$
\begin{aligned}
&\left\| \frac{\nabla f_k(\theta_t+\delta_t^j g_t^j)-\nabla f_k(\theta_t)}{\delta_t^j} - \frac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j} \right\|_2^2 \\
&\quad = \left\| \frac{\nabla f_k(\theta_t+\delta_t^j g_t^j)-\nabla f_k(\theta_t)}{\delta_t^j} - H_t^k g_t^j + H_t^k g_t^j - \frac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j} + H_t g_t^j - H_t g_t^j \right\|_2^2 \\
&\quad \leq 3\left(\left\| \left(H_t - H_t^k\right) g_t^j \right\|_2^2 + \left\| \frac{\nabla f_k(\theta_t+\delta_t^j g_t^j)-\nabla f_k(\theta_t)}{\delta_t^j} - H_t^k g_t^j \right\|_2^2 + \left\| \frac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j} - H_t g_t^j \right\|_2^2 \right) \\
&\quad \leq 3\left(\left\| H_t - H_t^k \right\|_2^2 + (\delta_t^j)^2 \left(\bar{h}^2 + h_k^2\right) \right) \left\| g_t^j \right\|_2^2 \\
&\quad \|H_t - H_t^k\|_2^2 \leq 2(\bar{\beta}^2 + \beta_k^2) \\
&\quad \leq 3\left(2\left(\bar{\beta}^2 + \beta_k^2\right) + (\delta_t^j)^2(\bar{h}^2 + h_k^2) \right) \left\| g_t^j \right\|_2^2 \\
&\quad \leq 6\left(2\left(\bar{\beta}^2 + \beta_k^2\right) + (\delta_t^j)^2(\bar{h}^2 + h_k^2) \right) \left(\left\| g_t^j - H_t^{-1} g_t^0 \right\|_2^2 + \left\| H_t^{-1} g_t^0 \right\|_2^2 \right)
\end{aligned}
$$

Taking the expectation over inner loop's random indices, for term [3], we have

$$
\begin{aligned}
\mathbb{E}_{I_i}\left[\left\| e_t^j \right\|_2^2 \middle| g_t^j, \theta_t \right] &\leq 6\left(\frac{1}{S_i} \cdot \left(1 - \frac{S_i - 1}{n-1}\right) \right) \left(\left(\delta_t^j \bar{h}\right)^2 + 2\bar{\beta}^2 + (\delta_t^j)^2 \bar{h}_2 + 2\bar{\beta}_2 \right) \left(\left\| g_t^j - H_t^{-1} g_t^0 \right\|_2^2 + \frac{1}{\alpha^2} \cdot \left\| g_t^0 \right\|_2^2 \right) \\
&\leq 18\left(\frac{1}{S_i} \left(1 - \frac{S_i - 1}{n-1}\right) \right) \left((\delta_t^j)^2 \bar{h}_2 + \bar{\beta}_2 \right) \left(\left\| g_t^j - H_t^{-1} g_t^0 \right\|_2^2 + \frac{1}{\alpha^2} \left\| g_t^0 \right\|_2^2 \right) \qquad (35)
\end{aligned}
$$

Combining all above, we have

$$
\begin{aligned}
\mathbb{E}\left[\left\| g_t^{j+1} - H_t^{-1} g_t^0 \right\|_2^2 \middle| g_t^j, \theta_t \right] &\leq \left\| g_t^j - H_t^{-1} g_t^0 \right\|_2^2 \\
&\quad - \tau_j \left(g_t^j - H_t^{-1} g_t^0 \right)^\top H_t \left(g_t^j - H_t^{-1} g_t^0 \right) + \frac{3\tau_j \delta_t^j \bar{h}}{2} \left\| g_t^j - H_t^{-1} g_t^0 \right\|_2^2 + \frac{\tau_j \delta_t^j \bar{h}}{\alpha^2} \left\| g_t^0 \right\|_2^2 \\
&\quad + \frac{4\tau_j^2 \delta_t^j \bar{h}}{\alpha^2} \cdot \left\| g_t^0 \right\|_2^2 + \tau_j^2 \left(1 + 5\delta_t^j \bar{h}\right) \left\| H_t g_t^j - g_t^0 \right\|_2^2 \\
&\quad + 18\tau_j^2 \left(\frac{1}{S_i} \left(1 - \frac{S_i - 1}{n-1}\right) \right) \left((\delta_t^j)^2 \bar{h}_2 + \bar{\beta}_2 \right) \left(\left\| g_t^j - H_t^{-1} g_t^0 \right\|_2^2 + \frac{1}{\alpha^2} \|g_t^0\|_2^2 \right)
\end{aligned}
$$

When we choose the Hessian vector product approximation scaling constant δ_t^j to be sufficiently small

$$
\begin{aligned}
\delta_t^j \bar{h} \leq \delta_t^j \sqrt{\bar{h}_2} &\leq 0.01, \\
\frac{3\delta_t^j \bar{h}}{2} &\leq 0.01\alpha, \\
\delta_t^j \bar{h} \leq \delta_t^j \sqrt{\bar{h}_2} &\leq \frac{0.01}{S_i}\left(1 - \frac{S_i - 1}{n-1}\right)\bar{\beta}^2 \leq \frac{0.01}{S_i}\left(1 - \frac{S_i - 1}{n-1}\right)\bar{\beta}_2, \\
\delta_t^j \bar{h} \leq \delta_t^j \sqrt{\bar{h}_2} &\leq \frac{0.01\tau_j}{S_i}\left(1 - \frac{S_i - 1}{n-1}\right)\bar{\beta}^2 \leq \frac{0.01\tau_j}{S_i}\left(1 - \frac{S_i - 1}{n-1}\right)\bar{\beta}_2,
\end{aligned}
$$

$$\delta_t^j \bar{h} \le \delta_t^j \sqrt{\bar{h}_2} \le 0.01\alpha \le 0.01\bar{\beta} \le 0.01\sqrt{\bar{\beta}_2},$$

we have

$$\mathbb{E}\left[\left\|g_t^{j+1} - H_t^{-1}g_t^0\right\|_2^2 \mid g_t^j, \theta_t\right] \le \left\|g_t^j - H_t^{-1}g_t^0\right\|_2^2 \underbrace{-\tau_j \left(g_t^j - H_t^{-1}g_t^0\right)^\top H_t \left(g_t^j - H_t^{-1}g_t^0\right)}_{[4]} + 1.05\tau_j^2 \|H_t g_t^j - g_t^0\|_2^2$$
$$+ 18.5\tau_j^2 \left(\frac{1}{S_i}\left(1 - \frac{S_i - 1}{n-1}\right)\right)\beta_2 \left\|g_t^j - H_t^{-1}g_t^0\right\|_2^2$$
$$+ 18.5\tau_j^2 \left(\frac{1}{S_i}\left(1 - \frac{S_i - 1}{n-1}\right)\right)\frac{\bar{\beta}_2}{\alpha^2} \cdot \|g_t^0\|_2^2.$$

For term [4], let us consider the α strongly convex and β smooth quadratic function

$$F(g) = \tfrac{1}{2} g^\top H_t g - \langle g_t^0, g\rangle,$$

who attains its minimum at $g = H_t^{-1}g_t^0$. Using a well known property of α strongly convex and β smooth functions (Lemma G.1), we have

$$-\left(g_t^j - H_t^{-1}g_t^0\right)^\top H_t \left(g_t^j - H_t^{-1}g_t^0\right) + \tfrac{1}{2\beta}\|H_t g_t^j - g_t^0\|_2^2 \le -\left(g_t^j - H_t^{-1}g_t^0\right)^\top H_t \left(g_t^j - H_t^{-1}g_t^0\right) + \tfrac{1}{\alpha+\beta}\|H_t g_t^j - g_t^0\|_2^2$$
$$\le -\tfrac{\alpha\beta}{\alpha+\beta}\|g_t^j - H_t^{-1}g_t^0\|_2^2$$
$$\le -\tfrac{\alpha}{2}\|g_t^j - H_t^{-1}g_t^0\|_2^2.$$

Thus, when we choose

$$\tau_j \le \tfrac{0.476}{\beta},$$

we have

$$-\tau_j \left(g_t^j - H_t^{-1}g_t^0\right)^\top H_t \left(g_t^j - H_t^{-1}g_t^0\right) + 1.05\tau_j^2 \cdot \left\|H_t g_t^j - g_t^0\right\|_2^2 \le -\tau_j \left(g_t^j - H_t^{-1}g_t^0\right)^\top H_t \left(g_t^j - H_t^{-1}g_t^0\right) + \tfrac{\tau_j}{2\beta}\left\|H_t g_t^j - g_t^0\right\|_2^2$$
$$\le -\tfrac{\tau_j \alpha}{2} \cdot \|g_t^j - H_t^{-1}g_t^0\|_2^2,$$

and we have

$$\mathbb{E}\left[\left\|g_t^{j+1} - H_t^{-1}g_t^0\right\|_2^2 \mid g_t^j, \theta_t\right] \le \left(1 - \tau_j\alpha + 18.5\tau_j^2\left(\frac{1}{S_i}\left(1 - \frac{S_i-1}{n-1}\right)\right)\beta_2\right)\left\|g_t^j - H_t^{-1}g_t^0\right\|_2^2$$
$$+ 18.5\tau_j^2\left(\frac{1}{S_i}\left(1 - \frac{S_i-1}{n-1}\right)\right)\frac{\bar{\beta}_2}{\alpha^2} \cdot \|g_t^0\|_2^2.$$

Next, we set

$$\tau_0 = \min\left\{\frac{0.476}{\beta}, \frac{0.025\cdot\alpha}{\frac{1}{S_i}\left(1-\frac{S_i-1}{n-1}\right)\bar{\beta}_2}\right\}\quad D_j = (j+1)^{-d_i}, \quad \tau_j = \tau_0 D_j, \tag{36}$$

where d_i is inner loop's step size decay rate, and we have:

$$\mathbb{E}\left[\left\|g_t^{j+1} - H_t^{-1}g_t^0\right\|_2^2 \mid \theta_t\right] \leq \left(1 - \min\left\{\frac{\alpha}{2\beta}, \frac{0.013\cdot\alpha^2}{\frac{1}{S_i}\left(1-\frac{S_i-1}{n-1}\right)\bar{\beta}_2}\right\} D_j\right)\cdot \mathbb{E}\left[\left\|g_t^j - H_t^{-1}g_t^0\right\|_2^2 \mid \theta_t\right] + 18.5 D_j^2\tau_0^2\left(\frac{1}{S_i}\left(1-\frac{S_i-1}{n-1}\right)\right)\frac{\bar{\beta}_2}{\alpha^2}\cdot\left\|g_t^0\right\|_2^2.$$

To satisfy the above requirements, for the Hessian vector product approximation scaling constant, we choose:

$$\delta_t^j = o\quad \min\left\{1, \frac{1}{h}\right\}\min\left\{1, \alpha, \min\left\{1, \tau_0^4\left(\frac{\tau_j}{\tau_0}\right)^4\right\}\frac{1}{S_i}\left(1-\frac{S_i-1}{n-1}\right)\right\}\cdot\delta_t^0 = o\left((j+1)^{-2}\right)\delta_t^0,$$
$$\delta_t^0 = O(\rho_t^4) = o((t+1)^{-2}) = o(1). \tag{37}$$

which is trivially satisfied for quadratic functions because all $h_i = 0$.

Note that:

$$18.5\tau_0^2\left(\frac{1}{S_i}\left(1-\frac{S_i-1}{n-1}\right)\right)\frac{\bar{\beta}_2}{\alpha^2} = \Theta\left(\min\left\{\left(\frac{1}{S_i}\left(1-\frac{S_i-1}{n-1}\right)\right)\left(\frac{\bar{\beta}_2}{\beta^2\alpha^2}, \frac{1}{\frac{1}{S_i}\left(1-\frac{S_i-1}{n-1}\right)\cdot\bar{\beta}_2}\right\}\right).$$

Applying Lemma G.2, we have:

$$\mathbb{E}\left[\left\|g_t^j - H_t^{-1}g_t^0\right\|_2^2 \mid \theta_t\right] = O\left(t^{-d_i}\cdot\|g_t^0\|_2^2\right) \tag{38}$$

where we have assumed that α, β, S_i, etc. are (data dependent) constants. Further, (38) implies:

$$\mathbb{E}\left[\left\|g_t^j\right\|_2^2\right] \leq 2\mathbb{E}\left[\left\|g_t^j - H_t^{-1}g_t^0\right\|_2^2 + \left\|H_t^{-1}g_t^0\right\|_2^2 \mid \theta_t\right] \lesssim \|g_t^0\|_2^2, \quad \text{for all } j. \tag{39}$$

In Algorithm 1, we have

$$g_t^{j+1} - H_t^{-1}g_t^0 = (I - \tau_j H_t)(g_t^j - H_t^{-1}g_t^0) + \tau_j\left(e_t^j - \frac{\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)}{\delta_t^j} + H_t g_t^j\right)$$

By unrolling the recursion we have:

$$g_t^{j+1} - H_t^{-1}g_t^0 = \sum_{k=0}^{j}\left(\prod_{l=k+1}^{j}(I - \tau_l H_t)\right)\left(\tau_k\cdot\left(e_t^k - \frac{\nabla f(\theta_t + \delta_t^k g_t^k) - \nabla f(\theta_t)}{\delta_t^k} + H_t g_t^k\right)\right. \tag{40}$$

For the average $\bar{g}_t$, we have:

$$\begin{aligned}
\bar{g}_t - H_t^{-1} g_t^0 &= \tfrac{1}{L+1} \sum_{j=0}^{L} (g_t^j - H_t^{-1} g_t^0) \\
&= \tfrac{1}{L+1} \sum_{j=0}^{L} \sum_{k=0}^{j-1} \left(\prod_{l=k+1}^{j-1} (I - \tau_l H_t) \right) \left(\tau_k \left(-e_t^k - \tfrac{\nabla f(\theta_t + \delta_t^k g_t^k) - \nabla f(\theta_t)}{\delta_t^k} + H_t g_t^k \right) \right. \\
&= \tfrac{1}{L+1} \sum_{k=0}^{L-1} \underbrace{\left(\tau_k \sum_{j=k+1}^{L} \prod_{l=k+1}^{j-1} (I - \tau_l H_t) \right)}_{[5]} \left(-e_t^k - \tfrac{\nabla f(\theta_t + \delta_t^k g_t^k) - \nabla f(\theta_t)}{\delta_t^k} + H_t g_t^k \right) \\
&= \underbrace{\tfrac{1}{L+1} \sum_{k=0}^{L-1} \left(\tau_k \sum_{j=k+1}^{L} \prod_{l=k+1}^{j-1} (I - \tau_l H_t) \right) \left(-e_t^k \right)}_{[6]} \\
&\quad + \underbrace{\tfrac{1}{L+1} \sum_{k=0}^{L-1} \left(\tau_k \sum_{j=k+1}^{L} \prod_{l=k+1}^{j-1} (I - \tau_l H_t) \right) \left(- \tfrac{\nabla f(\theta_t + \delta_t^k g_t^k) - \nabla f(\theta_t)}{\delta_t^k} + H_t g_t^k \right)}_{[7]}
\end{aligned} \tag{41}$$

For the term [5], we have:

$$\begin{aligned}
\left\| \left(\tau_k \sum_{j=k+1}^{L} \prod_{l=k+1}^{j-1} (I - \tau_l H_t) \right) \right\|_2 &\le \tau_k \sum_{j=k+1}^{L} \prod_{l=k+1}^{j-1} \|I - \tau_l H_t\|_2 \\
&\quad I - \tau_l H_t \text{ is positive definite by our choice of } \tau_l \text{ (36) and } \|I - \tau_l H_t\|_2 \le 1 - \tau_l \alpha \\
&\le \tau_k \sum_{j=k+1}^{L} \prod_{l=k+1}^{j-1} (1 - \tau_l \alpha) \\
&\le \tau_k \sum_{j=k+1}^{L} \prod_{l=k+1}^{j-1} \left(1 - \frac{1}{2} \tau_l \alpha \right)^2 \\
&\quad \tau_k \prod_{l=k+1}^{j-1} (1 - \frac{1}{2} \tau_l \alpha) \le \tau_k \exp(-\frac{1}{2} \alpha \sum_{l=k+1}^{j-1} \tau_l) \lesssim k^{-d_i} \exp(\Theta(-j^{1-d_i} + k^{1-d_i})) \lesssim j^{-d_i} \lesssim \tau_j \\
&\quad \text{because for a fixed } d_i \; x^{-d_i} e^{\Theta(x^{1-d_i})} \text{ is an increasing function when } x \text{ is sufficiently large} \\
&\lesssim \sum_{j=k+1}^{L} \tau_j \prod_{l=k+1}^{j-1} (1 - \tfrac{\tau_l \alpha}{2}) = \tfrac{2}{\alpha} \sum_{j=k+1}^{L} \left(\tfrac{1}{2} \tau_j \alpha \prod_{l=k+1}^{j-1} (1 - \tfrac{\tau_l \alpha}{2}) \right) \\
&= \tfrac{2}{\alpha} \left(1 - \prod_{j=k+1}^{L} (1 - \tfrac{\tau_l \alpha}{2}) \right) \le \tfrac{2}{\alpha} = O(1),
\end{aligned} \tag{42}$$

where we have assumed that α, β, S_i, etc. are (data-dependent) constants.

For the term [6], its norm is bounded by:

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{L+1}\sum_{k=0}^{L-1}\tau_k\sum_{j=k+1}^{L}\prod_{l=k+1}^{j-1}(I-\tau_l H_t)(-e_t^k)\right\|_2^2 \mid \theta_t\right] &= \frac{1}{(L+1)^2}\mathbb{E}\left[\sum_{k=0}^{L-1}\left\|\tau_k\sum_{j=k+1}^{L}\prod_{l=k+1}^{j-1}(I-\tau_l H_t)(-e_t^k)\right\|_2^2 \mid \theta_t\right] \\
&\quad \text{using (42)} \\
&\lesssim \frac{1}{(L+1)^2}\mathbb{E}\left[\sum_{k=0}^{L-1}\|e_t^k\|_2^2 \mid \theta_t\right] \\
&\quad \text{using (35) and (38)} \\
&\lesssim \frac{1}{L}\|g_t^0\|_2^2. \qquad (43)
\end{aligned}$$

where the first equality is due to $a < b$, $\mathbb{E}[{e_t^a}^\top e_t^b \mid \theta_t] = 0$, when we first condition on b.

For the term [7], its norm is bounded by:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{L+1}\sum_{k=0}^{L-1}\tau_k\sum_{j=k+1}^{L}\prod_{l=k+1}^{j-1}(I-\tau_l H_t)\left(\frac{\nabla f(\theta_t+\delta_t^k g_t^k)-\nabla f(\theta_t)}{\delta_t^k}+H_t g_t^k\right)\right\|_2^2 \mid \theta_t\right] \\
&= \frac{1}{(L+1)^2}\mathbb{E}\left[\sum_{0\le a,b,\le L-1}\left\langle \tau_a\sum_{j=a+1}^{L}\prod_{l=a+1}^{j-1}(I-\tau_l H_t)\left(\frac{\nabla f(\theta_t+\delta_t^a g_t^a)-\nabla f(\theta_t)}{\delta_t^a}+H_t g_t^a\right),\right.\right. \\
&\qquad \left.\left.\tau_b\sum_{j=b+1}^{L}\prod_{l=b+1}^{j-1}(I-\tau_l H_t)\left(\frac{\nabla f(\theta_t+\delta_t^b g_t^b)-\nabla f(\theta_t)}{\delta_t^b}+H_t g_t^b\right)\right\rangle \mid \theta_t\right] \\
&\le \frac{1}{(L+1)^2}\mathbb{E}\left[\sum_{0\le a,b,\le L-1}\left\|\tau_a\sum_{j=a+1}^{L}\prod_{l=a+1}^{j-1}(I-\tau_l H_t)\left(\frac{\nabla f(\theta_t+\delta_t^a g_t^a)-\nabla f(\theta_t)}{\delta_t^a}+H_t g_t^a\right)\right\|_2\right. \\
&\qquad \left.\cdot\left\|\tau_b\sum_{j=b+1}^{L}\prod_{l=b+1}^{j-1}(I-\tau_l H_t)\left(\frac{\nabla f(\theta_t+\delta_t^b g_t^b)-\nabla f(\theta_t)}{\delta_t^b}+H_t g_t^b\right)\right\|_2 \mid \theta_t\right] \\
&\quad \text{using (42) and Lemma F.1} \\
&\lesssim \frac{1}{(L+1)^2}\mathbb{E}\left[\sum_{0\le a,b,\le L-1}\delta_t^a\bar{h}\|g_t^a\|_2\delta_t^b\bar{h}\|g_t^b\|_2 \mid \theta_t\right] \le \frac{2\bar{h}^2}{(L+1)^2}\sum_{0\le a,b,\le L-1}\delta_t^a\delta_t^b\cdot\mathbb{E}\left[\|g_t^a\|_2^2+\|g_t^b\|_2^2 \mid \theta_t\right] \\
&\lesssim \frac{\|g_t^0\|_2^2}{(L+1)^2}\sum_{0\le a,b,\le L-1}\delta_t^a\delta_t^b \lesssim \frac{\|g_t^0\|_2^2}{L^2}\left(\sum_{k=0}^{L}\delta_t^k\right)^2 \qquad (44) \\
&\quad \text{using (39) and our choice of } \delta_t^k \text{ (37)}
\end{aligned}$$

$$\lesssim \frac{1}{L^2}\delta_t^{0^2}\left(\sum_{k=0}^{L}\tau_k\right)^2\cdot\|g_t^0\|_2^2 \lesssim \frac{1}{L^2}\delta_t^{0^2}\left(\sum_{k=0}^{L}(k+1)^{-d_i}\right)^2\cdot\|g_t^0\|_2^2$$

$$\text{because}\quad \left(\sum_{k=0}^{L}(k+1)^{-d_i}\right)^2 = O\left(L^{1-d_i}\right)\ \text{and}\ d_i\in\left(\tfrac{1}{2},1\right)$$

$$\ll \frac{1}{L}\|g_t^0\|_2^2. \tag{45}$$

Combining (43) and (45), we have

$$\|\bar{g}_t - H_t^{-1}g_t^0\|_2^2 = O\left(\tfrac{1}{L}\|g_t^0\|_2^2\right)$$

F.1.2. PROOF OF (9)

Using (32), we have

$$\begin{aligned}
&\mathbb{E}[\|g_t^{j+1} - H_t^{-1}g_t^0\|_2^4 \mid g_t^j]\\
&= \mathbb{E}[\|g_t^j - H_t^{-1}g_t^0 - \tau_j\tfrac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j} + \tau_j g_t^0 - \tau_j e_t^j\|_2^4 \mid g_t^j]\\
&= \mathbb{E}[(\|g_t^j - H_t^{-1}g_t^0 - \tau_j\tfrac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j} + \tau_j g_t^0\|_2^2\\
&- 2\langle\tau_j e_t^j, g_t^j - H_t^{-1}g_t^0 - \tau_j\tfrac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j} + \tau_j g_t^0\rangle + \tau_j^2\|e_t^j\|_2^2)^2 \mid g_t^j]\\
&= \mathbb{E}[\|g_t^j - H_t^{-1}g_t^0 - \tau_j\tfrac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j} + \tau_j g_t^0\|_2^4\\
&+ 4(\langle\tau_j e_t^j, g_t^j - H_t^{-1}g_t^0 - \tau_j\tfrac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j} + \tau_j g_t^0\rangle)^2 + \tau_j^4\|e_t^j\|_2^4\\
&+ 2\|g_t^j - H_t^{-1}g_t^0 - \tau_j\tfrac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j} + \tau_j g_t^0\|_2^2\tau_j^2\|e_t^j\|_2^2\\
&- 4\langle\tau_j e_t^j, g_t^j - H_t^{-1}g_t^0 - \tau_j\tfrac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j} + \tau_j g_t^0\rangle\|g_t^j - H_t^{-1}g_t^0 - \tau_j\tfrac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j} + \tau_j g_t^0\|_2^2\\
&- 4\langle\tau_j e_t^j, g_t^j - H_t^{-1}g_t^0 - \tau_j\tfrac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j} + \tau_j g_t^0\rangle\tau_j^2\|e_t^j\|_2^2 \mid g_t^j].
\end{aligned} \tag{46}$$

Because we have

$$\mathbb{E}[e_t^j \mid g_t^j] = 0,$$

$$\begin{aligned}
&\|g_t^j - H_t^{-1}g_t^0 - \tau_j\tfrac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j} + \tau_j g_t^0\|_2^4\\
&= \|(I-\tau_j H_t)(g_t^j - H_t^{-1}g_t^0) + \tau_j(-\tfrac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j} + H_t g_t^j)\|_2^4\\
&= (\|(I-\tau_j H_t)(g_t^j - H_t^{-1}g_t^0)\|_2^2
\end{aligned}$$

$$+ 2\tau_j\langle (I-\tau_j H_t)(g_t^j - H_t^{-1}g_t^0), -\tfrac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j} + H_t g_t^j\rangle$$
$$+ \tau_j^2 \| -\tfrac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j} + H_t g_t^j\|_2^2)^2$$

using Lemma F.1

$$\leq (\|(I-\tau_j H_t)(g_t^j - H_t^{-1}g_t^0)\|_2^2 + 2\tau_j \|I-\tau_j H_t\|_2 \|g_t^j - H_t^{-1}g_t^0\|_2 \delta_t^j \|g_t^j\|_2 + \tau_j^2 {\delta_t^j}^2 \|g_t^j\|_2^2)^2$$
$$= \|(I-\tau_j H_t)(g_t^j - H_t^{-1}g_t^0)\|_2^4$$
$$+ 2\tau_j \|(I-\tau_j H_t)(g_t^j - H_t^{-1}g_t^0)\|_2^2 (2\delta_t^j \|I-\tau_j H_t\|_2 \|g_t^j - H_t^{-1}g_t^0\|_2 \|g_t^j\|_2 + \tau_j {\delta_t^j}^2 \|g_t^j\|_2^2)$$
$$+ \tau_j^2 (2\delta_t^j \|I-\tau_j H_t\|_2 \|g_t^j - H_t^{-1}g_t^0\|_2 \|g_t^j\|_2 + \tau_j {\delta_t^j}^2 \|g_t^j\|_2^2)^2$$

by our choice of $\tau_j = \Theta((j+1)^{-d_i}) = o(1)$ (36)

and using $\|g_t^j\|_2 \leq \|g_t^j - H_t^{-1}g_t^0\|_2 + \|H_t^{-1}g_t^0\|_2 \lesssim \|g_t^j - H_t^{-1}g_t^0\|_2 +^4 \|g_t^0\|_2$

$$= (1-\Theta(\tau_j))\|g_t^j - H_t^{-1}g_t^0\|_2^4$$
$$+ O(\tau_j \delta_t^j (\|g_t^j - H_t^{-1}g_t^0\|_2^4 + \|g_t^j - H_t^{-1}g_t^0\|_2^3 \|g_t^0\|_2) + 2\tau_j^2 {\delta_t^j}^3 (\|g_t^j - H_t^{-1}g_t^0\|_2^4 + \|g_t^j - H_t^{-1}g_t^0\|_2^2 \|g_t^0\|_2^2)$$
$$+ \tau_j^2 {\delta_t^j}^2 (\|g_t^j - H_t^{-1}g_t^0\|_2^4 + \|g_t^j - H_t^{-1}g_t^0\|_2^2 \|g_t^0\|_2^2 + \tau_j \delta_t^j (\|g_t^j - H_t^{-1}g_t^0\|_2^4 + \|g_t^0\|_2^4))),$$

$$\mathbb{E}[\|e_t^j\|_2^4 \mid g_t^j]$$
$$= \mathbb{E}[\| \left(\frac{1}{S_i}\frac{1}{\delta_t^j} \sum_{k\in I_i} \left(\nabla f_k(\theta_t + \delta_t^j g_t^j) - \nabla f_k(\theta_t) \right) \right) - \tfrac{\nabla f(\theta_t+\delta_t^j g_t^j)-\nabla f(\theta_t)}{\delta_t^j} \|_2^4 \mid g_t^j]$$
$$= \mathbb{E}[\| \left(\frac{1}{S_i}\frac{1}{\delta_t^j} \sum_{k\in I_i} \left((\nabla f_k(\theta_t + \delta_t^j g_t^j) - \nabla f_k(\theta_t)) - H_t^k g_t^j + H_t^k g_t^j \right) \right) \Big($$
$$- (\frac{1}{\delta_t^j}(\nabla f(\theta_t + \delta_t^j g_t^j) - \nabla f(\theta_t)) - H_t g_t^j + H_t g_t^j) \|_2^4 \mid g_t^j]$$

using Lemma F.1 and repeatedly applying the AM-GM inequality

$$\lesssim (1 + {\delta_t^j}^4)\|g_t^j\|_2^4$$
$$\lesssim (1 + {\delta_t^j}^4){\delta_t^j}^4 (\|g_t^j - H_t^{-1}g_t^0\|_2^4 + \|g_t^0\|_2^4),$$

and by our choice of $\tau_j = \Theta((j+1)^{-d_i}) = o(1)$ (36) and $\delta_t^j = O(\tau_j^4)$ (37), after repeatedly applying the AM-GM inequality, Lemma F.1, triangle inequality, and (38), we can bound (46) by

$$\mathbb{E}[\|g_t^{j+1} - H_t^{-1}g_t^0\|_2^4 \mid g_t^j]$$
$$\leq (1-\Theta(\tau_j))\|g_t^j - H_t^{-1}g_t^0\|_2^4 + O(\tau_j^3 \|g_t^0\|_2^4)). \tag{47}$$

Applying Lemma G.2, we have

$$\mathbb{E}[\|g_t^{j+1} - H_t^{-1} g_t^0\|_2^4 \mid \theta_t] = O((j+1)^{-2d_i} \|g_t^0\|_2^4), \tag{48}$$

and using the AM-GM in equality we have

$$\mathbb{E}[\|g_t^{j+1}\|_2^4 \mid \theta_t] = O(\|g_t^0\|_2^4). \tag{49}$$

F.1.3. PROOF OF (6)

To prove bounds on $\|\theta_t - \widehat{\theta}\|_2^2$, we will use the following lemma

Lemma F.2

$$\begin{aligned}\mathbb{E}[\langle \nabla f(\theta_t), -g_t^L \rangle \mid \theta_t] \gtrsim & \rho_t \|\nabla f(\theta_t)\|_2^2 - \delta_t^0 \|\nabla f(\theta_t)\|_2 \|g_t^0\|_2 \\ \gtrsim & \rho_t \|\nabla f(\theta_t)\|_2^2 - {\delta_t^0}^2 \|g_t^0\|_2^2.\end{aligned}$$

Proof

Using (40), and because $\mathbb{E}[e_t^j \mid \theta_t = 0]$, we have

$$\begin{aligned}&\mathbb{E}[\langle \nabla f(\theta_t), -g_t^L \rangle \mid \theta_t] \\ &= \rho_t \nabla f(\theta_t)^\top H_t^{-1} \nabla f(\theta_t) - \mathbb{E}\left[\left\langle \nabla f(\theta_t), \sum_{k=0}^{L-1} \Big(\prod_{l=k+1}^{L-1} (I - \tau_l H_t) \Big) \tau_k \Big(\tfrac{\nabla f(\theta_t + \delta_t^k g_t^k) - \nabla f(\theta_t)}{\delta_t^k} - H_t g_t^k \Big) \right\rangle \,\middle|\, \theta_t \right] \\ &\text{using strong convexity and Lemma F.1} \\ &\geq \frac{1}{\beta} \rho_t \|\nabla f(\theta_t)\|_2^2 - \|\nabla f(\theta_t)\|_2 \underbrace{\mathbb{E}\left[\sum_{k=0}^{L-1} \prod_{l=k+1}^{L-1} \|I - \tau_l H_t\|_2 \tau_k \delta_t^k \|g_t^k\|_2 \,\middle|\, \theta_t \right]}_{[8]}\end{aligned}$$

By our choice of $\tau_j = \Theta((j+1)^{-d_i}) = o(1)$ (36) and $\delta_t^j = O(\delta_t^0 \tau_j^4)$ (37), and using (39), term [8] is bounded by

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=0}^{L-1} \prod_{l=k+1}^{L-1} \|I - \tau_l H_t\|_2 \tau_k \delta_t^k \|g_t^k\|_2 \mid \theta_t \right] \\ \lesssim & \sum_{k=0}^{L-1} \tau_k \delta_t^k \\ \lesssim & \|g_t^0\|_2 \delta_t^0 \underbrace{\sum_{k=0}^{L-1} \tau_k^5}_{=O(1)}.\end{aligned}$$

And we can conclude

$$\begin{aligned}
&\mathbb{E}[\langle \nabla f(\theta_t), -g_t^L \rangle \mid \theta_t] \\
&\geq C_1 \rho_t \|\nabla f(\theta_t)\|_2^2 - C_2 \delta_t^0 \|\nabla f(\theta_t)\|_2 \|g_t^0\|_2 \\
&= C_1 \rho_t \|\nabla f(\theta_t)\|_2^2 - \frac{C_1}{2} \delta_t^{0\,2} \left[2 \frac{\|\nabla f(\theta_t)\|_2}{\delta_t^0} \frac{C_2}{C_1} \|g_t^0\|_2 \right] \\
&\geq C_1 \rho_t \|\nabla f(\theta_t)\|_2^2 - \frac{C_1}{2} \delta_t^{0\,2} \left((\frac{\|\nabla f(\theta_t)\|_2}{\delta_t^0})^2 + (\frac{C_2}{C_1} \|g_t^0\|_2 \right)^2 \\
&= \frac{C_1}{2} \rho_t \|\nabla f(\theta_t)\|_2^2 - \frac{C_2^2}{2C_1} \delta_t^{0\,2} \|g_t^0\|_2^2,
\end{aligned}$$

for some (data dependent) positive constants C_1, C_2. ■

Now, we continue our proof of (6).

In Algorithm 1, because f is β smooth, we have

$$\begin{aligned}
&\mathbb{E}[f(\theta_{t+1}) - f(\widehat{\theta}) \mid \theta_t] \\
&= \mathbb{E}[f(\theta_t + g_t^L) - f(\widehat{\theta}) \mid \theta_t] \\
\leq & f(\theta_t) - f(\widehat{\theta}) + \mathbb{E}\left[\langle \nabla f(\theta_t), g_t^L \rangle + \frac{\beta}{2} \|g_t^L\|_2^2 \mid \theta_t \right] \\
& \text{using Lemma F.2 and (39)} \\
\leq & f(\theta_t) - f(\widehat{\theta}) - \Omega(\rho_t \|\nabla f(\theta_t)\|_2^2) + \mathbb{E}[O(\|g_t^0\|_2^2 + \delta_t^0 \|g_t^0\|_2 \|\nabla f(\theta_t)\|_2) \mid \theta_t].
\end{aligned} \tag{50}$$

For g_t^0, we have

$$\begin{aligned}
\frac{g_t^0}{\rho_t} &= \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\theta_t) \\
&= \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\widehat{\theta}) + \frac{1}{S_o} \sum_{i \in I_o} (\nabla f_i(\theta_t) - \nabla f_i(\widehat{\theta})),
\end{aligned} \tag{51}$$

which implies that

$$\begin{aligned}
&\mathbb{E}\left[\left\| \frac{g_t^0}{\rho_t} \right\|_2^2 \mid \theta_t \right] \\
\leq & 2\mathbb{E}\left[\left\| \frac{1}{S_o} \sum_{i \in I_o} \nabla f_i(\widehat{\theta}) \right\|_2^2 \mid \theta_t \right] + 2\mathbb{E}[\| \frac{1}{S_o} \sum_{i \in I_o} (\nabla f_i(\theta_t) - \nabla f_i(\widehat{\theta})) \|_2^2 \mid \theta_t] \\
& \text{because we sample uniformly with replacement and } \nabla f(\widehat{\theta}) = 0
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{2}{S_o}\sum_{i=1}^{n}\Big(\|\nabla f_i(\widehat{\theta})\|_2^2 + \mathbb{E}[\|\nabla f_i(\theta_t) - \nabla f_i(\widehat{\theta})\|_2^2 \mid \theta_t]\\
&\leq \frac{2}{S_o}\sum_{i=1}^{n}\Big(\|\nabla f_i(\widehat{\theta})\|_2^2 + \|\theta_t - \widehat{\theta}\|_2^2\mathbb{E}[\beta_i^2 \mid \theta_t]\\
&\lesssim 1 + \|\theta_t - \widehat{\theta}\|_2^2. \qquad (52)
\end{aligned}
$$

Thus, continuing (50), using (52) and strong convexity $\alpha^2\|\theta_t - \widehat{\theta}\|_2^2 \leq \|\nabla f(\theta_t)\|_2^2$, we have

$$
\begin{aligned}
&\mathbb{E}[f(\theta_{t+1}) - f(\widehat{\theta}) \mid \theta_t]\\
\leq& f(\theta_t) - f(\widehat{\theta}) - C_1\rho_t\|\nabla f(\theta_t)\|_2^2 + C_2\rho_t\delta_t^0(1 + \|\nabla f(\theta_t)\|_2)\|\nabla f(\theta_t)\|_2 + C_3\rho_t^2(1 + \|\nabla f(\theta_t)\|_2^2)\\
=& f(\theta_t) - f(\widehat{\theta}) - \rho_t(C_1 - C_2\delta_t^0 - C_3\rho_t)\|\nabla f(\theta_t)\|_2^2 + C_3\rho_t^2 + C_2\rho_t\delta_t^0\|\nabla f(\theta_t)\|_2\\
&\text{because we have } C_2\rho_t\delta_t^0\|\nabla f(\theta_t)\|_2 = \frac{1}{2}C_1\rho_t{\delta_t^0}^2 2\frac{\frac{C_2}{C_1}\|\nabla f(\theta_t)\|_2}{\delta_t^0} \leq \frac{1}{2}C_1\rho_t{\delta_t^0}^2\left(\left(\frac{C_2}{C_1}\right)^2 + \left(\frac{\|\nabla f(\theta_t)\|_2}{\delta_t^0}\right)^2\right)\\
\leq& f(\theta_t) - f(\widehat{\theta}) - \rho_t(\tfrac{1}{2}C_1 - C_2\delta_t^0 - C_3\rho_t)\|\nabla f(\theta_t)\|_2^2 + C_3\rho_t^2 + \tfrac{C_2^2}{C_1}\rho_t{\delta_t^0}^2\\
&\text{using strong convexity } \frac{1}{2\alpha}\|\nabla f(\theta_t)\|_2^2 \geq f(\theta_t) - f(\widehat{\theta}) \text{ and smoothness } \frac{1}{2\beta}\|\nabla f(\theta_t)\|_2^2 \leq f(\theta_t) - f(\widehat{\theta})\\
\leq& [f(\theta_t) - f(\widehat{\theta})] - \rho_t(\tfrac{1}{2}C_1 - C_2\delta_t^0 - C_3\rho_t)\tfrac{1}{2\alpha}[f(\theta_t) - f(\widehat{\theta})] + C_3\rho_t^2 + \tfrac{C_2^2}{C_1}\rho_t{\delta_t^0}^2\\
&\text{when we set } \delta_t^0 = O(\rho_t) \text{ in (37)}\\
\leq& [f(\theta_t) - f(\widehat{\theta})] - \rho_t(\tfrac{1}{2}C_1 - C_2\delta_t^0 - C_3\rho_t)\tfrac{1}{2\alpha}[f(\theta_t) - f(\widehat{\theta})] + (C_3 + O(1))\rho_t^2, \qquad (53)
\end{aligned}
$$

for some (data dependent) positive constants C_1, C_2, C_3.

In (53) we choose $\rho_t = \Theta((t+1)^{-d_o})$ for some $d_o \in (\frac{1}{2}, 1)$, and after applying Lemma G.2 we have

$$
\begin{aligned}
&\mathbb{E}[\|\theta_t - \widehat{\theta}\|_2^2]\\
\leq& \mathbb{E}[\tfrac{2}{\alpha}(f(\theta_t) - f(\widehat{\theta}))]\\
\lesssim& t^{-d_o} + e^{-\Theta(t^{1-d_o})}\|\theta_0 - \widehat{\theta}\|_2^2, \qquad (54)
\end{aligned}
$$

which is $O(t^{-d_o})$ when $\|\theta_0 - \widehat{\theta}\|_2 = O(1)$.

F.1.4. PROOF OF (7)

In Algorithm 1, because f is β smooth, and $\forall\theta\ f(\theta) - f(\widehat{\theta}) \geq 0$, we have

$$
\begin{aligned}
&(f(\theta_{t+1}) - f(\widehat{\theta}))^2\\
=& (f(\theta_t + g_t^L) - f(\widehat{\theta}))^2\\
\leq& (f(\theta_t) - f(\widehat{\theta}) + \langle\nabla f(\theta_t), g_t^L\rangle + \tfrac{\beta}{2}\|g_t^L\|_2^2)^2\\
=& (f(\theta_t) - f(\widehat{\theta}))^2 + 2\langle\nabla f(\theta_t), g_t^L\rangle(f(\theta_t) - f(\widehat{\theta}))
\end{aligned}
$$

$$+\langle\nabla f(\theta_t), g_t^L\rangle^2 + \tfrac{\beta^2}{4}\|g_t^L\|_2^4 + 2(f(\theta_t) - f(\widehat{\theta}) + \langle\nabla f(\theta_t), g_t^L\rangle)\tfrac{\beta}{2}\|g_t^L\|_2^2.$$

Because we have

$$\begin{aligned}&\mathbb{E}[\langle\nabla f(\theta_t), g_t^L\rangle(f(\theta_t) - f(\widehat{\theta})) \mid \theta_t]\\ \lesssim& -\rho_t\|\nabla f(\theta_t)\|_2^2(f(\theta_t) - f(\widehat{\theta})) + \delta_t^0\|g_t^0\|_2^2(f(\theta_t) - f(\widehat{\theta})),\end{aligned}$$

$$\begin{aligned}&\mathbb{E}\left[\left\|\frac{g_t^0}{\rho_t}\right\|_2^4 \mid \theta_t\right]\\ =&\,\mathbb{E}\left[\left\|\frac{1}{S_o}\sum_{i\in I_o}(\nabla f_i(\theta_t) - \nabla f_i(\widehat{\theta}) + \nabla f_i(\widehat{\theta}))\right\|_2^4 \mid \theta_t\right]\\ \lesssim& 1 + \|\theta_t - \widehat{\theta}\|_2^4,\end{aligned}$$

$$f(\theta_t) - f(\widehat{\theta}) = \Theta(\|\theta_t - \widehat{\theta}\|_2^2) = \Theta(\|\nabla f(\theta_t)\|_2^2),$$

and by our choice of $\rho_t = \Theta((t+1)^{-d_o}) = o(1)$ and $\delta_t^0 = O(\rho_t^4)$ (37), after repeatedly applying the AM-GM inequality and (54), we have

$$\begin{aligned}&\mathbb{E}[(f(\theta_{t+1}) - f(\widehat{\theta}))^2 \mid \theta_t]\\ \le&(1 - \Theta(\rho_t))(f(\theta_t) - f(\widehat{\theta}))^2 + O(\rho_t^3).\end{aligned}$$

Applying Lemma G.2, we have

$$\begin{aligned}&\mathbb{E}[\|\theta_t - \widehat{\theta}\|_2^4]\\ \le&\mathbb{E}\left[\tfrac{4}{\alpha^2}(f(\theta_t) - f(\widehat{\theta}))^2\right]\\ \lesssim& t^{-2d_o}.\end{aligned} \tag{55}$$

F.1.5. PROOF OF (10)

For $\frac{\bar{g}_t}{\rho_t}$, we have

$$\begin{aligned}\frac{\bar{g}_t}{\rho_t} =& \underbrace{-H^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta})}_{[1]}\\ &+ \underbrace{H^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta}) - H_t^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta}) + H_t^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta}) - H_t^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\theta_t)}_{[2]}\end{aligned}$$

$$\underbrace{-H_t^{-1}\frac{g_t^0}{\rho_t}+\frac{\bar{g}_t}{\rho_t}}_{[3]}. \tag{56}$$

Thus, for the "covariance" of our replicates, we have

$$\begin{aligned}
&\left\|H^{-1}GH^{-1}-\frac{S_o}{T}\sum_{t=1}^{T}\frac{\bar{g}_t\bar{g}_t^\top}{\rho_t^2}\right\|_2\\
\lesssim&\left\|H^{-1}GH^{-1}-\frac{S_o}{T}\sum_{t=1}^{T}[1]_t[1]_t^\top\right\|_2\\
&+\left\|\frac{S_o}{T}\sum_{t=1}^{T}[1]_t([2]_t+[3]_t)^\top\right\|_2+\left\|\frac{S_o}{T}\sum_{t=1}^{T}([2]_t+[3]_t)[1]_t^\top\right\|_2\\
&+\left\|\frac{S_o}{T}\sum_{t=1}^{T}([2]_t+[3]_t)([2]_t+[3]_t)^\top\right\|_2\\
&\text{because for two vectors } a,b \text{ the operator norm } \|ab^\top\|_2\le\|a\|_2\|b\|_2\\
\lesssim&\left\|H^{-1}GH^{-1}-\frac{S_o}{T}\sum_{t=1}^{T}[1]_t[1]_t^\top\right\|_2\\
&+\frac{1}{T}\sum_{t=1}^{T}\|[1]_t\|_2(\|[2]_t\|_2+\|[3]_t\|_2)\\
&+\frac{1}{T}\sum_{t=1}^{T}(\|[2]_t\|_2^2+\|[3]_t\|_2^2).
\end{aligned}$$

Because $\sum_{t=1}^{T}[1]_t$ consists of $S_o\cdot T$ i.i.d. samples from $\{H^{-1}\nabla f_i(htheta)\}_{i=1}^n$ and the mean $H^{-1}\nabla f(\widehat{\theta})=0$, using matrix concentration (Tropp, 2015), we know that

$$\mathbb{E}\left[\left\|H^{-1}GH^{-1}-\frac{S_o}{T}\sum_{t=1}^{T}[1]_t[1]_t^\top\right\|_2\right]\lesssim\frac{1}{\sqrt{T}}.$$

For term [3], using (41), because we have

$$\begin{aligned}
&\sum_{t=1}^{T}[3]_t\\
=&\underbrace{\sum_{t=1}^{T}\frac{1}{L+1}\sum_{k=0}^{L-1}\tau_k\sum_{j=k+1}^{L}\prod_{l=k+1}^{j-1}(I-\tau_lH_t)(-e_t^k)}_{\text{when } a\neq b\ \mathbb{E}[\langle e_t^a,e_t^b\rangle]=0}
\end{aligned}$$

$$+\sum_{t=1}^{T}\Bigg(\frac{1}{L+1}\sum_{k=0}^{L-1}\Bigg(k\sum_{j=k+1}^{L}\Bigg(\prod_{l=k+1}^{j-1}\Big((I-\tau_l H_t)(-\tfrac{\nabla f(\theta_t+\delta_t^k g_t^k)-\nabla f(\theta_t)}{\delta_t^k}+H_t g_t^k),$$

by using (42) and (44), we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}[\beta]_t\right\|_2^2\right]\\
\lesssim&\mathbb{E}\left[\frac{1}{T}\Big(\sum_{t=1}^{T}\frac{1}{L}+\Big(\sum_{t=1}^{T}\frac{\sum_{k=0}^{L}\delta_t^k}{L}\Big)^2\Big)\left\|\frac{g_t^0}{\rho_t}\right\|_2^2\right]\\
&\text{using (52), and by our choice of } \delta_t^k=\delta_t^0\,o((k+1)^{-2}) \text{ and } \delta_t^0=o((t+1)^{-2}) \text{ (37)}\\
\lesssim&\mathbb{E}\left[\left(\frac{1}{L}+\frac{\sum_{t=1}^{T}\delta_t^{0\,2}}{T}\right)\left(1+\|\theta_t-\widehat{\theta}\|_2^2\right)\right]\\
\lesssim&\frac{1}{L}+\frac{1}{T}.
\end{aligned}\tag{57}$$

And because we have

$$\mathbb{E}[\|[1]_t\|_2]=\mathbb{E}[\|-H^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta})\|_2]=O(1),$$

$$\begin{aligned}
&\mathbb{E}[\|[2]_t\|_2^2\mid\theta_t]\\
\lesssim&\mathbb{E}\left[\left\|(H^{-1}-H_t^{-1})\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta})\right\|_2^2+\left\|H_t^{-1}\frac{1}{S_o}\sum_{i\in I_o}(\nabla f_i(\widehat{\theta})-\nabla f_i(\theta_t))\right\|_2^2\mid\theta_t\right]\\
&\text{because } H^{-1}-H_t^{-1}=H^{-1}(H_t-H)H_t^{-1} \text{ and using Lemma F.1}\qquad\qquad (58)\\
\lesssim&\mathbb{E}[\|\theta_t-\widehat{\theta}\|_2^2\mid\theta_t]\\
\lesssim&(t+1)^{-d_o},\qquad\qquad (59)
\end{aligned}$$

by repeatedly applying Cauchy-Schwarz inequality and AM-GM inequality, we can conclude that

$$\begin{aligned}
&\mathbb{E}\left[\left\|H^{-1}GH^{-1}-\frac{S_o}{T}\sum_{t=1}^{T}\frac{\bar{g}_t\bar{g}_t^\top}{\rho_t^2}\right\|_2\right]\\
\lesssim&\frac{1}{\sqrt{T}}+\frac{1}{T}\sum_{t=1}^{T}(t+1)^{-\frac{d_o}{2}}+\frac{1}{T}\sum_{t=1}^{T}(t+1)^{-d_o}+\frac{1}{\sqrt{L}}+\frac{1}{L}\\
&\text{because } \sum_{t=1}^{T}(t+1)^{-\frac{d_o}{2}}=T^{1-\frac{d_o}{2}} \text{ for } d_o\in(\frac{1}{2},1)
\end{aligned}$$

$$\lesssim \frac{1}{T^{\frac{d_o}{2}}} + \frac{1}{\sqrt{L}}.$$

F.2. Proof of Corollary 2.1

For $\frac{g_t^L}{\rho_t}$, we have

$$\begin{aligned}\frac{g_t^L}{\rho_t} = &\underbrace{-H^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta})}_{[1]} \\ &+ \underbrace{H^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta}) - H_t^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta}) + H_t^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta}) - H_t^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\theta_t) + H_t^{-1}\nabla f(\theta_t)}_{[2]} \\ &\underbrace{-H_t^{-1}\nabla f(\theta_t) + (\theta_t - \widehat{\theta})}_{[3]} \underbrace{-H_t^{-1}\frac{g_t^0}{\rho_t} + \frac{g_t^L}{\rho_t}}_{[4]} - (\theta_t - \widehat{\theta}),\end{aligned} \tag{60}$$

which gives

$$\begin{aligned}&\theta_t - \widehat{\theta} \\ &= (1-\rho_{t-1})(\theta_{t-1} - \widehat{\theta}) + \rho_{t-1}([1]_{t-1} + [2]_{t-1} + [3]_{t-1} + [4]_{t-1}) \\ &= (\prod_{i=0}^{t-1}(1-\rho_i))(\theta_0 - \widehat{\theta}) + \sum_{i=0}^{t-1}(\prod_{j=i+1}^{t-1}(1-\rho_j))\rho_i([1]_i + [2]_i + [3]_i + [4]_i).\end{aligned}$$

And we have

$$\begin{aligned}&\sqrt{T}(\frac{\sum_{t=1}^T \theta_t}{T} - \widehat{\theta}) \\ &= \frac{1}{\sqrt{T}}(\sum_{t=1}^T\prod_{i=0}^{t-1}(1-\rho_i))(\theta_0 - \widehat{\theta}) + \frac{1}{\sqrt{T}}\sum_{t=1}^T\sum_{i=0}^{t-1}(\prod_{j=i+1}^{t-1}(1-\rho_j))\rho_i([1]_i + [2]_i + [3]_i + [4]_i) \\ &= \frac{1}{\sqrt{T}}(\sum_{t=1}^T\prod_{i=0}^{t-1}(1-\rho_i))(\theta_0 - \widehat{\theta}) + \frac{1}{\sqrt{T}}\sum_{i=0}^{T-1}\sum_{t=i+1}^{T}(\prod_{j=i+1}^{t-1}(1-\rho_j))\rho_i([1]_i + [2]_i + [3]_i + [4]_i).\end{aligned} \tag{61}$$

For the first term in (61), which is non-stochastic, we have

$$\left\|\frac{1}{\sqrt{T}}(\sum_{t=1}^T\prod_{i=0}^{t-1}(1-\rho_i))(\theta_0 - \widehat{\theta})\right\|_2 \lesssim \frac{1}{\sqrt{T}}.$$

For the second term in (61), which is stochastic, we first consider $\rho_i \sum_{t=i+1}^{T} \prod_{j=i+1}^{t-1}(1-\rho_j)$, which is $O(1)$ (similar to (42)) and satisfies

$$
\begin{aligned}
&\rho_i \sum_{t=i+1}^{T} \prod_{j=i+1}^{t-1}(1-\rho_j) \\
&= \sum_{t=i+1}^{T} \frac{\rho_i}{\rho_t}\rho_t \prod_{j=i+1}^{t-1}(1-\rho_j) \\
\leq &\frac{\rho_i}{\rho_s} \sum_{t=i+1}^{s} \rho_t \prod_{j=i+1}^{t-1}(1-\rho_j) + \rho_i(\prod_{j=i+1}^{s}(1-\rho_j)) \sum_{t=s+1}^{T} \prod_{j=s+1}^{t-1}(1-\rho_j) \\
&= (1+\frac{\rho_i-\rho_s}{\rho_s})(1-\prod_{t=i+1}^{s}(1-\rho_t)) + \rho_i(\prod_{j=i+1}^{s}(1-\rho_j)) \sum_{t=s+1}^{T} \prod_{j=s+1}^{t-1}(1-\rho_j) \\
\leq &(1+\frac{\rho_i-\rho_s}{\rho_s})(1-(1-\rho_s)^{s-i}) + \rho_i(1-\rho_s)^{s-i} \sum_{t=s+1}^{T} \prod_{j=s+1}^{t-1}(1-\rho_j) \\
\leq &1 + ((1+\frac{s-i}{i+1})^{d_o}-1) + \rho_i e^{-(s-i)\rho_s} \sum_{t=s+1}^{\infty} \prod_{j=s+1}^{t-1}(1-\rho_j) \\
\leq &1 + \frac{s-i}{i} + \rho_i e^{-(s-i)\rho_s} \sum_{t=s+1}^{\infty} \prod_{j=s+1}^{t-1}(1-\rho_j),
\end{aligned}
$$

for all $i \leq s \leq T$, and

$$
\begin{aligned}
&\rho_i \sum_{t=i+1}^{T} \prod_{j=i+1}^{t-1}(1-\rho_j) \\
&\geq \sum_{t=i+1}^{T} (\prod_{j=i+1}^{t-1}(1-\rho_j))\rho_t \\
&= 1 - \prod_{t=i+1}^{T}(1-\rho_t) \\
&\geq 1 - \exp(-\sum_{t=i+1}^{T}(\rho_t) \\
&\geq 1 - \exp(-\frac{1}{1-d_o}((T+2)^{1-d_o} - (i+2)^{1-d_o}))
\end{aligned}
$$

When we choose $s = i + \lceil (i+1)^{\frac{d_o+1}{2}} \rceil$, we have $\frac{s-i}{i} \lesssim i^{\frac{-1+d_o}{2}}$, $(s-i)\rho_s \gtrsim (i+1)^{\frac{1-d_o}{2}}$, and $\rho_i e^{-\frac{1}{2}(s-i)\rho_s} \lesssim \rho_s$. And these imply $|\rho_i \sum_{t=i+1}^{T} \prod_{j=i+1}^{t-1}(1-\rho_j) - 1| = O(\max\{(i+1)^{\frac{-1+d_o}{2}}, \exp(-\frac{1}{1-d_o}((T+2)^{1-d_o} - (i+2)^{1-d_o})\})$. Thus, for term

[1], we have

$$\frac{1}{\sqrt{T}}\sum_{i=0}^{T-1}\Big(\sum_{t=i+1}^{T}\prod_{j=i+1}^{t-1}\Big((1-\rho_j)\rho_i[1]_i = \frac{1}{\sqrt{T}}\sum_{i=0}^{T-1}[1]_i + \frac{1}{\sqrt{T}}\sum_{i=0}^{T-1}\Big(\sum_{t=i+1}^{T}\prod_{j=i+1}^{t-1}\Big((1-\rho_j)\rho_i - 1)[1]_i,$$

where the first term weakly converges to $\mathcal{N}(0, \frac{1}{S_o}H^{-1}GH^{-1})$ by Central Limit Theorem, and the second term satisfies $\mathbb{E}[\|\frac{1}{\sqrt{T}}\sum_{i=0}^{T-1}(\sum_{t=i+1}^{T}\prod_{j=i+1}^{t-1}(1-\rho_j))\rho_i - 1)[1]_i\|_2^2] = \mathbb{E}[\frac{1}{T}\sum_{i=0}^{T-1}|(\sum_{t=i+1}^{T}\prod_{j=i+1}^{t-1}(1-\rho_j))\rho_i - 1)|^2\|[1]_i\|_2^2] \lesssim T^{d_o-1} + \frac{1}{T}$.

For term [2], we have

$$\|[2]_t\|_2 \lesssim \|\theta_t - \widehat{\theta}\|_2,$$

and $\mathbb{E}[\langle[2]_a, [2]_b\rangle] = 0$ when $a \neq b$. Thus

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\sum_{i=0}^{T-1}\Big(\sum_{t=i+1}^{T}\prod_{j=i+1}^{t-1}\Big((1-\rho_j)\rho_i[2]_i\right\|_2^2\right] \lesssim \frac{1}{T}\sum_{i=0}^{T-1}\|\theta_t - \widehat{\theta}\|_2^2 \lesssim T^{-d_o}.$$

For term [3], we have

$$\|-H_t^{-1}\nabla f(\theta_t) + (\theta_t - \widehat{\theta})\|_2 \lesssim \|\theta_t - \widehat{\theta}\|_2^2.$$

By using (7) and Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\sum_{i=0}^{T-1}\Big(\sum_{t=i+1}^{T}\prod_{j=i+1}^{t-1}\Big((1-\rho_j)\rho_i[3]_i\right\|_2^2\right] \lesssim T^{1-2d_o}.$$

For term [4], similar to similar to (57), we have

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\sum_{i=0}^{T-1}\Big(\sum_{t=i+1}^{T}\prod_{j=i+1}^{t-1}\Big((1-\rho_j)\rho_i[4]_i\right\|_2^2\right] \lesssim \frac{1}{T} + \frac{1}{L}.$$

F.3. Proof of Corollary D.1

Using Theorem 6.5 of (Bubeck, 2015), we have

$$\mathbb{E}[\|\theta_t - \widehat{\theta}\|_2^2] \lesssim 0.9^t.$$

Similar to (8) in Theorem 2.1 (Appendix F.1.1), we have

$$\mathbb{E}\left[\left\|\frac{\bar{g}_t}{\rho_t} - [\nabla^2 f(\theta_t)]^{-1}g_t^0\right\|_2^2 \mid \theta_t\right] \lesssim \frac{1}{L}\|g_t^0\|_2^2.$$

Similar to the proof of (10) in Theorem 2.1 (Appendix F.1.5), using (56), we have

$$\mathbb{E}\left[\left\|H^{-1}GH^{-1}-\frac{S_o}{T}\sum_{t=1}^{T}\frac{\bar{g}_t\bar{g}_t^\top}{\rho_t^2}\right\|_2\right]\lesssim L^{-\frac{1}{2}}.$$

For $\frac{g_t^L}{\rho_t}$, we have

$$\begin{aligned}\frac{\bar{g}_t}{\rho_t}=&\underbrace{-H^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta})}_{[1]}\\&\underbrace{+H^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta})-H_t^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta})+H_t^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta})-H_t^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\theta_t)+H_t^{-1}\nabla f(\theta_t)}_{[2]}\\&\underbrace{-H_t^{-1}\nabla f(\theta_t)}_{[3]}\underbrace{-H_t^{-1}\frac{g_t^0}{\rho_t}+\frac{g_t^L}{\rho_t}}_{[4]}.\end{aligned}\tag{62}$$

For term [1], we have

$$\frac{1}{\sqrt{T}}\sum_{i=1}^{T}[1]_t=\frac{1}{\sqrt{T}}\sum_{i=1}^{T}\left(-H^{-1}\frac{1}{S_o}\sum_{i\in I_o}\nabla f_i(\widehat{\theta})\right)_t,$$

which consists of $S_o \cot T$ i.i.d samples from 0 mean set $\{H_{-1}\nabla f_i(\widehat{\theta})\}_{i=1}^n$, and weakly converges to $\mathcal{N}(0, \frac{1}{S_o}H^{-1}GH^{-1})$ by the Central Limit Theorem.

For term [2], similar to (58), we have

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\sum_{i=1}^{T}[2]_t\right\|_2^2\right]=\frac{1}{T}\mathbb{E}[\sum_{i=1}^{T}\|[2]_t\|_2^2]\lesssim\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\theta_t-\widehat{\theta}\|_2^2]\lesssim\frac{1}{T}.$$

For term [3], we have

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\sum_{i=1}^{T}[3]_t\right\|_2\right]\lesssim\frac{1}{\sqrt{T}}\mathbb{E}[\|\theta_t-\widehat{\theta}\|_2]\lesssim\frac{1}{\sqrt{T}}.$$

For term [4], similar to (57), we have

$$\mathbb{E}\left[\left\|\frac{1}{\sqrt{T}}\sum_{i=1}^{T}[4]_t\right\|_2\right]\lesssim\frac{1}{\sqrt{T}}+\frac{1}{\sqrt{L}}.$$

F.4. Proof of Theorem 3.1

The error bound proof is similar to standard LASSO proofs (Bühlmann and van de Geer, 2011; Negahban et al., 2012).

We will use Lemma F.3 for the covariance estimate using soft thresholding.

We denote "soft thresholding by ω" as an element-wise procedure $\mathbf{S}_\omega(A) = \text{sign}(A)(|A| - \omega)_+$, where A is an arbitrary number, vector, or matrix, and ω is non-negative.

Lemma F.3

Under our assumptions in Section 3, we choose soft threshold $\frac{1}{n}\sum_{i=1}^n X_i X_i^\top$ *using*

$$\omega = \Theta\left(\sqrt{\frac{\log p}{n}}\right)$$

When $n \gg \log p$*, the matrix max norm of* $\frac{1}{n}\sum_{i=1}^n x_i x_i^\top - \Sigma$ *is bounded by*

$$\max_{1\le i,j\le p}\left|\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)_{ij} - \Sigma_{ij}\right| \lesssim \sqrt{\frac{\log p}{n}},$$

with probability at least $1 - p^{-\Theta(1)}$.

Under this event, ℓ_2 *operator norm of* $\widehat{S} - \Sigma$ *satisfies*

$$\|\widehat{S} - \Sigma\|_2 \lesssim b\sqrt{\frac{\log p}{n}},$$

ℓ_1 *and* ℓ_∞ *operator norm of* $\widehat{S} - \Sigma$ *satisfies*

$$\|\widehat{S} - \Sigma\|_\infty = \|\widehat{S} - \Sigma\|_1 \lesssim b\sqrt{\frac{\log p}{n}}.$$

Proof

The proof is similar to that of Theorem 1, (Bickel and Levina, 2008).

Our assumption that Σ is well conditioned implies that each off diagonal entry is bounded, and each diagonal entry is $\Theta(1)$ and positive.

Omitting the subscript for the i^{th} sample, for each i.i.d. sample $x = [x(1), x(2), \ldots, x(p)]^\top \sim \mathcal{N}(0, \Sigma)$, each $x(j)x(k)$ satisfies

$$x(j)x(k) = \frac{1}{4}(x(j) + x(k))^2 - \frac{1}{4}(x(j) - x(k))^2,$$

where $x(j) \pm x(k)$ are Gaussian random variables with variance $\Sigma_{jj} \pm 2\Sigma_{jk} + \Sigma_{kk} = \Theta(1)$, because all of Σ's eigenvalues are upper and lower bounded. Thus, $x(j) \pm x(k)$ are χ_1^2 random variables scaled by $\Sigma_{jj} \pm 2\Sigma_{jk} + \Sigma_{kk} = \Theta(1)$, and they are

sub-exponential with parameters that are $\Theta(1)$ (Wainwright, 2017). And this implies that, $x(j)x(k) - \Sigma_{jk}$ is sub-exponential

$$\mathbb{P}[|x(j)x(k) - \Sigma_{jk}| > t] \lesssim \exp(-\Theta(\min\{t^2, t\})),$$

for all $1 \le j, k \le p$.

Using Bernstein inequality (Wainwright, 2017), we have

$$\mathbb{P}\left[\left|\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)_{jk} - \Sigma_{jk}\right| > t\right] \lesssim \exp(-n\Theta(\min\{t^2, t\})),$$

for all $1 \le j, k \le p$.

Taking a union bound over all matrix entries, and using $n \gg \log p$, we have

$$\max_{1\le j,k\le p}\left|\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)_{jk} - \Sigma_{jk}\right| \lesssim \sqrt{\frac{\log p + \log\frac{1}{\delta}}{n}},$$

with probability at least $1-\delta$.

Under this event, the soft thresholding estimate $\mathbf{S}_\omega(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top)_{ij}$ with $\omega = \Theta(\sqrt{\frac{\log p}{n}})$ is 0 when $\Sigma_{ij} = 0$, and $|\Sigma_{ij} - \mathbf{S}_\omega(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top)_{ij}| \le \omega$ (even when $|\Sigma_{ij}| \le \omega$). And this implies our bounds. ■

Lemma F.3 guarantees that the optimization problem (13) is well defined with high probability when $n \gg b\sqrt{\frac{\log p}{n}}$. Because the ℓ_2 operator norm $\|\widehat{S} - \Sigma\|_2 \lesssim b\sqrt{\frac{\log p}{n}} \ll 1$, and the positive definite matrix Σ's eigenvalues are all $\Theta(1)$, the symmetric matrix $\widehat{S}$ is positive definite, and $\widehat{S}$'s eigenvalues are all $\Theta(1)$, and for all $v \in \mathbb{R}^p$ we have

$$0 \le v^\top \widehat{S} v = \Theta(\|v\|_2^2). \tag{63}$$

Because $\widehat{\theta}$ attains the minimum, by definition, we have

$$\begin{aligned}
&\tfrac{1}{2}\widehat{\theta}^\top \left(\widehat{S} - \tfrac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)\widehat{\theta} + \tfrac{1}{n}\sum_{i=1}^{n}\tfrac{1}{2}(x_i^\top\widehat{\theta} - y_i)^2 + \lambda\|\widehat{\theta}\|_1 \\
\le &\tfrac{1}{2}\theta^{\star\top}\left(\widehat{S} - \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)\theta^\star + \tfrac{1}{n}\sum_{i=1}^{n}\tfrac{1}{2}\left(x_i^\top\theta^\star - y_i\right)^2 + \lambda\|\theta^\star\|_1,
\end{aligned}$$

which, after rearranging terms, is equivalent to

$$\tfrac{1}{2}(\widehat{\theta} - \theta^\star)^\top\widehat{S}(\widehat{\theta} - \theta^\star) + \left\langle\left(\widehat{S} - \tfrac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)\theta^\star + \tfrac{1}{n}\sum_{i=1} y_i x_i, \widehat{\theta} - \theta^\star\right\rangle \le \lambda(\|\theta^\star\|_1 - \|\widehat{\theta}\|_1). \tag{64}$$

Because $\widehat{S} = \mathbf{S}_\omega(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top)$ soft thresholds each entry of $\frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ with $\omega = \Theta(\sqrt{\frac{\log p}{n}})$, each entry of $\widehat{S} - \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ will lie in the interval $[-\omega, \omega]$. And this implies , with probability at least $1 - p^{-\Theta(1)}$, we have

$$\left\| \widehat{S} - \frac{1}{n}\sum_{i=1}^n x_i x_i^\top \right) \theta^\star \right\|_\infty \lesssim \sqrt{\frac{\log p}{n}} \|\theta^\star\|_1 \lesssim \sqrt{\frac{s \log p}{n}},$$

where we used the assumption that $\theta^\star$ is s sparse and $\|\theta^\star\|_2 = O(1)$, which implies $\|\theta^\star\|_1 \lesssim \sqrt{s}$.

For the j^{th} coordinate of $\epsilon_i x_i$, because ϵ_i and x_i are independent Gaussian random variables, we know that it is sub-exponential (Wainwright, 2017)

$$\mathbb{P}[|\epsilon_i x_i(j)| > t] \lesssim \exp\left(-\Theta\left(\min\left\{\frac{t^2}{\sigma^2}, \frac{t}{\sigma}\right\}\right)\right) \tag{65}$$

for all $1 \le i \le n$ and $1 \le j \le p$.

Using Bernstein inequality, we have

$$\mathbb{P}[|\frac{1}{n}\sum_{i=1}^n \epsilon_i x_i(j)| > t] \lesssim \exp\left(-\Theta\left(n \min\left\{\frac{t^2}{\sigma^2}, \frac{t}{\sigma}\right\}\right)\right)$$

for all $1 \le j \le p$.

Taking a union bound over all p coordinates, with probability at least $1 - p^{-\Theta(1)}$, we have

$$\|\frac{1}{n}\sum_{i=1}^n \epsilon_i x_i\|_\infty \lesssim \sigma\sqrt{\frac{\log p}{n}}, \tag{66}$$

when $n \gg \log p$.

Thus, we set the regularization parameter

$$\begin{aligned} \lambda =& \Theta\left((\sigma + \|\theta^\star\|_1)\sqrt{\frac{\log p}{n}}\right) \\ \ge& 2\left\| \widehat{S} - \frac{1}{n}\sum_{i=1}^n x_i x_i^\top \right) \theta^\star + \frac{1}{n}\sum_{i=1}^n \epsilon_i x_i \right\|_\infty, \end{aligned} \tag{67}$$

which holds under the events in Lemma F.3 and (66).

For a vector $v \in \mathbb{R}^p$, let v^S indicate the sub-vector of on the support of $\theta^\star$, and $v^{\bar{S}}$ the sub-vector not on the support of $\theta^\star$. (64) and (67) implies that

$$-\tfrac{1}{2}\lambda(\|(\theta - \theta^\star)^S\|_1 + \|\theta^{\bar{S}}\|_1) = -\tfrac{1}{2}\lambda\|\theta - \theta^\star\|_1 \le \lambda(\|\theta^\star\|_1 - \|\widehat{\theta}\|_1) \le \lambda(\|(\theta - \theta^\star)^S\|_1 - \|\theta^{\bar{S}}\|_1),$$

which is equivalent to

$$\|\theta^{\bar{S}}\|_1 \leq 3\|(\theta - \theta^\star)^S\|_1, \tag{68}$$

because $\lambda > 0$.

For any vector $v \in \mathbb{R}^p$, it satisfies $\|v\|_2^2 \geq \|v^S\|_2^2 \geq \frac{1}{s}\|v^S\|_1^2$. Using this in (64), we have

$$\tfrac{1}{s}\|(\theta - \theta^\star)^S\|_1^2 \lesssim \lambda\|(\theta - \theta^\star)^S\|_1,$$

which implies that

$$\|(\theta - \theta^\star)^S\|_1 \lesssim s(\sigma + \|\theta^\star\|_1)\sqrt{\frac{\log p}{n}}. \tag{69}$$

Combining (69) and (68), we have proven (14)

$$\|\theta - \theta^\star\|_1 \lesssim s\left(\sigma + \|\theta^\star\|_1\right)\sqrt{\frac{\log p}{n}} \lesssim s\left(\sigma + \sqrt{s}\right)\sqrt{\frac{\log p}{n}}.$$

In (64) because $\langle(\widehat{S} - \frac{1}{n}\sum_{i=1}^n x_i x_i^\top)\theta^\star + \frac{1}{n}\sum_{i=1}^n \epsilon_i x_i, \widehat{\theta} - \theta^\star\rangle \geq 0$ by convexity, and using (63), we have proven (15)

$$\|\theta - \theta^\star\|_2^2 \lesssim \lambda\|(\theta - \theta^\star)^S\|_1 \lesssim s\left(\sigma + \|\theta^\star\|_1\right)^2 \frac{\log p}{n} \lesssim s\left(\sigma + \sqrt{s}\right)^2 \frac{\log p}{n}.$$

F.5. Proof of Theorem 3.2

At the solution $\widehat{\theta}$ of the optimization problem (13), using the KKT condition, we have

$$\left(\widehat{S} - \frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)\widehat{\theta} + \frac{1}{n}\sum_{i=1}^n x_i(x_i^\top\widehat{\theta} - y_i) + \lambda\widehat{g} = 0, \tag{70}$$

where $\widehat{g} \in \partial\|\widehat{\theta}\|_1$. And this is equivalent to

$$\widehat{S}\widehat{\theta} - \frac{1}{n}\sum_{i=1}^n x_i(x_i^\top\theta^\star + \epsilon_i) + \lambda\widehat{g} = 0,. \tag{71}$$

By Lemma F.3, we know that $\widehat{S}$ is invertible when $n \gg b^2 \log p$.

Plugging (16) into (71), we have

$$\widehat{S}(\widehat{\theta}^{\mathrm{d}} - \widehat{S}^{-1}\left[\frac{1}{n}\sum_{i=1}^n y_i x_i - \left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)\widehat{\theta}\right] - \frac{1}{n}\sum_{i=1}^n x_i(x_i^\top\theta^\star + \epsilon_i) + \lambda\widehat{g} = 0,$$

which is equivalent to

$$\widehat{S}(\widehat{\theta}^{\mathrm{d}}-\theta^\star)-\frac{1}{n}\sum_{i=1}^n\left(\epsilon_i x_i+\quad\frac{1}{n}\sum_{i=1}^n\left(x_ix_i^\top-\widehat{S}\right)\right)\left(\widehat{\theta}-\theta^\star\right)=0, \tag{72}$$

where we used the fact that $\lambda\widehat{g}=-\widehat{S}\widehat{\theta}+\frac{1}{n}\sum_{i=1}^n x_i(x_i^\top\theta^\star+\epsilon_i)$.

Rewriting (72), we have

$$\widehat{\theta}^{\mathrm{d}}-\theta^\star=\widehat{S}^{-1}\frac{1}{n}\sum_{i=1}^n\left(\epsilon_i x_i+\quad I-\widehat{S}^{-1}\quad\frac{1}{n}\sum_{i=1}^n\left(x_ix_i^\top\right)\right)\left(\widehat{\theta}-\theta^\star\right). \tag{73}$$

For $\max_{1\le j,k\le p}\left|\left(I-\widehat{S}^{-1}\left(\frac{1}{n}\sum_{i=1}^n x_ix_i^\top\right)\right)_{jk}\right|$, we have

$$\begin{aligned}
&\max_{1\le j,k\le p}\left|\quad I-\widehat{S}^{-1}\quad\frac{1}{n}\sum_{i=1}^n x_ix_i^\top\right)\right)_{jk}\Bigg| \\
=&\max_{1\le j,k\le p}\left|\left(\widehat{S}^{-1}\quad S-\frac{1}{n}\sum_{i=1}^n\left(x_ix_i^\top\right)\right)\right)_{jk}\Bigg| \\
\le&\|\widehat{S}^{-1}\|_\infty\max_{1\le j,k\le p}\left|\quad S-\frac{1}{n}\sum_{i=1}^n\left(x_ix_i^\top\right)\right)_{jk}\Bigg|.
\end{aligned} \tag{74}$$

Under the event in Lemma F.3, we have

$$\max_{1\le j,k\le p}\left|\quad S-\frac{1}{n}\sum_{i=1}^n\left(x_ix_i^\top\right)\right)_{jk}\Bigg|\lesssim\sqrt{\frac{\log p}{n}}. \tag{75}$$

Also under the event in Lemma F.3, we have

$$\widehat{S}_{ii}-\sum_{j\ne i}|\widehat{S}_{ij}|\ge\Sigma_{ii}-\Theta\left(\sqrt{\tfrac{\log p}{n}}\right)\quad-\sum_{j\ne i}|\Sigma_{ij}|\ge D_\Sigma-\Theta\left(\sqrt{\tfrac{\log p}{n}}\right)$$

where we used $\widehat{S}_{ii}>0$ and $|\Sigma_{ij}|\ge|\widehat{S}_{ij}|$ by definition of the soft thresholding operation.

Thus, when $n\gg\frac{1}{{D_\Sigma}^2}\log p$, we have

$$\widehat{S}_{ii}-\sum_{j\ne i}|\widehat{S}_{ij}|\gtrsim D_\Sigma,$$

which implies that $\widehat{S}$ is also diagonally dominant. Thus, using Theorem 1, (Varah, 1975), when $n \gg \frac{1}{{D_\Sigma}^2} \log p$, we have

$$\|\widehat{S}\|_\infty \lesssim \frac{1}{D_\Sigma}, \tag{76}$$

with probability at least $1 - p^{-\Theta(1)}$

And using (75) and (76) in (74), we have

$$\max_{1\le j,k\le p} \left| I - \widehat{S}^{-1}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)\right|_{jk} \lesssim \frac{1}{D_\Sigma}\sqrt{\frac{\log p}{n}}. \tag{77}$$

Using (77) and the bound on $\|\widehat{\theta} - \theta^\star\|_1$ (14), in (73), we have

$$\left\| I - \widehat{S}^{-1}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)\right)\left(\widehat{\theta} - \theta^\star\right)\right\|_\infty \lesssim \frac{1}{D_\Sigma} s\left(\sigma + \|\theta^\star\|_1\right)\frac{\log p}{n} \lesssim \frac{1}{D_\Sigma} s\left(\sigma + \sqrt{s}\right)\frac{\log p}{n}. \tag{78}$$

Combining (78) and (73), we have proven Theorem 3.2, when $n \gg \max\{b^2, \frac{1}{{D_\Sigma}^2}\}\log p$, we have

$$\sqrt{n}(\widehat{\theta}^{\mathrm{d}} - \theta^\star) = Z + R,$$

where $Z \mid \{x_i\}_{i=1}^n \sim \mathcal{N}\left(0, \sigma^2 \widehat{S}^{-1}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)\widehat{S}^{-1}\right)$, and $\|R\|_\infty \lesssim \frac{1}{D_\Sigma} s\left(\sigma + \|\theta^\star\|_1\right)\frac{\log p}{\sqrt{n}} \lesssim \frac{1}{D_\Sigma} s\left(\sigma + \sqrt{s}\right)\frac{\log p}{\sqrt{n}}$ with probability at least $1 - p^{-\Theta(1)}$.

F.6. Proof of Theorem A.1

We analyze the optimization problem conditioned on the data set $\{x_i\}_{i=1}^n$, which satisfies Lemma F.3 with probability at least $1 - p^{\Theta(-1)}$ when $n \gg b^2 \log p$.

Here, we denote the objective function as

$$P(\theta) = \frac{1}{2}\theta^\top \left(\widehat{S} - \frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)\theta + \frac{1}{n}\sum_{i=1}^n \frac{1}{2}\left(x_i^\top \theta - y_i\right)^2 + \lambda\|\theta\|_1.$$

In Algorithm 2, lines 6 to 15 are using SVRG (Johnson and Zhang, 2013) to solve the Newton step

$$\min_\Delta \frac{1}{2}\Delta^\top \widehat{S}\Delta + \left\langle \frac{1}{S_o}\sum_{k\in I_o}\nabla f_k(\theta_t), \Delta\right\rangle \tag{79}$$

and using proximal SVRG (Xiao and Zhang, 2014) to solve the proximal Newton step

$$\min_\Delta \frac{1}{2}\Delta^\top \widehat{S}\Delta + \left\langle \frac{1}{n}\sum_{k=1}^n \nabla f_k(\theta_t), \Delta\right\rangle + \lambda\|\theta + \Delta\|_1. \tag{80}$$

The gradient of (79) is

$$\widehat{S}\Delta + \frac{1}{S_o}\sum_{k\in I_o}\nabla f_k(\theta_t) = \underbrace{\frac{1}{p}\sum_{k=1}^{p}\left[p\widehat{S}_k\right]\Delta(k)}_{\text{sample } \textit{by feature} \text{ in SVRG}} + \underbrace{\frac{1}{S_o}\sum_{k\in I_o}\nabla f_k(\theta_t)}_{\text{compute exactly in SVRG}},$$

where $\widehat{S}_k$ is the k^{th} column of $\widehat{S}$ and $\Delta(k)$ is the k^{th} coordinate of Δ.

Line 7 corresponds to SVRG's outer loop part that computes the full gradient. Line 12 corresponds to SVRG's inner loop update.

By Lemma F.3, when $n \gg b^2 \log p$, the ℓ_2 operator norm of $\|\widehat{S}\|_2 = O(1)$. And this implies $\|\widehat{S}^\top\widehat{S}\|_2 = O(1)$. Because $\|\widehat{S}_k\|_2^2$ is the k^{th} diagonal element of $\widehat{S}^\top\widehat{S}$, we have $\|\widehat{S}_k\|_2^2 = O(1)$ for all $1 \le k \le p$. Thus, each $\left[p\widehat{S}_k\right]\Delta(k)$ is a $O(p)$-Lipschitz function.

By Theorem 6.5 of (Bubeck, 2015), when conditioned on θ_t, and choosing

$$\tau = \Theta\left(\frac{1}{p}\right)$$
$$L_i \gtrsim p,$$

after L_o^t SVRG outer steps, we have

$$\mathbb{E}\left[\left\|\bar{g}_t + \widehat{S}^{-1}\left(\frac{1}{S_o}\sum_{k\in I_o}\nabla f_k(\theta_t)\right)\right\|_2^2 \middle|\, \theta_t, \{x_i\}_{i=1}^n\right] \lesssim 0.9^{L_o^t}\left\|\frac{1}{S_o}\sum_{k\in I_o}\nabla f_k(\theta_t)\right\|_2^2$$
$$\lesssim 0.9^{L_o^t}(1 + \|\theta_t - \widehat{\theta}\|_2),$$

where $\bar{g}_t = \frac{1}{L_o^t}\sum_{j=0}^{L_o^t} g_t^j$.

The gradient of the smooth component $\frac{1}{2}\Delta^\top\widehat{S}\Delta + \frac{1}{n}\sum_{k=1}^n\langle\nabla f_k(\theta_t), \Delta\rangle$ in (80) is

$$\widehat{S}\Delta + \frac{1}{n}\sum_{k=1}^{n}\nabla f_k(\theta_t) = \underbrace{\frac{1}{p}\sum_{k=1}^{p}\left[p\widehat{S}_k\right]\Delta(k)}_{\text{sample } \textit{by feature} \text{ in proximal SVRG}} + \underbrace{\frac{1}{n}\sum_{k=1}^{n}\nabla f_k(\theta_t)}_{\text{compute exactly in proximal SVRG}}.$$

Line 8 corresponds to proximal SVRG's outer loop part that computes the full gradient. Line 13 corresponds to proximal SVRG's inner loop update.

By Theorem 3.1 of (Xiao and Zhang, 2014), when conditioned on θ_t, and choosing

$$\eta = \Theta\left(\frac{1}{p}\right)$$
$$L_i \gtrsim p,$$

after L_o^t proximal SVRG outer steps, we have

$$\begin{aligned}\mathbb{E}[P(\theta_{t+1} - P(\widehat{\theta})) \mid \theta_t] =& \mathbb{E}\left[P(\theta_t + \bar{d}_t - \widehat{\theta}) - P(\widehat{\theta}) \,\middle|\, \theta_t, \{x_i\}_{i=1}^n\right] \\ \lesssim& 0.9^{L_o^t}(P(\theta_t) - P(\widehat{\theta})),\end{aligned}$$

where $\bar{d}_t = \frac{1}{L_o^t}\sum_{j=0}^{L_o^t} d_t^j$. And this implies

$$\mathbb{E}[\|\theta_t - \widehat{\theta}\|_2^2] \lesssim 0.9^{\sum_{i=0}^{t-1} L_o^t}(P(\theta_0) - P(\widehat{\theta})).$$

At each θ_t, we have

$$x_i(x_i^\top \theta_t - y_i) = x_i x_i^\top(\theta_t - \widehat{\theta}) + x_i(x_i^\top \widehat{\theta} - y_i).$$

For the first term, we have

$$\begin{aligned}\|x_i x_i^\top(\theta_t - \widehat{\theta})\|_\infty \leq& |x_i^\top(\theta_t - \widehat{\theta})|\|x_i\|_\infty \\ \leq& \|x_i\|_2\|\theta_t - \widehat{\theta}\|_2\|x_i\|_\infty \\ \leq& \sqrt{p}\|x_i\|_\infty^2\|\theta_t - \widehat{\theta}\|_2,\end{aligned}$$

which implies that

$$\max_{1\leq j,k\leq p}\left|\left[\left(x_i x_i^\top(\theta_t - \widehat{\theta})\right)\left(x_i x_i^\top(\theta_t - \widehat{\theta})\right)^\top\right]_{jk}\right| \leq \|x_i x_i^\top(\theta_t - \widehat{\theta})\|_\infty^2 \leq p\|x_i\|_\infty^4\|\theta_t - \widehat{\theta}\|_2^2.$$

For the second term, we have

$$\begin{aligned}\|x_i(x_i^\top \widehat{\theta} - y_i)\|_\infty \leq& \|x_i x_i^\top(\widehat{\theta} - \theta^\star)\|_\infty + \|x_i\epsilon_i\|_\infty \\ \leq& \|x_i\|_\infty^2\|\widehat{\theta} - \theta^\star\|_1 + |\epsilon_i|\|x_i\|_\infty\end{aligned}$$

Because when $n \gg \log p$, from (83) we have with probability at least $1 - p^{-\Theta(1)}$

$$\max_{1\leq i\leq n}\|x_i\|_\infty \lesssim \sqrt{\log p + \log n},$$

and from (85) we have with probability at least $1 - n^{-\Theta(1)}$

$$\max_{1\leq i\leq n}|\epsilon_i| \lesssim \sigma\sqrt{\log n},$$

when conditioned on θ_t (and the data set $\{x_i\}_{i=1}^n$) we have

$$\begin{aligned}
&\max_{1\le j,k\le p}\left|\left[\left(\widehat{S}^{-1}g_t^0\right)\left(\widehat{S}^{-1}g_t^0\right)^\top-\left(\widehat{S}^{-1}\tfrac{1}{S_o}\textstyle\sum_{k\in I_o}\nabla f_k(\theta_t)\right)\left(\widehat{S}^{-1}\tfrac{1}{S_o}\sum_{k\in I_o}\nabla f_k(\theta_t)\right)^\top\right]_{jk}\right|\\
\lesssim&\frac{1}{{D_\Sigma}^2}(\|x_ix_i^\top(\theta_t-\widehat{\theta})\|_\infty^2+2\|x_ix_i^\top(\theta_t-\widehat{\theta})\|_\infty\|x_i(x_i^\top\widehat{\theta}-y_i)\|_\infty)\\
\lesssim&\frac{1}{{D_\Sigma}^2}(p(\log p+\log n)^2\|\theta_t-\widehat{\theta}\|_2^2+\sqrt{p}(\log p+\log n)\|\theta_t-\widehat{\theta}\|_2((\log p+\log n)\|\widehat{\theta}-\theta^\star\|_1+\sigma\sqrt{(\log p+\log n)\log n}))\\
\lesssim&\frac{1}{{D_\Sigma}^2}(p\|\theta_t-\widehat{\theta}\|_2^2+\sqrt{p}\|\theta-\widehat{\theta}\|_2(\sigma+\|\widehat{\theta}-\theta^\star\|_1))\mathrm{polylog}(p,n)
\end{aligned}$$

under the events of (83), (76) , and (85), where we used the fact (76) that the ℓ_∞ operator norm $\|\widehat{S}^{-1}\|_\infty\lesssim\frac{1}{D_\Sigma}$ with probability at least $1-p^{-\Theta(1)}$ when $n\gg\max\{b^2,\frac{1}{{D_\Sigma}^2}\}\log p$.

Thus, we can conclude that, conditioned on the data set $\{x_i\}_{i=1}^n$, and the events (83), (85), and (76), we have we have an asymptotic normality result

$$\tfrac{1}{\sqrt{t}}\left(\textstyle\sum_{t=1}^T\sqrt{S_o}\bar{g}_t+\frac{1}{n}\sum_{i=1}^n x_i(x_i^\top\widehat{\theta}-y_i)\right)=W+R,$$

where W weakly converges to $\mathcal{N}\left(0,\widehat{S}^{-1}\left[\frac{1}{n}\sum_{i=1}^n(x_i^\top\widehat{\theta}-y_i)^2x_ix_i^\top-\left(\frac{1}{n}\sum_{i=1}^n x_i(x_i^\top\widehat{\theta}-y_i)\right)\left(\frac{1}{n}\sum_{i=1}^n x_i(x_i^\top\widehat{\theta}-y_i)\right)^\top\right]\widehat{S}^{-1}\right)$, and

$$\begin{aligned}
\|R\|_\infty\le&\frac{1}{\sqrt{t}}\sum_{t=1}^T\left(\|\bar{g}_t-\widehat{S}^{-1}g_t^0\|_\infty+\|\widehat{S}^{-1}g_t^0-\tfrac{1}{S_o}\textstyle\sum_{k\in I_o}\nabla f_k(\widehat{\theta})\|_\infty\right)\\
\le&\frac{1}{\sqrt{t}}\sum_{t=1}^T\left(\|\bar{g}_t-\widehat{S}^{-1}g_t^0\|_2+\|\widehat{S}^{-1}g_t^0-\tfrac{1}{S_o}\textstyle\sum_{k\in I_o}\nabla f_k(\widehat{\theta})\|_\infty\right)
\end{aligned}$$

which implies

$$\begin{aligned}
&\mathbb{E}\left[\|R\|_\infty\mid\{x_i\}_{i=1}^n,(83),(85),(76)\right]\\
\lesssim&\mathbb{E}\left[\frac{1}{\sqrt{t}}\sum_{t=1}^T\left(0.95^{L_o^t}(1+\|\theta_t-\widehat{\theta}\|_2)+\sqrt{p}(\log p+\log n)\|\theta_t-\widehat{\theta}\|_2\mid\{x_i\}_{i=1}^n,(83),(85),(76)\right]\right.\\
\lesssim&\frac{1}{\sqrt{T}}\sum_{t=1}^T\left(0.95^{L_o^t}(1+\sqrt{P(\theta_0)-P(\widehat{\theta})}0.95^{\sum_{i=0}^{t-1}L_o^t})+\sqrt{p}(\log p+\log n)\sqrt{P(\theta_0)-P(\widehat{\theta})}0.95^{\sum_{i=0}^{t-1}L_o^t}.\right.
\end{aligned}$$

And, because $\left(\frac{1}{S_o}\sum_{k\in I_o}\nabla f_k(\widehat{\theta})\right)_t$ are i.i.d., and bounded when conditioned on the data set $\{x_i\}_{i=1}^n$, and the events (83), (85), and (76), using a union bound over all matrix entries, and sub-Gaussian concentration inequalities (Wainwright, 2017) similar to Lemma A.1's proof, when $T\gg\left((\log p+\log n)\|\widehat{\theta}-\theta^\star\|_1+\sigma\sqrt{(\log p+\log n)\log n}\right)\log p$, we also have

$$\left\|\frac{S_o}{T}\textstyle\sum_{t=1}^T\bar{g}_t\bar{g}_t^\top-\widehat{S}^{-1}\left(\frac{1}{n}\sum_{i=1}^n(x_i^\top\widehat{\theta}-y_i)^2x_ix_i^\top\right)\widehat{S}^{-1}\right\|_{\max}$$

$$\lesssim\sqrt{\left(\left((\log p+\log n)\|\widehat{\theta}-\theta^\star\|_1+\sigma\sqrt{(\log p+\log n)\log n}\right)\sqrt{\frac{\log p}{T}}\right.}$$
$$\left.+\frac{1}{u}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}0.95^{L_o^t}(1+\sqrt{P(\theta_0)-P(\widehat{\theta})}0.95^{\sum_{i=0}^{t-1}L_o^t})+\sqrt{p}(\log p+\log n)\sqrt{P(\theta_0)-P(\widehat{\theta})}0.95^{\sum_{i=0}^{t-1}L_o^t}\right]\right)$$

with probability at least $1-p^{-\Theta(-1)}-u$, where we used Markov inequality for the remainder term.

F.7. Proof of Lemma A.1

We analyze the optimization problem conditioned on the data set $\{x_i\}_{i=1}^n$, which satisfies Lemma F.3 with probability at least $1-p^{\Theta(-1)}$ when $n\gg b^2\log p$.

Because we have

$$\begin{aligned}&(x_i^\top\widehat{\theta}-y_i)^2\\=&(x_i^\top(\widehat{\theta}-\theta^\star)-\epsilon_i)^2\\=&\epsilon_i^2-2\epsilon_i x_i^\top(\widehat{\theta}-\theta^\star)+(x_i^\top(\widehat{\theta}-\theta^\star))^2,\end{aligned}$$

we can write

$$\begin{aligned}&\sigma^2\tfrac{1}{n}\textstyle\sum_{i=1}^n x_ix_i^\top-\tfrac{1}{n}\sum_{i=1}^n(x_i^\top\widehat{\theta}-y_i)^2x_ix_i^\top\\=&\tfrac{1}{n}\textstyle\sum_{i=1}^n(\sigma^2-\epsilon_i^2)x_ix_i^\top+\tfrac{1}{n}\sum_{i=1}^n(2\epsilon_ix_i^\top(\widehat{\theta}-\theta^\star)-(x_i^\top(\widehat{\theta}-\theta^\star))^2)x_ix_i^\top.\end{aligned}\tag{81}$$

Conditioned on $\{x_i\}_{i=1}^n$, because $\epsilon_i\sim\mathcal{N}(0,\sigma^2)$ are i.i.d., and ϵ_i^2 is sub-exponential, using Bernstein inequality (Wainwright, 2017), we have

$$\begin{aligned}&\mathbb{P}\left[\left|\tfrac{1}{n}\textstyle\sum_{i=1}^n\left(1-\frac{\epsilon_i^2}{\sigma^2}\right)x_i(j)x_i(k)\right|>t\mid\{x_i\}_{i=1}^n\right]\\\lesssim&\exp\left(-n\min\left\{\frac{t}{\max_{1\le i\le n}|x_i(j)x_i(k)|},\left(\frac{t}{\max_{1\le i\le n}|x_i(j)x_i(k)|}\right)^2\right\}\right)\end{aligned}\tag{82}$$

for $1\le j,k\le p$, where $x_i(j)$ is the j^{th} coordinate of x_i.

Because each $x_i(j)$ is $\mathcal{N}(0,\Theta(1))$ by our assumptions, using a union bound over all samples' coordinates we have

$$\max_{\substack{1\le i\le n\\1\le j\le p}}|x_i(j)|\lesssim\sqrt{\log p+\log n},\tag{83}$$

with probability at least $1-(pn)^{-\Theta(1)}$.

Combining (82) and (83), and taking a union bound over all entries of the matrix $\frac{1}{n}\sum_{i=1}^n(\sigma^2-\epsilon_i^2)x_ix_i^\top$, when $n\gg\log p$, we have

$$\max_{1\le j,k\le p}|(\tfrac{1}{n}\textstyle\sum_{i=1}^n(\sigma^2-\epsilon_i^2)x_ix_i^\top)|_{jk}\lesssim\sigma^2(\log p+\log n)\sqrt{\frac{\log p}{n}},\tag{84}$$

with probability at least $(1-(pn)^{-\Theta(1)})(1-p^{-\Theta(1)}) = 1-(pn)^{-\Theta(1)}-p^{-\Theta(1)}$.

Because $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$, by a union bound, we have

$$\max_{1\le i\le n}|\epsilon_i| \lesssim \sigma\sqrt{\log n}, \tag{85}$$

with probability at least $1-n^{-\Theta(1)}$.

Using (83), we have

$$\begin{aligned}
&\max_{1\le i\le n}|x_i^\top(\widehat{\theta}-\theta^\star)|\\
\le&\|\widehat{\theta}-\theta^\star\|_1 \max_{1\le i\le n}\max_{1\le j\le p}|x_i(j)|\\
\lesssim& s\left(\sigma+\|\theta^\star\|_1\right)\sqrt{\tfrac{\log p}{n}(\log p+\log n)} \lesssim s\left(\sigma+\sqrt{s}\right)\sqrt{\tfrac{\log p}{n}(\log p+\log n)},
\end{aligned} \tag{86}$$

with probability at least $1-p^{-\Theta(1)}-(pn)^{-\Theta(1)}$.

Combining (83), (84), (85), (86), and using (81), when $n \gg \log p$, we have

$$\begin{aligned}
&\max_{1\le j,k\le p}\left|\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n(x_i^\top\widehat{\theta}-y_i)^2x_ix_i^\top-\sigma^2\tfrac{1}{n}\sum_{i=1}^n x_ix_i^\top\right)_{jk}\right|\\
\lesssim&\sigma^2(\log p+\log n)\sqrt{\tfrac{\log p}{n}}+\sigma s\left(\sigma+\|\theta^\star\|_1\right)(\log p+\log n)^{\frac{3}{2}}\sqrt{\tfrac{\log p\cdot\log n}{n}}\\
&+s^2\left(\sigma+\|\theta^\star\|_1\right)^2(\log p+\log n)^2\tfrac{\log p}{n},
\end{aligned} \tag{87}$$

with probability at least $1-p^{-\Theta(1)}-n^{-\Theta(1)}$.

Combining (87) and (76), when $n \gg \max\{b^2, \frac{1}{D_\Sigma{}^2}\}\log p$, we have

$$\begin{aligned}
&\max_{1\le j,k\le p}\left|\left(\widehat{S}^{-1}\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n(x_i^\top\widehat{\theta}-y_i)^2x_ix_i^\top\right)\widehat{S}^{-1}-\sigma^2\widehat{S}^{-1}\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n x_ix_i^\top\right)\widehat{S}^{-1}\right)_{jk}\right|\\
\lesssim&\tfrac{1}{D_\Sigma{}^2}\left(\sigma^2+\sigma s\left(\sigma+\|\theta^\star\|_1\right)\sqrt{\log p+\log n}\sqrt{\log n}+s^2\left(\sigma+\|\theta^\star\|_1\right)^2(\log p+\log n)\sqrt{\tfrac{\log p}{n}}\right)(\log p+\log n)\sqrt{\tfrac{\log p}{n}},
\end{aligned}$$

with probability at least $1-p^{-\Theta(1)}-n^{-\Theta(1)}$.

Appendix G. Technical lemmas

G.1. Lemma G.1

Next lemma is a well known property of convex functions (Lemma 3.11 of (Bubeck, 2015)).

Lemma G.1 *For a α strongly convex and β smooth function $F(x)$, we have*

$$\langle\nabla F(x_1)-\nabla F(x_2),x_1-x_2\rangle \ge \frac{\alpha\beta}{\alpha+\beta}\|x_1-x_2\|_2^2+\frac{1}{\beta+\alpha}\|\nabla F(x_1)-\nabla F(x_2)\|_2^2$$

$$\geq \frac{1}{2}\alpha\|x_1 - x_2\|_2^2 + \frac{1}{2\beta}\|\nabla F(x_1) - \nabla F(x_2)\|_2^2.$$

G.2. Lemma G.2

Next lemma provides a bound on a geometric-like sequence.

Lemma G.2

Suppose we have a sequence

$$a_{t+1} = (1 - \kappa t^{-d})a_t + Ct^{-pd},$$

where $a_1 \geq 0$, $0 < \kappa < 1$, $p \geq 2$ *and* $d \in (\frac{1}{2}, 1)$ *is the decaying rate.*

Then, $\forall 1 \leq s \leq t$ *we have*

$$a_t \leq C\frac{1}{pd-1}(1 - t^{1-pd})\exp\left(-\kappa\frac{1}{1-d}\left((t+1)^{1-d} - (s+1)^{1-d}\right)\right) + a_1 s^{-(p-1)d}\frac{1}{\kappa}.$$

When we assume that a_1, C, κ, p, d *are all constants, we have*

$$a_t = O(t^{-(p-1)d}).$$

Proof

Unrolling the recursion, we have

$$a_t = C\underbrace{\sum_{i=1}^{t-1}\left(\prod_{j=i+1}^{t-1}(1 - \kappa j^{-d})\right)i^{-pd}}_{[1]} + a_1\underbrace{\prod_{i=1}^{t-1}(1 - \kappa i^{-d})}_{[2]}.$$

Splitting term [1] into two parts, we have

$$\sum_{i=1}^{t-1}\left(\prod_{j=i+1}^{t-1}(1 - \kappa j^{-d})\right)i^{-pd}$$
$$= \sum_{i=1}^{s-1}\left(\prod_{j=i+1}^{t-1}(1 - \kappa j^{-d})\right)i^{-pd} + \sum_{i=s}^{t-1}\left(\prod_{j=i+1}^{t-1}(1 - \kappa j^{-d})\right)i^{-pd}.$$

For the first part, we have

$$\sum_{i=1}^{s-1}\left(\prod_{j=i+1}^{t-1}(1 - \kappa j^{-d})\right)i^{-pd}$$

$$\leq \left(\prod_{j=s}^{t-1}(1-\kappa j^{-d})\right)\left(\sum_{i=1}^{s-1} i^{-pd}\right)$$
$$\leq \frac{1}{pd-1}(1-t^{1-pd})\exp\left(-\kappa\frac{1}{1-d}((t+1)^{1-d}-(s+1)^{1-d})\right)($$

where we used

$$\sum_{i=r}^{s} i^{-pd}$$
$$\leq \int_{r}^{s+1} u^{-pd}\,du$$
$$\leq \frac{1}{pd-1}(r^{1-pd}-(s+1)^{1-pd}).$$

For term [2], notice that for $1 \leq r \leq s$, using $1-x \leq \exp(-x)$ when $x \in [0,1]$, we have

$$\prod_{i=r}^{s}(1-\kappa i^{-d}) \leq \exp(-\kappa\textstyle\sum_{i=r}^{s} i^{-d}),$$

and using the fact that

$$\sum_{i=r}^{s} i^{-d} \geq \int_{r}^{s+1}(u+1)^{-d}\,du$$
$$= \frac{1}{1-d}\left((s+2)^{1-d}-(r+1)^{1-d}\right)($$

we have

$$\prod_{i=1}^{t-1}(1-\kappa i^{-d}) \leq \exp\left(-\kappa\tfrac{1}{1-d}(t^{1-d}-2^{1-d})\right)($$

For the second part, we have

$$\sum_{i=s}^{t-1}\left(\prod_{j=i+1}^{t-1}(1-\kappa j^{-d})\right)($$
$$\leq s^{-(p-1)d}\sum_{i=s}^{t-1}\left(\prod_{j=i+1}^{t-1}(1-\kappa j^{-d})\right)($$
$$= s^{-(p-1)d}\frac{1}{\kappa}\sum_{i=s}^{t-1}\left(\prod_{j=i+1}^{t-1}(1-\kappa j^{-d})\right)(\kappa i^{-d}$$

$$
\begin{aligned}
&= s^{-(p-1)d}\frac{1}{\kappa}\quad 1-\prod_{i=s}^{t-1}\Big(1-\kappa i^{-d}\Big)\Bigg)\Bigg(\\
&\le s^{-(p-1)d}\frac{1}{\kappa},
\end{aligned}
$$

where we used the fact that

$$
\begin{aligned}
&\sum_{i=s}^{t-1}\kappa i^{-d}\prod_{j=i+1}^{t-1}\Big(1-\kappa j^{-d}\Big)\\
=&1-\prod_{i=s}^{t-1}\Big(1-\kappa j^{-d}\Big)\\
<&1.
\end{aligned}
$$

When we assume that a_1, C, κ, p, d are all constants, setting $s = \lfloor \frac{n}{2} \rfloor$, we have

$$a_t = O(t^{-(p-1)d}).$$

■

Appendix H. Experiments

H.1. Synthetic data

H.1.1. LOW DIMENSIONAL PROBLEMS

Here, we provide the exact configurations for linear/logistic regression examples provided in Table 1 and Table 2.

Linear regression. We consider the model $y = \langle [1, \cdots, 1]^\top/\sqrt{10}, x\rangle + \epsilon$, where $x \sim \mathcal{N}(0, \Sigma) \in \mathbb{R}^{10}$ and $\epsilon \sim \mathcal{N}(0, 0.7^2)$, with 100 i.i.d. data points.

Lin1: We used $\Sigma = I$. For Algorithm 1, we set $T = 100$, $d_o = d_i = 2/3$, $\rho_0 = 0.1$, $L = 200$, $\tau_0 = 20$, $S_o = S_i = 10$. In bootstrap we used 100 replicates. For averaged SGD, we used 100 averages each of length 50, with step size $0.7 \cdot (t+1)^{-2/3}$ and batch size 10.

Lin2: We used $\Sigma_{jk} = 0.4^{|j-k|}$. For Algorithm 1, we set $T = 100$, $d_o = d_i = 2/3$, $\rho_0 = 0.7$, $L = 100$, $\tau_0 = 1$, $S_o = S_i = 10$. In bootstrap we used 100 replicates. For averaged SGD, we used 100 averages each of length 50, with step size $(t+1)^{-2/3}$ and batch size 10.

Logistic regression. Although logistic regression does not satisfy strong convexity, experimentally Algorithm 1 still gives valid confidence intervals ((Gadat and Panloup, 2017) recently has shown that SGD in logistic regression behaves similar to strongly convex problems). We consider the model $\mathbb{P}[y=1] = \mathbb{P}[y=0] = 1/2$ and $x \mid y \sim \mathcal{N}(0.1/\sqrt{10} \cdot [1, \cdots, 1]^\top, \Sigma) \in \mathbb{R}^{10}$, with 100 i.i.d. data points. Because in bootstrap resampling the Hessian is singular for some replicates, we use jackknife and solve each replicate using Newton's method, which approximately needs 25 steps per replicate.

Approximate Newton	Bootstrap	Inverse Fisher information
(0.951, 0.224)	(0.946 0.205)	(0.966, 0.212)

Table 3: Average 95% confidence interval (coverage, length) after calibration

Log1: We used $\Sigma = I$. For Algorithm 1, we set $T = 50$, $d_o = d_i = 2/3$, $\rho_0 = 0.1$, $L = 100$, $\tau_0 = 2$, $S_o = S_i = 10$, $\delta_0 = 0.01$. For averaged SGD, we used 50 averages each of length 100, with step size $2 \cdot (t+1)^{-2/3}$ and batch size 10.

Log2: We used $\Sigma_{jk} = 0.4^{|j-k|}$. For Algorithm 1, we set $T = 50$, $d_o = d_i = 2/3$, $\rho_0 = 0.1$, $L = 100$, $\tau_0 = 5$, $S_o = S_i = 10$, $\delta_0 = 0.01$ For averaged SGD, we used 50 averages each of length 100, with step size $5 \cdot (t+1)^{-2/3}$ and batch size 10.

Calibration. Here, we give empirical results on calibrating confidence intervals ((Efron and Tibshirani, 1994), Ch.18; (Politis et al., 2012), Ch. 9) produced by our approximate Newton procedure. We consider the model $y = \langle [1, \cdots, 1]^\top/\sqrt{20}, x\rangle + \epsilon$, where $x \sim \mathcal{N}(0, \Sigma) \in \mathbb{R}^{20}$ and $\epsilon \sim \mathcal{N}(0, 0.7^2)$, with 200 i.i.d. data points. We ran 100 simulations. In each simulation, we bootstrapped the dataset 100 times, and computed confidence intervals on each bootstrap replicate using our approximate Newton procedure, bootstrap, and inverse Fisher information. For each method, we then used grid search to find a multiplier such that the empirical point estimate is covered by the bootstrap confidence intervals 95% of the time. Average 95% confidence interval coverage and length after calibration are given in Table 3.

H.1.2. HIGH DIMENSIONAL LINEAR REGRESSION

For comparison with de-biased LASSO (Javanmard and Montanari, 2015; van de Geer et al., 2014), we use the de-biased LASSO estimator with known covariance ("oracle" de-biased LASSO estimator)

$$\widehat{\theta}^{\mathrm{d}}_{\mathrm{oracle}} = \widehat{\theta}_{\mathrm{LASSO}} + \tfrac{1}{n} \cdot \Sigma^{-1} \quad \tfrac{1}{n}\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \Big(x_i x_i^\top \widehat{\theta}_{\mathrm{LASSO}} \Big) \Big($$

and its corresponding statistical error covariance estimate

$$\sigma^2 \cdot \Sigma^{-1} \quad \tfrac{1}{n}\sum_{i=1}^{n} \Big(x_i x_i^\top \Big) \Big(\Sigma^{-1},$$

which assumes that the true inverse covariance Σ^{-1} and observation noise variance σ^2 are known.

Confidence interval visualization. We use 600 i.i.d. samples from a model with $\Sigma = I, \sigma = 0.7, \theta^\star = [1/\sqrt{8}, \cdots, 1/\sqrt{8}, 0, \cdots, 0]^\top \in \mathbb{R}^{1000}$ which is 8-sparse. Figure 3 shows 95% confidence intervals for the first 20 coordinates. The average confidence interval length is 0.14 and average coverage is 0.83. Additional experimental results, including p-value distribution under the null hypothesis, are presented in Appendix H.1.2.

Comparison with de-biased LASSO. We use 600 i.i.d. samples from a model with $\Sigma = I, \sigma = 0.7, \theta^\star = [1/\sqrt{8}, \cdots, 1/\sqrt{8}, 0, \cdots, 0]$ $\mathbb{R}^{1000}$ which is 8-sparse.

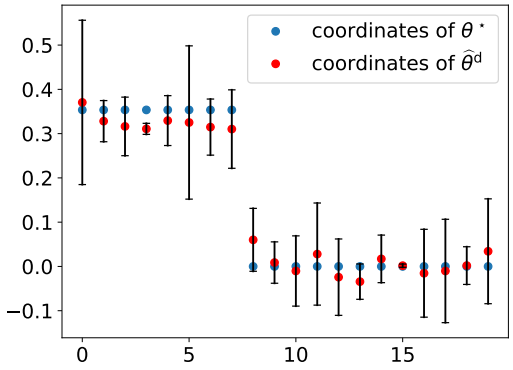

Figure 3: 95% confidence intervals

For our method, the average confidence interval length is 0.14 and average coverage is 0.83. For the de-biased LASSO estimator with known covariance, the average confidence interval length is 0.11 and average coverage is 0.98.

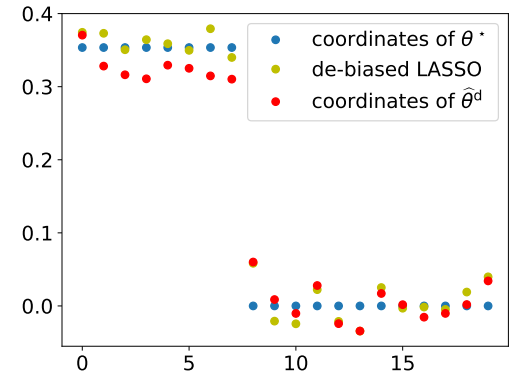

Figure 4: Comparison of our de-biased estimator and oracle de-biased LASSO estimator

H.1.3. TIME SERIES ANALYSIS

In our linear regression simulation, we generate i.i.d. random explanatory variables, and the observation noise is a 0-mean moving average (MA) process independent of the explanatory variables.

For the linear model

$$y_i = \langle x_i, \theta^\star \rangle + \epsilon_i,$$

$x_i \in \mathbb{R}^{20}$ are i.i.d. samples generated from $\mathcal{N}\left([1, 1, \ldots, 1]^\top/\sqrt{k}, I\right)$ and ϵ_i is a 0-mean moving average process

$$\epsilon_i = 0.6 \cdot z_i + 0.8 \cdot z_{i-1},$$

where z_i are i.i.d. $\mathcal{N}(0, 0.7^2)$.

We ran 10,000 simulations, with each time series containing $n = 10,000$ samples, and set the lag $\mathbf{l} = 32$. For our approximate Newton statistical inference procedure (Algorithm 5), average 95% confidence interval (coverage, length) is (0.958, 0.0142), and it matches our theory. For circular bootstrap, where each replicate contains $n - \mathbf{l}$ samples, average 95% confidence interval (coverage, length) is (0.949 , 0.0136).

H.2. Real data

H.2.1. NEURAL NETWORK ADVERSARIAL ATTACK DETECTION

The adversarial perturbation used in our experiments is shown in Figure 7. It is generated using the fast gradient sign method (Goodfellow et al., 2014) Figure 5 shows images in a "Shirt" example. Figure 6 shows images in a "T-shirt/top" example.

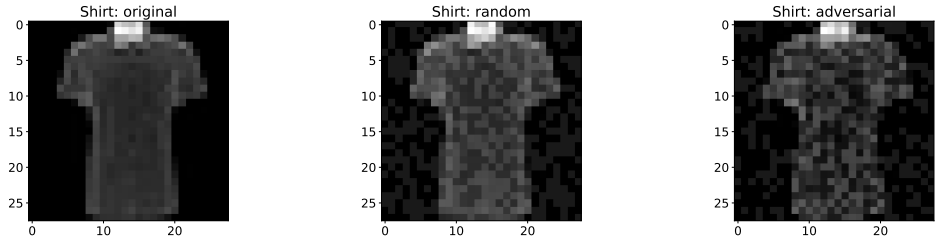

Figure 5: "Shirt" example

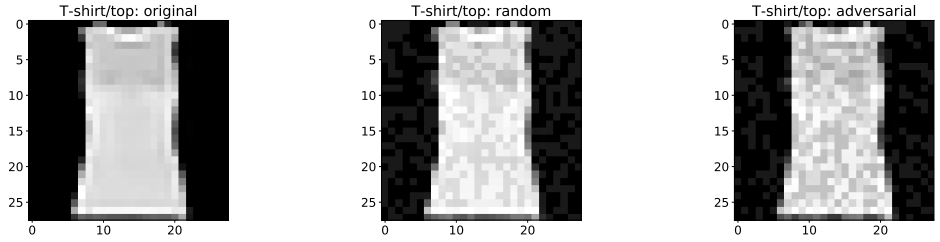

Figure 6: "T-shirt/top" example

H.2.2. HIGH DIMENSIONAL LINEAR REGRESSION

For both experiments, the hyper-parameters are chosen based on the results in Section 3, where we estimate the true parameter's ℓ_1 norm $\|\theta^\star\|_1$ and noise level σ by vanilla LASSO with cross validation, using the LASSO solution's ℓ_1 norm and LASSO residuals' 2^{nd} moment's square root. The covariance threshold is chosen so that it minimizes the thresholded covariance's condition number.

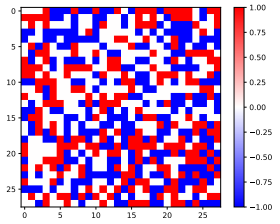

Figure 7: Adversarial perturbation generated using the fast gradient sign method (Goodfellow et al., 2014)

Drug		Mutations
PI	APV	10F
	ATV	33F, 43T, 84V
	IDV	48V, 84A
	LPV	46I
	NFV	46L
	RTV	10I, 54V
	SQV	20R, 84V
NRTI	3TC	184V
	ABC	41L
	AZT	41L, 210W
	D4T	41L, 215Y
	DDI	62V, 151M
	TDF	41L, 75M
NNRTI	DLV	228R
	EFV	74V, 103N
	NVP	103N, 181C

Table 4: HIV drug resistance related mutations detected by our high dimensional inference procedure

HIV drug resistance mutations dataset. We apply our high dimensional inference procedure to the dataset in (Rhee et al., 2006) to detect mutations related to HIV drug resistance. Our procedure is able to detect verified mutations in an expert dataset (Johnson et al., 2005), when we control the family-wise error rate (FWER) at 0.05.

Riboflavin (vitamin B2) production rate data set. For the vanilla LASSO estimate on the high-throughput genomic data set concerning riboflavin (vitamin B2) production rate (Bühlmann et al., 2014), we set $\lambda = 0.021864$. Figure 8, and we see that our point estimate is similar to the vanilla LASSO point estimate.

For statistical inference, in our method, we compute p-values using two-sided Z-test. Adjusting FWER to 5% signifi-cance level, our method does not find any significant gene. (Javanmard and Montanari, 2014; Bühlmann et al., 2014) report that (Bühlmann, 2013) also does not find any significant gene, whereas (Meinshausen et al., 2009) finds one significant gene (YXLD-at), and (Javanmard and Montanari, 2014) finds two significant genes (YXLD-at and YXLE-at). This indicates that our method is more conservative than (Javanmard and Montanari, 2014; Meinshausen et al., 2009).

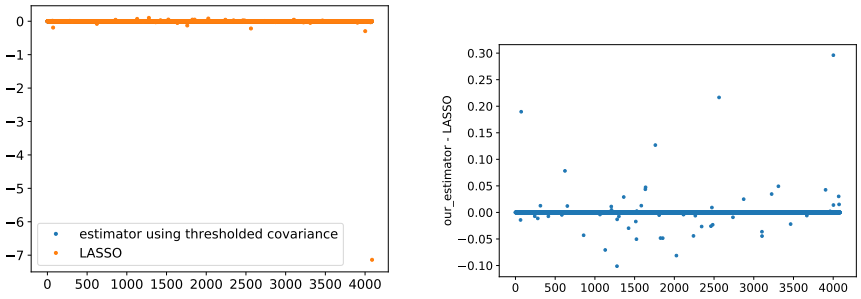

Figure 8: Comparison of our high dimensional linear regression point estimate with the vanilla LASSO estimate

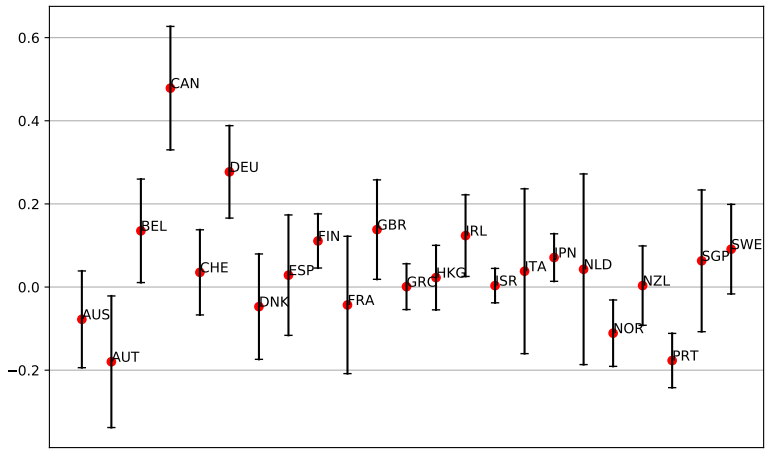

Figure 9: Exposure of US equities market to equities markets of other countries

H.2.3. TIME SERIES ANALYSIS

Using monthly equities returns data from (Frazzini and Pedersen, 2014), we use our approximate Newton statistical inference procedure to show that the correlation between US equities market returns and non-US global equities market returns is statistically significant, which validates the capital asset pricing model (CAPM) (Sharpe, 1964; Lintner, 1965; Fama and French, 2004).

We regress monthly US equities market returns from 1995 to 2018 against other countries' equities market returns, and each country's coefficient and its 95% confidence interval is shown in Figure 9. And we observe that the US market is highly positively correlated with Canada and other advanced economies such as Germany and UK.