Optimal Shrinkage for Distributed Second-Order Optimization

Fangzhao Zhang [1] **Mert Pilanci** [1]

Abstract

In this work, we address the problem of Hessian inversion bias in distributed second-order optimization algorithms. We introduce a novel shrinkage-based estimator for the resolvent of gram matrices that is asymptotically unbiased, and characterize its non-asymptotic convergence rate in the isotropic case. We apply this estimator to bias correction of Newton steps in distributed second-order optimization algorithms, as well as randomized sketching based methods. We examine the bias present in the naive averaging-based distributed Newton's method using analytical expressions and contrast it with our proposed bias-free approach. Our approach leads to significant improvements in convergence rate compared to standard baselines and recent proposals, as shown through experiments on both real and synthetic datasets.

1. Introduction

In a distributed setting, where multiple agents have access only to subsets of the entire dataset, accurate estimation of the Hessian inverse and Newton steps is crucial for the effective application of second-order optimization algorithms. A straightforward way for estimating Hessian inverse is to simply collect and average over all local Hessian inverses, however, this is usually not accurate due to the existence of inversion bias, i.e., in general

$$\lim_{m\to\infty}\left\|\frac{1}{m}\sum_{i=1}^{m}H_i^{-1}-H^{-1}\right\|\neq 0$$

where m represents the total number of agents, such as distributed workers, and H_i is the local Hessian computed by worker i (see Theorem 2.5 for more details). Therefore, a naive averaging of local Hessian inverses leads to a biased estimator of the global Hessian inverse. As a result, Newton steps computed by averaging local Newton steps can be far from exact.

Different ways to reduce the inversion bias mentioned above have already been studied by a line of prior work. A particularly related one is the determinantal averaging method proposed by Dereziński & Mahoney (2019). The authors show that $\sum_{i=1}^{m}\det(H_i)H_i^{-1}/\sum_{i=1}^{m}\det(H_i)$ serves as an unbiased estimator of the global Hessian inverse when the data is uniformly distributed to each agent. However, this method has shortcomings involving the overhead of computing local Hessian determinants, and potential numerical instabilities in computing determinants of local Hessian when data is of large dimension.

In this work, we borrow tools from random matrix theory and study the problem of estimation of the covariance resolvent when the data is randomly distributed. A key observation is that the inverses of positive semidefinite Hessians typically have the form of a covariance resolvent $(\frac{1}{n}X^TD^2X+\lambda I)^{-1}$ where $\frac{1}{n}X^TD^2X$ is the empirical covariance of an appropriately chosen matrix in which D is diagonal and X is a data matrix. We propose an asymptotic unbiased estimator of covariance resolvent in the form of a shrinkage formula (Theorem 2.2) with its informal version stated below, and we also characterize the non-asymptotic convergence rate (Section 2.1.1). Specifically, we find that under some weak assumptions on the data distribution,

Theorem. *(informal, see Theorem 2.2 for details)*

For a random data matrix $A\in\mathbb{R}^{n\times d}$, let d_λ denote the effective dimension of the true covariance Σ_n, and $\hat{\Sigma}_n$ denote the empirical covariance,

$$\lim_{\substack{n,d\to\infty\\ \frac{d}{n}\to y\in[0,1)}}\left\|\mathbb{E}\left[\left(\gamma\hat{\Sigma}_n+\lambda I\right)^{-1}\right]-(\Sigma_n+\lambda I)^{-1}\right\|=0$$

where $\gamma=\frac{1}{1-\frac{d_\lambda}{n}}$.

This result implies that a simple scaling of the Hessian by $\frac{1}{1-\frac{d_\lambda}{n}}$ removes the inversion bias, where d_λ is the effective dimension of the covariance. Since the Hessian inverse is related to covariance resolvents, this theorem can help reducing Hessian inversion bias in the large data regime. We study its application to distributed second-order optimiza-

[1]Department of Electrical Engineering, Stanford University. Correspondence to: Fangzhao Zhang <zfzhao@stanford.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

tion algorithms and randomized second-order optimization algorithms, where we observe a significant speedup in the convergence rate compared to baseline methods (see Figure 1 below, and more simulation results in Section 5 and Appendix C).

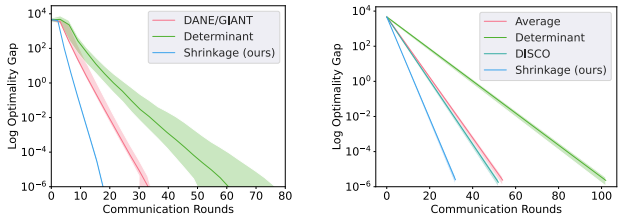

Figure 1: Synthetic data experiments on ridge regression. Total number of data $n = 30000$, data dimension $d = 150$, number of agents $m = 200$, regularizer $\lambda = 0.01$. The left plot shows the convergence of distributed Newton's method (Algorithm 1). The right plot shows the convergence of our distributed preconditioned conjugate gradient method (Algorithm 2). Step-sizes are chosen via line search in all methods. See Section 5 for details.

1.1. Prior Work

Early work on distributed Newton-type methods including DANE studied by Shamir et al. (2013), where the approximate Newton step is an average of local mirror descent steps. Later work such as AIDE (Reddi et al., 2016), standing for accelerated inexact DANE, solves an Nesterov accelerated version of DANE's local optimization problem inexactly to some degree of accuracy. Other works include COCOA (Ma et al., 2015) which solves a dual problem of logistic regression locally, DiSCO (Zhang & Lin, 2015) which considers inexact damped Newton's step solved with distributed preconditioned conjugate gradient method using the resolvent of first agent's local Hessian as the preconditioning matrix, and GIANT (Wang et al., 2017) which uses the average of local Newton's step as the global one. More recent work by Dobriban & Sheng (2018) analyzes performance loss in one-shot weighted averaging and iterative averaging for linear regression, and Dobriban & Sheng (2019) also study one-shot weighted averaging for distributed ridge regression in high dimensions.

While most of the work mentioned above does not address the Hessian inversion bias directly, another line of work considers this issue. For example, the determinantal averaging method (Dereziński & Mahoney, 2019) states that one can pass by the inversion bias as long as an unbiased estimator of Hessian matrix exists and computing determinants of local Hessian matrices is feasible. This method is unstable when data is of large dimension since local Hessian determinant computation is usually infeasible there. It also introduces computational overhead for computing local Hessian determinants. Zhang et al. (2012) proposed a bootstrap subsampling method to reduce bias in one-shot averaging that improves the approximated optimizer to some finite suboptimality. However, the improved approximated optimizer can still be much worse compared to the true optimizer (shown in (Shamir et al., 2013), Section 2).

Our work addresses bias correction in distributed optimization algorithms with analytical tools from random matrix theory. When data is independently and identically distributed, we introduce a shrinkage formula for estimating the resolvent of the covariance matrix of data which can serve as an inversion bias corrector in the large data regime. This bias correction method is more stable compared to the determinantal averaging method and achieves significantly better accuracy without any notable computational overhead. In the field of asymptotic random matrix theory, prior works studying similar shrinkage formulas exist only for statistical estimation of covariance and precision matrices. Ledoit & Wolf (2004) is one among the early works studying linear shrinkage estimator of large covariance matrices. Later work by Bodnar et al. (Bodnar et al., 2014) studies linear shrinkage estimator of large covariance matrices with almost sure smallest Frobenius loss. In Bodnar et al. (2016), they studied linear shrinkage estimator for the precision matrix. Bodnar et al. (2022) gives a comprehensive review of recent advancements in shrinkage-based high dimensional inference studies. These shrinkage-based methods have already been successfully applied to tests for weights for portfolios (Bodnar et al., 2019) and to robust adaptive beamforming (Xiao et al., 2018). Our work focuses on shrinkage-based estimation of the resolvent of covariance matrices, which is different from the shrinkage formula for both covariance and precision matrices, and we study its application to optimization algorithms.

Besides distributed second-order optimization, we find the shrinkage formula we studied is also useful for improving sketching methods. We note Dereziński et al. (2020) studied debiasing randomized optimization algorithms with surrogate sketching, where a non-standard carefully chosen sketching matrix is used. Also, in Bartan & Pilanci (2022), the authors exploited random matrix theory for bias elimination in distributed randomized ridge regression. They showed when the covariance is isotropic, an asymptotically unbiased estimator can be obtained by tuning local regularizers. Unlike their approach, our method works for general covariance matrices and sketching matrices.

1.2. Contribution

In this work, we propose an asymptotically unbiased shrinkage-based estimator of the resolvent of covariance matrices. Unlike most prior studies in the field of large-dimensional random matrix theory which consider only the asymptotic settings, we also characterize the non-asymptotic

convergence rate. Furthermore, we study the application of this shrinkage formula to distributed second-order optimization algorithms and randomized sketching methods. We carry out real data simulations where a significant convergence speedup is obtained compared to standard baselines.

To our best knowledge, there is no existing work on shrinkage-based estimation of the resolvent of covariance matrices, and we are also the first to derive a closed-form formula for optimal shrinkage and apply it to optimization.

2. Main Theorems

In this section, we establish our main theoretic results. A shrinkage-based asymptotically unbiased estimator of the resolvent of covariance matrices is studied with its convergence rate characterized and its variant in the small regularizer regime analyzed in Section 2.1. We then show that the commonly used averaging method has non-zero asymptotic bias and summarize different methods for estimating the resolvent of covariance matrices in Section 2.2.

We first introduce notations we use for stating our main theorems. We follow the classical Kolmogorov asymptotics. Consider a sequence of problems $\mathcal{B}_n = (\Sigma, \hat{\Sigma}, x, d)_n$ with $n, d \to \infty$, $d/n \to y \in [0, 1)$. $\Sigma_n \in \mathbb{R}^{d\times d}$ is the true covariance and data distribution satisfies $\mathbb{E}[x] = 0, \mathrm{Cov}(x) = \Sigma_n$. Empirical covariance is denoted as $\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$ where x_i's are i.i.d. samples of x. Consider any $\lambda > 0, \lambda \in \mathbb{R}$. $d_\lambda(\Sigma) = \mathrm{tr}(\Sigma(\Sigma + \lambda I)^{-1})$ is the effective dimension of Σ and we define $d_\lambda^n = d_\lambda(\Sigma_n)$. The empirical spectral distribution (e.s.d.) of Σ_n is $F_{\Sigma_n}(u) = d^{-1}\sum_{i=1}^d \mathbb{I}_{(\lambda_i \le u)}$ where λ_i's are eigenvalues of Σ_n. We use $F_{\Sigma_n}(u) \to F_\Sigma(u)$ to indicate that the e.s.d. converges almost surely to a density $F_\Sigma(u)$. We further define $M = \sup_{|e|=1}\mathbb{E}(e^T x)^4$, $\nu = \sup_{\|\Omega\|=1}\mathrm{Var}(x^T\Omega x/d)$.

We collect the main assumptions required below,

Assumption 2.1.

A1. $F_{\Sigma_n}(u) \to F_\Sigma(u)$ almost for any $u \ge 0$

A2. $M < \infty, \nu \to 0$

A3. eigenvalues of each Σ_n are in the interval $[\sigma_{\min}, \sigma_{\max}]$ where $\sigma_{\min} > 0$ and $\sigma_{\max}$ does not depend on n

2.1. Asymptotically Unbiased Shrinkage Formula for the Resolvent of Covariance

We propose an asymptotically unbiased estimator of the resolvent of covariance matrices in the form of a local shrinkage formula under notations defined at the beginning of this section, and we characterize an explicit convergence rate for the isotropic case in Section 2.1.1.

Theorem 2.2. *Under Assumption 2.1, for any $\lambda > 0$ and any data matrix with i.i.d. rows, assume additionally $d_\lambda^n < n$ for each n, d, then we have*

$$\mathbb{E}\left[\left(\frac{1}{1-\frac{d_\lambda^n}{n}}\hat{\Sigma}_n + \lambda I\right)^{-1}\right] = (\Sigma_n + \lambda I)^{-1} + \Omega_0$$

where $\|\Omega_0\| \to 0$ as $n, d \to \infty$, $d/n \to y \in [0, 1)$. [1]

Proof. See Appendix A.2. □

Note that this result universally holds for a large class of random data matrices with i.i.d. rows. In Assumption 2.1, we only require M bounded and ν vanishing as n, d tend to infinity. To see such constraints are quite mild, note requiring $M = \sup_{|e|=1}\mathbb{E}(e^T x)^4$ staying bounded is essentially bounding the dependence of x's components. When $x \sim \mathcal{N}(0, \Sigma)$, $M = 3\|\Sigma\|^2$. For $\nu = \sup_{\|\Omega\|=1}\mathrm{Var}(x^T\Omega x/d)$, when components of x are independent, $\nu \le M/d$, and when $x \sim \mathcal{N}(0, \Sigma), \nu \le 2\|\Sigma\|^2/d$. For more discussions on these constraints, see Serdobolskii (2008), Chapter 3.1.

2.1.1. Isotropic Convergence Rate

To evaluate the convergence rate in Theorem 2.2, note if there is no constraint on how fast d/n converges to y, then the convergence rate can be arbitrarily bad. To characterize the convergence rate, here we require $d/n = y$ always holds. Then from the expression for Ω_0 (see inequality (7) in Appendix A.2), we need to find the convergence rate of the Stieltjes transform of the spectral distribution of gram matrices. Such task is in general hard if no constraint on the covariance matrices is imposed. Prior work analyzing such bounds exists for covariance matrices with correlated Curie-Weiss entries (Fleermann & Heiny, 2019), sparse covariance matrices (Erdős et al., 2020), covariance matrices with independent but not necessarily identically distributed entries (Bai & Silverstein, 2010).

Here we focus on the isotropic covariance case where we can exploit isotropic local Marchenko-Pastur law to derive an explicit convergence rate for Theorem 2.2.

Theorem 2.3. *When $\Sigma_n = I$, under Assumption 2.1,*

$$\|\Omega_0\| \in \mathcal{O}\left(\frac{1}{\sqrt{n}} + \sqrt{\nu}\right)$$

Proof. See Appendix A.3. □

Remark. For $x \sim \mathcal{N}(0, I)$, Assumption 2.1 holds and $\|\Omega_0\| \in \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$

[1] Here Ω_0 depends on n and we write $\Omega_0 = \Omega_0(n)$ for notational simplicity. The same convention is used in later sections as well as in the appendix.

2.1.2. Small Regularizer Regime

In this subsection, we study the behavior of Theorem 2.2 when the regularizer diminishes to zero, which results in a simpler local shrinkage coefficient requiring no estimation of any effective dimension.

Theorem 2.4. *Under Assumption 2.1, assume $d < n$ always holds. If Σ_n is invertible and eigenvalues of $\hat{\Sigma}_n$ are bounded away from zero, then for $\epsilon > 0$,*

$$\mathbb{E}\left[\left(\frac{1}{1-\frac{d}{n}}\hat{\Sigma}_n + \epsilon I\right)^{-1}\right] = \Sigma_n^{-1} + \Omega_1$$

with $\|\Omega_1\| \to 0$ as $n, d \to \infty$ and $\epsilon \to 0$. If Σ_n is not invertible, replacing Σ_n^{-1} by $(\Sigma_n + \epsilon I)^{-1}$, the theorem still holds.

Proof. See Appendix A.4. □

Remark. With a small regularizer, the estimation of the resolvent of covariance matrices is close to the estimation of precision matrices. Theorem 2.4 parallels results on shrinkage estimators for the precision matrix as investigated in Theorem 3.2 of Bodnar et al. (2015). However, their work only considers the asymptotic setting.

2.2. Asymptotic Bias of the Naive Averaging Method

After introducing our asymptotic unbiased estimator for the resolvent of covariance matrices, we now analyze the asymptotic bias for simple averaging method without shrinkage. Theorem 2.5 below states that this bias is non-zero for small λ. See Appendix A.5 for analysis for general positive λ.

Theorem 2.5. *Under Assumption 2.1, and if $\lambda \in o(1)$,*

$$\lim_{\substack{n,d\to\infty \\ d/n\to y}} \left\| \mathbb{E}\left(\hat{\Sigma}_n + \lambda I\right)^{-1} - (\Sigma_n + \lambda I)^{-1} \right\| \geq \frac{y\sigma_{\min}}{\sigma_{\max}^2}$$

Proof. See Appendix A.5. □

This result demonstrates that the asymptotic bias for simple averaging method without shrinkage can be substantial, and our proposed shrinkage formula serves as an effective improvement to solve this issue.

As a summary, for random data matrices of growing size satisfying Assumption 2.1, Table 1 summarizes the bias of various methods for estimating the resolvent of covariance matrices under both non-asymptotic and asymptotic settings.

3. Application to Distributed Second-Order Optimization Algorithms

Now we are ready to utilize theorems studied in Section 2 in distributed second-order optimization algorithms. We outline the algorithms for distributed Newton's method with optimal shrinkage and its inexact version solved with distributed preconditioned conjugate gradient method with optimal shrinkage below. The convergence proofs for quadratic loss and general convex smooth loss are provided in Section 3.1, Section 3.2, and Appendix B.4. Finally, we analyze communication and computation complexity for the proposed algorithms in Section 3.3.

Let n denote the number of data samples and m denote the number of agents. Consider the ℓ_2 regularized loss function $f(x) = \frac{1}{m}\sum_{i=1}^m f_i(x) + \frac{\lambda}{2}\|x\|_2^2$ where $f_i(x) = \frac{1}{k}\sum_{j=1}^k \ell_{ij}(x)$ denotes loss function corresponding to agent i and k is the number of samples available to each agent. Here, we consider the case where the data is evenly split to all agents and thus $k = n/m$. Denote $\nabla^2 f$ as the Hessian and ∇f as the gradient of function f. In order to apply our results in Section 2, we need the effective dimension of the population Hessian $\mathbb{E}\frac{1}{m}\sum_{i=1}^m \nabla^2 f_i(x)$. We use the empirical effective dimension $d_{\lambda,i} = \operatorname{tr}\left(\nabla^2 f_i(\nabla^2 f_i + \lambda I)^{-1}\right)$ available at each agent as an approximation. Algorithm 1 gives a description of the proposed method.

Algorithm 1 Distributed Newton's method with optimal shrinkage

Initialize: starting point $x^{(0)}, t = 1$
repeat
  Gather local gradients $\nabla f_i(x^{(t-1)})$ from each agent i
  Compute global gradient $\nabla f(x^{(t-1)}) = \frac{1}{m}\sum_{i=1}^m \nabla f_i(x^{(t-1)}) + \lambda x^{(t-1)}$ and broadcast to all agents
  for $i = 1, 2, \ldots$ **do**
    agent i computes $x_i^{(t)} = \left(\frac{1}{1-\frac{md_{\lambda,i}}{n}}\nabla^2 f_i\left(x^{(t-1)}\right) + \lambda I\right)^{-1}\nabla f(x^{(t-1)})$
  end for
  Compute approximate Newton step $\Delta x^{(t)} = \frac{1}{m}\sum_{i=1}^m x_i^{(t)}$
  (Optional) Choose step size η by line search
  Update $x^{(t)} = x^{(t-1)} - \eta\Delta x^{(t)}$, $t = t + 1$
until convergence criterion or maximum iterates reached

We then provide a preconditioned conjugate gradient method that exploits optimal shrinkage for solving a single Newton step in Newton's method. Let v denote the current point at which the next Newton step needs to be performed. Algorithm 2 gives the distributed preconditioned conjugate gradient method for inexact Newton's method.

3.1. Convergence Analysis for Regularized Quadratic Loss

Given data matrix $A \in \mathbb{R}^{n\times d}$ with data i.i.d. with mean zero and covariance Σ, and label $b \in \mathbb{R}^n$, let $A^{(i)} \in \mathbb{R}^{(n/m)\times d}$

METHOD	SAMPLE SIZE	NON-ASYMPTOTIC BIAS	ASYMPTOTIC BIAS	COMPLEXITY
AVERAGING	$\forall k$	SEE (8)	$\geq y\sigma_{\min}/\sigma_{\max}^2$	0
DETERMINANTAL AVERAGING	$\forall k$	0	0	$\mathcal{O}(d^3)$
OPTIMAL SHRINKAGE	$k \geq d_\lambda$	$\leq \mathcal{O}\left(\frac{1}{\sqrt{n}} + \sqrt{\nu}\right)$	0	0

Table 1: The bias in the estimation of covariance resolvent with different methods under asymptotic/non-asymptotic settings. For any real positive λ, d_λ denotes the effective dimension of the covariance matrix. The asymptotic bias for the averaging method is given for $\lambda \in o(1)$, see Theorem 2.5 for details. The non-asymptotic bias for optimal shrinkage method is given for isotropic covariance, see Section 2.1.1 for details. The complexity column refers to agents' additional computational overhead compared to the averaging method. When $\lambda \in o(1)$, shrinkage coefficient requires no effective dimension estimation and thus no computational overhead, see Section 2.1.2 for details. When $\lambda \notin o(1)$, we assume the effective dimension of the covariance is either known or has been estimated in advance and is distributed to all agents .

Algorithm 2 Distributed preconditioned conjugate gradient with optimal shrinkage

Compute $b = \nabla f(v) = \frac{1}{m}\sum_{i=1}^m \nabla f_i(v) + \lambda v$ by gathering local gradients $\nabla f_i(v)$ from each agent
Initialize: $x = 0, r = b$
for $i = 1$ **to** m **do**
  agent i computes $z_i = \left(\frac{1}{1-\frac{md_{\lambda,i}}{n}}\nabla^2 f_i(v) + \lambda I\right)^{-1} r$
end for
$z = \frac{1}{m}\sum_{i=1}^m z_i, p = z, \rho_1 = r^T z$
for $t = 1$ **to** $t_{\max}$ **do**
  quit if stopping criterion achieved
  $\omega = \sum_{i=1}^m \nabla^2 f_i(v)p + \lambda p$ by gathering $\nabla^2 f_i(v)p$
  $\alpha = \frac{\rho_t}{\omega^T p}$
  $x = x + \alpha p$
  $r = r - \alpha\omega$
  for $i = 1$ **to** m **do**
    agent i computes z_i
    $= \left(\frac{1}{1-\frac{md_{\lambda,i}}{n}}\nabla^2 f_i(v) + \lambda I\right)^{-1} r$
  end for
  $z = \frac{1}{m}\sum_{i=1}^m z_i$
  $\rho_{t+1} = z^T r$
  $p = z + \frac{\rho_{t+1}}{\rho_t} p$
end for

denote agent i's local data. Consider the regularized quadratic loss function $f(x) = \frac{1}{2n}\|Ax-b\|_2^2 + \frac{\lambda}{2}\|x\|_2^2$ with gradient $g(x) := \nabla f(x)$ and Hessian $H := \nabla^2 f(x) = \frac{1}{n}A^T A + \lambda I$. Denote

$$\tilde{H} = \left(\frac{1}{m}\sum_{i=1}^m \left(\frac{1}{1-\frac{md_\lambda}{n}}\frac{m}{n}A^{(i)^T}A^{(i)} + \lambda I\right)^{-1}\right)^{-1}$$

with the true effective dimension of the covariance, $\Sigma := \mathbb{E}\frac{1}{n}A^T A$

$$d_\lambda = \text{tr}\left(\Sigma\left(\Sigma + \lambda I\right)^{-1}\right)$$

Theorem 3.1. *(Convergence of Newton's method with Shrinkage) Denote $\Delta_{t+1} = \omega_{t+1} - \omega^\star$ with $\omega^\star = \text{argmin } f(x)$ and $\omega_{t+1} = \omega_t - \tilde{H}^{-1}g(\omega_t)$. Then,*

$$\|\Delta_{t+1}\| \leq \beta\|\Delta_t\|$$

where $\beta = \frac{\sqrt{2}\alpha}{\sqrt{1-\alpha^2}}\sqrt{\frac{\sigma_{\max}+\lambda+\alpha_0}{\sigma_{\min}+\lambda-\alpha_0}}$ with $\alpha_0 = \|\Sigma - \frac{1}{n}A^T A\|, \alpha_1 = \|\tilde{H}^{-1} - \mathbb{E}[\tilde{H}^{-1}]\|$, and $\alpha = (\sigma_{\max} + \lambda + \alpha_0)\left(\frac{1}{\lambda^2}\alpha_0 + \alpha_1 + \|\Omega_0\|\right)$. $\sigma_{\min}$ and $\sigma_{\max}$ denote the smallest and largest eigenvalues of Σ correspondingly.

Proof. See Appendix B.2. □

The most important aspect of the above result is that the contraction rate β *vanishes to zero* as the number of workers increase and data dimensions grow asymptotically as we formalize next.

Remark. Consider a sequence of data matrices $\{A\}_n$ with $n/m \to \infty, d \to \infty, md/n \to y \in [0,1)$ and each $A^{(i)}$ satisfies Assumption 2.1, then $\|\Omega_0\| \to 0$ by Theorem 2.2. When data is Gaussian or sub-Gaussian, $\alpha_0 \to 0$ almost surely given $d/n \to 0$. By standard matrix concentration bounds, $\alpha_1 \leq \epsilon$ with probability $\geq 1 - 2d\exp\left(-\epsilon^2/\left((4\epsilon)/(3m\lambda) + 2/(m\lambda^2)\right)\right)$. Thus $\beta \to 0$ almost surely when each $A^{(i)}$ satisfies Assumption 2.1, data is Gaussian or subGaussian, and $m, d, n \to \infty, d, m = o(n), \log d = o(m), md/n \to y \in [0,1)$.

Theorem 3.2. *(Convergence of inexact Newton's method with Shrinkage) Let $\alpha, \alpha_0, \Delta_{t+1}, \sigma_{\min}, \sigma_{\max}$ defined as in Theorem 3.1. Then,*

$$\|\Delta_{t+1}\| \leq \frac{\sqrt{2}\alpha'}{\sqrt{1-\alpha'^2}}\sqrt{\frac{\sigma_{\max}+\lambda+\alpha_0}{\sigma_{\min}+\lambda-\alpha_0}}\|\Delta_t\|$$

where $\alpha' = \sqrt{4\left(\frac{1-\sqrt{1-\frac{\alpha}{1-\alpha}}}{1+\sqrt{1-\frac{\alpha}{1-\alpha}}}\right)^{s_t}}$ and s_t denotes the number of iterations in preconditioned conjugate gradient method.

Proof. The derivation follows by Lemma 14 in Dereziński & Mahoney (2019) and Lemma B.1. □

3.2. Convergence Analysis for Regularized General Convex Smooth Loss

Given data matrix $A \in \mathbb{R}^{n\times d}$ with data i.i.d. with mean zero. Let $A^{(i)} \in \mathbb{R}^{(n/m)\times d}$ denote agent i's local data and $A_j^{(i)}$ denote the jth piece of data held by agent i. Consider general convex smooth loss function f of the following form,

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x^T A_i) + \frac{\lambda}{2}\|x\|^2$$

with gradient $g(x) := \nabla f(x)$ and hessian $H(x) := \nabla^2 f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i''(x^T A_i) A_i A_i^T + \lambda I$. Assume f is twice differentiable and its hessian is L-Lipschitz. Denote

$$\tilde{H}(x)^{-1} = \frac{1}{m}\sum_{i=1}^{m}\left(\gamma \sum_{j=1}^{n/m} f_i''\left(x^T A_j^{(i)}\right) A_j^{(i)} A_j^{(i)^T} + \lambda I\right)^{-1}$$

where $\gamma = m/n\left(1 - \frac{m d_\lambda(x)}{n}\right)$ with $d_\lambda(x) = \mathrm{tr}\left(\Sigma(x)\left(\Sigma(x)+\lambda I\right)^{-1}\right)$ and $\Sigma(x)$ defined in Theorem 3.3.

Theorem 3.3. *(Convergence of Newton's method with Shrinkage) Assume ω_t independent of all $A_j^{(i)}$'s, denote $\Sigma(\omega_t) = Cov\left(f_i''\left(\omega_t^T A_j^{(i)}\right)^{\frac{1}{2}} A_j^{(i)}\right)$. Denote $\Delta_{t+1} = \omega_{t+1} - \omega^\star$ with $\omega^\star = argmin\ f(x)$ and $\omega_{t+1} = \omega_t - \tilde{H}^{-1}(\omega_t) g(\omega_t)$,*

$$\|\Delta_{t+1}\| \le \max\left\{\frac{2L}{\sigma_{\min}+\lambda-\alpha_0}\|\Delta_t\|^2, \frac{\sqrt{2}\alpha}{\sqrt{1-\alpha^2}}\sqrt{\frac{\sigma_{\max}+\lambda+\alpha_0}{\sigma_{\min}+\lambda-\alpha_0}}\|\Delta_t\|\right\}$$

where $\alpha = (\sigma_{\max}+\lambda+\alpha_0)\left(\frac{1}{\lambda^2}\alpha_0 + \alpha_1 + \|\Omega_0\|\right)$ with $\alpha_0 = \left\|\Sigma(\omega_t) - \frac{1}{n}\sum_{i=1}^{n} f_i''\left(\omega_t^T A_i\right) A_i A_i^T\right\|$ and $\alpha_1 = \left\|\tilde{H}(\omega_t)^{-1} - \mathbb{E}\left[\tilde{H}(\omega_t)^{-1}\right]\right\|$. $\sigma_{\min}$ and $\sigma_{\max}$ denote the smallest and largest eigenvalues of $\Sigma(\omega_t)$ correspondingly.

Proof. See Appendix B.3. □

3.3. Communication and Computation Complexity Analysis

In this section, we analyze the communication and computation complexity of distributed Newton's method with optimal shrinkage. The analysis for inexact Newton's method is similar and omitted.

On the communication side, in each Newton iteration[2], four rounds of communication between server and agents are required: the server broadcasts the current iterate, collects the local gradients, and then computes and broadcasts the global gradient to the agents and collects local approximate descent directions. Each communication involves $\mathcal{O}(d)$ words. In ridge regression, to achieve some fixed accuracy $\|\Delta_t\| \le \epsilon$, the number of iterations is bounded by $\mathcal{O}\left(\frac{\log(\sqrt{\kappa}/\epsilon)}{\log(\sqrt{1-\alpha^2}/\alpha)}\right)$ where κ is the condition number of the Hessian matrix and α is asymptotically vanishing (see Theorem 3.1 for definition). Therefore, the number of iterations decreases as the number of workers increases and vanishes asymptotically. This should be compared with GIANT's bound $\mathcal{O}\left(\frac{\log(d\kappa/\epsilon)}{\log(n/\mu dm)}\right)$ on the number of Newton iterations to achieve the same degree of accuracy, which does not vanish asymptotically due to the Hessian inversion bias. Note that our bound also gets rid of the dependency on the matrix coherence number μ, and we do not impose assumption such as $n > \mathcal{O}(\mu dm)$. Compared to DiSCO's bound $\tilde{\mathcal{O}}\left(\frac{d\kappa^{1/2}m^{3/4}}{n^{3/4}} + \frac{\kappa^{1/2}m^{1/4}}{n^{1/4}}\log\frac{1}{\epsilon}\right)$ which is also non-vanishing asymptotically due to the inversion bias, our bound only has log dependency on the square root of κ instead of polynomial dependency, which is an improvement[3].

The per-iteration computation complexity for each agent involves operations required for forming the local gradient, solving a linear equation involving the local Hessian matrix, and computing the effective dimension of the local Hessian matrix. The only overhead compared to GIANT is the computation of the effective dimension of the local Hessian matrix, which can be done in $\mathcal{O}(kd\min\{k,d\})$ by singular value decomposition or can be estimated much quicker by trace estimation method such as Hutch++ (Meyer et al., 2020). Note if the effective dimension of the global Hessian matrix is known beforehand, then it can be used in replace of the effective dimension of local Hessian matrices and there is no additional computational overhead for each agent. Another option is to use md/n in replace of $md_{\lambda i}/n$ in Algorithm 1, which should work well for small regularizer λ by Theorem 2.4. There is no significant computational complexity on the server's side since only some simple averaging and subtraction operations are required.

4. Application to Randomized Second-Order Optimization Algorithms

In this section, we study the application of our main theoretic result in Section 2 to randomized second-order op-

[2]Assuming fixed step size instead of line search for simplicity.

[3]The iteration bounds for GIANT and DiSCO are taken from Wang et al. (2017), Section 1.1. Note the number of iterations for DiSCO is required to achieve ϵ-accuracy in terms of function value evaluations.

timization algorithms. We mainly focus on the classic Iterative Hessian Sketch (IHS) method and discuss how our shrinkage formula can be used for bias correction there. Our results can also be used for randomized preconditioned conjugate gradient method and the more general Newton Sketch method.

We first give an introduction on randomized sketching which is used in the Iterative Hessian Sketch method. Randomized sketching is an important tool in randomized linear algebra for dealing large-scale data problems. Given a data matrix $A \in \mathbb{R}^{n\times d}$, consider a randomized sketching matrix $S \in \mathbb{R}^{m\times n}$, SA is referred to as a sketch of A and is of size m by d. It is usually the case $m < n$, thus storage is reduced when SA is stored instead of the original data matrix A, and with carefully chosen S, some properties of A can be preserved by considering only SA. When SA is composed of randomly selected rows of A, it is referred to as row subsampling and when SA is composed of random linear combination of rescaled rows of A, it is referred to as a random projection. A commonly used random projection is Gaussian projection, where S contains i.i.d. Gaussian entries $\mathcal{N}\left(0, \frac{1}{m}\right)$.

Consider regularized quadratic loss function defined in Section 3.1. With data matrix $A \in \mathbb{R}^{n\times d}$ and labels $b \in \mathbb{R}^n$, the Hessian matrix is $H = \frac{1}{n}A^TA+\lambda I$. Note we are not requiring the data to be i.i.d. with mean zero here as required in Section 3.1. Let $S \in \mathbb{R}^{m\times n}$ be Gaussian projection matrix. Iterative Hessian Sketch method is a preconditioned first-order method that replaces the Newton direction $H^{-1}g_t$ at the t-th Newton's step in Newton's method by $H_S^{-1}g_t$ where $H_S = \frac{1}{n}A^TS^TSA + \lambda I$ and g_t denotes the gradient at the t-th Newton's step.

Since H_S can be expressed as $\frac{1}{m}\sum_{i=1}^m x_i x_i^T + \lambda I$ where $x_i \sim \mathcal{N}\left(0, \frac{1}{n}A^TA\right)$, debiasing IHS reduces to minimizing $\|H_S^{-1} - H^{-1}\|$ and is exactly the problem of unbiased estimation of a covariance resolvent. We adapt Theorem 2.2 as below and obtain an asymptotically unbiased estimation of H^{-1} as a shrinkage formula involving H_S.

Consider a sequence of problems $\mathcal{B}_n = (A, S, d, m)_n$ with data matrix $A \in \mathbb{R}^{n\times d}, S \in \mathbb{R}^{m\times n}$ being Gaussian projection matrix, $m, d \to \infty$ and $d/m \to y \in [0, 1)$. Assume the empirical spectral distribution $F_{(\frac{1}{n}A^TA)}(u) \to F_\Sigma(u)$ almost surely for almost every $u \geq 0$, $\left\|n^{-1}A^TA\right\|^2 \leq \infty$, and eigenvalues of each $n^{-1}A^TA$ are located on a segment $[\sigma_{\min}, \sigma_{\max}]$ where $\sigma_{\min} > 0$ and $\sigma_{\max}$ does not depend on n. Denote $d_\lambda^n = \text{tr}\left(\frac{1}{n}A^TA\left(\frac{1}{n}A^TA + \lambda I\right)^{-1}\right)$.

Theorem 4.1. *Assume additionally $d_\lambda^n < m$ always holds. For any $\lambda > 0$,*

$$\lim_{m,d\to\infty}\left\|\mathbb{E}\left[\left(\frac{1}{1-\frac{d_\lambda^n}{m}}(H_S - \lambda I) + \lambda I\right)^{-1}\right] - H^{-1}\right\| = 0$$

Theorem 4.1 suggests to use $\left(\frac{1}{1-\frac{d_\lambda^n}{m}}(H_S - \lambda I) + \lambda I\right)^{-1}$ in place of H_S^{-1} when d_λ^n is available. For practicality, when only sketched data SA is available, $\tilde{d}_\lambda^n = \text{tr}\left(\frac{1}{n}A^TS^TSA\left(\frac{1}{n}A^TS^TSA + \lambda I\right)^{-1}\right)$ can be used as an approximation for d_λ^n. According to our real data simulation results in Section 5.4, using $\tilde{d}_\lambda^n$ in place of d_λ^n significantly improves the plain IHS method in most cases.

In Lacotte et al. (2021), Chapter 3.2, the authors studied preconditioned conjugate gradient method with H_S^{-1} as the preconditioning matrix. Using the shrinked version $\left(\frac{1}{1-\frac{d_\lambda^n}{m}}(H_S - \lambda I) + \lambda I\right)^{-1}$ in replace of H_S^{-1} there will also help.

Newton Sketch method generalizes IHS to regularized general convex smooth loss defined in Section 3.2. At tth Newton's step, the Hessian matrix can be expressed as $H_t = \frac{1}{n}A_t^TA_t + \lambda I$ with an appropriate choice of matrix A_t, and Newton Sketch method proposes to use $\left(\frac{1}{n}A_t^TS^TSA_t + \lambda I\right)^{-1}g_t$ as approximate Newton's descent direction with g_t denoting the gradient at the tth Newton's step. By replacing A with A_t in Theorem 4.1, the asymptotically unbiased estimation for H_t^{-1} can be derived.

Table 2 summarizes the bias to estimate the Hessian inverse for regularized quadratic loss with classic IHS paradigm and IHS with optimal shrinkage described above under both non-asymptotic setting and asymptotic setting.

5. Numerical Simulation

We now present synthetic and real data simulation results . All real datasets used in this section are public and available at https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/. For normalized real data plots, we experiment with ten random permutations. For sketched real data plots, we experiment ten random sketches. Median is plotted with 0.2/0.8 quantile shaded. We interpolate over $x-$axis whenever x ticks vary for different trials. We run all experiments on google cloud n1-standard-8 machine. One-hot embedding is used to transfer classification labels to regression labels when classification datasets are used for regression tasks. Code for experiments is included in the submission. According to our simulation results, optimal shrinkage helps speeding up both second-order optimization algorithms and sketching based algorithms significantly.

5.1. Estimation of the Effective Dimension

Our optimal shrinkage method requires the knowledge of the effective dimension d_λ of the true covariance matrix. In Algorithms 1 and 2, we employ the empirical effective dimension available at each worker as an approximation to the true effective dimension. Although this is a heuristic

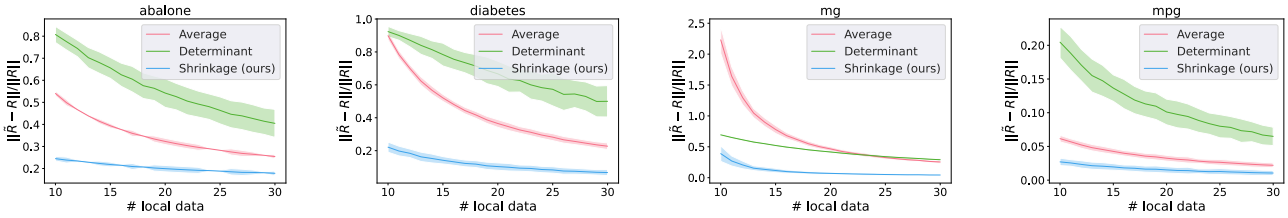

Figure 2: Experiments with real data in covariance resolvent estimation. The dataset is split evenly to each agent. We let $\lambda = 0.001$ and $\Sigma = \frac{1}{n}A^TA$. The relative matrix spectral norm difference between true covariance resolvent R and estimated covariance resolvent $\tilde{R}$ is plotted, see Section 5.2 for details.

METHOD	SKETCH METHOD	SKETCH SIZE	NON-ASYMPTOTIC BIAS	ASYMPTOTIC BIAS
IHS	GAUSSIAN PROJECTION	$\forall k$	SEE (8)	$\geq y\sigma_{\min}/\sigma^2_{\max}$
IHS WITH OPTIMAL SHRINKAGE	GAUSSIAN PROJECTION	$k \geq d_\lambda$	$\leq \mathcal{O}\left(\frac{1}{\sqrt{m}}\right)$	0

Table 2: Bias of estimation of Hessian inverse in IHS for regularized quadratic loss under asymptotic/non-asymptotic setting. For any real positive λ, d_λ denoting the effective dimension of A^TA/n. The asymptotic bias for the IHS method is for $\lambda \in o(1)$, see Theorem 2.5 for details. The non-asymptotic bias for IHS with optimal shrinkage method is for isotropic covariance, see Section 2.1.1 for details.

to approximate effective dimension in Theorem 2.2, our numerical results show that this approach works extremely well. Alternatively, the effective dimension can be estimated from a sketch of the data. We illustrate the effectiveness of this approach in the simulation for Iterative Hessian Sketch method in Section 5.4.

5.2. Experiments on Covariance Resolvent Estimation

In order to show that the shrinkage-based covariance resolvent estimation method studied in Section 2 helps improve the accuracy for covariance resolvent estimation compared to classic averaging method and determinantal averaging methods discussed in the introduction, we include simulation results for covariance resolvent esimation. Specifically, we let Σ denote the covariance matrix, we plot the relative matrix spectral norm difference $\|\tilde{R} - R\|_2/\|R\|_2$ where $R = (\Sigma + \lambda I)^{-1}$ is the resolvent of the true covariance matrix. The expressions for computing the estimated covariance resolvent $\tilde{R}$ with different methods are given in Appendix C.1.1. Since we require data rows to be i.i.d. with mean zero for the optimal shrinkage formula to hold, we standardize datasets by removing the mean and scaling to unit variance.

Figure 2 shows the relative matrix spectral norm difference between R and $\tilde{R}$ plotted over different number of local data, which suggests that our shrinkage method provides a more accurate estimate of the resolvent of covariance matrices than the naive averaging method and determinantal averaging over all the datasets we have tested on. The improvement is more pronounced when the local data size is small. Since the determinantal averaging method is also unbiased, this result also suggests that our shrinkage method does not need a large number of distributed agents to achieve an accurate estimation compared to the determinantal averaging. For simulations on additional datasets, see Appendix C.1.3. We also experiment with synthetic data and sketched real data in Appendix C.1.2 and Appendix C.1.4. We include the simulation results for the the small regularizer regime discussed in Section 2.1.2 in Appendix C.1.5.

5.3. Experiments on Distributed Second-Order Optimization

In this subsection, we include simulation results for distributed Newton's method and an inexact version where each Newton's step solved by the distributed preconditioned conjugate gradient method (see Section 3 for a description). We implement distributed line search for choosing step sizes in all methods. Datasets are standardized by removing the mean and scaling to unit variance. Due to limited space, we present additional results on different datasets in Appendix C.2.1, the inexact Newton's method applied to ridge regression in Appendix C.2.2 and experiments on logistic regression in Appendix C.2.3.

Figure 3 shows simulation results for distributed Newton's method applied to ridge regression. We compare with DANE and GIANT (see Section 1.1 for a discussion of these two methods). *Determinant* stands for the determinantal averaging method (see also Section 1.1 for details). Note DANE and GIANT coincides when regularized quadratic loss is minimized and both methods are simply taking the

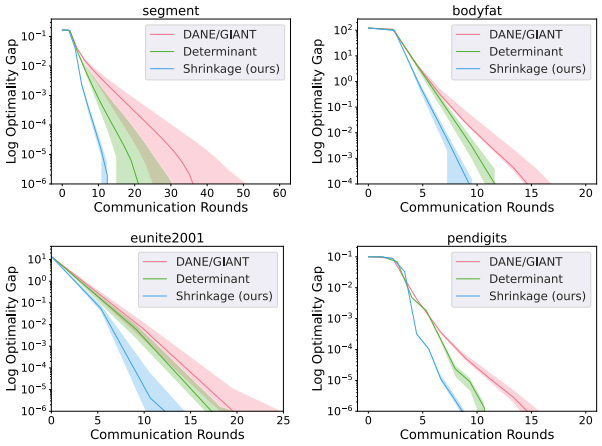

Figure 3: Experiments with real data for distributed Newton's method applied to ridge regression. Line search is used in all methods to determine the step sizes. Number of total samples is rounded down to a multiple of the number of agents and split evenly to each agent. We let $m = 100, \lambda = 0.1$ for segment, $m = 20, \lambda = 0.05$ for bodyfat, $m = 20, \lambda = 0.5$ for eunite2001, $m = 100, \lambda = 0.01$ for pendigits, where λ is the regularization parameter and m denotes the number of agents.

average of local descent direction as the global step, while the determinant averaging method adds a bias correction. The plots suggest that distributed Newton's method with optimal shrinkage achieves better log optimality gap within fewer communication rounds than other methods, which reveals that our shrinkage method is approximating the Hessian inverse more accurately. The ability of bias correction of determinantal averaging method can also be seen in the plots, but it is less effective compared to our approach.

5.4. Experiments on Iterative Hessian Sketch

In this subsection, we consider the Iterative Hessian Sketch paradigm, which can be seen as a stochastic Newton's method and incorporate our optimal shrinkage. This method is discussed in Section 4. We consider the ridge regression problem defined in Section 3.1. We use line search for the step size. Since the sketching matrix is generated from a random Gaussian ensemble, our required assumptions for the optimal shrinkage formula holds. We use the effective dimension of the sketched data as an approximation. We include additional simulation results with the effective dimension of the true covariance in Appendix C.2.4.

Figure 4 presents simulation results for the IHS method and IHS with optimal shrinkage. Although an inexact effective dimension is used for practicality, Iterative Hessian Sketch equipped with shrinkage still leads to significant speedup in the convergence rate, which once more confirms our shrinkage formula's ability for bias correction in Hessian inverse estimation.

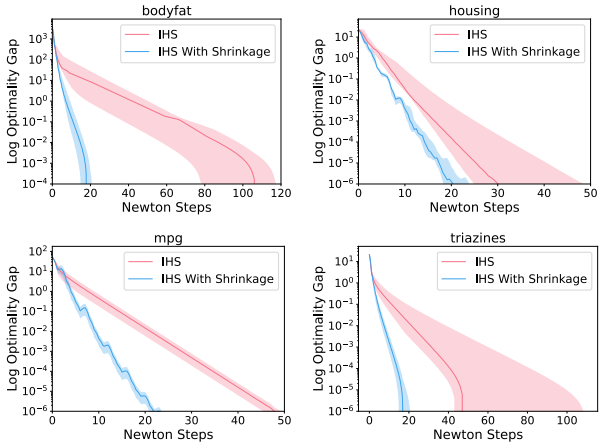

Figure 4: Experiments with real data for Iterative Hessian Sketch method applied to ridge regression. Line search is used to determine the step-sizes. We let $\lambda = 0.01$ for bodyfat, housing, mpg and $\lambda = 0.001$ for triazines; $m = 100$ for bodyfat, $m = 50$ for housing, $m = 30$ for mpg, $m = 300$ for triazines where λ denotes the regularization parameter and m denotes the sketch size.

6. Conclusion

In this work, we addressed bias correction in distributed second-order optimization and sketching based methods. Specifically, both types of algorithms require accurate estimation of a Hessian inverse. When either data can be modeled random or data sketching is used, this problem amounts to estimation of the resolvent of appropriately defined covariance matrix. We studied an asymptotically unbiased estimator for this resolvent, characterized its convergence rate, and leveraged it in Hessian inversion bias reduction, where significant convergence speedups are observed in real and synthetic datasets. One limitation of our theory is the need for prior knowledge of the effective dimension. Despite this limitation, we have shown that empirical approximations of the effective dimension yield highly accurate results. There exist other approaches to estimate the effective dimension, including trace estimation methods (Meyer et al., 2020), which can be used to further improve our scheme.

Acknowledgements

This work was supported in part by the National Science Foundation (NSF) under Grant ECCS-2037304 and Grant DMS-2134248; in part by the NSF CAREER Award under Grant CCF-2236829; in part by the U.S. Army Research Office Early Career Award under Grant W911NF-21-1-0242; in part by the Stanford Precourt Institute; and in part by the ACCESS—AI Chip Center for Emerging Smart Systems through InnoHK, Hong Kong, SAR.

References

Bai, Z. and Silverstein, J. W. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

Bartan, B. and Pilanci, M. Distributed sketching for randomized optimization: Exact characterization, concentration and lower bounds, 2022. URL https://arxiv.org/abs/2203.09755.

Bloemendal, A., Erdos, L., Knowles, A., Yau, H.-T., and Yin, J. Isotropic local laws for sample covariance and generalized wigner matrices. *Electronic Journal of Probability [electronic only]*, 19, 08 2013. doi: 10.1214/EJP.v19-3054.

Bodnar, O., Bodnar, T., and Parolya, N. Recent advances in shrinkage-based high-dimensional inference. *Journal of Multivariate Analysis*, 188:104826, 2022. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2021.104826. URL https://www.sciencedirect.com/science/article/pii/S0047259X21001044. 50th Anniversary Jubilee Edition.

Bodnar, T., Gupta, A. K., and Parolya, N. On the strong convergence of the optimal linear shrinkage estimator for large dimensional covariance matrix. *Journal of Multivariate Analysis*, 132:215–228, nov 2014. doi: 10.1016/j.jmva.2014.08.006. URL https://doi.org/10.1016%2Fj.jmva.2014.08.006.

Bodnar, T., Gupta, A., and Parolya, N. Direct shrinkage estimation of large dimensional precision matrix. *Journal of Multivariate Analysis*, 10 2015. doi: 10.1016/j.jmva.2015.09.010.

Bodnar, T., Gupta, A. K., and Parolya, N. Direct shrinkage estimation of large dimensional precision matrix. *J. Multivar. Anal.*, 146:223–236, 2016.

Bodnar, T., Dmytriv, S., Parolya, N., and Schmid, W. Tests for the weights of the global minimum variance portfolio in a high-dimensional setting. *IEEE Transactions on Signal Processing*, 67(17):4479–4493, 2019. doi: 10.1109/TSP.2019.2929964.

Dereziński, M. and Mahoney, M. W. Distributed estimation of the inverse hessian by determinantal averaging, 2019.

Dereziński, M., Bartan, B., Pilanci, M., and Mahoney, M. W. Debiasing distributed second order optimization with surrogate sketching and scaled regularization, 2020. URL https://arxiv.org/abs/2007.01327.

Dobriban, E. and Sheng, Y. Distributed linear regression by averaging, 2018. URL https://arxiv.org/abs/1810.00412.

Dobriban, E. and Sheng, Y. Wonder: Weighted one-shot distributed ridge regression in high dimensions, 2019. URL https://arxiv.org/abs/1903.09321.

Erdős, L., Götze, F., and Guionnet, A. Random matrices. *Oberwolfach Reports*, 16(4):3459–3527, 2020.

Fleermann, M. and Heiny, J. High-dimensional sample covariance matrices with curie-weiss entries. *arXiv preprint arXiv:1910.12332*, 2019.

Lacotte, J. and Pilanci, M. Fast convex quadratic optimization solvers with adaptive sketching-based preconditioners. *ArXiv*, abs/2104.14101, 2021.

Lacotte, J., Wang, Y., and Pilanci, M. Adaptive newton sketch: Linear-time optimization with quadratic convergence and effective hessian dimensionality. In *International Conference on Machine Learning*, pp. 5926–5936. PMLR, 2021.

Ledoit, O. and Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411, 2004.

Ma, C., Smith, V., Jaggi, M., Jordan, M., Richtarik, P., and Takac, M. Adding vs. averaging in distributed primal-dual optimization. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1973–1982, Lille, France, 07–09 Jul 2015. PMLR.

Meyer, R. A., Musco, C., Musco, C., and Woodruff, D. P. Hutch++: Optimal stochastic trace estimation, 2020. URL https://arxiv.org/abs/2010.09649.

Reddi, S. J., Konečný, J., Richtárik, P., Póczós, B., and Smola, A. Aide: Fast and communication efficient distributed optimization, 2016.

Serdobolskii, V. *Multiparametric Statistics*. Elsevier, 01 2008. doi: 10.1016/B978-0-444-53049-3.X5001-2.

Shamir, O., Srebro, N., and Zhang, T. Communication efficient distributed optimization using an approximate newton-type method, 2013.

Wang, S., Roosta-Khorasani, F., Xu, P., and Mahoney, M. W. Giant: Globally improved approximate newton method for distributed optimization, 2017.

Xiao, Z., Wang, J., Geng, L., Zhang, F., and Tong, J. On the robustness of covariance matrix shrinkage-based robust adaptive beamforming. In *Proceedings of the 2018 International Conference on Electronics and Electrical Engineering Technology*, EEET '18, pp. 196–201, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450365413. doi: 10.1145/

3277453.3277490. URL https://doi.org/10.1145/3277453.3277490.

Zhang, Y. and Lin, X. Disco: Distributed optimization for self-concordant empirical loss. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 362–370, Lille, France, 07–09 Jul 2015. PMLR.

Zhang, Y., Duchi, J. C., and Wainwright, M. Comunication-efficient algorithms for statistical optimization, 2012. URL https://arxiv.org/abs/1209.4129.

A. Proofs in Section 2

A.1. Technical Lemmas

Lemma A.1. *For any $z < 0, z \in \mathbb{R}$, if $F_{\Sigma_n} \to F_\Sigma$ almost everywhere, $M < \infty$, $\nu \to 0$. Then $\lim_{n,d\to\infty} \frac{1}{n}\mathbb{E}\left[tr\left(I - z\hat{\Sigma}_n\right)^{-1}\right]$ exists.*

Proof. Denote $\hat{\Sigma}_n' = \hat{\Sigma}_n - \frac{1}{n}x_n x_n^T, \psi_n = \frac{1}{n}x_n^T(I - z\hat{\Sigma}_n)^{-1}x_n$. By $y < 1$, without loss of generality consider the regime $d < n$ for all n. Denote $s_0(z) = 1 - \frac{d}{n} + \frac{1}{n}\mathbb{E}\left[\text{tr}\left(I - z\hat{\Sigma}_n\right)^{-1}\right]$, then $s_0(z) \in (0, 1]$. Follow Theorem 3.1 in book (Serdobolskii, 2008),

$$(I - zs_0(z)\Sigma_n)\,\mathbb{E}\left[\left(I - z\hat{\Sigma}_n\right)^{-1}\right] = I + \Omega$$

where

$$\Omega = z^2\mathbb{E}\left[\left(I - z\hat{\Sigma}_n'\right)^{-1} x_n x_n^T\left(\psi_n - \mathbb{E}[\psi_n]\right)\right] + zs_0(z)\mathbb{E}\left[\left(I - z\hat{\Sigma}_n'\right)^{-1} - \left(I - z\hat{\Sigma}_n\right)^{-1}\right]\Sigma_n$$

and

$$\|\Omega\| \le \frac{Mz^2}{n} + \sqrt{5M}z^2\sqrt{\frac{M^2d^2z^2}{n^3} + \frac{d^2\nu}{n^2}}$$

Thus,

$$\mathbb{E}\left[\left(I - z\hat{\Sigma}_n\right)^{-1}\right] = (I - zs_0(z)\Sigma_n)^{-1} + (I - zs_0(z)\Sigma_n)^{-1}\,\Omega$$

Since M bounded and ν diminishing,

$$\lim_{n,d\to\infty}\left\|\mathbb{E}\left[\left(I - z\hat{\Sigma}_n\right)^{-1}\right] - (I - zs_0(z)\Sigma_n)^{-1}\right\| = 0$$

which indicates

$$\lim_{n,d\to\infty}\left(\frac{1}{n}\mathbb{E}\left[\text{tr}\left(\left(I - z\hat{\Sigma}_n\right)^{-1}\right)\right] - \frac{1}{n}\text{tr}\left((I - zs_0(z)\Sigma_n)^{-1}\right)\right) = 0$$

Denote $e_n = \frac{1}{n}\mathbb{E}\left[\text{tr}\left(\left(I - z\hat{\Sigma}_n\right)^{-1}\right)\right], z_n = z(1 - \frac{d}{n})$,

$$\lim_{n,d\to\infty}\left(e_n - \frac{1}{n}\text{tr}\left((I - (z_n + ze_n)\Sigma_n)^{-1}\right)\right) = 0$$

Since $F_{\Sigma_n} \to F_\Sigma$ abd $\frac{d}{n} \to y$, denote the Stieltjes transform as $m_\mu(z) = \int_R \frac{1}{x-z}\mu(dx)$, define

$$f(e_n) = -\frac{y}{(z(1-y) + ze_n)}m_\Sigma\left(\frac{1}{z(1-y) + ze_n}\right)$$

with $f(e_n)$ monotone decreasing and

$$\lim_{n,d\to\infty}(e_n - f(e_n)) = 0$$

which indicates that for any $\epsilon > 0$, there exists $N > 0$ such that for any $n_1, n_2 > N$, $|(e_{n_1} - f(e_{n_1})) - (e_{n_2} - f(e_{n_2}))| < \epsilon$. Assume $\lim_{n,d\to\infty} e_n$ doesn't exist, then there exists $l_1, l_2 > N$ and $|e_{l_1} - e_{l_2}| > \epsilon$. Without loss of generality assume $e_{l_1} > e_{l_2}$. But then

$$|(e_{l_1} - f(e_{l_1})) - (e_{l_2} - f(e_{l_2}))| = |(e_{l_1} - e_{l_2}) - (f(e_{l_1}) - f(e_{l_2}))| > \epsilon$$

Contradiction. Therefore $\lim_{n,d\to\infty} e_n = \lim_{n,d\to\infty} \frac{1}{n}\mathbb{E}\left[\text{tr}\left(I - z\hat{\Sigma}_n\right)^{-1}\right]$ exists. □

Lemma A.2. *If $F_{\Sigma_n}(u) \to F_\Sigma(u)$ almost surely for almost every $u > 0$, $\lim_{n,d\to\infty} \frac{d_\lambda^n}{n}$ exists for any $\lambda > 0$.*

Proof.

$$\begin{aligned}\frac{d_\lambda^n}{n} &= \frac{\text{tr}(\Sigma_n(\Sigma_n+\lambda I)^{-1})}{n} \\ &= \frac{\text{tr}(I-\lambda(\Sigma_n+\lambda I)^{-1})}{n} \\ &= \frac{d}{n} - \frac{\lambda}{n}\text{tr}(\Sigma_n+\lambda I)^{-1}\end{aligned}$$

Denote the Stieltjes transform as $m_\mu(z) = \int_R \frac{1}{x-z}\mu(dx)$. Since $F_{\Sigma_n} \to F_\Sigma$ almost everywhere, $\lim_{n,d\to\infty}\frac{1}{n}\text{tr}(\Sigma_n+\lambda I)^{-1}$ exists and

$$\lim_{n,d\to\infty}\frac{1}{n}\text{tr}(\Sigma_n+\lambda I)^{-1} = ym_{F_\Sigma}(-\lambda) \in (0, \frac{y}{\lambda}]$$

Thus

$$\lim_{n,d\to\infty}\frac{d_\lambda^n}{n} = y - \lambda y m_{F_\Sigma}(-\lambda) \in [0, y) \text{ exists}$$

□

Lemma A.3. *If $y > 0$, $F_{\Sigma_n}(u) \to F_\Sigma(u)$ almost surely for almost every $u > 0$, eigenvalues of each Σ_n are located on a segment $[\sigma_{\min}, \sigma_{\max}]$. For any $z < 0, s > 0, z, s \in \mathbb{R}$, the fixed point equation in s*

$$s = 1 + \lim_{n,d\to\infty}\frac{1}{n}tr\left(zs\Sigma_n\left(I - zs\Sigma_n\right)^{-1}\right) \tag{1}$$

has at most one non-negative real solution.

Proof. First note zero is not a solution to (1). Assume (1) has two positive real solutions s_1 and s_2 such that $s_1 \neq s_2$. Without loss of generality assume $s_1 > s_2$. Then

$$s_1 = 1 - y + \lim_{n,d\to\infty}\frac{\text{tr}(I-zs_1\Sigma_n)^{-1}}{n} = 1 - y + \lim_{n,d\to\infty}\frac{\text{tr}(I-zs_2\Sigma_n)^{-1}}{n} = s_2$$

Therefore,

$$\lim_{n,d\to\infty}\frac{1}{n}\left(\text{tr}(I-zs_1\Sigma_n)^{-1} - \text{tr}(I-zs_2\Sigma_n)^{-1}\right) = 0$$

Apply resolvent identity,

$$\lim_{n,d\to\infty}\frac{z(s_1-s_2)}{n}\text{tr}\left(\Sigma_n(I-zs_2\Sigma_n)^{-1}(I-zs_1\Sigma_n)^{-1}\right) = 0$$

Equivalently,

$$\frac{1}{z}\left(\frac{1}{s_2}-\frac{1}{s_1}\right)\lim_{n,d\to\infty}\frac{1}{n}\sum_{i=1}^{d}\frac{\sigma_{ni}}{\left(\sigma_{ni}-\frac{1}{zs_2}\right)\left(\sigma_{ni}-\frac{1}{zs_1}\right)} = 0$$

where σ_{ni} is the ith eigenvalue of Σ_n. But

$$\frac{1}{z}\left(\frac{1}{s_2}-\frac{1}{s_1}\right)\lim_{n,d\to\infty}\frac{1}{n}\sum_{i=1}^{d}\frac{\sigma_{ni}}{\left(\sigma_{ni}-\frac{1}{zs_2}\right)\left(\sigma_{ni}-\frac{1}{zs_1}\right)} \leq \frac{y}{z}\left(\frac{1}{s_2}-\frac{1}{s_1}\right)\frac{\sigma_{\min}}{\left(\sigma_{\max}-\frac{1}{zs_2}\right)\left(\sigma_{\max}-\frac{1}{zs_1}\right)}$$

Thus, $s_1 = s_2$. □

Lemma A.4. *With $\Sigma_n = I$. Fix a small $\omega \in (0,1)$. Define the domain $\mathcal{D} = \{z = u + vi \in \mathbb{C} : \sqrt{y} + \frac{1}{\sqrt{y}} - 2 + |\frac{u}{\sqrt{y}}| \leq \omega^{-1}, d^{\omega-1} \leq \frac{v}{\sqrt{y}} \leq \omega^{-1}, |z| \geq \omega\sqrt{y}\}$. Then*

$$\left|\mathbb{E}\left[\frac{1}{n}tr(\hat{\Sigma}_n - zI)^{-1}\right] - \lim_{n,d\to\infty}\mathbb{E}\left[\frac{1}{n}tr(\hat{\Sigma}_n - zI)^{-1}\right]\right| \leq \frac{1}{v\sqrt{nd}}$$

for any $z \in \mathcal{D}$.

Proof. Lemma A.4 is a direct corollary from Theorem 2.4 in (Bloemendal et al., 2013). For sake of complementness, we restate the theorem here again together with the derivation of Lemma A.4.

Lemma A.5. *(Theorem 2.4 in (Bloemendal et al., 2013)) With $\Sigma_n = \frac{1}{\sqrt{y}}I$. Fix a small $\omega \in (0,1)$. Define the domain $\mathcal{D} = \{z = u + vi \in \mathbb{C} : \sqrt{y} + \frac{1}{\sqrt{y}} - 2 + |u| \leq \omega^{-1}, d^{\omega-1} \leq v \leq \omega^{-1}, |z| \geq \omega\}$. Then*

$$\left|\frac{1}{d}tr(\hat{\Sigma}_n - zI)^{-1} - m_\phi\right| \prec \frac{1}{dv}$$

uniformly for $z \in \mathcal{D}$, where $\prec$ denote stochastic dominance and m_ϕ is Stieltjes transform of the Marchenko-Pastur law with variance $\frac{1}{\sqrt{y}}$.

Note given $\left|\frac{1}{d}\text{tr}(\hat{\Sigma}_n - zI)^{-1} - m_\phi\right| \prec \frac{1}{dv}$, since $\frac{1}{d}\text{tr}(\hat{\Sigma}_n - zI)^{-1} \leq z$ for each n, by dominated convergence theorem, we can bound

$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{1}{d}\text{tr}(\hat{\Sigma}_n - zI)^{-1}\right] - \lim_{n,d\to\infty}\mathbb{E}\left[\frac{1}{d}\text{tr}(\hat{\Sigma}_n - zI)^{-1}\right]\right| = \left|\mathbb{E}\left[\frac{1}{d}\text{tr}(\hat{\Sigma}_n - zI)^{-1}\right] - \mathbb{E}[m_\phi]\right| \\
&\leq \mathbb{E}\left[\left|\frac{1}{d}\text{tr}(\hat{\Sigma}_n - zI)^{-1} - m_\phi\right|\right] \\
&< 2zd^{-D} + \frac{d^\epsilon}{dv}(1 - d^{-D})
\end{aligned}$$

for any $\epsilon > 0, D > 0$. Since ϵ can be taken arbitrarily close to 0 and D can be taken arbitrarily large,

$$\left|\mathbb{E}\left[\frac{1}{d}\text{tr}(\hat{\Sigma}_n - zI)^{-1}\right] - \lim_{n,d\to\infty}\mathbb{E}\left[\frac{1}{d}\text{tr}(\hat{\Sigma}_n - zI)^{-1}\right]\right| \leq \frac{1}{dv}$$

With some algebra, we get Lemma A.4 □

Lemma A.6. *Under Assumption 2.1, for any $z > 0, \epsilon_n > 0$, if*

$$\left|\mathbb{E}\left[\frac{tr(\hat{\Sigma}_n + (z - \epsilon_n i)I)^{-1}}{n}\right] - \lim_{n,d\to\infty}\mathbb{E}\left[\frac{tr(\hat{\Sigma}_n + (z - \epsilon_n i)I)^{-1}}{n}\right]\right| \leq \Omega(n, \epsilon_n)$$

then

$$\left|\mathbb{E}\left[\frac{tr(\hat{\Sigma}_n + zI)^{-1}}{n}\right] - \lim_{n,d\to\infty}\mathbb{E}\left[\frac{tr(\hat{\Sigma}_n + zI)^{-1}}{n}\right]\right| \leq \frac{2d|\epsilon_n|}{nz^2} + \Omega(n, \epsilon_n)$$

Proof. by resolvent identity,

$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{\text{tr}(\hat{\Sigma}_n + zI)^{-1}}{n}\right] - \mathbb{E}\left[\frac{\text{tr}(\hat{\Sigma}_n + (z - \epsilon_n i)I)^{-1}}{n}\right]\right| \\
&= \left|\frac{1}{n}\mathbb{E}\left[\text{tr}\left((\hat{\Sigma}_n + zI)^{-1}(-\epsilon_n iI)(\hat{\Sigma}_n + (z - \epsilon_n i)I)^{-1}\right)\right]\right| \\
&\leq \frac{d|\epsilon_n|}{nz^2}
\end{aligned}$$

Since the above inequality holds for each n,

$$\left|\lim_{n,d\to\infty}\mathbb{E}\left[\frac{\text{tr}(\hat{\Sigma}_n + zI)^{-1}}{n}\right] - \lim_{n,d\to\infty}\mathbb{E}\left[\frac{\text{tr}(\hat{\Sigma}_n + (z - \epsilon_n i)I)^{-1}}{n}\right]\right| \leq \frac{|\epsilon_n|d}{nz^2}$$

Thus,

$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{\mathrm{tr}(\hat{\Sigma}_n+zI)^{-1}}{n}\right]-\lim_{n,d\to\infty}\mathbb{E}\left[\frac{\mathrm{tr}(\Sigma_n+zI)^{-1}}{n}\right]\right| \\
\leq&\left|\left(\mathbb{E}\left[\frac{\mathrm{tr}(\hat{\Sigma}_n+zI)^{-1}}{n}\right]-\mathbb{E}\left[\frac{\mathrm{tr}(\hat{\Sigma}_n+(z-\epsilon_n i)I)^{-1}}{n}\right]\right)\right|+ \\
&\left|\left(\lim_{n,d\to\infty}\mathbb{E}\left[\frac{\mathrm{tr}(\hat{\Sigma}_n+zI)^{-1}}{n}\right]-\lim_{n,d\to\infty}\mathbb{E}\left[\frac{\mathrm{tr}(\hat{\Sigma}_n+(z-\epsilon_n i)I)^{-1}}{n}\right]\right)\right|+ \\
&\left|\left(\mathbb{E}\left[\frac{\mathrm{tr}(\hat{\Sigma}_n+(z-\epsilon_n i)I)^{-1}}{n}\right]-\lim_{n,d\to\infty}\mathbb{E}\left[\frac{\mathrm{tr}(\hat{\Sigma}_n+(z-\epsilon_n i)I)^{-1}}{n}\right]\right)\right| \\
\leq&\frac{2d|\epsilon_n|}{nz^2}+\Omega(n,\epsilon_n)
\end{aligned}$$

□

Lemma A.7. *Under Assumption 2.1, for any $\lambda>0$, denote $\overline{d_\lambda}=\liminf_{n,d\to\infty}n^{-1}d_\lambda^n$, then*

$$\mathbb{E}\left[\left(\frac{1}{1-\overline{d_\lambda}}\hat{\Sigma}_n+\lambda I\right)^{-1}\right]=(\Sigma_n+\lambda I)^{-1}+\Omega_2$$

where $\|\Omega_2\|\to 0$ as $n,d\to\infty, d/n\to y\in[0,1)$.

Proof. The existence of $\overline{d_\lambda}=\lim_{n,d\to\infty}\frac{d_\lambda^n}{n}$ has been established in Lemma A.2. Take any $z<0, z\in\mathbb{R}$, by Lemma A.1, $\lim_{n,d\to\infty}\mathbb{E}\left[\frac{\mathrm{tr}(I-z\hat{\Sigma}_n)^{-1}}{n}\right]$ exists. Denote $s(z)=1-y+\lim_{n,d\to\infty}\mathbb{E}\left[\frac{\mathrm{tr}(I-z\hat{\Sigma}_n)^{-1}}{n}\right]$, by Lemma A.1 again, $s(z)=\lim_{n,d\to\infty}s_0(z)\in[0,1]$. We can compute

$$(I-zs(z)\Sigma_n)\mathbb{E}\left[(I-z\hat{\Sigma}_n)^{-1}\right]=I+\Omega+\Omega' \tag{2}$$

with $\Omega'=z\left(s_0(z)-s(z)\right)\Sigma_n\mathbb{E}\left[(I-z\hat{\Sigma}_n)^{-1}\right]$. Thus,

$$\mathbb{E}\left[(I-z\hat{\Sigma}_n)^{-1}\right]=(I-zs(z)\Sigma_n)^{-1}+(I-zs(z)\Sigma_n)^{-1}(\Omega+\Omega') \tag{3}$$

We can bound

$$\begin{aligned}
\|(I-zs(z)\Sigma_n)^{-1}(\Omega+\Omega')\|&\leq\|\Omega+\Omega'\| \\
&\leq\|\Omega\|+\|\Omega'\| \\
&\leq\frac{Mz^2}{n}+\sqrt{5M}z^2\sqrt{\frac{M^2d^2z^2}{n^3}+\frac{d^2\nu}{n^2}}+\sigma_{\max}|z||s_0(z)-s(z)|
\end{aligned}$$

Since M stays bounded, $\nu\to 0$ and $s(z)=\lim_{n,d\to\infty}s_0(z)$,

$$\lim_{n,d\to\infty}\left\|\mathbb{E}\left[(I-z\hat{\Sigma}_n)^{-1}\right]-(I-zs(z)\Sigma_n)^{-1}\right\|=0$$

which indicates

$$\lim_{n,d\to\infty}\left(\frac{1}{n}\mathbb{E}\left[\mathrm{tr}(I-z\hat{\Sigma}_n)^{-1}\right]-\frac{1}{n}\mathrm{tr}(I-zs(z)\Sigma_n)^{-1}\right)=0$$

by Lemma A.1, $\lim_{n,d\to\infty}\frac{1}{n}\mathbb{E}\left[\mathrm{tr}(I-z\hat{\Sigma}_n)^{-1}\right]$ exists and therefore,

$$\lim_{n,d\to\infty}\left(\frac{1}{n}\mathbb{E}\left[\mathrm{tr}(I-z\hat{\Sigma}_n)^{-1}\right]\right)=\lim_{n,d\to\infty}\left(\frac{1}{n}\mathrm{tr}(I-zs(z)\Sigma_n)^{-1}\right) \tag{4}$$

substitute (4) back into expression for $s(z)$, we get

$$\begin{aligned} s(z) &= 1 - y + \lim_{n,d\to\infty} \frac{\text{tr}(I - zs(z)\Sigma_n)^{-1}}{n} \\ &= 1 - \lim_{n,d\to\infty} \left(\frac{d}{n} - \frac{\text{tr}(I - zs(z)\Sigma_n)^{-1}}{n} \right) \\ &= 1 + \lim_{n,d\to\infty} \frac{\text{tr}(zs(z)\Sigma_n(I - zs(z)\Sigma_n)^{-1})}{n} \end{aligned} \tag{5}$$

Set $z = -\frac{1}{\lambda(1-\overline{d_\lambda})}$. Since $\frac{d_\lambda^n}{n} \leq \frac{d}{n}$ for each n, $\overline{d_\lambda} \leq y < 1$, and therefore $z < 0$. Note $s(z) = 1 - \overline{d_\lambda}$ satisfies (5). When $y > 0$, by Lemma A.3, we conclude $s(z) = 1 - \overline{d_\lambda}$. When $y = 0, \overline{d_\lambda} = 0$. We get $s(z) = 1 + \lim_{n,d\to\infty} \frac{\text{tr}(I - zs(z)\Sigma_n)^{-1}}{n} \geq 1$ from (5), but since $s(z) \in [0, 1]$, $s(z) = 1 = 1 - \overline{d_\lambda}$. Thus, $s(z) = 1 - \overline{d_\lambda}$ for $y \in [0, 1)$. Substitute $(z = -\frac{1}{\lambda(1-\overline{d_\lambda})}, s(z) = 1 - \overline{d_\lambda})$ back into (3), we get

$$\begin{aligned} \mathbb{E}\left[\left(\frac{1}{1-\overline{d_\lambda}} \hat{\Sigma}_n + \lambda I \right)^{-1} \right] &= (\Sigma_n + \lambda I)^{-1} + (\Sigma_n + \lambda I)^{-1}(\Omega + \Omega') \\ &= (\Sigma_n + \lambda I)^{-1} + \Omega_2 \end{aligned}$$

with $\|\Omega_2\| = \|(\Sigma_n + \lambda I)^{-1}(\Omega + \Omega')\|$. Therefore, we can bound

$$\begin{aligned} \|\Omega_2\| &= \|(\Sigma_n + \lambda I)^{-1}(\Omega + \Omega')\| \\ &\leq \frac{1}{\lambda}\|\Omega + \Omega'\| \\ &\leq \frac{1}{\lambda}\left(\frac{Mz^2}{n} + \sqrt{5M} z^2 \sqrt{\frac{M^2 d^2 z^2}{n^3} + \frac{d^2\nu}{n^2}} + \sigma_{\max}|z|\,|s_0(z) - s(z)| \right) \\ &\leq \frac{1}{\lambda}\left(\frac{Mz^2}{n} + \sqrt{5M} z^2 \sqrt{\frac{M^2 d^2 z^2}{n^3} + \frac{d^2\nu}{n^2}} + \sigma_{\max}|z| \left| \left(y - \frac{d}{n} \right) + \left(\mathbb{E}\left[\frac{1}{n}\text{tr}\left(I - z\hat{\Sigma}_n \right)^{-1} \right] - \right. \right. \right. \\ &\quad \left. \left. \left. \lim_{n,d\to\infty} \mathbb{E}\left[\frac{1}{n}\text{tr}\left(I - z\hat{\Sigma}_n \right)^{-1} \right] \right) \right| \right) \end{aligned} \tag{6}$$

where $z = -\frac{1}{\lambda(1-\overline{d_\lambda})}$. Thus $\|\Omega_2\| \to \infty$ as $n, d \to \infty$. □

A.2. Proof of Theorem 2.2

Proof. Theorem 2.2 can be derived as a corollary of Lemma A.7. We include its derivation for sake of completeness. Build on proof of Theorem A.7,

$$\begin{aligned} \mathbb{E}\left[\left(\frac{1}{1-\frac{d_\lambda^n}{n}} \hat{\Sigma}_n + \lambda I \right)^{-1} \right] = (\Sigma_n + \lambda I)^{-1} + \left(\mathbb{E}\left[\left(\frac{1}{1-\frac{d_\lambda^n}{n}} \hat{\Sigma}_n + \lambda I \right)^{-1} \right] \right. \\ \left. -\mathbb{E}\left[\left(\frac{1}{1 - \lim_{n,d\to\infty} \frac{d_\lambda^n}{n}} \hat{\Sigma}_n + \lambda I \right)^{-1} \right] \right) + \Omega_2 \end{aligned}$$

Denote $\omega_n = \lim_{n,d\to\infty} \frac{d_\lambda^n}{n} - \frac{d_\lambda^n}{n}$, by resolvent identity,

$$\begin{aligned}
&\left\| \mathbb{E}\left[\left(\frac{1}{1-\frac{d_\lambda^n}{n}}\hat{\Sigma}_n + \lambda I\right)^{-1}\right] - \mathbb{E}\left[\left(\frac{1}{1-\lim_{n,d\to\infty}\frac{d_\lambda^n}{n}}\hat{\Sigma}_n + \lambda I\right)^{-1}\right]\right\| \\
&= \left\| \mathbb{E}\left[\left(\frac{1}{1-\frac{d_\lambda^n}{n}}\hat{\Sigma}_n + \lambda I\right)^{-1}\left(\frac{\omega_n}{\left(1-\lim_{n,d\to\infty}\frac{d_\lambda^n}{n}\right)\left(1-\frac{d_\lambda^n}{n}\right)}\hat{\Sigma}_n\right)\left(\frac{1}{1-\lim_{n,d\to\infty}\frac{d_\lambda^n}{n}}\hat{\Sigma}_n + \lambda I\right)^{-1}\right]\right\| \\
&\le \lambda^2 \frac{|\omega_n|}{\left(1-\lim_{n,d\to\infty}\frac{d_\lambda^n}{n}\right)\left(1-\frac{d_\lambda^n}{n}\right)} \|\mathbb{E}[\hat{\Sigma}_n]\| \\
&\le \frac{\lambda^2\sigma_{\max}|\omega_n|}{\left(1-\lim_{n,d\to\infty}\frac{d_\lambda^n}{n}\right)\left(1-\frac{d_\lambda^n}{n}\right)}
\end{aligned}$$

Since $\lim_{n,d\to\infty}\frac{d_\lambda^n}{n} \le \frac{y+1}{2} < 1$, therefore,

$$\left\| \mathbb{E}\left[\left(\frac{1}{1-\frac{d_\lambda^n}{n}}\hat{\Sigma}_n + \lambda I\right)^{-1}\right] - (\Sigma_n + \lambda I)^{-1}\right\| \le \frac{2\sigma_{\max}\lambda^2|\omega_n|}{(1-y)\left(1-\frac{d_\lambda^n}{n}\right)} + \|\Omega_2\| \tag{7}$$

with the right-hand side diminishing to 0. □

A.3. Proof of Theorem 2.3

Proof. When $\Sigma_n = I$, $\frac{d_\lambda^n}{n} = \frac{y}{\lambda+1}$ for each n and thus $\overline{d_\lambda} = \frac{y}{\lambda+1}$. Let $z = \lambda(1-\frac{y}{\lambda+1})$ and $\epsilon_n = \frac{1}{\sqrt{n}}$ in Lemma A.6, under Assumption 2.1 and with Lemma A.4,

$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{1}{n}\mathrm{tr}\left(\hat{\Sigma}_n + \lambda\left(1-\frac{y}{\lambda+1}\right)I\right)^{-1}\right] - \lim_{n,d\to\infty}\mathbb{E}\left[\frac{1}{n}\mathrm{tr}\left(\hat{\Sigma}_n + \lambda\left(1-\frac{y}{\lambda+1}\right)I\right)^{-1}\right]\right| \\
&\le \frac{2d\epsilon_n}{nz^2} + \frac{1}{\epsilon_n\sqrt{nd}} = \frac{2y}{z^2\sqrt{n}} + \frac{1}{\sqrt{d}}
\end{aligned}$$

Since $\|\Sigma_n\| = 1$ for each n, $\sigma_{\max} = 1$ and from (6), denote $z' = -\frac{1}{z}$,

$$\begin{aligned}
\|\Omega_2\| &\le \frac{1}{\lambda}\left(\frac{Mz'^2}{n} + \sqrt{5M}z'^2\sqrt{\frac{M^2d^2z'^2}{n^3} + \frac{d^2\nu}{n^2}} + \sigma_{\max}|z'|\,|s_0(z') - s(z')|\right) \\
&\le \frac{1}{\lambda}\left(\frac{Mz'^2}{n} + \sqrt{5M}z'^2\sqrt{\frac{M^2d^2z'^2}{n^3} + \frac{d^2\nu}{n^2}} + |z'|\left|\left(\mathbb{E}\left[\frac{1}{n}\mathrm{tr}\left(I - z'\hat{\Sigma}_n\right)^{-1}\right] - \right.\right.\right. \\
&\quad \left.\left.\left. \lim_{n,d\to\infty}\mathbb{E}\left[\frac{1}{n}\mathrm{tr}\left(I - z'\hat{\Sigma}_n\right)^{-1}\right]\right)\right|\right) \\
&\le \frac{1}{\lambda}\left(\frac{Mz'^2}{n} + \sqrt{5M}z'^2\sqrt{\frac{M^2d^2z'^2}{n^3} + \frac{d^2\nu}{n^2}} + \frac{1}{\sqrt{n}}\left(2yz'^2 + \frac{1}{\sqrt{y}}\right)\right) \\
&= \frac{1}{\lambda}\left(\frac{Mz'^2}{n} + z'^2y\sqrt{5M}\sqrt{\frac{M^2z'^2}{n} + \nu} + \frac{1}{\sqrt{n}}\left(2yz'^2 + \frac{1}{\sqrt{y}}\right)\right) \in \mathcal{O}\left(\frac{1}{\sqrt{n}} + \sqrt{\nu}\right)
\end{aligned}$$

In (7), since $|\omega_n| = 0$,

$$\left\| \mathbb{E}\left[\left(\frac{1}{1-\frac{d_\lambda^n}{n}}\hat{\Sigma}_n + \lambda I\right)^{-1}\right] - (\Sigma_n + \lambda I)^{-1}\right\| \le \|\Omega_2\| \in \mathcal{O}\left(\frac{1}{\sqrt{n}} + \sqrt{\nu}\right)$$

□

A.4. Proof of Theorem 2.4

Proof. Assume eigenvalues of $\hat{\Sigma}_n$ are no less than $\sigma > 0$. By resolvent identity,

$$\begin{aligned}
&\left\| \mathbb{E}\left[\left(\frac{1}{1-\frac{d_\epsilon^n}{n}}\hat{\Sigma}_n + \epsilon I\right)^{-1}\right] - \mathbb{E}\left[\left(\frac{1}{1-\frac{d}{n}}\hat{\Sigma}_n + \epsilon I\right)^{-1}\right] \right\| \\
&= \left\| \mathbb{E}\left[\epsilon\left(\frac{1}{1-\frac{d_\epsilon^n}{n}}\hat{\Sigma}_n + \epsilon I\right)^{-1}\left(\frac{\mathrm{tr}(\Sigma_n+\epsilon I)^{-1}}{n(1-\frac{d}{n})(1-\frac{d_\epsilon^n}{n})}\hat{\Sigma}_n\right)\left(\frac{1}{1-\frac{d}{n}}\hat{\Sigma}_n + \epsilon I\right)^{-1}\right] \right\| \\
&\le \frac{\epsilon}{\left(\frac{\sigma}{1-\frac{d_\epsilon^n}{n}}+\epsilon\right)^2}\left\| \mathbb{E}\left[\frac{\mathrm{tr}(\Sigma_n+\epsilon I)^{-1}}{n(1-\frac{d}{n})(1-\frac{d_\epsilon^n}{n})}\hat{\Sigma}_n\right] \right\| \\
&\le \frac{\epsilon}{\left(\frac{\sigma}{1-\frac{d_\epsilon^n}{n}}+\epsilon\right)^2}\cdot\frac{\sigma_{\max} d}{n(1-\frac{d}{n})^2(\sigma_{\min}+\epsilon)}
\end{aligned}$$

From Theorem 2.2, we know

$$\mathbb{E}\left[\left(\frac{1}{1-\frac{d}{n}}\hat{\Sigma}_n + \epsilon I\right)^{-1}\right] = (\Sigma_n+\epsilon I)^{-1} + \mathbb{E}\left[\left(\frac{1}{1-\frac{d}{n}}\hat{\Sigma}_n + \epsilon I\right)^{-1}\right] - \mathbb{E}\left[\left(\frac{1}{1-\frac{d_\epsilon^n}{n}}\hat{\Sigma} + \epsilon I\right)^{-1}\right] + \Omega_0$$

Therefore,

$$\lim_{n,d\to\infty,\epsilon\to 0}\left\| \mathbb{E}\left[\left(\frac{1}{1-\frac{d}{n}}\hat{\Sigma}_n + \epsilon I\right)^{-1}\right] - (\Sigma_n+\epsilon I)^{-1}\right\| = 0.$$

Furthermore, if Σ_n^{-1} exists for each n, by resolvent identity,

$$\|(\Sigma_n+\epsilon I)^{-1} - \Sigma_n^{-1}\| = \|\epsilon(\Sigma_n+\epsilon I)^{-1}\Sigma_n^{-1}\| \le \frac{\epsilon}{\sigma_{\min}^2}$$

Thus,

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{1-\frac{d}{n}}\hat{\Sigma}_n + \epsilon I\right)^{-1}\right] = \Sigma_n^{-1} + ((\Sigma_n+\epsilon I)^{-1} - \Sigma_n^{-1}) + \mathbb{E}\left[\left(\frac{1}{1-\frac{d}{n}}\hat{\Sigma}_n + \right.\right. \\
\left.\left.\epsilon I\right)^{-1}\right] - \mathbb{E}\left[\left(\frac{1}{1-\frac{d_\epsilon^n}{n}}\hat{\Sigma}_n + \epsilon I\right)^{-1}\right] + \Omega_0
\end{aligned}$$

and

$$\lim_{n,d\to\infty,\epsilon\to 0}\left\| \mathbb{E}\left[\left(\frac{1}{1-\frac{d}{n}}\hat{\Sigma}_n + \epsilon I\right)^{-1}\right] - \Sigma_n^{-1}\right\| = 0$$

□

A.5. Proof of Theorem 2.5

Proof. Built on equation (2) in proof of Lemma A.7, for any $z < 0, z \in \mathbb{R}$,

$$(I - z\Sigma_n)\mathbb{E}\left[\left(I - z\hat{\Sigma}_n\right)^{-1}\right] = (zs(z)\Sigma_n - z\Sigma_n)\,\mathbb{E}\left[\left(I - z\hat{\Sigma}_n\right)^{-1}\right] + I + \Omega + \Omega'$$

Take $z = -\frac{1}{\lambda}$, note $z < 0$, and thus

$$\mathbb{E}\left[\left(\hat{\Sigma}_n + \lambda I\right)^{-1}\right] = (\Sigma_n+\lambda I)^{-1} + (\Sigma_n+\lambda I)^{-1}\left[\left(1 - s\left(-\frac{1}{\lambda}\right)\right)\Sigma_n\mathbb{E}\left[(\hat{\Sigma}_n+\lambda I)^{-1}\right] + \Omega + \Omega'\right]$$

Therefore,

$$\begin{aligned}\left\|\mathbb{E}\left[\left(\hat{\Sigma}_n+\lambda I\right)^{-1}\right]-(\Sigma_n+\lambda I)^{-1}\right\| &= \left\|(\Sigma_n+\lambda I)^{-1}\left(1-s\left(-\frac{1}{\lambda}\right)\right)\Sigma_n\mathbb{E}\left[(\hat{\Sigma}_n+\lambda I)^{-1}\right]+\Omega_2\right\| \\ &\geq \frac{1}{\sigma_{\max}+\lambda}\left\|\left(1-s\left(-\frac{1}{\lambda}\right)\right)\Sigma_n\mathbb{E}\left[\left(\hat{\Sigma}_n+\lambda I\right)^{-1}\right]\right\|-\|\Omega_2\|\end{aligned} \tag{8}$$

Since $\left(1-s\left(-\frac{1}{\lambda}\right)\right)=y-\lambda\lim_{n,d\to\infty}\mathbb{E}\left[\frac{1}{n}\text{tr}\left(\hat{\Sigma}_n+\lambda I\right)^{-1}\right]$, with λ' satisfies $\lambda=\lambda'\left(1-\overline{d_{\lambda'}}\right)$, [4]

$$\begin{aligned}&\left\|\mathbb{E}\left[\left(\hat{\Sigma}_n+\lambda I\right)^{-1}\right]-(\Sigma_n+\lambda I)^{-1}\right\| \\ &\geq \frac{1}{\sigma_{\max}+\lambda}\left(\left|y-\lambda\lim_{n,d\to\infty}\mathbb{E}\left[\frac{1}{n}\text{tr}\left(\hat{\Sigma}_n+\lambda I\right)^{-1}\right]\right|\left\|\Sigma_n\mathbb{E}\left[\left(\hat{\Sigma}_n+\lambda I\right)^{-1}\right]\right\|\right)-\|\Omega_2\| \\ &\geq \frac{\sigma_{\min}}{\sigma_{\max}+\lambda}\left|y-\lambda\lim_{n,d\to\infty}\mathbb{E}\left[\frac{1}{n}\text{tr}\left(\hat{\Sigma}_n+\lambda I\right)^{-1}\right]\right|\left\|\mathbb{E}\left[\left(\hat{\Sigma}_n+\lambda I\right)^{-1}\right]\right\|-\|\Omega_2\| \\ &\geq \frac{\sigma_{\min}}{\sigma_{\max}+\lambda}\left|y-\lambda\lim_{n,d\to\infty}\mathbb{E}\left[\frac{1}{n}\text{tr}\left(\hat{\Sigma}_n+\lambda I\right)^{-1}\right]\right|\left\|\frac{\lambda'}{\lambda}(\Sigma_n+\lambda' I)^{-1}-\left(\frac{\lambda'}{\lambda}(\Sigma_n+\lambda' I)^{-1}-\mathbb{E}\left[\left(\hat{\Sigma}_n+\lambda I\right)^{-1}\right]\right)\right\| \\ &-\|\Omega_2\| \\ &\geq \frac{\sigma_{\min}}{\sigma_{\max}+\lambda}\left|y-\lambda\lim_{n,d\to\infty}\mathbb{E}\left[\frac{1}{n}\text{tr}\left(\hat{\Sigma}_n+\lambda I\right)^{-1}\right]\right|\left\|\frac{\lambda'}{\lambda}(\Sigma_n+\lambda' I)^{-1}\right\| \\ &-\frac{\sigma_{\min}}{\sigma_{\max}+\lambda}\left|y-\lambda\lim_{n,d\to\infty}\mathbb{E}\left[\frac{1}{n}\text{tr}\left(\hat{\Sigma}_n+\lambda I\right)^{-1}\right]\right|\left\|\frac{\lambda'}{\lambda}(\Sigma_n+\lambda' I)^{-1}-\mathbb{E}\left[\left(\hat{\Sigma}_n+\lambda I\right)^{-1}\right]\right\| \\ &-\|\Omega_2\|\end{aligned}$$

By Lemma A.7, $\lim_{n,d\to\infty}\|\Omega_2\|=0$, and also

$$\lim_{n,d\to\infty}\left\|\mathbb{E}\left[\left(\hat{\Sigma}_n+\lambda I\right)^{-1}\right]-\frac{\lambda'}{\lambda}(\Sigma_n+\lambda' I)^{-1}\right\|=0$$

thus

$$\lim_{n,d\to\infty}\mathbb{E}\left[\frac{1}{n}\text{tr}\left(\hat{\Sigma}_n+\lambda I\right)^{-1}\right]=\frac{\lambda'}{\lambda}\lim_{n,d\to\infty}\frac{1}{n}\text{tr}\left(\Sigma_n+\lambda' I\right)^{-1}$$

Therefore, since $\left|y-\lambda\lim_{n,d\to\infty}\mathbb{E}\left[\frac{1}{n}\text{tr}\left(\hat{\Sigma}_n+\lambda I\right)^{-1}\right]\right|\leq y$, denote the Stieltjes transform as $m_\mu(z)=\int_R\frac{1}{x-z}\mu(dx)$,

$$\begin{aligned}\lim_{n,d\to\infty}\left\|\mathbb{E}\left[\left(\hat{\Sigma}_n+\lambda I\right)^{-1}\right]-(\Sigma_n+\lambda I)^{-1}\right\| &\geq \frac{\sigma_{\min}}{\sigma_{\max}+\lambda}\cdot\frac{\lambda'}{\lambda(\sigma_{\max}+\lambda')}\left|y-\lambda'\lim_{n,d\to\infty}\frac{1}{n}\text{tr}\left(\Sigma_n+\lambda' I\right)^{-1}\right| \\ &= \frac{\lambda'\sigma_{\min}}{\lambda(\sigma_{\max}+\lambda)(\sigma_{\max}+\lambda')}\left|y-\lambda' y m_{F_\Sigma}(-\lambda')\right| \\ &\geq \frac{\lambda'\sigma_{\min}y}{\lambda(\sigma_{\max}+\lambda)(\sigma_{\max}+\lambda')}\left|\frac{\sigma_{\min}}{\sigma_{\min}+\lambda'}\right|\end{aligned} \tag{9}$$

When $\lambda\in o(1)$, since $1-\overline{d_{\lambda'}}\geq 1-y>0$, thus $\lambda'\in o(1)$ and therefore as $n,d\to\infty$,

$$\frac{\lambda'\sigma_{\min}y}{\lambda(\sigma_{\max}+\lambda)(\sigma_{\max}+\lambda')}\left|\frac{\sigma_{\min}}{\sigma_{\min}+\lambda'}\right|\to\frac{\lambda'\sigma_{\min}y}{\lambda\sigma_{\max}^2}=\frac{\sigma_{\min}y}{(1-\overline{d_{\lambda'}})\sigma_{\max}^2}\geq\frac{\sigma_{\min}y}{\sigma_{\max}^2}$$

□

[4] There exists λ' satisfying the equation for any $\lambda>0$. To see this, define $f:\mathbb{R}_+\to\mathbb{R}_+^*$ by $f(\lambda')=\lambda'(1-\overline{d_{\lambda'}})$, note $\lim_{\lambda'\to 0}f(\lambda')=0$ and $\lim_{\lambda'\to+\infty}f(\lambda')=+\infty$, and $f(\lambda')$ is continuous.

B. Proofs in Section 3

B.1. Technical Lemmas

Lemma B.1. *If data matrix A is random and satisfies Assumption 2.1, with A_i i.i.d. mean zero and covariance Σ. Denote $\tilde{p}_t$ as the result point of preconditioned conjugate gradient method, $p_t^\star$ as the true Newton's step. Then,*

$$\frac{\|\tilde{p}_t - p_t^\star\|_H^2}{\|p_t^\star\|_H^2} \leq 4\left(\frac{1-\sqrt{1-\frac{\alpha}{1-\alpha}}}{1+\sqrt{1-\frac{\alpha}{1-\alpha}}}\right)^{s_t}$$

where $\alpha = (\sigma_{\max}+\lambda+\alpha_0)\left(\frac{1}{\lambda^2}\alpha_0+\alpha_1+\|\Omega_0\|\right)$ with $\alpha_0 = \|\Sigma - \frac{1}{n}A^TA\|$, $\alpha_1 = \|\tilde{H}^{-1} - \mathbb{E}[\tilde{H}^{-1}]\|$ and $\sigma_{\max}$ being the largest eigenvalues of Σ. s_t denotes the number of iterations in preconditioned conjugate gradient method.

Proof. since

$$\begin{aligned}
&\left\|\tilde{H}^{-1} - H^{-1}\right\| \\
&= \left\|\tilde{H}^{-1} - \mathbb{E}\left[\tilde{H}^{-1}\right] + \mathbb{E}\left[\tilde{H}^{-1}\right] - (\Sigma+\lambda I)^{-1} + (\Sigma+\lambda I)^{-1} - H^{-1}\right\| \\
&\leq \frac{1}{\lambda^2}\alpha_0 + \alpha_1 + \|\Omega_0\|
\end{aligned} \tag{10}$$

thus

$$\begin{aligned}
\left\|H^{\frac{1}{2}}\tilde{H}^{-1}H^{\frac{1}{2}} - I\right\| &= \left\|H^{\frac{1}{2}}\left(\tilde{H}^{-1} - H^{-1}\right)H^{\frac{1}{2}}\right\| \\
&\leq \|H\|\left\|\tilde{H}^{-1} - H^{-1}\right\| \\
&\leq (\sigma_{\max}+\lambda+\alpha_0)\left(\frac{1}{\lambda^2}\alpha_0+\alpha_1+\|\Omega_0\|\right) = \alpha
\end{aligned}$$

Thus,

$$\left\|H^{-\frac{1}{2}}\tilde{H}H^{-\frac{1}{2}} - I\right\| \leq \frac{\alpha}{1-\alpha}$$

By equation (3.3) in (Lacotte & Pilanci, 2021), set the initial point in the preconditioned conjugate gradient method at 0,

$$\frac{\|\tilde{p}_t - p_t^\star\|_H^2}{\|p_t^\star\|_H^2} \leq 4\left(\frac{1-\sqrt{1-\frac{\alpha}{1-\alpha}}}{1+\sqrt{1-\frac{\alpha}{1-\alpha}}}\right)^{s_t}$$

□

Lemma B.2. *If data matrix A is random and satisfies Assumption 2.1, with A_i i.i.d. mean zero, assume ω_t independent of all $A_j^{(i)}$'s, denote $\Sigma(\omega_t) = Cov\left(f_i''\left(\omega_t^TA_j^{(i)}\right)^{\frac{1}{2}}A_j^{(i)}\right)$,*

$$\frac{\|\tilde{p}_t - p_t^\star\|_{H(\omega_t)}^2}{\|p_t^\star\|_{H(\omega_t)}^2} \leq 4\left(\frac{1-\sqrt{1-\frac{\alpha}{1-\alpha}}}{1+\sqrt{1-\frac{\alpha}{1-\alpha}}}\right)^{s_t}$$

where $\alpha = (\sigma_{\max}+\lambda+\alpha_0)\left(\frac{1}{\lambda^2}\alpha_0+\alpha_1+\|\Omega_0\|\right)$ with $\alpha_0 = \left\|\Sigma(\omega_t) - \frac{1}{n}\sum_{i=1}^n f_i''\left(\omega_t^TA_i\right)A_iA_i^T\right\|$, $\alpha_1 = \left\|\tilde{H}(\omega_t)^{-1} - \mathbb{E}\left[\tilde{H}(\omega_t)^{-1}\right]\right\|$ and $\sigma_{\max}$ being the largest eigenvalues of $\Sigma(\omega_t)$. s_t denotes the number of iterations in preconditioned conjugate gradient method.

Proof. Follow the same argument as in Lemma B.1 with H replaced by $H(\omega_t)$, $\tilde{H}$ replaced by $\tilde{H}(\omega_t)$, and Σ replaced by $\Sigma(\omega_t)$. □

B.2. Proof of Theorem 3.1

Proof. We build on proof in Dereziński & Mahoney (2019). Consider the tth Newton's step. Denote $p_t^\star = H^{-1}g_t$ and $\tilde{p}_t = \tilde{H}^{-1}g_t$ where g_t is the gradient and the current Newton's point, by Lemma 14 in Dereziński & Mahoney (2019), if

$$\|\tilde{p}_t - p_t^\star\|_H \leq \alpha \|p_t^\star\|_H$$

for some α, denote $\kappa = \text{cond}(H)$, then it holds

$$\|\Delta_{t+1}\| \leq \frac{\alpha\sqrt{2\kappa}}{\sqrt{1-\alpha^2}}\|\Delta_t\|$$

Since

$$\begin{aligned}\|\tilde{p}_t - p_t^\star\|_H &= \left\|H^{\frac{1}{2}}\left(\tilde{H}^{-1} - H^{-1}\right)H^{\frac{1}{2}}H^{-\frac{1}{2}}g_t\right\| \\ &\leq \|H\|\left\|\tilde{H}^{-1} - H^{-1}\right\|\left\|H^{-\frac{1}{2}}g_t\right\| \\ &= \|H\|\left\|\tilde{H}^{-1} - H^{-1}\right\|\|p_t^\star\|_H\end{aligned}$$

and

$$\sigma_{\max}(H) \leq \sigma_{\max} + \lambda + \alpha_0, \sigma_{\min}(H) \geq \sigma_{\min} + \lambda - \alpha_0$$

together with (10), we can derive

$$\|\Delta_{t+1}\| \leq \frac{\sqrt{2}\alpha}{\sqrt{1-\alpha^2}}\sqrt{\frac{\sigma_{\max}+\lambda+\alpha_0}{\sigma_{\min}+\lambda-\alpha_0}}\|\Delta_t\|$$

where $\alpha = (\sigma_{\max} + \lambda + \alpha_0)\left(\frac{1}{\lambda^2}\alpha_0 + \alpha_1 + \|\Omega_0\|\right)$ □

B.3. Proof of Theorem 3.3

Proof. Follow similar argument as in the convergence proof for Newton's method with quadratic loss (Appendix B.2), it can be derived that

$$\|H(\omega_t)^{-1} - \tilde{H}(\omega_t)^{-1}\| \leq \frac{1}{\lambda^2}\alpha_0 + \alpha_1 + \|\Omega_0\|$$

Let $\tilde{p}_t = \tilde{H}(\omega_t)^{-1}g(\omega_t)$ and $p_t^\star = H(\omega_t)^{-1}g(\omega_t)$. Since

$$\begin{aligned}\|\tilde{p}_t - p_t^\star\|_{H(\omega_t)} &= \left\|H(\omega_t)^{\frac{1}{2}}\left(\tilde{H}(\omega_t)^{-1} - H(\omega_t)^{-1}\right)H(\omega_t)^{\frac{1}{2}}H(\omega_t)^{-\frac{1}{2}}g(\omega_t)\right\| \\ &\leq \|H(\omega_t)\|\left\|\tilde{H}(\omega_t)^{-1} - H(\omega_t)^{-1}\right\|\left\|H(\omega_t)^{-\frac{1}{2}}g(\omega_t)\right\| \\ &= \|H(\omega_t)\|\left\|\tilde{H}(\omega_t)^{-1} - H(\omega_t)^{-1}\right\|\|p_t^\star\|_{H(\omega_t)} \\ &\leq (\sigma_{\max}+\lambda+\alpha_0)\left(\frac{1}{\lambda^2}\alpha_0 + \alpha_1 + \|\Omega_0\|\right)\|p_t^\star\|_{H(\omega_t)}\end{aligned}$$

By Lemma 14 in (Dereziński & Mahoney, 2019),

$$\|\Delta_{t+1}\| \leq \max\left\{\frac{\sqrt{2}\alpha}{\sqrt{1-\alpha^2}}\sqrt{\frac{\sigma_{\max}+\lambda+\alpha_0}{\sigma_{\min}+\lambda-\alpha_0}}\|\Delta_t\|, \frac{2L}{\sigma_{\min}+\lambda-\alpha_0}\|\Delta_t\|^2\right\}$$

where $\alpha = (\sigma_{\max} + \lambda + \alpha_0)\left(\frac{1}{\lambda^2}\alpha_0 + \alpha_1 + \|\Omega_0\|\right)$. □

B.4. More Convergence Analysis Results

B.4.1. Convergence of inexact Newton's method for Regularized General Convex Smooth Loss

Theorem B.3. *(Convergence of inexact Newton's method with Shrinkage) Assume ω_t independent of all $A_j^{(i)}$'s. Let $\alpha, \Sigma(\omega_t), \alpha_0, \alpha_1, \Delta_{t+1}, \sigma_{\min}, \sigma_{\max}$ as defined in Theorem 3.3,*

$$\|\Delta_{t+1}\| \leq \max\left\{\frac{2L}{\sigma_{\min}+\lambda-\alpha_0}\|\Delta_t\|^2, \frac{\sqrt{2}\alpha'}{\sqrt{1-\alpha'^2}}\sqrt{\frac{\sigma_{\max}+\lambda+\alpha_0}{\sigma_{\min}+\lambda-\alpha_0}}\|\Delta_t\|\right\}$$

where $\alpha' = \sqrt{4\left(\frac{1-\sqrt{1-\frac{\alpha}{1-\alpha}}}{1+\sqrt{1-\frac{\alpha}{1-\alpha}}}\right)^{s_t}}$ *with* s_t *being the number of iterations in preconditioned conjugate gradient method.*

Proof. The proof follows by a combination of Lemma 14 in (Dereziński & Mahoney, 2019) and Lemma B.2. □

C. Supplementary Simulation Results

C.1. Supplementary Simulation for Section 5.2

C.1.1. Formulas for Different Methods for Covariance Resolvent Estimation

Here we explicitly list the formulas used by different methods to compute $\tilde{R}$ which is used for the plots. Let n denote the number of data, m denote number of agents, create n, m in the way that n is divided by m and split the data evenly to each agent, let $A \in \mathbb{R}^{n\times d}$ denote the global data and A_i denote local data of size $\mathbb{R}^{(n/m)\times d}$, we compare three methods:

$$\text{Average: } \tilde{R}_a = \frac{1}{m}\sum_{i=1}^{m} R_{ai} \text{ where } R_{ai} = \left(\frac{m}{n}A_i^T A_i + \lambda I\right)^{-1}$$

$$\text{Shrinkage: } \tilde{R}_s = \frac{1}{m}\sum_{i=1}^{m} R_{si} \text{ where } R_{si} = \left(\frac{m}{n(1-\frac{md_\lambda}{n})}A_i^T A_i + \lambda I\right)^{-1}$$

$$\text{Determinant: } \tilde{R}_d = \frac{1}{m\det(\Sigma+\lambda I)}\sum_{i=1}^{m} R_{di} \text{ where } R_{di} = \det\left(\frac{m}{n}A_i^T A_i + \lambda I\right)\left(\frac{m}{n}A_i^T A_i + \lambda I\right)^{-1}$$

where $d_\lambda = \text{tr}\left(\Sigma(\Sigma+\lambda I)^{-1}\right)$.

C.1.2. Experiments on Covariance Resolvent Estimation with Synthetic Data

Here we give simulation results for covariance resolvent estimation with synthetic data. See Section 5.2 for a description of the setup and different methods being compared. Figure 5 shows that our shrinkage method gives more accurate estimation of covariance resolvent than both the averaging method and the determinantal averaging method for synthetic data created with difference covariance matrices.

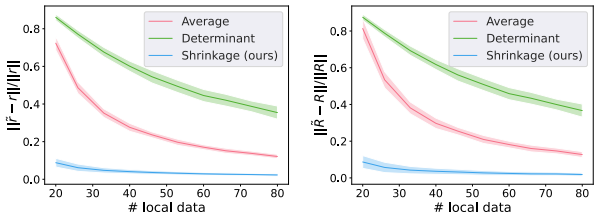

Figure 5: Synthetic data experiments for covariance resolvent estimation. Let m denote the number of agents, d denote data dimension, and λ denote the regularizer. We take $m = 100, d = 10, \lambda = 0.1$. Data is i.i.d. $\mathcal{N}(0, \Sigma)$ with $\Sigma = 0.1I$ in the left plot, and $\Sigma = 100C^T C, C_{ij} \sim U(0, 1)$ in the right plot.

C.1.3. More Experiments on Covariance Resolvent Estimation with Normalized Data

Here we give more simulation results for covariance resolvent estimation with normalized real datasets. See Section 5.2 for a description of the setup and different methods being compared. Figure 6 shows our simulation results, which confirms the shrinkage method's superiority over the averaging method and the determinantal averaging method in covariance resolvent estimation for normalized real data.

C.1.4. Experiments on Covariance Resolvent Estimation with Sketched Real Data

Here we give simulation results for covariance resolvent estimation with sketched real datasets. See Section 5.2 for a description of the setup and different methods being compared. Take a real dataset, we experiment with a sketch of it. See

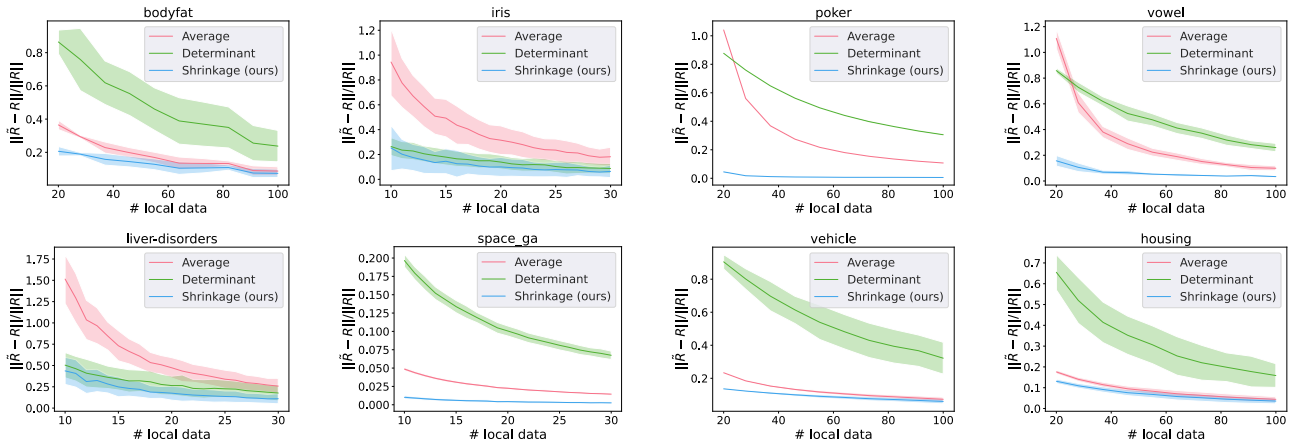

Figure 6: Normalized real data experiments for covariance resolvent estimation. $\lambda = 0.001$. Number of total data is rounded down as multiple of number of local data, number of agent is number of total data divided by number of local data. We use $\Sigma = \frac{1}{n} A^T A$.

Section 4 for an introduction to data sketching. We test with sketching matrix with both gaussian entries and uniform entries and they give similar results as what can be witnessed from Figure 7 and Figure 8. Figure 7 and Figure 8 demonstrate that the advantage of our method is not constrained to a specific data distribution.

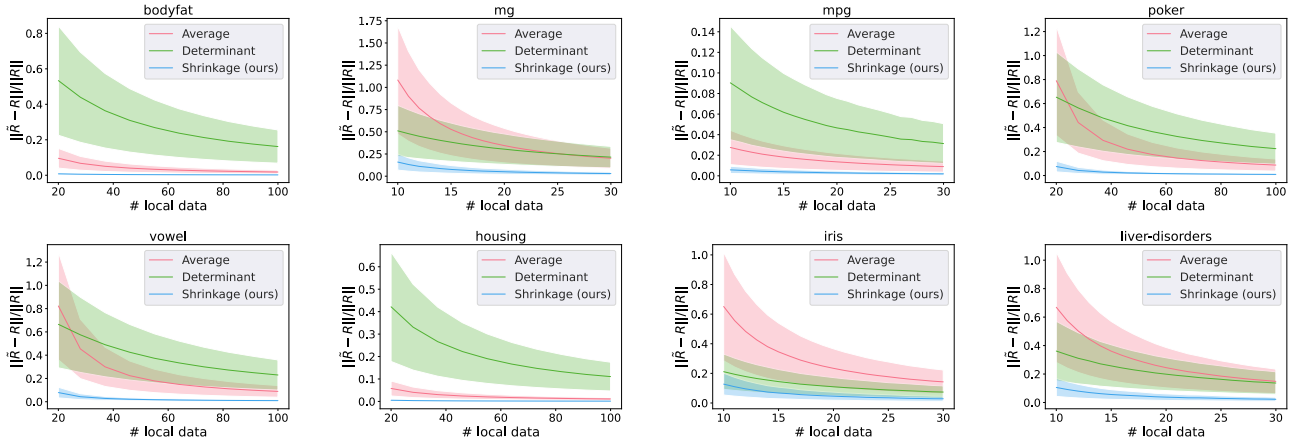

Figure 7: Sketched real data experiments for covariance resolvent estimation. We use sketching matrix with entry i.i.d. $\mathcal{N}(0, \frac{1}{m})$, sketch size $m = 10000$, regularizer $\lambda = 0.001$. We use $\Sigma = \frac{1}{n} A^T A$.

C.1.5. Experiments of small regularizer regime

We test our method in the small regularizer regime discussed in Section 2.1.2. Figure 9 shows the simulation results on synthetic data. The results suggest that when the regularizer is large, using $\frac{md}{n}$ is much worse than using $\frac{md_\lambda}{n}$, while when the regularizer is small, $\frac{md}{n}$ can be used in replace of $\frac{md_\lambda}{n}$ in computing $\tilde{R}_s$ and $\tilde{R}_s$ still approximates R well, which confirms Theorem 2.4 (see Section 5.2 for the definition of $\tilde{R}_s$ and R).

We also test with sketched real data. Figure 10 shows the result for sketched abalone dataset, which is similar to the synthetic data result.

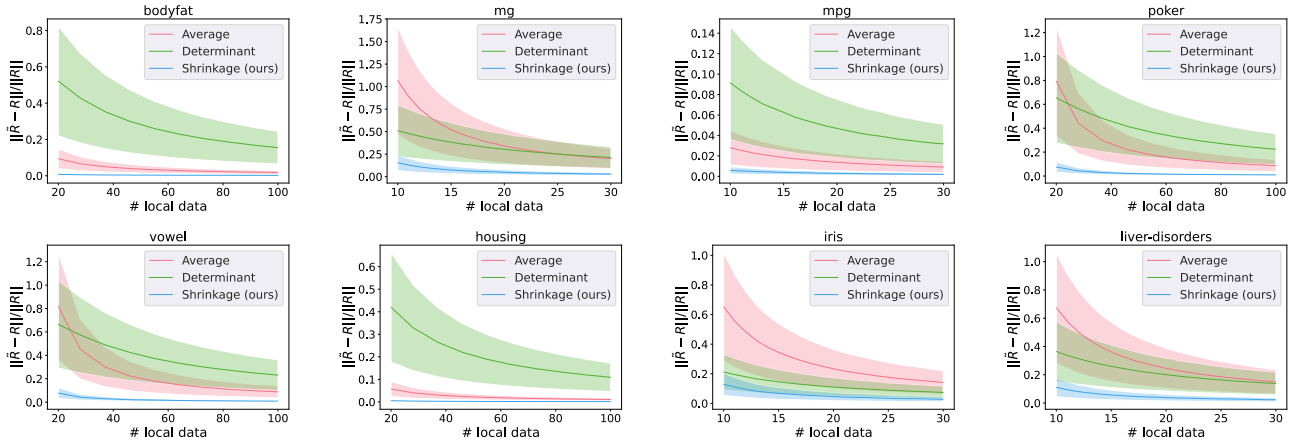

Figure 8: Sketched real data experiments for covariance resolvent estimation. We use sketching matrix with entry i.i.d. $\frac{1}{\sqrt{n}}U(-\frac{\sqrt{12}}{2}, \frac{\sqrt{12}}{2})$, sketch size $m = 10000$, regularizer $\lambda = 0.001$. We use $\Sigma = \frac{1}{n}A^TA$.

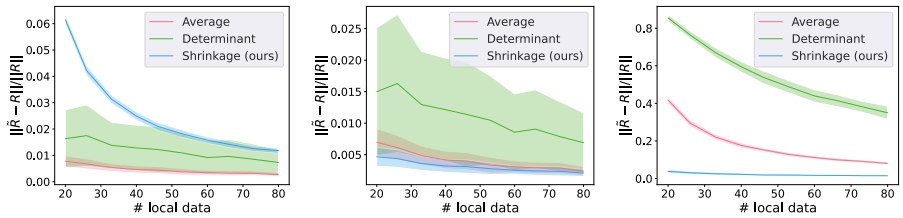

Figure 9: Synthetic data experiments for small regularizer regime. We experiment with $m = 100$ agents and data dimension $d = 10$. Data is i.i.d. $\mathcal{N}(0, \Sigma), \Sigma = 100C^TC, C_{ij} \sim U(0, 1)$. We take regularizer $\lambda = 2000$ in the first two plots and $\lambda = 1$ in the third plot. $\frac{md}{n}$ is used in replace of $\frac{md_\lambda}{n}$ in the first plot and the third plot. $\frac{md_\lambda}{n}$ is used in the second plot.

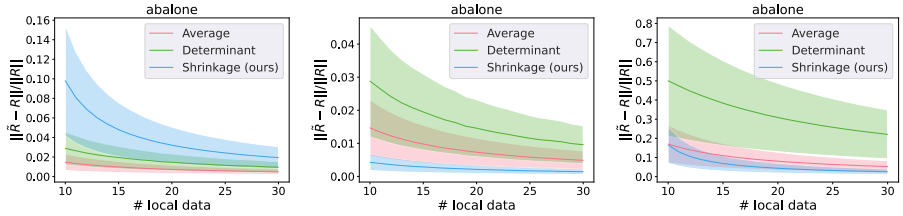

Figure 10: Sketched real data experiments for small regularizer regime. We use sketching matrix with entry i.i.d. $\mathcal{N}(0, \frac{1}{m})$, sketch size $m = 10000$. We use $\Sigma = \frac{1}{n}A^TA$. We take regularizer $\lambda = 1$ in the first two plots and $\lambda = 0.001$ in the third plot. $\frac{md}{n}$ is used in replace of $\frac{md_\lambda}{n}$ in the first plot and the third plot. $\frac{md_\lambda}{n}$ is used in the second plot.

C.2. Supplementary Simulation for Section 5.3

C.2.1. More Experiments on Distributed Newton's Method

Here we give more simulation results on distributed Newton's method for minimizing regularized quadratic loss with normalized real datasets. See Section 5.3 for a description of the setup and different methods being compared. Figure 11 shows the superiority of our method for saving communication rounds in distributed Newton's method.

C.2.2. Experiments on Distributed Inexact Newton's Method

Figure 12 shows simulation results for distributed inexact Newton's method. The algorithm of our method is given in Section 3. We compare with DiSCO, where Hessian resolvent of the first agent is used as the preconditioning matrix in

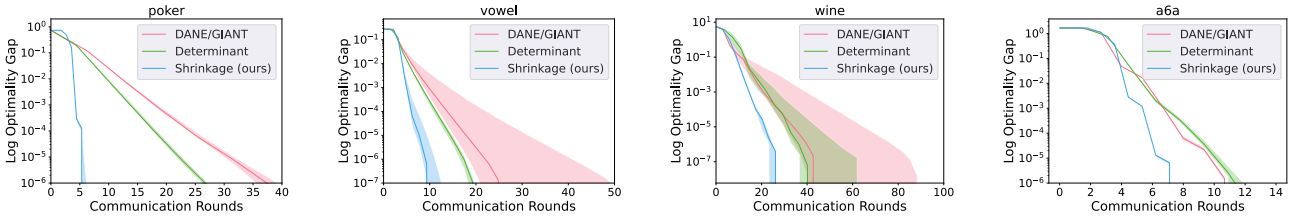

Figure 11: Normalized real data experiments for distributed Newton's method for ridge regression. Number of total data is rounded down as multiple of number of agents and number of local data is number of total data divided by number of agents. Let λ denote the regularizer and m denote the number of agents. We pick $m = 1000, \lambda = 0.001$ for poker, $m = 20, \lambda = 0.05$ for vowel, $m = 10, \lambda = 0.01$ for wine, $m = 50, \lambda = 0.5$ for a6a.

distributed preconditioned conjugate gradient method[5]. Average method is taking the average of local Hessian inverses as the preconditioning matrix. Determinant method is using exactly the same approximate Hessian inverse as in the determinantal averaging method discussed in Section 1.1 as the preconditioning matrix. We don't compare with conjugate gradient method since it usually takes many more steps to converge and involving it in the plot will make the difference between other methods less obvious. Determinant method is not plotted whenever there rises numerical issue.

From the plots, we see that compared to averaging method, DiSCO, and determinantal averaging method, our shrinkage method achieves smaller log optimality gap in fewer rounds of communication on the datasets we have tested, which suggests our shrinkage method is approximating Hessian inverse more accurately. Another takeaway is that determinantal averaging method is unstable when data dimension is large and computing determinant becomes infeasible, while our shrinkage method is always valid as long as the local data size is larger than effective dimension of local Hessian matrix. Note for minimizing quadratic loss, inexact Newton's method usually converges in one Newton step, and thus the discrepancy for different methods are smaller in these plots compared to distributed Newton's method's simulation plots.

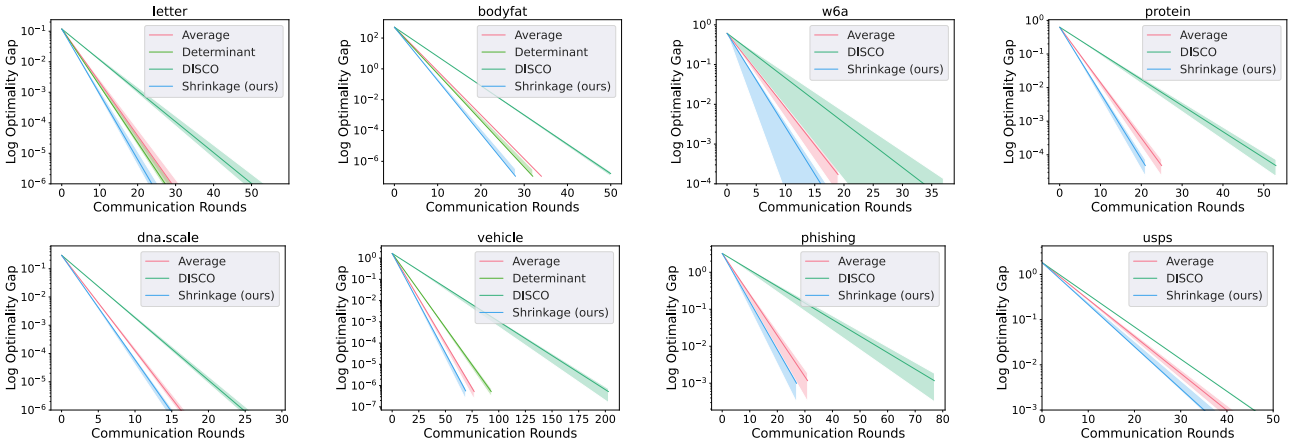

Figure 12: Normalized real data experiments for distributed inexact Newton's method for ridge regression. Number of total data is rounded down as multiple of number of agents and number of local data is number of total data divided by number of agents. Let λ denote the regularizer and m denote the number of agents. We pick $m = 1000, \lambda = 0.1$ for letter, $m = 20, \lambda = 0.01$ for bodyfat, $m = 30, \lambda = 1$ for w6a, $m = 50, \lambda = 0.1$ for protein and usps, $m = 50, \lambda = 10$ for dna.scale, $m = 40, \lambda = 1$ for vehicle, $m = 200, \lambda = 0.01$ for phishing.

[5]For the implementation of DiSCO, we only borrow its preconditioner and don't use its initialization step and step size choice since they are too specific.

C.2.3. Experiments on Logistic Regression

We give simulation results for distributed Newton's method and distributed inexact Newton's method for logistic regression on normalized real datasets. See Section 5.3 for a description of the setup and different methods being compared for distributed Newton's method. See Appendix C.2.2 for a description of methods being compared for distributed inexact Newton's method. According to our convergence analysis for non-quadratic loss in Section 3.2, we need each Newton's step to operate on data independent of previous Newton's steps. Therefore we are taking fresh data batches for computing each Newton's step. Determinantal averaging method does not appear in the inexact Newton's method plots since the method failes due to numerical issues.

According to Figure 13, our shrinkage method frequently saves communication rounds for both Newton's method and inexact Newton's method compared to other methods, though the discrepancy is not as significant for the case with quadratic loss. Moreover, a larger variance in the performance of Newton's method is observed. This might be due to a small batch size used in each worker. We believe that our shrinkage method can be optimized further for non-quadratic losses, which is left as future work.

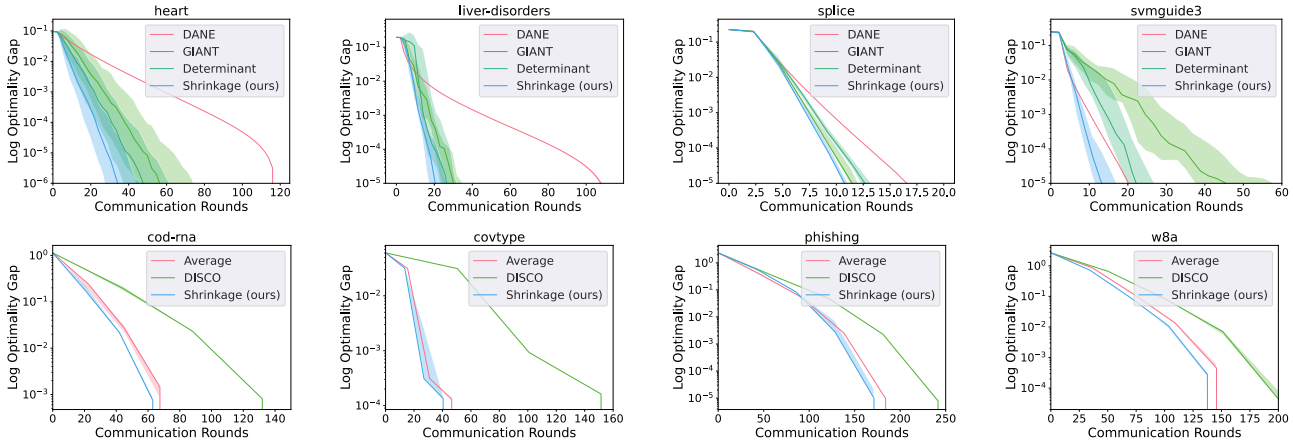

Figure 13: Normalized real data experiments for distributed second-order optimization algorithms for logistic regression. Top four plots are for distributed Newton's method and bottom four plots are for distributed inexact Newton's method. Number of total data is rounded down as multiple of number of agents and number of local data is number of total data divided by number of agents. For each Newton's step, a refreshed data batch containing the number of local data divided by max_iters pieces of data is used (we limit number of Newton's step to not exceed max_iters). Let λ denote the regularizer and m denote the number of agents. We pick $m = 5, \lambda = 0.01$,max_iters$= 20$ for heart, $m = 2, \lambda = 0.01$,max_iters$= 10$ for liver-disorders, $m = 3, \lambda = 0.1$,max_iters$= 5$ for splice, $m = 10, \lambda = 0.1$,max_iters$= 20$ for svmguide3, $m = 100, \lambda = 1e-5$,max_iters$= 50$ for cod-rna, $m = 200, \lambda = 1e-5$,max_iters$= 50$ for covtype, $m = 40, \lambda = 0.01$,max_iters$= 50$ for phishing, $m = 50, \lambda = 0.1$,max_iters$= 50$ for w8a.

C.2.4. Experiments on Iterative Hessian Sketch With Optimal Shrinkage

We only present the simulation plots for IHS and IHS with shrinkage where a heuristic shrinkage coefficient is used in the main text in Section 5.4. Here we provide more plots on these two methods and we also provide the exact version for IHS with optimal shrinkage. For sake of comparison, we include the plots in the main text here again.

Figure 14 presents our simulation results on IHS with shrinkage where the effective dimension of sketched data is used. From the figure, we see that IHS equipped with shrinkage method beats the classic IHS method in datasets we experimented with, though for datasets bodyfat, housing, mpg and triazines, the difference between these two methods are more obvious. Figure 15 presents the plots for IHS with the exact optimal shrinkage coefficient, on the same datasets choice and with the same parameter choice. Still, IHS with shrinkage method beats the classic IHS method in datasets we have tested on. But this time, the difference between the two methods are obvious for all datasets. If we look at Figure 14 and Figure 15 together, the performance of IHS with the exact shrinkage coefficient is at least as good as IHS with the heuristic shrinkage coefficient, which is as expected, though the heuristic version does not worsen the performance much such that IHS with heuristic shrinkage still remains superior compared to the classic IHS method.

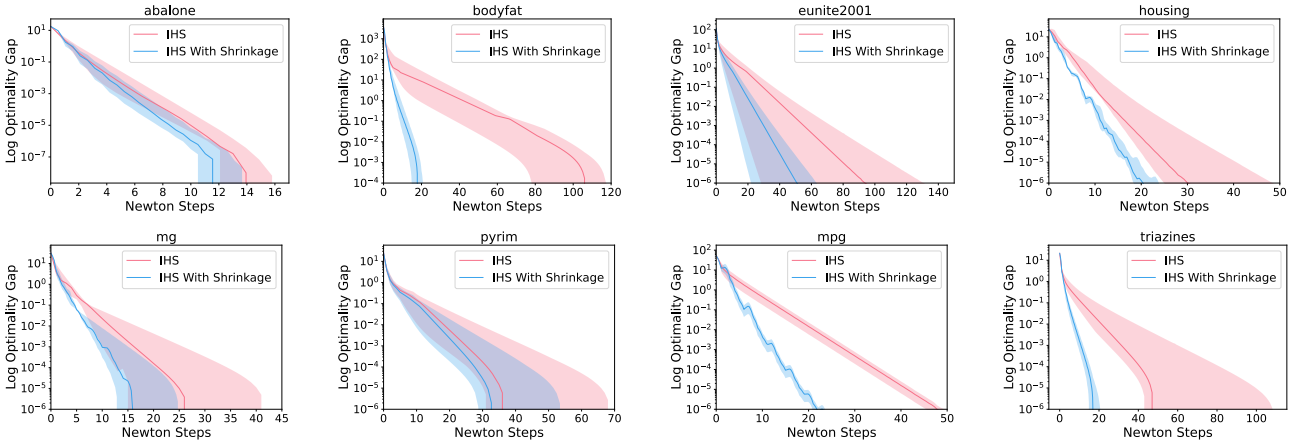

Figure 14: Real data experiments for Iterative Hessian Sketch method with heuristic shrinkage coefficient for ridge regression. Let λ denote the regularizer and m denote the sketch size. We pick $\lambda = 0.001$ for abalone,triazines and $\lambda = 0.01$ for bodyfat, eunite2001, housing, mg, pyrim, mpg . We pick $m = 50$ for abalone, $m = 100$ for bodyfat and pyrim, $m = 300$ for eunite2001, housing, and triazines, $m = 20$ for mg, and $m = 30$ for mpg.

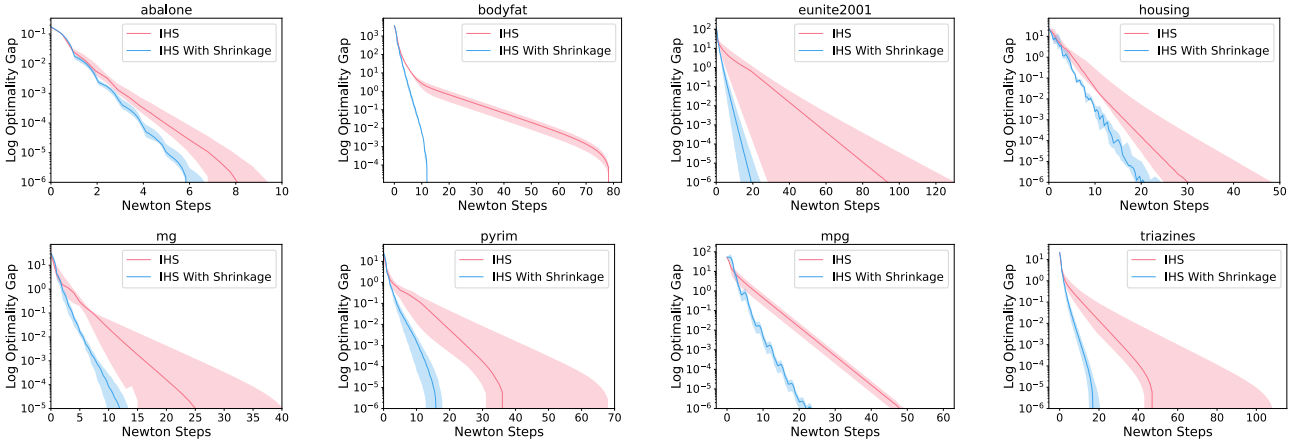

Figure 15: Real data experiments for Iterative Hessian Sketch method with exact shrinkage coefficient for ridge regression. Same parameter choice as in Figure 14.