Distributed estimation of the inverse Hessian by determinantal averaging

Michał Dereziński
Department of Statistics
University of California, Berkeley
mderezin@berkeley.edu

Michael W. Mahoney
ICSI and Department of Statistics
University of California, Berkeley
mmahoney@stat.berkeley.edu

Abstract

In distributed optimization and distributed numerical linear algebra, we often encounter an inversion bias: if we want to compute a quantity that depends on the inverse of a sum of distributed matrices, then the sum of the inverses does not equal the inverse of the sum. An example of this occurs in distributed Newton’s method, where we wish to compute (or implicitly work with) the inverse Hessian multiplied by the gradient. In this case, locally computed estimates are biased, and so taking a uniform average will not recover the correct solution. To address this, we propose determinantal averaging, a new approach for correcting the inversion bias. This approach involves reweighting the local estimates of the Newton’s step proportionally to the determinant of the local Hessian estimate, and then averaging them together to obtain an improved global estimate. This method provides the first known distributed Newton step that is asymptotically consistent, i.e., it recovers the exact step in the limit as the number of distributed partitions grows to infinity. To show this, we develop new expectation identities and moment bounds for the determinant and adjugate of a random matrix. Determinantal averaging can be applied not only to Newton’s method, but to computing any quantity that is a linear transformation of a matrix inverse, e.g., taking a trace of the inverse covariance matrix, which is used in data uncertainty quantification.

1 Introduction

Many problems in machine learning and optimization require that we produce an accurate estimate of a square matrix $H$ (such as the Hessian of a loss function or a sample covariance), while having access to many copies of some unbiased estimator of $H$, i.e., a random matrix $\tilde{H}$ such that $\mathbb{E}[\tilde{H}] = H$. In these cases, taking a uniform average of those independent copies provides a natural strategy for boosting the estimation accuracy, essentially by making use of the law of large numbers: $\frac{1}{m} \sum_{t=1}^{m} \tilde{H}_t \rightarrow H$. For many other problems, however, we are more interested in the inverse (Hessian/covariance) matrix $H^{-1}$, and it is necessary or desirable to work with $\tilde{H}^{-1}$ as the estimator. Here, a naïve averaging approach has certain fundamental limitations (described in more detail below). The basic reason for this is that $\mathbb{E}[\tilde{H}^{-1}] \neq H^{-1}$, i.e., that there is what may be called an inversion bias.

In this paper, we propose a method to address this inversion bias challenge. The method uses a weighted average, where the weights are carefully chosen to compensate for and correct the bias. Our motivation comes from distributed Newton’s method (explained shortly), where combining independent estimates of the inverse Hessian is desired, but our method is more generally applicable, and so we first state our key ideas in a more general context.
Theorem 1 Let $s_i$ be independent random variables and $Z_i$ be fixed square rank-1 matrices. If $\hat{H} = \sum_i s_i Z_i$ is invertible almost surely, then the inverse of the matrix $H = \mathbb{E}[\hat{H}]$ can be expressed as:

$$H^{-1} = \frac{\mathbb{E}[\det(\hat{H}) \hat{H}^{-1}]}{\mathbb{E}[\det(\hat{H})]}.$$ 

To demonstrate the implications of Theorem 1, suppose that our goal is to estimate $F(H^{-1})$ for some linear function $F$. For example, in the case of Newton’s method $F(H^{-1}) = H^{-1} g$, where $g$ is the gradient and $H$ is the Hessian. Another example would be $F(H^{-1}) = \text{tr}(H^{-1})$, where $H$ is the covariance matrix of a dataset and $\text{tr}(\cdot)$ is the matrix trace, which is useful for uncertainty quantification. For these and other cases, consider the following estimation of $F(H^{-1})$, which takes an average of the individual estimates $F(\hat{H}_i^{-1})$, each weighted by the determinant of $\hat{H}_i$, i.e.,

**Determinantal Averaging:**

$$\hat{F}_m = \frac{\sum_{i=1}^m a_t F(\hat{H}_i^{-1})}{\sum_{i=1}^m a_t}, \quad a_t = \det(\hat{H}_t).$$

By applying the law of large numbers (separately to the numerator and the denominator), Theorem 1 easily implies that if $\hat{H}_1, \ldots, \hat{H}_m$ are i.i.d. copies of $H$ then this determinantal averaging estimator is asymptotically consistent, i.e., $\hat{F}_m \rightarrow F(H^{-1})$, almost surely. This determinantal averaging estimator is particularly useful when problem constraints do not allow us to compute $F((1/m) \sum_i \hat{H}_i^{-1})$, e.g., when the matrices are distributed and not easily combined.

To establish finite sample convergence guarantees for estimators obtained via determinantal averaging, we establish the following matrix concentration result. We state it separately since it is technically interesting and since its proof requires novel bounds for the higher moments of the determinant of a random matrix, which is likely to be of independent interest. Below and throughout the paper, $C$ denotes an absolute constant and “$\leq$” is the Löwner order on positive semi-definite (psd) matrices.

**Theorem 2** Let $\hat{H} = \frac{1}{k} \sum_{i=1}^n b_i Z_i + B$ and $H = \mathbb{E}[\hat{H}]$, where $B$ is a positive definite $d \times d$ matrix and $b_i$ are i.i.d. Bernoulli($\frac{k}{n}$). Moreover, assume that all $Z_i$ are psd, $d \times d$ and rank-1. If $k \geq C \frac{\mu d^2}{\eta^2} \log \frac{3d}{\delta}$ for $\eta \in (0, 1)$ and $\mu = \max_i \|Z_i H^{-1}\|/d$, then

$$\left(1 - \frac{\eta}{\sqrt{m}}\right) \cdot H^{-1} \preceq \frac{\sum_{i=1}^m a_t \hat{H}_i^{-1}}{\sum_{i=1}^m a_t} \preceq \left(1 + \frac{\eta}{\sqrt{m}}\right) \cdot H^{-1} \quad \text{with probability } \geq 1 - \delta,$$

where $\hat{H}_1, \ldots, \hat{H}_m \overset{i.i.d.}{\sim} \hat{H}$ and $a_t = \det(\hat{H}_t)$.

### 1.1 Distributed Newton’s method

To illustrate how determinantal averaging can be useful in the context of distributed optimization, consider the task of batch minimization of a convex loss over vectors $w \in \mathbb{R}^d$, defined as follows:

$$L(w) \overset{def}{=} \frac{1}{n} \sum_{i=1}^n \ell_i(w^\top x_i) + \frac{\lambda}{2} \|w\|^2, \quad (1)$$

where $\lambda > 0$, and $\ell_i$ are convex, twice differentiable and smooth. Given a vector $w$, Newton’s method dictates that the correct way to move towards the optimum is to perform an update $\hat{w} = w - p$, with $p = \nabla^{-2}L(w) \nabla L(w)$, where $\nabla^{-2}L(w) = (\nabla^2 L(w))^{-1}$ denotes the inverse Hessian of $L$ at $w$. Here, the Hessian and gradient are:

$$\nabla^2 L(w) = \frac{1}{n} \sum_i \ell''_i(w^\top x_i) x_i x_i^\top + \lambda I, \quad \text{and} \quad \nabla L(w) = \frac{1}{n} \sum_i \ell'_i(w^\top x_i) x_i + \lambda w.$$  

For our distributed Newton application, we study a distributed computational model, where a single machine has access to a subsampled version of $L$ with sample size parameter $k \ll n$:

$$\hat{L}(w) \overset{def}{=} \frac{1}{k} \sum_{i=1}^n b_i \ell_i(w^\top x_i) + \frac{\lambda}{2} \|w\|^2, \quad \text{where } b_i \sim \text{Bernoulli}(k/n). \quad (2)$$

---

1 Clearly, one would not actually compute the inverse of the Hessian explicitly. We describe it this way for simplicity. Our results hold whether or not the inverse operator is computed explicitly.
We next establish how this error bound impacts the convergence guarantees offered by Newton’s method. Note that under our assumptions, the gradient has size $O(d)$ so it can be computed exactly within this communication budget (e.g., via map-reduce), however the Hessian has to be approximated locally by each machine. Note that other computational models can be considered, such as those where the global gradient is not computed (and local gradients are used instead).

Under the constraints described above, the most natural strategy is to use directly the Hessian of the locally subsampled loss $\tilde{L}$ (see, e.g., GIANT [WRKXM13]), resulting in the approximate Newton step $\hat{p} = \nabla^{-2}\tilde{L}(w) \nabla L(w)$. Suppose that we independently construct $m$ i.i.d. copies of this estimate: $\hat{p}_1, \ldots, \hat{p}_m$ (here, $m$ is the number of machines). Then, for sufficiently large $m$, taking a simple average of the estimates will stop converging to $p$ because of the inversion bias: $\frac{1}{m} \sum_{t=1}^m \hat{p}_t \to E[\hat{p}] \neq p$. Figure 1 shows this by plotting the estimation error (in Euclidean distance) of the averaged Newton step estimators, when the weights are uniform and determinantal (for more details and plots, see Appendix C).

The only way to reduce the estimation error beyond a certain point is to increase the local sample size $k$ (thereby reducing the inversion bias), which raises the computational cost per machine. Determinantal averaging corrects the inversion bias so that estimation error can always be decreased by adding more machines without increasing the local sample size. From the preceding discussion we can easily show that determinantal averaging leads to an asymptotically consistent estimator. This is a corollary of Theorem 1 as proven in Section B.

**Corollary 3** Let $\{\tilde{L}_i\}_{i=1}^\infty$ be i.i.d. samples of $\tilde{L}$ and define $a_t = \det(\nabla^2\tilde{L}_i(w))$. Then:

$$\frac{\sum_{t=1}^m a_t \hat{p}_t}{\sum_{t=1}^m a_t} \xrightarrow{m \to \infty} p,$$

where $\hat{p}_t = \nabla^{-2}\tilde{L}_i(w) \nabla L(w)$ and $p = \nabla^{-2}L(w) \nabla L(w)$.

The (unnormalized) determinantal weights can be computed locally in the same time as it takes to compute the Newton estimates so they do not add to the overall cost of the procedure. While this result is only an asymptotic statement, it holds with virtually no assumptions on the loss function (other than twice-differentiability) or the expected local sample size $k$. However, with some additional assumptions we will now establish a convergence guarantee with a finite number of machines $m$ by bounding the estimation error for the determinantal averaging estimator of the Newton step.

In the next result, we use Mahalanobis distance, denoted $\|v\|_M = \sqrt{v^T M v}$, to measure the error of the Newton step estimate (i.e., the deviation from optimum $p$), with $M$ chosen as the Hessian of $L$. This choice is motivated by standard convergence analysis of Newton’s method, discussed next. This is a corollary of Theorem 2 as explained in Section C.

**Corollary 4** For any $\delta, \eta \in (0, 1)$ if expected local sample size satisfies $k \geq C\eta^{-2} \mu^2 d \log^3 d$, then

$$\left\| \frac{\sum_{t=1}^m a_t \hat{p}_t}{\sum_{t=1}^m a_t} - p \right\|_{\nabla^2 L(w)} \leq \frac{\eta}{\sqrt{m}} \cdot \|p\|_{\nabla^2 L(w)}$$

with probability $\geq 1 - \delta$,

where $\mu = \frac{1}{d} \max_i \ell''_i(w^* x_i) \|x_i\| \|w^*\|$, and $a_t$, $\hat{p}_t$ and $p$ are defined as in Corollary 3.

We next establish how this error bound impacts the convergence guarantees offered by Newton’s method. Note that under our assumptions $L$ is strongly convex so there is a unique minimizer $w^* = \arg\min_w L(w)$. We ask how the distance from optimum, $\|w - w^*\|$, changes after we make an update $\hat{w} = w - \hat{p}$. For this, we have to assume that the Hessian matrix is $L$-Lipschitz as a function of $w$. After this standard assumption, a classical analysis of the Newton’s method reveals that Corollary 3 implies the following Corollary 5 (proof in Appendix D).

![Figure 1: Newton step estimation error versus number of machines, averaged over 100 runs (shading is standard error) for a libsvm dataset. More plots in Appendix C.](image-url)
Assumption 5 The Hessian is L-Lipschitz: \( \| \nabla^2 \mathcal{L}(w) - \nabla^2 \mathcal{L}(\bar{w}) \| \leq L \| w - \bar{w} \| \) for any \( w, \bar{w} \in \mathbb{R}^d \).

Corollary 6 For any \( \delta, \eta \in (0, 1) \) if expected local sample size satisfies \( k \geq C \eta^{-2} \mu d^2 \log^3 \frac{d}{\delta} \) then under Assumption 5 it holds with probability at least \( 1 - \delta \) that

\[
\| \bar{w} - w^* \| \leq \max \left\{ \frac{\eta}{\sqrt{m}} \sqrt{k} \| w - w^* \|, \frac{2L}{\lambda_{\min}} \| w - w^* \|^2 \right\} \quad \text{for} \quad \bar{w} = w - \sum_{t=1}^m a_t \bar{p}_t, 
\]

where \( C, \mu, a_t \) and \( \bar{p}_t \) are defined as in Corollaries 3 and 4, while \( \kappa \) and \( \lambda_{\min} \) are the condition number and smallest eigenvalue of \( \nabla^2 \mathcal{L}(w) \), respectively.

The bound is a maximum of a linear and a quadratic convergence term. As \( m \) goes to infinity and/or \( \eta \) goes to 0 the approximation coefficient \( \alpha = \frac{\eta}{\sqrt{m}} \) in the linear term disappears and we obtain exact Newton’s method, which exhibits quadratic convergence (at least locally around \( w^* \)). However, decreasing \( \eta \) means increasing \( k \) and with it the average computational cost per machine. Thus, to preserve the quadratic convergence while maintaining a computational budget per machine, as the optimization progresses we have to increase the number of machines \( m \) while keeping \( k \) fixed. This is only possible when we correct for the inversion bias, which is done by determinantal averaging.

1.2 Distributed data uncertainty quantification

Here, we consider another example of when computing a compressed linear representation of the inverse matrix is important. Let \( \mathbf{X} \) be an \( n \times d \) matrix where the rows \( \mathbf{x}_i^T \) represent samples drawn from a population for statistical analysis. The sample covariance matrix \( \Sigma = \frac{1}{n} \mathbf{X}^T \mathbf{X} \) holds the information about the relations between the features. Assuming that \( \Sigma \) is invertible, the matrix \( \Sigma^{-1} \), also known as the precision matrix, is often used to establish a degree of confidence we have in the variance information of each individual feature. Thus, efficiently estimating either the entire diagonal, its trace, or some subset of its entries, is of practical interest [Ste97, WLP16, BCF09].

We consider the distributed setting where data is separately stored in batches and each local covariance is modeled as:

\[
\hat{\Sigma} = \frac{1}{k} \sum_{i=1}^n b_i \mathbf{x}_i \mathbf{x}_i^T, \quad \text{where} \quad b_i \sim \text{Bernoulli}(k/n).
\]

For each of the local covariances \( \hat{\Sigma}_1, \ldots, \hat{\Sigma}_m \), we compute its compressed uncertainty information: \( F((\hat{\Sigma}_t + \frac{d}{\sqrt{m}} \mathbf{I})^{-1}) \), where we added a small amount of ridge to ensure invertibility.\(^2\) Here, \( F(\cdot) \) may for example denote the trace or the vector of diagonal entries. We arrive at the following asymptotically consistent estimator for \( F(\Sigma^{-1}) \):

\[
\hat{F}_m = \frac{\sum_{t=1}^m a_{t,m} F((\hat{\Sigma}_t + \frac{d}{\sqrt{m}} \mathbf{I})^{-1})}{\sum_{t=1}^m a_{t,m}}, \quad \text{where} \quad a_{t,m} = \det((\hat{\Sigma}_t + \frac{d}{\sqrt{m}} \mathbf{I})^{-1}).
\]

Note that the ridge term \( \frac{d}{\sqrt{m}} \mathbf{I} \) decays to zero as \( m \) goes to infinity, which is why \( \hat{F}_m \to F(\Sigma^{-1}) \).

Even though this limit holds for any local sample size \( k \), in practice we should choose \( k \) sufficiently large so that \( \hat{\Sigma} \) is well-conditioned. In particular, Theorem 2 implies that if \( k \geq 2C \eta^{-2} \mu d^2 \log^3 \frac{d}{\delta} \), where \( \mu = \frac{1}{d} \max_i \| \mathbf{x}_i \| \Sigma^{-1} \), then for \( F(\cdot) = \text{tr}(\cdot) \) we have \( |\hat{F}_m - \text{tr}(\Sigma^{-1})| \leq \frac{\eta}{\sqrt{m}} \cdot \text{tr}(\Sigma^{-1}) \) w.p. \( 1 - \delta \).

1.3 Related work

Many works have considered averaging strategies for combining distributed estimates, particularly in the context of statistical learning and optimization. This research is particularly important in federated learning [KBR16, KBY16], where data are spread out across a large network of devices with small local storage and severely constrained communication bandwidth. Using averaging to combine local estimates has been studied in a number of learning settings [MMS09, MHH10].

\(^2\)Since the ridge term vanishes as \( m \) goes to infinity, we are still estimating the ridge-free quantity \( F(\Sigma^{-1}) \).
well as for first-order stochastic optimization \cite{ZWLS10,AD11}. For example, \cite{ZDW13} examine
the effectiveness of simple uniform averaging of empirical risk minimizers and also propose a
booststrapping technique to reduce the bias.

More recently, distributed averaging methods gained considerable attention in the context of second-
order optimization, where the Hessian inversion bias is of direct concern. \cite{SSZ14} propose a
distributed approximate Newton-type method (DANE) which under certain assumptions exhibits
low bias. This was later extended and improved upon by \cite{ZLT13,RRK+16}. The GIANT method
\cite{WRKXM13} most closely follows our setup from Section 1.1, providing non-trivial guarantees
for uniform averaging of the Newton step estimates $\tilde{p}_i$ (except they use with-replacement uniform
sampling, whereas we use without-replacement, but that is typically a negligible difference). A
related analysis of this approach is provided in the context of ridge regression by \cite{WGM17}. Finally,\cite{ABH17,MLR17,BJK17} propose different estimates of the Newton step which exhibit low bias
under certain additional assumptions.

Our approach is related to recent developments in determinantal subsampling techniques (e.g.,
volume sampling), which have been shown to correct the inversion bias in the context of least squares
regression \cite{DWH17,DWH18}. However, despite recent progress \cite{DWH17,DWH18}, volume sampling
is still far too computationally expensive to be feasible for distributed optimization. Indeed, often
uniform sampling is the only practical choice in this context.

With the exception of the expensive volume sampling-based methods, all of the approaches discussed
above, even under favorable conditions, use biased estimates of the desired solution (e.g., the exact
Newton step). Thus, when the number of machines grows sufficiently large, with fixed local sample
size, the averaging no longer provides any improvement. This is in contrast to our determinantal
averaging, which converges exactly to the desired solution and requires no expensive subsampling.
Therefore, it can scale with an arbitrarily large number of machines.

2 Expectation identities for determinants and adjugates

In this section, we prove Theorem 1 and Corollary 3 establishing that determinantal averaging is
asymptotically consistent. To achieve this, we establish a lemma involving two expectation identities.

For a square $n \times n$ matrix $A$, we use $\text{adj}(A)$ to denote its adjugate, defined as an $n \times n$ matrix
whose $(i,j)$th entry is $(-1)^{i+j} \det(A_{j,i-1})$, where $A_{j,i-1}$ denotes $A$ without $j$th row and $i$th
column. The adjugate matrix provides a key connection between the inverse and the determinant
because for any invertible matrix $A$, we have $\text{adj}(A) = \det(A)A^{-1}$. In the following lemma,
we will also use a formula called Sylvester’s theorem, relating the adjugate and the determinant:
$\det(A + uv^T) = \det(A) + v^T \text{adj}(A)u$.

Lemma 7 For $A = \sum_i s_i Z_i$, where $s_i$ are independently random and $Z_i$ are square and rank-1,

(a) $\mathbb{E}[\det(A)] = \det(\mathbb{E}[A])$ and (b) $\mathbb{E}[\text{adj}(A)] = \det(\mathbb{E}[A])$.

Proof We use induction over the number of components in the sum. If there is only one component,
i.e., $A = sZ$, then $\det(A) = 0$ a.s., unless $Z$ is $1 \times 1$, in which case (a) is trivial, and (b) follows
similarly. Now, suppose we showed the hypothesis when the number of components is $n$ and let $A = \sum_{i=1}^{n+1} s_i Z_i$. Setting $Z_{n+1} = uv^T$, we have:

$$
\mathbb{E}[\det(A)] = \mathbb{E}\left[\det(\sum_{i=1}^{n} s_i Z_i + s_{n+1}uv^T)\right]
$$

(Sylvester’s Theorem) $=$ $\mathbb{E}\left[\det(\sum_{i=1}^{n} s_i Z_i) + s_{n+1}v^T \text{adj}(\sum_{i=1}^{n} s_i Z_i) u\right]$

(inductive hypothesis) $=$ $\det\left(\mathbb{E}\left[\sum_{i=1}^{n} s_i Z_i\right]\right) + \mathbb{E}[s_{n+1}] v^T \text{adj}\left(\mathbb{E}\left[\sum_{i=1}^{n} s_i Z_i\right]\right) u$

(Sylvester’s Theorem) $=$ $\det\left(\mathbb{E}\left[\sum_{i=1}^{n} s_i Z_i\right] + \mathbb{E}[s_{n+1}]uv^T\right) = \det(\mathbb{E}[A])$,

showing (a). Finally, (b) follows by applying (a) to each entry $\text{adj}(A)_{ij} = (-1)^{i+j} \det(A_{j,i-1})$. $\blacksquare$
We next prove Theorem 1 and Corollary 3 as consequences of Lemma 7. This concludes the proof.

The event that $B$ is invertible, its adjugate is given by $\text{adj}(A) = \det(A)A^{-1}$, so the lemma implies that

$$\mathbb{E}[\det(A)](\mathbb{E}[A])^{-1} = \det(\mathbb{E}[A]) = \mathbb{E}[\text{adj}(A)] = \mathbb{E}[\det(A)A^{-1}],$$

from which Theorem 1 follows immediately.

**Proof of Corollary 3** The subsampled Hessian matrix used in Corollary 3 can be written as:

$$\nabla^2 \mathcal{L}(w) = \frac{1}{k} \sum_{i} b_i k_{i} (w - x_i) x_i^T + \lambda \sum_{i=1}^d e_i e_i^T \overset{\text{def}}{=} \hat{H},$$

so letting $\hat{H} = \nabla^2 \mathcal{L}(w)$, Corollary 3 follows from Theorem 1 and the law of large numbers:

$$\begin{align*}
\frac{\sum_{i=1}^m a_i \hat{p}_i}{\sum_{i=1}^m a_i} = \frac{1}{m} \sum_{i=1}^m \det(\hat{H}_i) \hat{H}_i^{-1} \nabla \mathcal{L}(w) & \xrightarrow{m \to \infty} \frac{\mathbb{E}[\det(\hat{H}) \hat{H}^{-1}]}{\mathbb{E}[\det(\hat{H})]} \nabla \mathcal{L}(w) = \nabla^2 \mathcal{L}(w) \nabla \mathcal{L}(w),
\end{align*}$$

which completes the proof.

### 3 Finite-sample convergence analysis

In this section, we prove Theorem 2 and Corollary 4 establishing that determinantal averaging exhibits a $1/\sqrt{m}$ convergence rate, where $m$ is the number of sample matrices (or the number of machines in distributed Newton’s method). For this, we need a tool from random matrix theory.

**Lemma 8 (Matrix Bernstein [Tro02])** Consider a finite sequence $\{X_i\}$ of independent, random, self-adjoint matrices with dimension $d$ such that $\mathbb{E}[X_i] = 0$ and $\lambda_{\max}(X_i) \leq R$ almost surely. If the sequence satisfies $\|\sum_i \mathbb{E}[X_i^2]\| \leq \sigma^2$, then the following inequality holds for all $x \geq 0$:

$$\Pr(\lambda_{\max}(\sum_i X_i) \geq x) \leq \begin{cases} d e^{-\frac{x^2}{4\sigma^2}} & \text{for } x \leq \frac{\sigma^2}{R}; \\ d e^{-\frac{x^2}{4\sigma^2}} & \text{for } x \geq \frac{\sigma^2}{R}. \end{cases}$$

The key component of our analysis is bounding the moments of the determinant and adjugate of a certain class of random matrices. This has to be done carefully, because higher moments of the determinant grow more rapidly than, e.g., for a sub-gaussian random variable. For this reason, we do not require that the individual components $Z_i$ of matrix $A$ be rank-1, but we impose several additional boundedness assumptions. In the proof below we apply the concentration inequality of Lemma 8 twice: first to the random matrix $A$ itself, and then also to its trace, which allows finer control over the determinant.

**Lemma 9** Let $A = \frac{1}{d} \sum_i b_i Z_i + B$, where $b_i \sim \text{Bernoulli}(\gamma)$ are independent, whereas $Z_i$ and $B$ are $d \times d$ psd matrices such that $\|Z_i\| \leq \epsilon$ for all $i$ and $\mathbb{E}[A] = I$. If $\gamma \geq 8cd\eta^2(p + \ln d)$ for $0 < \eta \leq 0.25$ and $p \geq 2$, then

(a) $\mathbb{E} \left[ \left| \det(A) - 1 \right|^p \right] \leq 5\eta$ and (b) $\mathbb{E} \left[ \left\| \text{adj}(A) - I \right\|^p \right] \leq 9\eta$.

**Proof** We start by proving (a). Let $X = \det(A) - 1$ and denote $1_{[a,b]}$ as the indicator variable of the event that $X \in [a,b]$. Since $\det(A) \geq 0$, we have:

$$\mathbb{E} \left[ |X|^p \right] = \mathbb{E} \left[ (-X)^p \cdot 1_{[-1,0]} \right] + \mathbb{E} \left[ X^p \cdot 1_{[0,\infty]} \right] \leq \eta^p + \int_{-\eta}^{1} px^{p-1} \Pr(-X \geq x)dx + \int_{0}^{\infty} px^{p-1} \Pr(X \geq x)dx. \quad (3)$$
Thus it suffices to bound the two integrals. We will start with the first one. Let \( X_i = (1 - \frac{b_i}{\gamma})Z_i \). We use the matrix Bernstein inequality to control the extreme eigenvalues of the matrix \( I - A = \sum_i X_i \) (note that matrix \( B \) cancels out because \( I = E[A] = \sum_i Z_i + B \)). To do this, observe that \( \|X_i\| \leq \epsilon/\gamma \) and, moreover, \( E[(1 - \frac{b_i}{\gamma})^2] = \frac{1}{\gamma} - 1 \leq \frac{1}{\gamma} \), so:
\[
\left\| \sum_i E[X_i^2] \right\| = \left\| \sum_i E[(1 - \frac{b_i}{\gamma})^2] Z_i^2 \right\| \leq \frac{1}{\gamma} \cdot \left\| \sum Z_i^2 \right\| \leq \frac{\epsilon}{\gamma} \cdot \left\| \sum Z_i \right\| \leq \frac{\epsilon}{\gamma}.
\]
Thus, applying Lemma \( \text{[8]} \), we conclude that for any \( z \in [\frac{\gamma}{\sqrt{2p}}, 1] \):
\[
\Pr \left( \| I - A \| \geq z \right) \leq 2d e^{-z^2/\gamma} = 2e^{\ln(d) - z^2/\gamma} \leq 2e^{-z^2/2d} \leq 2e^{-z^2/2d} \leq 4 \sqrt{2p} \cdot \eta^p.
\]
(4) Conditioning on the high-probability event given by (4) leads to the lower bound \( \det(A) \geq (1 - z)^d \) which is very loose. To improve on it, we use the following inequality, where \( \delta_1, \ldots, \delta_d \) denote the eigenvalues of \( I - A \):
\[
det(A)e^{\text{tr}(I - A)} = \prod_i (1 - \delta_i) e^{\delta_i} \geq \prod_i (1 - \delta_i)(1 + \delta_i) = \prod_i (1 - \delta_i^2).
\]
Thus we obtain a tighter bound when \( \det(A) \) is multiplied by \( e^{\text{tr}(I - A)} \), and now it suffices to upper bound the latter. This is a simple application of the scalar Bernstein’s inequality (Lemma \( \text{[8]} \) with \( d = 1 \)) for the random variables \( X_i = \text{tr}(X_i) \leq \epsilon/\gamma \leq \frac{\eta^2}{8dp} \), which satisfy \( \sum_i E[X_i^2] \leq \frac{\epsilon}{\gamma} \cdot \text{tr}(\sum_i Z_i) \leq \frac{\epsilon}{\gamma} \cdot \frac{\epsilon}{\gamma} \leq \frac{\epsilon^2}{\gamma} \). Thus the scalar Bernstein’s inequality states that:
\[
\max \left\{ \Pr(\text{tr}(A - I) \geq y), \Pr(\text{tr}(A - I) \leq -y) \right\} \leq \begin{cases} e^{-y^2/2p} & \text{for } y \leq d; \\ e^{-y^2/2d} & \text{for } y \geq d. \end{cases}
\]
(5) Setting \( y = \frac{\gamma}{\sqrt{2p}} \) and \( z = \sqrt{\frac{2p}{\gamma}} \) and taking a union bound over the appropriate high-probability events given by (4) and (5), we conclude that for any \( x \in [\eta, 1] \):
\[
det(A) \geq (1 - z^2)^d \exp(\text{tr}(A - I)) \geq (1 - \frac{\gamma}{\sqrt{2p}})^d \geq 1 - x, \quad \text{with prob. } 1 - 3e^{-z^2/2d}.
\]
Thus, for \( X = \det(A) - 1 \) and \( x \in [\eta, 1] \) we obtain that \( \Pr(-X \geq x) \leq 3e^{-z^2/2d} \), and consequently:
\[
\int_{\eta}^{1} p x^{p-1} \Pr(-X \geq x) dx \leq 3p \int_{\eta}^{1} x^{p-1} e^{-x^2/2d} dx = 3p \sqrt{\frac{\eta^2}{p}} \int_{-\infty}^{\infty} \frac{x^{p-1} e^{-x^2/2d}}{\sqrt{2\pi\eta^2/p}} dx
\]
\[
\leq 3 \sqrt{2\pi\eta^2/p} \cdot \left( \frac{\eta^2}{p} \right)^{p/2} = 3\sqrt{2\pi\eta \cdot p^p}.
\]
We now move on to bounding the remaining integral from (4). Since determinant is the product of eigenvalues, we have \( \det(A) = \det(I + A - I) \leq e^{\text{tr}(A - I)} \), so we can use the Bernstein bound of (5) w.r.t. \( A - I \). It follows that:
\[
\int_{0}^{\infty} p x^{p-1} \Pr(X \geq x) dx \leq \int_{0}^{\infty} p x^{p-1} \Pr \left( e^{\text{tr}(A - I)} \geq 1 + x \right) dx
\]
\[
\leq \int_{0}^{e^d-1} p x^{p-1} e^{-\ln^2(1+x)\frac{2p}{\gamma^2}} dx + \int_{e^d-1}^{\infty} p x^{p-1} e^{-\ln(1+x)\frac{2dp}{\gamma^2}} dx
\]
\[
\leq \int_{0}^{e^d-1} p x^{p-1} e^{-\ln^2(1+x)\frac{2p}{\gamma^2}} dx + \int_{e^d-1}^{\infty} p x^{p-1} e^{-\ln(1+x)\frac{2dp}{\gamma^2}} dx,
\]
because \( \ln^2(1+x) \geq \ln(1+x) \) for \( x \geq e - 1 \). Note that \( \ln^2(1+x) \geq x^2/4 \) for \( x \in [0, e - 1] \), so
\[
\int_{0}^{e^d-1} p x^{p-1} e^{-\ln(1+x)\frac{2p}{\gamma^2}} dx \leq \int_{0}^{e^d-1} p x^{p-1} e^{-x^2/4} dx \leq \sqrt{2\pi p} \cdot \eta^p.
\]
In the interval \( x \in [e - 1, \infty) \), we have:
\[
\int_{e-1}^{\infty} p x^{p-1} e^{-\ln(1+x)\frac{2p}{\gamma^2}} dx = p \int_{e-1}^{\infty} e^{(p-1)\ln(x) - \ln(1+x)\frac{2p}{\gamma^2}} dx \leq p \int_{e-1}^{\infty} e^{-\ln(1+x)\frac{2p}{\gamma^2}} dx
\]
\[
\leq p \int_{\frac{1}{1+e}}^{\infty} \frac{\eta^p}{x^2 - 1} (\frac{1}{2})^{\frac{2p}{\gamma^2}-1} \leq p \cdot \left( \frac{1}{2} \right)^{\frac{1}{2}} \leq p \cdot (\eta^2)^p,
\]
where the last inequality follows because \( (\frac{1}{2})^{\frac{1}{\alpha}} \leq \eta^2 \). Noting that \( (1 + 4\sqrt{2\pi}p + p)^{\frac{1}{\alpha}} \leq 5 \) for any \( p \geq 2 \) concludes the proof of (a). The proof of (b), given in Appendix A, follows similarly as above because for any positive definite \( A \) we have \( \det(A) \leq \frac{\det(A)}{\min\{\lambda_{\min}(A)\}} \cdot I \). 

Having obtained bounds on the higher moments, we can now convert them to convergence with high probability for the average of determinants and the adjugates. Since determinant is a scalar variable, this follows by using standard arguments. On the other hand, for the adjugate matrix we require a somewhat less standard matrix extension of the Khintchine/Rosenthal inequalities (see Appendix A).

**Corollary 10** There is \( C > 0 \) s.t. for \( A \) as in Lemma 9 with all \( Z_i \) rank-1 and \( \gamma \geq C\epsilon d^{-\gamma} \log^3 \frac{n}{d} \).

\[
(a) \text{ Pr}\left( \left| \frac{1}{m} \sum_{t=1}^{m} \det(A_t) - 1 \right| \geq \frac{\eta}{\sqrt{m}} \right) \leq \delta \quad \text{and} \quad (b) \text{ Pr}\left( \left\| \frac{1}{m} \sum_{t=1}^{m} \text{adj}(A_t) - I \right\| \geq \frac{\eta}{\sqrt{m}} \right) \leq \delta,
\]

where \( A_1, \ldots, A_m \) are independent copies of \( A \).

We are ready to show the convergence rate of determinantal averaging, which follows essentially by upper/lower bounding the enumerator and denominator separately, using Corollary 10.

**Proof of Theorem 2** We will apply Corollary 10 to the matrices \( A_t = H^{-\frac{1}{2}} \hat{H}_t H^{-\frac{1}{2}} \). Note that \( A_t = \frac{1}{m} \sum b_i Z_i + \lambda H^{-1} \), where each \( Z_i = \frac{1}{m} \sum b_i Z_i H^{-\frac{1}{2}} \) satisfies \( \|Z_i\| \leq \mu \cdot d/n \). Therefore, Corollary 10 guarantees that for \( \frac{2}{\alpha} \geq C\epsilon d^{-\gamma} \log^3 \frac{n}{d} \), with probability \( 1 - \delta \) the following average of determinants is concentrated around 1:

\[
Z = \frac{1}{m} \sum_{t=1}^{m} \frac{\det(\hat{H}_t)}{\det(H)} = \frac{1}{m} \sum_{t=1}^{m} \det\left( H^{-\frac{1}{2}} \hat{H}_t H^{-\frac{1}{2}} \right) \in [1 - \alpha, 1 + \alpha] \quad \text{for} \quad \alpha = \frac{\eta}{\sqrt{m}},
\]

along with a corresponding bound for the adjugate matrices. We obtain that with probability \( 1 - 2\delta \),

\[
\left\| \frac{\sum_{t=1}^{m} \text{adj}(A_t)}{\sum_{t=1}^{m} \det(A_t)} - I \right\| \leq \left\| \frac{1}{m} \sum_{t=1}^{m} \text{adj}(A_t) - Z I \right\| /Z
\]

\[
\text{Corollary 10(a)} \quad \leq \frac{1}{1 - \alpha} \left\| \frac{1}{m} \sum_{t=1}^{m} \text{adj}(A_t) - I \right\| + \frac{\alpha}{1 - \alpha}
\]

\[
\text{Corollary 10(b)} \quad \leq \frac{\alpha}{1 - \alpha} + \frac{\alpha}{1 - \alpha},
\]

It remains to multiply the above expressions by \( H^{-\frac{1}{2}} \) from both sides to recover the desired estimator:

\[
\sum_{t=1}^{m} \frac{\det(\hat{H}_t)}{\det(H)} H^{-\frac{1}{2}} \leq H^{-\frac{1}{2}} \left( \sum_{t=1}^{m} \frac{\det(\hat{H}_t)}{\det(H)} \right) H^{-\frac{1}{2}} \leq H^{-\frac{1}{2}} \left( 1 + \frac{2\alpha}{1 - \alpha} \right) I H^{-\frac{1}{2}} = \left( 1 + \frac{2\alpha}{1 - \alpha} \right) H^{-1},
\]

and the lower bound follows identically. Appropriately adjusting the constants concludes the proof. 

As an application of the above result, we show how this allows us to bound the estimation error in distributed Newton’s method, when using determinantal averaging.

**Proof of Corollary 4** Follows from Theorem 2 by setting \( Z_t = \ell''(w^* x_t) x_t x_t^\top \) and \( B = \lambda I \). Note that the assumptions imply that \( \|Z_t\| \leq \mu \), so invoking the theorem and denoting \( g \) as \( \nabla \ell(w) \), with probability \( 1 - \delta \) we have

\[
\left\| \frac{\sum_{t=1}^{m} \alpha_t \hat{p}_t}{\sum_{t=1}^{m} \alpha_t} - p \right\|_H = \left\| H^{-\frac{1}{2}} \left( \frac{\sum_{t=1}^{m} \det(\hat{H}_t)}{\sum_{t=1}^{m} \det(H)} \right) H^{-\frac{1}{2}} g \right\|
\leq \left\| H^{-\frac{1}{2}} \left( \frac{\sum_{t=1}^{m} \det(\hat{H}_t)}{\sum_{t=1}^{m} \det(H)} \right) H^{-\frac{1}{2}} \right\| \cdot \|H^{-\frac{1}{2}} g\|
\leq \left\| H^{-\frac{1}{2}} \frac{\eta}{\sqrt{m}} H^{-\frac{1}{2}} \right\| \cdot \|p\|_H = \frac{\eta}{\sqrt{m}} \cdot \|p\|_H.
\]

which completes the proof of the corollary.
4 Conclusions and future directions

We proposed a novel method for correcting the inversion bias in distributed Newton’s method. Our approach, called determinantal averaging, can also be applied more broadly to distributed estimation of other linear functions of the inverse Hessian or an inverse covariance matrix. We show that estimators produced by determinantal averaging are asymptotically consistent, and we provide bounds on the estimation error by developing new moment bounds on the determinant of a random matrix.

Further empirical evaluation of determinantal averaging, both in the context of distributed optimization and other tasks involving inverse estimation, is an important direction for future work. Our preliminary experiments suggest that the bias-correction of determinantal averaging comes at a price of additional variance in the estimators. This leads to a natural open problem: find the optimal balance between bias and variance in weighted averaging for distributed inverse estimation. Finally, note that we construct our Newton estimates using local Hessian and global gradient. In some settings it is more practical to use local approximations for both the Hessian and the gradient. Whether or not determinantal averaging corrects the bias in this case remains open.

Acknowledgements

MWM would like to acknowledge ARO, DARPA, NSF and ONR for providing partial support of this work. Also, MWM and MD thank the NSF for funding via the NSF TRIPODS program. Part of this work was done while MD and MWM were visiting the Simons Institute for the Theory of Computing.

References


