Distributed estimation of the inverse Hessian by determinantal averaging

Michał Dereziński, Michael W. Mahoney

Berkeley UNIVERSITY OF CALIFORNIA

Distributed Newton's method

Task: Minimization of a convex loss:

$$\mathcal{L}(\mathbf{w}) \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} \ell_i(\mathbf{w}^\top \mathbf{x}_i) + \frac{\lambda}{2}\|\mathbf{w}\|^2 \quad \text{for } \mathbf{w} \in \mathbb{R}^d.$$

Goal: Find a good descent direction: $\widetilde{\mathbf{w}} = \mathbf{w} - \mathbf{p}$
Newton's method: use both *Hessian* and *gradient* information,

$$\mathbf{p} = \mathbf{H}^{-1}\mathbf{g}, \quad \text{where} \quad \mathbf{H} = \nabla^2\mathcal{L}(\mathbf{w}), \quad \mathbf{g} = \nabla\mathcal{L}(\mathbf{w}).$$

Distributed Newton: Avoid constructing the full Hessian by replacing it with local approximations computed on separate machines:

$$\widehat{\mathbf{p}}_t = [\underbrace{\nabla^2\widehat{\mathcal{L}}_t(\mathbf{w})}_{\text{local Hessian } \widehat{\mathbf{H}}_t}]^{-1} \underbrace{\nabla\mathcal{L}(\mathbf{w})}_{\text{global gradient } \mathbf{g}} \quad \text{for } t = 1, ..., m,$$

where $\widehat{\mathcal{L}}$ is based on a random sample of data,

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

Question: How to combine local Newton estimates $\widehat{\mathbf{p}}_1, ..., \widehat{\mathbf{p}}_m$?

Problem: Inversion bias

Standard averaging leads to biased estimates:

$$\lim_{m\to\infty} \frac{1}{m}\sum_{t=1}^{m} \widehat{\mathbf{p}}_t \neq \mathbf{p} \qquad (m \text{ is the number of machines})$$

For large m, adding more machines will not improve the accuracy

The reason for this is a general phenomenon, which we call *inversion bias*:

$$\mathbb{E}[\widehat{\mathbf{H}}^{-1}] \neq \mathbf{H}^{-1}, \qquad \text{even though} \quad \mathbb{E}[\widehat{\mathbf{H}}] = \mathbf{H}.$$

Other examples of inversion bias

Consider a data covariance matrix: $\mathbf{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top$.

In *uncertainty quantification* we wish to estimate:

- the trace of $\mathbf{\Sigma}^{-1}$,
- a subset of entries of $\mathbf{\Sigma}^{-1}$.

Again, we encounter inversion bias when averaging estimates.

Determinantal averaging

Goal: Estimate a linear function of inverse Hessian, $F(\mathbf{H}^{-1})$
Given: m independent local estimates $F(\widehat{\mathbf{H}}_1), ..., F(\widehat{\mathbf{H}}_m)$
Newton estimates: $F(\widehat{\mathbf{H}}_t^{-1}) = \widehat{\mathbf{H}}_t^{-1}\mathbf{g} = \widehat{\mathbf{p}}_t$, where $\mathbf{g}$ is the gradient.

Strategy: Weighted average of the estimates,

$$\hat{F}_m = \frac{\sum_{t=1}^{m} a_t F(\widehat{\mathbf{H}}_t^{-1})}{\sum_{t=1}^{m} a_t}$$

Uniform averaging ($a_t = \frac{1}{n}$) suffers from *inversion bias*: $\hat{F}_m \not\to F(\mathbf{H}^{-1})$

Determinantal averaging: use carefully-chosen non-uniform weights
$a_t = \det(\widehat{\mathbf{H}}_t) \quad \Rightarrow \quad$ no inversion bias!

Asymptotically consistent inverse estimator

Determinantal averaging is asymptotically consistent:

$$\lim_{m\to\infty} \frac{\sum_{t=1}^{m} \det(\widehat{\mathbf{H}}_t)\, F(\widehat{\mathbf{H}}_t^{-1})}{\sum_{t=1}^{m}\det(\widehat{\mathbf{H}}_t)} = F(\mathbf{H}^{-1}).$$

Adding more estimates always improves the accuracy

Key underlying expectation formula:

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

Main result: Distributed Newton without inversion bias

Theorem
If expected local sample size satisfies $k \geq C\eta^{-2}\mu d^2\log^3\frac{d}{\delta}$ then

$$\Big\| \frac{\sum_{t=1}^{m} a_t\,\widehat{\mathbf{p}}_t}{\sum_{t=1}^{m} a_t} - \mathbf{p}\Big\|_{\mathbf{H}} \leq \frac{\eta}{\sqrt{m}}\cdot\|\mathbf{p}\|_{\mathbf{H}} \quad \text{with probability } \geq 1-\delta,$$

where $\mu = \frac{1}{d}\max_i \ell_i''(\mathbf{w}^\top\mathbf{x}_i)\|\mathbf{x}_i\|^2_{[\nabla^2\mathcal{L}(\mathbf{w})]^{-1}}$ and $a_t = \det(\widehat{\mathbf{H}}_t)$.

Corollary
A convergence result for Distributed Newton:

$$\|\widetilde{\mathbf{w}} - \mathbf{w}^*\| \leq \max\Big\{\frac{\eta}{\sqrt{m}}\sqrt{\kappa}\,\|\mathbf{w}-\mathbf{w}^*\|,\ \frac{2L}{\lambda_{\min}}\|\mathbf{w}-\mathbf{w}^*\|^2\Big\}$$

$$\text{for} \quad \widetilde{\mathbf{w}} = \mathbf{w} - \frac{\sum_{t=1}^{m} a_t\,\widehat{\mathbf{p}}_t}{\sum_{t=1}^{m} a_t} \quad \text{and} \quad \mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}}\,\mathcal{L}(\mathbf{w}).$$

$L, \kappa, \lambda_{\min}$ - Lipschitz constant, condition number and smallest eigenvalue of $\mathbf{H}$.

Proof techniques

Key lemma
If $\widehat{\mathbf{H}} = \sum_i s_i\mathbf{Z}_i$, where s_i are independent random variables and $\mathbf{Z}_i$ are fixed square rank-1 matrices, then

$$\text{(a)} \quad \underbrace{\mathbb{E}[\det(\widehat{\mathbf{H}})] = \det(\mathbb{E}[\widehat{\mathbf{H}}])}_{\text{determinant commutes with expectation}} \quad \text{and} \quad \text{(b)} \quad \underbrace{\mathbb{E}[\operatorname{adj}(\widehat{\mathbf{H}})] = \operatorname{adj}(\mathbb{E}[\widehat{\mathbf{H}}])}_{\text{adjugate commutes with expectation}}.$$

Adjugate matrix: $\operatorname{adj}(\mathbf{A}) = \det(\mathbf{A})\,\mathbf{A}^{-1}$ for any invertible $\mathbf{A}$

Main result relies on showing an improved *matrix concentration inequality*:

$$\Big(1 - \frac{\eta}{\sqrt{m}}\Big)\cdot\mathbf{H}^{-1} \preceq \frac{\sum_{t=1}^{m}\det(\widehat{\mathbf{H}}_t)\,\widehat{\mathbf{H}}_t^{-1}}{\sum_{t=1}^{m}\det(\widehat{\mathbf{H}}_t)} \preceq \Big(1 + \frac{\eta}{\sqrt{m}}\Big)\cdot\mathbf{H}^{-1}$$

Experiment

Newton step estimation error versus number of machines m