# 1 Problem

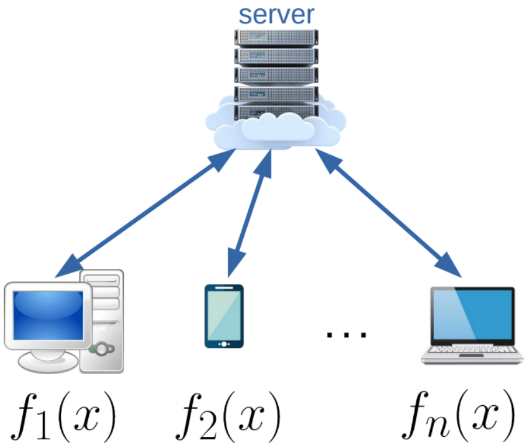

Figure 1: Model of centralized distributed framework.

We consider L2 regularized empirical risk minimization problems of the form

$$\min_{x\in\mathbb{R}^d}\left[P(x) := f(x) + \frac{\lambda}{2}\|x\|^2\right], \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a smooth[1] convex function of the "average of averages" structure

$$f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad f_i(x) := \frac{1}{m}\sum_{j=1}^{m} f_{ij}(x), \tag{2}$$

and $\lambda \geq 0$ is a regularization parameter. Here $n$ is the number of parallel workers (nodes), and $m$ is the number of training examples handled by each node[2]. The value $f_{ij}(x)$ denotes the loss of the model parameterized by vector $x \in \mathbb{R}^d$ on the $j^{\text{th}}$ example owned by the $i^{\text{th}}$ node. This example is denoted as $a_{ij} \in \mathbb{R}^d$, and the corresponding loss function is $\varphi_{ij} : \mathbb{R} \to \mathbb{R}$, and hence we have

$$f_{ij}(x) := \varphi_{ij}(a_{ij}^\top x). \tag{3}$$

Thus, $f$ represents the average loss/risk over all $nm$ training datapoints, and problem (1) seeks to find the model whose (L2 regularized) empirical risk is minimized. We make the following assumption.

**Assumption 1.1.** *Problem (1) has at least one optimal solution $x^*$. For all $i$ and $j$, the loss function $\varphi_{ij} : \mathbb{R} \to \mathbb{R}$ is twice differentiable, and its second derivative $\varphi''_{ij} : \mathbb{R} \to \mathbb{R}$ is $\nu$-Lipschitz continuous.*

[1] Function $\phi : \mathbb{R}^d \to \mathbb{R}$ is *smooth* if it is differentiable, and has $L_\phi$ Lipschitz gradient: $\|\nabla\phi(x) - \nabla\phi(y)\| \leq L_\phi\|x - y\|$ for all $x, y \in \mathbb{R}^d$. We say that $L_\phi$ is the *smoothness constant* of $\phi$.

[2] All our results can be extended in a straightforward way to the more general case when node $i$ contains $m_i$ training examples. We decided to present the results in the special case $m = m_i$ for all $i$ in order to simplify the notation.

Note that in view of (3), the Hessian of $f_{ij}$ at point $x$ is

$$\mathbf{H}_{ij}(x) := \nabla^2 f_{ij}(x) = h_{ij}(x) a_{ij} a_{ij}^\top, \tag{4}$$

where

$$h_{ij}(x) := \varphi_{ij}''(a_{ij}^\top x). \tag{5}$$

In view of Assumption 1.1, we have $|\varphi_{ij}''(t)| \leq \gamma$ for all $t \in \mathbb{R}$, and

$$|h_{ij}(x) - h_{ij}(y)| \leq \nu |a_{ij}^\top x - a_{ij}^\top y| \leq \nu \|a_{ij}\| \|x - y\| \tag{6}$$

for all $x, y \in \mathbb{R}^d$. Let $R := \max_{ij} \|a_{ij}\|$. The Hessian of $f_i$ is given by

$$\mathbf{H}_i(x) \overset{(2)}{=} \frac{1}{m} \sum_{j=1}^m \mathbf{H}_{ij}(x) \overset{(4)}{=} \frac{1}{m} \sum_{j=1}^m h_{ij}(x) a_{ij} a_{ij}^\top, \tag{7}$$

and the Hessian of $f$ is given by

$$\mathbf{H}(x) \overset{(2)}{=} \frac{1}{n} \sum_{i=1}^n \mathbf{H}_i(x) \overset{(7)}{=} \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m h_{ij}(x) a_{ij} a_{ij}^\top. \tag{8}$$

## 2 NEWTON-STAR: Newton's method with a single Hessian

We now introduce a simple idea which, surprisingly, enables us to *remove the need to iteratively communicate any coefficients altogether.* Assume, for the sake of argument, that we know the values $h_{ij}(x^*)$ for all $i, j$. That is, assume the server has access to coefficients $h_{ij}(x^*)$ for all $i, j$, and that each node $i$ has access to coefficients $h_{ij}(x^*)$ for $j = 1, \ldots, m$, i.e., to the vector

$$h_i(x) := (h_{i1}(x), \ldots, h_{im}(x)) \in \mathbb{R}^m \tag{9}$$

for $x = x^*$. Next, consider the following new Newton-like method which we call NEWTON-STAR (NS), where the "star" points to the method's reliance on the knowledge of the optimal solution $x^*$:

$$x^{k+1} \;=\; x^k - \left(\nabla^2 P(x^*)\right)^{-1} \nabla P(x^k) \overset{(1)}{=} x^k - \left(\mathbf{H}(x^*) + \lambda \mathbf{I}\right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \nabla f_i(x^k) + \lambda x^k \right). \tag{10}$$

Since the server knows $\mathbf{H}(x^*)$, all that the nodes need to communicate are the local gradients $\nabla f_i(x^k)$, which costs $O(d)$ per node. The server then computes $x^{k+1}$, broadcasts it back to the nodes, and the process is repeated. This method has the same per-iteration $O(d)$ communication complexity as GD. However, as we show next, the number of iterations (which is the same as the number of communications) of NEWTON-STAR does not depend on the condition number – a property it borrows from the classical Newton's method. The following theorem says that NEWTON-STAR enjoys *local quadratic convergence.*

**Theorem 2.1** (Local quadratic convergence)**.** *Let Assumption 1.1 hold, and assume that $\mathbf{H}(x^*) \succeq \mu^* \mathbf{I}$ for some $\mu^* \geq 0$ (for instance, this holds if $f$ is $\mu^*$-strongly convex) and that $\mu^* + \lambda > 0$. Then for any starting point $x^0 \in \mathbb{R}^d$, the iterates of* NEWTON-STAR *for solving problem* (1) *satisfy the following inequality:*

$$\|x^{k+1} - x^*\| \leq \frac{\nu}{2(\mu^* + \lambda)} \cdot \left( \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \|a_{ij}\|^3 \right) \cdot \|x^k - x^*\|^2. \tag{11}$$

*Proof.* By the first order optimality conditions, we have

$$\nabla f(x^*) + \lambda x^* = 0. \tag{12}$$

Let $\mathbf{H}_* := \mathbf{H}(x^*)$. Since $\mathbf{H}_* \succeq \mu^* \mathbf{I}$, we have $\mathbf{H}_* + \lambda \mathbf{I} \succeq (\mu^* + \lambda)\mathbf{I}$, and hence

$$\left\| (\mathbf{H}_* + \lambda \mathbf{I})^{-1} \right\| \le \frac{1}{\mu^* + \lambda}. \tag{13}$$

Using (12) and (13) and subsequently applying Jensen's inequality to the function $x \mapsto \|x\|$, we get

$$\begin{aligned}
\|x^{k+1} - x^*\| &= \left\| x^k - x^* - (\mathbf{H}_* + \lambda\mathbf{I})^{-1} \nabla P(x^k) \right\| \\
&\overset{(12)}{=} \left\| (\mathbf{H}_* + \lambda\mathbf{I})^{-1} \left[ (\mathbf{H}_* + \lambda\mathbf{I})(x^k - x^*) - \left( \nabla f(x^k) - \nabla f(x^*) + \lambda(x^k - x^*) \right) \right] \right\| \\
&\overset{(13)}{\le} \frac{1}{\mu^* + \lambda} \left\| (\mathbf{H}_* + \lambda\mathbf{I})(x^k - x^*) - \left( \nabla f(x^k) - \nabla f(x^*) \right) - \lambda(x^k - x^*) \right\| \\
&= \frac{1}{\mu^* + \lambda} \left\| \frac{1}{n} \sum_{i=1}^n \mathbf{H}_i(x^*)(x^k - x^*) - \frac{1}{n} \sum_{i=1}^n \left( \nabla f_i(x^k) - \nabla f_i(x^*) \right) \right\| \\
&\le \frac{1}{n(\mu^* + \lambda)} \sum_{i=1}^n \left\| \mathbf{H}_i(x^*)(x^k - x^*) - \left( \nabla f_i(x^k) - \nabla f_i(x^*) \right) \right\| \\
&\overset{(7)+(2)}{=} \frac{1}{n(\mu^* + \lambda)} \sum_{i=1}^n \left\| \frac{1}{m} \sum_{j=1}^m h_{ij}(x^*) a_{ij} a_{ij}^\top (x^k - x^*) - \frac{1}{m} \sum_{j=1}^m \left( \nabla f_{ij}(x^k) - \nabla f_{ij}(x^*) \right) \right\|. \qquad (14)
\end{aligned}$$

We now use the fundamental theorem of calculus to express difference of gradients $\nabla f_{ij}(x^k) - \nabla f_{ij}(x^*)$ in an integral, obtaining

$$\nabla f_{ij}(x^k) - \nabla f_{ij}(x^*) = \int_0^1 \nabla^2 f_{ij}(x^* + \tau(x^k - x^*))(x^k - x^*) d\tau. \tag{15}$$

Plugging this representation into (14) and noting that $\nabla^2 f_{ij}(x) \equiv \mathbf{H}_{ij}(x)$ (see (4)), we can continue:

$$\begin{aligned}
\|x^{k+1} - x^*\| &\overset{(14)+(15)}{\le} \frac{1}{n(\mu^* + \lambda)} \sum_{i=1}^n \left\| \frac{1}{m} \sum_{j=1}^m \left( h_{ij}(x^*) a_{ij} a_{ij}^\top (x^k - x^*) - \int_0^1 \mathbf{H}_{ij}(x^* + \tau(x^k - x^*))(x^k - x^*) d\tau \right) \right\| \\
&\overset{(4)}{=} \frac{1}{n(\mu^* + \lambda)} \sum_{i=1}^n \left\| \frac{1}{m} \sum_{j=1}^m \left( h_{ij}(x^*) a_{ij} a_{ij}^\top (x^k - x^*) - \int_0^1 h_{ij}(x^* + \tau(x^k - x^*)) a_{ij} a_{ij}^\top (x^k - x^*) d\tau \right) \right\| \\
&= \frac{1}{n(\mu^* + \lambda)} \sum_{i=1}^n \left\| \frac{1}{m} \sum_{j=1}^m a_{ij} a_{ij}^\top (x^k - x^*) \left( h_{ij}(x^*) - \int_0^1 h_{ij}(x^* + \tau(x^k - x^*)) d\tau \right) \right\| \\
&\le \frac{\|x^k - x^*\|}{(\mu^* + \lambda)} \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \|a_{ij}\|^2 \left| \int_0^1 h_{ij}(x^*) - h_{ij}(x^* + \tau(x^k - x^*)) d\tau \right|. \qquad (16)
\end{aligned}$$

In the last step we have again used Jensen's inequality applied to the function $x \mapsto \|x\|$, followed by inequalities of the form $\|\mathbf{A}_{ij} x t_{ij}\| \le \|\mathbf{A}_{ij}\| \, \|x\| \, |t_{ij}|$ for $\mathbf{A}_{ij} = a_{ij} a_{ij}^\top$, $x = x^k - x^*$ and $t_{ij} \in \mathbb{R}$.
From (6) we obtain $|h_{ij}(x^*) - h_{ij}(x^* + \tau(x^k - x^*))| \le \nu\tau \|a_{ij}\| \cdot \|x^k - x^*\|$, which implies that

$$\left| \int_0^1 h_{ij}(x^*) - h_{ij}(x^* + \tau(x^k - x^*)) d\tau \right| \le \int_0^1 \nu\tau \|a_{ij}\| \cdot \|x^k - x^*\| d\tau = \frac{\nu \|a_{ij}\|}{2} \cdot \|x^k - x^*\|.$$

Plugging this into (16), we finally arrive at (11). □

Note that we do not need to assume $f$ to be convex or strongly convex. All we need to assume is positive definiteness of the Hessian at the optimum. This implies local strong convexity, and since our convergence result is local, that is all we need.