Supplementary Information for Dual-LOC: Distributing Statistical Estimation Using Random Projections

A Supplementary Results

Here we introduce two lemmas. The first describes the random projection construction which we use in the distributed setting.

**Lemma 2** (Summing random features). Consider the singular value decomposition $X = USV^T$ where $U \in \mathbb{R}^{n \times r}$ and $V \in \mathbb{R}^{p \times r}$ have orthonormal columns and $\Sigma \in \mathbb{R}^{r \times r}$ is diagonal; $r = \text{rank}(X)$. $c_0$ is a fixed positive constant. In addition to the raw features, let $X_k \in \mathbb{R}^{n \times (r + \tau_{subs})}$ contain random features which result from summing the $K - 1$ random projections from the other workers. Furthermore, assume without loss of generality that the problem is permuted so that the raw features of worker $k$’s problem are the first $\tau$ columns of $X$ and $X_k$. Finally, let

$$\Theta_S = \begin{bmatrix} I_r & 0 \\ 0 & \Pi \end{bmatrix} \in \mathbb{R}^{p \times (r + \tau_{subs})}$$

such that $X_k = X \Theta_S$.

With probability at least $1 - (\delta + \frac{\rho - 1}{\rho})$

$$\|V^T \Theta_S ^2 V - V^T V\|_2 \leq \sqrt{\frac{c_0 \log(2r/\delta)r}{\tau_{subs}}}.$$  

**Proof.** See Appendix B.

**Definition 1.** For ease of exposition, we shall rewrite the dual problems so that we consider minimizing convex objective functions. More formally, the original problem is then given by

$$\alpha^* = \arg\min_{\alpha \in \mathbb{R}^n} \left\{ D(\alpha) := \sum_{i=1}^n f_i^*(\alpha_i) + \frac{1}{2n\lambda} \alpha^T K\alpha \right\}. \tag{9}$$

The problem worker $k$ solves is described by

$$\tilde{\alpha} = \arg\min_{\alpha \in \mathbb{R}^n} \left\{ \tilde{D}_k(\alpha) := \sum_{i=1}^n f_i^*(\alpha_i) + \frac{1}{2n\lambda} \tilde{K}_k\alpha \right\}. \tag{10}$$

Recall that $\tilde{K}_k = \tilde{X}_k \tilde{X}_k^T$, where $\tilde{X}_k$ is the concatenation of the $\tau$ raw features and $\tau_{subs}$ random features for worker $k$.

To proceed we need the following result which relates the solution of the original problem to that of the approximate problem solved by worker $k$.

**Lemma 3** (Adapted from Lemma 1 [17]). Let $\alpha^*$ and $\tilde{\alpha}$ be as defined in Definition 1. We obtain

$$\frac{1}{\lambda}(\tilde{\alpha} - \alpha^*)^T (K - \tilde{K}_k)\alpha^* \geq \frac{1}{\lambda}(\tilde{\alpha} - \alpha^*)^T \tilde{K}_k(\tilde{\alpha} - \alpha^*). \tag{11}$$

**Proof.** See [17].

For our main result, we rely heavily on the following variant of Theorem 1 in [17] which bounds the difference between the coefficients estimated by worker $k$, $\beta_k$ and the corresponding coordinates of the optimal solution vector $\beta^*_k$.

**Lemma 4** (Local optimization error. Adapted from [17]). For $\rho = \sqrt{\frac{c_0 \log(2r/\delta)r}{\tau_{subs}}}$, the following holds

$$\|\tilde{\beta}_k - \beta_k\|_2 \leq \frac{\rho}{1 - \rho} \|\beta^*_k\|_2$$

with probability at least $1 - (\delta + \frac{\rho - 1}{\rho})$.

The proof closely follows the proof of Theorem 1 in [17] which we restate here identifying the major differences.

**Proof.** Let the quantities $\tilde{D}_k(\alpha_k, \tilde{K}_k)$, be as in Definition 1. For ease of notation, we shall omit the subscript $k$ in $\tilde{D}_k(\alpha)$ and $\tilde{K}_k$ in the following.

By the SVD we have $X = USV^T$. So $K_k = U\Sigma_k U^T$ and $\tilde{K}_k = \tilde{U}\Sigma_k \tilde{U}^T$. We can make the following definitions

$$\gamma^* = \Sigma U^T \alpha^*, \quad \tilde{\gamma} = \Sigma U^T \tilde{\alpha}.$$ 

Defining $\tilde{M} = V^T \Pi \Pi^T V$ and plugging these into Lemma 3 we obtain

$$(\tilde{\gamma} - \gamma^*)^T (I - \tilde{M}) \gamma^* \geq (\tilde{\gamma} - \gamma^*)^T \tilde{M}(\tilde{\gamma} - \gamma^*). \tag{12}$$

We now bound the spectral norm of $I - \tilde{M}$ using Lemma 2. Recall that Lemma 2 bounds the difference between a matrix and its approximation by a distributed dimensionality reduction using the SRHT.

Using the Cauchy-Schwarz inequality we have for the l.h.s. of (12)

$$(\tilde{\gamma} - \gamma^*)^T (I - \tilde{M}) \gamma^* \leq \rho \|\gamma^*\|_2 \|\tilde{\gamma} - \gamma^*\|_2$$

For the r.h.s. of (12), we can write

$$(\tilde{\gamma} - \gamma^*)^T \tilde{M}(\tilde{\gamma} - \gamma^*)$$

$$= \|\tilde{\gamma} - \gamma^*\|_2^2 - (\tilde{\gamma} - \gamma^*)^T (I - \tilde{M})(\tilde{\gamma} - \gamma^*)$$

$$\geq \|\tilde{\gamma} - \gamma^*\|_2^2 - \rho \|\tilde{\gamma} - \gamma^*\|_2^2$$

$$= (1 - \rho) \|\tilde{\gamma} - \gamma^*\|_2^2.$$
Combining these two expressions and inequality (12) yields
\[
(1 - \rho)\|\hat{\gamma} - \gamma^*\|^2_2 \leq \rho\|\gamma^*\|_2\|\hat{\gamma} - \gamma^*\|_2 \\
(1 - \rho)\|\hat{\gamma} - \gamma^*\|_2 \leq \rho\|\gamma^*\|_2.
\] (13)

From the definition of \(\gamma^*\) and \(\hat{\gamma}\) above and \(\beta^*\) and \(\hat{\beta}\), respectively we have
\[
\beta^* = \frac{1}{n\lambda} V\gamma^*, \quad \hat{\beta} = \frac{1}{n\lambda} V\hat{\gamma}
\]
so \(\frac{1}{n\lambda}\|\gamma^*\|_2 = \|\beta^*\|_2\) and \(\|\hat{\beta} - \beta^*\|_2 = \frac{1}{n\lambda}\|\hat{\gamma} - \gamma^*\|_2\) due to the orthonormality of \(V\). Plugging this into (13) and using the fact that \(\|\beta^* - \hat{\beta}\|_2 \geq \|\beta_k^* - \hat{\beta}_k\|_2\) we obtain the stated result. \(\square\)

**B Proof of Row Summing Lemma**

*Proof of Lemma 2.* Let \(V_k\) contain the first \(\tau\) rows of \(V\) and let \(V_{(-k)}\) be the matrix containing the remaining rows. Decompose the matrix products as follows
\[
V^T V = V_k^T V_k + V_{(-k)}^T V_{(-k)}
\]
and
\[
V^T \Theta_S \Theta_S^T V = V_k^T V_k + \hat{V}_k^T \hat{V}_k
\]
with \(\hat{V}_k^T = V_{(-k)}^T \Pi\). Then
\[
\|V^T \Theta_S \Theta_S^T V - V^T V\|_2 \\
= \|V_k^T V_k + \hat{V}_k^T \hat{V}_k - V_k^T V_k - V_{(-k)}^T V_{(-k)}\|_2 \\
= \|V_{(-k)}^T \Pi \Pi^T V_{(-k)} - V_{(-k)}^T V_{(-k)}\|_2.
\]

Since \(\Theta_S\) is an orthogonal matrix, from Lemma 3.3 in [13] and Lemma 5, summing \((K - 1)\) independent SRHTs from \(\tau\) to \(\tau_{subs}\) is equivalent to applying a single SRHT from \(p - \tau\) to \(\tau_{subs}\). Therefore we can simply apply Lemma 1 of [15] to the above to obtain the result. \(\square\)

**Lemma 5 (Summed row sampling).** Let \(W\) be an \(n \times p\) matrix with orthonormal columns. Let \(W_1, \ldots, W_K\) be a balanced, random partitioning of the rows of \(W\) where each matrix \(W_k\) has exactly \(\tau = n/K\) rows. Define the quantity \(M := n \cdot \max_j \|w_j\|_2^2\). For a positive parameter \(\alpha\), select the subsample size
\[
l \cdot K \geq \alpha M \log(p).
\]
Let \(S_{T_k} \in \mathbb{R}^{l \times \tau}\) denote the operation of uniformly at random sampling a subset, \(T_k\) of the rows of \(W_k\) by sampling \(l\) coordinates from \(\{1, 2, \ldots, \tau\}\) without replacement. Now denote \(SW\) as the sum of the subsampled rows
\[
SW = \sum_{k=1}^K (S_{T_k} W_k).
\]

Then
\[
\sqrt{\frac{(1 - \delta)l \cdot K}{n}} \leq \sigma_p(SW)
\]
and
\[
\sigma_1(SW) \leq \sqrt{\frac{(1 + \eta)l \cdot K}{n}}
\]
with failure probability at most
\[
p \cdot \left[ \frac{e^{-\delta}}{(1 - \delta)^{1 - \delta}} \right]^{\alpha \log p} + p \cdot \left[ \frac{e^{\eta}}{(1 + \eta)^{1 + \eta}} \right]^{\alpha \log p}
\]

*Proof.* Define \(w_j^T\) as the \(j\)th row of \(W\) and \(M := n \cdot \max_j \|w_j\|_2^2\). Suppose \(K = 2\) and consider the matrix
\[
G_2 := (S_1 W_1 + S_2 W_2)^T (S_1 W_1 + S_2 W_2) \\
= (S_1 W_1)^T (S_1 W_1) + (S_2 W_2)^T (S_2 W_2) \\
+ (S_1 W_1)^T (S_2 W_2) + (S_2 W_2)^T (S_1 W_1).
\]

In general, we can express \(G := (SW)^T (SW)\) as
\[
G := \sum_{k=1}^K \sum_{j \in T_k} (w_j w_j^T + \sum_{k' \neq k, j' \in T_{k'}} w_j w_{j'}^T).
\]

By the orthonormality of \(W\), the cross terms cancel as \(w_j w_{j'}^T = 0\), yielding
\[
G := (SW)^T (SW) = \sum_{k=1}^K \sum_{j \in T_k} w_j w_j^T.
\]

We can consider \(G\) as a sum of \(l \cdot K\) random matrices
\[
X^{(1)}_1, \ldots, X^{(K)}_1, \ldots, X^{(1)}_{\tau}, \ldots, X^{(K)}_{\tau}
\]
sampled uniformly at random without replacement from the family \(X := \{w_i w_i^T : i = 1, \ldots, \tau \cdot K\}\).

To use the matrix Chernoff bound in Lemma 6, we require the quantities \(\mu_{\min}, \mu_{\max}\) and \(B\). Noticing that \(\lambda_{\max}(w_j w_j^T) = \|w_j\|_2^2 \leq \frac{M}{n}\), we can set \(B = \frac{M}{n}\).

Taking expectations with respect to the random partitioning \((\mathbb{E}_P)\) and the subsampling within each partition \((\mathbb{E}_S)\), using the fact that columns of \(W\) are orthonormal we obtain
\[
\mathbb{E} \left[ X^{(k)}_1 \right] = \mathbb{E}_P \mathbb{E}_S X^{(k)}_1 = \frac{1}{K \tau} \sum_{i=1}^{K \tau} w_i w_i^T = \frac{1}{n} W^T W = \frac{1}{n} I
\]

Recall that we take \(l\) samples in \(K\) blocks so we can define
\[
\mu_{\min} = \frac{l \cdot K}{n} \quad \text{and} \quad \mu_{\max} = \frac{l \cdot K}{n}.
\]
Plugging these values into Lemma 6, the lower and upper Chernoff bounds respectively yield

\[
P\left\{ \lambda_{\min}(G) \leq (1 - \delta) \frac{l \cdot K}{n} \right\} \leq p \cdot \left[ \frac{e^{-\delta}}{(1 - \delta)^{1+\delta}} \right]^{l \cdot K / M} \quad \text{for } \delta \in [0, 1], \text{ and}
\]

\[
P\left\{ \lambda_{\max}(G) \geq (1 + \delta) \frac{l \cdot K}{n} \right\} \leq p \cdot \left[ \frac{e^{\delta}}{(1 + \delta)^{1+\delta}} \right]^{l \cdot K / M} \quad \text{for } \delta \geq 0.
\]

Noting that \( \lambda_{\min}(G) = \sigma_p(G)^2 \), similarly for \( \lambda_{\max} \) and using the identity for \( G \) above obtains the desired result. \( \square \)

For ease of reference, we also restate the Matrix Chernoff bound from [13, 24] but defer its proof to the original papers.

**Lemma 6 (Matrix Chernoff from [13]):** Let \( \mathcal{X} \) be a finite set of positive-semidefinite matrices with dimension \( p \), and suppose that

\[
\max_{A \in \mathcal{X}} \lambda_{\max}(A) \leq B
\]

Sample \( \{A_1, \ldots, A_l\} \) uniformly at random from \( \mathcal{X} \) without replacement. Compute

\[
\mu_{\min} = l \cdot \lambda_{\min}(EX_1) \quad \text{and} \quad \mu_{\max} = l \cdot \lambda_{\max}(EX_1)
\]

Then

\[
P \left\{ \lambda_{\min} \left( \sum_i A_i \right) \leq (1 - \delta) \mu_{\min} \right\} \leq p \cdot \left[ \frac{e^{-\delta}}{(1 - \delta)^{1+\delta}} \right]^{\mu_{\min} / B} \quad \text{for } \delta \in [0, 1], \text{ and}
\]

\[
P \left\{ \lambda_{\max} \left( \sum_i A_i \right) \geq (1 + \delta) \mu_{\max} \right\} \leq p \cdot \left[ \frac{e^{\delta}}{(1 + \delta)^{1+\delta}} \right]^{\mu_{\max} / B} \quad \text{for } \delta \geq 0.
\]

### Table 1: Dogs vs Cats data: Normalized training and test MSE: mean and standard deviations (based on 5 repetitions).

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>K</th>
<th>TEST MSE</th>
<th>TRAIN MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEAL-LOCO 0.5</td>
<td>12</td>
<td>0.0300 (4.26e-03)</td>
<td>0.0300 (4.26e-03)</td>
</tr>
<tr>
<td>DEAL-LOCO 0.5</td>
<td>24</td>
<td>0.0300 (4.26e-03)</td>
<td>0.0300 (4.26e-03)</td>
</tr>
<tr>
<td>DEAL-LOCO 0.5</td>
<td>48</td>
<td>0.0300 (4.26e-03)</td>
<td>0.0300 (4.26e-03)</td>
</tr>
<tr>
<td>DEAL-LOCO 0.5</td>
<td>96</td>
<td>0.0300 (4.26e-03)</td>
<td>0.0300 (4.26e-03)</td>
</tr>
<tr>
<td>DEAL-LOCO 0.5</td>
<td>192</td>
<td>0.0300 (4.26e-03)</td>
<td>0.0300 (4.26e-03)</td>
</tr>
<tr>
<td>DEAL-LOCO 1</td>
<td>1</td>
<td>0.0310 (4.26e-03)</td>
<td>0.0310 (4.26e-03)</td>
</tr>
<tr>
<td>DEAL-LOCO 1</td>
<td>24</td>
<td>0.0300 (4.26e-03)</td>
<td>0.0300 (4.26e-03)</td>
</tr>
<tr>
<td>DEAL-LOCO 1</td>
<td>48</td>
<td>0.0300 (4.26e-03)</td>
<td>0.0300 (4.26e-03)</td>
</tr>
<tr>
<td>DEAL-LOCO 1</td>
<td>96</td>
<td>0.0300 (4.26e-03)</td>
<td>0.0300 (4.26e-03)</td>
</tr>
<tr>
<td>DEAL-LOCO 2</td>
<td>12</td>
<td>0.0300 (4.26e-03)</td>
<td>0.0300 (4.26e-03)</td>
</tr>
<tr>
<td>DEAL-LOCO 2</td>
<td>24</td>
<td>0.0300 (4.26e-03)</td>
<td>0.0300 (4.26e-03)</td>
</tr>
<tr>
<td>DEAL-LOCO 2</td>
<td>48</td>
<td>0.0300 (4.26e-03)</td>
<td>0.0300 (4.26e-03)</td>
</tr>
<tr>
<td>DEAL-LOCO 2</td>
<td>96</td>
<td>0.0300 (4.26e-03)</td>
<td>0.0300 (4.26e-03)</td>
</tr>
<tr>
<td>DEAL-LOCO 2</td>
<td>192</td>
<td>0.0300 (4.26e-03)</td>
<td>0.0300 (4.26e-03)</td>
</tr>
</tbody>
</table>

### Algorithm 2 Dual-LOCO – cross validation

**Input:** Data: \( X, Y \), no. workers: \( K \), no. folds: \( v \)

**Parameters:** \( \tau_{subs} \), \( \lambda_1, \ldots, \lambda_l \)

1. Partition \( \{p\} \) into \( K \) subsets of equal size \( \tau \) and distribute feature vectors in \( X \) accordingly over \( K \) workers.
2. Partition \( \{n\} \) into \( v \) folds of equal size.
3. for each fold \( f \) do
4. Communicate indices of training and test points.
5. for each worker \( k \in \{1, \ldots, K\} \) in parallel do
6. Compute and send \( X_k^{train}, Y_k^{train} \) to driver.
7. Receive random features and construct \( X_k^{test} \).
8. for each \( \lambda_j \in \{\lambda_1, \ldots, \lambda_l\} \) do
9. \( \hat{\alpha}_{k,f,j} \leftarrow \text{LocalDualSolver}(X_k^{train}, Y_k^{train}, \lambda_j) \)
10. \( \hat{\beta}_{k,f,j} = -\frac{1}{n\lambda_j} X_k^{test} \hat{\alpha}_{k,f,j} \)
11. \( Y_k^{test} = X_k^{test} \hat{\beta}_{k,f,j} \)
12. Send \( Y_k^{test} \) to driver.
13. end for
14. end for
15. for each \( \lambda_j \in \{\lambda_1, \ldots, \lambda_l\} \) do
16. Compute \( Y_k^{test} \) with \( Y_k^{test} \) and \( Y_k^{test} \).
17. Compute MSE with MSE and MSE.
18. end for
19. end for
20. for each \( \lambda_j \in \{\lambda_1, \ldots, \lambda_l\} \) do
21. Compute MSE\( \lambda_j \).
22. end for

**Output:** Parameter \( \lambda_j \) attaining smallest MSE\( \lambda_j \)