Supplementary Information for DUAL-LOCO: 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 $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ where $\mathbf{U} \in \mathbb{R}^{n\times r}$ and $\mathbf{V} \in \mathbb{R}^{p\times r}$ have orthonormal columns and $\mathbf{\Sigma} \in \mathbb{R}^{r\times r}$ is diagonal; $r = rank(\mathbf{X})$. c_0 is a fixed positive constant. In addition to the raw features, let $\bar{\mathbf{X}}_k \in \mathbb{R}^{n\times(\tau+\tau_{subs})}$ 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 τ columns of $\mathbf{X}$ and $\bar{\mathbf{X}}_k$. Finally, let*

$$\Theta_S = \begin{bmatrix} \mathbf{I}_\tau & 0 \\ 0 & \mathbf{\Pi} \end{bmatrix} \in \mathbb{R}^{p\times(\tau+\tau_{subs})}$$

such that $\bar{\mathbf{X}}_k = \mathbf{X}\Theta_S$.

With probability at least $1 - \left(\delta + \frac{p-\tau}{e^r}\right)$

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

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*

$$\boldsymbol{\alpha}^* = \underset{\boldsymbol{\alpha}\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ D(\boldsymbol{\alpha}) := \sum_{i=1}^n f_i^*(\alpha_i) + \frac{1}{2n\lambda}\boldsymbol{\alpha}^\top\mathbf{K}\boldsymbol{\alpha} \right\}. \tag{9}$$

The problem worker k solves is described by

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

Recall that $\tilde{\mathbf{K}}_k = \bar{\mathbf{X}}_k\bar{\mathbf{X}}_k^\top$, where $\bar{\mathbf{X}}_k$ is the concatenation of the τ raw *features and τ_{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 $\boldsymbol{\alpha}^*$ and $\tilde{\boldsymbol{\alpha}}$ be as defined in Definition 1. We obtain*

$$\frac{1}{\lambda}(\tilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*)^\top\left(\mathbf{K}-\tilde{\mathbf{K}}_k\right)\boldsymbol{\alpha}^* \ge \frac{1}{\lambda}(\tilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*)^\top\tilde{\mathbf{K}}_k(\tilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*). \tag{11}$$

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, $\widehat{\boldsymbol{\beta}}_k$ and the corresponding coordinates of the optimal solution vector $\boldsymbol{\beta}^*_k$.

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

$$\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}^*_k\|_2 \le \frac{\rho}{1-\rho}\|\boldsymbol{\beta}^*\|_2$$

with probability at least $1 - \left(\delta + \frac{p-\tau}{e^r}\right)$.

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

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

By the SVD we have $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. So $\mathbf{K} = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}\mathbf{U}^\top$ and $\tilde{\mathbf{K}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{\Pi}\mathbf{\Pi}^\top\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top$. We can make the following definitions

$$\gamma^* = \mathbf{\Sigma}\mathbf{U}^\top\boldsymbol{\alpha}^*, \qquad \tilde{\gamma} = \mathbf{\Sigma}\mathbf{U}^\top\tilde{\boldsymbol{\alpha}}.$$

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

$$(\tilde{\gamma}-\gamma^*)^\top(\mathbf{I}-\tilde{\mathbf{M}})\gamma^* \ge (\tilde{\gamma}-\gamma^*)^\top\tilde{\mathbf{M}}(\tilde{\gamma}-\gamma^*). \tag{12}$$

We now bound the spectral norm of $\mathbf{I} - \tilde{\mathbf{M}}$ 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)

$$(\tilde{\gamma}-\gamma^*)^\top\left(\mathbf{I}-\tilde{\mathbf{M}}\right)\gamma^* \le \rho\|\gamma^*\|_2\|\tilde{\gamma}-\gamma^*\|_2$$

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

$$\begin{aligned}
&(\tilde{\gamma}-\gamma^*)^\top\tilde{\mathbf{M}}(\tilde{\gamma}-\gamma^*) \\
&= \|\tilde{\gamma}-\gamma^*\|_2^2 - (\tilde{\gamma}-\gamma^*)^\top\left(\mathbf{I}-\tilde{\mathbf{M}}\right)(\tilde{\gamma}-\gamma^*) \\
&\ge \|\tilde{\gamma}-\gamma^*\|_2^2 - \rho\|\tilde{\gamma}-\gamma^*\|_2^2 \\
&= (1-\rho)\|\tilde{\gamma}-\gamma^*\|_2^2.
\end{aligned}$$

Combining these two expressions and inequality (12) yields

$$(1-\rho)\|\tilde{\gamma}-\gamma^*\|_2^2 \leq \rho\|\gamma^*\|_2\|\tilde{\gamma}-\gamma^*\|_2$$
$$(1-\rho)\|\tilde{\gamma}-\gamma^*\|_2 \leq \rho\|\gamma^*\|_2. \quad (13)$$

From the definition of γ^* and $\tilde{\gamma}$ above and $\boldsymbol{\beta}^*$ and $\tilde{\boldsymbol{\beta}}$, respectively we have

$$\boldsymbol{\beta}^* = -\frac{1}{n\lambda}\mathbf{V}\gamma^*, \qquad \tilde{\boldsymbol{\beta}} = -\frac{1}{n\lambda}\mathbf{V}\tilde{\gamma}$$

so $\frac{1}{n\lambda}\|\gamma^*\|_2 = \|\boldsymbol{\beta}^*\|_2$ and $\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2 = \frac{1}{n\lambda}\|\tilde{\gamma}-\gamma^*\|_2$ due to the orthonormality of $\mathbf{V}$. Plugging this into (13) and using the fact that $\|\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}\|_2 \geq \|\boldsymbol{\beta}_k^* - \widehat{\boldsymbol{\beta}}_k\|_2$ we obtain the stated result. □

B Proof of Row Summing Lemma

Proof of Lemma 2 . Let $\mathbf{V}_k$ contain the first τ rows of $\mathbf{V}$ and let $\mathbf{V}_{(-k)}$ be the matrix containing the remaining rows. Decompose the matrix products as follows

$$\mathbf{V}^\top\mathbf{V} = \mathbf{V}_k^\top\mathbf{V}_k + \mathbf{V}_{(-k)}^\top\mathbf{V}_{(-k)}$$

and

$$\mathbf{V}^\top\Theta_S\Theta_S^\top\mathbf{V} = \mathbf{V}_k^\top\mathbf{V}_k + \tilde{\mathbf{V}}_k^\top\tilde{\mathbf{V}}_k$$

with $\tilde{\mathbf{V}}_k^\top = \mathbf{V}_{(-k)}^\top\mathbf{\Pi}$. Then

$$\begin{aligned}&\|\mathbf{V}^\top\Theta_S\Theta_S^\top\mathbf{V} - \mathbf{V}^\top\mathbf{V}\|_2 \\ &= \|\mathbf{V}_k^\top\mathbf{V}_k + \tilde{\mathbf{V}}_k^\top\tilde{\mathbf{V}}_k - \mathbf{V}_k^\top\mathbf{V}_k - \mathbf{V}_{(-k)}^\top\mathbf{V}_{(-k)}\|_2 \\ &= \|\mathbf{V}_{(-k)}^\top\mathbf{\Pi}\mathbf{\Pi}^\top\mathbf{V}_{(-k)} - \mathbf{V}_{(-k)}^\top\mathbf{V}_{(-k)}\|_2.\end{aligned}$$

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

Lemma 5 (Summed row sampling)**.** *Let $\mathbf{W}$ be an $n\times p$ matrix with orthonormal columns. Let $\mathbf{W}_1,\ldots,\mathbf{W}_K$ be a balanced, random partitioning of the rows of $\mathbf{W}$ where each matrix $\mathbf{W}_k$ has exactly $\tau = n/K$ rows. Define the quantity $M := n\cdot\max_{j=1,\ldots n}\|e_j^\top\mathbf{W}\|_2^2$. For a positive parameter α, select the subsample size*

$$l\cdot K \geq \alpha M\log(p).$$

Let $\mathbf{S}_{T_k}\in\mathbb{R}^{l\times\tau}$ denote the operation of uniformly at random sampling a subset, T_k of the rows of $\mathbf{W}_k$ by sampling l coordinates from $\{1,2,\ldots\tau\}$ without replacement. Now denote $\mathbf{SW}$ as the sum of the subsampled rows

$$\mathbf{SW} = \sum_{k=1}^K \left(\mathbf{S}_{T_k}\mathbf{W}_k\right).$$

Then

$$\sqrt{\frac{(1-\delta)l\cdot K}{n}} \leq \sigma_p(\mathbf{SW})$$

and

$$\sigma_1(\mathbf{SW}) \leq \sqrt{\frac{(1+\eta)l\cdot K}{n}}$$

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 $\mathbf{w}_j^\top$ as the j^{th} row of $\mathbf{W}$ and $M := n\cdot\max_j\|\mathbf{w}_j\|_2^2$. Suppose $K=2$ and consider the matrix

$$\begin{aligned}\mathbf{G}_2 :&= (\mathbf{S}_1\mathbf{W}_1+\mathbf{S}_2\mathbf{W}_2)^\top(\mathbf{S}_1\mathbf{W}_1+\mathbf{S}_2\mathbf{W}_2)\\ &= (\mathbf{S}_1\mathbf{W}_1)^\top(\mathbf{S}_1\mathbf{W}_1) + (\mathbf{S}_2\mathbf{W}_2)^\top(\mathbf{S}_2\mathbf{W}_2)\\ &\quad + (\mathbf{S}_1\mathbf{W}_1)^\top(\mathbf{S}_2\mathbf{W}_2) + (\mathbf{S}_2\mathbf{W}_2)^\top(\mathbf{S}_1\mathbf{W}_1).\end{aligned}$$

In general, we can express $\mathbf{G} := (\mathbf{SW})^\top(\mathbf{SW})$ as

$$\mathbf{G} := \sum_{k=1}^K\sum_{j\in T_k}\left(\mathbf{w}_j\mathbf{w}_j^\top + \sum_{k'\neq k}\sum_{j'\in T_k'}\mathbf{w}_j\mathbf{w}_{j'}^\top\right).$$

By the orthonormality of $\mathbf{W}$, the cross terms cancel as $\mathbf{w}_j\mathbf{w}_{j'}^\top = \mathbf{0}$, yielding

$$\mathbf{G} := (\mathbf{SW})^\top(\mathbf{SW}) = \sum_{k=1}^K\sum_{j\in T_k}\mathbf{w}_j\mathbf{w}_j^\top.$$

We can consider $\mathbf{G}$ as a sum of $l\cdot K$ random matrices

$$\mathbf{X}_1^{(1)},\ldots,\mathbf{X}_1^{(K)},\ldots,\mathbf{X}_l^{(1)},\ldots,\mathbf{X}_l^{(K)}$$

sampled uniformly at random without replacement from the family $\mathcal{X} := \left\{\mathbf{w}_i\mathbf{w}_i^\top : i=1,\ldots,\tau\cdot K\right\}$.

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

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 $\mathbf{W}$ are orthonormal we obtain

$$\mathbb{E}\left[\mathbf{X}_1^{(k)}\right] = \mathbb{E}_P\mathbb{E}_S\mathbf{X}_1^{(k)} = \frac{1}{K}\frac{1}{\tau}\sum_{i=1}^{K\tau}\mathbf{w}_i\mathbf{w}_i^\top = \frac{1}{n}\mathbf{W}^\top\mathbf{W} = \frac{1}{n}\mathbf{I}$$

Recall that we take l samples in K blocks so we can define

$$\mu_{\min} = \frac{l\cdot K}{n} \qquad \text{and} \qquad \mu_{\max} = \frac{l\cdot K}{n}.$$

Plugging these values into Lemma 6, the lower and upper Chernoff bounds respectively yield

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

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

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

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*

$$\max_{\mathbf{A}\in\mathcal{X}} \lambda_{\max}(\mathbf{A}) \leq B$$

Sample $\{\mathbf{A}_1,\ldots,\mathbf{A}_l\}$ uniformly at random from $\mathcal{X}$ without replacement. Compute

$$\mu_{\min} = l\cdot\lambda_{\min}(\mathbb{E}\mathbf{X}_1) \qquad \textit{and} \qquad \mu_{\max} = l\cdot\lambda_{\max}(\mathbb{E}\mathbf{X}_1)$$

Then

$$\mathbb{P}\left\{\lambda_{\min}\left(\sum_i \mathbf{A}_i\right) \leq (1-\delta)\mu_{\min}\right\}$$
$$\leq p\cdot\left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu_{\min}/B} \quad \textit{for } \delta\in[0,1), \textit{ and}$$

$$\mathbb{P}\left\{\lambda_{\max}\left(\sum_i \mathbf{A}_i\right) \geq (1+\delta)\mu_{\max}\right\}$$
$$\leq p\cdot\left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu_{\max}/B} \quad \textit{for } \delta\geq 0.$$

Algorithm	K	TEST MSE	TRAIN MSE
DUAL-LOCO 0.5	12	0.0343 (3.75e-03)	0.0344 (2.59e-03)
DUAL-LOCO 0.5	24	0.0368 (4.22e-03)	0.0344 (3.05e-03)
DUAL-LOCO 0.5	48	0.0328 (3.97e-03)	0.0332 (2.91e-03)
DUAL-LOCO 0.5	96	0.0326 (3.13e-03)	0.0340 (2.67e-03)
DUAL-LOCO 0.5	192	0.0345 (3.82e-03)	0.0345 (2.69e-03)
DUAL-LOCO 1	12	0.0310 (2.89e-03)	0.0295 (2.28e-03)
DUAL-LOCO 1	24	0.0303 (2.87e-03)	0.0307 (1.44e-03)
DUAL-LOCO 1	48	0.0328 (1.92e-03)	0.0329 (1.55e-03)
DUAL-LOCO 1	96	0.0299 (1.07e-03)	0.0299 (7.77e-04)
DUAL-LOCO 2	12	0.0291 (2.16e-03)	0.0280 (6.80e-04)
DUAL-LOCO 2	24	0.0306 (2.38e-03)	0.0279 (1.24e-03)
DUAL-LOCO 2	48	0.0285 (6.11e-04)	0.0293 (4.77e-04)
CoCoA$^+$	12	0.0282 (4.25e-18)	0.0246 (2.45e-18)
CoCoA$^+$	24	0.0278 (3.47e-18)	0.0212 (3.00e-18)
CoCoA$^+$	48	0.0246 (6.01e-18)	0.0011 (1.53e-19)
CoCoA$^+$	96	0.0254 (5.49e-18)	0.0137 (1.50e-18)
CoCoA$^+$	192	0.0268 (1.23e-17)	0.0158 (6.21e-18)

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

C Supplementary Material for Section 5

Algorithm 2 DUAL-LOCO – cross validation

Input: Data: $\mathbf{X}$, Y, no. workers: K, no. folds: v
Parameters: τ_{subs}, $\lambda_1,\ldots\lambda_l$

Partition $\{p\}$ into K subsets of equal size τ and distribute feature vectors in $\mathbf{X}$ accordingly over K workers.
Partition $\{n\}$ into v folds of equal size.
for each fold f **do**
  Communicate indices of training and test points.
  for each worker $k\in\{1,\ldots K\}$ **in parallel do**
    Compute and send $\mathbf{X}_{k,f}^{train}\mathbf{\Pi}_{k,f}$.
    Receive random features and construct $\bar{\mathbf{X}}_{k,f}^{train}$.
    for each $\lambda_j\in\{\lambda_1,\ldots\lambda_l\}$ **do**
      $\tilde{\boldsymbol{\alpha}}_{k,f,\lambda_j} \leftarrow$ `LocalDualSolver`$(\bar{\mathbf{X}}_{k,f}^{train}, Y_f^{train}, \lambda_j)$
      $\widehat{\boldsymbol{\beta}}_{k,f,\lambda_j} = -\frac{1}{n\lambda_j}\mathbf{X}_{k,f}^{train\top}\tilde{\boldsymbol{\alpha}}_{k,f,\lambda_j}$
      $\hat{Y}_{k,f,\lambda_j}^{test} = \mathbf{X}_{k,f}^{test}\widehat{\boldsymbol{\beta}}_{k,f,\lambda_j}$
      Send $\hat{Y}_{k,f,\lambda_j}^{test}$ to driver.
    end for
  end for
  for each $\lambda_j\in\{\lambda_1,\ldots\lambda_l\}$ **do**
    Compute $\hat{Y}_{f,\lambda_j}^{test} = \sum_{k=1}^{K}\hat{Y}_{k,f,\lambda_j}^{test}$.
    Compute $\text{MSE}_{f,\lambda_j}^{test}$ with $\hat{Y}_{f,\lambda_j}^{test}$ and Y_f^{test}.
  end for
end for
for each $\lambda_j\in\{\lambda_1,\ldots\lambda_l\}$ **do**
  Compute $\text{MSE}_{\lambda_j} = \frac{1}{v}\sum_{f=1}^{v}\text{MSE}_{f,\lambda_j}$.
end for

Output: Parameter λ_j attaining smallest MSE_{λ_j}