Supplementary Material for One-shot distributed ridge regression in high dimensions

Edgar Dobriban [*1] **Yue Sheng** [*1 2]

A. Proof of the finite sample results (Theorem 1)

We can calculate the MSE of the weighted sum as

$$M(w) = \mathbb{E}\left\|\sum w_i\hat{\beta}_i - \beta\right\|^2 = \mathbb{E}(\sum w_i\hat{\beta}_i - \beta)^\top(\sum w_j\hat{\beta}_j - \beta)$$
$$= \sum_{ij} w_i w_j \cdot \mathbb{E}\hat{\beta}_i^\top\hat{\beta}_j - 2\sum_i w_i\mathbb{E}\hat{\beta}_i^\top\beta + \|\beta\|^2.$$

Let $\widehat{B}$ be the $p \times k$ matrix defined as $\widehat{B} = [\widehat{\beta}_1, \ldots, \widehat{\beta}_k]$. Then we can write the above MSE as

$$M(w) = w^\top\mathbb{E}\widehat{B}^\top\widehat{B}w - 2\mathbb{E}\beta^\top\widehat{B}w + \|\beta\|^2. \tag{1}$$

Let also

$$B = \mathbb{E}\widehat{B} = [\mathbb{E}\widehat{\beta}_1, \ldots, \mathbb{E}\widehat{\beta}_k]. \tag{2}$$

Since the local estimators are independent, we can write

$$M(w) = w^\top(B^\top B + R)w - 2\beta^\top Bw + \|\beta\|^2, \tag{3}$$

where R is a diagonal matrix with entries

$$R_i = \mathbb{E}\|\widehat{\beta}_i\|^2 - \|\mathbb{E}\widehat{\beta}_i\|^2 = \mathbb{E}\|\widehat{\beta}_i - \mathbb{E}\widehat{\beta}_i\|^2. \tag{4}$$

The objective function $M(w)$ can be viewed as corresponding to a k-parameter linear regression problem, with unknown parameters w_i, design matrix B and outcome vector β. Specifically, we regress β on $\mathbb{E}\widehat{B} = \mathbb{E}[\widehat{\beta}_1, \ldots, \widehat{\beta}_k]$. Therefore, the optimal weights are

$$w^* = (B^\top B + R)^{-1}B^\top\beta, \tag{5}$$

and the optimal risk equals

$$M^* = M(w^*) = \beta^\top\left[I_p - B(B^\top B + R)^{-1}B^\top\right]\beta. \tag{6}$$

Now, to find $B = \mathbb{E}\widehat{B}$, we need $\mathbb{E}\hat{\beta}_i$. The expectation of the ridge regression estimator for the full dataset is

$$\mathbb{E}\widehat{\beta}(\lambda) = \mathbb{E}(X^\top X + n\lambda I_p)^{-1}X^\top Y = (X^\top X + n\lambda I_p)^{-1}X^\top X\beta \tag{7}$$

Letting $\widehat{\Sigma} = n^{-1}X^\top X$, this equals $\mathbb{E}\widehat{\beta}(\lambda) = (\widehat{\Sigma} + \lambda I_p)^{-1}\widehat{\Sigma}\beta$. Similarly,

$$\mathbb{E}\widehat{\beta}_i(\lambda_i) = (X_i^\top X_i + n_i\lambda_i I_p)^{-1}X_i^\top X_i\beta. \tag{8}$$

[*]Equal contribution [1]Wharton Statistics Department, University of Pennsylvania, Philadelphia, PA, USA [2]Graduate Group in Applied Mathematics and Computational Science, University of Pennsylvania, Philadelphia, PA, USA. Correspondence to: Edgar Dobriban <dobriban@wharton.upenn.edu>, Yue Sheng <yuesheng@sas.upenn.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

Let $Q_i = Q_i(\lambda_i) = (X_i^\top X_i + n_i\lambda_i I_p)^{-1} X_i^\top X_i$ be those matrices and let $\widehat{\Sigma}_i = n^{-1} X_i^\top X_i$. Then the above equals $Q_i = (\widehat{\Sigma}_i + \lambda_i I_p)^{-1}\widehat{\Sigma}_i$, and

$$B = [Q_1\beta; \ldots; Q_k\beta]. \tag{9}$$

Therefore, $B^\top B$ has entries $\beta^\top Q_i Q_j \beta$, while $B^\top \beta$ has entries $\beta^\top Q_i \beta$. Moreover,

$$R_i = \mathbb{E}\|\widehat{\beta}_i - \mathbb{E}\widehat{\beta}_i\|^2 = \mathbb{E}\|(X_i^\top X_i + n_i\lambda_i I_p)^{-1} X_i^\top \varepsilon_i\|^2 = \sigma^2 \operatorname{tr}[(X_i^\top X_i + n_i\lambda_i I_p)^{-2} X_i^\top X_i] \tag{10}$$

We can also write this as $R_i = n_i^{-1}\sigma^2 \operatorname{tr}[(\widehat{\Sigma}_i + \lambda_i I_p)^{-2}\widehat{\Sigma}_i]$. To conclude the optimal risk, we have

$$M^*(k) = \|\beta\|^2 - v^\top (A + R)^{-1} v \tag{11}$$

where

$$\begin{aligned} v &= B^\top \beta = vec[\beta^\top Q_i \beta] \\ A &= mat[\beta^\top Q_i Q_j \beta] \\ R &= diag\left[n_i^{-1}\sigma^2 \operatorname{tr}[(\widehat{\Sigma}_i + \lambda_i I_p)^{-2}\widehat{\Sigma}_i]\right] \\ Q_i &= (\widehat{\Sigma}_i + \lambda_i I_p)^{-1}\widehat{\Sigma}_i \end{aligned}$$

Here we used the vectorization and to-matrix operators vec, mat. For the global MSE, we only need to consider the special case where $k = 1$, which gives us

$$\mathbb{E}||\widehat{\beta} - \beta||^2 = M^*(1) = \|\beta\|^2 - \frac{(\beta^\top Q\beta)^2}{\beta^\top Q^2\beta + \sigma^2 \operatorname{tr}[(X^\top X + n\lambda I_p)^{-2} X^\top X]},$$

where $Q = (\widehat{\Sigma} + \lambda I_p)^{-1}\widehat{\Sigma}$. This finishes the argument.

B. Proof of the asymptotical results for general covariance structure (Theorem 2)

In order to prove Theorem 2, we need to introduce the so-called *deterministic equivalents*.

B.1. Deterministic equivalents

The results about random matrix theory stated above in the paper can be expressed in a different, and perhaps slightly more modern language, using *deterministic equivalents* (Serdobolskii, 2007; Hachem et al., 2007; Couillet et al., 2011; Dobriban & Sheng, 2018). For instance, the Marchenko-Pastur law is a consequence of the following result. For any z where it is well-defined, consider the resolvent $(\widehat{\Sigma} - zI_p)^{-1}$. This random matrix is *equivalent* to a deterministic matrix $(x_p\Sigma - zI_p)^{-1}$ for a certain scalar $x_p = x(\Sigma, n, p, z)$, and we write

$$(\widehat{\Sigma} - zI_p)^{-1} \asymp (x_p\Sigma - zI_p)^{-1}.$$

Here two sequences of $n \times n$ matrices A_n, B_n (not necessarily symmetric) of growing dimensions are *equivalent*, and we write

$$A_n \asymp B_n$$

if

$$\lim_{n\to\infty} \operatorname{tr}[C_n(A_n - B_n)] = 0$$

almost surely, for any sequence C_n of $n \times n$ deterministic matrices (not necessarily symmetric) with bounded trace norm, i.e., such that $\limsup \|C_n\|_{tr} < \infty$ (Dobriban & Sheng, 2018). Informally, any linear combination of the entries of A_n can be approximated by the entries of B_n. This also can be viewed as a kind of *weak convergence* in the matrix space equipped with an inner product (trace). From this, it also follows that the traces of the two matrices are equivalent, from which we can recover the MP law.

(Dobriban & Sheng, 2018) collected some useful properties of the calculus of deterministic equivalents. In this work, we use those properties extensively.

B.2. Proof of Theorem 2

The first step is to use the well-known concentration of quadratic forms to reduce to trace functionals (See e.g. Lemma C.3 of (Dobriban & Wager, 2018) which is based on Lemma B.26 of (Bai & Silverstein, 2009)). Since β is independent of the data X with mean zero and finite variance, under the moment assumptions imposed in the theorem, we have

$$\begin{aligned}\beta^\top Q_i \beta - \sigma^2\alpha^2/p \cdot \operatorname{tr} Q_i \to_{a.s.} 0,\\ \beta^\top Q_i Q_j \beta - \sigma^2\alpha^2/p \cdot \operatorname{tr} Q_i Q_j \to_{a.s.} 0,\\ \beta^\top Q_i^2 \beta - \sigma^2\alpha^2/p \cdot \operatorname{tr} Q_i^2 \to_{a.s.} 0.\end{aligned}$$

Let us compute the limits of v, A and R respectively.

- Limit of v: First of all, we have already known that

$$\beta^\top Q_i \beta - \sigma^2\alpha^2/p \cdot \operatorname{tr} Q_i \to_{a.s.} 0, \tag{12}$$

so it is sufficient to consider the limit of $\operatorname{tr} Q_i/p$. Since

$$\operatorname{tr} Q_i/p = 1 - \lambda_i \operatorname{tr}[(\widehat{\Sigma}_i + \lambda_i I_p)^{-1}]/p, \tag{13}$$

assuming that the spectral distribution of $\widehat{\Sigma}_i$ converges almost surely to F_{γ_i}, we thus have

$$\operatorname{tr} Q_i/p \to_{a.s.} 1 - \lambda_i \mathbb{E}_{F_{\gamma_i}}(T+\lambda_i)^{-1} = 1 - \lambda_i m_{F_{\gamma_i}}(-\lambda_i). \tag{14}$$

Above we have introduced the Stieltjes transform $m_{F_{\gamma_i}}$ as a limiting object. So,

$$\beta^\top Q_i \beta \to_{a.s.} \sigma^2\alpha^2[1 - \lambda_i m_{F_{\gamma_i}}(-\lambda_i)]. \tag{15}$$

For the form in terms of the population spectral distribution H, if $p/n \to \gamma$ and the spectral distribution of Σ converges to H, we have by the general Marchenko-Pastur (MP) theorem of Rubio and Mestre (Rubio & Mestre, 2011), that

$$(\widehat{\Sigma} + \lambda I)^{-1} \asymp (x_p \Sigma + \lambda I)^{-1}, \tag{16}$$

where x_p is the unique positive solution of the fixed point equation

$$1 - x_p = \frac{x_p}{n} \operatorname{tr}\left[\Sigma(x_p\Sigma + \lambda I)^{-1}\right]. \tag{17}$$

When $n, p \to \infty$, $x_p \to x$ and x satisfies the equation

$$1 - x = \gamma\left[1 - \lambda \int_0^\infty \frac{dH(t)}{xt + \lambda}\right]. \tag{18}$$

We remark that the assumptions made in the theorem suffice for using the Rubio-Mestre result. Moreover, we only use a special case of their result, similar to (Dobriban & Sheng, 2018). Hence from the calculus of deterministic equivalents (Dobriban & Sheng, 2018), we can take the traces of the matrices in question to obtain

$$\operatorname{tr} Q_i/p = 1 - \lambda_i \operatorname{tr}[(\widehat{\Sigma}_i + \lambda_i I_p)^{-1}]/p \asymp 1 - \lambda_i \operatorname{tr}[(x_i\Sigma + \lambda_i I)^{-1}]/p \to_{a.s.} \mathbb{E}_H \frac{x_i T}{x_i T + \lambda_i}, \tag{19}$$

where $x_i = x(H, \gamma_i, -\lambda_i)$ is the unique solution of

$$1 - x_i = \gamma_i\left[1 - \lambda_i \int_0^\infty \frac{dH(t)}{x_i t + \lambda_i}\right]. \tag{20}$$

- Limit of A: Let us consider the cases $i \neq j$ and $i = j$ separately.

– $i \neq j$: We begin by

$$\beta^\top Q_i Q_j \beta - \sigma^2 \alpha^2 / p \cdot \operatorname{tr} Q_i Q_j \to_{a.s.} 0. \tag{21}$$

Based on the above expression for Q_i, we have

$$Q_i Q_j = I_p - \lambda_i (\widehat{\Sigma}_i + \lambda_i I_p)^{-1} - \lambda_j (\widehat{\Sigma}_j + \lambda_j I_p)^{-1} + \lambda_i \lambda_j (\widehat{\Sigma}_i + \lambda_i I_p)^{-1} (\widehat{\Sigma}_j + \lambda_j I_p)^{-1}. \tag{22}$$

So the key will be to find the limit of

$$E_{ij} = p^{-1} \operatorname{tr}\{(\widehat{\Sigma}_i + \lambda_i I_p)^{-1} (\widehat{\Sigma}_j + \lambda_j I_p)^{-1}\}. \tag{23}$$

From the general MP theorem, since $p/n_i \to \gamma_i$, we have for all i,

$$(\widehat{\Sigma}_i + \lambda_i I_p)^{-1} \asymp (x_{ip} \Sigma + \lambda_i I_p)^{-1}. \tag{24}$$

Here x_{ip} is the unique positive solution of the fixed point equation

$$1 - x_{ip} = \frac{x_{ip}}{n_i} \operatorname{tr}\left[\Sigma (x_{ip} \Sigma + \lambda_i I)^{-1}\right], \tag{25}$$

and $x_{ip} \to x_i$ as $n_i, p \to \infty$. By the product rule of the calculus of deterministic equivalents, we have for $i \neq j$

$$(\widehat{\Sigma}_i + \lambda_i I_p)^{-1} (\widehat{\Sigma}_j + \lambda_j I_p)^{-1} \asymp (x_{ip} \Sigma + \lambda_i I_p)^{-1} (x_{jp} \Sigma + \lambda_j I_p)^{-1}. \tag{26}$$

Hence by the trace rule of deterministic equivalents,

$$E_{ij} \asymp p^{-1} \operatorname{tr}[(x_{ip} \Sigma + \lambda_i I_p)^{-1} (x_{jp} \Sigma + \lambda_j I_p)^{-1}]. \tag{27}$$

Moreover, since the spectral distribution of Σ converges to H, we find for $i \neq j$

$$E_{ij} \to \mathbb{E}_H \frac{1}{(x_i T + \lambda_i)(x_j T + \lambda_j)}. \tag{28}$$

Putting it together,

$$Q_i Q_j \asymp I_p - \lambda_i (x_{ip} \Sigma + \lambda_i I_p)^{-1} - \lambda_j (x_{jp} \Sigma + \lambda_j I_p)^{-1} + \lambda_i \lambda_j (x_{ip} \Sigma + \lambda_i I_p)^{-1} (x_{jp} \Sigma + \lambda_j I_p)^{-1}. \tag{29}$$

So, again by the trace rule of deterministic equivalents, we have

$$\begin{aligned} p^{-1} \operatorname{tr}\{Q_i Q_j\} &\to_{a.s.} 1 - \mathbb{E}_H \frac{\lambda_i}{x_i T + \lambda_i} - \mathbb{E}_H \frac{\lambda_j}{x_j T + \lambda_j} + \mathbb{E}_H \frac{\lambda_i \lambda_j}{(x_i T + \lambda_i)(x_j T + \lambda_j)} \\ &= x_i x_j \mathbb{E}_H \frac{T^2}{(x_i T + \lambda_i)(x_j T + \lambda_j)}. \end{aligned}$$

Therefore, for $i \neq j$

$$A_{ij} \to \sigma^2 \alpha^2 \left[x_i x_j \mathbb{E}_H \frac{T^2}{(x_i T + \lambda_i)(x_j T + \lambda_j)} \right]. \tag{30}$$

– $i = j$: In this case,

$$\beta^\top Q_i^2 \beta - \sigma^2 \alpha^2 / p \cdot \operatorname{tr} Q_i^2 \to 0, \tag{31}$$

where $Q_i^2 = I_p - 2\lambda_i (\widehat{\Sigma}_i + \lambda_i I_p)^{-1} + \lambda_i^2 (\widehat{\Sigma}_i + \lambda_i I_p)^{-2}$. We can easily find the limit of $\operatorname{tr} Q_i^2 / p$ in terms of empirical quantities, based on our knowledge of the convergence of Stieltjes transforms and its derivatives:

$$\operatorname{tr} Q_i^2 / p \to 1 - 2\lambda_i m_{F_{\gamma_i}}(-\lambda_i) + \lambda_i^2 m'_{F_{\gamma_i}}(-\lambda_i). \tag{32}$$

Therefore, for $i = j$

$$A_{ii} \to \sigma^2 \alpha^2 [1 - 2\lambda_i m_{F_{\gamma_i}}(-\lambda_i) + \lambda_i^2 m'_{F_{\gamma_i}}(-\lambda_i)]. \tag{33}$$

- Limit of R: Recall that $R_i = n_i^{-1} \sigma^2 \operatorname{tr}[(\widehat{\Sigma}_i + \lambda_i I_p)^{-2} \widehat{\Sigma}_i]$. We note $p^{-1} \operatorname{tr}(\widehat{\Sigma} + \lambda I)^{-2} \to m'_{F_\gamma}(-\lambda)$ and $\widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-2} = (\widehat{\Sigma} + \lambda I)^{-1} - \lambda (\widehat{\Sigma} + \lambda I)^{-2}$, so

$$\frac{\operatorname{tr}[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-2}]}{n} \to \gamma [m_{F_\gamma}(-\lambda) - \lambda m'_{F_\gamma}(-\lambda)]. \tag{34}$$

Hence

$$R_{ii} \to \sigma^2 \left[\gamma_i [m_{F_{\gamma_i}}(-\lambda_i) - \lambda m'_{F_{\gamma_i}}(-\lambda_i)] \right]. \tag{35}$$

C. Proof of the asymptotical results for the identity covariance case (Theorem 3)

Here we will provide the proof of Theorem 3 and some explanation of the decoupling phenomenon via free probability theory.

C.1. Proof of Theorem 3 and some explanation

The proof for v and R is clear by Theorem 2. For the limit of A, the diagonal case is also direct. When $i \neq j$, recall that

$$E_{ij} = p^{-1} \operatorname{tr}\{(\widehat{\Sigma}_i + \lambda_i I_p)^{-1}(\widehat{\Sigma}_j + \lambda_j I_p)^{-1}\} \to \mathbb{E}_H \frac{1}{(x_i T + \lambda_i)(x_j T + \lambda_j)}. \tag{36}$$

For $H = \delta_1$, the expectation decouples, we find

$$E_{ij} \to \frac{1}{x_i + \lambda_i} \cdot \frac{1}{x_i + \lambda_j} = m_{\gamma_i}(-\lambda_i) m_{\gamma_j}(-\lambda_j). \tag{37}$$

Therefore,

$$A_{ij} \to \sigma^2 \alpha^2 [1 - \lambda_i m_{\gamma_i}(-\lambda_i)] \cdot [1 - \lambda_j m_{\gamma_j}(-\lambda_j)]. \tag{38}$$

Now let us put everything together. Recall that the optimal risk has the form $\mathrm{MSE}^*_{dist} = \|\beta\|^2 - v^\top (A + R)^{-1} v$. Based on the above discussion, we have

$$\sigma^2 \alpha^2 (A + R) \to \sigma^2 \alpha^2 (\mathcal{A} + \mathcal{R}) = V V^\top + D, \tag{39}$$

where D is a diagonal matrix with i-th diagonal entry $\sigma^2 \alpha^2 (\mathcal{R}_{ii} + \mathcal{A}_{ii}) - V_i^2$. Then, by using the Sherman$-$Morrison formula, we have

$$V^\top (V V^\top + D)^{-1} V = \frac{V^\top D^{-1} V}{1 + V^\top D^{-1} V}. \tag{40}$$

So the limiting distributed risk is

$$\mathcal{M}_k = \sigma^2 \alpha^2 - \sigma^2 \alpha^2 \frac{V^\top D^{-1} V}{1 + V^\top D^{-1} V} = \frac{\sigma^2 \alpha^2}{1 + V^\top D^{-1} V} = \frac{\sigma^2 \alpha^2}{1 + \sum_{i=1}^k \frac{V_i^2}{D_i}}, \tag{41}$$

which finishes the proof.

To explain in more detail the decoupling phenomenon, let us study how the local risks are related to the distributed risks. Define $V = V(\gamma, \lambda)$ to be the limiting scalar $V \in \mathbb{R}$ defind above, in the special case $k = 1$. Explicitly, this is the limit of the quantity $\beta^\top Q \beta$, where $Q = (\widehat{\Sigma} + \lambda I_p)^{-1} \widehat{\Sigma}$, as given in Theorem 1 applied for $k = 1$. Let D be the scalar expression $D(\gamma, \lambda) = \sigma^2 \alpha^2 (\mathcal{R} + \mathcal{A}) - V$ when $k = 1$. With these notations, the risk $\mathcal{M}_1$ of ridge regression when computed on the entire dataset equals

$$\mathcal{M}_1(\gamma, \lambda) = \frac{\sigma^2 \alpha^2}{1 + \frac{V(\gamma, \lambda)}{D(\gamma, \lambda)}}. \tag{42}$$

Moreover, the risk of optimally weighted one-shot distributed ridge over k subsets, with arbitrary regularization parameters λ_i, equals

$$\mathcal{M}_k(\gamma_1, \ldots, \gamma_k, \lambda_1, \ldots, \lambda_k) = \frac{\sigma^2 \alpha^2}{1 + \sum_{i=1}^k \frac{V_i^2(\gamma_i, \lambda_i)}{D_i(\gamma_i, \lambda_i)}}. \tag{43}$$

Then one can check that we have the following equations connecting the risk computed on the entire dataset and the distributed risk:

$$\frac{\sigma^2 \alpha^2}{\mathcal{M}_k(\gamma_1, \ldots, \gamma_k, \lambda_1, \ldots, \lambda_k)} - 1 = \sum_{i=1}^k \frac{\sigma^2 \alpha^2}{\mathcal{M}_1(\gamma_i, \lambda_i)} - k,$$

$$\mathcal{M}_k(\gamma_1, \ldots, \gamma_k, \lambda_1, \ldots, \lambda_k) = \frac{1}{\sum_{i=1}^k \frac{1}{\mathcal{M}_1(\gamma_i, \lambda_i)} + \frac{1-k}{\sigma^2 \alpha^2}}.$$

These equations are precisely what we mean by *decoupling*. The distributed risk can be written as a function of the type $1/(\sum_i 1/x_i + b)$ of the distributed risks. Therefore, there are no "interactions" between the different risk functions.

Next, we discuss in more depth why the limiting risk decouples. Mathematically, the key reason is that when $\Sigma = I$, the limit of A_{ij} for $i \neq j$ decouples into a product of two terms. Therefore, the distributed risk function involves a quadratic form with zero *off-diagonal* terms. This is not the case for general population covariance Σ. We provide an explanation via free probability theory.

C.2. Explaining decoupling via free probability theory

In this section, we provide an explanation via free probability theory for why the limiting distributed risk decouples. Specifically, we explain why the limit of the quantities $\beta^\top Q_i \beta \cdot \beta^\top Q_j \beta$ for $i \neq j$ becomes a product of terms depending on i, j.

We will use some basic notions from free probability theory (Voiculescu et al., 1992; Hiai & Petz, 2006; Nica & Speicher, 2006; Anderson et al., 2010; Couillet & Debbah, 2011). Let us define our non-commutative probability space as

$$\left(\mathcal{A} = (L^{\infty-} \otimes M_p(\mathbb{R})), \tau = \frac{1}{p}\operatorname{tr}\right),$$

where $L^{\infty-}$ denotes the collection of random variables with all moments finite and $M_p(\mathbb{R})$ is the space of $p \times p$ real matrices. Recall that, a sequence of random variables $\{a_{1,p}, a_{2,p}, \ldots, a_{k,p}\} \subset \mathcal{A}$ is said to be asymptotically free almost surely if

$$\tau[\prod_{j=1}^{m} P_j(a_{i_j,p} - \tau(P_j(a_{i_j,p})))] \to_{a.s.} 0,$$

for any positive integer m, any polynomials $P_1, \ldots, P_m$ and any indices $i_1, \ldots, i_m \in [k]$ with no two adjacent i_j equal. Suppose A_p, B_p are two sequences of independent random matrices and at least one of them is orthogonally invariant, then it is well-known that $\{A_p, B_p\} \subset \mathcal{A}$ is asymptotically free almost surely.

Now, let us assume that $X^\top X$ is orthogonally invariant, which is the case when $X^\top X$ follows the white Wishart distribution. Then clearly $X_i^\top X_i$ and $X_j^\top X_j$ are asymptotically free almost surely. It follows that Q_i and Q_j are also asymptotically free almost surely. By using the definition of asymptotic freeness, we have for $i \neq j$

$$\tau[(Q_i - \frac{1}{p}\operatorname{tr}(Q_i)I)(Q_j - \frac{1}{p}\operatorname{tr}(Q_j)I)] \to_{a.s.} 0,$$

which is equivalent to

$$\frac{1}{p}\operatorname{tr}(Q_i Q_j) - \frac{1}{p}\operatorname{tr}(Q_i)\frac{1}{p}\operatorname{tr}(Q_j) \to_{a.s.} 0.$$

Hence, under the random-effects assumption for β, the limit of $\beta^\top \beta \cdot \beta^\top Q_i Q_j \beta$ $(i \neq j)$ will decouple and is the same as the limit of $\beta^\top Q_i \beta \cdot \beta^\top Q_j \beta$.

D. Proof of Theorem 4

Recall that

$$\begin{aligned}\frac{V_i^2}{D_i} &= \frac{\sigma^4\alpha^4(1-\lambda_i m_{\gamma_i}(-\lambda_i))^2}{\sigma^4\alpha^4\lambda_i^2[m'_{\gamma_i}(-\lambda_i) - m^2_{\gamma_i}(-\lambda_i)] + \sigma^4\alpha^2\gamma_i[m_{\gamma_i}(-\lambda_i) - \lambda_i m'_{\gamma_i}(-\lambda_i)]} \\ &= \frac{\alpha^2(1-\lambda_i m_{\gamma_i}(-\lambda_i))^2}{\alpha^2\lambda_i^2[m'_{\gamma_i}(-\lambda_i) - m^2_{\gamma_i}(-\lambda_i)] + \gamma_i[m_{\gamma_i}(-\lambda_i) - \lambda_i m'_{\gamma_i}(-\lambda_i)]},\end{aligned}$$

and our goal is to find λ_i that maximizes V_i^2/D_i. Luckily, from (Dobriban & Wager, 2018) it follows that for $k = 1$, i.e. when there is only one machine, the optimal choice of the tuning parameter λ is γ/α^2. This means that the maximizer of V^2/D is $\lambda = \gamma/\alpha^2$. Now, due to the decoupled structure of $\mathcal{M}_k$, the optimal tuning parameters are $\lambda_i = \gamma_i/\alpha^2$. Plugging in the parameters, we have

$$\frac{V_i^2}{D_i} = \frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)} - 1.$$

Then the optimal risk can be simplified to

$$\mathcal{M}_k = \frac{\sigma^2\alpha^2}{1+\sum_{i=1}^k\left[\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)}-1\right]}. \tag{44}$$

When $k=1$, this equals to $\sigma^2\gamma m_\gamma(-\gamma/\alpha^2)$ which matches the known result from (Dobriban & Wager, 2018).

E. Proof of Theorem 5

For the first property, minimizing the ARE is equivalent to maximizing the following quantity

$$\sum_{i=1}^k \frac{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)}{\alpha^2} = \sum_{i=1}^k \frac{\phi(\gamma_i)}{\alpha^2}. \tag{45}$$

It is helpful to introduce $r(t)=\phi(\gamma)$, where $t=1/\gamma$. We can easily compute that

$$r'(t) = \frac{\alpha^2}{2}\left(-1+\frac{t-(1-1/\alpha^2)}{\sqrt{(t-(1-1/\alpha^2))^2+4/\alpha^2}}\right) < 0 \ \ , r''(t) = \frac{2}{[(t-(1-1/\alpha^2))^2+4/\alpha^2]^{3/2}} > 0.$$

Thus, $r(t)$ is a decreasing and convex function. We can show the ARE achieves minimum when the samples are equally distributed by considering the following optimization problem

$$\begin{aligned} \max_{t_i} \quad & \sum_{i=1}^k \frac{r(t_i)}{\alpha^2} \\ \text{subject to} \quad & \sum_{i=1}^k t_i = \frac{1}{\gamma}, \\ & t_i \geq 0, i=1,2,\ldots,k. \end{aligned}$$

We denote the objective by $R(t_1,\ldots,t_k)$, and the corresponding Lagrangian by $R_\xi = R - \xi(\sum_i t_i - 1/\gamma)$. Then it is easy to check that the condition $\frac{\partial R_\xi}{\partial t_i}=0$ reduces to

$$\frac{r'(t_i)}{\alpha^2} - \xi = 0, \ i=1,2,\ldots,k. \tag{46}$$

Since $r'(t)$ is also monotone, the unique solution to the stationary condition is $t_1=t_2=\cdots=t_k=1/(k\gamma)$. If some t_i equals to 0, then it reduces to a problem with $k-1$ machines. So it remains to check the boundary case where only one t_i is non-zero and equals to $1/\gamma$. Obviously, this is the trivial case where the ARE is 1. Therefore, we have shown that the ARE attains its minimum when the samples are equally distributed across k machines.

Next, for fixed α^2 and γ, we can check

$$\frac{\partial\psi}{\partial k} = \frac{\gamma m_\gamma(-\gamma/\alpha^2)}{\alpha^2}\left(\frac{\alpha^2}{2\gamma}\cdot\frac{\left(\gamma/\alpha^2+\gamma\right)^2 k+\gamma/\alpha^2-\gamma}{\sqrt{\left(\gamma/\alpha^2+\gamma\right)^2 k^2+2\left(\gamma/\alpha^2-\gamma\right)k+1}}-\frac{1+\alpha^2}{2}\right)\leq 0. \tag{47}$$

Moreover, the limit of ψ is

$$\begin{aligned} h(\alpha^2,\gamma) = \lim_{k\to\infty}\psi(k,\gamma,\alpha^2) &= \frac{\gamma m_\gamma(-\gamma/\alpha^2)}{\alpha^2}\left(1+\frac{\alpha^2}{\gamma(1+\alpha^2)}\right) \\ &= \frac{-\gamma/\alpha^2+\gamma-1+\sqrt{(-\gamma/\alpha^2+\gamma-1)^2+4\gamma^2/\alpha^2}}{2\gamma}\left(1+\frac{\alpha^2}{\gamma(1+\alpha^2)}\right). \end{aligned}$$

F. Proof of Theorem 6 and an intuitive explanation for why the weights are large

F.1. Proof of Theorem 6

Recall that the optimal weights are $w^* = (A+R)^{-1}v$ and $\sigma^2\alpha^2(A+R) \to VV^\top + D$. Denote the limit of the optimal weights by W, so that we have

$$W = \sigma^2\alpha^2(VV^\top + D)^{-1}V = \frac{\sigma^2\alpha^2 D^{-1}V}{1+V^\top D^{-1}V}.$$

When we choose $\lambda_i = \gamma_i/\alpha^2$ for each i, we can write the limiting optimal weights as

$$W = \mathcal{M}_k \cdot D^{-1}V.$$

So, it follows from the formulas of $\mathcal{M}_k, D$ and V that

$$W_i = \left(\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)}\right) \cdot \left(\frac{1}{1+\sum_{i=1}^k \left[\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)} - 1\right]}\right).$$

For the sum of the coordinates, we have

$$1^\top W = \frac{\sum_{i=1}^k \left(\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)}\right)}{1+\sum_{i=1}^k \left[\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)} - 1\right]} = \frac{\sum_{i=1}^k \left(\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)}\right)}{1-k+\sum_{i=1}^k \left(\frac{\alpha^2}{\gamma_i m_{\gamma_i}(-\gamma_i/\alpha^2)}\right)} \geq 1.$$

In the special case where all γ_i are equal, i.e., $\gamma_i = k\gamma$, we have all W_i equal to

$$W_i = \frac{\frac{\alpha^2}{k\gamma \cdot m_{k\gamma}(-k\gamma/\alpha^2)}}{1-k+\frac{\alpha^2}{\gamma \cdot m_{k\gamma}(-k\gamma/\alpha^2)}} = \frac{1}{k+(1-k)\cdot k\gamma/\alpha^2 \cdot m_{k\gamma}(-k\gamma/\alpha^2)}.$$

In terms of the optimal risk function $\phi(\gamma) = \phi(\gamma,\alpha) = \gamma m_\gamma(-\gamma/\alpha^2)$ defined before, this can also be written as the following optimal weight function

$$\mathcal{W}(k,\gamma,\alpha) = \frac{1}{k-(k-1)\cdot\phi(k\gamma)/\alpha^2}.$$

The monotonicity and the limits of $\mathcal{W}$ can be checked directly.

F.2. Intuitive explanation for the need to use weights summing to greater than unity

Consider a much more simplified problem, where we are estimating a scalar parameter θ. We have an estimator $\widehat{\theta}$, which is generally biased, and we would like to find the scale multiple $c \cdot \widehat{\theta}$ that minimizes the mean squared error. A calculation reveals that

$$M(c) = \mathbb{E}(c\cdot\widehat{\theta} - \theta)^2 = c^2\mathbb{E}(\widehat{\theta}^2) - 2c\cdot\mathbb{E}\widehat{\theta}\cdot\theta + \theta^2$$

Hence the optimal scale factor is $c = \mathbb{E}\widehat{\theta}\cdot\theta/\mathbb{E}(\widehat{\theta}^2)$.

We can achieve a better understanding of this optimal scale if we consider the bias-variance decomposition of $\widehat{\theta}$. Let us define the bias and the variance as

$$B = \mathbb{E}\widehat{\theta} - \theta$$
$$V = \mathbb{E}(\widehat{\theta} - \mathbb{E}\widehat{\theta})^2$$

We then see that the optimal scale factor is

$$c = \frac{B+\theta}{V+(B+\theta)^2}\theta = 1 - \frac{V+B(B+\theta)}{V+(B+\theta)^2}.$$

This quantity is an "inflation factor", i.e., greater than or equal to unity, if $V + B(B+\theta) \leq 0$. This can be written as

$$V + B^2 \leq -B\theta$$

Hence, this condition can only hold if the bias B has opposite sign with θ. This would be the case for a *shrinkage estimator* θ. In that case, the condition could hold if the parameter θ has a large magnitude.

Returning to our main problem, ridge regression is a shrinkage estimator, and averages of ridge regression estimators are still shrinkage estimators. Therefore, it makes sense that their weighted average should be inflated to minimize mean squared error. This provides an intuitive explanation for why the weights sum to greater than one.

G. Out-of-sample prediction

In real applications, out-of-sample prediction is also of interest. We consider a test datapoint (x_t, y_t), generated from the same model $y_t = x_t^\top \beta + \varepsilon_t$, where x_t, ε_t are independent of X, ε. We want to use $x_t^\top \widehat{\beta}$ to predict y_t, and the out-of-sample prediction error is defined as $\mathbb{E}(y_t - x_t^\top \widehat{\beta})^2$.

Under the conditions of Theorem 3, the limiting out-of-sample prediction error of the optimal distributed estimator $\widehat{\beta}_{dist}$ is

$$\mathcal{O}_k = \sigma^2 + \mathcal{M}_k.$$

Thus, the asymptotic out-of-sample relative efficiency, meaning the ratio of prediction errors, is

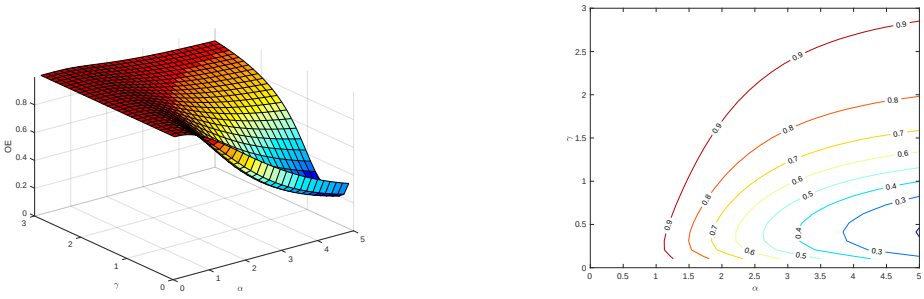
$$OE = \frac{\mathcal{O}_1}{\mathcal{O}_k} = \frac{\mathcal{M}_1 + \sigma^2}{\mathcal{M}_k + \sigma^2},$$

and the efficiency for prediction is higher than for estimation $OE \geq ARE$. Furthermore, when the samples are equally distributed, the relative efficiency has the form

$$\Psi(k, \gamma, \alpha^2) = \frac{1 + \gamma m_\gamma(-\gamma/\alpha^2)}{1 + \frac{\alpha^2 \gamma m_{k\gamma}(-k\gamma/\alpha^2)}{\alpha^2 + (1-k)\gamma m_{k\gamma}(-k\gamma/\alpha^2)}},$$

and the corresponding infinite-worker limit (taking $k \to \infty$) is

$$\mathcal{H}(\alpha^2, \gamma) = \frac{1 + \gamma m_\gamma(-\gamma/\alpha^2)}{1 + \frac{\gamma\alpha^2(1+\alpha^2)}{\alpha^2 + \gamma(1+\alpha^2)}}.$$



(a) k=5 (b) k=10

Figure 1. Limit of OE: (a) surface and (b) contour plots of $\mathcal{H}(\alpha^2, \gamma)$.

The proof is straightforward. We can see Figure 1 for some plots. For the identity covariance case, the efficiency loss of the distributed estimator in terms of the test error is always less than the loss in terms of the estimation error. When the signal-to-noise ratio α^2 is small, the relative efficiency is always very large and close to 1. This observation can be an encouragement to use our distributed methods for out-of-sample prediction.

H. Additional experimental results

H.1. How much does optimal weighting help?

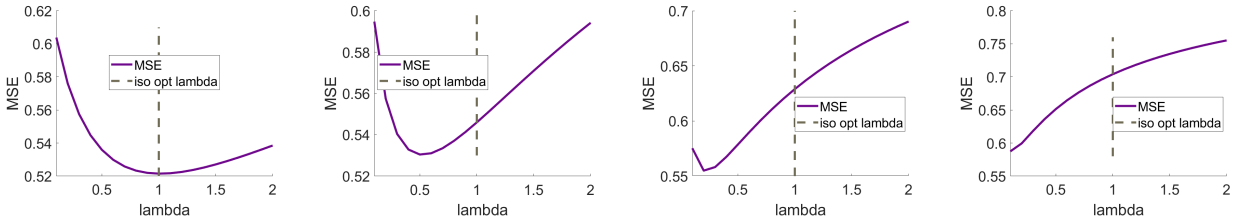
This allows us to compare our proposed weighting method to a "baseline". See Figure 2. We have plotted the risk of distributed ridge regression for both the optimally weighted version and the simple average, as a function of the regularization

(a) k=5 (b) k=10

Figure 2. Distributed risk as a function of the regularization parameter. We plot both the risk with optimal weights (MSE opt) and the risk obtained from sub-optimal averaging (MSE avg). We set $\alpha = 1$, $\gamma = 0.17$ and $k = 5, 10$.

parameter. We observe that *optimal weighting can lead to a 30-40% decrease in the risk*. Therefore, our proposed weighting scheme can sometimes lead to a substantial benefit.

H.2. Experiments on general covariance structures



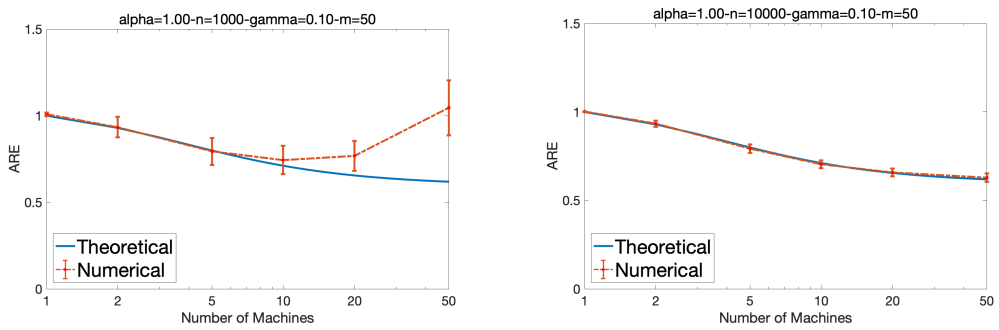
(a) k=1 (b) k=2 (c) k=5 (d) k=10

Figure 3. Distributed risk as a function of the regularization parameter. We plot the risk of the optimally weighted distributed estimator for an AR-1 covariance structure. We set $\alpha = 1$, $\gamma = 0.17$ and $k = 1, 2, 5, 10$.

How can we choose the optimal regularization parameters when the predictors have a general covariance structure Σ? In this case, our theoretical results do not give an explicit expression for the optimal regularization parameters. Here we present simulation results to shed light on this question. Here the regression model is $Y = X\beta + \varepsilon$. We generate the datapoints independently from an autoregressive model of order one (AR-1), i.e., each datapoint x_i is generated as $x_i \sim \mathcal{N}(0, \Sigma)$, where $\Sigma_{ij} = \rho^{|i-j|}$, and ρ is the autocorrelation parameter. β is a p-dimensional random vector with i.i.d. mean 0, variance α^2/p normal entries, and ε also has i.i.d standard normal entries. For each $k = 1, 2, 5, 10$, we split the data equally into k groups and do distributed ridge regression. We choose $\rho = 0.9$. We also choose $n = 3000$, $p = 500$, and report the results of a simulation where we average over $n_{mc} = 20$ independent realizations of β. Figure 3 shows the optimal distributed risk $M^*(k)$ as a function of the local regularization parameter λ. We set all local regularization parameters to equal values, which is reasonable, since the local problems are exchangeable. We also parametrize the regularization parameters as multiples of the optimal parameter for the isotropic case (which equals γ/α^2). We observe that for $k = 1$, the optimal parameter is the same as in the isotropic case. This makes sense, because the optimal regularization parameter for one machine is always the same, regardless of the structure of the design. However for $k > 1$, we generally observe that the regularization parameters are *smaller* than the isotropic ones. This is an insight that has apparently not been available before. In fact, further simulations show that similar results hold for more general covariance structures, i.e. the optimal λ tends to be smaller than the optimal one in the isotropic case. It is a topic of future work to develop an intuitive understanding.

H.3. Finite-sample comparison of relative efficiency for isotropic covariance

Figure 4 shows a comparison of the theoretical formulas for ARE and realized relative efficiency in a regression simulation. Here the regression model is $Y = X\beta + \varepsilon$, where X is $n \times p$ with i.i.d. standard normal entries, β is a p-dimensional random vector with i.i.d. mean 0, variance α^2/p normal entries, and ε also has i.i.d standard normal entries. For each $k = 1, 2, 5, 10, 20, 50$, we split the data equally into k groups and perform ridge regression on each group. For each group,



(a) n=1000 (b) n=10000

Figure 4. Realized relative efficiency in a regression simulation.

we choose the same tuning parameter $\lambda_i = p/(n_i\alpha^2)$. For the global regression on the entire dataset, we choose the tuning parameter $\lambda = p/(n\alpha^2)$ optimally.

We show the results of the expression for the realized relative efficiency $\|\widehat{\beta} - \beta\|^2/\|\widehat{\beta}_{dist} - \beta\|^2$ compared to the theoretical ARE. We generate 100 independent copies of ε, perform regression, recording the realized relative efficiency $||\widehat{\beta} - \beta||^2/||\widehat{\beta}_{dist} - \beta||^2$, as well as its overall Monte Carlo mean. For the first plot, we take $n = 1000$, $p = 100$, and $\alpha = \sigma = 1$.

As we can see in the plot, the theoretical formula is accurate only for a small number of machines. It turns out that this is due to finite-sample effects. In the second plot, we set $n = 10000$, $p = 1000$ and $\alpha = \sigma = 1$ such that the aspect ratio $\gamma = p/n$ is the same as before. In that case the theoretical formula becomes very accurate.

References

Anderson, G. W., Guionnet, A., and Zeitouni, O. *An Introduction to Random Matrices*. Number 118. Cambridge University Press, 2010.

Bai, Z. and Silverstein, J. W. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, 2009.

Couillet, R. and Debbah, M. *Random Matrix Methods for Wireless Communications*. Cambridge University Press, 2011.

Couillet, R., Debbah, M., and Silverstein, J. W. A deterministic equivalent for the analysis of correlated mimo multiple access channels. *IEEE Trans. Inform. Theory*, 57(6):3493–3514, 2011.

Dobriban, E. and Sheng, Y. Distributed linear regression by averaging. *arXiv preprint arxiv:1810.00412*, 2018.

Dobriban, E. and Wager, S. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

Hachem, W., Loubaton, P., and Najim, J. Deterministic equivalents for certain functionals of large random matrices. *The Annals of Applied Probability*, 17(3):875–930, 2007.

Hiai, F. and Petz, D. *The semicircle law, free random variables and entropy*. Number 77. American Mathematical Soc., 2006.

Nica, A. and Speicher, R. *Lectures on the combinatorics of free probability*, volume 13. Cambridge University Press, 2006.

Rubio, F. and Mestre, X. Spectral convergence for a general class of random matrices. *Statistics & Probability Letters*, 81 (5):592–602, 2011.

Serdobolskii, V. I. *Multiparametric Statistics*. Elsevier, 2007.

Voiculescu, D. V., Dykema, K. J., and Nica, A. *Free random variables*. Number 1. American Mathematical Soc., 1992.