Efficient Data-Dependent Random Projection for Least Square Regressions

Jacob Sturges
School of Computer Science
University of Oklahoma
Norman, USA
jacob.e.sturges-1@ou.edu

Luyuan Yang
School of Computer Science
University of Oklahoma
Norman, USA
luyuan.yang@ou.edu

Shayan Shafaei
School of Computer Science
University of Oklahoma
Norman, USA
shayan.shafaei@ou.edu

Chao Lan
School of Computer Science
University of Oklahoma
Norman, USA
clan@ou.edu

***Abstract*—This paper presents a new data-dependent random projection method D^{2}RP for least square regressions, which maps data into the row space of a randomly mapped training data matrix. Our theoretical analysis suggests D^{2}RP may not preserve pairwise data distance as well as its data-independent ancestors, but preserves enough information for reconstructing the training data. Our further analysis shows least square regression in the D^{2}RP projected space has an $O(e^{-k/n})$ empirical excess risk that decays exponentially faster as k increases, partly suggesting its high dimension efficiency. On the practical side, we apply D^{2}RP to speed up least square regression, kernel ridge regression and ensemble regression. Experimental results on real-world data sets show it achieves the best tradeoff between computation efficiency and dimension efficiency compared to multiple baselines methods.**

***Index Terms*—random projection, dimensionality reduction, least square regression**

I. Introduction

Random projection (RP) [1] is an efficient data dimensionality reduction technique widely used to accelerate machine learning tasks including regression [2], [3], [4], [5], [6], [7], [8], classification [9], [10], [11], [12], [13], [14], [15] and clustering [16], [17], [18], [19]. In these applications, data are first mapped into a lower dimensional space using a randomly generated projection matrix before being used to train models.

The success of learning in randomly projected space is often explained by distance preservation i.e., random projection only distorts the pairwise distance between n data points by $1 \pm \varepsilon$ if the projected dimension is $\tilde{O}(\ln n/\varepsilon^2)$ [1], therefore allowing models to preserve performance in the projected space should their performance depends on such distance.

In the literature, most random projection matrices are generated independently to the data to project. While this makes RP fairly computation-efficient, it also makes RP less dimension-efficient, meaning that one often has to map data into a higher dimensional space for preserving the same model performance as other methods like PCA [9]. Such inefficiency undermines the value of RP as a dimensionality reduction technique.

We argue the dimension efficiency of RP can be improved if its projection is not entirely data oblivious. Data-dependent random projection is rarely explored in the literature, and we notice a prior study [20] shows it allows models to perform better in the projected space compared to its data-independent counterparts, shedding light on this promising direction.

Research on data-dependent random projection is non-trivial due to an inherent tradeoff between computation efficiency and dimension efficiency i.e., a projection generated obliviously to data often takes less generation time but is also less dimension-efficient, whereas a projection learned from data can be more dimension-efficient but also takes more generation time. For example, the data-dependent random projection [20] improves dimension-efficiency but endures an $O(k^2)$ generation time for k-dimensional projected space, which is higher than the $O(k)$ time of its data-oblivious ancestors. How to improve the tradeoff remains an open research challenge.

This paper presents a new data-dependent random projection method D^{2}RP for least square regressions, which maps data into the row space of a randomly projected training data matrix with merely $O(k)$ projection generation time. Our theoretical analysis demystifies its effectiveness from two complementary views. We show D^{2}RP may not preserve data distance as well as its data-oblivious ancestors since it admits a larger relative data distance distortion bound, but its projected space contains enough information for reconstructing training data with an $O(\sqrt{n/k})$ error that matches its data-oblivious ancestor's. Our further analysis shows D^{2}RP allows least square regression to achieve an $O(e^{-k/n})$ empirical excess risk in the projected space, which can decay faster than the $O(\ln n/k)$ expected excess risk of its data-independent ancestor [2] and partly suggests the high dimension efficiency of D^{2}RP.

We also apply D^{2}RP to accelerate kernel ridge regression [21] and ensemble least square regression [22]. Experimental results on real-world data sets show D^{2}RP-based regressions achieve the best tradeoff between computation-efficiency and dimension-efficiency compared to multiple baseline methods.

II. D^{2}RP for Least Square Regression

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a labeled data set where $x_i \in \mathbb{R}^p$ is the i_{th} instance and $y_i \in \mathbb{R}$ is its label. Let $R \in \mathbb{R}^{p \times k}$ be a projection matrix mapping instances from $\mathbb{R}^p$ to $\mathbb{R}^k$ and let $\tilde{x}_i = R^T x_i$ be the mapping of x_i. Our goal is to randomly generate an R and find a least square regression model in the projected space based on $(\tilde{x}_1, y_1), \ldots, (\tilde{x}_n, y_n)$.

Algorithm 1 Projection Generation Mechanism of D^{2}RP

Input: training data $x_1, \ldots, x_n \in \mathbb{R}^p$, projected dimension k

1: Sample $W \in \mathbb{R}^{k \times n}$ with i.i.d. entries from $N(0, 1/k)$.

2: Compute projection matrix $R = (WX)^T \in \mathbb{R}^{p \times k}$, where

$$X = [x_1, \ldots, x_n]^T \in \mathbb{R}^{n \times p}, \tag{1}$$

is the training data matrix.

Most random projection methods generate R independently to the instances e.g., by sampling entries i.i.d. from $N(1/k)$ [1]. We present a data-dependent generation mechanism D^{2}RP in Algorithm 1, which maps instances into the row space of a randomly projected training data matrix. In the following, we theoretically analyze its properties.

A. Data in the D^{2}RP Projected Space

We first investigate how D^{2}RP projection may impact data. Our following theorem suggests it may not preserve pairwise data distance as well as its data-independent ancestors.

Theorem II.1. *In Algorithm 1, if $k \geq \tilde{O}(\frac{\ln n}{\varepsilon^2})$, with probability at least $1 - \frac{1}{n}$, the followings hold simultaneously*

$$||R^T x_i - R^T x_j||^2 \leq (1+\varepsilon)\sigma_1^2(X)||x_i - x_j||^2, \tag{2}$$

and

$$||R^T x_i - R^T x_j||^2 \geq (1-\varepsilon)\sigma_r^2(X)||x_i - x_j||^2, \tag{3}$$

for any i, j, where $||\cdot||$ is Frobenius norm, $\sigma_j(X)$ is the j_{th} largest singular value of X (including ties) and $r = rank(X)$.

Proof Sketch. Observe that

$$\begin{aligned} ||R^T x_i - R^T x_j||^2 &= ||WXx_i - WXx_j||^2 \\ &\leq (1+\varepsilon) \cdot ||Xx_i - Xx_j||^2 \\ &\leq (1+\varepsilon) \cdot ||x_i - x_j||^2 \cdot \sigma_1^2(X) \end{aligned} \tag{4}$$

where the first inequality holds with probability at least $1 - \frac{1}{n}$ if $k \geq \tilde{O}(\frac{\ln n}{\varepsilon^2})$ based on the well-known Johnson–Lindenstrauss lemma (e.g., [1, Lemma 1.3] plus a union bound), and the second follows a basic matrix norm bound (e.g., [23]).

Mirrored arguments give the lower bound. □

The above theorem suggests D^{2}RP distorts data distance by $[\sigma_r^2(X)(1-\varepsilon), \sigma_1^2(X)(1+\varepsilon)]$, which is a bigger (relatively) range than the $[1-\varepsilon, 1+\varepsilon]$ of its data-independent ancestors since the condition number of X is often bigger than 1. This suggests distance preservation is not why D^{2}RP is useful.

Our next theorem is an immediate result of Lemma 4.2 and Remark 4.1 in [24]. It suggests D^{2}RP is useful as the projected space contains enough information for data reconstruction.

Theorem II.2. *In Algorithm 1, if $k = O(\frac{r}{\varepsilon})$ for some $r > 0$, with probability at least 0.9, then the column space of XR contains an $(1+\varepsilon)$ rank-r approximation of X i.e., there exists a $Q_* \in \mathbb{R}^{k \times p}$ such that*

$$||XRQ_* - X|| \leq (1+\varepsilon)||X_r - X||, \tag{5}$$

where $X_r \in \mathbb{R}^{n \times p}$ is the best rank-r approximation of X with respect to the Frobenius norm.

In prior study [20], a $O(\sqrt{\frac{n}{k}})$ reconstruction error is suggested presuming the tail eigenvalues of X decay as fast as $\sqrt{\sum_{j>r}\sigma_j^2(X)} \ll \sqrt{n}\sigma_{r+1}(X)$. Theorem II.2 matches this error under the same condition plus $\sigma_{r+1}(X) = O(\sqrt{1/r})$.

B. Least Square Regression in the D^{2}RP Projected Space

We now investigate the performance of least square regression in the D^{2}RP projected space. Recall $X \in \mathbb{R}^{n \times p}$ is the training data matrix and $R \in \mathbb{R}^{p \times k}$ is a D^{2}RP projection; let $Y = [y_1, \ldots, y_n]^T \in \mathbb{R}^n$ be the label vector of X.

Let $\beta \in \mathbb{R}^p$ and $\tilde{\beta} \in \mathbb{R}^k$ denote the models for raw data and projected data respectively. We define the empirical risk of $\beta \in \mathbb{R}^p$ on the raw training data (X, Y) as

$$\widehat{L}(\beta) = ||X\beta - Y||/\sqrt{n}, \tag{6}$$

empirical risk of $\tilde{\beta}$ on the projected training data (XR, Y) as

$$\widehat{L}_R(\tilde{\beta}) = ||XR\tilde{\beta} - Y||/\sqrt{n}. \tag{7}$$

Let β_* and $\tilde{\beta}_*$ be the minimizers of $\widehat{L}(\beta)$ and $\widehat{L}_R(\tilde{\beta})$ respectively. Define the empirical excess risk of any $\tilde{\beta}$ as

$$||\tilde{\beta} - \beta_*||_P := \widehat{L}_R(\tilde{\beta}) - \widehat{L}(\beta_*). \tag{8}$$

Our empirical excess risk bound is stated as follows.

Theorem II.3. *Given any training data (X, Y) and D^{2}RP projection matrix R, we have*

$$||\tilde{\beta}_* - \beta_*||_P = O(e^{-k/n}), \tag{9}$$

with probability at least 0.9 over the random choice of R.

Proof. Fix any $\varepsilon \in (0, 1/2)$ and pick a proper $c > 0$ (to be specified later). Set $r = ck/\varepsilon$. We have

$$\begin{aligned} \sqrt{n}\widehat{L}_R(\tilde{\beta}_*) &= \min_{\tilde{\beta}} ||(XR)\tilde{\beta} - Y|| \\ &\leq \min_{Q} ||(XR)Q\beta_* - Y|| \\ &= \min_{Q} ||(XR)Q\beta_* - X\beta_* + X\beta_* - Y|| \\ &\leq \min_{Q} ||(XR)Q - X||\,||\beta_*|| + ||X\beta_* - Y|| \\ &\leq (1+\varepsilon)||X_r - X||\,||\beta_*|| + ||X\beta_* - Y|| \\ &= (1+\varepsilon)||X_r - X||\,||\beta_*|| + \sqrt{n}\widehat{L}(\beta_*). \end{aligned} \tag{10}$$

The first inequality holds because we restrict the search space of $\tilde{\beta}$ to those satisfying $\tilde{\beta} = Q\beta_*$ for some $Q \in \mathbb{R}^{k \times p}$, and the third follows Theorem II.2 plus our setting (which implies $k = O(r/\varepsilon)$) with probability at least 0.9. The above implies

$$\widehat{L}_R(\tilde{\beta}_*) - \widehat{L}(\beta_*) \leq (1+\varepsilon)||\beta_*||\frac{||X_r - X||}{\sqrt{n}}. \tag{11}$$

Further, the last ratio satisfies $\frac{||X_r - X||}{\sqrt{n}} = \sqrt{\frac{\sum_{j>r}\sigma_j^2(X)}{n}} \leq \sqrt{1 - \frac{r}{n}}\sigma_{r+1}(X) \leq e^{-\frac{r}{2n}}\sigma_{r+1}(X) \leq e^{-\frac{ck}{2\varepsilon n}}\sigma_{r+1}(X)$. Picking any $c \geq 2\varepsilon$ proves the theorem. □

The theorem suggests D^{2}RP allows least square regression to achieve an $O(e^{-k/n})$ empirical excess risk in the projected space. This exponential decay (w.r.t. the increase of k) appears faster than the $O(\ln n/k)$ expected excess risk decay of its data-independent ancestors (e.g. [2, Theorem 1] presuming all norms are bounded) and partly justifies the higher dimension efficiency of D^{2}RP for least square regression.

Several gaps remain in our result, however. First, we only analyze the empirical excess risk while prior study [2] analyzes the expected excess risk. It is possible to bridge the gap using concentration properties, although it is unclear whether or how the exponential decay may survive under the extension. Second, we consider unsquared ℓ_2 loss for technical convenience, while prior study considers squared loss. It may be possible to extend our result for squared loss, but again whether the exponential decay rate may survive remains an open question. Finally, in the proof we pick $c \geq 2\varepsilon$ for conciseness, and even without it the empirical excess risk still enjoys a $O(e^{-c'k/n})$ decay rate for some constant $c' > 0$.

C. Compare with the Prior Data-Dependent RP Method

We now make a comprehensive comparison between D^{2}RP and the prior data-dependent random projection method [20].

From an algorithm view, our projection matrix is generated in a similar way to theirs, with a notable difference that we do not perform SVD on WX as they do. This change reduces the projection generation time from $O(k^2)$ to $O(k)$, and results in a significantly better trade-off between computation efficiency and dimension efficiency. (This will be clearer in experiments.)

In theory, we analyze D^{2}RP from three novel perspectives. The $O(\sqrt{\frac{n}{k}})$ data reconstruction error in [20, Theorem 3] is derived using a deterministic bound for subspace obtained via orthogonal projection [25], while our $O(\sqrt{\frac{n}{k}})$ reconstruction error implied by Theorem II.2 is derived using a probabilistic bound for subspace obtained via random projection [24, Lemma 4.2]. Secondly, the $O(\frac{1}{\sqrt{n}})$ excess risk in [20, Theorem 1] is for least square model reconstructed in the input space, while our $O(e^{-\frac{k}{2n}})$ excess risk implied by Theorem II.3 is for least square model in the projected space. At last, our Theorem II.1 suggests that D^{2}RP may not preserve data distance as well as its data-independent ancestors, which is a new insight not discussed in the prior study [20].

III. Two More Applications of D^{2}RP

A. D^{2}RP for Kernel Ridge Regression

Recall kernel ridge regression based on a data mapping $\phi : \mathbb{R}^p \rightarrow \mathbb{R}^q$ learns a linear model β in the mapped space by solving $\min_{\beta \in \mathbb{R}^q} ||X_\phi \beta - Y||^2 + \lambda ||\beta||^2$, where $X_\phi = [\phi(x_1), \ldots, \phi(x_n)]^T$. The solution is $\beta = X_\phi \alpha$ where

$$\alpha = (K^2 + \lambda K)^{-1} K Y, \tag{12}$$

and $K \in \mathbb{R}^{n \times n}$ is a kernel matrix with $K_{ij} = \phi(x_i)^T \phi(x_j)$.

Random projection has been used to accelerate the computation of (12) in [21], [26], by replacing K with KS with a random projection matrix $R \in \mathbb{R}^{n \times k}$. The new solution is

$$\tilde{\alpha} = (R^T K^2 R + \lambda R^T K R)^{-1} R^T K Y. \tag{13}$$

It is clear that (13) is faster to compute than (12) as its inverse is taken over a smaller matrix. However, prior studies generate R independently to data, which is not dimension-efficient. We propose to apply D^{2}RP to generate R. The process is almost the same as Algorithm 1, except that X is replaced with K.

B. D^{2}RP for Ensemble Regression

Random projection has also been used to accelerate ensemble least square regression in [22]. The idea is to learn a set of models from different randomly projected spaces and average them. Let $R_1, \ldots, R_m \in \mathbb{R}^{p \times k}$ be independently generated random projection matrices and β_t be a model learned by solving $\min_{\beta_t \in \mathbb{R}^k} ||X R_t \beta_t - Y||^2$. The ensemble model is then $\beta = \frac{1}{m} \sum_{t=1}^{m} \beta_t$. In prior study, R_t's are generated independently to data, and we propose to use D^{2}RP instead.

IV. Experiment

We experiment on the public MNIST [27] and High Dimensional Datascape [28] data sets. Since our focus is on reducing feature dimension, we randomly downsampled the MNIST data set to a subset of 1k instances. On each data set, we use the first 50% instances for training and the rest for testing. All reported results for random projection based methods are averaged over 200 trials.

A. Projected Least Square Regression

We evaluate the projected least square regression described in Section II-B. We generate R using the following methods.

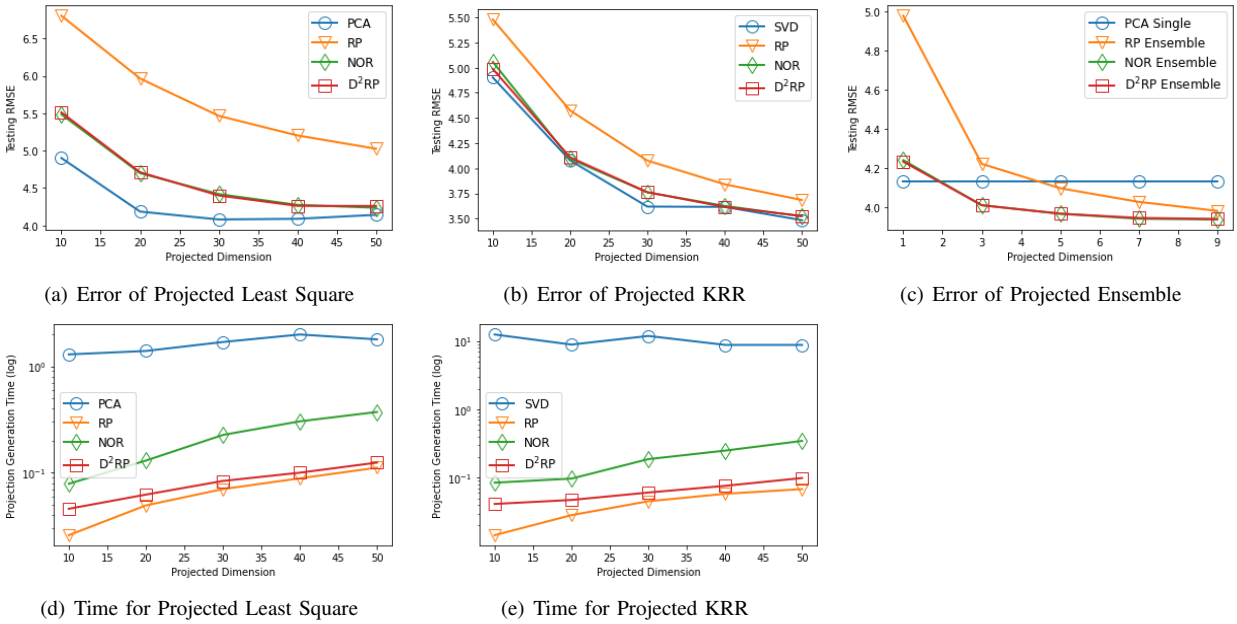
- PCA: R is learned from training data by PCA.
- RP: R has i.i.d. entries from $N(0, 1/k)$
- NOR: [20] where W has i.i.d. entries from $N(0, 1/k)$.
- D^{2}RP: the proposed method in Algorithm 1.

Results on MNIST are shown in Figure 1. In (a), we see both D^{2}RP and NOR give faster error convergence than RP, which verifies their superior dimension efficiency. We also see D^{2}RP has the same error rate as NOR in (a), but consumes much less generation time in (b). These show D^{2}RP achieves the best computation-efficiency and dimension-efficiency tradeoff. Similar patterns are observed on Datascape in Figure 2 (a)(d).

B. Projected Kernel Ridge Regression

We evaluate the projected kernel ridge regression described in Section III-A. For the kernel method, we use RBF kernel with γ set to 1e-7 on MNIST and 1e-3 on Datascape, and set the regularization coefficient to 0.1. Projection matrix R is generated by different methods, including RP, NOR and D^{2}RP. The 'SVD' method is a baseline that applies SVD on K and assign the top k left singular vectors to R.

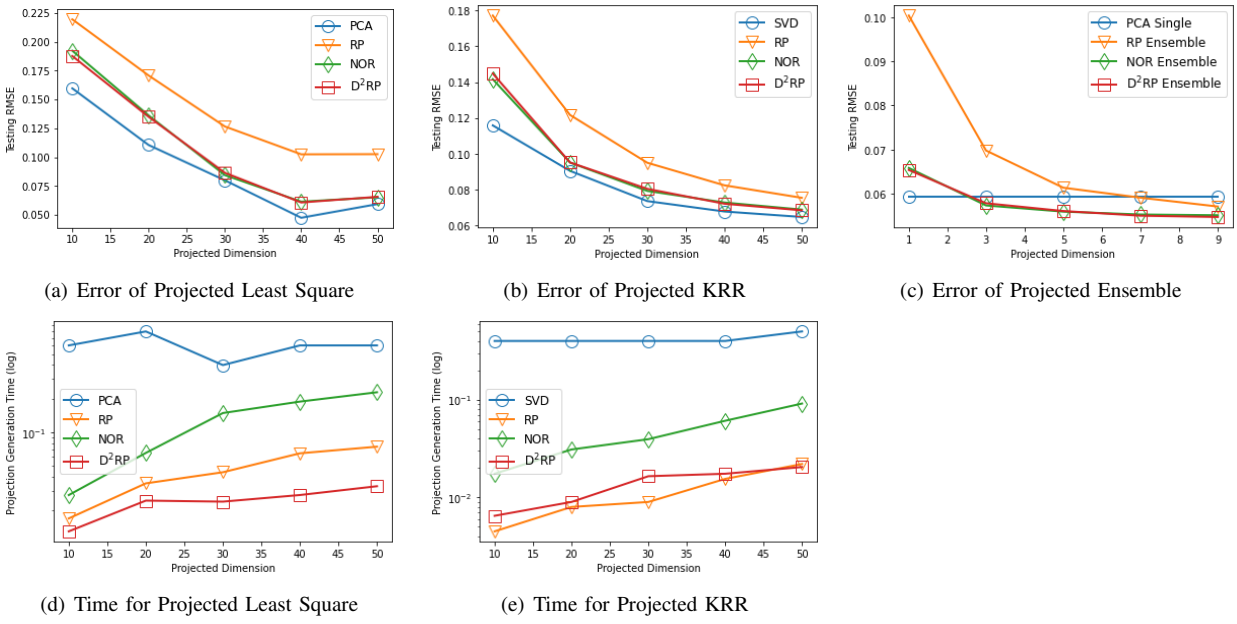
Results on MINST are shown in Figure 1. In (b), we see D^{2}RP and NOR have faster error convergence than RP, which verifies their superior dimension efficiency. We also see D^{2}RP has the same error rate as NOR in (b) but consumes much less generation time in (e). These, again, shows D^{2}RP achieves the best computation-efficiency and dimension-efficiency tradeoff. Similar patterns are found on Datascape in Figure 2 (b)(e).



(a) Error of Projected Least Square (b) Error of Projected KRR (c) Error of Projected Ensemble

(d) Time for Projected Least Square (e) Time for Projected KRR

Fig. 1. Performance on MNIST



(a) Error of Projected Least Square (b) Error of Projected KRR (c) Error of Projected Ensemble

(d) Time for Projected Least Square (e) Time for Projected KRR

Fig. 2. Performance on High Dimensional Datascape

C. *Projected Ensemble Regression*

Performance of the projected ensemble regressions are shown in Figures 1 (c) and 2 (c). The 'PCA Single' method learns a least square regression in a PCA projected space. The projected dimension for all methods are fixed to 50 on both data sets. We see both D^{2}RP and NOR errors converge faster than RP for the ensemble regression.

V. Conclusion and Future Work

This paper presents a new data-dependent random projection method D^{2}RP for least square regressions. Theoretical analysis suggests it preserves enough information for reconstructing the training data, and allows least square regression to achieve an $O(e^{-k/n})$ empirical excess risk in the projected space. We apply it to accelerate least square, kernel and ensemble regressions, and empirical results on real-world data sets show it achieves the best tradeoff between computation efficiency and dimension efficiency. In future, it may be interesting to explore other types of data-dependent random mappings (e.g., Fourier [29] or hypothesis [30]) or applications (e.g., data privacy [31] or matrix norm estimation [32]).

REFERENCES

[1] S. S. Vempala, *The random projection method*. American Mathematical Soc., 2005, vol. 65.

[2] O. Maillard and R. Munos, "Compressed least-squares regression," *Advances in neural information processing systems*, vol. 22, 2009.

[3] ——, "Linear regression with random projections," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 2735–2772, 2012.

[4] M. M. Fard, Y. Grinberg, J. Pineau, and D. Precup, "Compressed least-squares regression on sparse spaces," in *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[5] A. Kabán, "New bounds on compressive linear least squares regression," in *Artificial intelligence and statistics*. PMLR, 2014, pp. 448–456.

[6] B. McWilliams, C. Heinze, N. Meinshausen, G. Krummenacher, and H. P. Vanchinathan, "Loco: Distributing ridge regression with random projections," *stat*, vol. 1050, p. 26, 2014.

[7] R. Guhaniyogi and D. B. Dunson, "Bayesian compressed regression," *Journal of the American Statistical Association*, vol. 110, no. 512, pp. 1500–1514, 2015.

[8] M. Slawski, "Compressed least squares regression revisited," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1207–1215.

[9] D. Fradkin and D. Madigan, "Experiments with random projections for machine learning," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003, pp. 517–522.

[10] N. Goel, G. Bebis, and A. Nefian, "Face recognition experiments with random projection," in *Biometric Technology for Human Identification II*, vol. 5779. SPIE, 2005, pp. 426–437.

[11] S. Deegalla and H. Bostrom, "Reducing high-dimensional data by principal component analysis vs. random projection for nearest neighbor classification," in *2006 5th International Conference on Machine Learning and Applications (ICMLA'06)*. IEEE, 2006, pp. 245–250.

[12] R. J. Durrant and A. Kabán, "Compressed fisher linear discriminant analysis: Classification of randomly projected data," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 1119–1128.

[13] S. Paul, C. Boutsidis, M. Magdon-Ismail, and P. Drineas, "Random projections for support vector machines," in *Artificial intelligence and statistics*. PMLR, 2013, pp. 498–506.

[14] P. Li, M. Mitzenmacher, and A. Shrivastava, "Coding for random projections," in *International Conference on Machine Learning*. PMLR, 2014, pp. 676–684.

[15] K. Elkhalil, A. Kammoun, R. Calderbank, T. Y. Al-Naffouri, and M.-S. Alouini, "Asymptotic performance of linear discriminant analysis with random projections," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3472–3476.

[16] S. DASGUPTA, "Experiments with random projection," in *Uncertainty in Artificial Intelligence*, 2000, pp. 143–151.

[17] X. Z. Fern and C. E. Brodley, "Random projection for high dimensional data clustering: A cluster ensemble approach," in *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003, pp. 186–193.

[18] C. Boutsidis, A. Zouzias, and P. Drineas, "Random projections for k-means clustering," *Advances in neural information processing systems*, vol. 23, 2010.

[19] Y. Yang and P. Li, "Noisy l0-sparse subspace clustering on dimensionality reduced data," in *Uncertainty in Artificial Intelligence*. PMLR, 2022, pp. 2235–2245.

[20] Y. Xu, H. Yang, L. Zhang, and T. Yang, "Efficient non-oblivious randomized reduction for risk minimization with improved excess risk guarantee," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[21] Y. YANG, M. PILANCI, and M. J. WAINWRIGHT, "Randomized sketches for kernels: Fast and optimal nonparametric regression," *The Annals of Statistics*, vol. 45, no. 3, pp. 991–1023, 2017.

[22] G.-A. Thanei, C. Heinze, and N. Meinshausen, "Random projections for large-scale regression," *Big and Complex Data Analysis: Methodologies and Applications*, pp. 51–68, 2017.

[23] Y. Fang, K. A. Loparo, and X. Feng, "Inequalities for the trace of matrix product," *IEEE Transactions on Automatic Control*, vol. 39, no. 12, pp. 2489–2490, 1994.

[24] D. P. Woodruff *et al.*, "Sketching as a tool for numerical linear algebra," *Foundations and Trends® in Theoretical Computer Science*, vol. 10, no. 1–2, pp. 1–157, 2014.

[25] N. Halko, P.-G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM review*, vol. 53, no. 2, pp. 217–288, 2011.

[26] Y. Chen and Y. Yang, "Accumulations of projections—a unified framework for random sketches in kernel ridge regression," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 2953–2961.

[27] [Online]. Available: https://www.kaggle.com/datasets/oddrationale/mnist-in-csv

[28] [Online]. Available: https://www.kaggle.com/datasets/krishd123/high-dimensional-datascape?resource=download

[29] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in neural information processing systems*, 2008.

[30] Y. Cao and C. Lan, "A model-agnostic randomized learning framework based on random hypothesis subspace sampling," in *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[31] K. Kenthapadi, A. Korolova, I. Mironov, and N. Mishra, "Privacy via the johnson-lindenstrauss transform," *Journal of Privacy and Confidentiality*, vol. 5, no. 1, pp. 39–71, 2013.

[32] Y. Cao, S. Shafaei, L. Yang, and C. Lan, "Efficient estimation of kernel matrix spectral norm using random features," *International Conference on Acoustics, Speech, and Signal Processing*, 2025.