SYMMETRIC MATRIX PERTURBATION FOR DIFFERENTIALLY-PRIVATE PRINCIPAL COMPONENT ANALYSIS

Hafiz Imtiaz and Anand D. Sarwate

Rutgers, The State University of New Jersey

ABSTRACT

Differential privacy is a strong, cryptographically-motivated definition of privacy that has recently received a significant amount of research attention for its robustness to known attacks. The principal component analysis (PCA) algorithm is frequently used in signal processing, machine learning and statistics pipelines. In this paper, we propose a new algorithm for differentially-private computation of PCA and compare the performance empirically with some recent state-of-the-art algorithms on different data sets. We intend to investigate the performance of these algorithms with varying privacy parameters and database parameters. We show that our proposed algorithm, despite guaranteeing stricter privacy, provides very good utility for different data sets.

Index Terms— Differential privacy, dimensionality reduction, principal component analysis

1. INTRODUCTION

Analyzing private or sensitive data using machine learning and signal processing algorithms is a topic of increasing importance. Standard data analytics pipelines often use the singular value decomposition (SVD), or principal component analysis (PCA) to pre-process high-dimensional data by projecting it onto a lower dimensional subspace spanned by the singular vectors of the second-moment matrix of the data. For example, to save on the computational complexity of training a classifier, the algorithm may first project the data into lower dimension. In this paper, we propose an algorithm that approximates PCA while satisfying differential privacy [1].

Differential privacy (DP) measures privacy risk in terms of the probability of identifying individual data points in a data set from the results of computations performed on that data. There are several generic approaches to making DP approximations of algorithms [2, 3], including PCA. Input perturbation [4, 5] adds noise to the data prior to computing the SVD, whereas output perturbation [1] adds noise to the output of the desired algorithm. The Analyze Gauss algorithm of Dwork et al. [5] adds Gaussian noise to the data second-moment matrix. Hardt and Price [6] proposed a differentially private version of the power method that runs in near-linear time. Chaudhuri et al. [7] proposed a method based on the exponential mechanism [8], which samples random orthonormal basis using a utility function. Their implementation uses Markov Chain Monte Carlo (MCMC) sampling and is hence only approximate. Kapralov and Talwar [9] also used the exponential mechanism but sampled vectors sequentially; it runs in polynomial time but is intractable to implement for high dimensional data. Most recently, Sheffet [10] proposed adding noise from Wishart distribution to achieve differentially-private linear regression.

In this paper, we propose a new algorithm, SN, for differentially private principal component analysis. Our method also adds Wishart noise, but with parameters chosen to yield a better privacy guarantee. We compare SN with others [5–7] on the problem of computing and publishing a *private orthonormal subspace* using synthetic and real data sets. We analyze the variation of utility with different privacy level, number of samples and some other key parameters. Our results show that for strong privacy guarantees $(\epsilon, 0)$, SN outperforms other methods, and that weaker privacy guarantees (ϵ, δ) can yield significantly higher utility. We also show that despite guaranteeing stronger privacy, SN can achieve similar utility level as algorithms with weaker privacy guarantees. Due to space constraints, some details are deferred to the journal version of this work.

2. PROBLEM FORMULATION

Consider a dataset $\mathbb{D} = \{x_i \in \mathbb{R}^d : i = 1, 2 \ldots, n\}$ with n data samples corresponding to n individuals. We further assume $\|x_i\|_2 \leq 1$. Let $X = [x_1, x_2, \ldots, x_n]$ be the $d \times n$ data matrix whose i-th column is x_i. Define the $d \times d$ positive semi-definite second-moment matrix A:

$$A = \sum_{i=1}^{n} x_i x_i^\top = XX^\top. \tag{1}$$

For the setup of this paper, we have $\|\frac{1}{n} XX^\top\|_F \leq 1$, where $\|\cdot\|_F$ denotes the Frobenius norm. We define two

The work of the authors was supported by the NSF under award CCF-1453432 and by the NIH under award 1R01DA040487-01A1.

data sets to be *neighbors* if they differ in a single data point (column). If $X = [x_1, x_2, \ldots, x_{n-1}, x_n]$ and $X' = [x_1, x_2, \ldots, x_{n-1}, x'_n]$ are matrices corresponding to two neighboring data sets, then $A = XX^\top$ and $A' = X'X'^\top$ satisfy the condition $\|A - A'\|_2 \leq 1$. We will also use the relation $\|A\|_2 \leq \|A\|_F$ between the Frobenius norm and the $\mathscr{L}_2$ norm.

The Schmidt approximation theorem [11] characterizes the rank-k matrix A_k that minimizes the difference $\|A - A_k\|_F$ and shows that the minimizer can be found by taking the singular value decomposition of A:

$$A = V\Lambda V^\top, \tag{2}$$

where without loss of generality we assume Λ is a diagonal matrix $\mathrm{diag}(\lambda_1(A), \lambda_2(A), \ldots, \lambda_d(A))$ with $\lambda_1(A) \geq \lambda_2(A) \geq \ldots \geq \lambda_d(A) \geq 0$ and V is a matrix of eigenvectors corresponding to the eigenvalues. The *top-k PCA subspace* of A is the matrix $V_k(A) = [v_1, v_2, \ldots, v_k]$, where v_i is the i-th column of V. Given $V_k(A)$ and the eigenvalue matrix Λ, we can form an approximation $A_k = V_k(A)\Lambda_k V_k(A)^\top$ to A, where Λ_k contains the k largest eigenvalues in Λ. For a $d \times k$ matrix $\hat{V}$ with orthonormal columns, the quality of $\hat{V}$ in approximating $V_k(A)$ can be measured by the *captured variance* of A as

$$q(\hat{V}) = \mathrm{tr}(\hat{V}^\top A \hat{V}). \tag{3}$$

The $\hat{V}$, which maximizes $q(\hat{V})$ has columns equal to v_i for $i = 1, 2, ..., k$, corresponding to the top-k eigenvectors of A.

In this paper we study algorithms that approximate the top-k PCA subspace $V_k(A)$ while also guaranteeing differential privacy [1]. An algorithm $\mathscr{A}(\mathbb{B})$ taking values in a set $\mathbb{T}$ provides (ϵ, δ)-differential privacy if

$$\Pr(\mathscr{A}(\mathbb{D}) \in \mathbb{S}) \leq \exp(\epsilon)\Pr(\mathscr{A}(\mathbb{D}') \in \mathbb{S}) + \delta, \tag{4}$$

for all measurable $\mathbb{S} \subseteq \mathbb{T}$ and all data sets $\mathbb{D}$ and $\mathbb{D}'$ differing in a single entry. This definition essentially states that the probability of the output of an algorithm is not changed significantly if the corresponding database input is changed by just one entry. Here ϵ and δ are privacy parameters, where low ϵ and δ ensure more privacy. It should be noted here that the parameter δ can be interpreted as the probability that the algorithm fails. Therefore, an $(\epsilon, 0)$-differentially private algorithm guarantees much stronger privacy than an (ϵ, δ)-differentially private algorithm, where $\delta > 0$. We refer to $(\epsilon, 0)$ differential privacy as ϵ-differential privacy. For more details, see the recent survey [2] or monograph [12].

3. ALGORITHMS

Differentially private algorithms for approximating $V(A)$, the matrix containing the eigenvectors of A, either guarantee (ϵ, δ) or ϵ-differential privacy. Some of them approximate $V(A)$ by approximating A and then taking the SVD of the approximation $\hat{A}$. We propose a method for approximating A under ϵ-differential privacy.

Algorithm 1: SN Algorithm

Input : Data matrix $X \in \mathbb{R}^{d \times n}$ (with n samples of dimension d, each sample has bounded norm), privacy parameter ϵ

1 $A \leftarrow XX^\top$

2 Generate $d \times p$ matrix $Z = [z_1, z_2, \ldots, z_p]$ where $z_i \sim \mathscr{N}(0, \frac{1}{2\epsilon} I)$ and $p = d + 1$

3 $\hat{A} \leftarrow A + ZZ^\top$

Output: The private second-moment matrix $\hat{A}$. The private orthonormal basis matrix can be calculated by computing $SVD(\hat{A})$

Proposed Symmetric Noise (SN) Algorithm. Let z_i be a d-dimensional random vector drawn according to $\mathscr{N}(0, \Sigma)$, where $\Sigma = \frac{1}{2\epsilon} I$. We generate $p = d + 1$ iid samples and form a $d \times p$ noise matrix $Z = [z_1, z_2, \ldots, z_p]$. The random matrix $E = ZZ^\top$ is sample from a *Wishart* $W_d(\Sigma, p)$ distribution with p degrees of freedom [13]. Our algorithm outputs $\hat{A} = A + E$ as a private approximation to A.

Theorem 1 (Privacy of SN Algorithm) *Algorithm 1 computes an ϵ-differentially private approximation to A.*

Proof. The Wishart $W_d(\Sigma, p)$ with distribution $\Sigma = \frac{1}{2\epsilon} I$ and $p = d + 1$ has density

$$\begin{aligned} f_E(E) &\propto (\det(E))^{\frac{p-d-1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}E\right)\right) \\ &\propto \exp\left(-\epsilon\,\mathrm{tr}(E)\right), \end{aligned}$$

where the second proportionality is achieved by substituting the parameters Σ and p. Consider two neighboring databases with second moment matrices A and A' and an output Y from SN. The density of Y is $f_E(Y - A)$ under input A and $f_E(Y - A')$ under input A'. Therefore, using the assumption that the data is bounded,

$$\begin{aligned} \frac{f_E(Y - A)}{f_E(Y - A')} &= \frac{\exp\left(-\epsilon\,\mathrm{tr}(Y - A)\right)}{\exp\left(-\epsilon\,\mathrm{tr}(Y - A')\right)} \\ &= \exp\left(-\epsilon\,\mathrm{tr}(A' - A)\right) \\ &= \exp\left(\epsilon\,\mathrm{tr}\left(x_n x_n^\top - x'_n x'^\top_n\right)\right) \\ &\leq \exp(\epsilon). \end{aligned}$$

Thus, the addition of the positive semi-definite noise matrix E makes the algorithm ϵ-differentially private.

Theorem 2 (SN Approximation Guarantees) *If V_k is the top-k right singular subspace of X and $\hat{V}_k$ is the private*

subspace derived from computation of SVD on the output of Algorithm 1, then

$$\|\hat{V}_{k'}^\top X\|_F^2 \geq \|V_k^\top X\|_F^2 + O\left(k\left(\frac{d}{4\epsilon^2}\right)^2\right)$$

$$\|V_k V_k^\top - \hat{V}_{k'}\hat{V}_{k'}^\top\|_2 \leq O\left(\frac{\frac{d}{4\epsilon^2}}{\lambda_k - \lambda_{k+1}}\right)$$

$$\|A - \hat{A}(k)\|_2 \leq \|A - A_k\|_2 + O\left(\frac{d}{4\epsilon^2}\right).$$

Due to space limitations, we present only the sketch of the proof of the first inequality. We assume that $\lambda_k - \lambda_{k'+1} = \omega(\frac{\sqrt{d}}{2\epsilon})$ for $k' \geq k$. Then it follows that $\mathrm{tr}(\hat{V}_{k'}^T A \hat{V}_{k'}) = \mathrm{tr}(V_k^T A V_k) + \mathrm{tr}\left((P - \hat{P})E\right)$, where $P = V_k V_k^\top$ and $\hat{P} = \hat{V}_{k'}\hat{V}_{k'}^\top$. Using Von Neumann's trace inequality, we have $\left|\mathrm{tr}\left((P-\hat{P})E\right)\right| \leq \sqrt{2}k'\|E\|_2\left(\|P\hat{P}^\perp\|_2 + \|\hat{P}P^\perp\|_2\right)$, where $^\perp$ represents the orthogonal complement. Using the sin-θ theorem [14] and Weyl's inequality we have $\|P\hat{P}^\perp\|_2 = O\left(\frac{d/4\epsilon^2}{\lambda_k - \lambda_{k'+1}}\right)$. The inequality guarantee of captured variance follows from this.

In a simultaneous, independent work, Sheffet also proposed addition of Wishart noise to preserve differential privacy in the context of linear regression [10]. Our proposed method uses specific parameters to guarantee ϵ-differential privacy rather than his (ϵ, δ)-differential privacy. This distinction can be important for specific applications.

Previous algorithms. We empirically compare SN with three other algorithms: the Analyze Gauss (AG) algorithm of Dwork et al. [5], the private power method (PPM) of Hardt and Price [6], and the Private PCA (PPCA) algorithm of Chaudhuri et al. [7]. All of these methods have favorable theoretical guarantees but limited empirical validation. The AG method generates a symmetric noise matrix E of i.i.d. Gaussian entries with distribution $\mathcal{N}(0, \Delta_{\epsilon,\delta}^2)$ and publishes $A + E$, where $\Delta_{\epsilon,\delta}^2$ guarantees (ϵ, δ)-differential privacy. Unlike AG, our SN algorithm preserves the covariance structure of the perturbed matrix: SN's perturbation can be thought of as adding fictitious data points to X, whereas the output of AG may not even be positive semi-definite. The PPM algorithm adds noise in the iterations of the power method; this noise can be chosen to guarantee ϵ- or (ϵ, δ)-differential privacy. An open question is how to choose the number of iterations L. We chose the suggested scaling $L = O\left(\frac{\sigma_k}{\sigma_k - \sigma_{k+1}}\log(d)\right)$, where σ_i are the singular values of the data matrix X sorted in descending order. The variance of the Gaussian or Laplace noise to be added in each step of the power iteration depends on a parameter χ, which we set to $\frac{1}{\epsilon}\sqrt{4kL\log(\frac{1}{\delta})}$ for $\delta > 0$ and $\frac{10}{\epsilon}kL\sqrt{d}$ for $\delta = 0$. Finally, the PPCA algorithm samples a random orthonormal basis $\hat{V}$ using the exponential mechanism [8] with the utility function (3), which is a sample from the matrix Bingham distribution [15]. Our implementations of these algorithms are available [16].

4. EXPERIMENTAL RESULTS

Because these algorithms have a large parameter space, we focused on measuring how well the outputs of these algorithms approximate the true PCA subspace $V(A)$. Here we focus on the energy captured by the privately generated subspace. We studied the dependence of the energy on the privacy parameter ϵ and the sample size n, as well as the utility of the private subspace as preprocessing for a classification task.

We performed experiments using three data sets: a *synthetic data* set ($d = 100$, $n = 60000$, $k = 10$) generated with a pre-determined covariance matrix, the *Covertype* dataset ($d = 54$, $k = 10$) [17] (COVTYPE) and the *MNIST* ($d = 784$, $k = 50$) [18]. For the latter two we selected 20000 and 10000 samples at random, respectively, for our experiments. We preprocessed the data by subtracting the mean (centering) and normalizing the data with the maximum $\mathscr{L}_2$ norm in each set to enforce the condition $\|x_i\|_2 \leq 1$. We picked the reduced dimension k so that the top-k PCA subspace $V_k(A)$ has captured variance $q(V_k(A))$ that is at least 90% of $q(V(A))$. In all cases we show the average performance over 10 runs of each algorithm.

Dependence on Privacy Parameter ϵ. We first explored the *privacy-utility tradeoff* between ϵ and the captured variance. For the additive-noise algorithms, the standard deviation of the noise (Gaussian or Laplace) is inversely proportional to ϵ – smaller ϵ means more noise and lower privacy risk. For PPCA, an increase in ϵ means skewing the probability density function more towards the optimal subspace. In Fig. 1, we show the variation of percentage captured energy (with respect to SVD) with different values of ϵ. For all the data sets, we observed that as ϵ increases (higher privacy risk), the captured variance increases. The AG method vastly outperforms the PPM method; we believe this is because the noise stability for PPM may only hold for larger data sets or larger ϵ. Our new SN method also outperforms existing methods (PPM and PPCA) and for large enough ϵ it matches the performance of AG, despite providing a stronger privacy guarantee.

Dependence on Number of Samples n. Intuitively, it should be easier to guarantee smaller privacy risk ϵ and higher utility $q(\cdot)$ when the number of samples is large. Figure 2 shows how the captured variance increases as a function of sample size for the different algorithms. The variation with the sample size reinforces the results seen earlier with variation in ϵ: AG and SN have the best performance for $\delta > 0$ and $\delta = 0$, respectively, and PPM appears to suffer from too much noise. Interestingly, AG, SN, and PPCA all show a steep improvement with sample size, perhaps indicating a relationship between the convergence of the sample covariance as well as its private approximation.

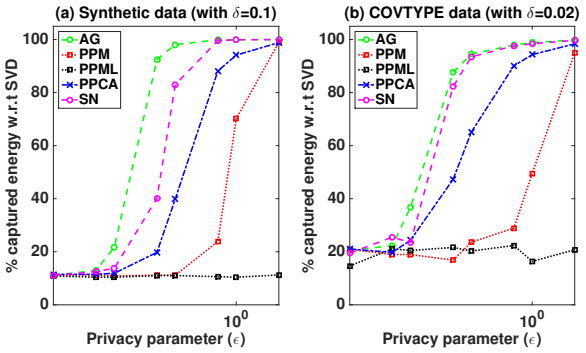

Fig. 1. Variation of the captured variance with different ϵ for (a) Synthetic data and (b) COVTYPE data

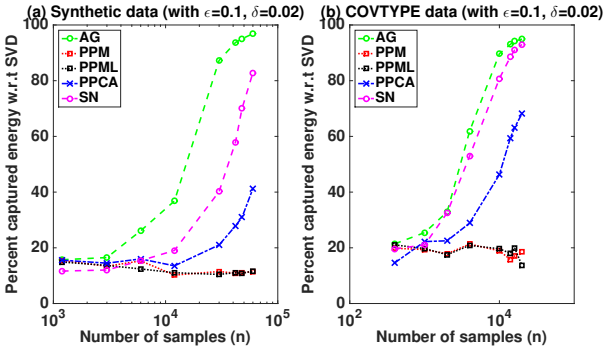

Fig. 2. Variation of the captured variance with different number of data samples n for (a) synthetic data, (b) COVTYPE data

Classification Performance. We also wanted to see how useful the differentially private subspace $\hat{V}$ was as a preprocessing step for a classification task. We projected the d-dimensional data samples onto the private k-dimensional subspace $\hat{V}$. Using an original training dataset $\mathbb{D}_{\text{tr}} = \{(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\} : i = 1, 2, \ldots, n\}$ and a private approximation $\hat{V}$ to $V_k(A)$ we created a projected data set $\{(\hat{V}^\top x_i, y_i) \in \mathbb{R}^k \times \{-1, 1\} : i = 1, 2, \ldots, n\}$ and trained a support vector machine (SVM) classifier to find the weight vector $f \in \mathbb{R}^k$ for a linear classifier $\text{sgn}(f^\top x_i^k)$, where x_i^k is the i-th k-dimensional sample. For this experiment, we formed our data sets slightly differently. The synthetic data set ($d = 100$, $n = 5000$) was generated i.i.d. Gaussian with one of two different means corresponding to the label y and a fixed covariance matrix with bounded spectral norm. The COVTYPE data set has 7 classes: we chose class 6 and class 7, with 10000 random samples from each class. Finally, for the MNIST data set, we chose two digits - digit 3 and digit 7, with 5000 samples selected randomly. We solved the optimization problem for classification using a built-in SVM classifier `svmtrain` in MATLAB. Table 1 shows the percentage errors of classification on the three data sets for all the three algorithms. We performed the experiments keeping the test sample size fixed at 4000 samples and varying the training sample size. Judging by the recognition accuracy compared to COVTYPE and MNIST, we note that the synthetic data set is a bit more difficult than the COVTYPE and MNIST due to the fact that the classes have comparatively smaller separation. For a particular privacy level (i.e. fixed ϵ and δ) with sufficient training samples, the AG algorithm performed consistently well. On the COVTYPE and MNIST data sets, our proposed SN algorithm outperformed all other private methods, even those with (ϵ, δ) guarantees. These observations certainly point out that the proposed algorithm provides a private subspace that not only can capture a significant amount of variance from the data second-moment matrix but also suited very well for projection and classification purposes.

Table 1. Percentage error in classification with varying training sample size

	SYNTHETIC		COVTYPE		MNIST	
$\lvert\mathbb{D}_{\text{tr}}\rvert = n$	4000	6000	4000	6000	4000	6000
SVD	5.65	5.80	0.025	0.025	0.575	0.25
AG	5.80	5.75	16.80	5.775	3.025	3.125
PPM	9.00	9.05	22.00	15.175	3.10	2.55
PPCA	6.375	6.125	31.85	19.875	2.725	2.80
SN	7.50	7.075	7.225	0.35	2.50	2.00

5. CONCLUSION AND FUTURE WORK

In this paper, we proposed a new algorithm SN for differentially private PCA. Comparing private feature learning methods may reveal their robustness to perturbations. We empirically compared SN algorithm with three recent state-of-the-art competitors on three different data sets. In general, the AG and the SN algorithms had the best performance among (ϵ, δ) and ϵ-private methods, respectively. In some regimes and on some data sets, SN achieved as much utility as AG, even though SN provides stricter privacy guarantee. We further examined the usefulness of the produced subspace for classification using SVM and showed that SN even outperformed non-private SVD for one data set. For data sets with a large *eigengap*, the SN algorithm provided a very close approximation to the subspace from SVD. Overall, SN and AG algorithms provided the best performance across data sets and privacy parameters. Our initial results suggest that the asymptotic guarantees for differentially private algorithms may not always reflect their empirical performance. We also note that because differential privacy is closed under post-processing, other feature extraction and classification techniques that use the second-moment matrix can use SN or AG to provide ϵ or (ϵ, δ)-differential privacy.

6. REFERENCES

[1] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proceedings of the Third Conference on Theory of Cryptography*, 2006, pp. 265–284.

[2] A. D. Sarwate and K. Chaudhuri, "Signal processing and machine learning with differential privacy: theory, algorithms, and challenges," *IEEE Signal Processing Magazine*, vol. 30, no. 5, pp. 86–94, September 2013.

[3] C. Dwork and A. Smith, "Differential privacy for statistics: What we know and what we want to learn," in *In Proceedings of the 33rd International Colloquium on Automata, Languages and Programming, volume 4052 of LECTURE NOTES IN COMPUTER SCIENCE*. 2009, p. 1, Springer.

[4] A. Blum, C. Dwork, F. McSherry, and K. Nissim, "Practical privacy: The sulq framework," in *Proceedings of the Twenty-fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 2005, pp. 128–138.

[5] C. Dwork, K. Talwar, A. Thakurta, and L. Zhang, "Analyze Gauss: Optimal bounds for privacy-preserving principal component analysis," in *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, 2014, pp. 11–20.

[6] M. Hardt and E. Price, "The noisy power method: A meta algorithm with applications," in *Advances in Neural Information Processing Systems 27*, pp. 2861–2869. Curran Associates, Inc., 2014.

[7] K. Chaudhuri, A. D. Sarwate, and K. Sinha, "A near-optimal algorithm for differentially-private principal components," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 2905–2943, Jan. 2013.

[8] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS '07)*, October 2007, pp. 94–103.

[9] M. Kapralov and K. Talwar, "On differentially private low rank approximation," in *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2013, pp. 1395–1414.

[10] O. Sheffet, "Private approximations of the 2nd-moment matrix using existing techniques in linear regression," Tech. Rep. arXiv:1507.00056 [cs.DS], ArXiV, June 2015.

[11] G. W. Stewart, "On the early history of the singular value decomposition," *SIAM Rev.*, vol. 35, no. 4, pp. 551–566, Dec. 1993.

[12] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2013.

[13] J. Wishart, "The generalised product moment distribution in samples from a normal multivariate population.," Biometrika 20, A, 32-52 (1928)., 1928.

[14] Frank Mcsherry, *Spectral Methods for Data Analysis*, Ph.D. thesis, 2004, AAI3118856.

[15] Y. Chikuse, *Statistics on special manifolds*, Springer - Lecture Notes in Statistics, 2003.

[16] H. Imtiaz and A. D. Sarwate, "`https://www.dropbox.com/s/dsfwijzw0jbrmql/Algorithms_Imtiaz_Sarwate_2015.zip?dl=0`.".

[17] M. Lichman, "UCI machine learning repository," 2013.

[18] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.