Principal Component Projection Without Principal Component Analysis

Roy Frostig RF@CS.STANFORD.EDU
Stanford University

Cameron Musco CNMUSCO@MIT.EDU
Christopher Musco CPMUSCO@MIT.EDU
MIT

Aaron Sidford ASID@MICROSOFT.COM
Microsoft Research, New England

Abstract

We show how to efficiently project a vector onto the top principal components of a matrix, *without explicitly computing these components*. Specifically, we introduce an iterative algorithm that provably computes the projection using few calls to any black-box routine for ridge regression.

By avoiding explicit principal component analysis (PCA), our algorithm is the first with no runtime dependence on the number of top principal components. We show that it can be used to give a fast iterative method for the popular principal component regression problem, giving the first major runtime improvement over the naive method of combining PCA with regression.

To achieve our results, we first observe that ridge regression can be used to obtain a "smooth projection" onto the top principal components. We then sharpen this approximation to true projection using a low-degree polynomial approximation to the matrix step function. Step function approximation is a topic of long-term interest in scientific computing. We extend prior theory by constructing polynomials with simple iterative structure and rigorously analyzing their behavior under limited precision.

1. Introduction

In machine learning and statistics, it is common—often essential—to represent data in a concise form that decreases noise and increases efficiency in downstream tasks. Perhaps the most widespread method for doing so is to project data onto the linear subspace spanned by its directions of highest variance—that is, onto the span of the top components given by principal component analysis (PCA). Computing principal components can be an expensive task, a challenge that prompts a basic algorithmic question:

> *Can we project a vector onto the span of a matrix's top principal components without performing principal component analysis?*

This paper answers that question in the affirmative, demonstrating that projection is much easier than PCA itself. We show that it can be solved using a simple iterative algorithm based on black-box calls to a ridge regression routine. The algorithm's runtime *does not depend* on the number of top principal components chosen for projection, a cost inherent to any algorithm for PCA, or even algorithms that just compute an orthogonal span for the top components.

1.1. Motivation: principal component regression

To motivate our projection problem, consider one of the most basic downstream applications for PCA: linear regression. Combined, PCA and regression comprise the *principal component regression* (PCR) problem:

Definition 1.1 (Principal component regression (PCR))**.** *Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ be a design matrix whose rows are data points and let $\mathbf{b} \in \mathbb{R}^n$ be a vector of data labels. Let $\mathbf{A}_\lambda \in \mathbb{R}^{n \times d}$ denote the result of projecting each row of $\mathbf{A}$ onto the span of the top principal components of $\mathbf{A}$, i.e. the eigenvectors of the covariance matrix $\frac{1}{n}\mathbf{A}^\mathsf{T}\mathbf{A}$ whose corresponding eigenvalue exceeds a threshold λ. The task of PCR is to find a minimizer of the squared loss $\|\mathbf{A}_\lambda \mathbf{x} - \mathbf{b}\|_2^2$. In other words, the goal is to compute $\mathbf{A}_\lambda^\dagger \mathbf{b}$, where $\mathbf{A}_\lambda^\dagger$ is the Moore-Penrose pseudoinverse of $\mathbf{A}_\lambda$.*

PCR is a key regularization method in statistics, numerical

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

linear algebra, and scientific disciplines including chemometrics (Hotelling, 1957; Hansen, 1987; Frank & Friedman, 1993). It models the assumption that small principal components represent noise rather than data signal. PCR is typically solved by first using PCA to compute $\mathbf{A}_\lambda$ and then applying linear regression. The PCA step dominates the algorithm's cost, especially if many principal components have variance above the threshold λ.

We remedy this issue by showing that our principal component *projection* algorithm yields a fast algorithm for *regression*. Specifically, full access to $\mathbf{A}_\lambda$ is unnecessary for PCR: $\mathbf{A}_\lambda^\dagger \mathbf{b}$ can be computed efficiently given only an approximate projection of the vector $\mathbf{A}^\mathsf{T}\mathbf{b}$ onto $\mathbf{A}$'s top principal components. By solving projection without PCA we obtain the first PCA-free algorithm for PCR.

1.2. A first approximation: ridge regression

Our approach to efficient principal component projection is actually based on a common alternative to PCR: ridge regression. This ubiquitous method computes a minimizer of $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_2^2$ for some regularization parameter λ (Tikhonov, 1963). The advantage of ridge regression is that it is a simple convex optimization problem that can be solved efficiently using many techniques (see Lemma 2.1).

Solving ridge regression is equivalent to applying the matrix $(\mathbf{A}^\mathsf{T}\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^\mathsf{T}$, an operation that can be viewed as a smooth relaxation of PCR. Adding the ℓ_2 norm penalty (i.e. $\lambda\mathbf{I}$) effectively "washes out" $\mathbf{A}$'s small principal components in comparison to its large ones and achieves an effect similar to PCR at the extreme ends of $\mathbf{A}$'s spectrum.

Accordingly, ridge regression gives access to a "smooth projection" operator, $(\mathbf{A}^\mathsf{T}\mathbf{A}+\lambda\mathbf{I})^{-1}\mathbf{A}^\mathsf{T}\mathbf{A}$, which approximates $\mathbf{P}_{\mathbf{A}_\lambda}$, the projection onto $\mathbf{A}$'s top row principal components. Both have the same singular vectors, but $\mathbf{P}_{\mathbf{A}_\lambda}$ has a singular value of 1 for each squared singular value $\sigma_i^2 \geq \lambda$ in $\mathbf{A}$ and a singular value of 0 for each $\sigma_i^2 < \lambda$, whereas $(\mathbf{A}^\mathsf{T}\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^\mathsf{T}\mathbf{A}$ has singular values equal to $\frac{\sigma_i^2}{\sigma_i^2+\lambda}$. This function approaches 1 when σ_i^2 is much greater than λ and 0 when it is smaller. Figure 1 shows the comparison.

Unfortunately, in many settings, ridge regression is a very crude approximation to PCR and projection and may perform significantly worse in certain data analysis applications (Dhillon et al., 2013). In short, while ridge regression algorithms are valuable tools, it has been unclear how to wield them for tasks like projection or PCR.

1.3. Main result: from ridge regression to projection

We show that it is possible to *sharpen* the weak approximation given by ridge regression. Specifically, there exists a low degree polynomial $p(\cdot)$ such that

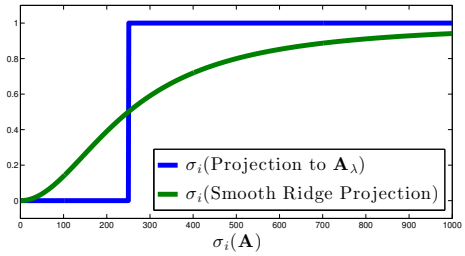

Figure 1: Singular values of the projection matrix $\mathbf{P}_{\mathbf{A}_\lambda}$ vs. those of the smooth projection operator $(\mathbf{A}^\mathsf{T}\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^\mathsf{T}\mathbf{A}$ obtained from ridge regression.

$p\left((\mathbf{A}^\mathsf{T}\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^\mathsf{T}\mathbf{A}\right)\mathbf{y}$ provides a very accurate approximation to $\mathbf{P}_{\mathbf{A}_\lambda}\mathbf{y}$ for any $\mathbf{y} \in \mathbb{R}^d$. Moreover, the polynomial can be evaluated as a recurrence, which translates into a simple iterative algorithm: we can apply the sharpened approximation to a vector by repeatedly applying any ridge regression routine a small number of times.

Theorem 1.2 (Principal component projection without PCA)**.** *Given $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\mathbf{y} \in \mathbb{R}^d$, Algorithm 1 uses $\tilde{O}(\gamma^{-2}\log(1/\epsilon))$ approximate applications of $(\mathbf{A}^\mathsf{T}\mathbf{A} + \lambda\mathbf{I})^{-1}$ and returns $\mathbf{x}$ with $\|\mathbf{x} - \mathbf{P}_{\mathbf{A}_\lambda}\mathbf{y}\|_2 \leq \epsilon\|\mathbf{y}\|_2$.*

Like all iterative PCA algorithms, our running time scales inversely with γ, the *spectral gap* around λ.[1] Notably, it does not depend on the number of principal components in $\mathbf{A}_\lambda$, a cost incurred by any method that applies the projection $\mathbf{P}_{\mathbf{A}_\lambda}$ directly, either by explicitly computing the top principal components of $\mathbf{A}$, or even by just computing an orthogonal span for these components.

As mentioned, the above theorem also yields an algorithm for principal component *regression* that computes $\mathbf{A}_\lambda^\dagger\mathbf{b}$ without finding $\mathbf{A}_\lambda$. We achieve this result by introducing a robust reduction, from projection to PCR, that again relies on ridge regression as a computational primitive.

Corollary 1.3 (Principal component regression without PCA)**.** *Given $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\mathbf{b} \in \mathbb{R}^n$, Algorithm 2 uses $\tilde{O}(\gamma^{-2}\log(1/\epsilon))$ approximate applications of $(\mathbf{A}^\mathsf{T}\mathbf{A} + \lambda\mathbf{I})^{-1}$ and returns $\mathbf{x}$ with $\|\mathbf{x} - \mathbf{A}_\lambda^\dagger\mathbf{b}\|_{\mathbf{A}^\mathsf{T}\mathbf{A}} \leq \epsilon\|\mathbf{b}\|_2$.*

Corollary 1.3 gives the first known algorithm for PCR that avoids the cost of principal component analysis.

1.4. Related work

A number of papers attempt to alleviate the high cost of principal component analysis when solving PCR. It has

[1] See Section 3.2 for a discussion of this gap dependence. Aside from a full SVD requiring $O(nd^2)$ time, any PCA algorithm giving the guarantee of Theorem 1.2 will have a dependence both on γ *and* on the number principal components in $\mathbf{A}_\lambda$. However, the γ dependence can be better – $\gamma^{-1/2}$ for Krylov methods (Musco & Musco, 2015), giving a runtime tradeoff.

been shown that an approximation to $\mathbf{A}_\lambda$ suffices for solving the regression problem (Chan & Hansen, 1990; Boutsidis & Magdon-Ismail, 2014). Unfortunately, even the fastest approximations are much slower than routines for ridge regression and inherently incur a linear dependence on the number of principal components above λ.

More closely related to our approach is work on the *matrix sign function*, an important operation in control theory, quantum chromodynamics, and scientific computing in general. Approximating the sign function often involves matrix polynomials similar to our "sharpening polynomial" that converts ridge regression to principal component projection. Significant effort addresses Krylov methods for applying such operators without computing them explicitly (van den Eshof et al., 2002; Frommer & Simoncini, 2008).

Our work differs from these methods in an important way: since we only assume access to an approximate ridge regression algorithm, it is essential that our sharpening step is robust to noise. Our iterative polynomial construction allows for a complete and rigorous noise analysis that is not available for Krylov methods, while at the same time eliminating space and post-processing costs. Iterative approximations to the matrix sign function have been proposed, but lack rigorous noise analysis (Higham, 2008).

1.5. Paper layout

Section 2: Mathematical and algorithmic preliminaries.

Section 3: Develop a PCA-free algorithm for principal component projection based on a ridge regression subroutine.

Section 4: Show how our approximate projection algorithm can be used to solve PCR, again without PCA.

Section 5: Detail our iterative approach to sharpening the smooth ridge regression projection towards true projection via a low degree sharpening polynomial.

Section 6: Empirical evaluation of our principal component projection and regression algorithms.

2. Preliminaries

Singular value decomposition. Any matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ of rank r has a singular value decomposition (SVD) $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}$, where $\mathbf{U} \in \mathbb{R}^{n\times r}$ and $\mathbf{V} \in \mathbb{R}^{d\times r}$ both have orthonormal columns and $\mathbf{\Sigma} \in \mathbb{R}^{r\times r}$ is a diagonal matrix. The columns of $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular vectors of $\mathbf{A}$. Moreover, $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1(\mathbf{A}), ..., \sigma_r(\mathbf{A}))$, where $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq ... \geq \sigma_r(\mathbf{A}) > 0$ are the singular values of $\mathbf{A}$ in decreasing order.

The columns of $\mathbf{V}$ are the eigenvectors of the covariance matrix $\mathbf{A}^\mathsf{T}\mathbf{A}$, i.e. the *principal components* of the data, and the eigenvalues of the covariance matrix are the squares of the singular values $\sigma_1, \ldots, \sigma_r$.

Functions of matrices. If $f : \mathbb{R} \to \mathbb{R}$ is a scalar function and $\mathbf{S} = \mathrm{diag}(s_1, \ldots, s_n)$ is a diagonal matrix, we define by $f(\mathbf{S}) \stackrel{\text{def}}{=} \mathrm{diag}(f(s_1), \ldots, f(s_n))$ the entrywise application of f to the diagonal. For a non-diagonal matrix $\mathbf{A}$ with SVD $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}$ we define $f(\mathbf{A}) \stackrel{\text{def}}{=} \mathbf{U}f(\mathbf{\Sigma})\mathbf{V}^\mathsf{T}$.

Matrix pseudoinverse. We define the *pseudoinverse* of $\mathbf{A}$ as $\mathbf{A}^\dagger = f(\mathbf{A})^\mathsf{T}$ where $f(x) = 1/x$. The pseudoinverse is essential in the context of regression, as the vector $\mathbf{A}^\dagger\mathbf{b}$ minimizes the squared error $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$.

Principal component projection. Given a threshold $\lambda > 0$ let k be the largest index with $\sigma_k(\mathbf{A})^2 \geq \lambda$ and define:

$$\mathbf{A}_\lambda \stackrel{\text{def}}{=} \mathbf{U}\,\mathrm{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)\mathbf{V}^\mathsf{T}.$$

The matrix $\mathbf{A}_\lambda$ contains $\mathbf{A}$'s rows projected to the span of all principal components having squared singular value at least λ. We sometimes write $\mathbf{A}_\lambda = \mathbf{A}\mathbf{P}_{\mathbf{A}_\lambda}$ where $\mathbf{P}_{\mathbf{A}_\lambda} \in \mathbb{R}^{d\times d}$ is the projection onto these top components. Here $\mathbf{P}_{\mathbf{A}_\lambda} = f(\mathbf{A}^\mathsf{T}\mathbf{A})$ where $f(x)$ is a step function: 0 if $x < \lambda$ and 1 if $x \geq \lambda$.

Miscellaneous notation. For any positive semidefinite $\mathbf{M}, \mathbf{N} \in \mathbb{R}^{d\times d}$ we use $\mathbf{N} \preceq \mathbf{M}$ to denote that $\mathbf{M} - \mathbf{N}$ is positive semidefinite. For any $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{x}\|_\mathbf{M} \stackrel{\text{def}}{=} \sqrt{\mathbf{x}^\mathsf{T}\mathbf{M}\mathbf{x}}$.

Ridge regression. Ridge regression is the problem of computing, given a regularization parameter $\lambda > 0$:

$$\mathbf{x}^\lambda \stackrel{\text{def}}{=} \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_2^2. \tag{1}$$

The solution to (1) is given by $\mathbf{x}^\lambda = \left(\mathbf{A}^\mathsf{T}\mathbf{A} + \lambda\mathbf{I}\right)^{-1}\mathbf{A}^\mathsf{T}\mathbf{b}$. Applying the matrix $\left(\mathbf{A}^\mathsf{T}\mathbf{A} + \lambda\mathbf{I}\right)^{-1}$ to $\mathbf{A}^\mathsf{T}\mathbf{b}$ is equivalent to solving the convex minimization problem:

$$\mathbf{x}^\lambda \stackrel{\text{def}}{=} \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}\mathbf{x}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x} - \mathbf{y}^\mathsf{T}\mathbf{x} + \lambda\|\mathbf{x}\|_2^2.$$

A vast literature studies solving problems of this form via (accelerated) gradient descent, stochastic variants, and random sketching (Nesterov, 1983; Nelson & Nguyên, 2013; Shalev-Shwartz & Zhang, 2014; Lin et al., 2014; Frostig et al., 2015; Cohen et al., 2015). We summarize a few, now standard, runtimes achievable by these iterative methods:

Lemma 2.1 (Ridge regression runtimes)**.** *Given* $\mathbf{y} \in \mathbb{R}^d$ *let* $\mathbf{x}^* = (\mathbf{A}^\mathsf{T}\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{y}$. *There is an algorithm,* $\textsc{ridge}(\mathbf{A}, \lambda, \mathbf{y}, \epsilon)$ *that, for any* $\epsilon > 0$*, returns* $\tilde{\mathbf{x}}$ *such that*

$$\|\tilde{\mathbf{x}} - \mathbf{x}^*\|_{\mathbf{A}^\mathsf{T}\mathbf{A}+\lambda\mathbf{I}} \leq \epsilon\|\mathbf{y}\|_{(\mathbf{A}^\mathsf{T}\mathbf{A}+\lambda\mathbf{I})^{-1}}.$$

It runs in time $T_{\textsc{ridge}}(\mathbf{A}, \lambda, \epsilon) = O\left(\mathrm{nnz}(\mathbf{A})\sqrt{\kappa_\lambda} \cdot \log(1/\epsilon)\right)$ *where* $\kappa_\lambda = \sigma_1^2(\mathbf{A})/\lambda$ *is*

the condition number of the regularized system and $\mathrm{nnz}(\mathbf{A})$ *is the number of nonzero entries in* $\mathbf{A}$*. There is a also stochastic algorithm that, for any* $\delta > 0$*, gives the same guarantee with probability* $1 - \delta$ *in time*

$$T_{\text{RIDGE}}(\mathbf{A}, \lambda, \epsilon, \delta) = O\left((\mathrm{nnz}(\mathbf{A}) + d\,\mathrm{sr}(\mathbf{A})\kappa_\lambda) \cdot \log(1/\delta\epsilon)\right),$$

where $\mathrm{sr}(\mathbf{A}) = \|\mathbf{A}\|_F^2/\|\mathbf{A}\|_2^2$ *is* $\mathbf{A}$*'s stable rank. When* $\mathrm{nnz}(\mathbf{A}) \geq d\,\mathrm{sr}(\mathbf{A})\kappa_\lambda$ *the runtime can be improved to*

$$T_{\text{RIDGE}}(\mathbf{A}, \lambda, \epsilon, \delta) = \tilde{O}(\sqrt{\mathrm{nnz}(\mathbf{A}) \cdot d\,\mathrm{sr}(\mathbf{A})\kappa_\lambda} \cdot \log(1/\delta\epsilon)),$$

where the $\tilde{O}$ *hides a factor of* $\log\left(\frac{d\,\mathrm{sr}(\mathbf{A})\kappa_\lambda}{\mathrm{nnz}(\mathbf{A})}\right)$.

Typically, the regularized condition number κ_λ will be significantly smaller than the full condition number of $\mathbf{A}^\mathsf{T}\mathbf{A}$.

3. From ridge regression to principal component projection

We now describe how to approximately apply $\mathbf{P}_{\mathbf{A}_\lambda}$ using any black-box ridge regression routine. The key idea is to first compute a soft step function of $\mathbf{A}^\mathsf{T}\mathbf{A}$ via ridge regression, and then to sharpen this step to approximate $\mathbf{P}_{\mathbf{A}_\lambda}$.

Let $\mathbf{B}\mathbf{x} = (\mathbf{A}^\mathsf{T}\mathbf{A} + \lambda\mathbf{I})^{-1}(\mathbf{A}^\mathsf{T}\mathbf{A})\mathbf{x}$ be the result of applying ridge regression to $(\mathbf{A}^\mathsf{T}\mathbf{A})\mathbf{x}$. In the language of functions of matrices, we have $\mathbf{B} = r(\mathbf{A}^\mathsf{T}\mathbf{A})$, where

$$r(x) \stackrel{\text{def}}{=} \frac{x}{x + \lambda}.$$

The function $r(x)$ is a smooth step about λ (see Figure 1). It primarily serves to map the eigenvalues of $\mathbf{A}^\mathsf{T}\mathbf{A}$ to the range $[0, 1]$, mapping those exceeding the threshold λ to a value above $1/2$ and the rest to a value below $1/2$.

To approximate the projection $\mathbf{P}_{\mathbf{A}_\lambda}$, it would now suffice to apply a simple symmetric step function:

$$s(x) = \begin{cases} 0 & \text{if } x < 1/2 \\ 1 & \text{if } x \geq 1/2 \end{cases}$$

It is easy to see that $s(\mathbf{B}) = s(r(\mathbf{A}^\mathsf{T}\mathbf{A})) = \mathbf{P}_{\mathbf{A}_\lambda}$. For $x \geq \lambda$, $r(x) \geq 1/2$ and so $s(r(x)) = 1$. Similarly for $x < \lambda$, $r(x) < 1/2$ and hence $s(r(x)) = 0$. That is, the symmetric step function exactly converts our smooth ridge regression step to the true projection operator.

3.1. Polynomial approximation to the step function

While computing $s(\mathbf{B})$ directly is expensive, requiring the SVD of $\mathbf{B}$, we show how to approximate this function with a *low-degree polynomial*. We also show how to apply this polynomial efficiently and stably using a simple iterative algorithm. Our main result, proven in Section 5, is:

Lemma 3.1 (Step function algorithm)**.** *Let* $\mathbf{S} \in \mathbb{R}^{d\times d}$ *be symmetric with every eigenvalue* σ *satisfying* $\sigma \in [0, 1]$ *and* $|\sigma - 1/2| \geq \gamma$*. Let* $\mathcal{A}$ *denote a procedure that on* $\mathbf{x} \in \mathbb{R}^d$ *produces* $\mathcal{A}(\mathbf{x})$ *with* $\|\mathcal{A}(\mathbf{x}) - \mathbf{S}\mathbf{x}\|_2 = O(\epsilon^2\gamma^2)\|\mathbf{x}\|_2$*. Given* $\mathbf{y} \in \mathbb{R}^d$ *set* $\mathbf{s}_0 := \mathcal{A}(\mathbf{y})$*,* $\mathbf{w}_0 := \mathbf{s}_0 - \frac{1}{2}\mathbf{y}$*, and for* $k \geq 0$ *set*

$$\mathbf{w}_{k+1} := 4\left(\frac{2k+1}{2k+2}\right)\mathcal{A}(\mathbf{w}_k - \mathcal{A}(\mathbf{w}_k))$$

and $\mathbf{s}_{k+1} := \mathbf{s}_k + \mathbf{w}_{k+1}$*. If all arithmetic operations are performed with* $\Omega(\log(d/\epsilon\gamma))$ *bits of precision then* $\|\mathbf{s}_q - s(\mathbf{S})\mathbf{y}\|_2 = O(\epsilon)\|\mathbf{y}\|_2$ *for* $q = \Theta(\gamma^{-2}\log(1/\epsilon))$*.*

Note that the output $\mathbf{s}_q$ is an approximation to a $2q$ degree polynomial of $\mathbf{S}$ applied to $\mathbf{y}$. In Algorithm 1, we give pseudocode for combining the procedure with ridge regression to solve principal component projection. Set $\mathbf{S} = \mathbf{B}$ and let $\mathcal{A}$ be an algorithm that approximately applies $\mathbf{B}$ to any $\mathbf{x}$ by applying approximate ridge regression to $\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x}$. As long as $\mathbf{B}$ has no eigenvalues falling within γ of $1/2$, the lemma ensures $\|\mathbf{s}_q - \mathbf{P}_{\mathbf{A}_\lambda}\mathbf{y}\|_2 = O(\epsilon)\|\mathbf{y}\|_2$. This requires γ on order of the *spectral gap*: $1 - \sigma_{k+1}^2(\mathbf{A})/\sigma_k^2(\mathbf{A})$, where k is the largest index with $\sigma_k^2(\mathbf{A}) \geq \lambda$.

Algorithm 1 (PC-PROJ) Principal component projection

input: $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{y} \in \mathbb{R}^d$, error ϵ, failure rate δ, threshold λ, gap $\gamma \in (0, 1)$

```
q := c_1 γ^{-2} log(1/ε)
ε' := c_2^{-1} ε^2 γ^2 / sqrt(κ_λ),   δ' := δ/(2q)
s := RIDGE(A, λ, A^T A y, ε', δ')
w := s − (1/2) y
for k = 0, ..., q − 1 do
    t := w − RIDGE(A, λ, A^T A w, ε', δ')
    w := 4 ((2k+1)/(2k+2)) RIDGE(A, λ, A^T A t, ε', δ')
    s := s + w
end for
return s
```

Theorem 3.2. *If* $\frac{1}{1-4\gamma}\sigma_{k+1}(\mathbf{A})^2 \leq \lambda \leq (1 - 4\gamma)\sigma_k(\mathbf{A})^2$ *and* c_1, c_2 *are sufficiently large constants,* PC-PROJ *(Algorithm 1) returns* $\mathbf{s}$ *such that with probability* $\geq 1 - \delta$*,*

$$\|\mathbf{s} - \mathbf{P}_{\mathbf{A}_\lambda}\mathbf{y}\|_2 \leq \epsilon\|\mathbf{y}\|_2.$$

The algorithm requires $O(\gamma^{-2}\log(1/\epsilon))$ *ridge regression calls, each costing* $T_{\text{RIDGE}}(\mathbf{A}, \lambda, \epsilon', \delta')$*. Lemma 2.1 yields total cost (with no failure probability)*

$$O\left(\mathrm{nnz}(\mathbf{A})\sqrt{\kappa_\lambda}\gamma^{-2}\log(1/\epsilon)\log\left(\kappa_\lambda/(\epsilon\gamma)\right)\right)$$

or, via stochastic methods,

$$\tilde{O}\left((\mathrm{nnz}(\mathbf{A}) + d\,\mathrm{sr}(\mathbf{A})\kappa_\lambda)\,\gamma^{-2}\log(1/\epsilon)\log(\kappa_\lambda/(\epsilon\gamma\delta))\right)$$

with acceleration possible when $\mathrm{nnz}(\mathbf{A}) > d\,\mathrm{sr}(\mathbf{A})\kappa_\lambda$.

Proof. We instantiate Lemma 3.1. Let $\mathbf{S} = \mathbf{B} = (\mathbf{A}^\mathsf{T}\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^\mathsf{T}\mathbf{A}$. As discussed, $\mathbf{B} = r(\mathbf{A}^\mathsf{T}\mathbf{A})$ and hence all its eigenvalues fall in $[0, 1]$. Specifically, $\sigma_i(\mathbf{B}) = \frac{\sigma_i(\mathbf{A})^2}{\sigma_i(\mathbf{A})^2+\lambda}$. Now, $\sigma_k(\mathbf{B}) \geq \frac{\lambda/(1-4\gamma)}{\lambda/(1-4\gamma)+\lambda} = \frac{1}{2-4\gamma} \geq \frac{1}{2} + \gamma$ and similarly $\sigma_{k+1}(\mathbf{B}) \leq \frac{\lambda(1-4\gamma)}{\lambda(1-4\gamma)+\lambda} = \frac{1-4\gamma}{2-4\gamma} \leq \frac{1}{2} - \gamma$, so all eigenvalues of $\mathbf{B}$ are at least γ far from $1/2$.

By Lemma 2.1, for any $\mathbf{x}$, with probability $\geq 1 - \delta'$:

$$\|\text{RIDGE}(\mathbf{A},\lambda,\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x},\epsilon',\delta') - \mathbf{B}\mathbf{x}\|_{\mathbf{A}^\mathsf{T}\mathbf{A}+\lambda\mathbf{I}} \leq \epsilon'\|\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x}\|_{(\mathbf{A}^\mathsf{T}\mathbf{A}+\lambda\mathbf{I})^{-1}} \leq \sigma_1(\mathbf{A})\epsilon'\|\mathbf{x}\|_2.$$

Since the minimum eigenvalue of $\mathbf{A}^\mathsf{T}\mathbf{A} + \lambda\mathbf{I}$ is λ:

$$\|\text{RIDGE}(\mathbf{A},\lambda,\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x},\epsilon',\delta') - \mathbf{B}\mathbf{x}\|_2 \leq \frac{\sigma_1(\mathbf{A})}{\sqrt{\lambda}}\epsilon'\|\mathbf{x}\|_2 \leq \frac{\sqrt{\kappa_\lambda}\epsilon^2\gamma^2}{c_2\sqrt{\kappa_\lambda}}\|\mathbf{x}\|_2 = O(\epsilon^2\gamma^2)\|\mathbf{x}\|_2.$$

Union bounding over all $2q$ calls of RIDGE, this bound holds for all calls with probability $\geq 1 - \delta' \cdot 2q = 1 - \delta$.

Overall, by Lemma 3.1, $\|\mathbf{s} - s(\mathbf{B})\mathbf{y}\|_2 = O(\epsilon)\|\mathbf{y}\|_2$ with probability $1 - \delta$. As discussed, $s(\mathbf{B}) = \mathbf{P}_{\mathbf{A}_\lambda}$. Adjusting constants on ϵ (via c_1, c_2) completes the proof. □

Note that the runtime of Theorem 3.2 depends on $\sqrt{\kappa_\lambda}$. In performing principal component projection, PC-PROC applies an *asymmetric step function* to $\mathbf{A}^\mathsf{T}\mathbf{A}$. The best polynomial approximating this step also has a $\sqrt{\kappa_\lambda}$ dependence (Eremenko & Yuditskii, 2011), so our reduction from projection to ridge regression is optimal in this regard.

3.2. Choosing λ and γ

Theorem 3.2 requires $\frac{\sigma_{k+1}(\mathbf{A})^2}{1-4\gamma} \leq \lambda \leq (1-4\gamma)\sigma_k(\mathbf{A})^2$. If λ is chosen approximately equidistant from the two eigenvalues, we need $\gamma = O(1 - \sigma_{k+1}^2(\mathbf{A})/\sigma_k^2(\mathbf{A}))$.

In practice, however, it is unnecessary to explicitly specify γ or to choose λ so precisely. With $q = O(\gamma^{-2}\log(1/\epsilon))$ our projection will be approximately correct on all singular values outside the range $[(1-\gamma)\lambda, (1+\gamma)\lambda]$. If there are any "intermediate" singular values in this range, as shown in Section 5, the approximate step function applied by Lemma 3.1 will map these values to $[0, 1]$ via a monotonically increasing soft step. That is, Algorithm 1 gives a slightly softened projection, removing any principal directions with value below $(1-\gamma)\lambda$, keeping any with value above $(1+\gamma)\lambda$, and partially projecting any in between.

4. From principal component projection to principal component regression

A major motivation for an efficient, PCA-free method for projecting a vector onto the span of top principal components is principal component regression (PCR). Recall that PCR solves the following problem:

$$\mathbf{A}_\lambda^\dagger\mathbf{b} = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^d} \|\mathbf{A}_\lambda\mathbf{x} - \mathbf{b}\|_2^2.$$

In *exact* arithmetic, $\mathbf{A}_\lambda^\dagger\mathbf{b}$ is equal to $(\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{P}_{\mathbf{A}_\lambda}\mathbf{A}^\mathsf{T}\mathbf{b}$. This identity suggests a method for computing the solution to ridge regression without finding $\mathbf{A}_\lambda$ explicitly: first apply a principal component projection algorithm to $\mathbf{A}^\mathsf{T}\mathbf{b}$ and then solve a linear system to apply $(\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}$.

Unfortunately, this approach is disastrously unstable, not only when $\mathbf{P}_{\mathbf{A}_\lambda}$ is applied approximately, but in any finite precision environment. We instead present a modified method for obtaining PCA-free regression from projection.

4.1. Stable inversion via ridge regression

Let $\mathbf{y} = \mathbf{P}_{\mathbf{A}_\lambda}\mathbf{A}^\mathsf{T}\mathbf{b}$ and suppose we have $\tilde{\mathbf{y}} \approx \mathbf{y}$ (e.g. from Algorithm 1). The issue with the above approach is that $(\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}$ can have a very large max eigenvalue, so we cannot guarantee $(\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\tilde{\mathbf{y}} \approx (\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{y}$. On the other hand, applying the ridge regression operator $(\mathbf{A}^\mathsf{T}\mathbf{A} + \lambda\mathbf{I})^{-1}$ is much more stable – it has max eigenvalue $1/\lambda$, so $(\mathbf{A}^\mathsf{T}\mathbf{A} + \lambda\mathbf{I})^{-1}\tilde{\mathbf{y}}$ will approximate $(\mathbf{A}^\mathsf{T}\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{y}$ well.

In short, it is more *stable* to apply $(\mathbf{A}^\mathsf{T}\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{y} = f(\mathbf{A}^\mathsf{T}\mathbf{A})\mathbf{y}$, where $f(x) = \frac{1}{x+\lambda}$, but the goal in PCR is to apply $(\mathbf{A}^\mathsf{T}\mathbf{A})^{-1} = h(\mathbf{A}^\mathsf{T}\mathbf{A})$ where $h(x) = 1/x$. So, in order to go from one function to the other, we use a correction function $g(x) = \frac{x}{1-\lambda x}$. By simple calculation,

$$\mathbf{A}_\lambda^\dagger\mathbf{b} = (\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{y} = g((\mathbf{A}^\mathsf{T}\mathbf{A} + \lambda\mathbf{I})^{-1})\mathbf{y}.$$

Additionally, we can stably approximate $g(x)$ with an iteratively computed low degree polynomial! Specifically, we use a truncation of the series $g(x) = \sum_{i=1}^{\infty}\lambda^{i-1}x^i$. Exact computation of $g(x)$ would exactly apply $(\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}$, which as discussed, is unstable due to large eigenvalues (corresponding to small eigenvalues of $\mathbf{A}^\mathsf{T}\mathbf{A}$). Our approximation to $g(x)$ is accurate on the large eigenvalues of $\mathbf{A}^\mathsf{T}\mathbf{A}$ but *inaccurate* on the small eigenvalues. This turns out to be the key to stability – by not "fully inverting" these eigenvalues, we avoid the instability of applying $(\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}$. We give a full analysis in Appendix B, yielding:

Lemma 4.1 (PCR approximation algorithm)**.** *Let $\mathcal{A}$ be a procedure that, given $\mathbf{x} \in \mathbb{R}^d$, produces $\mathcal{A}(\mathbf{x})$ with $\|\mathcal{A}(\mathbf{x}) - (\mathbf{A}^\mathsf{T}\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{x}\|_{\mathbf{A}^\mathsf{T}\mathbf{A}+\lambda\mathbf{I}} = O(\frac{\epsilon}{q^2\sigma_1(\mathbf{A})})\|\mathbf{x}\|_2$. Let $\mathcal{B}$ be a procedure that, given $\mathbf{x} \in \mathbb{R}^d$ produces $\mathcal{B}(\mathbf{x})$ with $\|\mathcal{B}(\mathbf{x}) - \mathbf{P}_{\mathbf{A}_\lambda}\mathbf{x}\|_2 = O(\frac{\epsilon}{q^2\sqrt{\kappa_\lambda}})\|\mathbf{x}\|_2$. Given $\mathbf{b} \in \mathbb{R}^n$ set $\mathbf{s}_0 := \mathcal{B}(\mathbf{A}^\mathsf{T}\mathbf{b})$ and $\mathbf{s}_1 := \mathcal{A}(\mathbf{s}_0)$. For $k \geq 1$ set:*

$$\mathbf{s}_{k+1} := \mathbf{s}_1 + \lambda \cdot \mathcal{A}(\mathbf{s}_k).$$

If all arithmetic operations are performed with $\Omega(\log(d/q\epsilon))$ bits of precision then $\|\mathbf{s}_q - \mathbf{A}_\lambda^\dagger\mathbf{b}\|_{\mathbf{A}^\mathsf{T}\mathbf{A}} = O(\epsilon)\|\mathbf{b}\|_2$ for $q = \Theta(\log(\kappa_\lambda/\epsilon))$.

We instantiate the iterative procedure above in Algorithm 2. PC-PROJ$(\mathbf{A}, \lambda, \mathbf{y}, \gamma, \epsilon, \delta)$ denotes a call to Algorithm 1.

Algorithm 2 (RIDGE-PCR) Ridge regression-based PCR

input: $\mathbf{A} \in \mathbb{R}^{n\times d}$, $\mathbf{b} \in \mathbb{R}^n$, error ϵ, failure rate δ, threshold λ, gap $\gamma \in (0,1)$

$q := c_1 \log(\kappa_\lambda/\epsilon)$
$\epsilon' := c_2^{-1}\epsilon/(q^2\sqrt{\kappa_\lambda}), \quad \delta' = \delta/2(q+1)$
$\mathbf{y} :=$ PC-PROJ$(\mathbf{A}, \lambda, \mathbf{A}^\mathsf{T}\mathbf{b}, \gamma, \epsilon', \delta/2)$
$\mathbf{s}_0 :=$ RIDGE$(\mathbf{A}, \lambda, \mathbf{y}, \epsilon', \delta')$, $\mathbf{s} := \mathbf{s}_0$
for $k = 1, ..., q$ **do**
  $\mathbf{s} := \mathbf{s}_0 + \lambda \cdot$ RIDGE$(\mathbf{A}, \lambda, \mathbf{s}, \epsilon', \delta')$
end for
return $\mathbf{s}$

Theorem 4.2. *If $\frac{1}{1-4\gamma}\sigma_{k+1}(\mathbf{A})^2 \le \lambda \le (1-4\gamma)\sigma_k(\mathbf{A})^2$ and c_1, c_2 are sufficiently large constants,* RIDGE-PCR *(Algorithm 2) returns* $\mathbf{s}$ *such that with probability* $\ge 1-\delta$,

$$\|\mathbf{s} - \mathbf{A}_\lambda^\dagger \mathbf{b}\|_{\mathbf{A}^\mathsf{T}\mathbf{A}} \le \epsilon \|\mathbf{b}\|_2.$$

The algorithm makes one call to PC-PROJ *and $O(\log(\kappa_\lambda/\epsilon))$ calls to ridge regression, each of which costs $T_{\text{RIDGE}}(\mathbf{A}, \lambda, \epsilon', \delta')$, so Lemma 2.1 and Theorem 3.2 imply a total runtime of*

$$\tilde{O}(\text{nnz}(\mathbf{A})\sqrt{\kappa_\lambda}\gamma^{-2}\log^2(\kappa_\lambda/(\epsilon\gamma))),$$

where $\tilde{O}$ hides $\log\log(1/\epsilon)$, or, with stochastic methods,

$$\tilde{O}((\text{nnz}(\mathbf{A}) + d\,\text{sr}(\mathbf{A})\kappa_\lambda)\gamma^{-2}\log^2(\kappa_\lambda/(\epsilon\gamma\delta))).$$

Proof. We apply Lemma 4.1; $\mathcal{A}$ is given by RIDGE$(\mathbf{A}, \lambda, \mathbf{x}, \epsilon', \delta')$. Since $\|(\mathbf{A}^\mathsf{T}\mathbf{A} + \lambda\mathbf{I})^{-1}\|_2 < 1/\lambda$, Lemma 2.1 states that with probability $1-\delta'$,

$$\|\mathcal{A}(\mathbf{x}) - (\mathbf{A}^\mathsf{T}\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{x}\|_{\mathbf{A}^\mathsf{T}\mathbf{A}+\lambda\mathbf{I}} \le \epsilon'\|\mathbf{x}\|_{(\mathbf{A}^\mathsf{T}\mathbf{A}+\lambda\mathbf{I})^{-1}} \le \frac{c_2^{-1}\epsilon}{q^2\sqrt{\kappa_\lambda\lambda}}\|\mathbf{x}\|_2 \le \frac{c_2^{-1}\epsilon}{q^2\sigma_1(\mathbf{A})}\|\mathbf{x}\|_2.$$

Now, $\mathcal{B}$ is given by PC-PROJ$(\mathbf{A}, \lambda, \mathbf{x}, \gamma, \epsilon', \delta/2)$. With probability $1-\delta/2$, if $\frac{1}{1-4\gamma}\sigma_{k+1}(\mathbf{A})^2 \le \lambda \le (1-4\gamma)\sigma_k(\mathbf{A})^2$ then by Theorem 3.2, $\|\mathcal{B}(\mathbf{x}) - \mathbf{P}_{\mathbf{A}_\lambda}\mathbf{x}\|_2 \le \epsilon'\|\mathbf{x}\|_2 = \epsilon/(c_2 q^2\sqrt{\kappa_\lambda})$. Applying the union bound over $q+1$ calls to $\mathcal{A}$ and a single call to $\mathcal{B}$, these bounds hold on every call with probability $\ge 1-\delta$. Adjusting constants on ϵ (via c_1 and c_2) proves the theorem. □

5. Approximating the matrix step function

We now return to proving our underlying result on iterative polynomial approximation of the matrix step function:

Lemma 3.1 (Step function algorithm)**.** *Let $\mathbf{S} \in \mathbb{R}^{d\times d}$ be symmetric with every eigenvalue σ satisfying $\sigma \in [0,1]$ and $|\sigma - 1/2| \ge \gamma$. Let $\mathcal{A}$ denote a procedure that on $\mathbf{x} \in \mathbb{R}^d$ produces $\mathcal{A}(\mathbf{x})$ with $\|\mathcal{A}(\mathbf{x}) - \mathbf{S}\mathbf{x}\| = O(\epsilon^2\gamma^2)\|\mathbf{x}\|_2$. Given $\mathbf{y} \in \mathbb{R}^d$ set $\mathbf{s}_0 := \mathcal{A}(\mathbf{y})$, $\mathbf{w}_0 := \mathbf{s}_0 - \frac{1}{2}\mathbf{y}$, and for $k \ge 0$ set*

$$\mathbf{w}_{k+1} := 4\left(\frac{2k+1}{2k+2}\right)\mathcal{A}(\mathbf{w}_k - \mathcal{A}(\mathbf{w}_k))$$

and $\mathbf{s}_{k+1} := \mathbf{s}_k + \mathbf{w}_{k+1}$. If all arithmetic operations are performed with $\Omega(\log(d/\epsilon\gamma))$ bits of precision and if $q = \Theta(\gamma^{-2}\log(1/\epsilon))$ then $\|s_q - s(\mathbf{S})\mathbf{y}\|_2 = O(\epsilon)\|\mathbf{y}\|_2$.

The derivation and proof of Lemma 3.1 is split into 3 parts. In Section 5.1 we derive a simple low degree polynomial approximation to the sign function:

$$\text{sgn}(x) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$$

In Section 5.2 we show how this polynomial can be computed with a stable iterative procedure. In Section 5.3 we use these pieces and the fact that the step function is simply a shifted and scaled sign function to prove Lemma 3.1. Along the way we give complementary views of Lemma 3.1 and show that there exist more efficient polynomial approximations. All proofs are in Appendix A.

5.1. Polynomial approximation to the sign function

We show that for sufficiently large k, the following polynomial is uniformly close to $\text{sgn}(x)$ on $[-1,1]$:

$$p_k(x) \stackrel{\text{def}}{=} \sum_{i=0}^{k} x(1-x^2)^i \prod_{j=1}^{i}\frac{2j-1}{2j}.$$

$p_k(x)$ can be derived in several ways. One follows from observing that $\text{sgn}(x)$ is odd and so $\text{sgn}(x)/x = 1/|x|$ is even. So, a good polynomial approximation for $\text{sgn}(x)$ should be odd and, when divided by x, should be even (i.e. a function of x^2). Specifically, given an approximation $q(x)$ to $1/\sqrt{x}$ on $(0,1]$ we can approximate $\text{sgn}(x)$ using $xq(x^2)$. Letting q be the k-th order Taylor approximation to $1/\sqrt{x}$ at $x=1$ yields $p_k(x)$. We first show:

Lemma 5.1. $\text{sgn}(x) = \lim_{k\to\infty} p_k(x)$ *for all* $x \in [-1,1]$.

Proof. Let $f(x) = x^{-1/2}$. By induction on k it is straightforward to show that the k-th derivative of f at $x > 0$ is $f^{(k)}(x) = (-1)^k x^{-\frac{1+2k}{2}}\prod_{i=1}^{k}\frac{2i-1}{2}$. Since $(-1)^i(x-1)^i = (1-x)^i$, the degree k Taylor approximation to $f(x)$ at $x=1$ is $q_k(x) = \sum_{i=0}^{k}(1-x)^i\prod_{j=1}^{i}\frac{2j-1}{2j}$. The integral form of Taylor series remainder gives $|f(x) - q_k(x)| = \left|\int_1^x \frac{f^{(k+1)}(t)}{k!}(x-t)^k dt\right|$ which is bounded by:

$$O\left(\int_1^x \frac{k}{t^{3/2}}\left(\frac{t-x}{t}\right)^k \prod_{j=1}^{k}\frac{2j-1}{2j}dt\right)$$

and consequently, the Taylor approximation converges, i.e. $\lim_{k\to\infty} q_k(x) = 1/\sqrt{x}$, for $x \in (0,1]$. Since $p_k(x) = x \cdot q_k(x^2)$, $\lim_{k\to\infty} p_k(x) = x/\sqrt{x^2} = \operatorname{sgn}(x)$ for $x \neq 0$ in $[-1,1]$. Since $p_k(0) = 0 = \operatorname{sgn}(0)$, the result follows. □

Alternatively, to derive $p_k(x)$ we can consider $(1-x^2)^k$, which is relatively large near 0 and small on the rest of $[-1,1]$. Integrating this function from 0 to x and normalizing yields a good step function. In Appendix A we prove:

Lemma 5.2. *For all $x \in \mathbb{R}$, $p_k(x) = \frac{\int_0^x (1-y^2)^k dy}{\int_0^1 (1-y^2)^k dy}$.*

Next, we bound the rate of convergence of $p_k(x)$ to $\operatorname{sgn}(x)$:

Lemma 5.3. *For $k \geq 1$ if $x \in (0,1]$ then $p_k(x) > 0$ and*

$$\operatorname{sgn}(x) - (x\sqrt{k})^{-1} e^{-kx^2} \leq p_k(x) \leq \operatorname{sgn}(x) \,. \quad (2)$$

If $x \in [-1,0)$ then $p_k(x) < 0$ and

$$\operatorname{sgn}(x) \leq p_k(x) \leq \operatorname{sgn}(x) + (x\sqrt{k})^{-1} e^{-kx^2} \,.$$

Proof. The claim is trivial for $x = 0$. Since $p_k(x)$ is odd it suffices to consider $x \in (0,1]$. It is direct that $p_k(x) > 0$, and $p_k(x) \leq \operatorname{sgn}(x)$ follows since $p_k(x)$ increases monotonically with k and $\lim_{k\to\infty} p_k(x) = \operatorname{sgn}(x)$ by Lemma 5.1. All that remains is the left-side inequality of (2). Using the Taylor series expansion of Lemma 5.1,

$$\operatorname{sgn}(x) - p_k(x) \leq x(1-x^2)^k \sum_{i=0}^{\infty} \left((1-x^2)^i \prod_{j=1}^{k} \frac{2j-1}{2j} \right) .$$

Now since $1 + x \leq e^x$ for all x and $\sum_{i=1}^{n} \frac{1}{i} \geq \ln n$,

$$\prod_{j=1}^{k} \frac{2j-1}{2j} \leq \exp\left(\sum_{j=1}^{k} \frac{-1}{2j} \right) \leq \exp\left(\frac{-\ln k}{2} \right) = \frac{1}{\sqrt{k}} .$$

Combining with $\sum_{i=0}^{\infty} (1-x^2)^i = x^{-2}$ and again that $1 + x \leq e^x$ proves the left hand side of (2). □

The lemma directly implies that $p_k(x)$ is a high quality approximation to $\operatorname{sgn}(x)$ for x bounded away from 0.

Corollary 5.4. *If $x \in [-1,1]$, with $|x| \geq \alpha > 0$ and $k = \alpha^{-2} \ln(1/\epsilon)$, then $|\operatorname{sgn}(x) - p_k(x)| \leq \epsilon$.*

We conclude by noting that this proof in fact implies the existence of a lower-degree polynomial approximation to $\operatorname{sgn}(x)$. Since the sum of coefficients in our expansion is small, we can replace each $(1-x^2)^q$ with Chebyshev polynomials of lower degree. In Appendix A, we prove:

Lemma 5.5. *There exists an $O(\alpha^{-1} \log(1/\alpha\epsilon))$ degree polynomial $q(x)$ such that $|\operatorname{sgn}(x) - q(x)| \leq \epsilon$ for all $x \in [-1,1]$ with $|x| \geq \alpha > 0$.*

Lemma 5.5 achieves, up to an additive $\log(1/\alpha)/\alpha$, the optimal trade off between degree and approximation of $\operatorname{sgn}(x)$ (Eremenko & Yuditskii, 2007). We have preliminary progress toward making this near-optimal polynomial algorithmic, a topic we leave to explore in future work.

5.2. Stable iterative algorithm for the sign function

We now provide an iterative algorithm for computing $p_k(x)$ that works when applied with limited precision. Our formula is obtained by considering each term of $p_k(x)$. Let

$$t_k(x) \stackrel{\text{def}}{=} x(1-x^2)^k \prod_{j=1}^{k} \frac{2j-1}{2j} .$$

Clearly $t_{k+1}(x) = t_k(x)(1-x^2)(2k+1)/(2k+2)$ and therefore we can compute the t_k iteratively. Since $p_k(x) = \sum_{i=0}^{k} t_i(x)$ we can compute $p_k(x)$ iteratively as well. We show this procedure works when applied to matrices, even if all operations are performed with limited precision:

Lemma 5.6. *Let $\mathbf{B} \in \mathbb{R}^{d\times d}$ be symmetric with $\|\mathbf{B}\|_2 \leq 1$. Let $\mathcal{C}$ be a procedure that given $\mathbf{x} \in \mathbb{R}^d$ produces $\mathcal{C}(\mathbf{x})$ with $\|\mathcal{C}(\mathbf{x}) - (\mathbf{I} - \mathbf{B}^2)\mathbf{x}\|_2 \leq \epsilon\|\mathbf{x}\|_2$. Given $\mathbf{y} \in \mathbb{R}^d$ suppose that we have $\mathbf{t}_0$ and $\mathbf{p}_0$ such that $\|\mathbf{t}_0 - \mathbf{B}\mathbf{y}\|_2 \leq \epsilon\|\mathbf{y}\|_2$ and $\|\mathbf{p}_0 - \mathbf{B}\mathbf{y}\|_2 \leq \epsilon\|\mathbf{y}\|_2$. For all $k \geq 1$ set*

$$\mathbf{t}_{k+1} := \left(\frac{2k+1}{2k+2} \right) \mathcal{C}(\mathbf{t}_k) \quad \text{and} \quad \mathbf{p}_{k+1} := \mathbf{p}_k + \mathbf{t}_{k+1} .$$

Then if operations are carried out with $\Omega(\log(d/\epsilon))$ bits of precision, for $1 \leq k \leq 1/(7\epsilon)$: $\|t_k(\mathbf{B})\mathbf{y} - \mathbf{t}_k\|_2 \leq 7k\epsilon\|\mathbf{y}\|_2$ and $\|p_k(\mathbf{B})\mathbf{y} - \mathbf{p}_k\|_2 \leq 7k\epsilon\|\mathbf{y}\|_2$.

Proof. Let $\mathbf{t}_k^* \stackrel{\text{def}}{=} t_k(\mathbf{B})\mathbf{y}$, $\mathbf{p}_k^* \stackrel{\text{def}}{=} p_k(\mathbf{B})\mathbf{y}$, and $\mathbf{C} \stackrel{\text{def}}{=} \mathbf{I} - \mathbf{B}^2$. Since $\mathbf{p}_0^* = \mathbf{t}_0^* = \mathbf{B}\mathbf{y}$ and $\|\mathbf{B}\|_2 \leq 1$ we see that even if $\mathbf{t}_0$ and $\mathbf{p}_0$ are truncated to the given bit precision we still have $\|\mathbf{t}_0 - \mathbf{t}_0^*\|_2 \leq \epsilon\|\mathbf{y}\|_2$ and $\|\mathbf{p}_0 - \mathbf{p}_0^*\|_2 \leq \epsilon\|\mathbf{y}\|_2$.

Now suppose that $\|\mathbf{t}_k - \mathbf{t}_k^*\|_2 \leq \alpha\|\mathbf{y}\|_2$ for some $\alpha \leq 1$. Since $|t_k(x)| \leq |p_k(x)| \leq |\operatorname{sgn}(x)| \leq 1$ for $x \in [-1,1]$ and $-\mathbf{I} \preceq \mathbf{B} \preceq \mathbf{I}$ we know that $\|\mathbf{t}_k^*\|_2 \leq \|\mathbf{y}\|_2$ and by reverse triangle inequality $\|\mathbf{t}_k\|_2 \leq (1+\alpha)\|\mathbf{y}\|_2$. Using our assumption on $\mathcal{C}$ and applying triangle inequality yields

$$\begin{aligned} \|\mathcal{C}(\mathbf{t}_k) - \mathbf{C}\mathbf{t}_k^*\|_2 &\leq \|\mathcal{C}(\mathbf{t}_k) - \mathbf{C}\mathbf{t}_k\|_2 + \|\mathbf{C}(\mathbf{t}_k - \mathbf{t}_k^*)\|_2 \\ &\leq \epsilon\|\mathbf{t}_k\|_2 + \|\mathbf{C}\|_2 \cdot \|(\mathbf{t}_k - \mathbf{t}_k^*)\|_2 \\ &\leq (\epsilon(1+\alpha) + \alpha)\|\mathbf{y}\|_2 \leq (2\epsilon + \alpha)\|\mathbf{y}\|_2 . \end{aligned}$$

In the last line we used $\|\mathbf{C}\|_2 \leq 1$ since $\mathbf{0} \preceq \mathbf{B}^2 \preceq \mathbf{I}$. Again, this gives $\|\mathbf{C}\mathbf{t}_k^*\|_2 \leq \|\mathbf{y}\|_2$ and by reverse triangle inequality $\|\mathcal{C}(\mathbf{t}_k)\|_2 \leq (1 + 2\epsilon + \alpha)\|\mathbf{y}\|_2$. Using $\mathcal{C}(\mathbf{t}_k)$ to compute $\mathbf{t}_{k+1}$ with bounded precision will then introduce an additional additive error of $\epsilon(1+2\epsilon+\alpha)\|\mathbf{y}\|_2 \leq 4\epsilon\|\mathbf{y}\|_2$. Putting all this together, $\|\mathbf{t}_k^* - \mathbf{t}_k\|_2$ grows by at most an

additive $6\epsilon\|\mathbf{y}\|_2$ every time k increases and by the same argument so does $\|\mathbf{p}_k - \mathbf{p}_k^*\|_2$. Including our initial error of $\epsilon\|\mathbf{y}\|_2$ on $\mathbf{t}_0$ and $\mathbf{p}_0$, we conclude that $\|\mathbf{t}_k^* - \mathbf{t}_k\|_2$ and $\|\mathbf{p}_k - \mathbf{p}_k^*\|_2$ are both bounded by $(6k\epsilon + \epsilon)\|\mathbf{y}\|_2 \leq 7k\epsilon\|\mathbf{y}\|_2$. □

5.3. Approximating the step function

We finally apply the results of Section 5.1 and Section 5.2 to approximate the step function and prove Lemma 3.1. We simply use $s(x) = \frac{1}{2}(1 + \text{sgn}(2x - 1))$, first showing how to compute $\frac{1}{2}(1 + p_k(2x - 1))$ via Lemma 5.6.

Lemma 5.7. *Let $\mathbf{S} \in \mathbb{R}^{d\times d}$ be symmetric with $\mathbf{0} \preceq \mathbf{S} \preceq \mathbf{I}$. Let $\mathcal{A}$ be a procedure that on $\mathbf{x} \in \mathbb{R}^d$ produces $\mathcal{A}(\mathbf{x})$ with $\|\mathcal{A}(\mathbf{x}) - \mathbf{S}\mathbf{x}\|_2 \leq \epsilon\|\mathbf{x}\|_2$. Given arbitrary $\mathbf{y} \in \mathbb{R}^d$ set $\mathbf{s}_0 := \mathcal{A}(\mathbf{y})$, $\mathbf{w}_0 := \mathbf{s}_0 - \frac{1}{2}\mathbf{y}$, and for all $k \geq 0$ set*

$$\mathbf{w}_{k+1} := 4\left(\frac{2k+1}{2k+2}\right)\mathcal{A}(\mathbf{w}_k - \mathcal{A}(\mathbf{w}_k))$$

and $\mathbf{s}_{k+1} := \mathbf{s}_k + \mathbf{w}_{k+1}$. If arithmetic operations are performed with $\Omega(\log(d/\epsilon))$ bits of precision and $k = O(1/\epsilon)$ then $\|\frac{1}{2}(\mathbf{I} + p_k(2\mathbf{S} - \mathbf{I}))\mathbf{y} - \mathbf{s}_k\|_2 = O(k\epsilon)\|\mathbf{y}\|_2$.

Proof. Since $\mathbf{M} \stackrel{\text{def}}{=} \mathbf{I} - (2\mathbf{S} - \mathbf{I})^2 = 4\mathbf{S}(\mathbf{I} - \mathbf{S})$, $\mathbf{w}_k$ is the same as $\frac{1}{2}\mathbf{t}_k$ in Lemma 5.6 with $\mathbf{B} = 2\mathbf{S} - \mathbf{I}$ and $\mathcal{C}(\mathbf{x}) = 4\mathcal{A}(\mathbf{x} - \mathcal{A}(\mathbf{x}))$, and $\mathbf{s}_k = \mathbf{s}_0 + \sum_{i=1}^k \mathbf{w}_i = \sum_{i=0}^k \frac{1}{2}\mathbf{t}_i + \frac{1}{2}\mathbf{y} = \frac{1}{2}(\mathbf{p}_k + \mathbf{y})$. Multiplying by $1/2$ everywhere does not increase error and $\|2\mathbf{S} - \mathbf{I}\|_2 \leq 1$ so we can invoke Lemma 5.6 to yield the result provided we can show $\|4\mathcal{A}(\mathbf{x} - \mathcal{A}(\mathbf{x})) - \mathbf{M}\mathbf{x}\|_2 = O(\epsilon)\|\mathbf{x}\|_2$. Computing $\mathcal{A}(\mathbf{x})$ and subtracting from $\mathbf{x}$ introduces at most additive error $2\epsilon\|\mathbf{x}\|_2$ Consequently by the error guarantee of $\mathcal{A}$, $\|4\mathcal{A}(\mathcal{A}(\mathbf{x}) - \mathbf{x}) - \mathbf{M}\mathbf{x}\|_2 = O(\epsilon)\|\mathbf{x}\|_2$ as desired. □

Using Lemma 5.7 and Corollary 5.4 we finally have:

Proof of Lemma 3.1. By assumption, $\mathbf{0} \preceq \mathbf{S} \preceq \mathbf{I}$ and $\epsilon\gamma^2 q = O(1)$. Invoking Lemma 5.7 with error $\epsilon' = \epsilon^2\gamma^2$, letting $\mathbf{a}_q \stackrel{\text{def}}{=} \frac{1}{2}(\mathbf{I} + p_q(2\mathbf{S} - \mathbf{I}))\mathbf{y}$ we have

$$\|\mathbf{a}_q - \mathbf{s}_q\|_2 = O(\gamma^2\epsilon^2 q)\|\mathbf{y}\|_2 = O(\epsilon)\|\mathbf{y}\|_2\,. \qquad (3)$$

Now, since $s(\mathbf{S}) = \frac{1}{2}(\mathbf{I} + \text{sgn}(2\mathbf{S} - \mathbf{I}))$ and every eigenvalue of $2\mathbf{S} - \mathbf{I}$ is in $[\gamma, 1]$, by assumption on $\mathbf{S}$ we can invoke Corollary 5.4 yielding $\|\mathbf{a}_q - s(\mathbf{S})\mathbf{y}\|_2 \leq \frac{1}{2}\|p_q(2\mathbf{S} - \mathbf{I}) - \text{sgn}(2\mathbf{S} - \mathbf{I})\|_2\|\mathbf{y}\|_2 \leq 2\epsilon\|\mathbf{y}\|_2$. The result follows from combining with (3) via triangle inequality. □

6. Empirical evaluation

We conclude with an empirical evaluation of PC-PROC and RIDGE-PCR (Algorithms 1 and 2). Since PCR has already been justified as a statistical technique, we focus on showing that, with few iterations, our algorithm recovers an accurate approximation to $\mathbf{A}_\lambda^\dagger\mathbf{b}$ and $\mathbf{P}_{\mathbf{A}_\lambda}\mathbf{y}$.

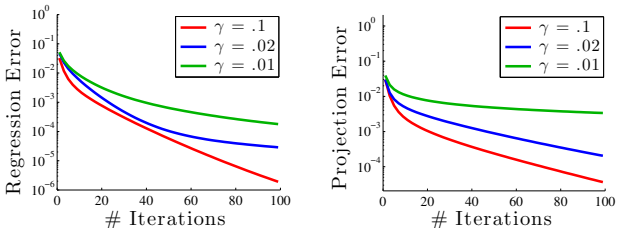

Figure 2: Relative log error of RIDGE-PCR (left) and PC-PROJ (right) for synthetically generated data.

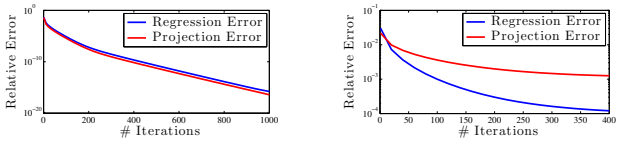

Figure 3: Extended log error plot on synthetic data with gap $\gamma = .1$ (left). Relative log error of RIDGE-PCR and PC-PROJ for an MNIST-based regression problem (right).

We begin with synthetic data, which lets us control the spectral gap γ that dominates our iteration bounds (see Theorem 3.2). Data is generated randomly by drawing top singular values uniformly from the range $[.5(1+\gamma), 1]$ and tail singular values from $[0, .5(1-\gamma)]$. λ is set to .5 and $\mathbf{A}$ (500 rows, 200 columns) is formed via the SVD $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T}$ where $\mathbf{U}$ and $\mathbf{V}$ are random bases and $\mathbf{\Sigma}$ contains our random singular values. To model a typical application, $\mathbf{b}$ is generated by adding noise to the response $\mathbf{A}\mathbf{x}$ of a random $\mathbf{x}$ that correlates with $\mathbf{A}$'s top principal components.

As apparent in Figure 2, our algorithm performs very well for regression, even for small γ. Error is measured via the natural $\mathbf{A}^\mathsf{T}\mathbf{A}$-norm and we plot $\|\text{RIDGE-PCR}(\mathbf{A}, \mathbf{b}, \lambda) - \mathbf{A}_\lambda^\dagger\mathbf{b}\|_{\mathbf{A}^\mathsf{T}\mathbf{A}}^2 / \|\mathbf{A}_\lambda^\dagger\mathbf{b}\|_{\mathbf{A}^\mathsf{T}\mathbf{A}}^2$. The figure shows similar convergence for projection (given with respect to the more natural 2-norm), although we do notice a stronger effect of a small gap γ in this case. Both plots confirm the linear convergence predicted by our analysis (Theorems 3.2 and 4.2). To illustrate stability, we include an extended plot for the $\gamma = .1$ data which shows arbitrarily high accuracy as iterations increase (Figure 3).

Finally, we consider a 60K-point regression problem constructed from MNIST classification data (LeCun et al., 2015), with the goal of distinguishing handwritten digits {1,2,4,5,7} from the rest. Input is normalized and 1000 random Fourier features are generated according to a unit RBF kernel (Rahimi & Recht, 2007). Our final data set is both of larger scale and condition number than the original. The MNIST principal component regression was run with $\lambda = .01\sigma_1^2$. Although the gap γ is very small around this cutoff point (just .006), we see fast convergence for PCR. Convergence for projection is slowed more notably by the gap, but it is still possible to obtain 0.01 relative error with only 20 iterations (i.e. invocations of ridge regression).

References

Boutsidis, Christos and Magdon-Ismail, Malik. Faster SVD-truncated regularized least-squares. In *Proceedings of the 2014 IEEE International Symposium on Information Theory (ISIT)*, pp. 1321–1325, 2014.

Chan, Tony F. and Hansen, Per Christian. Computing truncated singular value decomposition least squares solutions by rank revealing QR-factorizations. *SIAM Journal on Scientific and Statistical Computing*, 11(3):519–530, 1990.

Cohen, Michael B., Lee, Yin Tat, Musco, Cameron, Musco, Christopher, Peng, Richard, and Sidford, Aaron. Uniform sampling for matrix approximation. In *Proceedings of the 6th Conference on Innovations in Theoretical Computer Science (ITCS)*, pp. 181–190, 2015.

Dhillon, Paramveer S., Foster, Dean P., Kakade, Sham M., and Ungar, Lyle H. A risk comparison of ordinary least squares vs ridge regression. *The Journal of Machine Learning Research*, 14(1):1505–1511, 2013.

Eremenko, Alexandre and Yuditskii, Peter. Uniform approximation of $sgn(x)$ by polynomials and entire functions. *Journal d'Analyse Mathmatique*, 101(1):313–324, 2007.

Eremenko, Alexandre and Yuditskii, Peter. Polynomials of the best uniform approximation to $sgn(x)$ on two intervals. *Journal d'Analyse Mathématique*, 114(1):285–315, 2011.

Frank, Ildiko E. and Friedman, Jerome H. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

Frommer, Andreas and Simoncini, Valeria. *Model Order Reduction: Theory, Research Aspects and Applications*, chapter Matrix Functions, pp. 275–303. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

Frostig, Roy, Ge, Rong, Kakade, Sham M., and Sidford, Aaron. Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

Hansen, Per Christian. The truncated SVD as a method for regularization. *BIT Numerical Mathematics*, 27(4): 534–553, 1987.

Higham, Nicholas J. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, 2008.

Hotelling, Harold. The relations of the newer multivariate statistical methods to factor analysis. *British Journal of Statistical Psychology*, 10(2):69–79, 1957.

LeCun, Yann, Cortes, Corinna, and Burges, Christopher J.C. MNIST handwritten digit database. 2015. URL http://yann.lecun.com/exdb/mnist/.

Lin, Qihang, Lu, Zhaosong, and Xiao, Lin. An accelerated proximal coordinate gradient method. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pp. 3059–3067, 2014.

Musco, Cameron and Musco, Christopher. Randomized block krylov methods for stronger and faster approximate singular value decomposition. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pp. 1396–1404, 2015.

Nelson, Jelani and Nguyên, Huy L. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 117–126, 2013.

Nesterov, Yurii. A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$. In *Soviet Mathematics Doklady*, volume 27, pp. 372–376, 1983.

Rahimi, Ali and Recht, Benjamin. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20 (NIPS)*, pp. 1177–1184. 2007.

Sachdeva, Sushant and Vishnoi, Nisheeth K. Faster algorithms via approximation theory. *Foundations and Trends in Theoretical Computer Science*, 9(2):125–210, 2014.

Shalev-Shwartz, Shai and Zhang, Tong. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, pp. 1–41, 2014.

Tikhonov, Andrey. Solution of incorrectly formulated problems and the regularization method. In *Soviet Mathematics Doklady*, volume 4, pp. 1035–1038, 1963.

van den Eshof, Jasper, Frommer, Andreas, Lippert, Thomas, Schilling, Klaus, and van der Vorst, Henk A. Numerical methods for the QCD overlap operator I: Sign-function and error bounds. *Computer physics communications*, 146(2):203–224, 2002.