PCA Denoising Layer

Anonymous Author(s)
Affiliation
Address
email

Abstract

1 PCA Denoising Layer

Options to implement PCA Norm Layer.

- Given input covariance matrix $\mathbf{M}$, use **Eigen Decomposition** or **Singular Value Decomposition (SVD)** Operation as forward computation, and use the analytic solution of its gradient for backward propogation.
- Given input covariance matrix $\mathbf{M}$, vectors with random values $[\mathbf{v}_1^1, \mathbf{v}_2^1, ...]$, use **Power Iteration** as forward computation, and use its gradient for backward propogation.
- Given input covariance matrix $\mathbf{M}$, vectors with random values $[\mathbf{v}_1^1, \mathbf{v}_2^1, ...]$, use **Eigen Decomposition** or **SVD** Operation as forward computation, and use **Power Iteration** to approximate the analytic solutions of the gradient for backward propogation.

Usually, people choose either option 1 or option 2 to implement PCA Norm Layer, but both of them have problems. In option 1, **Eigen Decomposition** or **Singular Value Decomposition (SVD)**, the analytic solutions of the gradient sometimes causes NaN problem when there are two or more eigenvalues are too close to each other. In option 2, if the two eigenvalues are very close, eigenvectors could not be computed precisely with limited power iteration number. Thus, during backprorogation, the derivatives will be very inaccurate and destroy the parameters of model, and cause numerical instability in the training process.

In this paper, we propose to using option 3. During forward pass, we use **SVD** to compute the eigenvalues. SVD is numerically more stable than eigendecomposition [1] as SVD implementation employs a divide-and-conquer strategy, while the eigendecomposition uses QR algorithm. During backpropogation, we employ **Power Iteration** method to compute the numerical solutions of the covariance matrix $\mathbf{M}$ gradient. In **sections** 2.1 & 2.3, we will prove that when the iteration number goes to infinite, the accumulated gradients (*i.e.* numerical solution) from the **Power Iteration** method is exactly the same with the analytic solution of the gradient.

2 Approximate SVD gradient with Power Iteration in backpropogation

In the following 2 subsections, we will prove that when the gradient computed from Power Iteration equals to the gradients computed from SVD.

2.1 Gradient of Power Iteration

To compute the leading eigenvector $\mathbf{v}$ of $\mathbf{M}$, Power Iteration uses the following standard formula,

$$\mathbf{v}^{(k)} = \frac{\mathbf{M}\mathbf{v}^{(k-1)}}{\|\mathbf{M}\mathbf{v}^{(k-1)}\|}, \tag{1}$$

in which $\|\cdot\|$ denotes the l_2 norm, and $v^{(0)}$ is usually initialized randomly with $\|v^{(0)}\|$=1. Its gradient formula is as follows [2],

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{M}} &= \sum_k \frac{\left(\mathbf{I} - \mathbf{v}^{(k+1)}\mathbf{v}^{(k+1)\top}\right)}{\left\|\mathbf{M}\mathbf{v}^{(k)}\right\|} \frac{\partial L}{\partial \mathbf{v}^{(k+1)}} \mathbf{v}^{(k)\top} \\ \frac{\partial L}{\partial \mathbf{v}^{(k)}} &= \mathbf{M}\frac{\left(\mathbf{I} - \mathbf{v}^{(k+1)}\mathbf{v}^{(k+1)\top}\right)}{\left\|\mathbf{M}\mathbf{v}^{(k)}\right\|} \frac{\partial L}{\partial \mathbf{v}^{(k+1)}}\end{aligned} \tag{2}$$

Using 3 power iteration steps for demonstration.

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{v}^{(2)}} &= \mathbf{M}\frac{\left(\mathbf{I} - \mathbf{v}^{(3)}\mathbf{v}^{(3)\top}\right)}{\left\|\mathbf{M}\mathbf{v}^{(2)}\right\|} \frac{\partial L}{\partial \mathbf{v}^{(3)}} \\ \frac{\partial L}{\partial \mathbf{v}^{(1)}} &= \mathbf{M}\frac{\left(\mathbf{I} - \mathbf{v}^{(2)}\mathbf{v}^{(2)\top}\right)}{\left\|\mathbf{M}\mathbf{v}^{(1)}\right\|} \frac{\partial L}{\partial \mathbf{v}^{(2)}} = \mathbf{M}\frac{\left(\mathbf{I} - \mathbf{v}^{(2)}\mathbf{v}^{(2)\top}\right)}{\left\|\mathbf{M}\mathbf{v}^{(1)}\right\|}\mathbf{M}\frac{\left(\mathbf{I} - \mathbf{v}^{(3)}\mathbf{v}^{(3)\top}\right)}{\left\|\mathbf{M}\mathbf{v}^{(2)}\right\|} \frac{\partial L}{\partial \mathbf{v}^{(3)}}\end{aligned} \tag{3}$$

Then the $\frac{\partial L}{\partial \mathbf{M}}$ should be like the following, for the reason that we use eigenvalue decomposition (ED)'s result, denoted as $\mathbf{v}$ as initial value, then $\mathbf{v}$=$\mathbf{v}^{(0)}\approx\mathbf{v}^{(1)}\approx\mathbf{v}^{(2)}\approx\cdots\approx\mathbf{v}^{(k)}$.

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{M}} =& \frac{\left(\mathbf{I} - \mathbf{v}^{(3)}\mathbf{v}^{(3)\top}\right)}{\left\|\mathbf{M}\mathbf{v}^{(2)}\right\|} \frac{\partial L}{\partial \mathbf{v}^{(3)}}\mathbf{v}^{(2)\top} + \frac{\left(\mathbf{I} - \mathbf{v}^{(2)}\mathbf{v}^{(2)\top}\right)}{\left\|\mathbf{M}\mathbf{v}^{(1)}\right\|} \frac{\partial L}{\partial \mathbf{v}^{(2)}}\mathbf{v}^{(1)\top} + \frac{\left(\mathbf{I} - \mathbf{v}^{(1)}\mathbf{v}^{(1)\top}\right)}{\left\|\mathbf{M}\mathbf{v}^{(0)}\right\|} \frac{\partial L}{\partial \mathbf{v}^{(1)}}\mathbf{v}^{(0)\top} \\ =& \left(\frac{\left(\mathbf{I}-\mathbf{v}\mathbf{v}^\top\right)}{\|\mathbf{M}\mathbf{v}\|} + \frac{\left(\mathbf{I}-\mathbf{v}\mathbf{v}^\top\right)\mathbf{M}\left(\mathbf{I}-\mathbf{v}\mathbf{v}^\top\right)}{\|\mathbf{M}\mathbf{v}\|^2} + \frac{\left(\mathbf{I}-\mathbf{v}\mathbf{v}^\top\right)\mathbf{M}\left(\mathbf{I}-\mathbf{v}\mathbf{v}^\top\right)\mathbf{M}\left(\mathbf{I}-\mathbf{v}\mathbf{v}^\top\right)}{\|\mathbf{M}\mathbf{v}\|^3}\right)\frac{\partial L}{\partial \mathbf{v}^{(3)}}\mathbf{v}^\top\end{aligned} \tag{4}$$

Known that $\mathbf{v}\mathbf{v}^\top$ and $\mathbf{M}$ are symmetric and $\mathbf{M}\mathbf{v} = \lambda\mathbf{v}$, we have

$$\mathbf{v}\mathbf{v}^\top\mathbf{M} = (\mathbf{M}^\top\mathbf{v}\mathbf{v}^\top)^\top = (\mathbf{M}\mathbf{v}\mathbf{v}^\top)^\top = (\lambda\mathbf{v}\mathbf{v}^\top)^\top = \lambda\mathbf{v}\mathbf{v}^\top = \mathbf{M}\mathbf{v}\mathbf{v}^\top.$$

Introducing the equation above into the numerator in the second term of Eq.4, we can obtain:

$$\begin{aligned}&\left(\mathbf{I} - \mathbf{v}\mathbf{v}^\top\right)\mathbf{M}\left(\mathbf{I} - \mathbf{v}\mathbf{v}^\top\right) = \left(\mathbf{M} - \mathbf{v}\mathbf{v}^\top\mathbf{M}\right)\left(\mathbf{I} - \mathbf{v}\mathbf{v}^\top\right) = \left(\mathbf{M} - \mathbf{M}\mathbf{v}\mathbf{v}^\top\right)\left(\mathbf{I} - \mathbf{v}\mathbf{v}^\top\right) \\ &= \mathbf{M}\left(\mathbf{I} - \mathbf{v}\mathbf{v}^\top\right)\left(\mathbf{I} - \mathbf{v}\mathbf{v}^\top\right) = \mathbf{M}\left(\mathbf{I} - 2\mathbf{v}\mathbf{v}^\top + \mathbf{v}\cancel{\left(\mathbf{v}^\top\mathbf{v}\right)}\mathbf{v}^\top\right) = \mathbf{M}\left(\mathbf{I} - \mathbf{v}\mathbf{v}^\top\right).\end{aligned} \tag{5}$$

Similarly, for the numerator in the third term in Eq.4, we have:

$$\left(\mathbf{I} - \mathbf{v}\mathbf{v}^\top\right)\mathbf{M}\left(\mathbf{I} - \mathbf{v}\mathbf{v}^\top\right)\mathbf{M}\left(\mathbf{I} - \mathbf{v}\mathbf{v}^\top\right) = \mathbf{M}\mathbf{M}\left(\mathbf{I} - \mathbf{v}\mathbf{v}^\top\right). \tag{6}$$

Introducing Eq.5 and Eq.6 into Eq.4, we can obtain

$$\frac{\partial L}{\partial \mathbf{M}} = \left(\frac{\left(\mathbf{I} - \mathbf{v}\mathbf{v}^\top\right)}{\|\mathbf{M}\mathbf{v}\|} + \frac{\mathbf{M}\left(\mathbf{I} - \mathbf{v}\mathbf{v}^\top\right)}{\|\mathbf{M}\mathbf{v}\|^2} + \frac{\mathbf{M}\mathbf{M}\left(\mathbf{I} - \mathbf{v}\mathbf{v}^\top\right)}{\|\mathbf{M}\mathbf{v}\|^3}\right)\frac{\partial L}{\partial \mathbf{v}^{(3)}}\mathbf{v}^\top \tag{7}$$

Extending the iteration number from 3 to k, Eq.4 will be extended as

$$\frac{\partial L}{\partial \mathbf{M}} = \left(\frac{\left(\mathbf{I} - \mathbf{v}\mathbf{v}^\top\right)}{\|\mathbf{M}\mathbf{v}\|} + \frac{\mathbf{M}\left(\mathbf{I} - \mathbf{v}\mathbf{v}^\top\right)}{\|\mathbf{M}\mathbf{v}\|^2} + \cdots + \frac{\mathbf{M}^{k-1}\left(\mathbf{I} - \mathbf{v}\mathbf{v}^\top\right)}{\|\mathbf{M}\mathbf{v}\|^k}\right)\frac{\partial L}{\partial \mathbf{v}^{(k)}}\mathbf{v}^\top \tag{8}$$

Eq.8 is the form we adopt to compute the gradients of SVD, and we set k=19.

2.2 Proof of Gradient Equivalence Between Power Iteration and SVD

In this subsection, we are going to prove that the gradients of SVD and Power Iteration are equivalent. In the end of this section, we can observe that when the number of the iterations goes to infinity, the gradients of Power Iteration can be written as the same form as the one of SVD.

$$
\begin{aligned}
\|\mathbf{M}\mathbf{v}\| &= \lambda,\ \|\mathbf{M}\mathbf{v}\|^2 = \lambda^2, \cdots,\ \|\mathbf{M}\mathbf{v}\|^k = \lambda^k \\
\mathbf{M} &= \lambda_1 \mathbf{v}_1 \mathbf{v}_1^\top + \lambda_2 \mathbf{v}_2 \mathbf{v}_2^\top + \cdots + \lambda_n \mathbf{v}_n \mathbf{v}_n^\top \\
\mathbf{M}^2 &= (\lambda_1 \mathbf{v}_1 \mathbf{v}_1^\top + \lambda_2 \mathbf{v}_2 \mathbf{v}_2^\top + \cdots)(\lambda_1 \mathbf{v}_1 \mathbf{v}_1^\top + \lambda_2 \mathbf{v}_2 \mathbf{v}_2^\top + \cdots) \\
&= \lambda_1^2 \mathbf{v}_1 \mathbf{v}_1^\top \mathbf{v}_1 \mathbf{v}_1^\top + \lambda_2^2 \mathbf{v}_2 \mathbf{v}_2^\top \mathbf{v}_2 \mathbf{v}_2^\top + \cdots \cancel{\lambda_1 \lambda_2 \mathbf{v}_1 \mathbf{v}_1^\top \mathbf{v}_2 \mathbf{v}_2^\top} + \cancel{\lambda_1 \lambda_2 \mathbf{v}_2 \mathbf{v}_2^\top \mathbf{v}_1 \mathbf{v}_1^\top} + \cdots \\
&= \lambda_1^2 \mathbf{v}_1 \mathbf{v}_1^\top + \lambda_2^2 \mathbf{v}_2 \mathbf{v}_2^\top + \cdots + \lambda_n^2 \mathbf{v}_n \mathbf{v}_n^\top, \\
\mathbf{M}^k &= \lambda_1^k \mathbf{v}_1 \mathbf{v}_1^\top + \lambda_2^k \mathbf{v}_2 \mathbf{v}_2^\top + \cdots + \lambda_n^k \mathbf{v}_n \mathbf{v}_n^\top,
\end{aligned}
\tag{9}
$$

in which $\mathbf{v} = \mathbf{v}_1$ is the leading eigenvector and $\lambda = \lambda_1$ is the leading eigenvalue. By introducing Eq.9 into Eq.12, the derivative can be further formulated as

$$
\begin{aligned}
\frac{\partial L}{\partial \mathbf{M}} &= \left(\frac{\left(\mathbf{I} - \mathbf{v}_1 \mathbf{v}_1^\top\right)}{\|\mathbf{M}\mathbf{v}_1\|} + \frac{\mathbf{M}\left(\mathbf{I} - \mathbf{v}_1 \mathbf{v}_1^\top\right)}{\|\mathbf{M}\mathbf{v}_1\|^2} + \cdots + \frac{\mathbf{M}^{k-1}\left(\mathbf{I} - \mathbf{v}_1 \mathbf{v}_1^\top\right)}{\|\mathbf{M}\mathbf{v}_1\|^k} \right) \frac{\partial L}{\partial \mathbf{v}_1^{(k)}} \mathbf{v}_1^\top \\
&= \left(\frac{\left(\mathbf{I} - \mathbf{v}_1 \mathbf{v}_1^\top\right)}{\|\mathbf{M}\mathbf{v}_1\|} + \frac{\left(\mathbf{M} - \lambda \mathbf{v}_1 \mathbf{v}_1^\top\right)}{\|\mathbf{M}\mathbf{v}_1\|^2} + \cdots + \frac{\left(\mathbf{M}^{k-1} - \lambda^{k-1} \mathbf{v}_1 \mathbf{v}_1^\top\right)}{\|\mathbf{M}\mathbf{v}\|^k} \right) \frac{\partial L}{\partial \mathbf{v}_1^{(k)}} \mathbf{v}_1^\top \\
&= \left(\frac{\left(\sum_{i=2}^n \mathbf{v}_i \mathbf{v}_i^\top\right)}{\lambda_1} + \frac{\left(\sum_{i=2}^n \lambda_i \mathbf{v}_i \mathbf{v}_i^\top\right)}{\lambda_1^2} + \cdots + \frac{\left(\sum_{i=2}^n \lambda_i^{k-1} \mathbf{v}_i \mathbf{v}_i^\top\right)}{\lambda_1^k} \right) \frac{\partial L}{\partial \mathbf{v}_1^{(k)}} \mathbf{v}^\top \\
&= \left(\sum_{i=2}^n \left(\frac{1}{\lambda_1} + \frac{1}{\lambda_1} \left(\frac{\lambda_i}{\lambda_1}\right)^1 + \frac{1}{\lambda_1} \left(\frac{\lambda_i}{\lambda_1}\right)^2 + \cdots + \frac{1}{\lambda_1} \left(\frac{\lambda_i}{\lambda_1}\right)^{k-1} \right) \mathbf{v}_i \mathbf{v}_i^\top \right) \frac{\partial L}{\partial \mathbf{v}_1^{(k)}} \mathbf{v}_1^\top
\end{aligned}
\tag{10}
$$

In Eq.10, we have a geometric progression series. Given that

$$
1 - (\frac{\lambda_i}{\lambda_1})^k \to 1, \text{when } k \to \infty, |\frac{\lambda_i}{\lambda_1}| < 1,
$$

then we have

$$
\frac{1}{\lambda_1} + \frac{1}{\lambda_1} \left(\frac{\lambda_i}{\lambda_1}\right)^1 + \frac{1}{\lambda_1} \left(\frac{\lambda_i}{\lambda_1}\right)^2 + \cdots + \frac{1}{\lambda_1} \left(\frac{\lambda_i}{\lambda_1}\right)^{k-1} = \frac{\frac{1}{\lambda_1}(1 - (\frac{\lambda_i}{\lambda_1})^k)}{1 - \frac{\lambda_i}{\lambda_1}} \to \frac{\frac{1}{\lambda_1}}{1 - \frac{\lambda_i}{\lambda_1}}, \text{when } k \to \infty.
\tag{11}
$$

Introducing Eq.11 to Eq.10, we can obtain

$$
\frac{\partial L}{\partial \mathbf{M}} = \left(\sum_{i=2}^n \left(\frac{\frac{1}{\lambda_1}}{1 - \frac{\lambda_i}{\lambda_1}} \right) \mathbf{v}_i \mathbf{v}_i^\top \right) \frac{\partial L}{\partial \mathbf{v}_1^{(k)}} \mathbf{v}_1^\top = \left(\sum_{i=2}^n \frac{\mathbf{v}_i \mathbf{v}_i^\top}{\lambda_1 - \lambda_i} \right) \frac{\partial L}{\partial \mathbf{v}_1^{(k)}} \mathbf{v}_1^\top
\tag{12}
$$

2.3 Matrix Back-propagation

The analytic soltions of the gradients are from matrix back-propagation [3].

$$
\frac{\partial L}{\partial M} = V \left\{ \left(\tilde{K}^\top \circ \left(V^\top \frac{\partial L}{\partial V} \right) \right) + \left(\frac{\partial L}{\partial \Sigma} \right)_{diag} \right\} V^\top
\tag{13}
$$

$$
\tilde{K}_{ij} = \begin{cases} \frac{1}{\lambda_i - \lambda_j}, & i \neq j \\ 0, & i = j \end{cases}
\tag{14}
$$

$$\tilde{K} = \begin{bmatrix} 0 & \frac{1}{\lambda_1 - \lambda_2} & \frac{1}{\lambda_1 - \lambda_3} & \cdots & \frac{1}{\lambda_1 - \lambda_n} \\ \frac{1}{\lambda_2 - \lambda_1} & 0 & \frac{1}{\lambda_2 - \lambda_3} & \cdots & \frac{1}{\lambda_2 - \lambda_n} \\ \frac{1}{\lambda_3 - \lambda_1} & \frac{1}{\lambda_3 - \lambda_2} & 0 & \cdots & \frac{1}{\lambda_3 - \lambda_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\lambda_n - \lambda_1} & \frac{1}{\lambda_n - \lambda_2} & \frac{1}{\lambda_n - \lambda_3} & \cdots & 0 \end{bmatrix} \tag{15}$$

where λ_i is the eigen-value.

$$V = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \mathbf{v}_3 \quad \cdots \quad \mathbf{v}_n] \tag{16}$$

where v_i is the eigen-vector.

$$\frac{\partial L}{\partial V} = \begin{bmatrix} \frac{\partial L}{\partial \mathbf{v}_1} & \frac{\partial L}{\partial \mathbf{v}_2} & \frac{\partial L}{\partial \mathbf{v}_3} & \cdots & \frac{\partial L}{\partial \mathbf{v}_n} \end{bmatrix}^\top \tag{17}$$

$$V^\top \frac{\partial L}{\partial V} = \begin{bmatrix} \mathbf{v}_1^\top \frac{\partial L}{\partial \mathbf{v}_1} & \mathbf{v}_1^\top \frac{\partial L}{\partial \mathbf{v}_2} & \mathbf{v}_1^\top \frac{\partial L}{\partial \mathbf{v}_3} & \cdots & \mathbf{v}_1^\top \frac{\partial L}{\partial \mathbf{v}_n} \\ \mathbf{v}_2^\top \frac{\partial L}{\partial \mathbf{v}_1} & \mathbf{v}_2^\top \frac{\partial L}{\partial \mathbf{v}_2} & \mathbf{v}_2^\top \frac{\partial L}{\partial \mathbf{v}_3} & \cdots & \mathbf{v}_2^\top \frac{\partial L}{\partial \mathbf{v}_n} \\ \mathbf{v}_3^\top \frac{\partial L}{\partial \mathbf{v}_1} & \mathbf{v}_3^\top \frac{\partial L}{\partial \mathbf{v}_2} & \mathbf{v}_3^\top \frac{\partial L}{\partial \mathbf{v}_3} & \cdots & \mathbf{v}_3^\top \frac{\partial L}{\partial \mathbf{v}_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{v}_n^\top \frac{\partial L}{\partial \mathbf{v}_1} & \mathbf{v}_n^\top \frac{\partial L}{\partial \mathbf{v}_2} & \mathbf{v}_n^\top \frac{\partial L}{\partial \mathbf{v}_3} & \cdots & \mathbf{v}_n^\top \frac{\partial L}{\partial \mathbf{v}_n} \end{bmatrix} \tag{18}$$

$$\tilde{K} \circ V^\top \frac{\partial L}{\partial V} = \begin{bmatrix} 0 & \frac{1}{\lambda_2 - \lambda_1} \mathbf{v}_1^\top \frac{\partial L}{\partial \mathbf{v}_2} & \frac{1}{\lambda_3 - \lambda_1} \mathbf{v}_1^\top \frac{\partial L}{\partial \mathbf{v}_3} & \cdots & \frac{1}{\lambda_n - \lambda_1} \mathbf{v}_1^\top \frac{\partial L}{\partial \mathbf{v}_n} \\ \frac{1}{\lambda_1 - \lambda_2} \mathbf{v}_2^\top \frac{\partial L}{\partial \mathbf{v}_1} & 0 & \frac{1}{\lambda_3 - \lambda_2} \mathbf{v}_2^\top \frac{\partial L}{\partial \mathbf{v}_3} & \cdots & \frac{1}{\lambda_n - \lambda_2} \mathbf{v}_2^\top \frac{\partial L}{\partial \mathbf{v}_n} \\ \frac{1}{\lambda_1 - \lambda_3} \mathbf{v}_3^\top \frac{\partial L}{\partial \mathbf{v}_1} & \frac{1}{\lambda_2 - \lambda_3} \mathbf{v}_3^\top \frac{\partial L}{\partial \mathbf{v}_2} & 0 & \cdots & \frac{1}{\lambda_n - \lambda_3} \mathbf{v}_3^\top \frac{\partial L}{\partial \mathbf{v}_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\lambda_1 - \lambda_n} \mathbf{v}_n^\top \frac{\partial L}{\partial \mathbf{v}_1} & \frac{1}{\lambda_2 - \lambda_n} \mathbf{v}_n^\top \frac{\partial L}{\partial \mathbf{v}_2} & \frac{1}{\lambda_3 - \lambda_n} \mathbf{v}_n^\top \frac{\partial L}{\partial \mathbf{v}_3} & \cdots & 0 \end{bmatrix} \tag{19}$$

We do not use eigenvalues in the forward pass, so that it has no gradients, which means $\frac{\partial L}{\partial \Sigma} = 0$. Now let's consider the partial derivative *w.r.t* $\mathbf{v}_i$ and ignore $\frac{\partial L}{\partial \mathbf{v}_i}, i \neq 1$. Then $\frac{\partial L}{\partial M}$ would be,

$$\begin{aligned} \frac{\partial L}{\partial M} &= \begin{bmatrix} \sum_{i=2}^n \frac{1}{\lambda_1 - \lambda_i} \mathbf{v}_i \mathbf{v}_i^\top \frac{\partial L}{\partial \mathbf{v}_1} & \cancel{term_2} & \cancel{term_3} & \cdots & \cancel{term_n} \end{bmatrix} V^\top + \cancel{V \left(\frac{\partial L}{\partial \Sigma} \right)_{diag} V^\top} \\ &= \sum_{i=2}^n \frac{1}{\lambda_1 - \lambda_i} \mathbf{v}_i \mathbf{v}_i^\top \frac{\partial L}{\partial \mathbf{v}_1} \mathbf{v}_1^\top \end{aligned} \tag{20}$$

Now we have shown that the partial derivative of *e.g.*, $\mathbf{v}_1$ computed from Power Iteration and SVD share the same form when $k \to \inf$. Similar deductions could be done for $\mathbf{v}_i, i = 2, 3, ...$ This justifies that we could use power iteration method during backpropogation to approximate the gradients of SVD, but we need to choose an approximate iteration number.

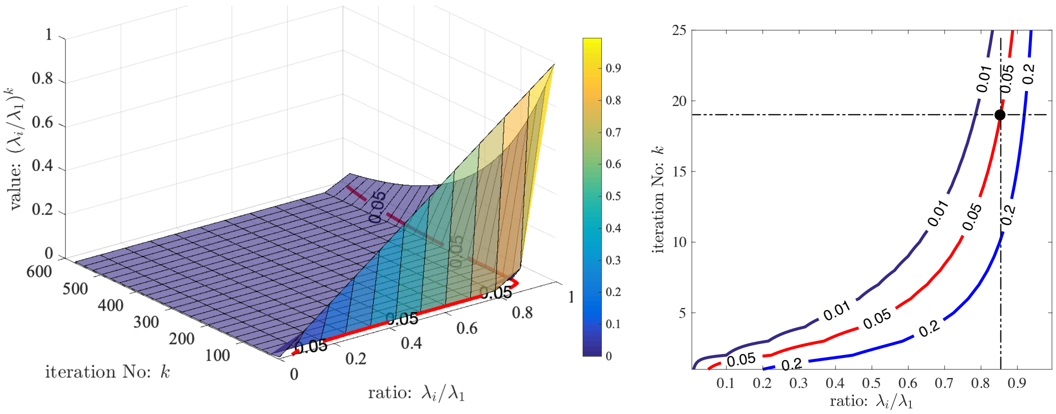
2.4 Number of Power Iterations

Fig. 1 shows how the value of $(\lambda_i/\lambda_1)^k$ evolves with different power iteration number k and ratio λ_i/λ_1. We need to select appropriate k for different λ_i/λ_1 given $(0 < \lambda_i/\lambda_1 \leq 1)$.

Let's assume $(\lambda_i/\lambda_1)^k < 0.05$ being a good approximation to $(\lambda_i/\lambda_1)^k = 0$. Then we have

$$(\lambda_i/\lambda_1)^k < 0.05 \Leftrightarrow k \ \ln(\lambda_i/\lambda_1) < \ln(0.05) \Leftrightarrow k \geq \frac{\ln(0.05)}{\ln(\lambda_i/\lambda_1)}. \tag{21}$$

The minimum value of k to satisfy $(\lambda_i/\lambda_1)^k < 0.05$ is $k = \lceil \frac{\ln(0.05)}{\ln(\lambda_i/\lambda_1)} \rceil$.



(a) (b)

Figure 1: (a) shows how the value of $(\lambda_k/\lambda_1)^k$ changes *w.r.t.* the eigenvalue ratio λ_k/λ_1 and iteration number k. (b) shows the contour of curved surface in (a).

λ_i/λ_1	0.2	0.4	0.6	0.8	**0.85**	0.9	0.95	0.99	0.995	0.999
$k = \lceil \frac{\ln(0.01)}{\ln(\lambda_i/\lambda_1)} \rceil$	2	4	6	14	**19**	29	59	299	598	2995

Table 1: The minimum value of k we need to guarantee $(\lambda_i/\lambda_1)^k < 0.05$.

Table 1 shows the minimum number of iterations we need to guarantee that the assumption holds. We can observe that when the two eigenvalues are very close to each other *e.g.*, $\lambda_i/\lambda_1 = 0.999$, we need about 3000 iterations to achieve a good approximation. However, in practice, the case is very rare, and we set power iteration number to be 19. This will satisfy most of the cases. Besides, two very close eigenvalues usually leads to overflow according to Eq.12 as the denominator $\lambda_1 - \lambda_i$ would be close to 0, but with our approximation, this problem could be avoided, and our method is more numerical stable.

3 Experiment

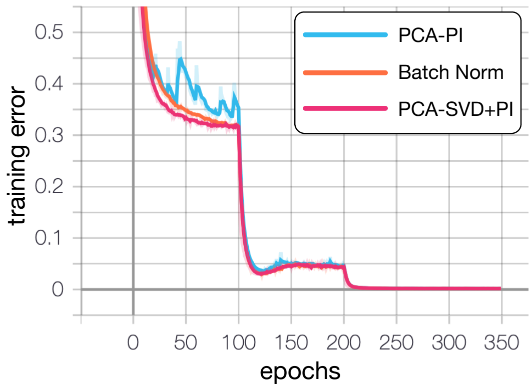
In the experiment, within the PCA denoising layer, we remove the noise in the feature maps by only selecting its top-k eigenvectors to reconstruct the input feature maps. We first reshape the input feature maps $X_{n\times c\times h\times w}$ to $X_{c\times nhw}$, and compute the covariance matrix $Var(X) = \frac{XX^\top}{nhw-1}$. The constraint for k is that $\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{n}\lambda_i} \geq 0.95$, which means that 95% of the information is preserved and the rest of the information which is lower than 5% is removed. In practice, k is relatively small compared with channel number c. For instance, in the first convolutional layer in ResNet18 which has the channel 64, we observe that 95% of the information could be preserved when $k = 8$, or 9.

Norm Methods	BN	PCA(PI)	PCA(SVD)	PCA(SVD+PI)
Minimum Error	4.66	5.05	NaN	**4.58**
Mean Error (4)	4.81±0.19	5.35±0.25	NaN	**4.67±0.06**

Table 2: CIFAR-10 test errors using ResNet18 (single PCA/ZCA normalization layer).

Norm Methods	BN	PCA(1 layer)	ZCA(1 layer)	PCA(1 block)	ZCA(1 block)
Minimum Error	4.66	4.58	4.91	5.14	
Mean Error (4)	4.81±0.19	4.67±0.06	5.02±0.28	5.30±0.12	

Table 3: CIFAR-10 test errors using ResNet18 (single PCA/ZCA normalization layer).



References

[1] Yuji Nakatsukasa and Nicholas J Higham. Stable and efficient spectral divide and conquer algorithms for the symmetric eigenvalue decomposition and the svd. *SIAM Journal on Scientific Computing*, 35(3):A1325–A1349, 2013.

[2] Mang Ye, Andy J Ma, Liang Zheng, Jiawei Li, and Pong C Yuen. Dynamic label graph matching for unsupervised video re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5142–5150, 2017.

[3] Catalin Ionescu, Orestis Vantzos, and Cristian Sminchisescu. Matrix backpropagation for deep networks with structured layers. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2965–2973, 2015.