**CMSC 35900 (Spring 2009) Large Scale Learning** **Lecture: 4**

# Dimensionality Reduction and Learning, continued

*Instructors: Sham Kakade and Greg Shakhnarovich*

## 1 PCA Projections and MLEs

Fix some $\lambda$. Consider the following 'keep or kill' estimator, which uses the MLE estimate if $\lambda_i \geq \lambda$ and 0 otherwise:

$$[\hat{\beta}_{PCA,\lambda}]_j = \begin{cases} [\hat{\beta}_0]_j & \text{if } \lambda_i \geq \lambda \\ 0 & \text{else} \end{cases}$$

where $\hat{\beta}_0$ is the MLE estimator. This estimator is 0 for the small values of $\lambda_i$ (those in which we are effectively regularizing more anyways).

**Theorem 1.1.** *(Risk Inflation of $\hat{\beta}_{PCA,\lambda}$)*
*Assume* $\mathrm{Var}(Y_i) = 1$, *then*

$$\mathbb{E}_Y[R(\hat{\beta}_{PCA,\lambda})] \leq 4\mathbb{E}_Y[R(\hat{\beta}_\lambda)]$$

Note that the the actual risk (not just an upper bound) of the simple PCA estimate is within a factor of 4 of the ridge regression risk on a wide class of problems.

*Proof.* Recall that:

$$\mathbb{E}_Y[R(\hat{\beta}_\lambda)] = \frac{1}{n}\sum_j (\frac{\lambda_j}{\lambda_j + \lambda})^2 + \sum_j \beta_j^2 \frac{\lambda_j}{(1+\lambda_j/\lambda)^2}$$

Since we can write the risk as:

$$\mathbb{E}_Y[R(\hat{\beta})] = \mathbb{E}_Y \|\hat{\beta} - \overline{\beta}\|_\Sigma^2 + \|\overline{\beta} - \beta\|_\Sigma^2$$

we have that:

$$\mathbb{E}_Y[R(\hat{\beta}_{PCA,\lambda})] = \frac{1}{n}\sum_j \mathbb{I}(\lambda_j > \lambda) + \sum_{j:\lambda_j < \lambda} \lambda_j \beta_j^2$$

where $\mathbb{I}$ is the indicator function.

We now show that each term in the risk of $\hat{\beta}_{PCA,\lambda}$ is within a factor of 4 for each term in $\hat{\beta}_\lambda$. If $\lambda_j > \lambda$, then the ratio of the $j-th$ terms is:

$$\begin{aligned} \frac{\frac{1}{n}}{\frac{1}{n}(\frac{\lambda_j}{\lambda_j+\lambda})^2 + \beta_j^2 \frac{\lambda_j}{(1+\lambda_j/\lambda)^2}} &\leq \frac{\frac{1}{n}}{\frac{1}{n}(\frac{\lambda_j}{\lambda_j+\lambda})^2} \\ &= \frac{(\lambda_j+\lambda)^2}{\lambda_j^2} \\ &\leq (1+\frac{\lambda}{\lambda_j})^2 \\ &\leq 4 \end{aligned}$$

Similarly, if $\lambda_j \leq \lambda$, then the ratio of the $j$-th terms is:

$$\begin{aligned}\frac{\lambda_j \beta_j^2}{\frac{1}{n}(\frac{\lambda_j}{\lambda_j+\lambda})^2 + \frac{\lambda_j \beta_j^2}{(1+\lambda_j/\lambda)^2}} &\leq \frac{\lambda_j \beta_j^2}{\frac{\lambda_j \beta_j^2}{(1+\lambda_j/\lambda)^2}} \\ &= (1+\lambda_j/\lambda)^2 \\ &\leq 4\end{aligned}$$

Since each term is within a factor of $4$, the proof is completed. □

# 2 Random Projections of Common Kernels

Let us say we have a Kernel $K(x, y)$, which represent an inner product between $x$ and $y$. When can we find random $\phi_w(x)$ and $\phi_w(y)$ such that:

$$\mathbb{E}_w[\phi_w(x)\phi_w(y)] = K(x, y)$$

and such that these have the concentration properties such that if we take many projections, then the average inner product will converge rapidly to $K(x, y)$ (so that we have an analogue of the inner product preserving lemma). For kernels of the form $K(x - y)$, Bochner's theorem provides a form of $\phi_w(x)$ and $p(w)$.

## 2.1 Bochner's theorem and shift invariant kernels

First, let us provide some background on Fourier Series. Given a positive finite measure $\mu$ on the real line $\mathbb{R}$ (i.e. if $\mu$ integrates to 1 then it would be a density), the Fourier transform $Q(t)$ of $\mu$ is the continuous function:

$$Q(t) = \int_{\mathbb{R}} e^{-itx} d\mu(x)$$

Clearly, the function $e^{-itx}$ is continuous and periodic. Also, $Q$ is continuous since for a fixed $x$, and the function $Q$ is a positive definite function. In particular, the kernel $K(x, y) := Q(y - x)$ is positive definite, which can be checked via a direct calculation.

Bochner's theorem says the converse is true:

**Theorem 2.1.** *(Bochner) Every positive definite function $Q$ is the Fourier transform of a positive finite Borel measure. This implies that for any shift invariant kernel, $K(x - y)$, we have that there exists a positive measure $\mu$ s.t.*

$$K(x - y) = \int_{\mathbb{R}} e^{-iw(x-y)} d\mu(w)$$

*and the measure $p = \mu / \int d\mu(w)$ is a probability measure.*

To see the implications of this, take a kernel $K(x - y)$ and it's corresponding measure $\mu(w)$ under Bochner's theorem. Let $\alpha = \int d\mu(w)$ and let $p = \mu/\alpha$. Now consider independently sampling $w_1, w_2, \ldots w_k \sim p$. Consider the random projection vector of $x$ to be:

$$\phi(x) = \frac{\alpha}{\sqrt{k}}(e^{-iw_1 x}, \ldots e^{-iw_k x})$$

Note that:

$$\phi(x) \cdot \phi(y) = \frac{1}{k} \sum_i e^{-iw_i(x-y)} \to K(x - y)$$

for large $k$. We can clearly ignore the imaginary components (as these have expectation 0), so it suffices to consider the projection vector:

$$\psi(x) = \frac{\alpha}{\sqrt{k}}(\cos(w_1 \cdot x), \sin(w_1 \cdot x), \ldots \cos(w_k \cdot x), \sin(w_k \cdot x))$$

which also is correct in expectation.

### 2.1.1 Example: Radial Basis Kernel

Consider the radial basis kernel:

$$K(x-y) = e^{-\frac{\|x-y\|^2}{2}}$$

It is easy to see that if choose $p$ to be the gaussian measure with variance and $\mu =$, then:

$$\int e^{-iw\cdot(x-y)} \frac{1}{(2\pi)^{d/2}} e^{-\|w\|^2/2} dw = e^{-\frac{\|x-y\|^2}{2}} = K(x-y)$$

Hence the sampling distribution for $w$ is Gaussian. If the bandwidth of the RBF kernel is not 1 then we scale the variance of the Gaussians.

For other scale invariant Kernels, there are different corresponding sampling measures $p$. However, we always use the fourier features.

### 2.1.2 High Probability Inner Product Preservation

Recall that to prove a risk bound using random features, the key was the inner product preserving lemma, which characterizes how fast the inner products in the projected space converge to the truth as a function of $k$.

We can do this here as well:

**Lemma 2.2.** *Let $x, y \in \mathbb{R}^d$ and let $K(x-y)$ be a (shift invariant) kernel. If $k = \frac{\alpha^2}{\epsilon^2} \log \frac{1}{\delta}$, and if $\phi$ is created using independently sampled $w_1, w_2, \ldots w_k \sim p$ (as discussed above), then with probability greater than $1-\delta$:*

$$|\phi(x)\cdot\phi(y) - K(x-y)| \leq \epsilon$$

*where $\phi(\cdot)$ uses the* cos *and* sin *features.*

*Proof.* Note that the cos and sin functions are bounded by 1. Now we can apply Hoeffdings directly to the random variables

$$\frac{\alpha}{k}\sum_i \cos(w_i\cdot x)\cos(w_i\cdot y), \qquad \frac{\alpha}{k}\sum_i \sin(w_i\cdot x)\sin(w_i\cdot y)$$

to show that these are $\epsilon$ close to their mean with the $k$ chosen above. This proves the result. □

### 2.1.3 Polynomial Kernels

Now say our we have a polynomial Kernel degree $l$ of the form:

$$K(x,y) = \sum_{i=1}^{l} c_i (x\cdot y)^l$$

Consider randomly sampling a set of projections $W = \{w_{i,j} : 1 \leq i \leq l, 1 \leq j \leq i\}$ of size $l(l-1)/2$. Now consider the random projetion (down to one dimension) of $K(x,.)$:

$$\phi_W(x) = \sum_{i=1}^{l} \sqrt{c_i}\Pi_{j=1}^{l} w_{i,j}\cdot x$$

One can verify that:

$$\mathbb{E}_W[\phi_W(x)\phi_W(y)] = K(x,y)$$

to see this, note

$$\mathbb{E}_W[(\Pi_{j=1}^{l} w_{i,j}\cdot x)(\Pi_{j'=1}^{l} w_{i,j'}\cdot y)] = \mathbb{E}_W[\Pi_{j=1}^{l}(w_{i,j}\cdot x)(w_{i,j}\cdot y)] = (x\cdot y)^l$$

where second step is due to independence.

Again consider independently sampling $W_1, W_2, \ldots W_k$ (note each $W_i$ is $O(l^2)$ random vectors). Consider the random projection vector of $x$ to be:

$$\frac{1}{\sqrt{k}}(\phi_{W_1}(x), \phi_{W_2}(x), \ldots \phi_{W_k}(x))$$

Concentration properties (for the an innerproduct preserving lemma) should be possible to prove as well (again using tail properties of Gaussians).

This random projection is not as convenient to use unless $l$ is small.