Problem 1. General view of GMM [Bi] Ex. 9.9

Recall that The expected value of the complete-data log likelihood function for GMM is given by

$$\mathbb{E}_{\boldsymbol{Z}}[\ln p(\boldsymbol{X},\boldsymbol{Z} \mid \boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi})]=\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\left\{\ln\pi_k+\ln\mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right\}$$

With a fixed $\gamma(z_{nk})$, find the maximizer $\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k$ for $\mathbb{E}_{\boldsymbol{Z}}[\ln p(\boldsymbol{X},\boldsymbol{Z} \mid \boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi})]$.

Solution. To find the minimum, we take the derivative with respect to $\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k$,

$$\begin{aligned}\frac{\partial\mathbb{E}_{\boldsymbol{Z}}[\ln p(\boldsymbol{X},\boldsymbol{Z} \mid \boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi})]}{\partial\boldsymbol{\mu}_k}&=\frac{\partial}{\partial\boldsymbol{\mu}_k}\sum_{n=1}^{N}\gamma(z_{nk})\ln\mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\\&=\frac{\partial}{\partial\boldsymbol{\mu}_k}\sum_{n=1}^{N}\gamma(z_{nk})\left(-\frac{1}{2}\ln(\det(2\pi\boldsymbol{\Sigma}_k))-\frac{1}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\right)\\&=\sum_{n=1}^{N}\gamma(z_{nk})\left(-\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\right)\end{aligned}$$

$$\begin{aligned}\frac{\partial\mathbb{E}_{\boldsymbol{Z}}[\ln p(\boldsymbol{X},\boldsymbol{Z} \mid \boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi})]}{\partial\boldsymbol{\Sigma}_k}&=\frac{\partial}{\partial\boldsymbol{\Sigma}_k}\sum_{n=1}^{N}\gamma(z_{nk})\ln\mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\\&=\frac{\partial}{\partial\boldsymbol{\Sigma}_k}\sum_{n=1}^{N}\gamma(z_{nk})\left(-\frac{1}{2}\ln(\det(2\pi\boldsymbol{\Sigma}_k))-\frac{1}{2}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\right)\\&=\sum_{n=1}^{N}\gamma(z_{nk})\left(-\frac{1}{2}\boldsymbol{\Sigma}_k^{-1}+\frac{1}{2}\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}\right)\end{aligned}$$

By the first order condition of the maximum, we set the derivatives to 0 and get,

$$\sum_{n=1}^{N}\gamma(z_{nk})\left(-\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)\right)=0\Rightarrow\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{x}_n-\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{\mu}_k=0$$

$$\tilde{\boldsymbol{\mu}}_k=\frac{\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{x}_n}{\sum_{n=1}^{N}\gamma(z_{nk})}$$

$$\sum_{n=1}^{N}\gamma(z_{nk})\left(-\frac{1}{2}\boldsymbol{\Sigma}_k^{-1}+\frac{1}{2}\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_k)(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}\right)=0\Rightarrow\sum_{n=1}^{N}\gamma(z_{nk})\left(-1+(\boldsymbol{x}_n-\boldsymbol{\mu}_k)(\boldsymbol{x}_n-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}_k^{-1}\right)=0$$

$$\tilde{\boldsymbol{\Sigma}}_k=\frac{\sum_{n=1}^{N}\gamma(z_{nk})(\boldsymbol{x}_n-\tilde{\boldsymbol{\mu}}_k)(\boldsymbol{x}_n-\tilde{\boldsymbol{\mu}}_k)^\top}{\sum_{n=1}^{N}\gamma(z_{nk})}$$

■

Problem 2. K-means as the limit of EM cf. [Bi] Ch.9.3.2

Consider the EM algorithm where the covariance matrices of the mixture components are all given by $\boldsymbol{\Sigma}_k=\epsilon\boldsymbol{I},k=1,\cdots,K$.

1. Write $p(\boldsymbol{x} \mid \boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$.

   Solution.

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)=\det(2\pi\boldsymbol{\Sigma}_k)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right)=(2\pi\epsilon)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\epsilon}\|\boldsymbol{x}_n-\boldsymbol{\mu}_k\|^2\right\}$$

■

2. Show that $\gamma\left(z_{nk}\right) \to r_{nk}$ as $\epsilon \to 0$, where $r_{nk} = 1$ if $k = \operatorname{argmin}_j \left\|\boldsymbol{x}_n - \boldsymbol{\mu}_j\right\|^2$ and $r_{nk} = 0$ otherwise.

Solution.

$$\gamma(z_{nk}) = \frac{\pi_k \exp\left\{-\frac{1}{2\epsilon}\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2\right\}}{\sum_j \pi_j \exp\left\{-\frac{1}{2\epsilon}\|\boldsymbol{x}_n - \boldsymbol{\mu}_j\|^2\right\}}$$

When $\epsilon \to 0$, in the denominator the term for which $\|\boldsymbol{x}_n - \boldsymbol{\mu}_j\|^2$ is smallest will go to zero most slowly. Therefore, $\gamma(z_{nk})$ will go to zero except for term j, for which it will go to 1.

Note that this holds independent of the value of $\boldsymbol{\pi}$ (as long as it is not 0). Each data point is thereby assigned to the cluster having the closest mean. ■

3. Show that as $\epsilon \to 0$,

$$\mathbb{E}_{\boldsymbol{Z}}[\ln p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] \to -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\right\|^2 + \text{ const.}$$

Solution.

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{Z}}[\ln p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] &= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma\left(z_{nk}\right)\left\{\ln \pi_k + \ln \mathcal{N}\left(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right\} \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left\{\ln \pi_k - \frac{1}{2}\ln(2\pi\epsilon) - \frac{1}{2\epsilon}\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2\right\} \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left(-\frac{1}{2\epsilon}\|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|^2\right) + \text{ const.}
\end{aligned}$$

■

Problem 3. Rayleigh quotient

The Rayleigh quotient for a real symmetric matrix $\boldsymbol{A}$ and a nonzero vector $\boldsymbol{v}$ is given by

$$\rho(\boldsymbol{v}, \boldsymbol{A}) = \frac{\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v}}{\boldsymbol{v}^\top \boldsymbol{v}}.$$

Prove that the $\rho(\boldsymbol{v}, \boldsymbol{A}) \in [\lambda_{\min}, \lambda_{\max}]$ where $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and largest eigenvalues of $\boldsymbol{A}$, respectively. For what $\boldsymbol{v}$ does $\rho(\boldsymbol{v}, \boldsymbol{A})$ achieve the min and the max, respectively?

Solution. Note that the Rayleigh quotient is scaling invariant, i.e. $\rho(\boldsymbol{v}, \boldsymbol{A}) = \rho(\alpha\boldsymbol{v}, \boldsymbol{A})$. Without the loss of generality, we consider the following constrained problem:

$$\max_{\boldsymbol{v}\in\mathbb{R}^n : \|\boldsymbol{v}\|=1} \boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v}$$

Let $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$ be the eigenvalue decomposition of $\boldsymbol{A}$, where $\boldsymbol{Q} = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n]$ are orthogonal eigenvectors, $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ are eigenvalues. Then for any unit vector $\boldsymbol{v}$,

$$\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} = \boldsymbol{v}^\top(\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top)\boldsymbol{v} = (\boldsymbol{v}^\top\boldsymbol{Q})\boldsymbol{\Lambda}(\boldsymbol{Q}^\top\boldsymbol{v}) = \boldsymbol{y}^\top\boldsymbol{\Lambda}\boldsymbol{y}$$

where $\boldsymbol{y} = \boldsymbol{Q}^\top\boldsymbol{v}$ is also a unit vector:

$$\|\boldsymbol{y}\|^2 = \boldsymbol{y}^T\boldsymbol{y} = \left(\boldsymbol{Q}^T\boldsymbol{v}\right)^T\left(\boldsymbol{Q}^T\boldsymbol{v}\right) = \boldsymbol{v}^T\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{v} = \boldsymbol{v}^T\boldsymbol{v} = 1$$

So the original optimization problem becomes the following one:

$$\max_{\boldsymbol{y}\in\mathbb{R}^n : \|\boldsymbol{y}\|=1} \boldsymbol{y}^T\boldsymbol{\Lambda}\boldsymbol{y}$$

To solve this new problem, write $\boldsymbol{y} = (y_1, \ldots, y_n)^T$. It follows that

$$\boldsymbol{y}^T \boldsymbol{\Lambda} \boldsymbol{y} = \sum_{i=1}^{n} \lambda_i y_i^2 \quad \text{(subject to } y_1^2 + y_2^2 + \cdots + y_n^2 = 1 \text{)}$$

Because $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, when $y_1^2 = 1, y_2^2 = \cdots = y_n^2 = 0$ (i.e., $\boldsymbol{y} = \pm \boldsymbol{e}_1$), the objective function attains its maximum value $\boldsymbol{y}^T \boldsymbol{\Lambda} \boldsymbol{y} = \lambda_1$.

In terms of the original variable $\boldsymbol{v}$, the maximizer is

$$\boldsymbol{v}_{\max} = \boldsymbol{Q} \boldsymbol{y}_{\max} = \boldsymbol{Q}(\pm \boldsymbol{e}_1) = \pm \boldsymbol{q}_1.$$

The minimum is the same procedure, resulting in $\boldsymbol{v}_{\min} = \pm \boldsymbol{q}_n$ ■

Problem 4. Graph Laplacian

1. Prove that all the eigenvalues of the graph Laplacian $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ are non-negative.

Solution.

$$\begin{aligned}
\boldsymbol{z}^\top \boldsymbol{L} \boldsymbol{z} &= \boldsymbol{z}^\top (\boldsymbol{D} - \boldsymbol{W}) \boldsymbol{z} \\
&= \sum_{n=1}^{N} z_n d_n z_n - \sum_{n=1}^{N} \sum_{m=1}^{N} z_n W_{nm} z_m \\
&= \frac{1}{2} \sum_{n=1}^{N} z_n^2 \sum_{m=1}^{N} W_{nm} + \frac{1}{2} \sum_{m=1}^{N} z_m^2 \sum_{n=1}^{N} W_{nm} - \sum_{n=1}^{N} \sum_{m=1}^{N} z_n z_m W_{nm} \\
&= \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} (z_n^2 W_{nm} + z_m^2 W_{nm} - 2 z_n z_m W_{nm}) \\
&= \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} |z_n - z_m|^2 W_{nm} \geq 0
\end{aligned}$$

Therefore, $\boldsymbol{L}$ is positive semidefinite and therefore its eigenvalues are nonnegative. ■

2. Prove that all the eigenvalues of the normalized graph Laplacian $\boldsymbol{L}_{\text{sym}} = \boldsymbol{I} - \boldsymbol{D}^{-1/2} \boldsymbol{W} \boldsymbol{D}^{-1/2}$ are in $[0, 2]$.

Solution. We could see that $\boldsymbol{L}_{\text{sym}} = \boldsymbol{I} - \boldsymbol{D}^{-1/2} \boldsymbol{W} \boldsymbol{D}^{-1/2} = \boldsymbol{D}^{-1/2} (\boldsymbol{D} - \boldsymbol{W}) \boldsymbol{D}^{-1/2} = \boldsymbol{D}^{-1/2} \boldsymbol{L} \boldsymbol{D}^{-1/2}$.

First, we show that 0 is an eigenvalue of $\boldsymbol{L}_{\text{sym}}$ using $\boldsymbol{x} = \boldsymbol{D}^{1/2} \boldsymbol{e}$,

$$\boldsymbol{L}_{\text{sym}} \boldsymbol{D}^{1/2} \boldsymbol{e} = \boldsymbol{D}^{-1/2} \boldsymbol{L} \boldsymbol{D}^{-1/2} \boldsymbol{D}^{1/2} \boldsymbol{e} = \boldsymbol{D}^{-1/2} \boldsymbol{L} \boldsymbol{e} = 0$$

since $\boldsymbol{D}\boldsymbol{e} - \boldsymbol{W}\boldsymbol{e} = 0$. Therefore, $\boldsymbol{x}$ is an eigenvector of $\boldsymbol{L}_{\text{sym}}$ with eigenvalue 0. To show that it is the smallest, note that $\boldsymbol{L}_{\text{sym}}$ is also positive semidefinite,

$$\boldsymbol{z}^\top \boldsymbol{L}_{\text{sym}} \boldsymbol{z} = \boldsymbol{z}^\top \boldsymbol{D}^{-1/2} \boldsymbol{L} \boldsymbol{D}^{-1/2} \boldsymbol{z} = \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \frac{|z_n - z_m|^2 W_{nm}}{\sqrt{d_n d_m}} \geq 0$$

Thus, the eigenvalues are non-negative and 0 is the smallest eigenvalue.

Similarly, we can show that $\boldsymbol{I} + \boldsymbol{D}^{-1/2} \boldsymbol{W} \boldsymbol{D}^{-1/2}$ is also positive semidefinite.

$$\boldsymbol{z}^\top (\boldsymbol{I} + \boldsymbol{D}^{-1/2} \boldsymbol{W} \boldsymbol{D}^{-1/2}) \boldsymbol{z} = \boldsymbol{z}^\top \boldsymbol{D}^{-1/2} (\boldsymbol{D} + \boldsymbol{W}) \boldsymbol{D}^{-1/2} \boldsymbol{z} = \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \frac{|z_n + z_m|^2 W_{nm}}{\sqrt{d_n d_m}} \geq 0$$

Therefore, $\boldsymbol{z}^\top(\boldsymbol{I}+\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2})\boldsymbol{z} \geq 0$ and we have

$$-\boldsymbol{z}^\top\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}\boldsymbol{z} \leq \boldsymbol{z}^\top\boldsymbol{z} \Rightarrow \boldsymbol{z}^\top\boldsymbol{I}\boldsymbol{z} - \boldsymbol{z}^\top\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}\boldsymbol{z} \leq 2\boldsymbol{z}^\top\boldsymbol{z} \Rightarrow \frac{\boldsymbol{z}^\top\boldsymbol{L}_{\text{sym}}\boldsymbol{z}}{\boldsymbol{z}^\top\boldsymbol{z}} \leq 2$$

By Rayleigh quotient, $\lambda_{\max} \leq 2$. ■

Problem 5. One-class SVM

The optimization problem for one-class SVM is

$$\begin{array}{ll} \min & R^2 + C\sum_{n=1}^N \xi_n \\ \text{s.t.} & \left\|\phi\left(\boldsymbol{x}_n\right)-\boldsymbol{a}\right\|^2 \leq R^2+\xi_n \text{ for all } n \\ & \xi_n \geq 0 \text{ for all } n \end{array}$$

Write the Lagrangian and express it using only the Lagrange multipliers and the kernel $K\left(\boldsymbol{x}_n, \boldsymbol{x}_m\right) = \phi\left(\boldsymbol{x}_n\right)^\top \phi\left(\boldsymbol{x}_m\right)$.

Solution. The Lagrangian is,

$$L(R, a, \alpha_n, \xi_n) = R^2 + C\sum_{n=1}^N \xi_n - \sum_{n=1}^N \gamma_n\xi_n - \sum_{n=1}^N \alpha_n\left(R^2+\xi_n-(\phi(\boldsymbol{x}_n)-\boldsymbol{a})^\top(\phi(\boldsymbol{x}_n)-\boldsymbol{a})\right)$$

with Lagrange multipliers $\alpha_i, \gamma_i \geq 0$. Then, we take the derivative with respect to the primal variables $\boldsymbol{a}$, ξ_i and R,

$$\begin{aligned} \frac{\partial L(R, a, \alpha_n, \xi_n)}{\partial \boldsymbol{a}} &= 2\sum_{n=1}^N \alpha_n(\boldsymbol{a}-\phi(\boldsymbol{x}_n)) \\ \frac{\partial L(R, a, \alpha_n, \xi_n)}{\partial \xi_n} &= C - \gamma_n - \alpha_n \\ \frac{\partial L(R, a, \alpha_n, \xi_n)}{\partial R} &= 2R - 2R\sum_{n=1}^N \alpha_n \end{aligned}$$

Set them to zero and we get $\boldsymbol{a} = \sum_{n=1}^N \alpha_n\phi(\boldsymbol{x}_n)$, $\gamma_n = C - \alpha_n$, $0 \leq \alpha_n \leq C$, and $\sum_{n=1}^N \alpha_n = 1$. Substituting them into the Lagrangian we obtain the following dual problem where we maximize with respect to α_i:

$$\begin{aligned} L(R, a, \alpha_n, \xi_n) &= R^2 + C\sum_{n=1}^N \xi_n - \sum_{n=1}^N \gamma_n\xi_n - \sum_{n=1}^N \alpha_n\left(R^2+\xi_n-(\phi(\boldsymbol{x}_n)-\boldsymbol{a})^\top(\phi(\boldsymbol{x}_n)-\boldsymbol{a})\right) \\ &= R^2 + C\sum_{n=1}^N \xi_n - \sum_{n=1}^N (C-\alpha_n)\xi_n - R^2\sum_{n=1}^N \alpha_n - \sum_{n=1}^N \alpha_n\xi_n + \sum_{n=1}^N \alpha_n(\phi(\boldsymbol{x}_n)-\boldsymbol{a})^\top(\phi(\boldsymbol{x}_n)-\boldsymbol{a}) \\ &= \sum_{n=1}^N \alpha_n(\phi(\boldsymbol{x}_n)-\sum_{m=1}^N \alpha_m\phi(\boldsymbol{x}_m))^\top(\phi(\boldsymbol{x}_n)-\sum_{m=1}^N \alpha_m\phi(\boldsymbol{x}_m)) \\ &= \sum_{n=1}^N \alpha_n(\phi(\boldsymbol{x}_n)^\top\phi(\boldsymbol{x}_n)) - \sum_{n=1}^N\sum_{m=1}^N \alpha_n\alpha_m(\phi(\boldsymbol{x}_n)^\top\phi(\boldsymbol{x}_m)) \\ &= \sum_{n=1}^N \alpha_n K\left(\boldsymbol{x}_n, \boldsymbol{x}_n\right) - \sum_{n=1}^N\sum_{m=1}^N \alpha_n\alpha_m K\left(\boldsymbol{x}_n, \boldsymbol{x}_m\right) \end{aligned}$$

with constrains $0 \leq \alpha_n \leq C$, $\sum_{n=1}^N \alpha_n = 1$. ■

Problem 6. RKHS cf. [HaTF] Ex.5.16

Recall that $K(x, y) = \sum_{j=1}^\infty \gamma_j\phi_j(x)\phi_j(y)$ for which we can order $\gamma_1 \geq \gamma_2 \geq \cdots$ and $\{\phi_j\}_{j=1}^\infty$ is orthonormal:

$\langle \phi_i, \phi_j \rangle = \delta_{ij}$. Consider the ridge regression problem

$$\min_{\{c_j\}_{j=1}^{\infty}} \sum_{n=1}^{N} \left(y_n - \sum_{j=1}^{\infty} c_j \phi_j (x_n) \right)^2 + \lambda \sum_{j=1}^{\infty} \frac{c_j^2}{\gamma_j},$$

1. Explain why the problem is equivalent to

$$\min_{\boldsymbol{\alpha}} (\boldsymbol{y} - \boldsymbol{K\alpha})^\top (\boldsymbol{y} - \boldsymbol{K\alpha}) + \lambda \boldsymbol{\alpha}^\top \boldsymbol{K\alpha}.$$

Solution. In this setting, we have $f(x) = \sum_{i=1}^{\infty} c_i \phi_i(x)$, $\|f\|_{\mathcal{H}_K}^2 = \sum_{i=1}^{\infty} c_i^2/\gamma_i$.

The solution have the form $f(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i)$. By HW3, we have $\|f\|_{\mathcal{H}_K}^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} K(x_i, x_j)\alpha_i \alpha_j$. Substitute them into the problem yield the results. ■

2. Assume $K(x, y) = \sum_{m=1}^{M} h_m(x) h_m(y)$ and $M \geq N$. Prove:

$$\boldsymbol{h}(x) = \boldsymbol{V} \boldsymbol{D}_\gamma^{1/2} \boldsymbol{\phi}(x)$$

where $\boldsymbol{h}(x) = [h_1(x), \cdots, h_M(x)]^\top$ and $\boldsymbol{\phi}(x) = [\phi_1(x), \cdots, \phi_M(x)]^\top$; $\boldsymbol{V}$ is an $M \times M$ orthogonal matrix and $\boldsymbol{D}_\gamma = \operatorname{diag}(\gamma_1, \cdots, \gamma_M)$. What are $\boldsymbol{V}$ and $\boldsymbol{D}_\gamma$? (Hint: $h_m = \sum_{j=1}^{M} \langle h_m, \phi_j \rangle \phi_j$).

Solution. From the definition of the kernel, we have

$$K(x, y) = \sum_{m=1}^{M} h_m(x) h_m(y) = \sum_{j=1}^{\infty} \gamma_j \phi_j(x) \phi_j(y)$$

Multiply both side by $\phi_k(x)$ yields,

$$\sum_{m=1}^{M} \langle h_m(x), \phi_k(x) \rangle h_m(y) = \sum_{j=1}^{\infty} \gamma_j \langle \phi_j(x), \phi_k(x) \rangle \phi_j(y) = \sum_{j=1}^{\infty} \gamma_j \delta_{jk} \phi_j(y) = \gamma_k \phi_j(y)$$

Let $g_{km} = \langle h_m(x), \phi_k(x) \rangle$ and multiply both side by $\phi_l(y)$ yields,

$$\sum_{m=1}^{M} g_{km} \langle h_m(y), \phi_l(y) \rangle = \gamma_k \langle \phi_j(y), \phi_l(y) \rangle \quad \Rightarrow \quad \sum_{m=1}^{M} g_{km} g_{lm} = \gamma_k \delta_{kl}$$

Let $\boldsymbol{G}_M = \{g_{nm}\} \in \mathbb{R}^{M \times N}$, we have

$$\boldsymbol{G}_M \boldsymbol{G}_M^\top = \operatorname{diag}\{\gamma_1, \gamma_2, \ldots, \gamma_M\} = \boldsymbol{D}_\gamma$$

Let $\boldsymbol{V}^\top = \boldsymbol{D}_\gamma^{-\frac{1}{2}} \boldsymbol{G}_M$, then

$$\boldsymbol{V}\boldsymbol{V}^\top = \boldsymbol{G}_M \boldsymbol{D}_\gamma^{-1} \boldsymbol{G}_M^\top = \boldsymbol{I}$$

Then, we have

$$\sum_{m=1}^{M} g_{km} h_m(y) = \gamma_k \phi_j(y) \Rightarrow \boldsymbol{G}_M h(x) = \boldsymbol{D}_\gamma \boldsymbol{\phi}(x) \Rightarrow \boldsymbol{V}\boldsymbol{D}_\gamma^{-\frac{1}{2}} \boldsymbol{G}_M h(x) = \boldsymbol{V}\boldsymbol{D}_\gamma^{-\frac{1}{2}} \boldsymbol{D}_\gamma \boldsymbol{\phi}(x) \Rightarrow h(x) = \boldsymbol{V}\boldsymbol{D}^{\frac{1}{2}} \boldsymbol{\phi}(x)$$

■