Kernel Based Support Vector Machines
Proof: Weights Lie in Space Spanned by $\boldsymbol{\phi}(\boldsymbol{x}^j)$

Background: Classifier training features and labels are $\boldsymbol{x}^i, d^i, i = 1, 2, \ldots N$. Classification in a high-dimensional feature space is performed using the mapping $\boldsymbol{\phi}(\boldsymbol{x})$ as $\hat{d}(\boldsymbol{x}) = \text{sign}\left\{\boldsymbol{\phi}^T(\boldsymbol{x})\boldsymbol{w}\right\}$.

Claim: The weights $\boldsymbol{w}$ that satisfy

$$\min_{\boldsymbol{w}} f(\boldsymbol{w}) = \min_{\boldsymbol{w}} \sum_{i=1}^{N} \left(1 - d^i \boldsymbol{\phi}^T(\boldsymbol{x}^i)\boldsymbol{w}\right)_+ + \lambda||\boldsymbol{w}||_2^2 \tag{1}$$

are of the form

$$\boldsymbol{w} = \sum_{j=1}^{N} \boldsymbol{\phi}(\boldsymbol{x}^j)\alpha_j$$

Proof: The proof proceeds by adding a component to $\boldsymbol{w}$ that is orthogonal to the space spanned by the vectors $\{\boldsymbol{\phi}(\boldsymbol{x}^j), j = 1, 2, \ldots, N\}$ and then showing that component must be zero at the minimum of Eq. 1.

Suppose

$$\boldsymbol{w} = \sum_{j=1}^{N} \boldsymbol{\phi}(\boldsymbol{x}^j)\alpha_j + \boldsymbol{\phi}^\perp$$

where $\boldsymbol{\phi}^\perp$ is orthogonal to the space spanned by $\{\boldsymbol{\phi}(\boldsymbol{x}^j), j = 1, 2, \ldots, N\}$, that is, $\boldsymbol{\phi}^T(\boldsymbol{x}^j)\boldsymbol{\phi}^\perp = 0, j = 1, 2, \ldots, N\}$. Note that any vector $\boldsymbol{w}$ can be expressed as a sum of a component in the space spanned by the $\boldsymbol{\phi}(\boldsymbol{x}^j)$ and a component orthogonal to that same space.

The optimization problem Eq. 1 may be rewritten

$$\min_{\boldsymbol{w}} f(\boldsymbol{w}) = \min_{\boldsymbol{\alpha}, \boldsymbol{\phi}^\perp} \sum_{i=1}^{N} \left(1 - d^i \boldsymbol{\phi}^T(\boldsymbol{x}^i)\left(\sum_{j=1}^{N} \boldsymbol{\phi}(\boldsymbol{x}^j)\alpha_j + \boldsymbol{\phi}^\perp\right)\right)_+ + \lambda \left|\left|\sum_{j=1}^{N} \boldsymbol{\phi}(\boldsymbol{x}^j)\alpha_j + \boldsymbol{\phi}^\perp\right|\right|_2^2 \tag{2}$$

$$= \min_{\boldsymbol{\alpha}, \boldsymbol{\phi}^\perp} \sum_{i=1}^{N} \left(1 - d^i \left(\sum_{j=1}^{N} \boldsymbol{\phi}^T(\boldsymbol{x}^i)\boldsymbol{\phi}(\boldsymbol{x}^j)\alpha_j + \boldsymbol{\phi}^T(\boldsymbol{x}^i)\boldsymbol{\phi}^\perp\right)\right)_+$$

$$+ \lambda \left(\sum_{i=1}^{N}\sum_{j=1}^{N} \boldsymbol{\phi}^T(\boldsymbol{x}^i)\boldsymbol{\phi}(\boldsymbol{x}^j)\alpha_i\alpha_j + 2\sum_{i=1}^{N} \alpha_i \boldsymbol{\phi}^T(\boldsymbol{x}^i)\boldsymbol{\phi}^\perp + {\boldsymbol{\phi}^\perp}^T\boldsymbol{\phi}^\perp\right) \tag{3}$$

where in the second line we have used the identity

$$\left|\left|\sum_{j=1}^{N}\boldsymbol{\phi}(\boldsymbol{x}^j)\alpha_j + \boldsymbol{\phi}^{\perp}\right|\right|_2^2 = \left(\sum_{i=1}^{N}\boldsymbol{\phi}(\boldsymbol{x}^i)\alpha_i + \boldsymbol{\phi}^{\perp}\right)^T \left(\sum_{j=1}^{N}\boldsymbol{\phi}(\boldsymbol{x}^j)\alpha_j + \boldsymbol{\phi}^{\perp}\right)$$

and multiplied out the terms in the product.

Now use the fact that $\boldsymbol{\phi}^T(\boldsymbol{x}^i)\boldsymbol{\phi}^{\perp} = 0$ to reexpress the optimization problem as

$$\min_{\boldsymbol{w}} f(\boldsymbol{w}) = \min_{\boldsymbol{\alpha},\boldsymbol{\phi}^{\perp}} \sum_{i=1}^{N}\left(1 - d^i \sum_{j=1}^{N}\boldsymbol{\phi}^T(\boldsymbol{x}^i)\boldsymbol{\phi}(\boldsymbol{x}^j)\alpha_j\right)_+ + \lambda\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{\phi}^T(\boldsymbol{x}^i)\boldsymbol{\phi}(\boldsymbol{x}^j)\alpha_i\alpha_j + \lambda{\boldsymbol{\phi}^{\perp}}^T\boldsymbol{\phi}^{\perp} \tag{4}$$

$$= \min_{\boldsymbol{\alpha},\boldsymbol{\phi}^{\perp}} \sum_{i=1}^{N}\left(1 - d^i \sum_{j=1}^{N}\boldsymbol{\phi}^T(\boldsymbol{x}^i)\boldsymbol{\phi}(\boldsymbol{x}^j)\alpha_j\right)_+ + \lambda\left|\left|\sum_{j=1}^{N}\boldsymbol{\phi}(\boldsymbol{x}^j)\alpha_j\right|\right|_2^2 + \lambda\left|\left|\boldsymbol{\phi}^{\perp}\right|\right|_2^2 \tag{5}$$

The only term containing $\boldsymbol{\phi}^{\perp}$ is the last one, $\lambda\left|\left|\boldsymbol{\phi}^{\perp}\right|\right|_2^2$, which is nonnegative since $\lambda > 0$. Consequently, we conclude the minimum is attained when $\boldsymbol{\phi}^{\perp} = \mathbf{0}$ and thus

$$\boldsymbol{w} = \sum_{j=1}^{N}\boldsymbol{\phi}(\boldsymbol{x}^j)\alpha_j$$