Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact subset. Assume $x_1, \ldots, x_n$ are i.i.d. and $y_1, \ldots, y_n = \pm 1$ for classification and $[-1, 1]$ for regression. Assume we have a kernel $K(x, y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(y)$, $\lambda_i > 0$.

Consider a map

$$x \in \mathcal{X} \mapsto \phi(x) = (\sqrt{\lambda_1}\phi_1(x), \ldots, \sqrt{\lambda_k}\phi_k(x), \ldots) = (\sqrt{\lambda_k}\phi_k(x))_{k\geq 1} \in \mathcal{H}$$

where $\mathcal{H}$ is a Hilbert space.

Consider the scalar product in $\mathcal{H}$: $(u, v)_{\mathcal{H}} = \sum_{i=1}^{\infty} u_i v_i$ and $\|u\|_{\mathcal{H}} = \sqrt{(u, v)_{\mathcal{H}}}$.

For $x, y \in \mathcal{X}$,

$$(\phi(x), \phi(y))_{\mathcal{H}} = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(y) = K(x, y).$$

Function ϕ is called *feature map*

Family of classifiers:

$$\mathcal{F}_{\mathcal{H}} = \{(w, z)_{\mathcal{H}} : \|w\|_{\mathcal{H}} \leq 1\}.$$

$$\mathcal{F} = \{(w, \phi(x))_{\mathcal{H}} : \|w\|_{\mathcal{H}} \leq 1\} \ni f : \mathcal{X} \mapsto \mathbb{R}.$$

Algorithms:

(1) **SVMs**

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) = (\underbrace{\sum_{i=1}^{n} \alpha_i \phi(x_i)}_{w}, \phi(x))_{\mathcal{H}}$$

Here, instead of taking any w, we only take w as a linear combination of images of data points. We have a choice of Loss function $\mathcal{L}$:

- $\mathcal{L}(y, f(x)) = I(yf(x) \leq 0)$ – classification
- $\mathcal{L}(y, f(x)) = (y - f(x))^2$ – regression

(2) **Square-loss regularization**

Assume an algorithm outputs a classifier from $\mathcal{F}$ (or $\mathcal{F}_{\mathcal{H}}$), $f(x) = (w, \phi(x))_{\mathcal{H}}$. Then, as in Lecture 18,

$$\begin{aligned}
\mathbb{P}\left(yf(x) \le 0\right) \le \mathbb{E}\varphi_\delta\left(yf(x)\right) &= \frac{1}{n}\sum_{i=1}^n \varphi_\delta\left(y_i f(x_i)\right) + \left(\mathbb{E}\varphi_\delta\left(yf(x)\right) - \frac{1}{n}\sum_{i=1}^n \varphi_\delta\left(y_i f(x_i)\right)\right) \\
&\le \frac{1}{n}\sum_{i=1}^n I(y_i f(x_i) \le \delta) + \sup_{f\in\mathcal{F}}\left(\mathbb{E}\varphi_\delta\left(yf(x)\right) - \frac{1}{n}\sum_{i=1}^n \varphi_\delta\left(y_i f(x_i)\right)\right)
\end{aligned}$$

By McDiarmid's inequality, with probability at least $1 - e^{-t}$

$$\sup_{f\in\mathcal{F}}\left(\mathbb{E}\varphi_\delta\left(yf(x)\right) - \frac{1}{n}\sum_{i=1}^n \varphi_\delta\left(y_i f(x_i)\right)\right) \le \mathbb{E}\sup_{f\in\mathcal{F}}\left(\mathbb{E}\varphi_\delta\left(yf(x)\right) - \frac{1}{n}\sum_{i=1}^n \varphi_\delta\left(y_i f(x_i)\right)\right) + \sqrt{\frac{2t}{n}}$$

Using the symmetrization technique,

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left(\mathbb{E}(\varphi_\delta\left(yf(x)\right) - 1) - \frac{1}{n}\sum_{i=1}^n (\varphi_\delta\left(y_i f(x_i)\right) - 1)\right) \le 2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\big(\varphi_\delta\left(y_i f(x_i)\right) - 1\big)\right|.$$

Since $\delta \cdot (\varphi_\delta - 1)$ is a contraction,

$$\begin{aligned}
&2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\big(\varphi_\delta\left(y_i f(x_i)\right) - 1\big)\right| \le \frac{2}{\delta}2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i y_i f(x_i)\right| \\
&= \frac{4}{\delta}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\right| = \frac{4}{\delta}\mathbb{E}\sup_{\|w\|\le 1}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i (w, \phi(x_i))_{\mathcal{H}}\right| \\
&= \frac{4}{\delta n}\mathbb{E}\sup_{\|w\|\le 1}\left|(w, \sum_{i=1}^n \varepsilon_i \phi(x_i))_{\mathcal{H}}\right| = \frac{4}{\delta n}\mathbb{E}\sup_{\|w\|\le 1}\left\|\sum_{i=1}^n \varepsilon_i \phi(x_i)\right\|_{\mathcal{H}} \\
&= \frac{4}{\delta n}\mathbb{E}\sqrt{\left(\sum_{i=1}^n \varepsilon_i \phi(x_i), \sum_{i=1}^n \varepsilon_i \phi(x_i)\right)_{\mathcal{H}}} = \frac{4}{\delta n}\mathbb{E}\sqrt{\sum_{i,j} \varepsilon_i \varepsilon_j (\phi(x_i), \phi(x_i))_{\mathcal{H}}} \\
&= \frac{4}{\delta n}\mathbb{E}\sqrt{\sum_{i,j} \varepsilon_i \varepsilon_j K(x_i, x_j)} \le \frac{4}{\delta n}\sqrt{\mathbb{E}\sum_{i,j} \varepsilon_i \varepsilon_j K(x_i, x_j)} \\
&= \frac{4}{\delta n}\sqrt{\sum_{i=1}^n \mathbb{E}K(x_i, x_i)} = \frac{4}{\delta}\sqrt{\frac{\mathbb{E}K(x_1, x_1)}{n}}
\end{aligned}$$

Putting everything together, with probability at least $1 - e^{-t}$,

$$\mathbb{P}\left(yf(x) \leq 0\right) \leq \frac{1}{n}\sum_{i=1}^{n} I(y_i f(x_i) \leq \delta) + \frac{4}{\delta}\sqrt{\frac{\mathbb{E}K(x_1, x_1)}{n}} + \sqrt{\frac{2t}{n}}.$$

Before the contraction step, we could have used Martingale method again to have $\mathbb{E}_\varepsilon$ only. Then $\mathbb{E}K(x_1, x_1)$ in the above bound will become $\frac{1}{n}\sum_{i=1}^{n} K(x_i, x_i)$.