(1.) $k(x,x') = \exp(-\|x-x'\|^2/2\sigma^2)$
$= \exp(-x^Tx/2\sigma^2)\exp(x^Tx'/\sigma^2)\exp(-(x')^Tx'/2\sigma^2)$

$\star$ $\exp(x^Tx'/\sigma^2) = \sum_{i=1}^{\infty}\left(\frac{x^Tx'}{\sigma^2}\right)^i/i!$
$= \sum_{i=1}^{\infty}\left(\frac{x_1x_1'+x_2x_2'+\dots+x_nx_n'}{\sigma^2}\right)^i/i!$

Claim: Every term in $\exp(x^Tx'/\sigma^2)$ can be written as
$C_{a_1a_2\dots a_n}(x_1x_1')^{a_1}(x_2x_2')^{a_2}\dots(x_nx_n')^{a_n}$ with $C_{a_1\dots a_n}\in\mathbb{R}^+$.

Proof)

It suffices to show that every term in $(x_1x_1'+\dots+x_nx_n')^i$ has the form $K_{a_1\dots a_n}(x_1x_1')^{a_1}\dots(x_nx_n')^{a_n}$, because terms will have the same form after addition and scalar multiples. (with $K_{a_1\dots a_n}\in\mathbb{R}^+$)

By the binomial theorem, (multinomial)

$$(x_1x_1'+\dots+x_nx_n')^i$$
$$= \sum_{i=0}^{n}\sum_{j=0}^{n-i}\sum_{k=0}^{n-i-j}\dots(x_1x_1')^i(x_2x_2')^j(x_3x_3')^k\dots\underbrace{\frac{n!}{i!\,j!\,k!\dots}}_{\in\mathbb{R}^+}$$
$$= (x_1x_1')^n + n(x_1x_1')^{n-1}(x_2x_2') + \dots$$

$\Rightarrow$ This proves the result.

Thus, any term in $\exp(x^Tx'/\sigma^2)$, denoted by

$C_{a_1a_2\dots a_n}(x_1x_1')^{a_1}(x_2x_2')^{a_2}\dots(x_nx_n')^{a_n}$

is equal to $\hat{\phi}(x)\hat{\phi}(x')$

with $\hat{\phi}(x) = \sqrt{C_{a_1\dots a_n}}\,x_1^{a_1}x_2^{a_2}\dots x_n^{a_n}$ for any x.

∞-dim feature space with components $\hat{\phi}$

Thus, $\exp(x^Tx'/\sigma^2)$ can be written as $\phi(x)^T\phi(x')$, and therefore $k(x,x') = \exp(-x^Tx/2\sigma^2)\,\phi(x)^T\phi(x')\exp(-x'^Tx'/2\sigma^2)$
$= \alpha(x)^T\alpha(x')$
(with $\alpha(x) = \exp(-x^Tx/2\sigma^2)\phi(x)$)
= an inner product of an ∞-dimensional feature space defined by α.

Hilroy

(2.) Claim: For $t \geq 0$, w^t = a linear combination of the vectors $y_n \phi(x_n)$.

Proof) We proceed by induction.

Base case $t = 0$.

$w^0 = 0$, so the result trivially holds.

Assume the result holds for some t.

Then $w^{t+1} = w^t + y_n \phi(x_n) T(y_n w^{t^T} \phi(x_n))$ with $T(x) = \begin{cases} 1 & x \leq 0 \\ 0 & x > 0 \end{cases}$

$= \sum_{i=1}^{d} b_i y_i \phi(x_i) + y_n \phi(x_n) T(y_n w^{t^T} \phi(x_n))$

$= \sum_{i=1}^{d} c_i y_i \phi(x_i)$ with $c_i = \begin{cases} b_i & i \neq n \\ b_i + T(y_n w^{t^T} \phi(x_n)) & i = n \end{cases}$

$\Rightarrow w^{t+1}$ is a linear combination

$\Rightarrow$ The result holds for w^{t+1}.

$\Rightarrow$ Induction follows.

Thus, $\forall t \geq 0$, $w^t = \sum_{i=1}^{d} \alpha_i y_i \phi(x_i)$ $\quad \alpha_i \in \mathbb{R}$

$\Rightarrow w = \sum_{i=1}^{d} a_i y_i \phi(x_i)$ $\quad a_i \in \mathbb{R}$

$\Rightarrow$ w can be written as a linear combination of the $y_n \phi(x_n)$'s.

(a) Let d be the number of data points in the training set.

Initially, the linear combination is zero

$\Rightarrow a_i^0 = 0 \quad \forall 1 \leq i \leq d \qquad (a^0 = \vec{0})$.

When we consider a new data point $y_n \phi(x_n)$, the coefficient a_n gets updated by 1 if $y_n w^T \phi(x_n) \leq 0$ $\quad (w = \sum_{i=1}^{d} a_i^t y_i \phi(x_i))$.

Thus, the rule is as follows:

$a_i^0 = 0 \quad \forall 1 \leq i \leq d$.

When considering $y_n \phi(x_n)$,

$$a_i^{t+1} = \begin{cases} a_i^t & i \neq n \\ a_i^t + 1 & i = n \text{ and } y_n \left(\sum_{j=1}^{d} a_j^t y_j \phi^T(x_j) \right) \phi(x_n) \leq 0 \\ a_i^t & \text{otherwise} \end{cases}$$

$i = n$

The feature vector $\phi(x)$ enters in $y_n \left(\sum_{j=1}^{d} a_j^t y_j \phi^T(x_j) \right) \phi(x_n)$

$$= y_n \sum_{j=1}^{d} a_j^+ y_j \, \phi^T(x_j) \, \phi(x_n)$$

$$= y_n \sum_{j=1}^{d} a_j^+ y_j \, k(x_j, x_n)$$

$\Rightarrow \phi(x)$ only enters in the form of the kernel function

d) $$w^T \phi(x) > 0 \iff \left(\sum_{i=1}^{d} a_i y_i \phi(x_i) \right)^T \phi(x) > 0$$

$$\iff \sum_{i=1}^{d} a_i y_i \phi(x_i)^T \phi(x) > 0$$

$$\iff \sum_{i=1}^{d} a_i y_i k(x_i, x) > 0$$

Thus, the learning rule is

$$y = \begin{cases} 1 & \text{if} \quad \sum_{i=1}^{d} a_i y_i k(x_i, x) > 0 \\ -1 & \text{otherwise} \end{cases}$$