1. Since $A$ is OAROS and AERM we know the following relations hold

$$\mathbb{E}_{S\sim D^m}\big[L_D(A(S)) - L_S(A(S))\big] = \mathbb{E}_{(S,z')\sim P^{m+1},i\sim U[m]}\big[\tilde{\ell}(A(S^{(i)}); z_i) - \tilde{\ell}(A(S); z_i)\big] \leq \epsilon_1(m)$$
$$\mathbb{E}_{S\sim D^m}\big[L_S(A(S)) - \min_{h\in\mathcal{H}} L_S(h)\big] \leq \epsilon_2(m)$$

Combining the two inequalities, we get

$$\mathbb{E}_{S\sim D^m}\big[L_D(A(S)) - \min_{h\in\mathcal{H}} L_S(h)\big] \leq \epsilon_1(m) + \epsilon_2(m)$$

Note that

$$\mathbb{E}\big[\min_{h\in H} L_S(h)\big] = \mathbb{E}\big[\min_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \tilde{\ell}(h; z_i)\big] = \min_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \mathbb{E}\big[\tilde{\ell}(h; z_i)\big] = \min_{h\in H} L_D(h)$$

and so we can make a substitution in the last inequality

$$\mathbb{E}_{S\sim D^m}\big[L_D(A(S)) - \min_{h\in\mathcal{H}} L_D(h)\big] \leq \epsilon_1(m) + \epsilon_2(m)$$

showing that $A$ learns $\mathcal{H}$ at with rate $\epsilon_1(m) + \epsilon_2(m)$.

2. The symmetry of both functions is obvious. The first can be represented as an inner product as follows.

$$\begin{aligned}
k_1(x,z)k_2(x,z) &= \phi_1^T(x)\phi_1(z)\phi_2^T(x)\phi_2(z)\\
&= \sum_i \phi_{1i}(x)\phi_{1i}(z)(\phi_{21}(x)\phi_{21}(z) + \cdots + \phi_{2n}(x)\phi_{2n}(x))\\
&= \sum_{i,j} \phi_{1i}(x)\phi_{2j}(x)\phi_{1i}(z)\phi_{2j}(z) = \phi^T(x)\phi(z)
\end{aligned}$$

where

$$\phi(x) = (\phi_{11}(x)\phi_{21}(x), \ldots, \phi_{11}(x)\phi_{2n}(x), \phi_{12}(x)\phi_{21}(x), \ldots, \phi_{1i}(x)\phi_{2j}(x), \ldots, \phi_{1n}(x)\phi_{2n}(x))^T$$

By induction we can infer that powers, $h^k(x,z)$, of kernels are also kernels. Thus we can decompose the second expression into the following product

$$e^{h(x,z)} = \sum_{k=0}^{\infty} \frac{h^k(x,z)}{k!} = \sum_{k=0}^{\infty} \frac{\phi_k(x)^T\phi_k(z)}{k!} = \phi(x)^T\phi(z)$$

where

$$\phi(x) = \sum_{k=0}^{\infty} \frac{\phi_k(x)^T}{\sqrt{k!}}$$

and $\phi_k(x)$ is the map associated with kernel $h^k(x,z)$.

3. (a) Since $\{v_a\}$ and $\{\alpha^{(a)}\}$ are eigenvalues of $C$ and $K_0$ respectively we can write

$$CX\alpha^{(a)} = \frac{1}{N}XX^TX\alpha^{(a)} = X(\frac{1}{X}^T X)\alpha^{(a)} = XK_0\alpha^{(a)} = \sigma_a^2 X\alpha^{(a)}$$

This implies that $X\alpha^{(a)}$ is an eigenvector of $C$ with corresponding eigenvalue $\sigma_a^2$. In other words,

$$v_a = \sum_i \alpha_i^{(a)} x_i.$$

(b) Applying the transform $\phi$ and performing PCA on the new space, the objective becomes to express $\phi(x_{\text{test}})$ in terms of the left singular vectors of $\frac{1}{\sqrt{N}}\Phi$, which themselves form an orthonormal basis fort the new space. This can be done by computing the vector

$$\bar{x}_{\text{test}} = [\phi(x_{\text{test}})^T v_1, \dots, \phi(x_{\text{test}})^T v_d]^T$$

Given that we've already demonstrated $v_k = \sum_i \beta_i^{(k)} \phi(x_i)$, we can rewrite $\bar{x}_{\text{test}}$ as

$$\begin{aligned}\bar{x}_{\text{test}} &= \big[\sum_i \beta_i^{(1)} \phi(x_{\text{test}})^T \phi(x_i), \dots, \sum_i \beta_i^{(d)} \phi(x_{\text{test}})^T \phi(x_i)\big]^T \\ &= \big[\sum_i \beta_i^{(1)} K(x_i, x_{\text{test}}), \dots, \sum_i \beta_i^{(d)} K(x_i, x_{\text{test}}\big]^T\end{aligned}$$

4. (a) SVM's over a spearable set of points satisfy Slater's condition in general. Thus strong duality holds and we can find an optimal $w$ by solving the dual problem. In the case of hard-SVM over the class of homogeneous hyperplanes, the corresponding optimization problem is

$$\begin{aligned}&\min \frac{1}{2}\|w\|_2^2 \\ &\text{s.t. } 1 - y_i w^T x_i \leq 0\end{aligned}$$

and the associated Lagrangian is

$$L = \frac{1}{2}\|w\|_2^2 + \sum_i \alpha_i (1 - y_i w^T x_i)$$

$L$ is is a convex, differentiable function of $w$ so we can find its minimum with respect to $w$ by setting its gradient to zero. Thus

$$\nabla_w L = w - \sum_i \alpha_i x_i = 0 \implies w - \sum_i \alpha_i x_i$$

and we see that the optimal $w$ can always be expressed as a linear combination of the support vectors $\{x_i\}_{i \in \mathcal{I}}$ for which $\alpha_i$ may be non-zero.

(b) Removing samples does not change which points $x_i$ are support vectors. The dual problem involves a maximization over the constrianed vector $\alpha$. Removing points $x_i$ for which $\alpha_i = 0$ at the optimal point will not change the elements of $\alpha$ associated with the remaining points. Thus $w = \sum_i \alpha_i x_i$ will remain the same.

(c) An anologous argument holds in the soft-SVM case. The initial Lagrangian is slightly different

$$L = \frac{1}{2}\|w\|_2^2 + \sum_i \alpha_i (1 - y_i w^T x_i - \zeta_i) - \sum_i \lambda_i \zeta_i$$

but it's graient with respect to $w$ remains the same, meaning $w = \sum_i \alpha_i x_i$. Furthermore, removing points associated with inactive constraints will not change the non-zero elements of the optimal vector $\alpha*$. Thus $w$ is not affected by the removal of non-supporting vectors.

(d) In the dual problem of the hard-SVM case, the only constraint on $\alpha$ is that each element must be greater than zero. In the soft case, we also have that $0 \leq \alpha_i \leq C$. Noting that the soft objective function can be re-written as $\frac{1}{2}\|w\|_2^2 + \frac{1}{2\lambda m}\sum_i \zeta_i$, this must mean $0 \leq \alpha_i \leq \frac{1}{2\lambda m}$.

Given a finite solution for each $\alpha_i$ in the hard case, let $\alpha_{\max} = \max_i\{\alpha_i\}_{i\in\mathcal{I}}$. If we set $\lambda \leq \frac{1}{2\alpha_{\max} m}$ then the extra constraint on $\alpha$ won't prevent the soft-SVM problem from selecting the same optimal vector $\alpha^*$ as in the hard case.

However, this is only for a specific sample set $\mathcal{D}$. In the more general case, where we don't know $D$ ahead of time, for fixed $\lambda$ we can alwasy pick a sample $\mathcal{D}$ which induces some $\alpha_i$ that violates our constraint on $C$. So no, the learning rules can, but will not always, return the same weight vector.

5. We'll prove the first result by induction.

**Base Case:** $(x_1y_1 + c) = (x_1, \sqrt{c})(y_1, \sqrt{c})^T$. Here we see that the feature space associated to $K$ is two-dimensional, since $\phi(x) = (x, \sqrt{c})^T$. This agrees with the formula $\mathrm{Dim}(\phi(x)) = \binom{1+1}{1} = 2$.

**Induction 1:** Assume that the kernal $K_d$ associated with $(x^Ty + c)^d$ maps to a space of dimension $\binom{N+d}{d}$. We can decompose $(x^Ty + c)^{d+1}$ as

$$(x^Ty + c)^{d+1} = (x^Ty + c)^d(x^Ty + c) = \sum_i x_iy_i\phi_d(x)^T\phi_d(x) + \sqrt{c}\sqrt{c}\phi_d(x)^T\phi_d(y).$$

Appealing of the multinomial choice theorem, $\phi_{d+1}$ associated with $K_{d+1}$ maps into a feature space of size $\binom{N+d+1}{d+1}$

**Induction 2:** Assume that the kernal $K_d$ associated with $(x^Ty + c)^d$ maps to a space of dimension $\binom{N+d}{d}$. We can decompose $(x^Ty + c + x_{N+1}y_{N+1})^d$ as

$$\begin{aligned}(x^Ty + c + x_{N+1}y_{N+1})^d &= \sum_{i=0}^{d}\binom{d}{i}(x^Ty + c)^{d-i}(x_{N+1}y_{N+1})^i \\ &= \sum_{i=0}^{d}\binom{d}{i}\phi_{d-i}(x)^T\phi_{d-i}(y)(x_{N+1}y_{N+1})^i\end{aligned}$$

Appealing of the multinomial choice theorem, $\phi_{d+1}$ associated with $K_{d+1}$ maps into a feature space of size $\binom{N+d+1}{d}$

$K$ can be written in terms of kernels as

$$K(x,y) = \sum_{i=0}^{d}\binom{d}{i}(x^Ty)^ic^{d-i} = \sum_{i=0}^{d}\binom{d}{i}c^{d-i}k_i$$

the coefficient of each $k-i$ being $\binom{d}{i}c^{d-i}$. Each coefficient is then proportional to the $i^{\text{th}}$ power of $c$.

6. (a) Documents $D_1$ and $D_2$ contain $(m-k+1)$ and $(n-k+1)$ words of size $k$ respectively. Filling each vector $\phi(D_i)$ has complexity on the order of the number of words in the largest document.

(b) Taking the inner product has complexity on the order of $|\Sigma|^k$.

(c) If we manually compare each word in $D_1$ to every work in $D_2$, keeping a sum of matches, it will take $(m-k+1)(n-k+1)$ steps. We could also say it's on the order of the square of the number of words in the largest document.

(d) Assuming that $|\Sigma|^k$ is way larger than the work count in either dictionary (which is likely) the algorithm in (c) is more efficient.