A. Proof of Theorem 3

Although $f^*_{\mathcal{P}}$ is a feasible solution, it is not a local optimum for $\theta \in [0, 1)$ and $s \le 0$ because

$$\alpha_i \le C\theta \quad \text{for} \quad i \in \tilde{\mathcal{I}} \cap \mathcal{O}, \tag{12a}$$
$$\alpha_i \ge C \quad \text{for} \quad i \in \tilde{\mathcal{O}} \cap \mathcal{I}, \tag{12b}$$

violate the KKT conditions (7) for $\tilde{\mathcal{P}}$. These feasibility and *sub*-optimality indicates that

$$J_{\tilde{\mathcal{P}}}(f^*_{\tilde{\mathcal{P}}}; \theta) < J_{\mathcal{P}}(f^*_{\mathcal{P}}; \theta), \tag{13}$$

we arrive at (9). **Q.E.D.**

B. Proof of Theorem 4

Sufficiency: If (10e) is true, i.e., if there are NO instances with $y_i f^*_{\mathcal{P}}(\boldsymbol{x}_i) = s$, then any convex problems defined by different partitions $\tilde{\mathcal{P}} \neq \mathcal{P}$ do not have feasible solutions in the neighborhood of $f^*_{\mathcal{P}}$. This means that if $f^*_{\mathcal{P}}$ is a conditionally optimal solution, then it is locally optimal. (10a)-(10d) are sufficient for $f^*_{\mathcal{P}}$ to be conditionally optimal for the given partition $\mathcal{P}$. Thus, (10) is sufficient for $f^*_{\mathcal{P}}$ to be locally optimal.

Necessity: From Theorem 3, if there exists an instance such that $y_i f^*_{\mathcal{P}}(\boldsymbol{x}_i) = s$, then $f^*_{\mathcal{P}}$ is a feasible but not locally optimal. Then (10e) is necessary for $f^*_{\mathcal{P}}$ to be locally optimal. In addition, (10a)-(10d) are also necessary for local optimality, because of every local optimal solutions are conditionally optimal for the given partition $\mathcal{P}$. Thus, (10) is necessary for $f^*_{\mathcal{P}}$ to be locally optimal.

Q.E.D.

C. Implementation of D-step

In D-step, we work with the following convex problem

$$f^*_{\tilde{\mathcal{P}}} := \underset{f \in \mathrm{pol}(\tilde{\mathcal{P}}; s)}{\operatorname{argmin}} J_{\tilde{\mathcal{P}}}(f; \theta). \tag{14}$$

where, $\tilde{\mathcal{P}}$ is updated from $\mathcal{P}$ as (8).

Let us define a partition $\Pi := \{\mathcal{R}, \mathcal{E}, \mathcal{L}, \tilde{\mathcal{I}}', \tilde{\mathcal{O}}', \hat{\mathcal{O}}''\}$ of $\mathbb{N}_n$ such that

$$i \in \mathcal{R} \;\Rightarrow\; y_i f(\boldsymbol{x}_i) > 1, \tag{15a}$$
$$i \in \mathcal{E} \;\Rightarrow\; y_i f(\boldsymbol{x}_i) = 1, \tag{15b}$$
$$i \in \mathcal{L} \;\Rightarrow\; s < y_i f(\boldsymbol{x}_i) < 1, \tag{15c}$$
$$i \in \tilde{\mathcal{I}}' \;\Rightarrow\; y_i f(\boldsymbol{x}_i) = s \text{ and } i \in \tilde{\mathcal{I}}, \tag{15d}$$
$$i \in \tilde{\mathcal{O}}' \;\Rightarrow\; y_i f(\boldsymbol{x}_i) = s \text{ and } i \in \tilde{\mathcal{O}}, \tag{15e}$$
$$i \in \tilde{\mathcal{O}}'' \;\Rightarrow\; y_i f(\boldsymbol{x}_i) < s. \tag{15f}$$

If we write the conditionally optimal solution as

$$f^*_{\tilde{\mathcal{P}}}(x) := \sum_{j \in \mathbb{N}_n} \alpha^*_j y_j K(x, \boldsymbol{x}_j), \tag{16}$$

$\{\alpha^*_j\}_{j \in \mathbb{N}_n}$ must satisfy the following KKT conditions

$$y_i f^*_{\tilde{\mathcal{P}}}(\boldsymbol{x}_i) > 1 \;\Rightarrow\; \alpha^*_i = 0 \tag{17a}$$
$$y_i f^*_{\tilde{\mathcal{P}}}(\boldsymbol{x}_i) = 1 \;\Rightarrow\; \alpha^*_i \in [0, C], \tag{17b}$$
$$s < y_i f^*_{\tilde{\mathcal{P}}}(\boldsymbol{x}_i) < 1 \;\Rightarrow\; \alpha^*_i = C \tag{17c}$$
$$y_i f^*_{\tilde{\mathcal{P}}}(\boldsymbol{x}_i) = s, i \in \tilde{\mathcal{I}}' \;\Rightarrow\; \alpha^*_i \ge C, \tag{17d}$$
$$y_i f^*_{\tilde{\mathcal{P}}}(\boldsymbol{x}_i) = s, i \in \tilde{\mathcal{O}}' \;\Rightarrow\; \alpha^*_i \le C\theta, \tag{17e}$$
$$y_i f^*_{\tilde{\mathcal{P}}}(\boldsymbol{x}_i) < s, i \in \tilde{\mathcal{O}}'' \;\Rightarrow\; \alpha^*_i = C\theta. \tag{17f}$$

At the beginning of the D-step, $f^*_{\tilde{\mathcal{P}}}(\boldsymbol{x}_i)$ violates the KKT conditions by

$$\Delta f_i := y_i \begin{bmatrix} \boldsymbol{K}_{i,\Delta_{\mathcal{I}\to\mathcal{O}}} & \boldsymbol{K}_{i,\Delta_{\mathcal{O}\to\mathcal{I}}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}^{(\text{bef})}_{\Delta_{\mathcal{I}\to\mathcal{O}}} - \mathbf{1}C\theta \\ \boldsymbol{\alpha}^{(\text{bef})}_{\Delta_{\mathcal{O}\to\mathcal{I}}} - \mathbf{1}C \end{bmatrix}.$$

where $\boldsymbol{\alpha}^{(\text{bef})}$ is the corresponding $\boldsymbol{\alpha}$ at the beginning of the D-step, while $\Delta_{\mathcal{I}\to\mathcal{O}}$ and $\Delta_{\mathcal{O}\to\mathcal{I}}$ denote the difference in $\tilde{\mathcal{P}}$ and $\mathcal{P}$ defined as

$$\Delta_{\mathcal{I}\to\mathcal{O}} := \{i \in \mathcal{I} \mid y_i f_{\mathcal{P}}(\boldsymbol{x}_i) = s\},$$
$$\Delta_{\mathcal{O}\to\mathcal{I}} := \{i \in \mathcal{O} \mid y_i f_{\mathcal{P}}(\boldsymbol{x}_i) = s\}.$$

Then, we consider the following another parametrized problem with a parameter $\mu \in [0, 1]$:

$$f_{\tilde{\mathcal{P}}}(\boldsymbol{x}_i; \mu) := f_{\tilde{\mathcal{P}}}(\boldsymbol{x}_i) + \mu \Delta f_i \;\forall\, i \in \mathbb{N}_n.$$

In order to always satisfy the KKT conditions for $f_{\tilde{\mathcal{P}}}(\boldsymbol{x}_i; \mu)$, we solve the following linear system

$$\boldsymbol{Q}_{\mathcal{A},\mathcal{A}} \begin{bmatrix} \boldsymbol{\alpha}_{\mathcal{E}} \\ \boldsymbol{\alpha}_{\tilde{\mathcal{I}}'} \\ \boldsymbol{\alpha}_{\tilde{\mathcal{O}}'} \end{bmatrix} = \begin{bmatrix} \mathbf{1} \\ \boldsymbol{s} \\ \boldsymbol{s} \end{bmatrix} - \boldsymbol{Q}_{\mathcal{A},\mathcal{L}} \mathbf{1}C - \boldsymbol{Q}_{\mathcal{A},\tilde{\mathcal{O}}''} \mathbf{1}C\theta$$
$$- \begin{bmatrix} \boldsymbol{Q}_{\mathcal{A},\Delta_{\mathcal{I}\to\mathcal{O}}} & \boldsymbol{Q}_{\mathcal{A},\Delta_{\mathcal{O}\to\mathcal{I}}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}^{(\text{bef})}_{\Delta_{\mathcal{I}\to\mathcal{O}}} - \mathbf{1}C\theta \\ \boldsymbol{\alpha}^{(\text{bef})}_{\Delta_{\mathcal{O}\to\mathcal{I}}} - \mathbf{1}C \end{bmatrix} \mu,$$

where $\mathcal{A} := \{\mathcal{E}, \tilde{\mathcal{I}}', \tilde{\mathcal{O}}'\}$. This linear system can also be solved by using the piecewise-linear parametric programming while the scalar parameter μ is continuously moved from 1 to 0.

In this parametric problem, we can show that $f^*_{\tilde{\mathcal{P}}}(\boldsymbol{x}_i; \mu) = f^*_{\mathcal{P}}(\boldsymbol{x}_i)$ if $\mu = 1$ and $f^*_{\tilde{\mathcal{P}}}(\boldsymbol{x}_i; \mu) = f^*_{\tilde{\mathcal{P}}}(\boldsymbol{x}_i)$ if $\mu = 0$ for all $i \in \mathbb{N}_n$.

Since the number of elements in $\Delta_{\mathcal{I}\to\mathcal{O}}$ and $\Delta_{\mathcal{O}\to\mathcal{I}}$ are typically small, the D-step can be efficiently implemented by a technique used in the context of incremental learning (Cauwenberghs & Poggio, 2001).