A. Proof of Theorem 3

Although \( f_p \) is a feasible solution, it is not a local optimum for \( \theta \in [0, 1] \) and \( s \leq 0 \) because
\[
\begin{align*}
\alpha_i &\leq C \theta & \text{for } i \in \tilde{I} \cap O, \\
\alpha_i &\geq C & \text{for } i \in \tilde{O} \cap I,
\end{align*}
\]
violate the KKT conditions (7) for \( \tilde{P} \). These feasibility and sub-optimality indicates that
\[
J_p(f_p; \theta) < J_p(f^*_p; \theta),
\]
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 \( y_i f_p(x_i) = s \), then any convex problems defined by different partitions \( \tilde{P} \neq P \) do not have feasible solutions in the neighborhood of \( f_p \). This means that if \( f_p \) is a conditionally optimal solution, then it is locally optimal. (10a)-(10d) are sufficient for \( f_p \) to be conditionally optimal for the given partition \( P \). Thus, (10) is sufficient for \( f_p \) to be locally optimal.

Necessity: From Theorem 3, if there exists an instance such that \( y_i f_p(x_i) = s \), then \( f_p \) is a feasible but not locally optimal. Then (10e) is necessary for \( f_p \) to be locally optimal. In addition, (10a)-(10d) also necessary for local optimality, because of every local optimal solutions are conditionally optimal for the given partition \( P \). Thus, (10) is necessary for \( f_p \) to be locally optimal.
Q.E.D.

C. Implementation of D-step

In D-step, we work with the following convex problem
\[
f_p := \arg\min_{f_p \in \mathcal{P}(\mathcal{P})} J_p(f; \theta),
\]
where \( \tilde{P} \) is updated from \( P \) as (8).

Let us define a partition \( \Pi := \langle \mathcal{R}, \mathcal{E}, \mathcal{L}, \tilde{I}, \tilde{O}, \tilde{O}' \rangle \) of \( \mathbb{N} \) such that
\[
\begin{align*}
\mathcal{R} &\Rightarrow y_i f(x_i) > 1, \\
\mathcal{E} &\Rightarrow y_i f(x_i) = 1, \\
\mathcal{L} &\Rightarrow s < y_i f(x_i) < 1, \\
\tilde{I} &\Rightarrow y_i f(x_i) = s \quad \text{and } i \in \tilde{I}, \\
\tilde{O} &\Rightarrow y_i f(x_i) = s \quad \text{and } i \in \tilde{O}, \\
\tilde{O}' &\Rightarrow y_i f(x_i) < s.
\end{align*}
\]

If we write the conditionally optimal solution as
\[
f_p(x) := \sum_{j \in \mathbb{N}} \alpha_j f_j K(s, x),
\]
[\( \alpha_j \)] \( j \in \mathbb{N} \) must satisfy the following KKT conditions
\[
\begin{align*}
y_i f_p^*(x_i) > 1 &\Rightarrow \alpha_j = 0 \quad (17a) \\
y_i f_p^*(x_i) = 1 &\Rightarrow \alpha_j \in [0, C], \quad (17b) \\
\alpha_i < y_i f_p^*(x_i) < 1 &\Rightarrow \alpha_j = C \quad (17c) \\
y_i f_p^*(x_i) = s, i \in \tilde{I} &\Rightarrow \alpha_j \geq C, \quad (17d) \\
y_i f_p^*(x_i) = s, i \in \tilde{O} &\Rightarrow \alpha_j \leq C \theta, \quad (17e) \\
y_i f_p^*(x_i) < s, i \in \tilde{O}' &\Rightarrow \alpha_j = C \theta. \quad (17f)
\end{align*}
\]

At the beginning of the D-step, \( f_p^*(x_i) \) violates the KKT conditions by
\[
\Delta_f := y_i \begin{bmatrix} K_{I, \Delta_{\sim \theta}} & K_{I, \Delta_{\sim \theta}} \end{bmatrix} \begin{bmatrix} \alpha_{\text{opt}} & \alpha_{\text{eff}} - 1C \theta \end{bmatrix},
\]
where \( \alpha_{\text{opt}} \) is the corresponding \( \alpha \) at the beginning of the D-step, while \( \Delta_{I, \theta} \) and \( \Delta_{O, \theta} \) denote the difference in \( \tilde{P} \) and \( P \) defined as
\[
\Delta_{I,\theta} := \{ i \in \mathcal{I} | y_i f_p(x_i) = s \}, \quad \Delta_{O,\theta} := \{ i \in \mathcal{O} | y_i f_p(x_i) = s \}.
\]

Then, we consider the following another parametrized problem with a parameter \( \mu \in [0, 1] \):
\[
f_p(x_i; \mu) := f_p(x_i) + \mu \Delta_f \quad \forall i \in \mathbb{N}.
\]

In order to always satisfy the KKT conditions for \( f_p(x_i; \mu) \), we solve the following linear system
\[
\mathcal{A}(\mathcal{A}) \begin{bmatrix} \alpha_E \\ \alpha_I \\ \alpha_O \end{bmatrix} = \begin{bmatrix} 1 \\ s \end{bmatrix} - Q_{\mathcal{A}, \mathcal{A}} IC - Q_{\mathcal{A}, \Delta_{\sim \theta}} IC \\
- Q_{\mathcal{A}, \Delta_{\sim \theta}} Q_{\mathcal{A}, \Delta_{\sim \theta}} \begin{bmatrix} \alpha_{\text{opt}} & \alpha_{\text{eff}} - 1C \theta \end{bmatrix} \mu,
\]
where \( \mathcal{A} := \{ E, I, O \} \). This linear system can also be solved by using the piecewise-linear parametric programming while the scalar parameter \( \mu \) is continuously moved from 1 to 0.

In this parametric problem, we can show that \( f_p(x_i; \mu) = f_p(x_i) \) if \( \mu = 1 \) and \( f_p(x_i; \mu) = f_p(x_i) \) if \( \mu = 0 \) for all \( i \in \mathbb{N} \).

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