A. Proof of Lemma 3 and Theorem 5

First, the KKT optimality conditions of a conditionally optimal solution is written as follows:

Lemma 6 *For a given $\hat{y} \in \{-1, 1\}^{|\mathcal{U}|}$, the necessary and sufficient conditions for a f to be the optimal solution of the convex problem (5) is (8) and*

$$\hat{y}_i \alpha_i \geq C^* \text{ for } i \in \{i \in \mathcal{U} | \hat{y}_i f(x_i) = 0\}. \tag{13}$$

We omit the proof of this lemma because they are straightforwardly derived by using Lagrange multiplier theory (Boyd & Vandenberghe, 2004). Here, we just note that the derivation is almost same as the standard SVM case because the predicted labels $\hat{y}$ are fixed here.

Based on Lemma 6, we first prove Theorem 5.

Proof of Theorem 5 Let $f^*_{\hat{y}}$ and $f^*_{\hat{y}'}$ be two conditionally optimal solutions defined in $\mathrm{pol}(\hat{y})$ and $\mathrm{pol}(\hat{y}')$, respectively, and consider a situation that the former $f^*_{\hat{y}}$ is at a boundary of $\mathrm{pol}(\hat{y})$. To prove the theorem, we suppose for the moment that it is also conditionally optimal in the next polytope $\mathrm{pol}(\hat{y}')$, i.e., $f^*_{\hat{y}} = f^*_{\hat{y}'}$.

Since $f^*_{\hat{y}}$ and $f^*_{\hat{y}'}$ are conditionally optimal, they satisfy the optimality condition (13):

$$\hat{y}_i \alpha_i \geq C^* \text{ for } i \in \{i \in \mathcal{U} | \hat{y}_i f^*_{\hat{y}}(x_i) = 0\}, \tag{14}$$

and

$$\hat{y}'_i \alpha_i \geq C^* \text{ for } i \in \{i \in \mathcal{U} | \hat{y}'_i f^*_{\hat{y}'}(x_i) = 0\}, \tag{15}$$

respectively. From our current assumption that $f^*_{\hat{y}} = f^*_{\hat{y}'}$ and the fact that

$$y'_i = -y_i, i \in \{i \in \mathcal{U} | y_i f^*_{\hat{y}}(x_i) = 0\}, \tag{16}$$

(14) is rewritten as

$$\hat{y}'_i \alpha_i \leq -C^* \text{ for } i \in \{i \in \mathcal{U} | \hat{y}'_i f^*_{\hat{y}'}(x_i) = 0\}. \tag{17}$$

Now, it is clear that the two conditions (15) and (17) cannot be satisfied at the same time, and it disprove our assumption that $f^*_{\hat{y}} = f^*_{\hat{y}'}$.

Noting that $f^*_{\hat{y}} \in \mathrm{pol}(\hat{y}')$ and that it is not the conditionally optimal solution in $\mathrm{pol}(\hat{y}')$, we immediately arrive at the conclusion that $f^*_{\hat{y}'}$ is a better S^{3}VM solutions than $f^*_{\hat{y}}$. **Q.E.D.**

Next, we prove Lemma 3, which is immediately obtained from Lemma 6 and Theorem 5.

Proof of Lemma 3 First, if the conditionally optimal solution $f^*_{\hat{y}}$ is in the strict interior of the convex polytope $\mathrm{pol}(\hat{y})$, it is clear that there is no better solution in the arbitrary neighborhood of $f^*_{\hat{y}}$. It suggests that $f^*_{\hat{y}}$ is a local optimal solution of S^{3}VM is it is in the strict interior of $\mathrm{pol}(\hat{y})$. On the other hand, from Theorem 5, $f^*_{\hat{y}}$ is not a local optimal solution of S^{3}VM because there exists a strictly better solution in the adjacent convex polytope $\mathrm{pol}(\hat{y}')$. Combining the fact that f is conditionally optimal if and only if (8) and (13) are satisfied, it is clear that (8) and (9) are the necessary and sufficient conditions of a local optimal solution. **Q.E.D.**

B. Computational Complexity of S^{3}VM Algorithm

The computational cost of the entire algorithm (from $C^* = 0$ to C) depends on the number of so-called *breakpoints* in the CP-step and the number of movements to adjacent polytopes in the DJ-step. It has been reported in many empirical studies (Efron & Tibshirani, 2004; Hastie et al., 2004) that the number of breakpoints is $\mathcal{O}(n)$, where n is the training set size. We also observed in our experiments that the total number of breakpoints in all CP steps scales almost linearly with respect to $\mathcal{O}(|\mathcal{L}| + |\mathcal{U}|)$.

The main computational cost in each breakpoint is the same as that in the SVM regularization path (Hastie et al., 2004). That is, at each breakpoint, we need to solve a rank-one update problem of a linear system of equations of size $O(|\mathcal{M}|)$, which costs $\mathcal{O}(|\mathcal{M}|^2)$. On the other hand, the number of movements between two polytopes depends on the number of unlabeled instances. In our experience, this number also scales linearly with respect to $\mathcal{O}(|\mathcal{U}|)$. Note that, if we use a warm-start strategy from the previous conditionally optimal solution, the computational cost of the DJ-step is negligibly small compared with the CP-step.