Due to the page limit, in the main paper, we omitted some details of Algorithms 1 and 2, and all the proof of the main conclusions. In this supplementary file, we first present the omitted details of Algorithms 1 and 2, and then present the derivations of the main conclusions in the paper.

A. More Details about Algorithm 1 and Algorithm 2

For convenience, we first present the stopping conditions of Algorithm 2.

A.1. Stopping Condition of Algorithm 2

It is usually non-trivial to set a proper stopping condition for stochastic optimization algorithms. Usually, we can stop an algorithm when the objective value does not change significantly. For example, we can stop Algorithm 2 if the primal objective value cannot decrease significantly. Unfortunately, computing the primal objective value $f_h$ is very expensive. Moreover, the primal objective value does not monotonically decrease w.r.t. $h$. Therefore, we propose to stop Algorithm 2 if $h > 5$ and

$$\frac{|f_h - f_{h-5}|}{f_{h-5}} \leq \epsilon.$$ 

Here, $f_h$ is computed as in Algorithm 2, and it approximates the primal objective value of the SSVM subproblem. In our implementation, we choose and fix $\epsilon = 0.005$.

A.2. Inequality Constraint Handling in Subproblem Optimization

Note that the conjugate dual of the subproblem in (12) is

$$\max_{\alpha} -\lambda \Omega^* \left( \frac{1}{\lambda n} \sum_{i=1}^{n} \sum_{y \neq y_i} \alpha_{iy} w^\top \Phi_{d,y_i,y}(x_i) \right) - \frac{1}{n} \sum_{i=1}^{n} L^*_i (-\alpha_{iy}),$$

subject to

$$\sum_{y \neq y_i} \alpha_{iy} \leq 1, \forall i \in [n].$$

where $\alpha_{iy} = [\alpha_{iy}]_{y \neq y_i}$ and $L^*_i$ denotes the conjugate of the loss function $L_i$. Note that in (15), we have an inequality constraint $\sum_{y \neq y_i} \alpha_{iy} \leq 1$ on $\alpha_{iy}$.

In Algorithm 2, we do not store $\alpha_{iy}$ explicitly, thus we cannot handle the inequality constraint directly. When the inequality constraint is ignored, the update rule $\delta_{iy} = \frac{\lambda n (\alpha_{iy} - d)}{(a^2 + \nu)}$ may be too aggressive (e.g. $\delta_{iy}$ may be too large). To address this, we use a scaled update rule $\delta_{iy} = \frac{\lambda n (\alpha_{iy} - d)}{\theta (a^2 + \nu)}$ instead, where $\theta > 1$. In our implementation, we initialize $\theta = 2$ and update $\theta := 2\theta$ when $f_h$ does not decrease (See Algorithm 2).

A.3. Stopping Condition of Algorithm 1

Similar in Algorithm 2, we stop Algorithm 1 when the primal objective value does not decrease significantly. Let $f^t = f_h$, where $f_h$ is the approximated primal objective value obtained from Algorithm 2. Then, we stop Algorithm 1 if $t > 2$ and

$$\frac{|f^t - f^{t-2}|}{f_{t-2}} \leq \epsilon_o.$$ 

In our implementation, we choose and fix $\epsilon_o = 0.05$.  

978-1-4673-6964-0/15/$31.00 ©2015 IEEE
B. Lagrangian Dual in (4) of Problem (3)

Proof. Note that $\Phi^V_{y,i,y}(x_i) := \Psi^V_{y}(y_i, x_i) - \Psi^V_{y}(y, x_i)$, $\Phi^C_{y,i,y}(x_i) := \Psi^C_{y}(y_i, x_i; \eta) - \Psi^C_{y}(y, x_i; \eta)$.

The Lagrangian function of the inner minimization problem in (3) can be written as:

$$\mathcal{L}(w, \xi, \alpha, \beta) = \frac{\lambda}{2} \|w\|^2 + \frac{1}{n} \sum_{i=1}^{n} \xi_i - \beta^\top \xi + \sum_{i=1}^{n} \sum_{y \neq y_i} \alpha_{iy} (\Delta(y, y_i) - u^\top \Phi^V_{y,i,y}(x_i) - v^\top \Phi^C_{y,i,y}(x_i; \eta) - \xi_i).$$  \hspace{1cm} (16)

Let $\alpha := [\alpha_{iy}, ..., \alpha_{ny}]^\top$. The KKT condition of (16) can be written as

$$\begin{align*}
\frac{\partial \mathcal{L}(w, \xi, \alpha, \beta)}{\partial u} = 0 & \Rightarrow u = \frac{1}{\lambda} \sum_{i=1}^{n} \sum_{y \neq y_i} \alpha_{iy} \Phi^V_{y,i,y}(x_i); \\
\frac{\partial \mathcal{L}(w, \xi, \alpha, \beta)}{\partial v} = 0 & \Rightarrow v = \frac{1}{\lambda} \sum_{i=1}^{n} \sum_{y \neq y_i} \alpha_{iy} \Phi^C_{y,i,y}(x_i; \eta); \\
\frac{\partial \mathcal{L}(w, \xi, \alpha, \beta)}{\partial \xi_i} = 0 & \Rightarrow \frac{1}{n} = \sum_{y \neq y_i} \alpha_{iy} + \beta_i; \\
\alpha \succeq 0, \text{ and } \beta \succeq 0.
\end{align*}$$  \hspace{1cm} (17-19)

Let $\mathcal{A} = \{\alpha \in \mathbb{R}^l | \alpha \succeq 0, \sum_{y \neq y_i} \alpha_{iy} \leq \frac{1}{n}\}$ be the domain of $\alpha$. Define

$$u(\alpha) := \sum_{i=1}^{n} \sum_{y \neq y_i} \alpha_{iy} \Phi^V_{y,i,y}(x_i) \text{ and } v(\alpha, \eta) := \sum_{i=1}^{n} \sum_{y \neq y_i} \alpha_{iy} \Phi^C_{y,i,y}(x_i; \eta).$$  \hspace{1cm} (20)

Substituting the above relations into (16), the Lagrangian dual of the inner problem of (3) can be written as

$$\max_{\alpha \in \mathcal{A}} - \frac{1}{2\lambda} \|u(\alpha)\|^2 - \frac{1}{2\lambda} \|v(\alpha, \eta)\|^2 + b^\top \alpha.$$  \hspace{1cm} (21)

C. Proof of Theorem 1

The proof of Theorem 1 can be adapted from the proof of Theorem 2 in [34].

D. Proof of Proposition 1

Proof. Let $\Omega(\omega) = \frac{1}{2} (\sum_{k=1}^{t} \|\omega_k\|^2)^2$. Define a cone $Q_r = \{(u, v) \in \mathbb{R}^{r+1}, \|u\|_2 \leq v\}$. Let $z_k = \|\omega_k\|$, we have $\Omega(v) = \frac{1}{2} (\sum_{k=1}^{t} \|\omega_k\|)^2 = \frac{1}{2} z^2$, where $z = \sum_{k=1}^{t} z_k$, $z_k \geq 0$ and $z \geq 0$. Then, problem (10) can be transformed to the following problem:

$$\begin{align*}
\min_{z,u,v} & \frac{\lambda}{2} \|u\|^2 + \frac{\lambda}{2} z^2 + \frac{1}{n} \sum_{i=1}^{n} \xi_i, \hspace{0.5cm} \text{s.t.} \hspace{0.5cm} \sum_{k=1}^{t} z_k \leq z, \hspace{0.5cm} (\omega_k, z_k) \in Q_r, \\
w^\top \Phi^V_{y,i,y}(x_i) - \xi_i & \geq 0 \hspace{0.5cm} \forall i, \forall y \in \mathcal{Y} \setminus y_i.
\end{align*}$$

where $\omega = [\omega_1', ..., \omega_t']$. The Lagrangian function of (23) can be written as:

$$L(z, v, \xi, h, \gamma, \zeta, \omega) = \frac{\lambda}{2} \|u\|^2 + \frac{\lambda}{2} z^2 + \frac{1}{n} \sum_{i=1}^{n} \xi_i + \gamma \sum_{k=1}^{t} z_k - z - \sum_{k=1}^{t} \zeta_k (\omega_k + \omega_k z_k)$$

$$- \frac{n}{\sum_{i=1}^{n} \sum_{y \neq y_i} \xi_i} \left( \Delta(y, y_i) - \xi_i - \left( \Phi^V_{y,i,y}(x_i) + \frac{t}{\sum_{k=1}^{t} \omega_k \Phi^V_{y,i,y}(x_i)} \right) \right) - \beta^\top \xi.$$
where $\alpha, \gamma, \zeta_k$ and $w_k$ are the Lagrangian dual variables to the corresponding constraints. The KKT condition can be expressed as

\begin{align*}
\nabla_z L &= \lambda z - \gamma = 0 \\
\nabla_u L &= \gamma - w_k = 0 \\
\nabla_u L &= \lambda u + \sum_{j=1}^{n} \sum_{y \neq y_i} \alpha_{iy} \Phi_{y, y_i}(x_i) \\
\nabla_{\omega_k} L &= -\sum_{j=1}^{n} \sum_{y \neq y_i} \alpha_{iy} \left( \Phi_{y, y_i}(x_i) \right) - \zeta_k = 0 \\
\n\|\zeta_k\| &\leq w_k \\
\beta_i &\geq 0
\end{align*}

\[\Rightarrow z = \frac{\gamma}{\lambda}; \quad w_k = \gamma; \quad u = -\frac{1}{\lambda} \sum_{j=1}^{n} \sum_{y \neq y_i} \alpha_{iy} \Phi_{y, y_i}(x_i) \]

\[\Rightarrow \zeta_k = -\sum_{j=1}^{n} \sum_{y \neq y_i} \alpha_{iy} \left( \Phi_{y, y_i}(x_i) \right); \quad \frac{1}{\pi} = \sum_{y \neq y_i} \alpha_{iy} + \beta_i; \quad \|\zeta_k\| \leq \gamma; \quad \sum_{y \neq y_i} \alpha_{iy} \leq \frac{1}{\pi}.
\]

By substituting the above equations into the Lagrangian function, we have

\[L(z, v, \xi, b, \alpha, \gamma, \zeta, w) = -\frac{1}{2\lambda} \gamma^2 - \frac{1}{2\lambda} \|\omega(\alpha)\|^2 + \sum_{i=1}^{n} \sum_{y \neq y_i} \alpha_{iy} \Delta(y, y_i).
\]

Hence the dual problem of the $\ell^2_{2,1}$-regularized problem can be written as:

\[
\begin{align*}
\max_{\gamma, \alpha} & \quad -\frac{1}{2\lambda} \gamma^2 - \frac{1}{2\lambda} \|u(\alpha)\|^2 + \sum_{i=1}^{n} \sum_{y \neq y_i} \alpha_{iy} \Delta(y, y_i) \\
\text{s.t.} & \quad \left| \sum_{i=1}^{n} \sum_{y \neq y_i} \alpha_{iy} \Phi_{y, y_i}(x_i) - \zeta_k \right| \leq \gamma, \quad k = 1, \ldots, t, \\
& \quad \alpha_i \geq 0, \sum_{y \neq y_i} \alpha_{iy} \leq \frac{1}{n}, \quad i = 1, \ldots, n.
\end{align*}
\]

Let $\theta := -\frac{1}{2\lambda} \gamma^2 - \frac{1}{2\lambda} \|u(\alpha)\|^2 + \sum_{i=1}^{n} \sum_{y \neq y_i} \alpha_{iy} \Delta(y, y_i), \quad \omega_k(\alpha, \eta_k) := \sum_{i=1}^{n} \sum_{y \neq y_i} \alpha_{iy} \Phi_{y, y_i}(x_i)$

and $g(\alpha, \eta_k) = -\frac{1}{2\lambda} \|\omega_k(\alpha, \eta_k)\|^2 - \frac{1}{2\lambda} \|\omega(\alpha)\|^2 + \sum_{i=1}^{n} \sum_{y \neq y_i} \alpha_{iy} \Delta(y, y_i)$. We have

\[
\begin{align*}
\max_{\theta, \alpha} & \quad \theta, \\
\text{s.t.} & \quad \theta \leq g(\alpha, \eta_k), \quad k = 1, \ldots, t, \\
& \quad \alpha_i \geq 0, \quad i = 1, \ldots, n.
\end{align*}
\]

which indeed is in the form of problem (8) by letting $A$ be the domain of $\alpha$. This completes the proof and brings the connection between the primal and dual formulation.

\[\square\]

E. Computation of $\Omega^*(z)$

The conjugate of $\Omega(w)$ is defined as

\[\Omega^*(z) = \max_u \omega^T z - \left( \frac{1}{2} \|u\|^2 + \frac{\sigma}{2\lambda} \|\omega\|^2 + \frac{1}{2} \sum_{k=1}^{t} \|\omega_k\|^2 \right).
\]
Let \( \mathbf{z} = [\mathbf{z}_u; \mathbf{z}_v] \), where \( \mathbf{z}_u \) and \( \mathbf{z}_v \) are vectors corresponding to \( \mathbf{u} \) and \( \mathbf{w} \), respectively. Let \( \Upsilon(\mathbf{w}) = \left( \frac{\sigma}{2\lambda} \| \mathbf{w} - \frac{\lambda\mathbf{z}_v}{\sigma} \|^2 + \frac{1}{2} \left( \sum_{k=1}^{t} \| \mathbf{w}_k \| \right)^2 \right) \). \( \Omega^*(\mathbf{z}) \) can be computed by

\[
\Omega^*(\mathbf{z}) = \arg\max_{\mathbf{u}, \mathbf{w}} \mathbf{u}^\top \mathbf{z}_u + \mathbf{w}^\top \mathbf{z}_v - \left( \frac{1}{2} \| \mathbf{u} \|^2 + \frac{\sigma}{2\lambda} \| \mathbf{w} \|^2 + \frac{1}{2} \left( \sum_{k=1}^{t} \| \mathbf{w}_k \| \right)^2 \right)
\]

In other words, we just need to solve the following problem

\[
\min_{\mathbf{w}} \quad \frac{\sigma}{2\lambda} \| \mathbf{w} - \frac{\lambda\mathbf{z}_v}{\sigma} \|^2 + \frac{1}{2} \left( \sum_{k=1}^{t} \| \mathbf{w}_k \| \right)^2 . \tag{23}
\]

This is a strictly convex problem, and a unique minimizer can be computed in closed-form [24].

**Proposition 3.** Let \( \bar{\mathbf{w}} \) be an optimal solution of problem (23). Then, \( \bar{\mathbf{w}} \) is unique, and can be cheaply calculated by Algorithm 3.

**Algorithm 3 Computation of \( \Omega^*(\mathbf{z}) \).**

Given \( \mathbf{z} = [\mathbf{z}_u; \mathbf{z}_v] \), parameter \( s = \frac{\lambda}{\sigma} \) and scalar \( T \). Let \( \bar{\mathbf{w}} = \frac{\lambda\mathbf{z}_v}{\sigma} \).

1: Calculate \( \bar{\omega}_k = \| \mathbf{w}_k \| \), where \( \mathbf{w}_k \) is associated with \( \mathbf{w}_k \) for all \( k = 1, ..., T \).
2: Sort \( \bar{\omega} \) to obtain \( \bar{\omega}_1 \geq \cdots \geq \bar{\omega}_T \).
3: Find \( \rho = \max \left\{ t \bar{\omega}_k - \frac{s}{1+\rho s} \sum_{i=1}^{k} \bar{\omega}_i > 0, \quad k = 1, ..., T \right\} \).
4: Calculate a threshold value \( \varsigma = \frac{s}{1+\rho s} \sum_{i=1}^{\rho} \bar{\omega}_i \).
5: Compute \( \bar{\omega}_k \), where \( \bar{\omega}_k \) = \( \begin{cases} \bar{\omega}_k - \varsigma, & \text{if } \bar{\omega}_k > \varsigma, \\ 0, & \text{Otherwise.} \end{cases} \)
6: Compute \( \bar{\mathbf{w}}_k = \begin{cases} \frac{\rho}{1+\rho s} \mathbf{w}_k, & \text{if } \bar{\omega}_k > 0, \\ 0, & \text{otherwise,} \end{cases} \)
7: Let \( \bar{\mathbf{w}} = [\mathbf{w}_k]_{k \in [T]} \). Output \( \Omega^*(\mathbf{z}) = [\mathbf{z}_u; \bar{\mathbf{w}}] \).

**Proof.** Please refer the proof in Appendix F of [24].

**F. Proof of Proposition 2**

**Proof.** Let \( P(\mathbf{w}) = \frac{1}{2} \| \mathbf{u} \|^2 + \frac{3}{2} (\sum_{k=1}^{t} \| \mathbf{w}_k \|)^2 + \frac{1}{n^2} \sum_{i=1}^{n} \xi_i \), \( Q(\mathbf{w}) = P(\mathbf{w}) + \frac{\sigma}{2} \| \mathbf{w} \|^2 \) and \( \Theta = \frac{1}{n} \sum_{i=1}^{n} (\max_{y \neq y_i} \Delta(\mathbf{y}, \mathbf{y}_i)) = P(\mathbf{0}) \). Suppose \( \mathbf{w} \) is a minimizer of \( P(\mathbf{w}) \). Then, we have \( P(\mathbf{w}) \leq P(\mathbf{0}) \).

Accordingly, we have \( \frac{1}{2} \| \mathbf{w} \|^2 \leq \frac{3}{2} (\sum_{k=1}^{t} \| \mathbf{w}_k \|)^2 \leq \frac{1}{2} (\sum_{k=1}^{t} \| \mathbf{w}_k \|)^2 + \frac{1}{2} \| \mathbf{u} \|^2 \leq P(\mathbf{w}) \leq \Theta \), which implies that \( \frac{1}{2} \| \mathbf{w} \|^2 \leq \Theta \). Let \( \mathbf{w}^* \) be an \( \frac{\varepsilon}{2} \)-accurate solution of (11). Then, we have \( Q(\mathbf{w}^*) \leq Q(\mathbf{w}) + \frac{\varepsilon}{2} \).

It follows that

\[
P(\mathbf{w}^*) \leq Q(\mathbf{w}^*) \leq Q(\mathbf{w}) + \frac{\varepsilon}{2} = P(\mathbf{w}) + \frac{\sigma}{2} \| \mathbf{w} \|^2 + \frac{\varepsilon}{2}.
\]

By setting \( \sigma \leq \lambda \varepsilon / 2\Theta \), we have \( \frac{\sigma}{2} \| \mathbf{w} \|^2 \leq \frac{\varepsilon}{2} \), and \( \mathbf{w}^* \) is an \( \varepsilon \)-accurate solution of (10).

**G. Proof of Theorem 2**

The proof can be adapted from the proof of Corollary 3 in [28].