Supplementary Materials for Paper 1915

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

$$\max_{\boldsymbol{\alpha}} \quad -\lambda\Omega^*\left(\frac{1}{\lambda n}\sum_{i=1}^{n}\sum_{\mathbf{y}\neq\mathbf{y}_i}\alpha_{i\mathbf{y}}\mathbf{w}^\top\,\Phi^{\mathcal{U}^t}_{\mathbf{y}_i,\mathbf{y}}(\mathbf{x}_i)\right) - \frac{1}{n}\sum_{i=1}^{n}L_i^*(-\boldsymbol{\alpha}_{i\mathbf{y}}), \qquad \text{s.t.} \sum_{\mathbf{y}\neq\mathbf{y}_i}\alpha_{i\mathbf{y}} \leq 1, \forall i \in [n]. \tag{15}$$

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

In Algorithm 2, we do not store $\alpha_{i\mathbf{y}}$ explicitly, thus we can not handle the inequality constraint directly. When the inequality constraint is ignored, the update rule $\delta_{i\mathbf{y}} = \frac{\lambda n(\triangle(\mathbf{y},\mathbf{y}_i)-d)}{(a^2+\nu)}$ may be too aggressive (e.g. $\delta_{i\mathbf{y}}$ may be too large). To address this, we use a scaled update rule $\delta_{i\mathbf{y}} = \frac{\lambda n(\triangle(\mathbf{y},\mathbf{y}_i)-d)}{\theta(a^2+\nu)}$ 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 $\Phi^{\mathcal{V}}_{\mathbf{y}_i,\mathbf{y}}(\mathbf{x}_i) := \mathbf{\Psi}_{\mathcal{V}}(\mathbf{y}_i,\mathbf{x}_i) - \mathbf{\Psi}_{\mathcal{V}}(\mathbf{y},\mathbf{x}_i)$, $\Phi^{\mathcal{C}}_{\mathbf{y}_i,\mathbf{y}}(\mathbf{x}_i;\boldsymbol{\eta}) := \mathbf{\Psi}_{\mathcal{C}}(\mathbf{y}_i,\mathbf{x}_i;\boldsymbol{\eta}) - \mathbf{\Psi}_{\mathcal{C}}(\mathbf{y},\mathbf{x}_i;\boldsymbol{\eta})$. The Lagrangian function of the inner minimization problem in (3) can be written as:

$$\mathcal{L}(\mathbf{w},\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{\lambda}{2}||\mathbf{w}||^2 + \frac{1}{n}\sum_{i=1}^{n}\xi_i - \boldsymbol{\beta}^\top\boldsymbol{\xi} + \sum_{i=1}^{n}\sum_{\mathbf{y}\neq\mathbf{y}_i}\alpha_{i\mathbf{y}}\left(\Delta(\mathbf{y},\mathbf{y}_i) - \mathbf{u}^\top\,\Phi^{\mathcal{V}}_{\mathbf{y}_i,\mathbf{y}}(\mathbf{x}_i) - \mathbf{v}^\top\,\Phi^{\mathcal{C}}_{\mathbf{y}_i,\mathbf{y}}(\mathbf{x}_i;\boldsymbol{\eta}) - \xi_i\right). \quad (16)$$

Let $\boldsymbol{\alpha} := [\alpha_{1\mathbf{y}},...,\alpha_{n\mathbf{y}}]^\top$. The KKT condition of (16) can be written as

$$\frac{\partial\mathcal{L}(\mathbf{w},\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial\mathbf{u}} = 0 \Rightarrow \mathbf{u} = \frac{1}{\lambda}\sum_{i=1}^{n}\sum_{\mathbf{y}\neq\mathbf{y}_i}\alpha_{i\mathbf{y}}\,\Phi^{\mathcal{V}}_{\mathbf{y}_i,\mathbf{y}}(\mathbf{x}_i); \quad (17)$$

$$\frac{\partial\mathcal{L}(\mathbf{w},\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial\mathbf{v}} = 0 \Rightarrow \mathbf{v} = \frac{1}{\lambda}\sum_{i=1}^{n}\sum_{\mathbf{y}\neq\mathbf{y}_i}\alpha_{i\mathbf{y}}\,\Phi^{\mathcal{C}}_{\mathbf{y}_i,\mathbf{y}}(\mathbf{x}_i;\boldsymbol{\eta}); \quad (18)$$

$$\frac{\partial\mathcal{L}(\mathbf{w},\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial\xi_i} = 0 \Rightarrow \frac{1}{n} = \sum_{\mathbf{y}\neq\mathbf{y}_i}\alpha_{i\mathbf{y}} + \beta_i; \quad (19)$$

$$\boldsymbol{\alpha}\succeq\mathbf{0}, \text{ and } \boldsymbol{\beta}\succeq\mathbf{0}. \quad (20)$$

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

$$\mathbf{u}(\boldsymbol{\alpha}) := \sum_{i=1}^{n}\sum_{\mathbf{y}\neq\mathbf{y}_i}\alpha_{i\mathbf{y}}\,\Phi^{\mathcal{V}}_{\mathbf{y}_i,\mathbf{y}}(\mathbf{x}_i) \text{ and } \mathbf{v}(\boldsymbol{\alpha},\boldsymbol{\eta}) := \sum_{i=1}^{n}\sum_{\mathbf{y}\neq\mathbf{y}_i}\alpha_{i\mathbf{y}}\,\Phi^{\mathcal{C}}_{\mathbf{y}_i,\mathbf{y}}(\mathbf{x}_i;\boldsymbol{\eta}). \quad (21)$$

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

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}} \quad -\frac{1}{2\lambda}||\mathbf{u}(\boldsymbol{\alpha})||^2 - \frac{1}{2\lambda}||\mathbf{v}(\boldsymbol{\alpha},\boldsymbol{\eta})||^2 + \mathbf{b}^\top\boldsymbol{\alpha}. \quad (22)$$

□

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 $\Omega(\boldsymbol{\omega}) = \frac{1}{2}(\sum_{k=1}^{t}\|\boldsymbol{\omega}_k\|)^2$. Define a cone $\mathcal{Q}_r = \{(\mathbf{u},v)\in\mathbb{R}^{r+1}, \|\mathbf{u}\|_2 \leq v\}$. Let $z_k = \|\boldsymbol{\omega}_k\|$, we have $\Omega(\mathbf{v}) = \frac{1}{2}(\sum_{k=1}^{t}\|\boldsymbol{\omega}_k\|)^2 = \frac{1}{2}z^2$, 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:

$$\min_{z,\mathbf{u},\mathbf{v}} \frac{\lambda}{2}||\mathbf{u}||^2 + \frac{\lambda}{2}z^2 + \frac{1}{n}\sum_{i=1}^{n}\xi_i, \quad \text{s.t.} \sum_{k=1}^{t}z_k \leq z, \quad (\boldsymbol{\omega}_k, z_k)\in\mathcal{Q}_r,$$
$$\mathbf{w}^\top\,\Phi^{\mathcal{U}^t}_{\mathbf{y}_i,\mathbf{y}}(\mathbf{x}_i) \geq \Delta(\mathbf{y},\mathbf{y}_i) - \xi_i, \;\; \xi_i \geq 0\,\forall i, \forall\mathbf{y}\in\mathcal{Y}\backslash\mathbf{y}_i.$$

where $\boldsymbol{\omega} = [\boldsymbol{\omega}'_1,...,\boldsymbol{\omega}'_t]'$. The Lagrangian function of (23) can be written as:

$$\mathcal{L}(z,\mathbf{v},\boldsymbol{\xi},b,\boldsymbol{\alpha},\gamma,\boldsymbol{\zeta},\boldsymbol{\varpi}) = \frac{\lambda}{2}||\mathbf{u}||^2 + \frac{\lambda}{2}z^2 + \frac{1}{n}\sum_{i=1}^{n}\xi_i + \gamma(\sum_{k=1}^{t}z_k - z) - \sum_{k=1}^{t}(\boldsymbol{\zeta}'_k\boldsymbol{\omega}_k + \varpi_k z_k)$$
$$-\sum_{i=1}^{n}\sum_{\mathbf{y}\neq\mathbf{y}_i}\alpha_{i\mathbf{y}}\left(\Delta(\mathbf{y},\mathbf{y}_i) - \xi_i - \left(\Phi^{\mathcal{V}}_{\mathbf{y}_i,\mathbf{y}}(\mathbf{x}_i) + \sum_{k=1}^{t}\boldsymbol{\omega}_k^\top\,\Phi^{\Gamma_k}_{\mathbf{y}_i,\mathbf{y}}(\mathbf{x}_i)\right)\right) - \boldsymbol{\beta}^\top\boldsymbol{\xi},$$

where $\boldsymbol{\alpha}$, γ, $\boldsymbol{\zeta}_k$ and ϖ_k are the Lagrangian dual variables to the corresponding constraints. The KKT condition can be expressed as

$$
\begin{array}{ll}
\nabla_z \mathcal{L} = \lambda z - \gamma = 0 & \Rightarrow z = \frac{\gamma}{\lambda}; \\
\nabla_{z_k} \mathcal{L} = \gamma - \varpi_k = 0 & \Rightarrow \varpi_k = \gamma; \\
\nabla_{\mathbf{u}} \mathcal{L} = \lambda \mathbf{u} + \sum_{i=1}^n \sum_{\mathbf{y} \neq \mathbf{y}_i} \alpha_{i\mathbf{y}} \, \Phi^{\mathcal{V}}_{\mathbf{y}_i, \mathbf{y}}(\mathbf{x}_i) & \Rightarrow \mathbf{u} = -\frac{1}{\lambda} \sum_{i=1}^n \sum_{\mathbf{y} \neq \mathbf{y}_i} \alpha_{i\mathbf{y}} \, \Phi^{\mathcal{V}}_{\mathbf{y}_i, \mathbf{y}}(\mathbf{x}_i) \\
\nabla_{\boldsymbol{\omega}_k} \mathcal{L} = -\sum_{i=1}^n \sum_{\mathbf{y} \neq \mathbf{y}_i} \alpha_{i\mathbf{y}} \left(\Phi^{\Gamma_k}_{\mathbf{y}_i, \mathbf{y}}(\mathbf{x}_i) \right) - \boldsymbol{\zeta}_k = 0 & \Rightarrow \boldsymbol{\zeta}_k = -\sum_{i=1}^n \sum_{\mathbf{y} \neq \mathbf{y}_i} \alpha_{i\mathbf{y}} \left(\Phi^{\Gamma_k}_{\mathbf{y}_i, \mathbf{y}}(\mathbf{x}_i) \right); \\
\nabla_{\xi_i} \mathcal{L} = \frac{1}{n} - \sum_{\mathbf{y} \neq \mathbf{y}_i} \alpha_{i\mathbf{y}} - \beta_i = 0 & \Rightarrow \frac{1}{n} = \sum_{\mathbf{y} \neq \mathbf{y}_i} \alpha_{i\mathbf{y}} + \beta_i; \\
\|\boldsymbol{\zeta}_k\| \leq \varpi_k & \Rightarrow \|\boldsymbol{\zeta}_k\| \leq \gamma; \\
\beta_i \geq 0 & \Rightarrow \sum_{\mathbf{y} \neq \mathbf{y}_i} \alpha_{i\mathbf{y}} \leq \frac{1}{n}.
\end{array}
$$

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

$$
\mathcal{L}(z, \mathbf{v}, \boldsymbol{\xi}, b, \boldsymbol{\alpha}, \gamma, \boldsymbol{\zeta}, \boldsymbol{\varpi}) = -\frac{1}{2\lambda}\gamma^2 - \frac{1}{2\lambda}||\omega(\boldsymbol{\alpha})||^2 + \sum_{i=1}^{n} \sum_{\mathbf{y} \neq \mathbf{y}_i} \alpha_{i\mathbf{y}} \Delta(\mathbf{y}, \mathbf{y}_i).
$$

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

$$
\begin{aligned}
\max_{\gamma, \boldsymbol{\alpha}} \quad & -\frac{1}{2\lambda}\gamma^2 - \frac{1}{2\lambda}||\mathbf{u}(\boldsymbol{\alpha})||^2 + \sum_{i=1}^{n} \sum_{\mathbf{y} \neq \mathbf{y}_i} \alpha_{i\mathbf{y}} \Delta(\mathbf{y}, \mathbf{y}_i) \\
\text{s.t} \quad & \left\| \sum_{i=1}^{n} \sum_{\mathbf{y} \neq \mathbf{y}_i} \alpha_{i\mathbf{y}} \, \Phi^{\Gamma_k}_{\mathbf{y}_i, \mathbf{y}}(\mathbf{x}_i) - \boldsymbol{\zeta}_k \right\| \leq \gamma, \;\; k = 1, \cdots, t, \\
& \alpha_i \geq 0, \sum_{\mathbf{y} \neq \mathbf{y}_i} \alpha_{i\mathbf{y}} \leq \frac{1}{n}, \;\; i = 1, \cdots, n.
\end{aligned}
$$

Let $\theta := -\frac{1}{2\lambda}\gamma^2 - \frac{1}{2\lambda}||\mathbf{u}(\boldsymbol{\alpha})||^2 + \sum_{i=1}^n \sum_{\mathbf{y} \neq \mathbf{y}_i} \alpha_{i\mathbf{y}} \Delta(\mathbf{y}, \mathbf{y}_i)$, $\boldsymbol{\omega}_k(\boldsymbol{\alpha}, \boldsymbol{\eta}_k) := \sum_{i=1}^n \sum_{\mathbf{y} \neq \mathbf{y}_i} \alpha_{i\mathbf{y}} \, \Phi^{\Gamma_k}_{\mathbf{y}_i, \mathbf{y}}(\mathbf{x}_i)$ and $g(\boldsymbol{\alpha}, \boldsymbol{\eta}_k) = -\frac{1}{2\lambda}\|\boldsymbol{\omega}_k(\boldsymbol{\alpha}, \boldsymbol{\eta}_k)\|^2 - \frac{1}{2\lambda}||\omega(\boldsymbol{\alpha})||^2 + \sum_{i=1}^n \sum_{\mathbf{y} \neq \mathbf{y}_i} \alpha_{i\mathbf{y}} \Delta(\mathbf{y}, \mathbf{y}_i)$. We have

$$
\begin{aligned}
\max_{\theta, \boldsymbol{\alpha}} \quad & \theta, \\
\text{s.t.} \quad & \theta \leq g(\boldsymbol{\alpha}, \boldsymbol{\eta}_k), \;\; k = 1, \cdots, t, \\
& \alpha_i \geq 0, \; i = 1, \cdots, n.
\end{aligned}
$$

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

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

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

$$
\Omega^*(\mathbf{z}) = \max_{\mathbf{u}, \boldsymbol{\omega}} \;\; \mathbf{w}^\top \mathbf{z} \; - \left(\frac{1}{2}\|\mathbf{u}\|^2 + \frac{\sigma}{2\lambda}\|\boldsymbol{\omega}\|^2 + \frac{1}{2}(\sum_{k=1}^{t} \|\boldsymbol{\omega}_k\|)^2 \right).
$$

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

$$\begin{aligned}
\Omega^*(\mathbf{z}) &= \arg\max_{\mathbf{u},\boldsymbol{\omega}} \ \mathbf{u}^\top\mathbf{z}_u + \boldsymbol{\omega}^\top\mathbf{z}_v - \left(\frac{1}{2}\|\mathbf{u}\|^2 + \frac{\sigma}{2\lambda}\|\boldsymbol{\omega}\|^2 + \frac{1}{2}(\sum_{k=1}^t \|\boldsymbol{\omega}_k\|)^2\right) \\
&= \left[\arg\min_{\mathbf{u}} \ \left(\frac{1}{2}\|\mathbf{u}\|^2 - \mathbf{u}^\top\mathbf{z}_u\right); \arg\min_{\boldsymbol{\omega}} \Upsilon(\boldsymbol{\omega})\right] \\
&= \left[\mathbf{z}_u; \arg\min_{\boldsymbol{\omega}} \Upsilon(\boldsymbol{\omega})\right],
\end{aligned}$$

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

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

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

Proposition 3. *Let $\widehat{\boldsymbol{\omega}}$ be an optimal solution of problem (23). Then, $\widehat{\boldsymbol{\omega}}$ 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 $\boldsymbol{\omega} = \frac{\lambda\mathbf{z}_v}{\sigma}$.
1: Calculate $\widehat{o}_k = \|\boldsymbol{\omega}_k\|$, where $\boldsymbol{\omega}_k$ is associated with $\boldsymbol{\omega}_k$ for all $k = 1, ..., T$.
2: Sort $\widehat{\mathbf{o}}$ to obtain $\bar{\mathbf{o}}$ such that $\bar{o}_{(1)} \geq ... \geq \bar{o}_{(T)}$.
3: Find $\rho = \max\left\{t \middle| \bar{o}_k - \frac{s}{1+ks}\sum_{i=1}^k \bar{o}_i > 0, k = 1, ..., T\right\}$.
4: Calculate a threshold value $\varsigma = \frac{s}{1+\rho s}\sum_{i=1}^{\rho} \bar{o}_i$.
5: Compute $\mathbf{o}$, where $o_k = \begin{cases} \widehat{o}_k - \varsigma, & \text{if} \quad \widehat{o}_k > \varsigma, \\ 0, & \text{Otherwise.} \end{cases}$.
6: Compute $\widehat{\boldsymbol{\omega}}_k = \begin{cases} \frac{o_k}{\|\boldsymbol{\omega}_k\|}\boldsymbol{\omega}_k, & \text{if } o_k > 0, \\ \mathbf{0}, & \text{otherwise,} \end{cases}$
7: Let $\widehat{\boldsymbol{\omega}} = [\boldsymbol{\omega}_k]_{k\in[T]}$. Output $\Omega^*(\mathbf{z}) = [\mathbf{z}_u; \widehat{\boldsymbol{\omega}}]$.

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

F. Proof of Proposition 2

Proof. Let $P(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{u}\|^2 + \frac{\lambda}{2}(\sum_{k=1}^t \|\boldsymbol{\omega}_k\|)^2 + \frac{1}{n}\sum_{i=1}^n \xi_i$, $Q(\mathbf{w}) = P(\mathbf{w}) + \frac{\sigma}{2}\|\boldsymbol{\omega}\|^2$ and $\Theta = \frac{1}{n}\sum_{i=1}^n(\max_{\mathbf{y}\neq\mathbf{y}_i}\Delta(\mathbf{y}, \mathbf{y}_i)) = P(\mathbf{0})$. Suppose $\bar{\mathbf{w}}$ is a minimizer of $P(\mathbf{w})$. Then, we have $P(\bar{\mathbf{w}}) \leq P(\mathbf{0})$. Accordingly, we have $\frac{\lambda}{2}||\bar{\boldsymbol{\omega}}||^2 \leq \frac{\lambda}{2}(\sum_{k=1}^t \|\bar{\boldsymbol{\omega}}_k\|)^2 \leq \frac{\lambda}{2}(\sum_{k=1}^t \|\bar{\boldsymbol{\omega}}_k\|)^2 + \frac{\lambda}{2}\|\bar{\mathbf{u}}\|^2 \leq P(\bar{\mathbf{w}}) \leq \Theta$, which implies that $\frac{\lambda}{2}||\bar{\boldsymbol{\omega}}||^2 \leq \Theta$. Let $\mathbf{w}^*$ be an $\frac{\epsilon}{2}$-accurate solution of (11). Then, we have $Q(\mathbf{w}^*) \leq Q(\bar{\mathbf{w}}) + \frac{\epsilon}{2}$. It follows that

$$P(\mathbf{w}^*) \leq Q(\mathbf{w}^*) \leq Q(\bar{\mathbf{w}}) + \frac{\epsilon}{2} = P(\bar{\mathbf{w}}) + \frac{\sigma}{2}\|\bar{\boldsymbol{\omega}}\|^2 + \frac{\epsilon}{2}.$$

By setting $\sigma \leq \lambda\epsilon/2\Theta$, we have $\frac{\sigma}{2}\|\bar{\boldsymbol{\omega}}\|^2 \leq \frac{\epsilon}{2}$, and $\mathbf{w}^*$ is an ϵ-accurate solution of (10). □

G. Proof of Theorem 2

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