Supplementary materials for
An Exact Solver for the Weston-Watkins SVM Subproblem

Contents

A.1. Regarding offsets

In this section, we review the literature on SVMs in particular with regard to offsets.

For binary kernel SVMs, Steinwart et al. (2011) demonstrates that kernel SVMs without offset achieve comparable classification accuracy as kernel SVMs with offset. Furthermore, they propose algorithms that solve kernel SVMs without offset that are significantly faster than solvers for kernel SVMs with offset.

For binary linear SVMs, Hsieh et al. (2008) introduced coordinate descent for the dual problem associated to linear SVMs without offsets, or with the bias term included in the $\mathbf{w}$ term. Chiu et al. (2020) studied whether the method of Hsieh et al. (2008) can be extended to allow offsets, but found evidence that the answer is negative. For multiclass linear SVMs, Keerthi et al. (2008) studied block coordinate descent for the CS-SVM and WW-SVM, both without offsets. We are not aware of a multiclass analogue to Chiu et al. (2020) although the situation should be similar.

The previous paragraph discussed coordinate descent in relation to the offset. Including the offset presents challenges to primal methods as well. In Section 6 of Shalev-Shwartz et al. (2011), the authors argue that including an unregularized offset term in the primal objective leads to slower convergence guarantee. Furthermore, Shalev-Shwartz et al. (2011) observed that including an unregularized offset did not significantly change the classification accuracy.

The original Crammer-Singer (CS) SVM was proposed without offsets (Crammer & Singer, 2001). In Section VI of (Hsu & Lin, 2002), the authors show the CS-SVM with offsets do *not* perform better than CS-SVM without offsets. Furthermore, CS-SVM with offsets requires twice as many iterations to converge than without.

A.2. Proof of Proposition 3.2

Below, let $i \in [n]$ be arbitrary. First, we note that $-\boldsymbol{\pi}' = \begin{bmatrix} -\mathbb{1}' \\ \mathbf{I}_{k-1} \end{bmatrix}$ and so

$$\boldsymbol{\pi}'\beta_i = \begin{bmatrix} -\mathbb{1}'\beta_i \\ \beta_i \end{bmatrix}. \tag{14}$$

Now, let $j \in [k]$, we have by (3) that

$$[\alpha_i]_j = [-\boldsymbol{\sigma}_{y_i}\boldsymbol{\pi}'\beta_i]_j = [-\boldsymbol{\pi}'\beta_i]_{\boldsymbol{\sigma}_{y_i}(j)}. \tag{15}$$

Note that if $j \neq y_i$, then $\boldsymbol{\sigma}_{y_i}(j) \neq 1$ and so $[\alpha_i]_j = [-\boldsymbol{\pi}'\beta_i]_{\boldsymbol{\sigma}_{y_i}(j)} = [\beta_i]_{\boldsymbol{\sigma}_{y_i}(j)-1} \in [0, C]$. On the other hand, if $j = y_i$, then $\boldsymbol{\sigma}_{y_i}(y_i) = 1$ and $[\alpha_i]_{y_i} = [-\boldsymbol{\pi}'\beta_i]_1 = -\mathbb{1}'\beta_i = -\sum_{t\in[k-1]}[\beta_i]_t = -\sum_{t\in[k]:t\neq y_i}[\beta_i]_{\boldsymbol{\sigma}_{y_i}(t)-1} = -\sum_{t\in[k]:t\neq y_i}[\alpha_i]_t$. Thus, $\boldsymbol{\alpha} \in \mathcal{F}$. This proves that $\Psi(\mathcal{G}) \subseteq \mathcal{F}$.

Next, let us define another map $\Xi : \mathcal{F} \to \mathbb{R}^{(k-1)\times n}$ as follows: For each $\boldsymbol{\alpha} \in \mathcal{F}$, define $\boldsymbol{\beta} := \Xi(\boldsymbol{\alpha})$ block-wise by

$$\beta_i := \mathtt{proj}_{2:k}(\boldsymbol{\sigma}_{y_i}\alpha_i) \in \mathbb{R}^{k-1}$$

where

$$\mathtt{proj}_{2:k} = \begin{bmatrix}\mathbb{0} & \mathbf{I}_{k-1}\end{bmatrix} \in \mathbb{R}^{(k-1)\times k}.$$

By construction, we have for each $j \in [k-1]$ that $[\beta_i]_j = [\boldsymbol{\sigma}_{y_i}\alpha_i]_{j+1} = [\boldsymbol{\sigma}_{y_i}\alpha_i]_{j+1} = [\alpha_i]_{\boldsymbol{\sigma}_{y_i}(j+1)}$ Since $j+1 \neq 1$ for any $j \in [k-1]$, we have that $\boldsymbol{\sigma}_{y_i}(j+1) \neq y_i$ for any $j \in [k-1]$. Thus, $[\beta_i]_j = [\alpha_i]_{\boldsymbol{\sigma}_{y_i}(j+1)} \in [0, C]$. This proves that $\Xi(\mathcal{F}) \subseteq \mathcal{G}$.

Next, we prove that for all $\boldsymbol{\alpha} \in \mathcal{F}$ and $\boldsymbol{\beta} \in \mathcal{G}$, we have $\Xi(\Psi(\boldsymbol{\beta})) = \boldsymbol{\beta}$ and $\Psi(\Xi(\boldsymbol{\alpha})) = \boldsymbol{\alpha}$.

By construction, the i-th block of $\Xi(\Psi(\boldsymbol{\beta}))$ is given by

$$\begin{aligned}
\mathtt{proj}_{2:k}(\boldsymbol{\sigma}_{y_i}(-\boldsymbol{\sigma}_{y_i}\boldsymbol{\pi}'\beta_i)) &= -\mathtt{proj}_{2:k}(\boldsymbol{\sigma}_{y_i}\boldsymbol{\sigma}_{y_i}\boldsymbol{\pi}'\beta_i) \\
&= -\mathtt{proj}_{2:k}(\boldsymbol{\pi}'\beta_i) \\
&= -\begin{bmatrix}\mathbb{0} & \mathbf{I}_{k-1}\end{bmatrix}\begin{bmatrix}\mathbb{1}' \\ -\mathbf{I}_{k-1}\end{bmatrix}\beta_i \\
&= \mathbf{I}_{k-1}\beta_i = \beta_i.
\end{aligned}$$

For the second equality, we used the fact that $\boldsymbol{\sigma}_y^2 = \mathbf{I}$ for all $y \in [k]$. Thus, $\Xi(\Psi(\boldsymbol{\beta})) = \boldsymbol{\beta}$.

Next, note that the i-th block of $\Psi(\Xi(\boldsymbol{\alpha}))$ is, by construciton,

$$-\boldsymbol{\sigma}_{y_i}\boldsymbol{\pi}'\mathtt{proj}_{2:k}(\boldsymbol{\sigma}_{y_i}\alpha_i) = -\boldsymbol{\sigma}_{y_i}\pi'\begin{bmatrix}\mathbb{0} & \mathbf{I}_{k-1}\end{bmatrix}\boldsymbol{\sigma}_{y_i}\alpha_i = -\boldsymbol{\sigma}_{y_i}\begin{bmatrix}\mathbb{0} & \pi'\end{bmatrix}\boldsymbol{\sigma}_{y_i}\alpha_i \tag{16}$$

Recall that $\pi' = \begin{bmatrix}\mathbb{1}' \\ -\mathbf{I}_{k-1}\end{bmatrix}$ and so $\begin{bmatrix}\mathbb{0} & \pi'\end{bmatrix} = \begin{bmatrix}0 & \mathbb{1}' \\ \mathbb{0} & -\mathbf{I}_{k-1}\end{bmatrix}$. Therefore,

$$\left[\begin{bmatrix}\mathbb{0} & \pi'\end{bmatrix}\boldsymbol{\sigma}_{y_i}\alpha_i\right]_1 = \sum_{j=2}^{k}[\boldsymbol{\sigma}_{y_i}\alpha_i] = \sum_{j\in[k]:j\neq y_i}[\alpha_i]_j = -[\alpha_i]_{y_i} = -[\boldsymbol{\sigma}_{y_i}\alpha_i]_1$$

and, for $j = 2, \ldots, k$,

$$\left[\begin{bmatrix}\mathbb{0} & \pi'\end{bmatrix}\boldsymbol{\sigma}_{y_i}\alpha_i\right]_j = -[\boldsymbol{\sigma}_{y_i}\alpha_i]_j.$$

Hence, we have just shown that $\begin{bmatrix}\mathbb{0} & \pi'\end{bmatrix}\boldsymbol{\sigma}_{y_i}\alpha_i = -\boldsymbol{\sigma}_{y_i}\alpha_i$. Continuing from (16), we have

$$-\boldsymbol{\sigma}_{y_i}\boldsymbol{\pi}'\mathtt{proj}_{2:k}(\boldsymbol{\sigma}_{y_i}\alpha_i) = -\boldsymbol{\sigma}_{y_i}(-\boldsymbol{\sigma}_{y_i}\alpha_i) = \boldsymbol{\sigma}_{y_i}\boldsymbol{\sigma}_{y_i}\alpha_i = \alpha_i.$$

This proves that $\Psi(\Xi(\boldsymbol{\alpha})) = \boldsymbol{\alpha}$. Thus, we have shown that Ψ and Ξ are inverses of one another. This proves that Ψ is a bijection.

Finally, we prove that

$$f(\Psi(\boldsymbol{\beta})) = g(\boldsymbol{\beta}).$$

Recall that

$$f(\boldsymbol{\alpha}) := \frac{1}{2}\sum_{i,s\in[n]} x_s'x_i\alpha_i'\alpha_s - \sum_{i\in[k]}\sum_{\substack{j\in[k]:\\ j\neq y_i}}\alpha_{ij}$$

Thus,

$$\alpha_i'\alpha_s = (-\boldsymbol{\sigma}_{y_i}\boldsymbol{\pi}'\beta_i)'(-\boldsymbol{\sigma}_{y_s}\boldsymbol{\pi}'\beta_s) = \beta_i'\boldsymbol{\pi}\boldsymbol{\sigma}_{y_i}\boldsymbol{\sigma}_{y_s}'\boldsymbol{\pi}'\beta_s$$

On the other hand, (3) implies that $\boldsymbol{\sigma}_{y_i}\alpha_i = -\boldsymbol{\pi}'\beta_i$. Hence

$$\sum_{j\in[k]\setminus\{y_i\}}\alpha_{ij} = \sum_{j\in[k]:j\neq 1}[\alpha_i]_{\boldsymbol{\sigma}_{y_i}(j)} = \sum_{j\in[k]:j\neq 1}[\boldsymbol{\sigma}_{y_i}\alpha_i]_j = \sum_{j\in[k]:j\neq 1}[-\boldsymbol{\pi}'\beta_i]_j = \sum_{j\in[k-1]}[\beta_i]_j = \mathbb{1}'\beta_i.$$

Thus,

$$f(\boldsymbol{\alpha}) := \frac{1}{2}\sum_{i,s\in[n]} x_s'x_i\alpha_i'\alpha_s - \sum_{i\in[k]}\sum_{\substack{j\in[k]:\\ j\neq y_i}}\alpha_{ij} = \frac{1}{2}\sum_{i,s\in[n]} x_s'x_i\beta_i'\boldsymbol{\pi}\boldsymbol{\sigma}_{y_i}\boldsymbol{\sigma}_{y_s}'\boldsymbol{\pi}'\beta_s - \sum_{i\in[k]}\mathbb{1}'\beta_i = g(\boldsymbol{\beta})$$

as desired. Finally, we note that $\boldsymbol{\sigma}_y = \boldsymbol{\sigma}_y'$ for all $y \in [k]$. This concludes the proof of Proposition 3.2. □

A.3. Proof of Proposition 3.5

We prove the following lemma which essentially unpacks the succinct Proposition 3.5:

Lemma A.1. *Recall the situation of Corollary 3.3: Let $\boldsymbol{\beta} \in \mathcal{G}$ and $i \in [n]$. Let $\boldsymbol{\alpha} = \Psi(\boldsymbol{\beta})$. Consider*

$$\min_{\widehat{\boldsymbol{\beta}} \in \mathcal{G}} \; g(\widehat{\boldsymbol{\beta}}) \text{ such that } \; \widehat{\beta}_s = \beta_s, \, \forall s \in [n] \setminus \{i\}. \tag{17}$$

Let $\mathbf{w}$ *be as in* (1)*, i.e.,* $\mathbf{w} = -\sum_{i \in [n]} x_i \alpha_i'$*. Then a solution to* (17) *is given by* $[\beta_1, \ldots, \beta_{i-1}, \widetilde{\beta}_i, \beta_{i+1}, \ldots, \beta_n]$ *where* $\widetilde{\beta}_i$ *is a minimizer of*

$$\min_{\widehat{\beta}_i \in \mathbb{R}^{k-1}} \; \frac{1}{2} \widehat{\beta}_i' \boldsymbol{\Theta} \widehat{\beta}_i - \widehat{\beta}_i' \left((\mathbb{1} - \boldsymbol{\pi} \boldsymbol{\sigma}_{y_i} \mathbf{w}' x_i) / \|x_i\|_2^2 + \boldsymbol{\Theta} \beta_i \right) \text{ such that } 0 \le \widehat{\beta}_i \le C.$$

Furthermore, the above optimization has a unique minimizer which is equal to the minimizer of (4) *where*

$$v := (\mathbb{1} - \boldsymbol{\rho}_{y_i} \boldsymbol{\pi} \mathbf{w}' x_i + \boldsymbol{\Theta} \beta_i \|x_i\|_2^2) / \|x_i\|_2^2$$

and $\mathbf{w}$ *is as in* (1)*.*

Proof. First, we prove a simple identity:

$$\boldsymbol{\pi} \boldsymbol{\pi}' = \begin{bmatrix} \mathbb{1} & -\mathbf{I}_{k-1} \end{bmatrix} \begin{bmatrix} \mathbb{1}' \\ -\mathbf{I}_{k-1} \end{bmatrix} = \mathbf{I} + \mathbf{O} = \boldsymbol{\Theta}. \tag{18}$$

Next, recall that by definition, we have

$$g(\boldsymbol{\beta}) := \left(\frac{1}{2} \sum_{s,t \in [n]} x_s' x_t \beta_t' \boldsymbol{\pi} \boldsymbol{\sigma}_{y_t} \boldsymbol{\sigma}_{y_s} \boldsymbol{\pi}' \beta_s \right) - \left(\sum_{s \in [n]} \mathbb{1}' \beta_s \right).$$

Let us group the terms of $g(\boldsymbol{\beta})$ that depends on β_i:

$$\begin{aligned}
g(\boldsymbol{\beta}) = & \frac{1}{2} x_i' x_i \beta_i' \boldsymbol{\pi} \boldsymbol{\sigma}_{y_i} \boldsymbol{\sigma}_{y_i} \boldsymbol{\pi}' \beta_i \\
& + \frac{1}{2} \sum_{s \in [n] : s \neq i} x_s' x_i \beta_i' \boldsymbol{\pi} \boldsymbol{\sigma}_{y_i} \boldsymbol{\sigma}_{y_s} \boldsymbol{\pi}' \beta_s \\
& + \frac{1}{2} \sum_{t \in [n] : t \neq i} x_i' x_t \beta_t' \boldsymbol{\pi} \boldsymbol{\sigma}_{y_t} \boldsymbol{\sigma}_{y_i} \boldsymbol{\pi}' \beta_i \\
& + \frac{1}{2} \sum_{s,t \in [n]} x_s' x_t \beta_t' \boldsymbol{\pi} \boldsymbol{\sigma}_{y_t} \boldsymbol{\sigma}_{y_s} \boldsymbol{\pi}' \beta_s - \sum_{s \in [n]} \mathbb{1}' \beta_s \\
= & \frac{1}{2} x_i' x_i \beta_i' \boldsymbol{\Theta} \beta_i \qquad \because \boldsymbol{\sigma}_{y_i}^2 = \mathbf{I} \text{ and (18)} \\
& + \sum_{s \in [n] : s \neq i} x_s' x_i \beta_i' \boldsymbol{\pi} \boldsymbol{\sigma}_{y_i} \boldsymbol{\sigma}_{y_s} \boldsymbol{\pi}' \beta_s \\
& - \mathbb{1}' \beta_i \\
& + \underbrace{\frac{1}{2} \sum_{s,t \in [n]} x_s' x_t \beta_t' \boldsymbol{\pi} \boldsymbol{\sigma}_{y_t} \boldsymbol{\sigma}_{y_s} \boldsymbol{\pi}' \beta_s - \sum_{s \in [n] : s \neq i} \mathbb{1}' \beta_s}_{=: C_i}
\end{aligned}$$

where C_i is a scalar quantity which does not depend on β_i. Thus, plugging in $\widehat{\boldsymbol{\beta}}$, we have

$$g(\widehat{\boldsymbol{\beta}}) = \frac{1}{2} \|x_i\|_2^2 \widehat{\beta}_i' \boldsymbol{\Theta} \widehat{\beta}_i + \sum_{s \in [n] : s \neq i} x_s' x_i \widehat{\beta}_i' \boldsymbol{\pi} \boldsymbol{\sigma}_{y_i} \boldsymbol{\sigma}_{y_s} \boldsymbol{\pi}' \beta_s - \mathbb{1}' \widehat{\beta}_i + C_i. \tag{19}$$

Furthermore,

$$\begin{aligned}
\sum_{s\in[n]:s\neq i} x_s' x_i \widehat{\beta}_i' \boldsymbol{\pi}\boldsymbol{\sigma}_{y_i}\boldsymbol{\sigma}_{y_s}\boldsymbol{\pi}'\beta_s &= \sum_{s\in[n]:s\neq i} \widehat{\beta}_i' \boldsymbol{\pi}\boldsymbol{\sigma}_{y_i}\boldsymbol{\sigma}_{y_s}\boldsymbol{\pi}'\beta_s x_s' x_i \\
&= \widehat{\beta}_i' \boldsymbol{\pi}\boldsymbol{\sigma}_{y_i} \left(\sum_{s\in[n]:s\neq i} \boldsymbol{\sigma}_{y_s}\boldsymbol{\pi}'\beta_s x_s' \right) x_i \\
&= \widehat{\beta}_i' \boldsymbol{\pi}\boldsymbol{\sigma}_{y_i} \left(-\boldsymbol{\sigma}_{y_i}\boldsymbol{\pi}'\beta_i x_i' + \sum_{s\in[n]} \boldsymbol{\sigma}_{y_s}\boldsymbol{\pi}'\beta_s x_s' \right) x_i \\
&= \widehat{\beta}_i' \boldsymbol{\pi}\boldsymbol{\sigma}_{y_i} \left(-\boldsymbol{\sigma}_{y_i}\boldsymbol{\pi}'\beta_i x_i' - \sum_{s\in[n]} \alpha_s x_s' \right) x_i \qquad \because (3) \\
&= \widehat{\beta}_i' \boldsymbol{\pi}\boldsymbol{\sigma}_{y_i} \left(-\boldsymbol{\sigma}_{y_i}\boldsymbol{\pi}'\beta_i x_i' + \mathbf{w}' \right) x_i \qquad \because (1) \\
&= \widehat{\beta}_i' \left(-\boldsymbol{\pi}\boldsymbol{\sigma}_{y_i}\boldsymbol{\sigma}_{y_i}\boldsymbol{\pi}'\beta_i \|x_i\|_2^2 + \boldsymbol{\pi}\boldsymbol{\sigma}_{y_i}\mathbf{w}' x_i \right) \\
&= \widehat{\beta}_i' \left(\boldsymbol{\pi}\boldsymbol{\sigma}_{y_i}\mathbf{w}' x_i - \boldsymbol{\pi}\boldsymbol{\pi}'\beta_i \|x_i\|_2^2 \right) \qquad \because \boldsymbol{\sigma}_{y_i}^2 = \mathbf{I} \\
&= \widehat{\beta}_i' \left(\boldsymbol{\pi}\boldsymbol{\sigma}_{y_i}\mathbf{w}' x_i - \boldsymbol{\Theta}\beta_i \|x_i\|_2^2 \right) \qquad \because (18)
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
g(\widehat{\boldsymbol{\beta}}) &= \frac{1}{2}\|x_i\|_2^2 \widehat{\beta}_i' \boldsymbol{\Theta} \widehat{\beta}_i + \widehat{\beta}_i' \left(\boldsymbol{\pi}\boldsymbol{\sigma}_{y_i}\mathbf{w}' x_i - \boldsymbol{\Theta}\beta_i \|x_i\|_2^2 - \mathbb{1} \right) + C_i \\
&= \frac{1}{2}\|x_i\|_2^2 \widehat{\beta}_i' \boldsymbol{\Theta} \widehat{\beta}_i - \widehat{\beta}_i' \left(\mathbb{1} - \boldsymbol{\pi}\boldsymbol{\sigma}_{y_i}\mathbf{w}' x_i + \boldsymbol{\Theta}\beta_i \|x_i\|_2^2 \right) + C_i
\end{aligned}$$

Thus, (17) is equivalent to

$$\begin{aligned}
\min_{\widehat{\boldsymbol{\beta}}\in\mathcal{G}} \quad & \frac{1}{2}\|x_i\|_2^2 \widehat{\beta}_i' \boldsymbol{\Theta} \widehat{\beta}_i - \widehat{\beta}_i' \left(\mathbb{1} - \boldsymbol{\pi}\boldsymbol{\sigma}_{y_i}\mathbf{w}' x_i + \boldsymbol{\Theta}\beta_i \|x_i\|_2^2 \right) + C_i \\
s.t. \quad & \widehat{\beta}_s = \beta_s, \, \forall s \in [n] \setminus \{i\}.
\end{aligned}$$

Dropping the constant C_i and dividing through by $\|x_i\|_2^2$ does not change the minimizers. Hence, (17) has the same set of minimizers as

$$\begin{aligned}
\min_{\widehat{\boldsymbol{\beta}}\in\mathcal{G}} \quad & \frac{1}{2} \widehat{\beta}_i' \boldsymbol{\Theta} \widehat{\beta}_i - \widehat{\beta}_i' \left((\mathbb{1} - \boldsymbol{\pi}\boldsymbol{\sigma}_{y_i}\mathbf{w}' x_i)/\|x_i\|_2^2 + \boldsymbol{\Theta}\beta_i \right) \\
s.t. \quad & \widehat{\beta}_s = \beta_s, \, \forall s \in [n] \setminus \{i\}.
\end{aligned}$$

Due to the equality constraints, the only free variable is $\widehat{\beta}_i$. Note that the above optimization, when restricted to $\widehat{\beta}_i$, is equivalent to the optimization (4) with

$$v := (\mathbb{1} - \boldsymbol{\pi}\boldsymbol{\sigma}_{y_i}\mathbf{w}' x_i)/\|x_i\|_2^2 + \boldsymbol{\Theta}\beta_i$$

and $\mathbf{w}$ is as in (1). The uniqueness of the minimizer is guaranteed by Theorem 3.4. □

A.4. Global linear convergence

Wang & Lin (2014) established the global linear convergence of the so-called *feasible descent method* when applied to a certain class of problems. As an application, they prove global linear convergence for coordinate descent for solving the dual problem of the binary SVM with the hinge loss. Wang & Lin (2014) considered optimization problems of the following form:

$$\min_{x\in\mathcal{X}} f(x) := g(\mathbf{E}x) + b'x \tag{20}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a function such that ∇f is Lipschitz continuous, $\mathcal{X} \subseteq \mathbb{R}^n$ is a polyhedral set, $\arg\min_{x\in\mathcal{X}} f(x)$ is nonempty, $g : \mathbb{R}^m \to \mathbb{R}$ is a strongly convex function such that ∇g is Lipschitz continuous, and $\mathbf{E} \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^n$ are fixed matrix and vector, respectively.

Below, let $\mathcal{P}_{\mathcal{X}} : \mathbb{R}^n \to \mathcal{X}$ denote the orthogonal projection on $\mathcal{X}$.

Definition A.2. In the context of (20), an iterative algorithm that produces a sequence $\{x^0, x^1, x^2, \dots\} \subseteq \mathcal{X}$ is a *feasible descent method* if there exists a sequence $\{\epsilon^0, \epsilon^1, \epsilon^2, \dots\} \subseteq \mathbb{R}^n$ such that for all $t \geq 0$

$$x^{t+1} = \mathcal{P}_{\mathcal{X}}\left(x^t - \nabla f(x^t) + \epsilon^t\right) \tag{21}$$

$$\|\epsilon^t\| \leq B\|x^t - x^{t+1}\| \tag{22}$$

$$f(x^t) - f(x^{t+1}) \geq \Gamma\|x^t - x^{t+1}\|^2 \tag{23}$$

where $B, \Gamma > 0$.

One of the main result of (Wang & Lin, 2014) is

Theorem A.3 (Theorem 8 from (Wang & Lin, 2014))**.** *Suppose an optimization problem* $\min_{x\in\mathcal{X}} f(x)$ *is of the form* (20) *and* $\{x^0, x^1, x^2, \dots\} \subseteq \mathcal{X}$ *is a sequence generated by a feasible descent method. Let* $f^* := \min_{x\in\mathcal{X}} f(x)$*. Then there exists* $\Delta \in (0, 1)$ *such that*

$$f(x^{t+1}) - f^* \leq \Delta(f(x^t) - f^*), \quad \forall t \geq 0.$$

Now, we begin verifying that the WW-SVM dual optimization and the BCD algorithm for WW-SVM satisfies the requirements of Theorem A.3.

Given $\boldsymbol{\beta} \in \mathbb{R}^{(k-1)\times n}$, define its vectorization

$$\mathrm{vec}(\boldsymbol{\beta}) = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} \in \mathbb{R}^{(k-1)n}.$$

Define the matrix $\mathbf{P}_{is} = \boldsymbol{\pi}\boldsymbol{\sigma}_{y_i} x_i' x_s \boldsymbol{\sigma}_{y_s}\boldsymbol{\pi}' \in \mathbb{R}^{(k-1)\times(k-1)}$, and $\mathbf{Q} \in \mathbb{R}^{(k-1)n\times(k-1)n}$ by

$$\mathbf{Q} = \begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \cdots & \mathbf{P}_{1n} \\ \mathbf{P}_{21} & \mathbf{P}_{22} & \cdots & \mathbf{P}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{P}_{n1} & \mathbf{P}_{n2} & \cdots & \mathbf{P}_{nn} \end{bmatrix}.$$

Let

$$\mathbf{E} = \begin{bmatrix} x_1\boldsymbol{\sigma}_{y_1}\boldsymbol{\pi}' \\ x_2\boldsymbol{\sigma}_{y_2}\boldsymbol{\pi}' \\ \vdots \\ x_n\boldsymbol{\sigma}_{y_n}\boldsymbol{\pi}' \end{bmatrix}.$$

We observe that $\mathbf{Q} = \mathbf{E}'\mathbf{E}$. Thus, $\mathbf{Q}$ is symmetric and positive semi-definite. Let $\|\mathbf{Q}\|_{op}$ be the operator norm of $\mathbf{Q}$.

Proposition A.4. *The optimization* (D2) *is of the form* (20)*. More precisely, the optimization* (D2) *can be expressed as*

$$\min_{\boldsymbol{\beta}\in\mathcal{G}} \quad g(\boldsymbol{\beta}) = \varphi(\mathbf{E}\mathrm{vec}(\boldsymbol{\beta})) - \mathbb{1}'\mathrm{vec}(\boldsymbol{\beta}) \tag{24}$$

where the feasible set $\mathcal{G}$ *is a nonempty polyhedral set (i.e., defined by a system of linear inequalities, hence convex),* φ *is strongly convex, and* ∇g *is Lipschitz continuous with Lipschitz constant* $L := \|\mathbf{Q}\|_{op}$*. Furthermore,* (24) *has at least one minimizer.*

Proof. Observe

$$\begin{aligned} g(\boldsymbol{\beta}) &= \frac{1}{2}\sum_{i,s\in[n]} x_s' x_i \beta_i' \boldsymbol{\pi}\boldsymbol{\sigma}_{y_i}\boldsymbol{\sigma}_{y_s}\boldsymbol{\pi}'\beta_s - \sum_{i\in[n]} \mathbb{1}'\beta_i \\ &= \frac{1}{2}\mathrm{vec}(\boldsymbol{\beta})'\mathbf{Q}\mathrm{vec}(\boldsymbol{\beta}) - \mathbb{1}'\mathrm{vec}(\boldsymbol{\beta}) \\ &= \frac{1}{2}(\mathbf{E}\mathrm{vec}(\boldsymbol{\beta}))'(\mathbf{E}\mathrm{vec}(\boldsymbol{\beta})) - \mathbb{1}'\mathrm{vec}(\boldsymbol{\beta}) \\ &= \varphi(\mathbf{E}\mathrm{vec}(\boldsymbol{\beta})) - \mathbb{1}'\mathrm{vec}(\boldsymbol{\beta}) \end{aligned}$$

where $\varphi(\bullet) = \frac{1}{2}\|\bullet\|^2$. Note that $\mathrm{vec}(\nabla g(\boldsymbol{\beta})) = \mathbf{Q}\mathrm{vec}(\boldsymbol{\beta}) - \mathbb{1}$. Hence, the Lipschitz constant of g is $\|\mathbf{Q}\|_{op}$. For the "Furthermore" part, note that the above calculation shows that (24) is a quadratic program where the second order term is positive semi-definite and the constraint set is convex. Hence, (24) has at least one minimizer. □

Let $B = [0, C]^{k-1}$. Let $\boldsymbol{\beta}^t$ be $\boldsymbol{\beta}$ at the end of the t-iteration of the outer loop of Algorithm 1. Define

$$\boldsymbol{\beta}^{t,i} := [\beta_1^{t+1}, \cdots, \beta_i^{t+1}, \beta_{i+1}^t, \cdots, \beta_n^t].$$

By construction, we have

$$\beta_i^{t+1} = \operatorname*{arg\,min}_{\beta\in B} g\left([\beta_1^{t+1}, \cdots, \beta_{i-1}^{t+1}, \beta, \beta_{i+1}^t, \cdots, \beta_n^t]\right) \tag{25}$$

For each $i = 1, \ldots, n$, let

$$\nabla_i g(\boldsymbol{\beta}) = \left[\frac{\partial g}{\partial \beta_{1i}}(\boldsymbol{\beta}), \frac{\partial g}{\partial \beta_{2i}}(\boldsymbol{\beta}), \ldots, \frac{\partial g}{\partial \beta_{(k-1)i}}(\boldsymbol{\beta})\right]'.$$

By Lemma 24 (Wang & Lin, 2014), we have

$$\beta_i^{t+1} = \mathcal{P}_B(\beta_i^{t+1} - \nabla_i g(\boldsymbol{\beta}^{t,i}))$$

where $\mathcal{P}_B$ denotes orthogonal projection on to B. Now, define $\boldsymbol{\epsilon}^t \in \mathbb{R}^{(k-1)\times n}$ such that

$$\epsilon_i^t = \beta_i^{t+1} - \beta_i^t - \nabla_i g(\boldsymbol{\beta}^{t,i}) + \nabla_i g(\boldsymbol{\beta}^t).$$

Proposition A.5. *The BCD algorithm for the WW-SVM is a feasible descent method. More precisely, the sequence* $\{\boldsymbol{\beta}^0, \boldsymbol{\beta}^1, \ldots\}$ *satisfies the following conditions:*

$$\boldsymbol{\beta}^{t+1} = \mathcal{P}_{\mathcal{G}}\left(\boldsymbol{\beta}^t - \nabla g(\boldsymbol{\beta}^t) + \boldsymbol{\epsilon}^t\right) \tag{26}$$

$$\|\boldsymbol{\epsilon}^t\| \leq (1 + \sqrt{n}L)\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^{t+1}\| \tag{27}$$

$$g(\boldsymbol{\beta}^t) - g(\boldsymbol{\beta}^{t+1}) \geq \Gamma\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^{t+1}\|^2 \tag{28}$$

where L is as in Proposition A.4, $\Gamma := \min_{i\in[n]} \frac{\|x_i\|^2}{2}$, $\mathcal{G}$ is the feasible set of (D2)*, and $\mathcal{P}_{\mathcal{G}}$ is the orthogonal projection onto $\mathcal{G}$.*

The proof of Proposition A.5 essentially generalizes Proposition 3.4 of (Luo & Tseng, 1993) to the higher dimensional setting:

Proof. Recall that $\mathcal{G} = B^{\times n} := B \times \cdots \times B$. Note that the i-th block of $\boldsymbol{\beta}^t - \nabla g(\boldsymbol{\beta}^t) + \boldsymbol{\epsilon}^t$ is

$$\beta_i^t - \nabla_i g(\boldsymbol{\beta}^t) + \epsilon_i^t = \beta_i^t - \nabla_i g(\boldsymbol{\beta}^t) + (\beta_i^{t+1} - \beta_i^t - \nabla_i g(\boldsymbol{\beta}^{t,i}) + \nabla_i g(\boldsymbol{\beta}^t)) = \beta_i^{t+1} - \nabla_i g(\boldsymbol{\beta}^{t,i}).$$

Thus, the i-th block of $\mathcal{P}_{\mathcal{G}}(\boldsymbol{\beta}^t - \nabla g(\boldsymbol{\beta}^t) + \boldsymbol{\epsilon}^t)$ is

$$\mathcal{P}_B(\beta_i^{t+1} - \nabla_i g(\boldsymbol{\beta}^{t,i})) = \beta_i^{t+1}.$$

This is precisely the identity (26).

Next, we have

$$\begin{aligned}\|\epsilon_i^t\| &\le \|\beta_i^{t+1} - \beta_i^t\| + \|\nabla_i g(\boldsymbol{\beta}^{t,i}) - \nabla_i g(\boldsymbol{\beta}^t)\| \\ &\le \|\beta_i^{t+1} - \beta_i^t\| + L\|\boldsymbol{\beta}^{t,i} - \boldsymbol{\beta}^t\| \\ &\le \|\beta_i^{t+1} - \beta_i^t\| + L\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|.\end{aligned}$$

From this, we get that

$$\begin{aligned}\|\boldsymbol{\epsilon}^t\| &= \sqrt{\sum_{i=1}^n \|\epsilon_i^t\|^2} \\ &\le \sqrt{\sum_{i=1}^n (\|\beta_i^{t+1} - \beta_i^t\| + L\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|)^2} \\ &\le \sqrt{\sum_{i=1}^n \|\beta_i^{t+1} - \beta_i^t\|^2} + \sqrt{\sum_{i=1}^n L^2\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|^2} \\ &= \|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\| + \sqrt{n}L\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\| \\ &= (1 + \sqrt{n}L)\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|.\end{aligned}$$

Thus, we conclude that $\|\boldsymbol{\epsilon}^t\| \le (1 + \sqrt{n}L)\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|$ which is (27).

Finally, we show that

$$g(\boldsymbol{\beta}^{t,i-1}) - g(\boldsymbol{\beta}^{t,i}) + \nabla_i g(\boldsymbol{\beta}^{t,i})'(\beta_i^{t+1} - \beta_i^t) \ge \Gamma\|\beta_i^{t+1} - \beta_i^t\|^2$$

where $\Gamma := \min_{i\in[n]} \frac{\|x_i\|^2}{2}$.

Lemma A.6. *Let $\beta_1, \cdots, \beta_{i-1}, \beta, \beta_{i+1}, \cdots, \beta_n \in \mathbb{R}^{k-1}$ be arbitrary. Then there exist $v \in \mathbb{R}^{k-1}$ and $C \in \mathbb{R}$ which depend only on $\beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_n$, but not on β, such that*

$$g\left([\beta_1, \cdots, \beta_{i-1}, \beta, \beta_{i+1}, \cdots, \beta_n]\right) = \frac{1}{2}\|x_i\|^2\beta'\beta - v'\beta - C.$$

In particular, we have

$$\nabla_i g\left([\beta_1, \cdots, \beta_{i-1}, \beta, \beta_{i+1}, \cdots, \beta_n]\right) = \|x_i\|^2\beta - v.$$

Proof. The result follows immediately from the identity (19). □

Lemma A.7. *Let $\beta_1, \cdots, \beta_{i-1}, \beta, \eta, \beta_{i+1}, \cdots, \beta_n \in \mathbb{R}^{k-1}$ be arbitrary. Then we have*

$$\begin{aligned}&g\left([\beta_1, \cdots, \beta_{i-1}, \eta, \beta_{i+1}, \cdots, \beta_n]\right) - g\left([\beta_1, \cdots, \beta_{i-1}, \beta, \beta_{i+1}, \cdots, \beta_n]\right) \\ &\quad + \nabla_i g\left([\beta_1, \cdots, \beta_{i-1}, \beta, \beta_{i+1}, \cdots, \beta_n]\right)'(\beta - \eta) \\ &= \frac{\|x_i\|^2}{2}\|\eta - \beta\|^2\end{aligned}$$

Proof. Let v, C be as in Lemma A.6. We have

$$\begin{aligned}&g\left([\beta_1, \cdots, \beta_{i-1}, \eta, \beta_{i+1}, \cdots, \beta_n]\right) - g\left([\beta_1, \cdots, \beta_{i-1}, \beta, \beta_{i+1}, \cdots, \beta_n]\right) \\ &= \frac{\|x_i\|^2}{2}\|\eta\|^2 - v'\eta - \frac{\|x_i\|^2}{2}\|\beta\|^2 + v'\beta \\ &= \frac{\|x_i\|^2}{2}(\|\eta\|^2 - \|\beta\|^2) + v'(\beta - \eta)\end{aligned}$$

and

$$\nabla_i g\left([\beta_1, \cdots, \beta_{i-1}, \beta, \beta_{i+1}, \cdots, \beta_n]\right)'(\beta - \eta) = (\|x_i\|^2\beta - v)'(\beta - \eta) = \|x_i\|^2(\|\beta\|^2 - \beta'\eta) - v'(\beta - \eta).$$

Thus,

$$\begin{aligned}
&g\left([\beta_1, \cdots, \beta_{i-1}, \eta, \beta_{i+1}, \cdots, \beta_n]\right) - g\left([\beta_1, \cdots, \beta_{i-1}, \beta, \beta_{i+1}, \cdots, \beta_n]\right) \\
&\quad + \nabla_i g\left([\beta_1, \cdots, \beta_{i-1}, \beta, \beta_{i+1}, \cdots, \beta_n]\right)'(\beta - \eta) \\
&= \frac{\|x_i\|^2}{2}(\|\eta\|^2 - \|\beta\|^2) + v'(\beta - \eta) + \|x_i\|^2(\|\beta\|^2 - \beta'\eta) - v'(\beta - \eta) \\
&= \frac{\|x_i\|^2}{2}(\|\eta\|^2 - \|\beta\|^2) + \|x_i\|^2(\|\beta\|^2 - \beta'\eta) \\
&= \|x_i\|^2\left(\frac{1}{2}(\|\eta\|^2 - \|\beta\|^2) + (\|\beta\|^2 - \beta'\eta)\right) \\
&= \|x_i\|^2\left(\frac{1}{2}(\|\eta\|^2 + \|\beta\|^2) - \beta'\eta\right) \\
&= \frac{\|x_i\|^2}{2}\|\eta - \beta\|^2
\end{aligned}$$

as desired. □

Applying Lemma A.7, we have

$$g(\boldsymbol{\beta}^{t,i-1}) - g(\boldsymbol{\beta}^{t,i}) + \nabla_i g(\boldsymbol{\beta}^{t,i})'(\beta_i^{t+1} - \beta_i^t) \geq \frac{\|x_i\|^2}{2}\|\beta_i^{t+1} - \beta_i^t\|^2.$$

Since (25) is true, we have by Lemma 24 of (Wang & Lin, 2014) that

$$\nabla_i g(\boldsymbol{\beta}^{t,i})'(\beta_i^t - \beta_i^{t+1}) \geq 0$$

Equivalently, $\nabla_i g(\boldsymbol{\beta}^{t,i})'(\beta_i^{t+1} - \beta_i^t) \leq 0$. Thus, we deduce that

$$g(\boldsymbol{\beta}^{t,i-1}) - g(\boldsymbol{\beta}^{t,i}) \geq \frac{\|x_i\|^2}{2}\|\beta_i^{t+1} - \beta_i^t\|^2 \geq \Gamma\|\beta_i^{t+1} - \beta_i^t\|^2$$

Summing the above identity over $i \in [n]$, we have

$$g(\boldsymbol{\beta}^{t,0}) - g(\boldsymbol{\beta}^{t,n}) = \sum_{i=1}^{n} g(\boldsymbol{\beta}^{t,i-1}) - g(\boldsymbol{\beta}^{t,i}) \geq \Gamma\sum_{i=1}^{n}\|\beta_i^{t+1} - \beta_i^t\|^2 = \Gamma\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|^2$$

Since $(\boldsymbol{\beta}^{t,0}) = \boldsymbol{\beta}^t$ and $\boldsymbol{\beta}^{t,n} = \boldsymbol{\beta}^{t+1}$, we conclude that $g(\boldsymbol{\beta}^t) - g(\boldsymbol{\beta}^{t+1}) \geq \Gamma\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^t\|^2$. □

To conclude the proof of Theorem 3.6, we note that Proposition A.5 and Proposition A.4 together imply that the requirements of Theorem 8 from (Wang & Lin, 2014) (restated as Theorem A.3 here) are satisfied for the BCD algorithm for WW-SVM. Hence, we are done. □

A.5. Proof of Theorem 3.4

The goal of this section is to prove Theorem 3.4. The time complexity analysis has been carried out at the end of Section 4 of the main article. Below, we focus on the part of the theorem on the correctness of the output. Throughout this section, $k \geq 2$, $C > 0$ and $v \in \mathbb{R}^{k-1}$ are assumed to be fixed. Additional variables used are summarized in Table 2.

Table 2. Variables used in Section A.5

VARIABLE(S)	DEFINED IN	NOTA BENE
t	ALGORITHM 2	ITERATION INDEX
$\ell, \texttt{vals}, \delta_t, \gamma_t$	SUBROUTINE 3	$t \in [\ell]$ IS AN ITERATION INDEX
$\texttt{up}, \texttt{dn}$	SUBROUTINE 3	SYMBOLS
$\widetilde{b}, \widetilde{\gamma}, v_{\max}$	LEMMA A.9	
$\langle 1 \rangle, \ldots, \langle k-1 \rangle$	ALGORITHM 2	
$n_{\mathtt{m}}^t, n_{\mathtt{u}}^t, S^t, \widehat{\gamma}^t, \widehat{b}^t$	ALGORITHM 2	$t \in [\ell]$ IS AN ITERATION INDEX
$\llbracket k \rrbracket, I_{\mathtt{u}}^\gamma, I_{\mathtt{m}}^\gamma, n_{\mathtt{u}}^\gamma, n_{\mathtt{m}}^\gamma$	DEFINITION A.10	$\gamma \in \mathbb{R}$ IS A REAL NUMBER
$S^{(n_{\mathtt{m}}, n_{\mathtt{u}})}, \widehat{\gamma}^{(n_{\mathtt{m}}, n_{\mathtt{u}})}, \widehat{b}^{(n_{\mathtt{m}}, n_{\mathtt{u}})}$	DEFINITION A.13	$(n_{\mathtt{m}}, n_{\mathtt{u}}) \in \llbracket k \rrbracket^2$
$\texttt{vals}^+$	DEFINITION A.18	
$\mathtt{u}(j), \mathtt{d}(j)$	DEFINITION A.19	$j \in [k-1]$ IS AN INTEGER
$\texttt{crit}_1, \texttt{crit}_2$	DEFINITION A.20	
$\texttt{KKT\_cond}()$	SUBROUTINE 5	

A.5.1. THE CLIPPING MAP

First, we recall the clipping map:

Definition A.8. The *clipping map* $\texttt{clip}_C : \mathbb{R}^{k-1} \to [0, C]^{k-1}$ is the function defined as follows: for $w \in \mathbb{R}^{k-1}$, $[\texttt{clip}_C(w)]_i := \max\{0, \min\{C, w_i\}\}$.

Lemma A.9. *Let* $v_{\max} = \max_{i \in [k-1]} v_i$. *The optimization* (4) *has a unique global minimum* $\widetilde{b}$ *satisfying the following:*

1. $\widetilde{b} = \texttt{clip}_C(v - \widetilde{\gamma}\mathbb{1})$ *for some* $\widetilde{\gamma} \in \mathbb{R}$
2. $\widetilde{\gamma} = \sum_{i=1}^{k-1} \widetilde{b}_i$. *In particular,* $\widetilde{\gamma} \geq 0$.
3. *If* $v_i \leq 0$, *then* $\widetilde{b}_i = 0$. *In particular, if* $v_{\max} \leq 0$, *then* $\widetilde{b} = \mathbb{0}$.
4. *If* $v_{\max} > 0$, *then* $0 < \widetilde{\gamma} < v_{\max}$.

Proof. We first prove part 1. The optimization (4) is a minimization over a convex domain with strictly convex objective, and hence has a unique global minimum $\widetilde{b}$. For each $i \in [k-1]$, let $\lambda_i, \mu_i \in \mathbb{R}$ be the dual variables for the constraints $0 \geq b_i - C$ and $0 \geq -b_i$, respectively. The Lagrangian for the optimization (4) is

$$\mathcal{L}(b, \lambda, \mu) = \frac{1}{2} b'(\mathbf{I} + \mathbf{O})b - v'b + (b - C)'\lambda + (-b)'\mu.$$

Thus, the stationarity (or gradient vanishing) condition is

$$0 = \nabla_b \mathcal{L}(b, \lambda, \mu) = (\mathbf{I} + \mathbf{O})b - v + \lambda - \mu.$$

The KKT conditions are as follows:

for all $i \in [k-1]$, the following holds:

$$[(\mathbf{I} + \mathbf{O})b]_i + \lambda_i - \mu_i = v_i \quad \text{stationarity} \tag{29}$$
$$C \geq b_i \geq 0 \quad \text{primal feasibility} \tag{30}$$
$$\lambda_i \geq 0 \quad \text{dual feasibility} \tag{31}$$
$$\mu_i \geq 0 \quad \text{"} \tag{32}$$
$$\lambda_i(C - b_i) = 0 \quad \text{complementary slackness} \tag{33}$$
$$\mu_i b_i = 0 \quad \text{"} \tag{34}$$

(29) to (34) are satisfied if and only if $b = \widetilde{b}$ is the global minimum.

Let $\widetilde{\gamma} \in \mathbb{R}$ be such that $\widetilde{\gamma}\mathbb{1} = \mathbf{O}\widetilde{b}$. Note that by definition, part 2 holds. Furthermore, (29) implies

$$\widetilde{b} = v - \widetilde{\gamma}\mathbb{1} - \lambda + \mu. \tag{35}$$

Below, fix some $i \in [k-1]$. Note that λ_i or μ_i cannot both be nonzero. Otherwise, (33) and (34) would imply that $C = \widetilde{b}_i = 0$, a contradiction. We claim the following:

1. If $v_i - \widetilde{\gamma} \in [0, C]$, then $\lambda_i = \mu_i = 0$ and $\widetilde{b}_i = v_i - \widetilde{\gamma}$.
2. If $v_i - \widetilde{\gamma} > C$, then $\widetilde{b}_i = C$.
3. $v_i - \widetilde{\gamma} < 0$, then $\widetilde{b}_i = 0$.

We prove the first claim. To this end, suppose $v_i - \widetilde{\gamma} \in [0, C]$. We will show $\lambda_i = \mu_i = 0$ by contradiction. Suppose $\lambda_i > 0$. Then we have $C = \widetilde{b}_i$ and $\mu_i = 0$. Now, (35) implies that $C = \widetilde{b}_i = v_i - \widetilde{\gamma} - \lambda_i$. However, we now have $v_i - \widetilde{\gamma} - \lambda_i \leq C - \lambda_i < C$, a contradiction. Thus, $\lambda_i = 0$. Similarly, assuming $\mu_i > 0$ implies

$$0 = \widetilde{b}_i = v_i - \lambda + \mu_i \geq 0 + \mu_i > 0,$$

a contradiction. This proves the first claim.

Next, we prove the second claim. Note that

$$C \geq \widetilde{b}_i = v_i - \widetilde{\gamma} - \lambda_i + \mu_i > C - \lambda_i + \mu_i \implies 0 > -\lambda_i + \mu_i \geq -\lambda_i.$$

In particular, we have $\lambda_i > 0$ which implies $C = \widetilde{b}_i$ by complementary slackness.

Finally, we prove the third claim. Note that

$$0 \leq \widetilde{b}_i = v_i - \widetilde{\gamma} - \lambda_i + \mu_i < -\lambda_i + \mu_i \leq \mu_i$$

Thus, $\mu_i > 0$ and so $0 = \widetilde{b}_i$ by complementary slackness. This proves that $\widetilde{b} = \texttt{clip}_C(v - \widetilde{\gamma}\mathbb{1})$, which concludes the proof of part 1.

For part 2, note that $\widetilde{\gamma} = \sum_{i=1}^{k-1} \widetilde{b}_i$ holds by definition. The "in particular" portion follows immediately from $\widetilde{b} \geq 0$.

We prove part 3 by contradiction. Suppose there exists $i \in [k-1]$ such that $v_i \leq 0$ and $\widetilde{b}_i > 0$. Thus, by (34), we have $\mu_i = 0$. By (29), we have $b_i + \widetilde{\gamma} \leq b_i + \widetilde{\gamma} + \lambda_i = v_i \leq 0$. Thus, we have $-\widetilde{\gamma} \geq b_i > 0$, or equivalently, $\widetilde{\gamma} < 0$. However, this contradicts part 2. Thus, $\widetilde{b}_i = 0$ whenever $v_i \leq 0$. The " in particular" portion follows immediately from the observation that $v_{\max} \leq 0$ implies that $v_i \leq 0$ for all $i \in [k-1]$.

For part 4, we first prove that $\widetilde{\gamma} < v_{\max}$ by contradiction. Suppose that $\widetilde{\gamma} \geq v_{\max}$. Then we have $v - \widetilde{\gamma}\mathbb{1} \leq v - v_{\max}\mathbb{1} \leq \mathbb{0}$. Thus, by part 1, we have $\widetilde{b} = \texttt{clip}_C(v - \widetilde{\gamma}\mathbb{1}) = \mathbb{0}$. By part 2, we must have that $\widetilde{\gamma} = \sum_{i=1}^{k-1} \widetilde{b}_i = 0$. However, $\widetilde{\gamma} \geq v_{\max} > 0$, which is a contradiction.

Finally, we prove that $\widetilde{\gamma} > 0$ again by contradiction. Suppose that $\widetilde{\gamma} = 0$. Then part 2 and the fact that $\widetilde{b} \geq \mathbb{0}$ implies that $\widetilde{b} = \mathbb{0}$. However, by part 1, we have $\widetilde{b} = \texttt{clip}_C(v)$. Now, let i^* be such that $v_{i^*} = v_{\max}$. This implies that $\widetilde{b}_{i^*} = \texttt{clip}_C(v_{\max}) > 0$, a contradiction. □

A.5.2. Recovering $\widetilde{\gamma}$ from discrete data

Definition A.10. For $\gamma \in \mathbb{R}$, let $b^\gamma := \texttt{clip}_C(v - \gamma\mathbb{1}) \in \mathbb{R}^{k-1}$. Define

$$\begin{aligned} I_{\mathfrak{u}}^\gamma &:= \{i \in [k-1] : b_i^\gamma = C\} \\ I_{\mathfrak{m}}^\gamma &:= \{i \in [k-1] : b_i^\gamma \in (0, C)\} \\ n_{\mathfrak{u}}^\gamma &:= |I_{\mathfrak{u}}^\gamma|, \quad \text{and} \quad n_{\mathfrak{m}}^\gamma := |n_{\mathfrak{m}}^\gamma|. \end{aligned}$$

Let $[\![k]\!] := \{0\} \cup [k-1]$. Note that by definition, $n_{\mathfrak{m}}^\gamma, n_{\mathfrak{u}}^\gamma \in [\![k]\!]$.

Note that $I_{\mathtt{u}}^{\gamma}$ and $I_{\mathtt{m}}^{\gamma}$ are determined by their cardinalities. This is because

$$I_{\mathtt{u}}^{\gamma} = \{\langle 1\rangle, \langle 2\rangle, \ldots, \langle n_{\mathtt{u}}^{\gamma}\rangle\}$$
$$I_{\mathtt{m}}^{\gamma} = \{\langle n_{\mathtt{u}}^{\gamma}+1\rangle, \langle n_{\mathtt{u}}^{\gamma}+2\rangle, \ldots, \langle n_{\mathtt{u}}^{\gamma}+n_{\mathtt{m}}^{\gamma}\rangle\}.$$

Definition A.11. Define

$$\mathtt{disc}^{+} := \{v_i : i \in [k-1],\, v_i > 0\} \cup \{v_i - C : i \in [k-1], v_i - C > 0\} \cup \{0\}.$$

Note that $\mathtt{disc}^{+}$ is slightly different from $\mathtt{disc}$ as defined in the main text.

Lemma A.12. *Let $\gamma', \gamma'' \in \mathtt{disc}^{+}$ be such that $\gamma \notin \mathtt{disc}^{+}$ for all $\gamma \in (\gamma', \gamma'')$. The functions*

$$(\gamma', \gamma'') \ni \gamma \mapsto I_{\mathtt{m}}^{\gamma}$$
$$(\gamma', \gamma'') \ni \gamma \mapsto I_{\mathtt{u}}^{\gamma}$$

are constant.

Proof. We first prove $I_{\mathtt{m}}^{\lambda} = I_{\mathtt{m}}^{\rho}$. Let $\lambda, \rho \in (\gamma', \gamma'')$ be such that $\lambda < \rho$. Assume for the sake of contradiction that $I_{\mathtt{m}}^{\lambda} \neq I_{\mathtt{m}}^{\rho}$. Then either 1) $i \in [k-1]$ such that $v_i - \lambda \in (0, C)$ but $v_i - \rho \notin (0, C)$ or 2) $i \in [k-1]$ such that $v_i - \lambda \notin (0, C)$ but $v_i - \rho \in (0, C)$. This implies that there exists some $\gamma \in (\lambda, \rho)$ such that $v_i - \gamma \in \{0, C\}$, or equivalently, $\gamma \in \{v_i, v_i - C\}$. Hence, $\gamma \in \mathtt{disc}^{+}$, which is a contradiction. Thus, for all $\lambda, \rho \in (\gamma', \gamma'')$, we have $I_{\mathtt{m}}^{\lambda} = I_{\mathtt{m}}^{\rho}$.

Next, we prove $I_{\mathtt{u}}^{\lambda} = I_{\mathtt{u}}^{\rho}$. Let $\lambda, \rho \in (\gamma', \gamma'')$ be such that $\lambda < \rho$. Assume for the sake of contradiction that $I_{\mathtt{u}}^{\lambda} \neq I_{\mathtt{u}}^{\rho}$. Then either 1) $i \in [k-1]$ such that $v_i - \lambda \geq C$ but $v_i - \rho < C$ or 2) $i \in [k-1]$ such that $v_i - \lambda < C$ but $v_i - \rho \geq C$. This implies that there exists some $\gamma \in (\lambda, \rho)$ such that $v_i - \gamma = C$, or equivalently, $\gamma = v_i = C$. Hence, $\gamma \in \mathtt{disc}^{+}$, which is a contradiction. Thus, for all $\lambda, \rho \in (\gamma', \gamma'')$, we have $I_{\mathtt{u}}^{\lambda} = I_{\mathtt{u}}^{\rho}$. □

Definition A.13. For $(n_{\mathtt{m}}, n_{\mathtt{u}}) \in [\![k]\!]^2$, define $S^{(n_{\mathtt{m}}, n_{\mathtt{u}})}, \widehat{\gamma}^{(n_{\mathtt{m}}, n_{\mathtt{u}})} \in \mathbb{R}$ by

$$S^{(n_{\mathtt{m}}, n_{\mathtt{u}})} := \sum_{i=n_{\mathtt{u}}+1}^{n_{\mathtt{u}}+n_{\mathtt{m}}} v_{\langle i\rangle},$$
$$\widehat{\gamma}^{(n_{\mathtt{m}}, n_{\mathtt{u}})} := \left(C \cdot n_{\mathtt{u}} + S^{(n_{\mathtt{m}}, n_{\mathtt{u}})}\right) / (n_{\mathtt{m}} + 1).$$

Furthermore, define $\widehat{b}^{(n_{\mathtt{m}}, n_{\mathtt{u}})} \in \mathbb{R}^{k-1}$ such that, for $i \in [k-1]$, the $\langle i\rangle$-th entry is

$$\widehat{b}_{\langle i\rangle}^{(n_{\mathtt{m}}, n_{\mathtt{u}})} := \begin{cases} C & : i \leq n_{\mathtt{u}} \\ v_{\langle i\rangle} - \gamma^{(n_{\mathtt{m}}, n_{\mathtt{u}})} & : n_{\mathtt{u}} < i \leq n_{\mathtt{u}} + n_{\mathtt{m}} \\ 0 & : n_{\mathtt{u}} + n_{\mathtt{m}} < i. \end{cases}$$

Below, recall ℓ as defined on Subroutine 3-line 2.

Lemma A.14. *Let $t \in [\ell]$. Let $n_{\mathtt{m}}^t$, $n_{\mathtt{u}}^t$, and $\widehat{b}^t$ be as in the for loop of Algorithm 2. Then $\widehat{\gamma}^{(n_{\mathtt{m}}^t, n_{\mathtt{u}}^t)} = \widehat{\gamma}^t$ and $\widehat{b}^{(n_{\mathtt{m}}^t, n_{\mathtt{u}}^t)} = \widehat{b}^t$.*

Proof. It suffices to show that $S^t = S^{(n_{\mathtt{m}}^t, n_{\mathtt{u}}^t)}$ where the former is defined as in Algorithm 2 and the latter is defined as in Definition A.13. In other words, it suffices to show that

$$S^t = \sum_{j \in [k-1] \,:\, n_{\mathtt{u}}^t < j \leq n_{\mathtt{u}}^t + n_{\mathtt{m}}^t} v_{\langle j\rangle}. \tag{36}$$

We prove (36) by induction. The base case $t = 0$ follows immediately due to the initialization in Algorithm 2-line 4.

Now, suppose that (36) holds for S^{t-1}:

$$S^{t-1} = \sum_{j \in [k-1] \,:\, n_{\mathtt{u}}^{t-1} < j \leq n_{\mathtt{u}}^{t-1} + n_{\mathtt{m}}^{t-1}} v_{\langle j\rangle}. \tag{37}$$

Consider the first case that $\delta_t = \texttt{up}$. Then we have $n_{\mathtt{u}}^t + n_{\mathtt{m}}^t = n_{\mathtt{u}}^{t-1} + n_{\mathtt{m}}^{t-1}$ and $n_{\mathtt{u}}^t = n_{\mathtt{u}}^{t-1} + 1$. Thus, we have

$$\begin{aligned} S^t &= S^{t-1} - v_{\langle n_{\mathtt{u}}^{t-1} \rangle} \quad \because \text{Subroutine 4-line 3,} \\ &= \sum_{j \in [k-1] \,:\, n_{\mathtt{u}}^{t-1}+1 < j \le n_{\mathtt{u}}^{t-1}+n_{\mathtt{m}}^{t-1}} v_{\langle j \rangle} \quad \because (37) \\ &= \sum_{j \in [k-1] \,:\, n_{\mathtt{u}}^{t} < j \le n_{\mathtt{u}}^{t}+n_{\mathtt{m}}^{t}} v_{\langle j \rangle} \end{aligned}$$

which is exactly the desired identity in (36).

Consider the second case that $\delta_t = \texttt{dn}$. Then we have $n_{\mathtt{u}}^t + n_{\mathtt{m}}^t = n_{\mathtt{u}}^{t-1} + n_{\mathtt{m}}^{t-1} + 1$ and $n_{\mathtt{u}}^t = n_{\mathtt{u}}^{t-1}$. Thus, we have

$$\begin{aligned} S^t &= S^{t-1} + v_{\langle n_{\mathtt{u}}^{t} + n_{\mathtt{m}}^{t} \rangle} \quad \because \text{Subroutine 4-line 6,} \\ &= \sum_{j \in [k-1] \,:\, n_{\mathtt{u}}^{t-1}+1 < j \le n_{\mathtt{u}}^{t-1}+n_{\mathtt{m}}^{t-1}+1} v_{\langle j \rangle} \quad \because (37) \\ &= \sum_{j \in [k-1] \,:\, n_{\mathtt{u}}^{t} < j \le n_{\mathtt{u}}^{t}+n_{\mathtt{m}}^{t}} v_{\langle j \rangle} \end{aligned}$$

which, again, is exactly the desired identity in (36). □

Lemma A.15. *Let $\widetilde{\gamma}$ be as in Lemma A.9. Then we have*

$$\widetilde{b} = \widehat{b}^{(n_{\mathtt{m}}^{\widetilde{\gamma}}, n_{\mathtt{u}}^{\widetilde{\gamma}})} = \mathtt{clip}_C(v - \widehat{\gamma}^{(n_{\mathtt{m}}^{\widetilde{\gamma}}, n_{\mathtt{u}}^{\widetilde{\gamma}})} \mathbb{1}).$$

Proof. It suffices to prove that $\widetilde{\gamma} = \widehat{\gamma}^{(n_{\mathtt{m}}^{\widetilde{\gamma}}, n_{\mathtt{u}}^{\widetilde{\gamma}})}$. To this end, let $i \in [k-1]$. If $i \in I_{\mathtt{m}}^{\widetilde{\gamma}}$, then $\widetilde{b}_i = v_i - \widetilde{\gamma}$. If $i \in I_{\mathtt{u}}^{\widetilde{\gamma}}$, then $\widetilde{b}_i = C$. Otherwise, $\widetilde{b}_i = 0$. Thus

$$\widetilde{\gamma} = \mathbb{1}'\widetilde{b} = C \cdot n_{\mathtt{u}}^{\widetilde{\gamma}} + S^{(n_{\mathtt{m}}^{\widetilde{\gamma}}, n_{\mathtt{u}}^{\widetilde{\gamma}})} - \widetilde{\gamma} \cdot n_{\mathtt{m}}^{\widetilde{\gamma}}$$

Solving for $\widetilde{\gamma}$, we have

$$\widetilde{\gamma} = \left(C \cdot n_{\mathtt{u}}^{\widetilde{\gamma}} + S^{(n_{\mathtt{m}}^{\widetilde{\gamma}}, n_{\mathtt{u}}^{\widetilde{\gamma}})} \right) / (n_{\mathtt{m}}^{\widetilde{\gamma}} + 1) = \widehat{\gamma}^{(n_{\mathtt{m}}^{\widetilde{\gamma}}, n_{\mathtt{u}}^{\widetilde{\gamma}})},$$

as desired. □

A.5.3. Checking the KKT conditions

Lemma A.16. *Let $(n_{\mathtt{m}}, n_{\mathtt{u}}) \in [\![k]\!]^2$. To simplify notation, let $b := \widehat{b}^{(n_{\mathtt{m}}, n_{\mathtt{u}})}$, $\gamma := \widehat{\gamma}^{(n_{\mathtt{m}}, n_{\mathtt{u}})}$. We have $\mathbf{O}b = \gamma \mathbb{1}$ and for all $i \in [k-1]$ that*

$$[(\mathbf{I} + \mathbf{O})b]_{\langle i \rangle} = \begin{cases} C + \gamma & : i \le n_{\mathtt{u}} \\ v_{\langle i \rangle} & : n_{\mathtt{u}} < i \le n_{\mathtt{u}} + n_{\mathtt{m}} \\ \gamma & : n_{\mathtt{u}} + n_{\mathtt{m}} < i. \end{cases} \tag{38}$$

Furthermore, b satisfies the KKT conditions (29) *to* (34) *if and only if, for all $i \in [k-1]$,*

$$v_{\langle i \rangle} \begin{cases} \ge C + \gamma & : i \le n_{\mathtt{u}} \\ \in [\gamma, C + \gamma] & : n_{\mathtt{u}} < i \le n_{\mathtt{u}} + n_{\mathtt{m}} \\ \le \gamma & : n_{\mathtt{u}} + n_{\mathtt{m}} < i. \end{cases} \tag{39}$$

Proof. First, we prove $\mathbf{O}b = \gamma\mathbb{1}$ which is equivalent to $[\mathbf{O}b]_j = \gamma$ for all $j \in [k-1]$. This is a straightforward calculation:

$$\begin{aligned}
[\mathbf{O}b]_j = \mathbb{1}'b &= \sum_{i\in[k-1]} b_{\langle i\rangle} \\
&= \sum_{i\in[k-1]\,:\,i\le n_{\mathrm{u}}} b_{\langle i\rangle} + \sum_{i\in[k-1]\,:\,n_{\mathrm{u}}<i\le n_{\mathrm{u}}+n_{\mathrm{m}}} b_{\langle i\rangle} + \sum_{i\in[k-1]\,:\,n_{\mathrm{u}}+n_{\mathrm{m}}<i} b_{\langle i\rangle} \\
&= \sum_{i\in[k-1]\,:\,i\le n_{\mathrm{u}}} C + \sum_{i\in[k-1]\,:\,n_{\mathrm{u}}<i\le n_{\mathrm{u}}+n_{\mathrm{m}}} v_{\langle i\rangle} - \gamma \\
&= C\cdot n_{\mathrm{u}} + S^{(n_{\mathrm{m}}^t, n_{\mathrm{u}}^t)} - n_{\mathrm{m}}\gamma \\
&= \gamma.
\end{aligned}$$

Since $[(\mathbf{I}+\mathbf{O})b]_i = [\mathbf{I}b]_i + [\mathbf{O}b]_i$, the identity (38) now follows immediately.

Next, we prove the "Furthermore" part. First, we prove the "only if" direction. By assumption, we have $b = \widetilde{b}$ and so $\gamma = \widetilde{\gamma}$. Furthermore, from Lemma A.9 we have $\widetilde{b} = \mathtt{clip}_C(v - \widetilde{\gamma}\mathbb{1})$ and so $b = \mathtt{clip}_C(v - \gamma\mathbb{1})$. To proceed, recall that by construction, we have

$$b_{\langle i\rangle} = \begin{cases} C & : i \le n_{\mathrm{u}} \\ v - \gamma & : n_{\mathrm{u}} < i \le n_{\mathrm{u}} + n_{\mathrm{m}} \\ 0 & : n_{\mathrm{u}} + n_{\mathrm{m}} < i \end{cases}$$

Thus, if $i \le n_{\mathrm{u}}$, then $C = b_{\langle i\rangle} = [\mathtt{clip}_C(v-\gamma\mathbb{1})]_{\langle i\rangle}$ implies that $v_{\langle i\rangle} - \gamma \ge C$. If $n_{\mathrm{u}} < i \le n_{\mathrm{u}} + n_{\mathrm{m}}$, then $b_{\langle i\rangle} = v_{\langle i\rangle} - \gamma$. Since $b_j \in [0, C]$ for all $j \in [k-1]$, we have in particular that $v_{\langle i\rangle} - \gamma \in [0, C]$. Finally, if $n_{\mathrm{u}} + n_{\mathrm{m}} < i$, then $0 = b_{\langle i\rangle} = [\mathtt{clip}_C(v - \gamma\mathbb{1})]_{\langle i\rangle}$ implies that $v - \gamma \le 0$. In summary,

$$v_{\langle i\rangle} - \gamma \begin{cases} \ge C & : i \le n_{\mathrm{u}} \\ \in [0, C] & : n_{\mathrm{u}} < i \le n_{\mathrm{u}} + n_{\mathrm{m}} \\ \le 0 & : n_{\mathrm{u}} + n_{\mathrm{m}} < i. \end{cases}$$

Note that the above identity immediately implies (39).

Next, we prove the "if" direction. Using (38) and (39), we have

$$[(\mathbf{I}+\mathbf{O})b]_{\langle i\rangle} - v_{\langle i\rangle} \begin{cases} \le 0 & : i \le n_{\mathrm{u}} \\ = 0 & : n_{\mathrm{u}} < i \le n_{\mathrm{u}} + n_{\mathrm{m}} \\ \ge 0 & : n_{\mathrm{u}} + n_{\mathrm{m}} < i. \end{cases}$$

For each $i \in [k-1]$, define $\lambda_i, \mu_i \in \mathbb{R}$ where

$$\lambda_{\langle i\rangle} = \begin{cases} -([(\mathbf{I}+\mathbf{O})b]_{\langle i\rangle} - v_{\langle i\rangle}) & : i \le n_{\mathrm{u}} \\ 0 & : n_{\mathrm{u}} < i \le n_{\mathrm{u}} + n_{\mathrm{m}} \\ 0 & : n_{\mathrm{u}} + n_{\mathrm{m}} < i \end{cases}$$

and

$$\mu_{\langle i\rangle} = \begin{cases} 0 & : i \le n_{\mathrm{u}} \\ 0 & : n_{\mathrm{u}} < i \le n_{\mathrm{u}} + n_{\mathrm{m}} \\ [(\mathbf{I}+\mathbf{O})b]_{\langle i\rangle} - v_{\langle i\rangle} & : n_{\mathrm{u}} + n_{\mathrm{m}} < i. \end{cases}$$

It is straightforward to verify that all of (29) to (34) are satisfied for all $i \in [k-1]$, i.e., the KKT conditions hold at b. □

Recall that we use indices with angle brackets $\langle 1\rangle, \langle 2\rangle, \ldots, \langle k-1\rangle$ to denote a fixed permutation of $[k-1]$ such that

$$v_{\langle 1\rangle} \ge v_{\langle 2\rangle} \ge \cdots \ge v_{\langle k-1\rangle}.$$

Corollary A.17. *Let $t \in [\ell]$ and $\widetilde{b}$ be the unique global minimum of the optimization* (4)*. Then $\widehat{b}^t = \widetilde{b}$ if and only if* `KKT_cond()` *returns true during the t-th iteration of Algorithm 2.*

Proof. First, by Lemma A.9 we have $\widehat{b}^t = \widetilde{b}$ if and only if $\widehat{b}^t$ satisfies the KKT conditions (29) to (34). From Lemma A.14, we have $\widehat{b}^{(n_{\mathtt{m}}^t, n_{\mathtt{u}}^t)} = \widehat{b}^t$ and $\widehat{\gamma}^{(n_{\mathtt{m}}^t, n_{\mathtt{u}}^t)} = \widehat{\gamma}^t$. To simplify notation, let $\gamma = \widehat{\gamma}^{(n_{\mathtt{m}}^t, n_{\mathtt{u}}^t)}$. By Lemma A.16, $\widehat{b}^{(n_{\mathtt{m}}^t, n_{\mathtt{u}}^t)}$ satisfies the KKT conditions (29) to (34) if and only if the following are true:

$$v_{\langle i \rangle} \begin{cases} \geq C + \gamma & : i \leq n_{\mathtt{u}}^t \\ \in [\gamma, C + \gamma] & : n_{\mathtt{u}}^t < i \leq n_{\mathtt{u}}^t + n_{\mathtt{m}}^t \\ \leq \gamma & : n_{\mathtt{u}}^t + n_{\mathtt{m}}^t < i. \end{cases}$$

Since $v_{\langle 1 \rangle} \geq v_{\langle 2 \rangle} \geq \cdots$, the above system of inequalities holds for all $i \in [k-1]$ if and only if

$$\begin{cases} C + \gamma \leq v_{\langle n_{\mathtt{u}}^t \rangle} & : \text{if } n_{\mathtt{u}}^t > 0. \\ \gamma \leq v_{\langle n_{\mathtt{u}}^t + n_{\mathtt{m}}^t \rangle} \text{ and } v_{\langle n_{\mathtt{u}}^t + 1 \rangle} \leq C + \gamma & : \text{if } n_{\mathtt{m}}^t > 0, \\ v_{\langle n_{\mathtt{u}}^t + n_{\mathtt{m}}^t + 1 \rangle} \leq \gamma & : \text{if } n_{\mathtt{u}}^t + n_{\mathtt{m}}^t < k - 1. \end{cases}$$

Note that the above system holds if and only if `KKT_cond()` returns true. □

A.5.4. The variables $n_{\mathtt{m}}^t$ and $n_{\mathtt{u}}^t$

Definition A.18. Define the set $\mathtt{vals}^+ = \{(v_j, \mathtt{dn}, j) : v_j > 0,\ j = 1, \ldots, k-1\} \cup \{(v_j - C, \mathtt{up}, j) : v_j > C,\ j = 1, \ldots, k-1\}$. Sort the set $\mathtt{vals}^+ = \{(\gamma_1, \delta_1, j_1), \ldots, (\gamma_\ell, \delta_\ell, j_\ell)\}$ so that the ordering of $\{(\gamma_1, \delta_1), \ldots, (\gamma_\ell, \delta_\ell)\}$ is identical to `vals` from Subroutine 3-line 2.

To illustrate the definitions, we consider the following running example

$\langle j \rangle$	=	$\langle 1 \rangle$	$\langle 2 \rangle$	$\langle 3 \rangle$	$\langle 4 \rangle$	$\langle 5 \rangle$	$\langle 6 \rangle$	$\langle 7 \rangle$	$\langle 8 \rangle$	$\langle 9 \rangle$	$\langle 10 \rangle$
$v_{\langle j \rangle}$	=	1.8	1.4	1.4	1.4	1.2	0.7	0.4	0.4	0.1	−0.2

t	=	1	2	3	4	5	6	7	8	9	10	11	12	13	14
γ_t	=	1.8	1.4	1.4	1.4	1.2	0.8	0.7	0.4	0.4	0.4	0.4	0.4	0.2	0.1
δ_t	=	dn	dn	dn	dn	dn	up	dn	up	up	up	dn	dn	up	dn

Definition A.19. Define

$$\mathtt{u}(j) := \max\{\tau \in [\ell] : v_{\langle j \rangle} - C = \gamma_\tau\}, \quad \text{and} \quad \mathtt{d}(j) := \max\{\tau \in [\ell] : v_{\langle j \rangle} = \gamma_\tau\}, \tag{40}$$

where $\max \emptyset = \ell + 1$.

Below, we compute $\mathtt{d}(3), \mathtt{d}(6)$ and $\mathtt{u}(3)$ for our running example.

					$\mathtt{d}(3)$ ↓			$\mathtt{d}(6)$ ↓					$\mathtt{u}(3)$ ↓		
t	=	1	2	3	4	5	6	7	8	9	10	11	12	13	14
γ_t	=	1.8	1.4	1.4	1.4	1.2	0.8	0.7	0.4	0.4	0.4	0.4	0.4	0.2	0.1
δ_t	=	dn	dn	dn	dn	dn	up	dn	up	up	up	dn	dn	up	dn

Definition A.20. Define the following sets

$$\mathtt{crit}_1(v) = \{\tau \in [\ell] : \gamma_\tau > \gamma_{\tau+1}\}$$
$$\mathtt{crit}_2(v) = \{\tau \in [\ell] : \gamma_\tau = \gamma_{\tau+1},\ \delta_\tau = \mathtt{up},\ \delta_{\tau+1} = \mathtt{dn}\}$$

where $\gamma_{\ell+1} = 0$.

Below, we illustrate the definition in our running example. The arrows $\downarrow$ and $\Downarrow$ point to elements of $\texttt{crit}_1(v)$ and $\texttt{crit}_2(v)$, respectively.

		$\downarrow$			$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$			$\Downarrow$		$\downarrow$	$\downarrow$	$\downarrow$
t	=	1	2	3	4	5	6	7	8	9	10	11	12	13	14
γ_t	=	1.8	1.4	1.4	1.4	1.2	0.8	0.7	0.4	0.4	0.4	0.4	0.4	0.2	0.1
δ_t	=	dn	dn	dn	dn	dn	up	dn	up	up	up	dn	dn	up	dn

Later, we will show that Algorithm 2 will halt and output the global optimizer $\widetilde{b}$ on or before the t-th iteration where $t \in \texttt{crit}_1(v) \cup \texttt{crit}_2(v)$.

Lemma A.21. *Suppose that $t \in \texttt{crit}_1(v)$. Then*

$$\#\{j \in [k-1] : \texttt{d}(j) \le t\} = \#\{\tau \in [t] : \delta_\tau = \texttt{dn}\}, \quad \textit{and} \quad \#\{j \in [k-1] : \texttt{u}(j) \le t\} = \#\{\tau \in [t] : \delta_\tau = \texttt{up}\}.$$

Proof. First, we observe that

$$\#\{\tau \in [t] : \delta_\tau = \texttt{up}\} = \#\{(\gamma, \delta, j') \in \texttt{vals}^+ : \delta = \texttt{up},\, \gamma \ge \gamma_t\}$$

Next, note that $j \mapsto (\gamma_{\texttt{d}(j)}, \texttt{up}, \langle j \rangle)$ is a bijection from $\{j \in [k-1] : \texttt{d}(j) \le t\}$ to $\{(\gamma, \delta, j') \in \texttt{vals}^+ : \delta = \texttt{up},\, \gamma \ge \gamma_t\}$. To see this, we view the permutation $\langle 1 \rangle, \langle 2 \rangle, \ldots$ viewed as a bijective mapping $\langle \cdot \rangle : [k-1] \to [k-1]$ given by $j \mapsto \langle j \rangle$. Denote by $\rangle \cdot \langle$ the inverse of $\langle \cdot \rangle$. Then the (two-sided) inverse to $j \mapsto (\gamma_{\texttt{d}(j)}, \texttt{up}, \langle j \rangle)$ is clearly given by $(\gamma, \texttt{up}, j') \mapsto \rangle j' \langle$. This proves the first identity of the lemma.

The proof of the second identity is completely analogous. □

Lemma A.22. *The functions* $\texttt{u}$ *and* $\texttt{d} : [k-1] \to [\ell+1]$ *are non-decreasing. Furthermore, for all $j \in [k-1]$, we have $\texttt{u}(j) < \texttt{d}(j)$.*

Proof. Let $j', j'' \in [k-1]$ be such that $j' < j''$. By the sorting, we have $v_{\langle j' \rangle} \ge v_{\langle j'' \rangle}$. Now, suppose that $\texttt{d}(j') > \texttt{d}(j'')$, then by construction we have $\gamma_{\texttt{d}(j')} < \gamma_{\texttt{d}(j'')}$. On the other hand, we have

$$\gamma_{\texttt{d}(j')} = v_{\langle j' \rangle} \ge v_{\langle j'' \rangle} = \gamma_{\texttt{d}(j'')}$$

which is a contradiction.

For the "Furthermore" part, suppose the contrary that $\texttt{u}(j) \ge \texttt{d}(j)$. Then we have $\gamma_{\texttt{u}(j)} \le \gamma_{\texttt{d}(j)}$. However, by definition, we have $\gamma_{\texttt{u}(j)} = v_{\langle j \rangle} > v_{\langle j \rangle} - C = \gamma_{\texttt{d}(j)}$. This is a contradiction. □

Lemma A.23. *Let $t \in \texttt{crit}_1(v)$. Then $n^t_{\texttt{u}} = \#\{j \in [k-1] : \texttt{u}(j) \le t\}$. Furthermore, $[n^t_{\texttt{u}}] = \{j \in [k-1] : \texttt{u}(j) \le t\}$. Equivalently, for each $j \in [k-1]$, we have $j \le n^t_{\texttt{u}}$ if and only if $\texttt{u}(j) \le t$.*

Proof. First, we note that

$$\begin{aligned} n^t_{\texttt{u}} &= \#\{\tau \in [t] : \delta_\tau = \texttt{up}\} \quad \because \text{Subroutine 4-line 2} \\ &= \#\{j \in [k-1] : \texttt{u}(j) \le t\} \quad \because \text{Lemma A.21} \end{aligned}$$

This proves the first part. For the "Furthermore" part, let $N := \#\{j \in [k-1] : \texttt{u}(j) \le t\}$. Since $\texttt{u}$ is monotonic non-decreasing (Lemma A.22), we have $\{j \in [k-1] : \texttt{u}(j) \le t\} = [N]$. Since $N = n^t_{\texttt{u}}$ by the first part, we are done. □

Lemma A.24. *Let $\hat{t}, \check{t} \in \texttt{crit}_1(v)$ be such that there exists $t \in [\ell]$ where*

$$n^t_{\texttt{m}} = \#\{j \in [k-1] : \texttt{d}(j) \le \check{t}\,\} - \#\{j \in [k-1] : \texttt{u}(j) \le \hat{t}\,\}. \tag{41}$$

Then $\texttt{d}(j) \le \check{t}$ and $\hat{t} < \texttt{u}(j)$ if and only if $n^{\hat{t}}_{\texttt{u}} < j \le n^{\hat{t}}_{\texttt{u}} + n^t_{\texttt{m}}$.

Proof. By Lemma A.23 and (41), we have $\#\{j \in [k-1] : \mathtt{d}(j) \le \check{t}\} = n_{\mathtt{u}}^{\hat{t}} + n_{\mathtt{m}}^{t}$. By Lemma A.22, $\mathtt{d}$ is monotonic non-decreasing and so $[n_{\mathtt{u}}^{\hat{t}} + n_{\mathtt{m}}^{t}] = \{j \in [k-1] : \mathtt{d}(j) \le \check{t}\}$. Now,

$$\begin{aligned}
&\{j \in [k-1] : \mathtt{d}(j) \le \check{t},\, \hat{t} < \mathtt{u}(j)\} \\
&= \{j \in [k-1] : \mathtt{d}(j) \le \check{t}\} \cap \{j \in [k-1] : \hat{t} < \mathtt{u}(j)\} \\
&= \{j \in [k-1] : \mathtt{d}(j) \le \check{t}\} \setminus \{j \in [k-1] : \mathtt{u}(j) \le \hat{t}\} \\
&= [n_{\mathtt{u}}^{\hat{t}} + n_{\mathtt{m}}^{t}] \setminus [n_{\mathtt{u}}^{\hat{t}}],
\end{aligned}$$

where in the last equality, we used Lemma A.23. □

Corollary A.25. *Let $t \in \mathtt{crit}_1(v)$. Then $\mathtt{d}(j) \le t$ and $t < \mathtt{u}(j)$ if and only if $n_{\mathtt{u}}^{t} < j \le n_{\mathtt{u}}^{t} + n_{\mathtt{m}}^{t}$.*

Proof. We apply Lemma A.24 with $t = \hat{t} = \check{t}$, which requires checking that

$$n_{\mathtt{m}}^{t} = \#\{j \in [k-1] : \mathtt{d}(j) \le t\} - \#\{j \in [k-1] : \mathtt{u}(j) \le t\}.$$

This is true because from Subroutine 4-line 2 and 5, we have

$$n_{\mathtt{m}}^{t} = \#\{\tau \in [t] : \delta_\tau = \mathtt{dn}\} - \#\{\tau \in [t] : \delta_\tau = \mathtt{up}\}.$$

Applying Lemma A.21, we are done. □

Lemma A.26. *Let $t \in \mathtt{crit}_1(v)$. Let $\varepsilon > 0$ be such that for all $\tau, \tau' \in \mathtt{crit}_1(v)$ where $\tau' < \tau$, we have $\gamma_{\tau'} - \varepsilon > \gamma_\tau$. Then $(n_{\mathtt{m}}^{t}, n_{\mathtt{u}}^{t}) = (n_{\mathtt{m}}^{\gamma_t - \varepsilon}, n_{\mathtt{u}}^{\gamma_t - \varepsilon})$.*

Proof. We claim that

$$v_{\langle j \rangle} - \gamma_t + \varepsilon \begin{cases} < 0 & : t < \mathtt{d}(j) \\ \in (0, C) & : \mathtt{d}(j) \le t < \mathtt{u}(j) \\ > C & : \mathtt{u}(j) \le t. \end{cases} \tag{42}$$

To prove the $t < \mathtt{d}(j)$ case of (42), we have

$$\begin{aligned}
v_{\langle j \rangle} - \gamma_t + \varepsilon &= \gamma_{\mathtt{d}(j)} - \gamma_t + \varepsilon \quad \because (40) \\
&< -\varepsilon + \varepsilon = 0 \quad \because\ t < \mathtt{d}(j) \text{ implies that } \gamma_t - \varepsilon > \gamma_{\mathtt{d}(j)}.
\end{aligned}$$

To prove the $\mathtt{d}(j) \le t < \mathtt{u}(j)$ case of (42), we note that

$$\begin{aligned}
v_{\langle j \rangle} - \gamma_t + \varepsilon &= \gamma_{\mathtt{d}(j)} - \gamma_t + \varepsilon \quad (40) \\
&\ge \varepsilon > 0 \quad \because \mathtt{d}(j) \le t \text{ implies } \gamma_{\mathtt{d}(j)} \ge \gamma_t.
\end{aligned}$$

For the other inequality,

$$\begin{aligned}
v_{\langle j \rangle} - \gamma_t + \varepsilon &= \gamma_{\mathtt{u}(j)} + C - \gamma_t + \varepsilon \quad \because (40) \\
&< -\varepsilon + C + \varepsilon = C \quad \because t < \mathtt{u}(j) \text{ implies } \gamma_t - \varepsilon > \gamma_{\mathtt{u}(j)}.
\end{aligned}$$

Finally, we prove the $\mathtt{u}(j) \le t$ case of (42). Note that

$$\begin{aligned}
v_{\langle j \rangle} - \gamma_t + \varepsilon &= \gamma_{\mathtt{u}(j)} + C - \gamma_t + \varepsilon \quad \because (40) \\
&\ge C + \varepsilon > C \quad \because \mathtt{u}(j) \le t \text{ implies that } \gamma_{\mathtt{u}(j)} \ge \gamma_t.
\end{aligned}$$

Thus, we have proven (42). By Lemma A.23 and Corollary A.25, (42) can be rewritten as

$$v_{\langle j \rangle} - \gamma_t + \varepsilon \begin{cases} < 0 & : n_{\mathtt{u}}^{t} + n_{\mathtt{m}}^{t} < j, \\ \in (0, C) & : n_{\mathtt{u}}^{t} < j \le n_{\mathtt{u}}^{t} + n_{\mathtt{m}}^{t}, \\ > C & : j \le n_{\mathtt{u}}^{t}. \end{cases} \tag{43}$$

Thus, we have $I_{\mathtt{u}}^{\gamma_t - \varepsilon} = \{\langle 1 \rangle, \ldots, \langle n_{\mathtt{u}}^{t} \rangle\}$ and $I_{\mathtt{m}}^{\gamma_t - \varepsilon} = \{\langle n_{\mathtt{u}}^{t} + 1 \rangle, \ldots, \langle n_{\mathtt{u}}^{t} + n_{\mathtt{m}}^{t} \rangle\}$. By the definitions of $n_{\mathtt{u}}^{\gamma_t - \varepsilon}$ and $n_{\mathtt{m}}^{\gamma_t - \varepsilon}$, we are done. □

								$\check{t}$			t		$\hat{t}$		$\mathtt{d}(9)$
								$\downarrow$			$\Downarrow$		$\downarrow$		$\downarrow$
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
γ_t	=	1.8	1.4	1.4	1.4	1.2	0.8	0.7	0.4	0.4	0.4	0.4	0.4	0.2	0.1
δ_t	=	dn	dn	dn	dn	dn	up	dn	up	up	up	dn	dn	up	dn

Figure 3. Example of a critical iterate type 2. The first case that $\check{t} < \mathtt{d}(j)$ where $j = 9$.

Lemma A.27. *Let* $t \in \mathtt{crit}_2(v)$. *Then* $(n_{\mathtt{m}}^t, n_{\mathtt{u}}^t) = (n_{\mathtt{m}}^{\gamma_t}, n_{\mathtt{u}}^{\gamma_t})$.

Proof. Let $\hat{t} \in \mathtt{crit}_1(v)$ be such that $\gamma_{\hat{t}} = \gamma_t$, and $\check{t} = \max\{\tau \in \mathtt{crit}_1(v) : \gamma_\tau > \gamma_t\}$. We claim that

$$v_{\langle j \rangle} - \gamma_{\hat{t}} \begin{cases} \leq 0 & : \check{t} < \mathtt{d}(j), \\ \in (0, C) & : \mathtt{d}(j) \leq \check{t},\ \hat{t} < \mathtt{u}(j), \\ \geq C & : \mathtt{u}(j) \leq \hat{t}. \end{cases} \tag{44}$$

Note that by definition, we have $\gamma_{\check{t}} > \gamma_{\hat{t}}$, which implies that $\check{t} < \hat{t}$.

Consider the first case of (44) that $\check{t} < \mathtt{d}(j)$. See the running example Figure 3. We have by construction that $v_{\langle j \rangle} = \gamma_{\mathtt{d}(j)}$ and so $v_{\langle j \rangle} - \gamma_t = \gamma_{\mathtt{d}(j)} - \gamma_{\check{t}} \leq 0$.

Next, consider the case when $\mathtt{d}(j) \leq \check{t}$ and $\hat{t} < \mathtt{u}(j)$. Thus,

$$\begin{aligned} v_{\langle j \rangle} - \gamma_{\hat{t}} &> v_{\langle j \rangle} - \gamma_{\check{t}} && \because \gamma_{\check{t}} > \gamma_t \\ &= \gamma_{\mathtt{d}(j)} - \gamma_{\check{t}} && \because \text{definition of } \mathtt{d}(j) \\ &\geq 0 && \because \mathtt{d}(j) \leq \check{t} \implies \gamma_{\mathtt{d}(j)} \geq \gamma_{\check{t}}. \end{aligned}$$

On the other hand

$$\begin{aligned} v_{\langle j \rangle} - \gamma_{\hat{t}} &= \gamma_{\mathtt{u}(j)} + C - \gamma_{\hat{t}} && \because \text{definition of } \mathtt{u}(j) \\ &< C && \because \hat{t} < \mathtt{u}(j) \implies \gamma_{\hat{t}} > \gamma_{\mathtt{u}(j)} \end{aligned}$$

Thus, we've shown that in the second case, we have $v_{\langle j \rangle} - \gamma_{\hat{t}} \in (0, C)$.

We consider the final case that $\mathtt{u}(j) \leq \hat{t}$. We have

$$\begin{aligned} v_{\langle j \rangle} - \gamma_{\hat{t}} &= \gamma_{\mathtt{u}(j)} + C - \gamma_{\hat{t}} && \because \text{definition of } t \\ &\geq C && \because \mathtt{u}(j) \leq \hat{t} \implies \gamma_{\mathtt{u}(j)} \geq \gamma_{\hat{t}}. \end{aligned}$$

Thus, we have proven (44).

Next, we claim that $t, \hat{t}, \check{t}$ satisfy the condition (41) of Lemma A.24, i.e.,

$$n_{\mathtt{m}}^t = \#\{j \in [k-1] : \mathtt{d}(j) \leq \check{t}\,\} - \#\{j \in [k-1] : \mathtt{u}(j) \leq \hat{t}\,\}.$$

To this end, we first recall that

$$n_{\mathtt{m}}^t = \#\{\tau \in [t] : \delta_\tau = \mathtt{dn}\} - \#\{\tau \in [t] : \delta_\tau = \mathtt{up}\}.$$

By assumption on t, for all τ such that $\check{t} < \tau \leq t$, we have $\delta_\tau = \mathtt{up}$. Thus,

$$\#\{\tau \in [t] : \delta_\tau = \mathtt{dn}\} = \#\{\tau \in [\check{t}\,] : \delta_\tau = \mathtt{dn}\} = \#\{j \in [k-1] : \mathtt{d}(j) \leq \check{t}\,\}$$

where for the last equality, we used Lemma A.21. Similarly, for all τ such that $t < \tau \leq \hat{t}$, we have $\delta_\tau = \mathtt{dn}$. Thus, we get that analogous result

$$n_{\mathtt{u}}^t = \#\{\tau \in [t] : \delta_\tau = \mathtt{up}\} = \#\{\tau \in [\hat{t}\,] : \delta_\tau = \mathtt{up}\} = \#\{j \in [k-1] : \mathtt{u}(j) \leq \hat{t}\,\} = n_{\mathtt{u}}^{\hat{t}}. \tag{45}$$

Thus, we have verified the condition (41) of Lemma A.24. Now, applying Lemma A.23 and Lemma A.24, we get

$$v_{\langle j\rangle} - \gamma_{\hat{t}} \begin{cases} \leq 0 & : n_{\mathtt{u}}^{\hat{t}} + n_{\mathtt{m}}^{t} < j, \\ \in (0, C) & : n_{\mathtt{u}}^{\hat{t}} < j \leq n_{\mathtt{u}}^{\hat{t}} + n_{\mathtt{m}}^{t} \\ \geq C & : j \leq n_{\mathtt{u}}^{\hat{t}}. \end{cases} \tag{46}$$

By (45) and that $\gamma_t = \gamma_{\hat{t}}$, the above reduces to

$$v_{\langle j\rangle} - \gamma_{t} \begin{cases} \leq 0 & : n_{\mathtt{u}}^{t} + n_{\mathtt{m}}^{t} < j, \\ \in (0, C) & : n_{\mathtt{u}}^{t} < j \leq n_{\mathtt{u}}^{t} + n_{\mathtt{m}}^{t} \\ \geq C & : j \leq n_{\mathtt{u}}^{t}. \end{cases} \tag{47}$$

Thus, $I_{\mathtt{u}}^{\gamma_t} = \{\langle 1\rangle, \ldots, \langle n_{\mathtt{u}}^t\rangle\}$ and $I_{\mathtt{m}}^{\gamma_t} = \{\langle n_{\mathtt{u}}^t + 1\rangle, \ldots, \langle n_{\mathtt{u}}^t + n_{\mathtt{m}}^t\rangle\}$. By the definitions of $n_{\mathtt{u}}^{\gamma_t}$ and $n_{\mathtt{m}}^{\gamma_t}$, we are done. □

A.5.5. PUTTING IT ALL TOGETHER

If $v_{\max} \leq 0$, then Algorithm 2 returns $\mathbb{0}$.

Otherwise, by Lemma A.9, we have $\widetilde{\gamma} \in (0, v_{\max})$.

Lemma A.28. *Let $t \in [\ell]$ be such that $(n_{\mathtt{m}}^t, n_{\mathtt{u}}^t) = (n_{\mathtt{m}}^{\widetilde{\gamma}}, n_{\mathtt{u}}^{\widetilde{\gamma}})$. Then during the t-th loop of Algorithm 2 we have $\widetilde{b} = \widehat{b}^t$ and* `KKT_cond()` *returns true. Consequently, Algorithm 2 returns the optimizer $\widetilde{b}$ on or before the t-th iteration.*

Proof. We have

$$\begin{aligned} \widetilde{b} &= \widehat{b}^{(n_{\mathtt{m}}^{\widetilde{\gamma}}, n_{\mathtt{u}}^{\widetilde{\gamma}})} && \because \text{ Lemma A.15} \\ &= \widehat{b}^{(n_{\mathtt{m}}^{t}, n_{\mathtt{u}}^{t})} && \because \text{Assumption} \\ &= \widehat{b}^{t} && \because \text{ Lemma A.14.} \end{aligned}$$

Thus, by Corollary A.17 `KKT_cond()` returns true on the t-th iteration. This means that Algorithm 2 halts on or before iteration t. Let $\tau \in [\ell]$ be the iteration where Algorithm 2 halts and outputs $\widehat{b}^\tau$. Then $\tau \leq t$. Furthermore, by Corollary A.17, $\widehat{b}^\tau = \widetilde{b}$, which proves the "Consequently" part of the lemma. □

By Lemma A.28, it suffices to show that $(n_{\mathtt{m}}^t, n_{\mathtt{u}}^t) = (n_{\mathtt{m}}^{\widetilde{\gamma}}, n_{\mathtt{u}}^{\widetilde{\gamma}})$ for some $t \in [\ell]$.

We first consider the case when $\widetilde{\gamma} \neq \gamma_t$ for any $t \in \mathtt{crit}_1(v)$. Thus, there exists $t \in \mathtt{crit}_1(v)$ such that $\gamma_{t+1} < \widetilde{\gamma} < \gamma_t$, where we recall that $\gamma_{\ell+1} := 0$.

Now, we return to the proof of Theorem 3.4.

$$\begin{aligned} (n_{\mathtt{m}}^t, n_{\mathtt{u}}^t) &= (n_{\mathtt{m}}^{\gamma_t - \varepsilon}, n_{\mathtt{u}}^{\gamma_t - \varepsilon}) && \because \text{ Lemma A.26} \\ &= (n_{\mathtt{m}}^{\widetilde{\gamma}}, n_{\mathtt{u}}^{\widetilde{\gamma}}) && \because \text{ Lemma A.12, and that both } \widetilde{\gamma} \text{ and } \gamma_i - \varepsilon \in (\gamma_{t+1}, \gamma_t). \end{aligned}$$

Thus, Lemma A.28 implies the result of Theorem 3.4 under the assumption that $\widetilde{\gamma} \neq \gamma_t$ for any $t \in \mathtt{crit}_1(v)$.

Next, we consider when $\widetilde{\gamma} = \gamma_t$ for some $t \in \mathtt{crit}_1(v)$. There are three possibilities:

1. There does not exist $j \in [k-1]$ such that $v_{\langle j\rangle} = \gamma_t$,
2. There does not exist $j \in [k-1]$ such that $v_{\langle j\rangle} - C = \gamma_t$,
3. There exist $j_1, j_2 \in [k-1]$ such that $v_{\langle j_1\rangle} = \gamma_t$ and $v_{\langle j_2\rangle} - C = \gamma_t$.

First, we consider case 1. We claim that

$$(n_{\mathtt{m}}^{\gamma_t}, n_{\mathtt{u}}^{\gamma_t}) = (n_{\mathtt{m}}^{\gamma_t - \varepsilon'}, n_{\mathtt{u}}^{\gamma_t - \varepsilon'}) \quad \text{for all } \varepsilon' > 0 \text{ sufficiently small.} \tag{48}$$

We first note that $n_{\mathrm{u}}^{\gamma_t} = n_{\mathrm{u}}^{\gamma_t - \varepsilon'}$ for all $\varepsilon' > 0$ sufficiently small. To see this, let $i \in [k-1]$ be arbitrary. Note that

$$\begin{aligned} i \in I_{\mathrm{u}}^{\gamma_t} &\iff v_i - \gamma_t \geq C \iff v_i - \gamma_t + \varepsilon' \geq C,\ \forall \epsilon' > 0,\ \text{sufficiently small} \\ &\iff i \in I_{\mathrm{u}}^{\gamma_t - \varepsilon'}, \forall \epsilon' > 0,\ \text{sufficiently small}. \end{aligned}$$

Next, we show that $n_{\mathrm{m}}^{\gamma_t} = n_{\mathrm{m}}^{\gamma_t - \varepsilon'}$ for all $\varepsilon' > 0$ sufficiently small. To see this, let $i \in [k-1]$ be arbitrary. Note that

$$\begin{aligned} i \in I_{\mathrm{m}}^{\gamma_t} &\iff v_i - \gamma_t \in (0, C) \overset{\dagger}{\iff} v_i - \gamma_t + \varepsilon' \in (0, C),\ \forall \epsilon' > 0,\ \text{sufficiently small} \\ &\iff i \in I_{\mathrm{m}}^{\gamma_t - \varepsilon'}, \forall \epsilon' > 0,\ \text{sufficiently small} \end{aligned}$$

where at "$\overset{\dagger}{\iff}$", we used the fact that $v_i - \gamma_t \neq 0$ for any $i \in [k-1]$. Thus, we have proven (48). Taking $\varepsilon' > 0$ so small so that both (48) and the condition in Lemma A.26 hold, we have

$$(n_{\mathrm{m}}^t, n_{\mathrm{u}}^t) = (n_{\mathrm{m}}^{\gamma_t - \varepsilon'}, n_{\mathrm{u}}^{\gamma_t - \varepsilon'}) = (n_{\mathrm{m}}^{\gamma_t}, n_{\mathrm{u}}^{\gamma_t}) = (n_{\mathrm{m}}^{\widetilde{\gamma}}, n_{\mathrm{u}}^{\widetilde{\gamma}}).$$

This proves Theorem 3.4 under case 1.

Next, we consider case 2. We claim that

$$(n_{\mathrm{m}}^{\gamma_t}, n_{\mathrm{u}}^{\gamma_t}) = (n_{\mathrm{m}}^{\gamma_t + \varepsilon''}, n_{\mathrm{u}}^{\gamma_t + \varepsilon''}) \quad \text{for all } \varepsilon'' > 0 \text{ sufficiently small.} \tag{49}$$

We first note that $n_{\mathrm{u}}^{\gamma_t} = n_{\mathrm{u}}^{\gamma_t - \varepsilon''}$ for all $\varepsilon'' > 0$ sufficiently small. To see this, let $i \in [k-1]$ be arbitrary. Note that

$$\begin{aligned} i \in I_{\mathrm{u}}^{\gamma_t} &\iff v_i - \gamma_t \geq C \overset{\ddagger}{\iff} v_i - \gamma_t - \varepsilon'' \geq C,\ \forall \epsilon'' > 0,\ \text{sufficiently small} \\ &\iff i \in I_{\mathrm{u}}^{\gamma_t + \varepsilon''}, \forall \epsilon'' > 0,\ \text{sufficiently small}. \end{aligned}$$

where at "$\overset{\ddagger}{\iff}$", we used the fact that $v_i - \gamma_t \neq C$ for any $i \in [k-1]$. Next, we show that $n_{\mathrm{m}}^{\gamma_t} = n_{\mathrm{m}}^{\gamma_t - \varepsilon''}$ for all $\varepsilon'' > 0$ sufficiently small. To see this, let $i \in [k-1]$ be arbitrary. Note that

$$\begin{aligned} i \in I_{\mathrm{m}}^{\gamma_t} &\iff v_i - \gamma_t \in (0, C) \overset{\ddagger}{\iff} v_i - \gamma_t - \varepsilon'' \in (0, C),\ \forall \epsilon'' > 0,\ \text{sufficiently small} \\ &\iff i \in I_{\mathrm{m}}^{\gamma_t + \varepsilon''}, \forall \epsilon'' > 0,\ \text{sufficiently small} \end{aligned}$$

where again at "$\overset{\ddagger}{\iff}$", we used the fact that $v_i - \gamma_t \neq C$ for any $i \in [k-1]$. Thus, we have proven (49). Since $\widetilde{\gamma} = \gamma_t \in (0, v_{\max})$ and $\gamma_1 = v_{\max}$, we have in particular that $\gamma_t < \gamma_1$. Thus, there exists $\tau \in \texttt{crit}_1(v)$ such that $\tau < t$ and $\gamma_t < \gamma_\tau$. Furthermore, we can choose τ such that for all $\gamma \in (\gamma_t, \gamma_\tau)$, $\gamma \notin \texttt{crit}_1(v)$. Let $\varepsilon'' > 0$ be so small that $\gamma_t + \varepsilon'', \gamma_\tau - \varepsilon'' \in (\gamma_t, \gamma_\tau)$, and furthermore both (49) and the condition in Lemma A.26 hold. We have

$$\begin{aligned} (n_{\mathrm{m}}^\tau, n_{\mathrm{u}}^\tau) &= (n_{\mathrm{m}}^{\gamma_\tau - \varepsilon''}, n_{\mathrm{u}}^{\gamma_\tau - \varepsilon''}) \quad \because \text{ Lemma A.26} \\ &= (n_{\mathrm{m}}^{\gamma_t + \varepsilon''}, n_{\mathrm{u}}^{\gamma_t + \varepsilon''}) \quad \because \text{ Lemma A.12 and } \gamma_t + \varepsilon'', \gamma_\tau - \varepsilon'' \in (\gamma_t, \gamma_\tau) \\ &= (n_{\mathrm{m}}^{\gamma_t}, n_{\mathrm{u}}^{\gamma_t}) \quad \because (49) \\ &= (n_{\mathrm{m}}^{\widetilde{\gamma}}, n_{\mathrm{u}}^{\widetilde{\gamma}}) \quad \because \text{Assumption.} \end{aligned}$$

This proves Theorem 3.4 under case 2.

Finally, we consider the last case. Under the assumptions, we have $t \in \texttt{crit}_2(v)$. Then Lemma A.27 $(n_{\mathrm{m}}^t, n_{\mathrm{u}}^t) = (n_{\mathrm{m}}^{\gamma_t}, n_{\mathrm{u}}^{\gamma_t}) = (n_{\mathrm{m}}^{\widetilde{\gamma}}, n_{\mathrm{u}}^{\widetilde{\gamma}})$. Thus, we have proven Theorem 3.4 under case 3. □

A.6. Experiments

The Walrus solver is available at:

`https://github.com/YutongWangUMich/liblinear`

The actual implementation is in the file `linear.cpp` in the class `Solver_MCSVM_WW`.

All code for downloading the datasets used, generating the train/test split, running the experiments and generating the figures are included. See the `README.md` file for more information.

All experiments are run on a single machine with the following specifications:

Operating system and kernel:

```
4.15.0-122-generic #124-Ubuntu SMP Thu Oct 15 13:03:05 UTC 2020 x86_64 GNU/Linux
```

Processor:

```
Intel(R) Core(TM) i7-6850K CPU @ 3.60GH
```

Memory:

```
31GiB System memory
```

A.6.1. On Sharks linear WW-SVM solver

Shark's linear WW-SVM solver is publicly available in the GitHub repository `https://github.com/Shark-ML`. Specifically, the C++ code is in `Algorithms/QP/QpMcLinear.h` in the class `QpMcLinearWW`. Our reimplementation follows their implementation with two major differences. In our implementations, neither Shark nor Walrus use the shrinking heuristic. Furthermore, we use a stopping criterion based on duality gap, following (Steinwart et al., 2011).

We also remark that Shark solves the following variant of the WW-SVM which is equivalent to ours after a change of variables. Let $0 < A \in \mathbb{R}$ be a hyperparameter.

$$\min_{\mathbf{u}\in\mathbb{R}^{d\times k}} F_A(\mathbf{u}) := \frac{1}{2}\|\mathbf{u}\|_F^2 + A\sum_{i=1}^{n}\sum_{\substack{j\in[k]:\\ j\neq y_i}} \texttt{hinge}\left((u'_{y_i}x_i - u'_j x_i)/2\right). \tag{50}$$

Recall the formulation (P) that we consider in this work, which we repeat here:

$$\min_{\mathbf{w}\in\mathbb{R}^{d\times k}} G_C(\mathbf{w}) := \frac{1}{2}\|\mathbf{w}\|_F^2 + C\sum_{i=1}^{n}\sum_{\substack{j\in[k]:\\ j\neq y_i}} \texttt{hinge}(w'_{y_i}x_i - w'_j x_i). \tag{51}$$

The formulation (50) is used by Weston & Watkins (1999), while the formulation (51) is used by Vapnik (1998). These two formulations are equivalent under the change of variables $\mathbf{w} = \mathbf{u}/2$ and $A = 4C$. To see this, note that

$$\begin{aligned}
G_C(\mathbf{w}) &= G_C(\mathbf{u}/2)\\
&= \frac{1}{2}\|\mathbf{u}/2\|_F^2 + C\sum_{i=1}^{n}\sum_{\substack{j\in[k]:\\ j\neq y_i}} \texttt{hinge}((u'_{y_i}x_i - u'_j x_i)/2)\\
&= \frac{1}{8}\|\mathbf{u}\|_F^2 + C\sum_{i=1}^{n}\sum_{\substack{j\in[k]:\\ j\neq y_i}} \texttt{hinge}((u'_{y_i}x_i - u'_j x_i)/2)\\
&= \frac{1}{4}\left(\frac{1}{2}\|\mathbf{u}\|_F^2 + 4C\sum_{i=1}^{n}\sum_{\substack{j\in[k]:\\ j\neq y_i}} \texttt{hinge}((u'_{y_i}x_i - u'_j x_i)/2)\right)\\
&= \frac{1}{4}F_{4C}(\mathbf{u}) = \frac{1}{4}F_A(\mathbf{u}).
\end{aligned}$$

Thus, we have proven

Proposition A.29. *Let $C > 0$ and $\mathbf{u} \in \mathbb{R}^{d\times k}$. Then $\mathbf{u}$ is a minimizer of F_{4C} if and only if $\mathbf{u}/2$ is a minimizer of G_C.*

In our experiments, we use the above proposition to rescale the variant formulation to the standard formulation.