Supplementary Material to Efficient Large-Scale Structured Learning

1. Multi-Sample Updates

In this section, we derive fast approximate optimization algorithms that meet the properties of Theorem 2.1.

1.1. Notation and Review

Recall that $\max_\alpha \mathcal{D}_t(\alpha) = \min_{\mathbf{w}} F_T(\mathbf{w})$, where $F_T(\mathbf{w})$ is the structured SVM training error and is defined as

$$\begin{aligned} F_T(\mathbf{w}) &= \sum_{t=1}^{T} f(\mathbf{w}; Z_t) && (1) \\ f(\mathbf{w}; Z_t) &= \frac{\lambda}{2}\|\mathbf{w}\|^2 + \ell(\mathbf{w}, Z_t) && (2) \\ \ell(\mathbf{w}, Z_t) &= \max_Y \left(\langle \mathbf{w}, \Psi(X_t, Y)\rangle + \Delta(Y, Y_t)\right) \\ &\qquad - \langle \mathbf{w}, \Psi(X_t, Y_t)\rangle && (3) \end{aligned}$$

and $\mathcal{D}_t(\alpha)$ is the equivalent dual problem and is defined as

$$\begin{aligned} \mathcal{D}_T(\alpha) &= -\frac{\lambda T}{2}\|\mathbf{w}^T\|^2 + \sum_{i,Y} \alpha_i^Y \Delta(Y, Y_i) && (4) \\ \mathbf{w}^T &= -\frac{1}{\lambda T}\sum_{i=1}^{T} \mathbf{u}_i && (5) \\ \mathbf{u}_i &= \sum_Y \alpha_i^Y \mathbf{v}(X_i, Y) && (6) \\ \mathbf{v}(X_i, Y) &= \Psi(X_i, Y) - \Psi(X_i, Y_i) && (7) \end{aligned}$$

Our goal is to derive exact or approximate solvers to the problem

$$\begin{aligned} &\arg\max_{\alpha_t} \mathcal{D}_t(\alpha_1 ... \alpha_{t-1}, \alpha_t) \\ &\qquad \text{s.t., } \forall_j, \alpha_t^{Y_t^j} \geq 0, \quad \sum_j \alpha_t^{Y_t^j} \leq 1 && (8) \end{aligned}$$

which is equivalent to the following QP problem:

$$\begin{aligned} &\min_{\alpha_t} \frac{1}{2}\alpha_t^T Q \alpha_t + \mathbf{c}^T \alpha_t \\ &\qquad \text{s.t., } \forall_j, \alpha_t^{Y_t^j} \geq 0, \quad \sum_j \alpha_t^{Y_t^j} \leq 1 && (9) \end{aligned}$$

where Q is a $K \times K$ matrix and $\mathbf{c}$ is a K-vector with elements

$$\begin{aligned} Q_{ij} &= \left\langle \mathbf{v}(X_t, \bar{Y}_t^i), \mathbf{v}(X_t, \bar{Y}_t^j) \right\rangle && (10) \\ c_j &= -\lambda t \left(\langle \mathbf{w}^{t-1}, \mathbf{v}_t^{\bar{Y}_t^j} \rangle + \Delta(\bar{Y}_t^j, Y_t) \right) && (11) \end{aligned}$$

1.2. A Fast, Approximate Multi-Sample Update Algorithm

In this section, derive an approximate update step that occurs over a set of samples $\bar{\mathbf{Y}}_{\mathbf{i}} = \bar{Y}_i^1, \bar{Y}_i^2 ... \bar{Y}_i^K$. For convergence guarantees (see Theorem 2.1), we assume that this set includes the subgradient as the first sample: $\mathbf{v}(X_t, \bar{Y}_t^1) = \nabla \ell(\mathbf{w}^{t-1}; Z_t)$. The update provides an approximate solution to the optimization problem:

$$
\begin{aligned}
&\arg\max_{\alpha_t} \mathcal{D}_t(\alpha_1 ... \alpha_{t-1}, \alpha_t) \\
&\qquad\qquad \text{s.t., } \forall_j, \alpha_t^{Y_t^j} \geq 0, \quad \sum_j \alpha_t^{Y_t^j} \leq 1
\end{aligned} \tag{12}
$$

Pseudo-code for the algorithm is shown in Algorithm 2. In this algorithm, each sample $\bar{Y}_i^j$ is iterated over once. We define an iterative procedure, such that each $\alpha_i^{\bar{Y}_i^j}$ is updated in order $j = 1...K$. In each iteration, we solve (in closed form) for the optimal value to set $\alpha_t^{\bar{Y}_t^j}$ which maximizes $\Delta\mathcal{D}_{t,j}^{expand}$:

$$
\begin{aligned}
\Delta\mathcal{D}_{t,j}^{expand} &= \mathcal{D}_t(\alpha_{\mathbf{1}} ... \alpha_{\mathbf{t-1}}, \alpha_t^{\bar{Y}_t^j}) - \mathcal{D}_t(\alpha_{\mathbf{1}} ... \alpha_{\mathbf{t-1}}, 0) \\
&= \frac{-\lambda t}{2} \left\| \mathbf{w}^t - \frac{\alpha_t^{\bar{Y}_t^j}}{\lambda t} \mathbf{v}(X_t, \bar{Y}_t^j) \right\|^2 + \frac{\lambda t}{2} \left\| \mathbf{w}^t \right\|^2 + \alpha_t^{\bar{Y}_t^j} \Delta(\bar{Y}_t^j, Y_t) \\
&= \alpha_t^{\bar{Y}_t^j} \left(\langle \mathbf{w}^{t,j-1}, \mathbf{v}(X_t, \bar{Y}_t^j) \rangle + \Delta(\bar{Y}_t^j, Y_t) \right) - \frac{(\alpha_t^{\bar{Y}_t^j})^2}{2\lambda t} \|\mathbf{v}(X_t, \bar{Y}_t^j)\|^2
\end{aligned}
$$

This is maximized (by setting the derivative equal to 0) by choosing

$$
\alpha_t^{\bar{Y}_t^j} = \frac{\lambda t \left(\langle \mathbf{w}^{t,j-1}, \mathbf{v}(X_t, \bar{Y}_t^j) \rangle + \Delta(\bar{Y}_t^j, Y_t) \right)}{\|\mathbf{v}(X_t, \bar{Y}_t^j)\|^2}
$$

Unfortunately, due to the constraint $\sum_{j=1}^K \alpha_i^{\bar{Y}_i^j} \leq 1$ in Eq 12, this update becomes invalid once $\sum_{j=1}^K \alpha_i^{\bar{Y}_i^j} = 1$. We therefore consider an alternate "swap" move, which works by setting $\alpha_i^{\bar{Y}_i^j}$ while simultaneously scaling all parameters $\alpha_i^{\bar{Y}_i^1} ... \alpha_i^{\bar{Y}_i^{j-1}}$ by $s = 1 - \alpha_i^{\bar{Y}_i^j}$. This swap move preserves the constraint $\sum_{j=1}^K \alpha_i^{\bar{Y}_i^j} = 1$. The subsequent change in the dual objective is:

$$
\begin{aligned}
&\Delta\mathcal{D}_{t,j}^{swap} \\
= \; &\mathcal{D}_t(\alpha_{\mathbf{1}} ... \alpha_{\mathbf{t-1}}, s\alpha_t^{\bar{Y}_t^1}, ..., s\alpha_t^{\bar{Y}_t^{j-1}}, \alpha_t^{\bar{Y}_t^j}) - \mathcal{D}_t(\alpha_{\mathbf{1}} ... \alpha_{\mathbf{t-1}}, s\alpha_t^{\bar{Y}_t^1}, ..., s\alpha_t^{\bar{Y}_t^{j-1}}, 0) \\
= \; &\frac{-\lambda t}{2} \left\| \mathbf{w}^t - \frac{(s-1)\mathbf{u}_t^{j-1} + \alpha_t^{\bar{Y}_t^j} \mathbf{v}(X_t, \bar{Y}_t^j)}{\lambda t} \right\|^2 + \frac{\lambda t}{2} \left\| \mathbf{w}^t \right\|^2 + (s-1) D_t^{j-1} + \alpha_t^{\bar{Y}_t^j} \Delta(\bar{Y}_t^j, Y_t) \\
= \; &\frac{-\lambda t}{2} \left\| \mathbf{w}^t - \frac{-\alpha_t^{\bar{Y}_t^j} \mathbf{u}_t^{j-1} + \alpha_t^{\bar{Y}_t^j} \mathbf{v}(X_t, \bar{Y}_t^j)}{\lambda t} \right\|^2 + \frac{\lambda t}{2} \left\| \mathbf{w}^t \right\|^2 - \alpha_t^{\bar{Y}_t^j} D_t^{j-1} + \alpha_t^{\bar{Y}_t^j} \Delta(\bar{Y}_t^j, Y_t) \\
= \; &\alpha_t^{\bar{Y}_t^j} \left[\left(\langle \mathbf{w}^t, \mathbf{v}(X_t, \bar{Y}_t^j) \rangle + \Delta(\bar{Y}_t^j, Y_t) \right) - \left(\langle \mathbf{w}^t, \mathbf{u}_t^{j-1} \rangle + D_t^{j-1} \right) \right] + \frac{\left(\alpha_t^{\bar{Y}_t^j} \right)^2}{2\lambda t} \left\| \mathbf{u}_t^{j-1} + \mathbf{v}(X_t, \bar{Y}_t^j) \right\|^2
\end{aligned}
$$

where D_t^{j-1} is shorthand for

$$
D_t^{j-1} = \sum_{k=1}^{j-1} \alpha_t^{\bar{Y}_t^k} \Delta(\bar{Y}_t^k, Y_t) \tag{13}
$$

and can be interpreted as a (weighted) average loss over all samples. The value of $\alpha_t^{\bar{Y}_t^j}$ which maximizes $\Delta\mathcal{D}_{t,j}^{swap}$ is:

$$
\alpha_t^{\bar{Y}_t^j} = \frac{\lambda t \left[\left(\langle \mathbf{w}^t, \mathbf{v}(X_t, \bar{Y}_t^j) \rangle + \Delta(\bar{Y}_t^j, Y_t) \right) - \left(\langle \mathbf{w}^t, \mathbf{u}_t^{j-1} \rangle + D_t^{j-1} \right) \right]}{\left\| \mathbf{u}_t^{j-1} + \mathbf{v}(X_t, \bar{Y}_t^j) \right\|^2} \tag{14}
$$

Algorithm 1 MULTISAMPLEUPDATE

Input: New example X_t, Y_t, current weights $\mathbf{w}^{t-1}$
Output: New weights $\mathbf{w}^{t,K}$

1: $\bar{Y}_t^1 ... \bar{Y}_t^K \leftarrow \text{IMPORTANCESAMPLE}(X_t, Y_t, \mathbf{w}^{t-1})$
2: Initialize $\mathbf{w}^{t,0} \leftarrow \frac{t-1}{t}\mathbf{w}^{t-1}$, $\mathbf{u}_t^0 \leftarrow 0$, $D_t^0 \leftarrow 0$, $\alpha_t \leftarrow 0$
3: **for** $j = 1$ to K **do**
4: **if** $\alpha_t < 1$ **then**
5: $\alpha_t^{\bar{Y}_t^j} \leftarrow \min\left(1 - \alpha_t, \max\left(0, \frac{\lambda t(l_j)}{\|\mathbf{v}(X_t, \bar{Y}_t^j)\|^2}\right)\right)$
6: $s \leftarrow 1$
7: $\alpha_t \leftarrow \alpha_t + \alpha_t^{\bar{Y}_t^j}$
8: **else**
9: $\alpha_t^{\bar{Y}_t^j} \leftarrow \min\left(1, \max\left(0, \frac{\lambda t\left[\left(\langle \mathbf{w}^t, \mathbf{v}(X_t, \bar{Y}_t^j)\rangle + \Delta(\bar{Y}_t^j, Y_t)\right) - \left(\langle \mathbf{w}^t, \mathbf{u}_t^{j-1}\rangle + D_t^{j-1}\right)\right]}{\left\|\mathbf{u}_t^{j-1} + \mathbf{v}(X_t, \bar{Y}_t^j)\right\|^2}\right)\right)$
10: $s \leftarrow 1 - \alpha_t^{\bar{Y}_t^j}$
11: **end if**
12: $\mathbf{u}_t^j \leftarrow s\, \mathbf{u}_t^{j-1} + \alpha_t^{\bar{\mathbf{Y}}_t^j} \mathbf{v}(X_t, \bar{\mathbf{Y}}_t^j)$
13: $\mathbf{w}^{t,j} \leftarrow \mathbf{w}^{t,j-1} - \frac{(s-1)\mathbf{u}_t^{j-1} + \alpha_t^{\bar{Y}_t^j} \mathbf{v}(X_t, \bar{Y}_t^j)}{\lambda t}$
14: $D_t^j \leftarrow s D_t^{j-1} + \alpha_t^{\bar{Y}_t^j} \Delta(\bar{Y}_t^j, Y_t)$
15: **end for**
16: Optionally repeat steps 3-16 multiple times

We can compute $\alpha_t^{\bar{Y}_t^j}$ in $O(d)$ time, where d is the dimensionality of the feature space Ψ, if we maintain updated values for $\mathbf{w}_t^j$, $\mathbf{u}_t^j$, and D_t^j. The appropriate updates if we were to scale α_t by s and then set $\alpha_t^{\bar{Y}_t^j}$ are:

$$\mathbf{u}_t^j \leftarrow s\, \mathbf{u}_t^{j-1} + \alpha_t^{\bar{Y}_t^j} \mathbf{v}(X_t, \bar{Y}_t^j) \tag{15}$$

$$\mathbf{w}^{t,j} \leftarrow \mathbf{w}^{t,j-1} - \frac{(s-1)\mathbf{u}_t^{j-1} + \alpha_t^{\bar{Y}_t^j} \mathbf{v}(X_t, \bar{Y}_t^j)}{\lambda t} \tag{16}$$

$$D_t^j \leftarrow s D_t^{j-1} + \alpha_t^{\bar{Y}_t^j} \Delta(\bar{Y}_t^j, Y_t) \tag{17}$$

1.3. A Fast, Exact Multi-Sample Update Algorithm For Multiclass Problems

Recall that a cost-sensitive multiclass SVM can be represented by a structured SVM (Eq 1) with a feature space that concatenates features for each class:

$$\Psi(X, Y) = [\psi_1(X, Y) ... \psi_C(X, Y)] \tag{18}$$

$$\psi_c(X, Y) = \begin{cases} \phi(X) & \text{if } Y = c \\ \mathbf{0} & \text{otherwise} \end{cases} \tag{19}$$

This feature space has the property that

$$\langle \Psi(X, c_1), \Psi(X, c_2)\rangle = \begin{cases} \phi^2(X) & \text{if } c_1 = c_2 \\ \mathbf{0} & \text{otherwise} \end{cases} \tag{20}$$

It follows that the matrix $\mathbf{Q}$ has entries (Eq 10)

$$Q_{ij} = \begin{cases} 2\,|\phi(X)\|^2 & \text{if } i = j \\ |\phi(X)\|^2 & \text{otherwise} \end{cases} \tag{21}$$

It follows that $\mathbf{Q} = |\phi(X)\|^2 (\mathbf{1}_{K\times K} + \mathbf{I}_{K\times K})$, where $\mathbf{I}_{K\times K}$ is the identity matrix and $\mathbf{1}_{K\times K}$ is a matrix of ones. Since Eq 9 is a quadratic program, if we ignore the constraints, a Newton-update

$$\alpha_t \leftarrow -\mathbf{H}^{-1}\nabla \tag{22}$$

Algorithm 2 MULTICLASSMULTISAMPLEUPDATE

Input: New example X_t, Y_t, current weights $\mathbf{w}^{t-1}$
Output: New weights $\mathbf{w}^{t,K}$

For $i = 1...K$, $c_i \leftarrow -\lambda t \left(\langle \mathbf{w}^{t-1}, \mathbf{v}(X_t, \bar{Y}_t^i)\rangle + \Delta(\bar{Y}_t^i, Y_t)\right)$
Let $c_{j(1)}, c_{j(2)}, ..., c_{j(K)}$ be the entries of $\mathbf{c}$ in ascending order
$s \leftarrow 0, n = 1$
while $n \leq K$ and $c_{j(n)} \leq \min\left(\frac{s}{n+1}, \frac{1+s}{n}\right)$ **do**
    $s \leftarrow s + c_{j(n)}$
    $n \leftarrow n + 1$
end while
for $i = 1$ to n **do**
    $\alpha_{j(i)} \leftarrow -\left(c_{j(i)} - \min\left(\frac{s}{n+1}, \frac{1+s}{n}\right)\right)$
    $\mathbf{w}^t \leftarrow \mathbf{w}^t - \frac{\alpha_{j(i)}\mathbf{v}(X_t, j(i))}{\lambda t}$
end for

Algorithm 3

for $t = 1$ to T **do**
    Receive an example $Z_t \leftarrow (X_i, Y_i)$ where $i = t \mod n$
    Suffer loss $f(\mathbf{w}^{t-1}; Z_t) = \frac{\lambda}{2}\|\mathbf{w}^{t-1}\|^2 + \ell(\mathbf{w}^{t-1}; Z_t)$
    Update $\mathbf{w}^t$ using Algorithm 2
    If $\|\mathbf{w}^t\| > \frac{1}{\sqrt{\lambda}}$, $\mathbf{w}^t \leftarrow \frac{1/\sqrt{\lambda}}{\|\mathbf{w}^t\|}\mathbf{w}^t$
end for

with entries of the Hessian matrix $\mathbf{H}$ being $H_{ij} = Q_{ij}$, and entries of the gradient vector ∇ being

$$\mathbf{H} = \mathbf{Q} = |\phi(X)\|^2 \left(\mathbf{1}_{K\times K} + \mathbf{I}_{K\times K}\right) \tag{23}$$
$$\nabla_i = c_i = -\lambda t \left(\langle \mathbf{w}, \mathbf{v}(X_t, \bar{Y}_t^i)\rangle + \Delta(\bar{Y}_t^i, Y_t)\right) \tag{24}$$

Since the inverse of the sum of two matrices satisfies, $(A+B)^{-1} = A^{-1} - \frac{A^{-1}BA^{-1}}{1+\text{trace}(BA^{-1})}$, it follows that the inverse Hessian $\mathbf{H}^{-1}$ is:

$$\mathbf{H}^{-1} = \frac{1}{|\phi(X)\|^2}\left(\mathbf{I}_{K\times K} - \frac{1}{K+1}\mathbf{1}_{K\times K}\right) \tag{25}$$

The constraint $\alpha_t^{Y_t^j} \geq 0$ can be handled by sorting c_i in increasing and considering a smaller matrix $\mathbf{Q}$ that excludes entries where $\alpha_t^{Y_t^j}$ would like to go below zero. The constraints $\sum_j \alpha_t^{Y_t^j} \leq 1$ can be handled by reducing the step size.

2. Bounds For Structured Learning

The results of [**?**] can be applied to structured SVMs and can be used to provide generalization guarantees to customizable loss functions $\Delta(g(X_t; \mathbf{w}), Y)$, and to bound the convergence rate of online or sequential optimization algorithms:

Theorem 2.1 *Let $f(\mathbf{w}; Z)$ be the structured SVM objective defined in Eqn 2. Let $L = \sqrt{\lambda} + 2R$, where $\|\Psi(X,Y)\| \leq R$ is a bound on the image of Ψ. Then using Algorithm 1:*

1. **Online Regret:** *The average loss accumulated by Algorithm 2 over T iterations can be bounded in relation to the minimum achievable training error:*

$$\frac{1}{T}\sum_{t=1}^{T}\Delta(g(X_t; \mathbf{w}^{t-1}), Y_t) \leq \frac{1}{T}\min_{\mathbf{w}} F_T(\mathbf{w}) + \frac{L^2(\log(T)+1)}{2\lambda T}$$

2. **Generalization Error:** *Let $Z_1 ... Z_{n+1}$ be examples selected independently at random from some distribution $\mathbb{P}(Z)$. With $T = n$, the expected loss when training on $Z_1 ... Z_n$ and testing on Z_{n+1} can be bounded in relation to the Bayes optimal solution:*

$$\mathbb{E}_{Z_1 ... Z_n}[\mathbb{E}_Z[\Delta(g(X; \mathbf{w}^n), Y)]] \leq \min_{\mathbf{w}} \mathbb{E}_Z[f(\mathbf{w}; Z)] + \frac{L^2(\log(n) + 1)}{2\lambda n} \tag{26}$$

3. **Empirical error:** *Let $\bar{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{w}^{t-1}$. If Algorithm 1 is run for $T = mn$ iterations, passing over each example $m \geq 1$ times, the average training error can be bounded in relation to the minimal achievable training error:*

$$\frac{1}{n} F_n(\bar{\mathbf{w}}) \leq \left(\frac{1}{n} \min_{\mathbf{w}} F_n(\mathbf{w}) \right) + \frac{L^2(\log(mn) + 1)}{2\lambda mn} \tag{27}$$

A similar bound holds for arbitrary strongly convex loss functions, and our proof closely follows the one presented in [**?**]. Similar results were presented in [**?**]. One implication is that if we take only a single pass through the training set (setting $m = 1$), where train time equals test time, the empirical error bound (*e.g.*, error because we haven't spent enough computation time) converges at the same asymptotic rate as the generalization error (*e.g.*, error because we don't have enough training examples). Thus when computation time is a bottleneck, it is better to use as many training examples as possible (use a large value for n and a small value for m). A second implication is that the convergence rates for structured SVMs are the same as for linear SVMs, with the same constants involved [**?**], and therefore we don't sacrifice theoretical guarantees by using an application specific loss function $\Delta(g(X_t; \mathbf{w}), Y)$.

Proof The main result follows because the dual objective in each step of Algorithm 1 increases by a predictable amount (see Lemma 2.2). Summing over each iteration and using weak duality, Lemma 2.4 shows that

$$\frac{1}{T} \sum_{t=1}^{T} f(\mathbf{w}^{t-1}; Z_t) \quad \leq \quad \frac{1}{T} \min_{\mathbf{w}} F_T(\mathbf{w}) + \frac{G^2 \left(\log(T) + 1 \right)}{2\lambda T}$$

where G is a bound on the gradient $\nabla f(\mathbf{w}^{t-1}; Z_t)$. Lemma 2.5 establishes a numerical bound on the gradient $G = L = \sqrt{\lambda} + 2R$ for structured SVMs. Finally, Lemma 2.6 establishes that $f(\mathbf{w}; Z)$ is an upper bound on $\Delta(g(X; \mathbf{w}), Y)$. It therefore follows that

$$\frac{1}{T} \sum_{t=1}^{T} \Delta(g(X_t; \mathbf{w}^{t-1}), Y_t) \quad \leq \quad \frac{1}{T} \min_{\mathbf{w}} F_T(\mathbf{w}) + \frac{L^2 \left(\log(T) + 1 \right)}{2\lambda T}$$

thus proving Thm 2.1.1. Thm 2.1.2 is directly applicable from this regret bound and theorem 2 in [**?**]. Lastly, by Jensen's inequality $\sum_{t=1}^{T} f(\bar{\mathbf{w}}; Z_t) \leq \sum_{t=1}^{T} f(\mathbf{w}^{t-1}; Z_t)$. Since Algorithm 2 iterates over each example m times, $F_n(\bar{\mathbf{w}}; Z_i) = \sum_{t=1}^{T} f(\bar{\mathbf{w}}; Z_t)$. Thm 2.1.3 follows from Thm 2.1.1.

Lemma 2.2 *The change in the dual objective in each step in Algorithm 2 is at least*

$$\mathcal{D}_t(\alpha_{\mathbf{1}} ... \alpha_{\mathbf{t-1}}, \alpha_t^{\bar{Y}_t}) - \mathcal{D}_{t-1}(\alpha_{\mathbf{1}} ... \alpha_{\mathbf{t-1}}) \geq f(\mathbf{w}^{t-1}; Z_t) - \frac{1}{2\lambda t} \|\nabla f(\mathbf{w}^{t-1}; Z_t)\|^2$$

Proof Consider the simpler case where each step of Algorithm 1 simply adds a new dual variable in the direction of the subgradient $\mathbf{v}_t^{\bar{Y}_t} = \nabla\ell(\mathbf{w}^{t-1}; Z_t)$ with weight $\alpha_t^{\bar{Y}_t} = 1$. The change in the dual objective is exactly

$$\begin{aligned}
& \mathcal{D}_t(\alpha_{\mathbf{1}}...\alpha_{\mathbf{t-1}}, \alpha_t^{\bar{Y}_t}) - \mathcal{D}_{t-1}(\alpha_{\mathbf{1}}...\alpha_{\mathbf{t-1}}) \\
= & \left(\frac{-\lambda t}{2}\left\|\mathbf{w}^t\right\|^2 + \sum_{t',Y_{t'}}^{t} \alpha_{t'}^{\bar{Y}_{t'}} \Delta(Y, Y_{t'})\right) - \left(\frac{-\lambda(t-1)}{2}\left\|\mathbf{w}^{t-1}\right\|^2 + \sum_{t',Y_{t'}}^{t-1} \alpha_{t'}^{\bar{Y}_{t'}} \Delta(Y, Y_{t'})\right) \\
= & \frac{-\lambda t}{2}\left\|\frac{t-1}{t}\mathbf{w}^{t-1} - \frac{\mathbf{v}_t^{\bar{Y}_t}}{\lambda t}\right\|^2 + \frac{\lambda(t-1)}{2}\left\|\mathbf{w}^{t-1}\right\|^2 + \Delta(Y, \bar{Y}_t) \\
= & \frac{\lambda t - 2\lambda}{2t}\left\|\mathbf{w}^{t-1}\right\|^2 + \frac{t-1}{t}\langle \mathbf{w}^{t-1}, \mathbf{v}_t^{\bar{Y}_t}\rangle - \frac{(\mathbf{v}_t^{\bar{Y}_t})^2}{2\lambda t} + \Delta(Y, \bar{Y}_t) \\
= & \left(\frac{\lambda}{2}\left\|\mathbf{w}^{t-1}\right\|^2 + \langle \mathbf{w}^{t-1}, \mathbf{v}_t^{\bar{Y}_t}\rangle + \Delta(Y, \bar{Y}_t)\right) - \left(\frac{\lambda}{2t}\left\|\mathbf{w}^{t-1}\right\|^2 - \frac{1}{t}\langle \mathbf{w}^{t-1}, \mathbf{v}_t^{\bar{Y}_t}\rangle - \frac{(\mathbf{v}_t^{\bar{Y}_t})^2}{2\lambda t}\right) \\
= & \left(\frac{\lambda}{2}\left\|\mathbf{w}^{t-1}\right\|^2 + \ell(\mathbf{w}^{t-1}, Z_t)\right) - \frac{1}{2\lambda t}\left\|\lambda \mathbf{w}^{t-1} + \mathbf{v}_t^{\bar{Y}_t}\right\|^2 \\
= & f(\mathbf{w}^{t-1}, Z_t) - \frac{1}{2\lambda t}\left\|\nabla f(\mathbf{w}^{t-1}, Z_t)\right\|^2
\end{aligned}$$

Since Algorithm 1, maximizes $\mathcal{D}_t(\alpha_{\mathbf{1}}...\alpha_{\mathbf{t-1}}, \alpha_t^{\bar{Y}_t}) - \mathcal{D}_{t-1}(\alpha_{\mathbf{1}}...\alpha_{\mathbf{t-1}})$ over a set of samples that includes $\nabla\ell(\mathbf{w}^{t-1}; Z_t)$, the increase in the dual objective must be at least as much. This completes the proof as long as Line 5 of Algorithm 2 does not reduce the dual objective (which we prove in Lemma 2.3):

Lemma 2.3 *The projection step in Line 5 of Algorithm 2 cannot decrease the dual objective.*

Proof Line 5 of Algorithm 2 projects $\mathbf{w}^t$ onto the L_2 ball $\|\mathbf{w}^t\|^2 \le \frac{1}{\lambda}$. It checks if $\|\mathbf{w}^t\| > \frac{1}{\sqrt{\lambda}}$, and if so scales $\mathbf{w}^t$ by

$$s \leftarrow \frac{1/\sqrt{\lambda}}{\|\mathbf{w}^t\|} \tag{28}$$

where $0 < s < 1$. The corresponding change in the dual objective is

$$\begin{aligned}
\Delta\mathcal{D}_t^{proj} &= \mathcal{D}_t(s\alpha) - \mathcal{D}_t(\alpha) && (29)\\
&= \left[-\frac{t\lambda}{2}\|s\mathbf{w}^t\|^2 + s\sum_{i,Y}\alpha_i^Y \Delta(Y, Y_i)\right] - \mathcal{D}_t(\alpha) && (30)\\
&= \left[-\frac{t\lambda}{2}\|s\mathbf{w}^t\|^2 + s\left(\frac{t\lambda}{2}\|\mathbf{w}^t\|^2 + \mathcal{D}_t(\alpha)\right)\right] - \mathcal{D}_t(\alpha) && (31)\\
&= \left[-\frac{t\lambda}{2}\frac{s^2}{\lambda s^2} + s\left(\frac{t\lambda}{2}\frac{1}{\lambda s^2} + \mathcal{D}_t(\alpha)\right)\right] - \mathcal{D}_t(\alpha) && (32)\\
&= -\frac{t}{2} + \frac{t}{2s} + (s-1)\mathcal{D}_t(\alpha) && (33)\\
&= (1-s)\left(\frac{t}{2s} - \mathcal{D}_t(\alpha)\right) && (34)\\
&\ge 0 && (35)
\end{aligned}$$

where the last line follows because $1 - s > 0$ and $\mathcal{D}_t(\alpha) \leq \frac{t}{2s}$:

$$\begin{aligned}
\mathcal{D}_t(\alpha) &= -\frac{t\lambda}{2}\|\mathbf{w}^t\|^2 + \sum_{i,Y} \alpha_i^Y \Delta(Y, Y_i) && (36)\\
&= -\frac{t\lambda}{2}\frac{1}{\lambda s^2} + \sum_{i,Y} \alpha_i^Y \Delta(Y, Y_i) && (37)\\
&= -\frac{t}{2s^2} + \sum_{i,Y} \alpha_i^Y \Delta(Y, Y_i) && (38)\\
&\leq -\frac{t}{2s^2} + t \leq -\frac{t}{2} + t \leq \frac{t}{2s} && (39)
\end{aligned}$$

where we have assumed $\Delta(Y, Y_i) \leq 1$.

Lemma 2.4 *Let G be a bound on $\nabla f(\mathbf{w}^{t-1}; Z_t)$. The average loss accumulated by Algorithm 2 over T iterations is bounded by*

$$\frac{1}{T}\sum_{t=1}^{T} f(\mathbf{w}^{t-1}; Z_t) \quad \leq \quad \frac{1}{T}\min_{\mathbf{w}} F_T(\mathbf{w}) + \frac{G^2\left(\log(T) + 1\right)}{2\lambda T}$$

Proof By Lemma 2.2 and the definition of G,

$$\mathcal{D}_t(\alpha_{\mathbf{1}}...\alpha_{\mathbf{t-1}}, \alpha_t^{\bar{Y}_t}) - \mathcal{D}_{t-1}(\alpha_{\mathbf{1}}...\alpha_{\mathbf{t-1}}) \geq f(\mathbf{w}^{t-1}; Z_t) - \frac{G^2}{2\lambda t}$$

Summing over $t = 1...T$:

$$\begin{aligned}
\mathcal{D}_T(\alpha_{\mathbf{1}}...\alpha_{\mathbf{T}}) &= \sum_{t=1}^{T}\Big(\mathcal{D}_t(\alpha_{\mathbf{1}}...\alpha_{\mathbf{t-1}}, \alpha_t^{\bar{Y}_t}) - \mathcal{D}_{t-1}(\alpha_{\mathbf{1}}...\alpha_{\mathbf{t-1}})\Big)\\
&\geq \sum_{t=1}^{T} f(\mathbf{w}^{t-1}; Z_t) - \frac{G^2}{2\lambda}\sum_{t=1}^{T}\frac{1}{t}
\end{aligned}$$

Since $\sum_{t=1}^{T}\frac{1}{t} \leq \log(T) + 1$

$$\mathcal{D}_T(\alpha_{\mathbf{1}}...\alpha_{\mathbf{T}}) \quad \geq \quad \sum_{t=1}^{T} f(\mathbf{w}^{t-1}; Z_t) - \frac{G^2(\log T + 1)}{2\lambda}$$

Applying weak duality $\mathcal{D}_T(\alpha_{\mathbf{1}}...\alpha_{\mathbf{T}}) \leq \min_{\mathbf{w}} F_T(\mathbf{w})$ and rearranging terms completes the proof.

Lemma 2.5 *The magnitude of the gradient of the structured SVM error $\nabla f(\mathbf{w}^{t-1}; Z_i)$ in Algorithm 2 is bounded $\|\nabla f(\mathbf{w}^{t-1}; Z_t)\| \leq \sqrt{\lambda} + 2R$, where $\|\Psi(X, Y)\| \leq R$ is a bound on the image of Ψ.*

Proof Since $\nabla \ell(\mathbf{w}^{t-1}; Z_t) = \Psi(X_t, \bar{Y}_t) - \Psi(X_t, Y_t)$ and $\|\Psi(X, Y)\| \leq R$ for all X, Y, it must be the case that $\|\nabla \ell(\mathbf{w}^{t-1}; Z_t)\| \leq 2R$. Line 5 of Algorithm 2 ensures that $\|\mathbf{w}^{t-1}\| \leq \frac{1}{\sqrt{\lambda}}$. Since $\nabla f(\mathbf{w}^{t-1}; Z_t) = \lambda \mathbf{w}^{t-1} + \nabla \ell(\mathbf{w}^{t-1}; Z_t)$, by the triangle inequality, $\|\nabla f(\mathbf{w}^{t-1}; Z_t)\| \leq \sqrt{\lambda} + 2R$.

Lemma 2.6 *Let $g(X; \mathbf{w}) = \arg\max_Y \langle \mathbf{w}, \Psi(X, Y)\rangle$ be the label predicted by model parameters $\mathbf{w}$. The loss associated with this prediction is upper-bounded by the structured hinge loss: $\ell(\mathbf{w}; Z) \geq \Delta(g(X; \mathbf{w}), Y)$*

Proof

$$\begin{aligned}
\max_{Y'}\left(\langle \mathbf{w}, \Psi(X, Y')\rangle + \Delta(Y', Y)\right) &\geq \langle \mathbf{w}, \Psi(X, g(X; \mathbf{w}))\rangle + \Delta(g(X; \mathbf{w}), Y)\\
\max_{Y'}\left(\langle \mathbf{w}, \Psi(X, Y')\rangle + \Delta(Y', Y)\right) &\geq \langle \mathbf{w}, \Psi(X, Y)\rangle + \Delta(g(X; \mathbf{w}), Y)\\
\max_{Y'}\left(\langle \mathbf{w}, \Psi(X, Y')\rangle + \Delta(Y', Y)\right) - \langle \mathbf{w}, \Psi(X, Y)\rangle &\geq \Delta(g(X; \mathbf{w}), Y)\\
\ell(\mathbf{w}; Z) &\geq \Delta(g(X; \mathbf{w}), Y)
\end{aligned}$$

References