Supplementary Material
Efficient Training of Structured SVMs via Soft Constraints

A Dual of Soft Problem

In this section we show that the problems Eq. (5) and Eq. (6) are Lagrange duals. We start from a formulation equivalent to Eq. (6):

$$\begin{aligned}
\min_{w,\xi,\delta} \;& \frac{\lambda}{2}\|w\|^2 + \frac{\rho}{2}\sum_m \|\delta^{(m)}\|^2 + \sum_m \sum_\alpha \xi_\alpha^{(m)} \\
\text{s.t. }\; & \xi_i^{(m)} \geq \frac{1}{M}\left(\theta_i^{(m)}(y_i;w) + \sum_{c:i\in c}\delta_{ci}^{(m)}(y_i)\right) \quad \text{for all } m,i,y_i \\
& \xi_c^{(m)} \geq \frac{1}{M}\left(\theta_c^{(m)}(y_c;w) - \sum_{i:i\in c}\delta_{ci}^{(m)}(y_i)\right) \quad \text{for all } m,c,y_c
\end{aligned}$$

The Lagrangian is:

$$\begin{aligned}
L(w,\xi,\delta,\mu\geq 0) = & \frac{\lambda}{2}\|w\|^2 + \frac{\rho}{2}\sum_m \|\delta^{(m)}\|^2 + \sum_m\sum_\alpha \xi_\alpha^{(m)} \\
& - \sum_m\sum_i\sum_{y_i}\mu_i^{(m)}(y_i)\left(\xi_i^{(m)} - \frac{1}{M}\theta_i^{(m)}(y_i;w) - \frac{1}{M}\sum_{c:i\in c}\delta_{ci}^{(m)}(y_i)\right) \\
& - \sum_m\sum_c\sum_{y_c}\mu_c^{(m)}(y_c)\left(\xi_c^{(m)} - \frac{1}{M}\theta_c^{(m)}(y_c;w) + \frac{1}{M}\sum_{i:i\in c}\delta_{ci}^{(m)}(y_i)\right)
\end{aligned}$$

The optimality conditions entail:

$$\begin{aligned}
& w = \frac{1}{\lambda M}\sum_m\sum_\alpha\sum_{y_\alpha}\mu_\alpha^{(m)}(y_\alpha)\left(\phi_\alpha(x^{(m)},y_\alpha^{(m)}) - \phi_\alpha(x^{(m)},y_\alpha)\right) = \Psi\mu \\
& \sum_{y_\alpha}\mu_\alpha(y_\alpha) = 1 \quad \text{for all } m, \alpha=\{c,i\} \\
& \delta_{ci}^{(m)}(y_i) = \frac{1}{\rho M}\left(\mu_c^{(m)}(y_i) - \mu_i^{(m)}(y_i)\right) \quad \text{for all } m,c,i\in c,y_i \quad \Rightarrow \delta = A\mu
\end{aligned}$$

Using those in the Lagrangian yields the dual problem of Eq. (5).

B Proof of Theorem 4.1

In this section we prove Theorem 4.1, which is restated here for convenience.

Theorem 4.1 *Let g_ρ^* be the optimal value of G_ρ, and let g^* be the optimal value of G. Then $g_\rho^* - \frac{\rho}{2}h \leq g^* \leq g_\rho^*$, where $h = M(8Y_{\max}q(BR+L))^2$.*

Proof. Denote by (w^*,δ^*) an optimal solution to g, and by (w_ρ^*,δ_ρ^*) an optimal solution to g_ρ.

For the first direction, we have:

$$\begin{aligned} g^* &= \min_{w,\delta} g(w,\delta) \\ &\le g(w^*_\rho, \delta^*_\rho) \\ &\le g(w^*_\rho, \delta^*_\rho) + \frac{\rho}{2}\|\delta^*_\rho\|^2 \\ &= g^*_\rho \end{aligned}$$

Using the bound $\|\delta^*\|^2 \le h$, we can prove the other direction:

$$\begin{aligned} g^*_\rho &= \min_{w,\delta} \left(g(w,\delta) + \frac{\rho}{2}\|\delta\|^2 \right) \\ &\le g(w^*, \delta^*) + \frac{\rho}{2}\|\delta^*\|^2 \\ &= g^* + \frac{\rho}{2}\|\delta^*\|^2 \\ &\le g^* + \frac{\rho}{2} h \end{aligned}$$

To conclude the proof, we next show that $\|\delta^*\|^2 \le h$ by bounding $\|\delta^{(m)^*}\| \le 8Y_{\max} q(BR+L)$.

□

B.1 Bounding $\|\delta\|^2$

In this section we prove the bound[11] $\|\delta^*\|^2 \le h(\theta)$, where $h(\theta) = (4Y_{\max} q \|\theta\|_\infty)^2$. Since $\|\theta\|_\infty \le 2BR + L$, this concludes the proof of Theorem 4.1. The proof here is the zero-temperature limit of the proof in Meshi et al. (2012) [see Lemma 1.2 in the appendix therein].

We actually prove this bound for any δ such that $\sigma(\delta) \le \sigma(0) \equiv \kappa(\theta)$, where $\sigma(\delta) = \sum_i \max_{y_i} \left(\theta_i(y_i; w) + \sum_{c:i\in c} \delta_{ci}(y_i)\right) + \sum_c \max_{y_c} \left(\theta_c(y_c; w) - \sum_{i:i\in c} \delta_{ci}(y_i)\right)$. This obviously holds at the optimum δ^*. Our goal is to bound $\|\delta\|^2$ under this constraint. Since shifting $\delta_{ci}(\cdot)$ by a constant does not change the value of the solution, but changes the norm arbitrarily, we need to add some constraints.
In particular, we require that:

$$\sum_{y_i} \delta_{ci}(y_i) = 0 \quad \text{for all } c, i$$

We will actually find:

$$\max_\delta \|\delta\|_1 \quad \text{s.t. } \sigma(\delta) \le \kappa(\theta), \text{ and } \sum_{y_i} \delta_{ci}(y_i) = 0 \;\; \forall c, i \tag{7}$$

Since $\|\delta\|_2 \le \|\delta\|_1$ this implies a bound on $\|\delta\|_2^2$.

We begin by formulating an equivalent optimization problem to Eq. (7):

$$\begin{aligned} \max_{\delta,\bar{\delta}} \;& \frac{1}{2}\sum_c \sum_{i:i\in c} \sum_{y_i} u_{ci}(y_i)\delta_{ci}(y_i) + \frac{1}{2}\sum_c \sum_{i:i\in c} \sum_{y_i} u_{ci}(y_i)\bar{\delta}_{ci}(y_i) \\ \text{s.t. } & \sigma(\delta,\bar{\delta}) \le \kappa(\theta) \\ & \sum_{y_i} \bar{\delta}_{ci}(y_i) = 0 \;\; \forall c, i \\ & \delta = \bar{\delta} \end{aligned} \tag{8}$$

maximizing externally over $u_{ci}(y_i) \in \{-1, +1\}$, and where:

$$\sigma(\delta,\bar{\delta}) = \sum_c \max_{y_c} \left(\theta_c(y_c) - \sum_{i:i\in c} \delta_{ci}(y_i) \right) + \sum_i \max_{y_i} \left(\theta_i(y_i) + \sum_{c:i\in c} \bar{\delta}_{ci}(y_i) \right)$$

[11] To simplify notation we drop the sample index m and the dependence on w.

We will upper bound the dual of this problem.

The Lagrangian is:

$$\begin{aligned}
L(\delta,\bar{\delta},\tau,\eta,\beta) = &\frac{1}{2}\sum_c\sum_{i:i\in c}\sum_{y_i} u_{ci}(y_i)\delta_{ci}(y_i) + \frac{1}{2}\sum_c\sum_{i:i\in c}\sum_{y_i} u_{ci}(y_i)\bar{\delta}_{ci}(y_i) \\
&+ \tau\kappa(\theta) - \tau\sum_i \max_{y_i}\left(\theta_i(y_i) + \sum_{c:i\in c}\bar{\delta}_{ci}(y_i)\right) - \tau\sum_c\max_{y_c}\left(\theta_c(y_c) - \sum_{i:i\in c}\delta_{ci}(y_i)\right) \\
&+ \sum_c\sum_{i:i\in c}\sum_{y_i}\eta_{ci}(y_i)(\delta_{ci}(y_i) - \bar{\delta}_{ci}(y_i)) \\
&+ \sum_c\sum_{i:i\in c}\beta_{ci}\sum_{y_i}\bar{\delta}_{ci}(y_i)
\end{aligned}$$

with $\tau \geq 0$.

Rearranging terms we obtain:

$$\begin{aligned}
= &-\tau\sum_i\max_{y_i}\left(\theta_i(y_i) + \sum_{c:i\in c}\left(\bar{\delta}_{ci}(y_i) - \sum_{y_i'}\frac{1}{\tau}\bar{\delta}_{ci}(y_i')\left(\frac{1}{2}u_{ci}(y_i') - \eta_{ci}(y_i') + \beta_{ci}\right)\right)\right) \\
&-\tau\sum_c\max_{y_c}\left(\theta_c(y_c) - \sum_{i:i\in c}\left(\delta_{ci}(y_i) - \sum_{y_i'}\frac{1}{\tau}\delta_{ci}(y_i')\left(\frac{1}{2}u_{ci}(y_i') + \eta_{ci}(y_i')\right)\right)\right) \\
&+\tau\kappa(\theta)
\end{aligned}$$

The Lagrangian dual is therefore:

$$\begin{aligned}
= \min_{\tau\geq 0,\eta,\beta} &-\tau\sum_i\min_{\bar{\delta}_{\cdot i}(\cdot)}\max_{y_i}\left(\theta_i(y_i) + \sum_{c:i\in c}\left(\bar{\delta}_{ci}(y_i) - \sum_{y_i'}\frac{1}{\tau}\bar{\delta}_{ci}(y_i')\left(\frac{1}{2}u_{ci}(y_i') - \eta_{ci}(y_i') + \beta_{ci}\right)\right)\right) \\
&-\tau\sum_c\min_{\delta_{c\cdot}(\cdot)}\max_{y_c}\left(\theta_c(y_c) - \sum_{i:i\in c}\left(\delta_{ci}(y_i) - \sum_{y_i'}\frac{1}{\tau}\delta_{ci}(y_i')\left(\frac{1}{2}u_{ci}(y_i') + \eta_{ci}(y_i')\right)\right)\right) \\
&+\tau\kappa(\theta) \qquad (9)
\end{aligned}$$

We next replace the local singleton/factor problems with their dual problems. This yields the dual problem of (8):

$$\begin{aligned}
\min_{\tau\geq 0,\eta,\beta}\ &\tau\left(\kappa(\theta) - \sum_i\max_{\mu_i}\sum_{y_i}\mu_i(y_i)\theta_i(y_i) - \sum_c\max_{\mu_c}\sum_{y_c}\mu_c(y_c)\theta_c(y_c)\right) \\
\text{s.t}\ \ &\mu_i \geq 0,\ \ \mu_c \geq 0,\ \ \sum_{y_i}\mu_i(y_i) = 1,\ \ \sum_{y_c}\mu_c(y_c) = 1 \\
&\mu_i(y_i) = \frac{\frac{1}{2}u_{ci}(y_i) - \eta_{ci}(y_i) + \beta_{ci}}{\tau} \quad \text{for all } i, c : i \in c, y_i \\
&\mu_c(y_i) = -\frac{\frac{1}{2}u_{ci}(y_i) + \eta_{ci}(y_i)}{\tau} \quad \text{for all } c, i : i \in c, y_i \qquad (10)
\end{aligned}$$

Next, consider the objective in Eq. (10):

$$f(\tau,\eta,\beta) = \tau\left(\kappa(\theta) + \sum_i\min_{\mu_i}\sum_{y_i}\mu_i(y_i)(-\theta_i(y_i)) + \sum_c\min_{\mu_c}\sum_{y_c}\mu_c(y_c)(-\theta_c(y_c))\right)$$

For feasible μ (satisfies the constraints in Eq. (10)), it holds that:

$$f(\tau,\eta,\beta) \leq \tau\left(\kappa(\theta) + \sum_i\max_{y_i}|\theta_i(y_i)| + \sum_c\max_{y_c}|\theta_c(y_c)|\right)$$

(of course, this is true for the optimal μ as well).

Therefore, for all δ satisfying the constraints of Eq. (7), if we can find $\tau \geq 0, \eta, \beta$ such that the constraints of Eq. (10) are satisfied, then by weak duality we have:

$$\begin{aligned}
\|\delta\|_1 \leq\ & \max_u \sum_c \sum_{i:i\in c} \sum_{y_i} u_{ci}(y_i)\delta_{ci}(y_i) \\
\leq\ & f(\tau,\eta,\beta) \\
\leq\ & \tau \left(\kappa(\theta) + \sum_i \max_{y_i} |\theta_i(y_i)| + \sum_c \max_{y_c} |\theta_c(y_c)| \right)
\end{aligned} \tag{11}$$

So now we need to find $\tau \geq 0$, η and β such that μ is feasible.
Notice that in order to tighten the bound we want τ to be as small as possible.

Finally, choosing:

$$\begin{aligned}
\tau &= 2 \max_i |Y_i| \\
\eta_{ci}(y_i) &= \frac{1}{2} u_{ci}(y_i) - \frac{1}{|Y_i|} \sum_{y_i'} u_{ci}(y_i') - \frac{\tau}{|Y_i|} \\
\beta_{ci} &= -\frac{1}{|Y_i|} \sum_{y_i} u_{ci}(y_i)
\end{aligned}$$

yields:

$$\mu_i(y_i) = \frac{1}{|Y_i|}$$

So the singletons are uniform (and feasible!).
As for the factor variables:

$$\mu_c(y_i) = \frac{\frac{1}{|Y_i|}\sum_{y_i'} u_{ci}(y_i') - u_{ci}(y_i)}{2\max_{i'} |Y_{i'}|} + \frac{1}{|Y_i|}$$

Notice that if we sum this over y_i we get 1, as required. Also notice that since $-1 \leq u_{ci}(y_i) \leq 1$ then:

$$\mu_c(y_i) \geq \frac{-1-1}{2\max_{i'}|Y_{i'}|} + \frac{1}{|Y_i|} \geq -\frac{1}{|Y_i|} + \frac{1}{|Y_i|} = 0$$

as required.
So if we set:

$$\begin{aligned}
\hat{\mu}_i(y_i) &= \frac{\frac{1}{|Y_i|}\sum_{y_i'} u_{ci}(y_i') - u_{ci}(y_i)}{2\max_{i'} |Y_{i'}|} + \frac{1}{|Y_i|} \\
\mu_c(x_c) &= \prod_{i:i\in c} \hat{\mu}_i(y_i)
\end{aligned}$$

we obtain the desired (feasible!) factor marginals.
To conclude, we can use $\tau = 2\max_i |Y_i|$ in the bound of Eq. (11) to get:

$$\begin{aligned}
\|\delta\|_2 \leq \|\delta\|_1 \leq\ & 2\max_i |Y_i| \left(\kappa(\theta) + \sum_i \max_{y_i} |\theta_i(y_i)| + \sum_c \max_{y_c} |\theta_c(y_c)| \right) \\
=\ & 2\max_i |Y_i| \left(\sigma(0) + \sum_i \max_{y_i} |\theta_i(y_i)| + \sum_c \max_{y_c} |\theta_c(y_c)| \right) \\
\leq\ & 4\max_i |Y_i| \left(\sum_i \max_{y_i} |\theta_i(y_i)| + \sum_c \max_{y_c} |\theta_c(y_c)| \right) \\
\leq\ & 4 Y_{\max} q \|\theta\|_\infty \equiv \sqrt{h(\theta)}
\end{aligned}$$

C Proof of Theorem 4.2

In this section we prove Theorem 4.2. For simplicity, we denote $w_\rho^\epsilon = w(\mu_\rho^\epsilon)$ and $\delta_\rho^\epsilon = \delta(\mu_\rho^\epsilon)$.

$$\begin{aligned}
\epsilon &\geq g_\rho(w_\rho^\epsilon, \delta_\rho^\epsilon) - f_\rho(\mu_\rho^\epsilon) && \text{[duality gap bound]} \\
&\geq g_\rho(w_\rho^\epsilon, \delta_\rho^\epsilon) - g_\rho^* && [f_\rho(\mu_\rho) \leq g_\rho^* \quad \forall \mu_\rho] \\
&\geq g(w_\rho^\epsilon, \delta_\rho^\epsilon) - g_\rho^* && [g_\rho(w,\delta) \geq g(w,\delta) \quad \forall w, \delta] \\
&\geq g(w_\rho^\epsilon, \delta_\rho^\epsilon) - g^* - \frac{\rho}{2} h && \text{[Theorem 4.1]}
\end{aligned}$$

D Efficient Implementation

In this section we provide details on the implementation of Algorithm 1. Specifically, the update in line 15 of Algorithm 1 maintains primal quantities: $w = \Psi\mu$ and $\delta = A\mu$. In order to do this efficiently, we exploit the fact that at each iteration only a single $\mu_\alpha^{(m)}$ block is changed. This means that only w_α and $\delta^{(m)}$ variables that depend on $\mu_\alpha^{(m)}$ need to be updated. In particular, for the weights we obtain:

$$w_\alpha \leftarrow w_\alpha + \gamma \Psi_{m,\alpha}(s_\alpha - \mu_\alpha^{(m)}) \ ,$$

where $\mu_\alpha^{(m)}$ is the value before applying the update. Notice that only parameters pertaining to factor α are changed, so the cost is often much smaller than the full dimension d. As mentioned in Section 5, the algorithm can be implemented in terms of primal quantities. This requires storing a weight vector for each sample and factor $w_{m,\alpha} = \Psi_{m,\alpha}\mu_\alpha^{(m)}$. Again, only weights related to the specific factor α need to be stored, so the required space is often smaller than d. We can then carry out the update above in terms of $w_{m,\alpha}$ instead of $\Psi_{m,\alpha}\mu_\alpha^{(m)}$.

Similarly, for the agreement variables δ we have the update:

$$\begin{aligned}
&\text{Factor } c \text{ updated:} && \delta_{ci}^{(m)} \leftarrow \delta_{ci}^{(m)} + \frac{\gamma}{\rho M} A_{ci}\left(s_c - \mu_c^{(m)}\right) && \forall i : i \in c \\
&\text{Variable } i \text{ updated:} && \delta_{ci}^{(m)} \leftarrow \delta_{ci}^{(m)} + \frac{\gamma}{\rho M}\left(s_i - \mu_i^{(m)}\right) && \forall c : i \in c
\end{aligned}$$

where, as before, $\mu_\alpha^{(m)}$ is the value before updating. Notice that the computational cost of this update depends on the degree of the factor graph. When a factor c contains many variables in its scope, storing the marginal distribution μ_c may be prohibitive. In that case we can store instead only the marginals $\mu_{ci}^{(m)} = A_{ci}\mu_c^{(m)}$, which only requires $|Y_i|$ space (this has the same dimension as δ_{ci}, so we never have to store higher dimensional variables than the ones already stored). As before, the updates can then be implemented in terms of the compact $\mu_{ci}^{(m)}$ and $\mu_i^{(m)}$ values.

Finally, notice that we can compute the optimal step size γ in Algorithm 1 using only the auxiliary variables $w_{m,\alpha}$, $\mu_{ci}^{(m)}$ and $\mu_i^{(m)}$.

E Computing the Curvature Constant

To complete the convergence rate analysis in Section 5.1 we need to compute the curvature constant $C_{f_\rho}^{\otimes}$. It is shown in Lacoste-Julien et al. (2013) that for product domains the global curvature constant is a sum of the block-wise curvature constants: $C_{f_\rho}^{\otimes} = \sum_{m,\alpha} C_{f_\rho}^{(m,\alpha)}$. Furthermore, the curvature constant of a single block is bounded in terms of the Hessian as follows:

$$C_{f_\rho}^{(m,\alpha)} \leq \sup_{\substack{\mu,\mu' \in S, \\ (\mu'-\mu) \in S_\alpha^{(m)}, \\ z \in [\mu,\mu'] \subseteq S}} (\mu' - \mu)^\top \nabla^2 f(z)(\mu' - \mu) \ ,$$

To use this bound, we compute the Hessian for our problem[12] Eq. (5): $\nabla^2_\mu = \lambda \Psi^\top \Psi + \rho A^\top A$, which is constant in μ. Using arguments similar to Lemma A.2 in Lacoste-Julien et al. (2013), we obtain:

$$\begin{aligned}
C_{f_\rho}^{(m,\alpha)} &\leq \sup_{\substack{\mu,\mu' \in S, \\ (\mu'-\mu) \in S_\alpha^{(m)}}} (\mu' - \mu)^\top \left(\lambda \Psi^\top \Psi + \rho A^\top A\right) (\mu' - \mu) \\
&\leq \lambda \sup_{\substack{\mu,\mu' \in S, \\ (\mu'-\mu) \in S_\alpha^{(m)}}} \|\Psi(\mu' - \mu)\|_2^2 + \rho \sup_{\substack{\mu,\mu' \in S, \\ (\mu'-\mu) \in S_\alpha^{(m)}}} \|A(\mu' - \mu)\|_2^2 \\
&\leq 4\lambda \sup_{u \in \Psi S_\alpha^{(m)}} \|u\|_2^2 + 4\rho \sup_{v \in A S_\alpha^{(m)}} \|v\|_2^2 \\
&\leq \frac{16R^2}{\lambda M^2} + \frac{4\hat{R}^2}{\rho M^2}
\end{aligned}$$

where $\max_{m,\alpha,y_\alpha} \|\phi_\alpha(x^{(m)}, y_\alpha) - \phi_\alpha(x^{(m)}, y_\alpha^{(m)})\|_2 \leq 2R$ is the maximal feature difference, and $\hat{R}^2 = 1 + \max_{m,\alpha,y_\alpha} \frac{|Y_c|}{|Y_i|}$ is the maximal number of marginalized assignments.

Finally, we have:

$$C_{f_\rho}^{\otimes} = \sum_{m,\alpha} C_{f_\rho}^{(m,\alpha)} \leq 4Mq \left(\frac{4R^2}{\lambda M^2} + \frac{\hat{R}^2}{\rho M^2} \right) = O\left(\frac{q}{M} \left(\frac{1}{\lambda} + \frac{1}{\rho} \right) \right)$$

References

S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. In *ICML*, pages 53–61, 2013.

O. Meshi, T. Jaakkola and A. Globerson. Convergence rate analysis of MAP coordinate minimization algorithms. In *Advances in Neural Information Processing Systems*. 2012.

[12]Here we actually use the *negative* of Eq. (5) and treat this as a minimization problem.