Supplementary Material:
Learning the Right Model: Efficient Max-Margin Learning in Laplacian CRFs

Dhruv Batra
TTI Chicago
dbatra@ttic.edu

Ashutosh Saxena
Cornell University
asaxena@cs.cornell.edu

Abstract

In this supplementary document, we describe how the edge-weights in LCRFs may be learnt in an analogous fashion to the node-weights (Section 1); provide proofs of Theorem 1 (Section 2), Theorem 2 (Section 3) and Theorem 3 (Section 4).

1. Learning Edge-Weights

In this section, we describe how the approach in Sections 4.2 and 4.3 of the main paper can be generalized to also learn edge-weights. Recall that LCRF energy is given by:

$$E(\boldsymbol{y}\,|\,\mathcal{X},\theta) = ||\boldsymbol{y} - \mathcal{X}\theta||_1 + \sum_{(u,v)\in\mathcal{E}} |w_{uv}(y_u - y_v)|, \tag{1}$$

where w_{uv} are edge-weights. These edge-weights are themselves functions of edge-features, *i.e.* $w_{uv} = \boldsymbol{x}_{\boldsymbol{uv}}^T\beta$, where $\boldsymbol{x}_{uv}$ is a feature extracted at edge (u,v), and β is the (shared) edge parameter vector.

Similar to Section 4.2 in the main paper, let us first consider a single training sample $(\mathcal{X},\boldsymbol{y}^*)$, where $\boldsymbol{y}^*$ is the ground-truth labeling. An SSVM formulation for learning $\{\theta,\beta\}$ can be given by:

$$\min_{\theta,\beta,\xi} \quad \frac{1}{2}||\theta||_2^2 + \frac{1}{2}||\beta||_2^2 + C\,\xi \tag{2a}$$

$$\text{s.t.} \quad ||\boldsymbol{y}^i - \mathcal{X}\theta||_1 + \sum_{(u,v)} |\boldsymbol{x}_{\boldsymbol{uv}}^T\beta|\,|y_u^i - y_v^i|$$

$$-||\boldsymbol{y}^* - \mathcal{X}\theta||_1 - \sum_{(u,v)} |\boldsymbol{x}_{\boldsymbol{uv}}^T\beta|\,|y_u^* - y_v^*| \;\geq\; 1-\xi \tag{2b}$$

$$\xi \geq 0 \qquad \forall i \in \tilde{\mathcal{I}}. \tag{2c}$$

Using the same linearization tricks as used in Section 4.2 of the main paper, we now show how the above program can be relaxed into a QP with auxiliary variables: $\boldsymbol{d}^*, \{\boldsymbol{d}^i\} \in \mathbb{R}^n, \boldsymbol{e} \in \mathbb{R}^m$. Here $n = |\mathcal{V}|$, the number of nodes, and $m = |\mathcal{E}|$, the number of edges.

$$\min_{\theta,\beta,\xi,\boldsymbol{d}^*,\{\boldsymbol{d}^i\},\boldsymbol{e}} \quad \frac{1}{2}||\theta||_2^2 + \frac{1}{2}||\beta||_2^2 + C\xi$$

$$+ C_1\sum_{j=1}^n d_j^* + C_2\sum_{i\in\tilde{\mathcal{I}}}\sum_{j=1}^n d_j^i + C_3\sum_{uv} e_{uv} \tag{3a}$$

$$\text{s.t.} \quad \sum_{j=1}^n d_j^* - \sum_{j=1}^n d_j^i + \sum_{(u,v)}\Big(|y_u^i - y_v^i| - |y_u^* - y_v^*|\Big)\,e_{uv}$$

$$\geq\; 1-\xi \qquad \forall i \in \tilde{\mathcal{I}} \tag{3b}$$

$$e_{uv} \geq +\boldsymbol{x}_{uv}^T\beta, \quad e_{uv} \geq -\boldsymbol{x}_{uv}^T\beta \tag{3c}$$

$$\boldsymbol{d}^* \geq +(\boldsymbol{y}^* - \mathcal{X}\theta), \quad \boldsymbol{d}^* \geq -(\boldsymbol{y}^* - \mathcal{X}\theta) \tag{3d}$$

$$\boldsymbol{d}^i \geq +(\boldsymbol{y}^i - \mathcal{X}\theta), \quad \boldsymbol{d}^i \geq -(\boldsymbol{y}^i - \mathcal{X}\theta) \tag{3e}$$

$$\xi \geq 0. \tag{3f}$$

Note that during parameter learning, variables $\boldsymbol{y}^*, \{\boldsymbol{y}^i\}$ are known constants. Thus, similar to Section 4.2 in the main paper, all constraints in the above program are linear in $\theta,\beta,\xi,\boldsymbol{d}^*,\{\boldsymbol{d}^i\}$, and this program is a convex quadratic program, solvable by standard techniques. The extension to multiple images via Lagrangian decomposition is analogous to Section 4.3, where polytope $\mathcal{P}^{(t)}$ now refers to the linear constraints (3b)-(3e).

2. Proof of Theorem 1

Theorem 1 *Hinge Loss for the LCRF model,* i.e. $HLoss(\theta) = \max\Big\{0, ||\boldsymbol{y}^* - \mathcal{X}\theta||_1 + ||Q\boldsymbol{y}^*||_1 - \min_{\boldsymbol{y}^i}\Big(||\boldsymbol{y}^i - \mathcal{X}\theta||_1 + ||Q\boldsymbol{y}^i||_1 - \Delta(\boldsymbol{y}^i,\boldsymbol{y}^*)\Big)\Big\}$, *is non-convex in* θ.

Proof. First, let us define two functions $f(\theta), g(\theta)$:

$$f(\theta) \triangleq ||\boldsymbol{y}^* - \mathcal{X}\theta||_1 + ||Q\boldsymbol{y}^*||_1 - \min_{\boldsymbol{y}^i}\Big(||\boldsymbol{y}^i - \mathcal{X}\theta||_1 + ||Q\boldsymbol{y}^i||_1 - \Delta(\boldsymbol{y}^i,\boldsymbol{y}^*)\Big) \tag{4}$$

$$g(\theta) \triangleq \min_{\boldsymbol{y}^i}\Big(||\boldsymbol{y}^i - \mathcal{X}\theta||_1 + ||Q\boldsymbol{y}^i||_1 - \Delta(\boldsymbol{y}^i,\boldsymbol{y}^*)\Big). \tag{5}$$

Thus, we can express $f(\theta)$ and $HLoss(\theta)$ as:

$$f(\theta) = ||\boldsymbol{y}^* - \mathcal{X}\theta||_1 + ||Q\boldsymbol{y}^*||_1 - g(\theta) \tag{6}$$

$$HLoss(\theta) = \max\Big\{0, f(\theta)\Big\} \tag{7}$$

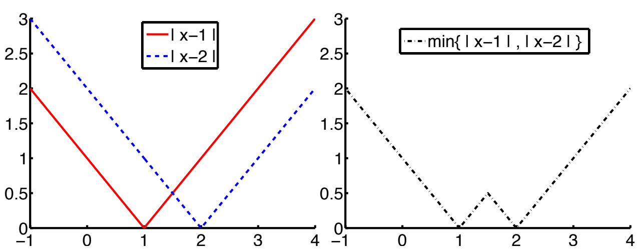
Now, we will prove that $g(\theta)$ is non-convex in θ, thus making $f(\theta)$ and $HLoss(\theta)$ non-convex as well. To this end, let us rewrite $g(\theta)$ as:

$$g(\theta) = \min_{\boldsymbol{y}^i} \Big(||\mathcal{X}\theta - \boldsymbol{y}^i||_1 + h(\boldsymbol{y}^i)\Big), \qquad \text{where} \tag{8}$$

$$h(\boldsymbol{y}^i) = ||Q\boldsymbol{y}^i||_1 - \Delta(\boldsymbol{y}^i, \boldsymbol{y}^*). \tag{9}$$

Now clearly $||\mathcal{X}\theta - \boldsymbol{y}^i||_1$ is convex in θ since ℓ_1-norm $||\cdot||$ is a convex function and composition with a linear map $\mathcal{X}\theta - \boldsymbol{y}^i$ preserves convexity [1]. Unfortunately, pointwise *mins* of a collection of convex functions is not guaranteed to be convex. Specifically, here is a simple counter-example that results in $g(\theta)$ being non-convex in θ:

$$g(\theta) = \min\Big\{|\theta - 1|, |\theta - 2|\Big\} \tag{10}$$



(a) (b)

Figure 1: Non-convexity of pointwise min of convex functions.

Clearly, the function in Fig. 1b is non-convex in θ. This completes the proof. ∎

3. Proof of Theorem 2

Let us first recall from the main manuscript the max-margin problem we are interested in solving:

$$(MM:\tilde{\mathcal{I}}) \quad \min_{\theta,\xi} \frac{1}{2}||\theta||_2^2 + C\,\xi \tag{11a}$$

$$s.t. \quad ||\boldsymbol{y}^i - \mathcal{X}\theta||_1 + ||Q\boldsymbol{y}^i||_1 - ||\boldsymbol{y}^* - \mathcal{X}\theta||_1 - ||Q\boldsymbol{y}^*||_1 \geq 1 - \xi \quad \forall i \in \tilde{\mathcal{I}} \tag{11b}$$

$$\xi \geq 0. \tag{11c}$$

In the manuscript, we showed how the non-convex program $(MM:\tilde{\mathcal{I}})$ can be *approximated* by a convex QP using auxiliary variables: $\boldsymbol{d}^*, \{\boldsymbol{d}^i\} \in \mathbb{R}^n$.

$(MMQP:\tilde{\mathcal{I}})$

$$\min_{\theta,\xi,\{\boldsymbol{d}^*\},\{\boldsymbol{d}^i\}} \quad \frac{1}{2}||\theta||_2^2 + C\,\xi + C_1\sum_{j=1}^n d_j^* + C_2\sum_{i\in\tilde{\mathcal{I}}}\sum_{j=1}^n d_j^i \tag{12a}$$

$$s.t. \quad \sum_{j=1}^n d_j^i - \sum_{j=1}^n d_j^* \geq 1 + ||Q\boldsymbol{y}^*||_1 - ||Q\boldsymbol{y}^i||_1 - \xi \tag{12b}$$

$$\xi \geq 0 \qquad \forall i \in \tilde{\mathcal{I}} \tag{12c}$$

$$\boldsymbol{d}^* \geq +(\boldsymbol{y}^* - \mathcal{X}\theta), \quad \boldsymbol{d}^* \geq -(\boldsymbol{y}^* - \mathcal{X}\theta) \tag{12d}$$

$$\boldsymbol{d}^i \geq +(\boldsymbol{y}^i - \mathcal{X}\theta), \quad \boldsymbol{d}^i \geq -(\boldsymbol{y}^i - \mathcal{X}\theta) \tag{12e}$$

Theorem 2 *If $\{\hat{\theta}, \hat{\xi}, \hat{\boldsymbol{d}}^*, \hat{\boldsymbol{d}}^i\}$ is the optimum solution of $MMQP:\tilde{\mathcal{I}}$ (12), then $\hat{\xi}$ is equal to the LCRF hinge-loss $HLoss(\hat{\theta})$, and thus an upper-bound on the loss incurred by the MAP solution,* i.e. $\hat{\xi} = HLoss(\hat{\theta}) \geq \Delta(\hat{\boldsymbol{y}}(\hat{\theta}), \boldsymbol{y}^*)$.

Proof. Since $\{\hat{\theta}, \hat{\xi}, \hat{d}^*, \hat{d}^i\}$ is the optimum solution of $MMQP:\tilde{\mathcal{I}}$ (12), it must be a feasible solution, *i.e.*:

$$\sum_{j=1}^n \hat{d}_j^i - \sum_{j=1}^n \hat{d}_j^* \geq 1 + ||Q\boldsymbol{y}^*||_1 - ||Q\boldsymbol{y}^i||_1 - \hat{\xi} \tag{13a}$$

$$\hat{d}_j^* \geq |y_j^* - \boldsymbol{x}_j^T\hat{\theta}| \qquad \forall j \in [n] \tag{13b}$$

$$\hat{d}_j^i \geq |y_j^i - \boldsymbol{x}_j^T\hat{\theta}| \qquad \forall j \in [n] \tag{13c}$$

$$\hat{\xi} \geq 0. \tag{13d}$$

Claim 1: If $C_2 > 0$, (13b) must be tight, *i.e.* hold with equality.

Proof of Claim 1: Suppose the claim is false. Thus we can reduce $\hat{d}_j^*$, *i.e.* $\hat{d}_j^* \leftarrow \hat{d}_j^* - \delta_j$ for some $\delta_j > 0$ without violating (13b). Note that this reduced $\hat{d}_j^*$ would also satisfy (13a), and thus is a feasible solution that reduces the objective function by $\sum_j \delta_j$. This is a contradiction to the optimality of $\{\hat{\theta}, \hat{\xi}, \hat{d}^*, \hat{d}^i\}$ and hence the claim must hold.

The above claim implies that $\hat{d}_j^* = |y_j^* - \boldsymbol{x}_j^T\hat{\theta}|$. Thus we can rewrite (13a)-(13c) as follows:

$$\sum_{j=1}^n \hat{d}_j^i + \hat{\xi} \geq 1 + ||Q\boldsymbol{y}^*||_1 - ||Q\boldsymbol{y}^i||_1 + |y_j^* - \boldsymbol{x}_j^T\hat{\theta}| \tag{14a}$$

$$\hat{d}_j^* = |y_j^* - \boldsymbol{x}_j^T\hat{\theta}| \qquad \forall j \in [n] \tag{14b}$$

$$\hat{d}_j^i \geq |y_j^i - \boldsymbol{x}_j^T;\hat{\theta}| \qquad \forall j \in [n] \tag{14c}$$

Claim 2: If $C_2 > C$, then (14c) must be tight, *i.e.* hold with equality.

Proof of Claim 2: This claim is a little trickier to prove. We cannot directly apply the proof of claim 1 because decreasing $\hat{d}_j^i$ might violate (14a). However, note that at least one of (14a),(14c) must be tight because if neither are tight them we can directly apply proof of claim 1 to show a contradiction.

Let us assume (14a) is tight, (14c) is not tight and try to show a contradiction. Since (14c) is not tight, we can decrease $\hat{d}^i_j$, *i.e.* $\hat{d}^i_j \leftarrow \hat{d}^i_j - \delta_j$ for some $\delta_j > 0$ without violating (14c). Since (14a) was tight, this reduced $\hat{d}^i_j$ is not part of a feasible solution as it now violates (14a). However, we can simply increase ξ_i to fix this, *i.e.* $\xi_i \leftarrow \xi_i + \sum_j \delta_j$. Clearly, the effect of the $\{\delta_j\}$ cancels outs in (14a) and the new solution $\{\xi_i + \sum_j \delta_j, \ \hat{d}^i_j - \delta_j\}$ is again a feasible solution of Program $MMQP : \tilde{\mathcal{I}}$ (12). However, if $C_2 > C$, this new solution actually *reduces* the objective function by $(C_2 - C)\sum_j \delta_j$. This is a contradiction since we started with the optimal solution, and thus claim 2 must hold.

Combining Claim 1 and Claim 2 together, we can rewrite (13a)-(13b) as follows:

$$\begin{aligned} & ||\boldsymbol{y}^i - \mathcal{X}\theta||_1 + ||Q\boldsymbol{y}^i||_1 \\ & -||\boldsymbol{y}^* - \mathcal{X}\theta||_1 - ||Q\boldsymbol{y}^*||_1 \geq 1 - \hat{\xi} && \text{(15a)} \\ & \hat{d}^*_j = |y^*_j - \boldsymbol{x}_j^T\hat{\theta}| \quad \forall j \in [n] && \text{(15b)} \\ & \hat{d}^i_j = |y^i_j - \boldsymbol{x}_j^T;\hat{\theta}| \quad \forall j \in [n] && \text{(15c)} \end{aligned}$$

Comparing (15a) with (11), we can see that constraint (15a) is equivalent to constraint (11b), which expresses the structured hinge-loss. Thus, $\hat{\xi} = HLoss(\hat{\theta}) \geq \Delta(\hat{\boldsymbol{y}}(\hat{\theta}), \boldsymbol{y}^*)$. This completes the proof. ∎

4. Proof of Theorem 3

Let us first recall from the manuscript the generalization of the parameter learning problem $(MMQP : \tilde{\mathcal{I}})$ to multiple training images:

$$(MMQP : \tilde{\mathcal{I}}^{\mathcal{T}})$$

$$\begin{aligned} & \min_{\theta, \{\xi^{(t)}, \boldsymbol{D}^{(t)}\}} \frac{1}{2}||\theta||_2^2 + \frac{C}{T}\sum_{t\in\mathcal{T}} \xi^{(t)} + \frac{C'}{T}\sum_{t\in\mathcal{T}} \boldsymbol{D}^{(t)} \cdot \mathbf{1} && \text{(16a)} \\ & s.t. \quad \{\theta, \xi^{(t)}, \boldsymbol{D}^{(t)}\} \in \mathcal{P}^{(t)} \quad \forall t \in \mathcal{T}. && \text{(16b)} \end{aligned}$$

Also recall that we presented a Lagrangian relaxation based dual-decomposition algorithm that first allocated to each training image its own copy of the parameters $\theta^{(t)}$:

$$(MMQP : \tilde{\mathcal{I}}^{\mathcal{T}}2)$$

$$\begin{aligned} & \min_{\tilde{\theta}, \{\theta^{(t)}, \xi^{(t)}, \boldsymbol{D}^{(t)}\}} \frac{1}{2T}\sum_{t\in\mathcal{T}} ||\theta^{(t)}||_2^2 + \frac{C}{T}\sum_{t\in\mathcal{T}} \xi^{(t)} \\ & \qquad + \frac{C'}{T}\sum_{t\in\mathcal{T}} \boldsymbol{D}^{(t)} \cdot \mathbf{1} && \text{(17a)} \\ & s.t. \quad \{\theta^{(t)}, \xi^{(t)}, \boldsymbol{D}^{(t)}\} \in \mathcal{P}^{(t)} && \text{(17b)} \\ & \qquad \theta^{(t)} = \tilde{\theta} \quad \forall t \in \mathcal{T}. && \text{(17c)} \end{aligned}$$

The Lagrangian dual of $(MMQP : \tilde{\mathcal{I}}^{\mathcal{T}})$ was derived as:

$$\begin{aligned} & (LD : \tilde{\mathcal{I}}^{\mathcal{T}}) \quad \max_{\{\boldsymbol{\lambda}^{(t)}\}} \sum_{t\in\mathcal{T}} \mathcal{F}^{(t)}(\boldsymbol{\lambda}^{(t)}) && \text{(18a)} \\ & \qquad s.t. \quad \sum_{t\in\mathcal{T}} \boldsymbol{\lambda}^{(t)} = 0, && \text{(18b)} \end{aligned}$$

where $\mathcal{F}^{(t)}$ are independent sub-problems that are functions of the dual variables $(\boldsymbol{\lambda}^{(t)})$:

$$\begin{aligned} \mathcal{F}^{(t)}(\boldsymbol{\lambda}^{(t)}) = & \min_{\theta^{(t)}, \xi^{(t)}, \boldsymbol{D}^{(t)}} \frac{1}{2T}||\theta^{(t)}||_2^2 + \boldsymbol{\lambda}^{(t)} \cdot \theta^{(t)} \\ & + \frac{C}{T}\xi^{(t)} + \frac{C'}{T}\boldsymbol{D}^{(t)} \cdot \mathbf{1} && \text{(19a)} \\ & s.t. \quad \{\theta^{(t)}, \xi^{(t)}, \boldsymbol{D}^{(t)}\} \in \mathcal{P}^{(t)}. && \text{(19b)} \end{aligned}$$

Algorithm 1. We solve this dual problem via projected gradient ascent.

Theorem 3 *$LD : \tilde{\mathcal{I}}^{\mathcal{T}}$ (18) has zero duality gap and Algorithm 1 converges to the optimum of $MMQP : \tilde{\mathcal{I}}^{\mathcal{T}}$ (16).*

Proof. Our proof consists of the following steps:

1. **Convexity of $LD : \tilde{\mathcal{I}}^{\mathcal{T}}$ (18):** First, note that by construction, a Lagrangian dual is always concave in multipliers $\{\boldsymbol{\lambda}^{(t)}\}$ since it is a point-wise minimum of concave (linear) functions of $\{\boldsymbol{\lambda}^{(t)}\}$.

2. **Optimality of Algorithm 1 for $LD : \tilde{\mathcal{I}}^{\mathcal{T}}$ (18):** Thus, projected gradient ascent converges to the solution of (18).

3. **Zero Duality of $LD : \tilde{\mathcal{I}}^{\mathcal{T}}$ (18):** To show this, we note that $MMQP : \tilde{\mathcal{I}}^{\mathcal{T}}2$ (17) is a convex problem because it has convex (linear) constraints and the Hessian of objective is positive-definite:

$$\begin{aligned} & \frac{\partial^2 f}{\partial\theta^{(t)}\partial\theta^{(t)}} = \frac{1}{2T} I_{k\times k} \succeq 0 && \text{(20a)} \\ & \frac{\partial^2 f}{\partial\xi^{(t)}\partial\xi^{(t)}} = 0 && \text{(20b)} \\ & \frac{\partial^2 f}{\partial\boldsymbol{D}^{(t)}\partial\boldsymbol{D}^{(t)}} = 0 && \text{(20c)} \\ & \frac{\partial^2 f}{\partial\theta^{(t)}\partial\xi^{(t)}} = 0 && \text{(20d)} \\ & \frac{\partial^2 f}{\partial\xi^{(t)}\partial\boldsymbol{D}^{(t)}} = 0 && \text{(20e)} \\ & \frac{\partial^2 f}{\partial\theta^{(t)}\partial\boldsymbol{D}^{(t)}} = 0 && \text{(20f)} \\ & && \text{(20g)} \end{aligned}$$

where,

$$f(\theta^{(t)}, \xi^{(t)}, \boldsymbol{D^{(t)}}) = \frac{1}{2T}||\theta^{(t)}||_2^2 + \boldsymbol{\lambda}^{(t)} \cdot \theta^{(t)} + \frac{C}{T}\xi^{(t)} + \frac{C'}{T}\boldsymbol{D^{(t)}} \cdot \mathbf{1} \quad (21)$$

Moreover, $MMQP : \tilde{\mathcal{I}}^{\mathcal{T}}2$ (17) satisfies Slater's condition [1], which is a sufficient condition for zero duality gap in convex problems, since it has a non-empty feasible set.

Thus, $(LD : \tilde{\mathcal{I}}^{\mathcal{T}})$ (18) achieves the same value as $MMQP : \tilde{\mathcal{I}}^{\mathcal{T}}2$ (17), which in turn has the same value as $MMQP : \tilde{\mathcal{I}}^{\mathcal{T}}$ (16). This completes the proof. ∎

References

[1] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004. 2, 4