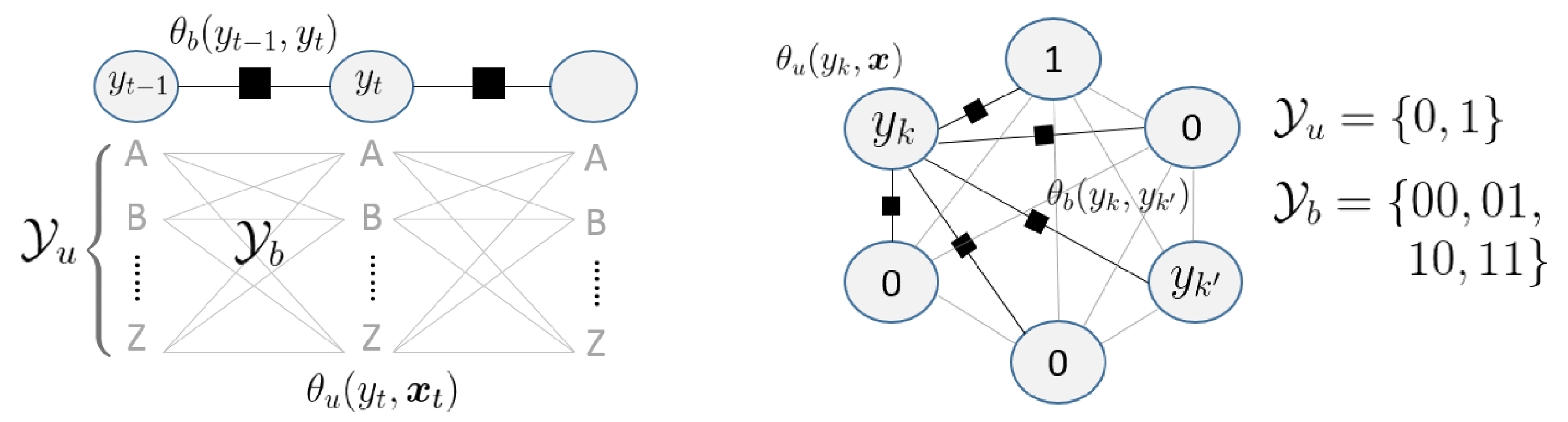
Dual-Decomposed Learning with Factorwise Oracles for Structural SVMs of Large Output Domain

Ian E.H. Yen[1], Xiangru Huang[2], Kai Zhong[2], Ruohan Zhang[2], Pradeep Ravikumar[1] and Inderjit S. Dhillon[2].
[1] Carnegine Mellon University. [2] University of Texas at Austin.

Abstract

- Many applications of machine learning involve structured outputs with large domains, such as Translation, Alignment, and Parsing.
- Learning of a structured predictor is prohibitive due to repetitive calls to an expensive inference oracle.
- We propose decomposing training of a structural SVM into factorwise multiclass SVMs connected with messages, replacing structured oracles with factorwise oracles.
- The proposed algorithm, *Greedy Direction Method of Multiplier (GDMM)*, guarantees ϵ-suboptimality in $O(\log(1/\epsilon))$ iterations, and shows orders-of-magnitude speedup over state of the art on large-domain problems.

Structured Prediction of Large Output Domain



- We considers Structured Predictor of the form
$$h(x; w) = arg \max_{y \in \mathcal{Y}(x)} \langle w, \phi(x, y) \rangle.$$
obtained by solving the regularized Empirical Risk Minimization problem
$$\min_{w} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} L(w; x_i, \bar{y}_i) .$$
- For Structural SVM, we use the structured hinge loss
$$L(w; x, \bar{y}) = \max_{y \in \mathcal{Y}(x)} \langle w, \ \phi(x, y) - \phi(x, \bar{y}) \rangle + \delta(y, \bar{y}),$$
where the inner product allows factor decomposition of the form
$$\langle w, \ \phi(x, y) \rangle = \sum_{F \in \mathcal{T}} \sum_{f \in F(x)} \langle w_F, \phi_F(x_f, y_f) \rangle,$$
and $\delta(y, \bar{y}_i)$ is a task-dependent error function (usually Hamming Error).
- Evaluation of the loss $L(w; x, \bar{y})$ (and its derivative) requires maximization over the structured domain $\mathcal{Y}(x)$ (i.e. structured oracle), a very expensive inference procedure when domain $|\mathcal{Y}_f|$ or #factor $|\mathcal{F}|$ is large.

Existing Approaches

- Approximate inference via Beam Search (suboptimal due to local decision).
- Pseudolikelihood (high-variance estimator downgrades testing performance).
- Generative model + Discriminative Re-ranking k-best.

Dual Decomposition: Struct-SVM to Multiclass SVMs

- **Key Insight:** the Factorwise Oracle
$$y_f^* := argmax_{y_f} \langle w_F, \phi(x_f, y_f) \rangle$$
can be solved cheaply, even in sublinear time.
- Replace the maximization domain $\mathcal{Y}(x)$ with its Linear-Program (LP) relaxation $\mathcal{M}_L$, giving the LP-relaxed loss
$$L^{LP}(w; x, \bar{y}) \geq L(w; x, \bar{y}),$$
which is tight for tree-structured factor graph.
- Apply strong duality to the LP relaxation gives the dual-decomposed loss:
$$L^{LP}(w; x, \bar{y}) = \min_{\lambda \in \Lambda} \sum_{f \in \mathcal{F}(x)} L_f(w; x_f, \bar{y}_f, \lambda_f).$$
where $\lambda : \sum_f \lambda_{jf} = 0, \ j \in \mathcal{V}(x)$ plays the role of messages and
$$L_f(w_F, \lambda_f) := \max_{y_f \in \mathcal{Y}_f} \langle w_F, \bar{\phi}_F(x_f, y_f) \rangle + \sum_{j \in \mathcal{N}(f)} \lambda_{jf}([y_f]_j)$$
is a multiclass SVM loss augmented with messages λ_f.
- The dual problem comprises independent multiclass SVM problems (in dual forms):
$$\min_{\alpha_f \in \Delta^{|\mathcal{Y}_f|}} G(\alpha) := \frac{1}{2} \sum_{F \in \mathcal{T}} \left\| \sum_{f \in F} \Phi_f^T \alpha_f \right\|^2 - \sum_{j \in \mathcal{V}} \delta_j^T \alpha_j$$
connected by consistency constraints $M_{jf}\alpha_f = \alpha_j, \ (j, f) \in \mathcal{E}$.

Greedy Direction Method of Multiplier (GDMM)

- Use Augmented Lagrangian Method:
$$\mathcal{L}(\alpha, \lambda) := G(\alpha) + \frac{\rho}{2} \sum_{(j,f) \in \mathcal{E}} \|m_{jf}(\alpha, \lambda)\|^2 \qquad (1)$$
where $m_{jf}(\alpha, \lambda^t) = M_{jf}\alpha_f - \alpha_j + \lambda_{jf}^t$ are the messages between factors.

GDMM Algorithm:

for $t = 0, 1, \ldots$ **do**
1. Compute $(\alpha^{t+1}, \mathcal{A}^{t+1})$ via one pass of Algorithm 1 or 2.
2. $\lambda_{jf}^{t+1} = \lambda_{jf}^t + \eta \left(M_{jf}\alpha_f^{t+1} - \alpha_j^{t+1} \right), \ j \in \mathcal{N}(f), \ \forall f \in \mathcal{F}$.

end for

When Factor Domain $|\mathcal{Y}_f|$ is Large.

Algorithm 1 Block-Coordinate Frank-Wolfe (BCFW)

for $s = 1$ to $|\mathcal{F}|$ **do**
1. Draw $f \in \mathcal{F}$ uniformly at random.
2. Find the incorrect label y_f^* by factorwise oracle:
$$v_f^+ := \underset{v_f \in \Delta^{|\mathcal{Y}_f|}}{argmin} \langle \nabla_{\alpha_f} \mathcal{L}(\alpha^t, \lambda^t), v_f \rangle = C(e_{\bar{y}_f} - e_{y_f^*}).$$
3. $\mathcal{A}_f^{s+1} = \mathcal{A}_f^s \cup \{v_f^+\}$.
4. Minimize $\mathcal{L}(\alpha, \lambda^t)$ w.r.t. the active set $\mathcal{A}_f^{s+1}$.

end for

- Messages $m_{jf}(\alpha, \lambda)$ have size bounded by active label size $|\mathcal{A}_{f'}|$ of neighboring factor f'.
- A pairwise factorwise oracle can be realized in time $O(|\mathcal{A}_i|^2)$ instead of $O(|\mathcal{Y}_i|^2)$ by maintaining priority queues for $w_F(\alpha)$.

When Number of Factors $|\mathcal{F}|$ is Large.

Algorithm 2 Block-Greedy Coordinate Descent (BGCD)

for $i \in [n]$ **do**
1. $f^* := \underset{f \in \mathcal{F}(x_i)}{argmin} \left(\min_{\alpha_f + d \in \Delta^{|\mathcal{Y}_f|}} \langle \nabla_{\alpha_f} \mathcal{L}(\alpha^t, \lambda^t), d \rangle + \frac{Q_{\max}}{2} \|d\|^2 \right)$.
2. $\mathcal{A}_i^{s+1} = \mathcal{A}_i^s \cup \{f^*\}$.
3. Minimize $\mathcal{L}(\alpha, \lambda^t)$ w.r.t. $\{\alpha_f\}_{f \in \mathcal{A}_i^{s+1}}$.

end for

- The number of active factors of sample i is bounded by $|\mathcal{A}_i|$.
- Only $O(|\mathcal{A}_i|^2)$ pairwise factors require gradient computation (others can be compared using priority queues maintained on $w_F(\alpha)$).

Convergence Analysis

Let $d(\lambda) = \min_{\alpha} \ \mathcal{L}(\alpha, \lambda)$ and
$$\Delta_d^t := d^* - d(\lambda^t), \ \ \Delta_p^t := \mathcal{L}(\alpha^t, \lambda^t) - d(\lambda^t)$$
be the dual and primal suboptimality respectively. The *GDMM* algorithm has
$$E[\Delta_p^t + \Delta_d^t] \leq \epsilon \ \ for \ \ t \geq \omega \log(\frac{1}{\epsilon}) \qquad (2)$$
for some constant $\omega > 0$.

Experiments

- **Sequence Labeling:** POS ($|\mathcal{Y}_i| = 45$), ChineseOCR ($|\mathcal{Y}_i| = 3039$).
- Structural oracle uses Viterbi Algorithm.
- **Multilabel with Pairwise Interaction:** RCV1 ($|\mathcal{Y}_i| = 228$).
- Structural oracle solves a Linear Program.