### Variable-Order CRFs

**Goal:** Define a good conditional distribution over tag sequences.

\[
p_\theta(x) = \frac{1}{Z_\theta(x)} \exp \left( \sum_{t=1}^{n-1} \theta^T f(x, t, y_{t-k-1} \ldots y_t) \right)
\]

Certain combinations go well together, some don’t.

- **label-word:** D-the \( \vee \), V-the \( \wedge \)
- **label-label:** D-N \( \vee \), D-V \( \wedge \)

Sometimes, it’s useful to look at larger label combinations.

**The problem:** For features to look at output contexts of size \( k \), we need \( O(n|\mathcal{Y}|^k) \) time for inference even if most combinations don’t improve the model, e.g., combinations that are easily ruled out by local features.

\[
p_\theta(y|x) = \frac{1}{Z_\theta(x)} \exp \left( \sum_{t=1}^{n-1} \theta^T f(x, t, y_{t-k-1} \ldots y_t) \right)
\]

**The VoCRF idea:** Remove output contexts that aren’t necessary!

Turn \( O(n|\mathcal{Y}|^k) \) into \( O(n|\mathcal{W}|) \)

**Technical details:**
- Need closure under prefixes & last-character substitution
- Can lift assumption with \( \theta \)-arcs.

**Very flexible**
- No need to specify a fixed size.
- Covers semi-Markov & higher-order
- (Ye et al., 2009, Cuong et al., 2014)

**"Correcting" prior work**

Original algorithm for computing gradients and expectations was uncessarily slow and complicated. Our revised algorithm is \( O(|\mathcal{W}|) \) times faster \( O(a \text{ few pages}) \) simpler!

- Just run autodiff on their forward algorithm!
- Protip: Evaluating the gradient should be as fast as the function!
- Check out Jason’s paper at the structured prediction workshop for more on the connection between autodiff and inference.

### Structured Sparsity

**Goal:** Select higher-order features \( \mathcal{W} \), which gives us the best possible accuracy under a budget for runtime.

**How:** Augmenting the training objective with a penalty for runtime!

\[
\sum_{i=1}^{m} \log p_\theta(y^{(i)}|x^{(i)}) + \lambda \| \theta \|_2^2 + \gamma R(\theta)
\]

\( \theta \) implicitly encodes \( \mathcal{W} \) in its nonzero entries.

**Sparsity -> Speed**

**Dependencies among features**
- prefix closure
- last tag subst. closure

**Ideal runtime**

\[
\mathcal{R}(\theta) = |\mathcal{W}| = \sum_{w \in \mathcal{Y}^*} \| \theta_{G_w} \|_2^2
\]

**Convex surrogate**

\[
\mathcal{R}(\theta) = \sum_{w \in \mathcal{Y}^*} \| \theta_{G_w} \|_2^2
\]

**Active set**

- Grow \( \mathcal{W}^{\text{active}} \)
- Unavailable

---

### Experiments

- Part of speech tagging with Universal Tags in 5 languages.
- Best system in **bold**.
- Superscript \( k \) indicates a significant difference from the k-CRF’s accuracy (paired-permutation \( p < 0.5 \), color indicates **better** or **worse**.
- Underlined system is the fastest “statistically indistinguishable” \( \mathcal{W} \) model compared to the 2-CRF.