Structure Regularization for Structured Prediction

Xu SUN
xusun@pku.edu.cn

Peking University
Structured prediction

Structured prediction methods are useful for many areas
- Natural language processing (NLP)
- Vision recognition
- Signal processing
- Bioinformatics
- Speech recognition
- Etc.
For example, many natural language processing (NLP) tasks are **structured prediction** tasks

- Parsing
- SMT
- POS tagging
- Word segmentation
- Named entity recognition
- Chunking
Given a structured prediction task, is the scale of structure matters?

He reckons the current account deficit will narrow to only #1.8 billion in September.

Structured prediction model (e.g., CRF, HMM, MEMM, or perceptron)
Given a structured prediction task, is the scale of structure matters?

He reckons the current account deficit will narrow to only # 1.8 billion in September.
Given a structured prediction task, is the scale of structure matters?

How about this scale?

structured prediction model (e.g., CRF, HMM, MEMM, or perceptron)
Basic question

- **Sub-question-1**
  - Given a structured prediction task, is the scale of structure matters?

- **Sub-question-2:**
  - If it matters, which scale is the best?
    - E.g., most of the tasks are based on sentence level, but is it really a good choice?

- **Sub-question-3:**
  - How to find the best scale of complexity in practice?
Current research trend → using more and more complex structures
- E.g., long distance features, high order dependencies, global information
- This is helpful to some tasks, but also helpless (even harmful) to some other tasks, Why??

Our study
- Theoretical analysis:
  - Complex structures is not always good
  - → it can be harmful to generalization ability
  - → we need to find an optimal scale of complexity
- Proposed a solution: structure regularization (SR)
Theoretical analysis: Overfitting risk

**Theorem 4 (Generalization vs. structure regularization)** Let the structured prediction objective function of $G$ be penalized by structure regularization with factor $\alpha \in [1, n]$ and $L_2$ weight regularization with factor $\lambda$, and the penalized function has a minimizer $f$:

$$f = \arg\min_{g \in \mathcal{F}} R_{\alpha, \lambda}(g) = \arg\min_{g \in \mathcal{F}} \left( \frac{1}{mn} \sum_{j=1}^{m\alpha} \mathcal{L}_\tau(g, z'_j) + \frac{\lambda}{2} \|g\|_2^2 \right)$$  \hspace{1cm} (8)

Assume the point-wise loss $\ell_\tau$ is convex and differentiable, and is bounded by $\ell_\tau(f, z, k) \leq \gamma$. Assume $f(x, k)$ is $\rho$-admissible. Let a local feature value be bounded by $v$ such that $x_{(k, q)} \leq v$ for $q \in \{1, \ldots, d\}$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the random draw of the training set $S$, the generalization risk $R(f)$ is bounded by

$$R(f) \leq R_e(f) + \frac{2d\tau^2 \rho^2 v^2 n^2}{m\lambda\alpha} + \left( \frac{(4m - 2)d\tau^2 \rho^2 v^2 n^2}{m\lambda\alpha} + \gamma \right) \sqrt{\frac{\ln \delta^{-1}}{2m}}$$  \hspace{1cm} (9)

- **Expected risk** (risk on test data)
- **Empirical risk** (risk on training data)
- **Overfitting risk** (risk of overfitting from training data to test data)
Theoretical analysis: Overfitting risk

**Theorem 4 (Generalization vs. structure regularization)** Let the structured prediction objective function of $G$ be penalized by structure regularization with factor $\alpha \in [1, n]$ and $L_2$ weight regularization with factor $\lambda$, and the penalized function has a minimizer $f$:

$$
f = \arg\min_{g \in \mathcal{F}} R_{\alpha, \lambda}(g) = \arg\min_{g \in \mathcal{F}} \left( \frac{1}{mn} \sum_{j=1}^{m\alpha} \mathcal{L}_\tau(g, z'_j) + \frac{\lambda}{2} \|g\|_2^2 \right) \quad (8)
$$

Assume the point-wise loss $\ell_\tau$ is convex and differentiable, and is bounded by $\ell_\tau(f, z, k) \leq \gamma$. Assume $f(x, k)$ is $\rho$-admissible. Let a local feature value be bounded by $v$ such that $x_{(k,q)} \leq v$ for $q \in \{1, \ldots, d\}$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the random draw of the training set $S$, the generalization risk $R(f)$ is bounded by

$$
R(f) \leq R_e(f) + \frac{2d^2 \rho^2 v^2 n^2}{m\lambda\alpha} + \left( \frac{(4m - 2)d^2 \rho^2 v^2 n^2}{m\lambda\alpha} + \gamma \right) \sqrt{\frac{\ln \delta^{-1}}{2m}} \quad (9)
$$

**Complexity of structure (nodes of a training sample with structured dependencies)**

$\Rightarrow$ Complex structure leads to higher overfitting risk
Theoretical analysis: Overfitting risk

Theorem 4 (Generalization vs. structure regularization) Let the structured prediction objective function of $G$ be penalized by structure regularization with factor $\alpha \in [1, n]$ and $L_2$ weight regularization with factor $\lambda$, and the penalized function has a minimizer $f$:

$$f = \arg\min_{g \in F} R_{\alpha, \lambda}(g) = \arg\min_{g \in F} \left( \frac{1}{mn} \sum_{j=1}^{m\alpha} \mathcal{L}_\tau(g, z'_j) + \frac{\lambda}{2} \|g\|_2^2 \right)$$

Assume the point-wise loss $\ell_\tau$ is convex and differentiable, and is bounded by $\ell_\tau(f, z, k) \leq \gamma$. Assume $f(x, k)$ is $\rho$-admissible. Let a local feature value be bounded by $v$ such that $x_{(k,q)} \leq v$ for $q \in \{1, \ldots, d\}$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the random draw of the training set $S$, the generalization risk $R(f)$ is bounded by

$$R(f) \leq R_e(f) + \frac{2d\tau^2 \rho^2 v^2 n^2}{m\lambda\alpha} + \left( \frac{(4m - 2)d\tau^2 \rho^2 v^2 n^2}{m\lambda\alpha} + \gamma \right) \sqrt{\frac{\ln \delta^{-1}}{2m}}$$

Strength of structure regularization (strength of decomposition)

$\rightarrow$ Stronger SR leads to reduction of overfitting risk
Theoretical analysis: Overfitting risk

Theorem 4 (Generalization vs. structure regularization) Let the structured prediction objective function of $G$ be penalized by structure regularization with factor $\alpha \in [1, n]$ and $L_2$ weight regularization with factor $\lambda$, and the penalized function has a minimizer $f$:

$$f = \arg\min_{g \in \mathcal{F}} R_{\alpha, \lambda}(g) = \arg\min_{g \in \mathcal{F}} \left( \frac{1}{mn} \sum_{j=1}^{m\alpha} \mathcal{L}_\tau(g, z'_j) + \frac{\lambda}{2} ||g||_2^2 \right) \quad (8)$$

Assume the point-wise loss $\ell_\tau$ is convex and differentiable, and is bounded by $\ell_\tau(f, z, k) \leq \gamma$. Assume $f(x, k)$ is $\rho$-admissible. Let a local feature value be bounded by $v$ such that $x_{(k, q)} \leq v$ for $q \in \{1, \ldots, d\}$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the random draw of the training set $S$, the generalization risk $R(f)$ is bounded by

$$R(f) \leq R_e(f) + \frac{2d\tau^2 \rho^2 v^2 n^2}{m\lambda\alpha} + \left( \frac{(4m - 2)d\tau^2 \rho^2 v^2 n^2}{m\lambda\alpha} + \gamma \right) \sqrt{\frac{\ln \delta^{-1}}{2m}} \quad (9)$$

Number of training samples

→ More training samples leads to reduction of overfitting risk
Theoretical analysis: Overfitting risk

**Theorem 4 (Generalization vs. structure regularization)** Let the structured prediction objective function of $G$ be penalized by structure regularization with factor $\alpha \in [1, n]$ and $L_2$ weight regularization with factor $\lambda$, and the penalized function has a minimizer $f$:

$$f = \arg\min_{g \in \mathcal{F}} R_{\alpha, \lambda}(g) = \arg\min_{g \in \mathcal{F}} \left( \frac{1}{mn} \sum_{j=1}^{m\alpha} \mathcal{L}_\tau(g, z'_j) + \frac{\lambda}{2} \|g\|_2^2 \right)$$  \hspace{1cm} (8)

Assume the point-wise loss $\ell_\tau$ is convex and differentiable, and is bounded by $\ell_\tau(f, z, k) \leq \gamma$. Assume $f(x, k)$ is $\rho$-admissible. Let a local feature value be bounded by $v$ such that $x_{(k, q)} \leq v$ for $q \in \{1, \ldots, d\}$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the random draw of the training set $S$, the generalization risk $R(f)$ is bounded by

$$R(f) \leq R_e(f) + \frac{2d\tau^2 \rho^2 v^2 n^2}{\lambda \alpha} + \left( \frac{(4m - 2)d\tau^2 \rho^2 v^2 n^2}{\lambda \alpha} + \gamma \right) \sqrt{\frac{\ln \delta^{-1}}{2m}}$$  \hspace{1cm} (9)

✓ **Conclusions from our analysis:**

1. **Complex structure** $\rightarrow$ low empirical risk & high overfitting risk
2. **Simple structure** $\rightarrow$ high empirical risk & low overfitting risk
3. Need a balanced complexity of structures
Theoretical analysis: Overfitting risk

**Theorem 4 (Generalization vs. structure regularization)** Let the structured prediction objective function of $G$ be penalized by structure regularization with factor $\alpha \in [1, n]$ and $L_2$ weight regularization with factor $\lambda$, and the penalized function has a minimizer $f$:

$$
\begin{align*}
   f &= \arg\min_{g \in \mathcal{F}} R_{\alpha, \lambda}(g) = \arg\min_{g \in \mathcal{F}} \left( \frac{1}{mn} \sum_{j=1}^{m\alpha} \mathcal{L}(g, z'_j) + \frac{\lambda}{2} \|g\|_2^2 \right) \\
   \text{Assume the point-wise loss } \mathcal{L} \text{ is convex and differentiable, and is bounded by } \mathcal{L}(f, z, k) \leq \gamma. \text{ Assume } f(x, k) \text{ is } \rho \text{-admissible. Let a local feature value be bounded by } v \text{ such that } x_{(k, q)} \leq v \text{ for } q \in \{1, \ldots, d\}. \text{ Then, for any } \delta \in (0, 1), \text{ with probability at least } 1 - \delta \text{ over the random draw of the training set } S, \text{ the generalization risk } R(f) \text{ is bounded by}
\end{align*}
$$

$$
R(f) \leq R_e(f) + \frac{2d\tau^2 \rho^2 v^2 n^2}{m\lambda\alpha} + \left( \frac{(4m - 2)d\tau^2 \rho^2 v^2 n^2}{m\lambda\alpha} + \gamma \right) \sqrt{\frac{\ln \delta^{-1}}{2m}} 
$$

- **In other words, more intuitively:**
  1. Too complex structure $\rightarrow$ high accuracy on training + very easy to overfit $\rightarrow$ low accuracy on testing
  2. Too simple structure $\rightarrow$ very low accuracy on training + not easy to overfit $\rightarrow$ low accuracy on testing

Proper structure $\rightarrow$ good accuracy on training + not easy to overfit $\rightarrow$ high accuracy on testing
Theoretical analysis: Overfitting risk

**Theorem 4 (Generalization vs. structure regularization)** Let the structured prediction objective function of $G$ be penalized by structure regularization with factor $\alpha \in [1, n]$ and $L_2$ weight regularization with factor $\lambda$, and the penalized function has a minimizer $f$:

$$ f = \arg \min_{g \in \mathcal{F}} R_{\alpha, \lambda}(g) = \arg \min_{g \in \mathcal{F}} \left( \frac{1}{mn} \sum_{j=1}^{m\alpha} \mathcal{L}_\tau(g, z'_j) + \frac{\lambda}{2} \|g\|_2^2 \right) $$

(8)

Assume the point-wise loss $\ell_\tau$ is convex and differentiable, and is bounded by $\ell_\tau(f, z, k) \leq \gamma$. Assume $f(x, k)$ is $\rho$-admissible. Let a local feature value be bounded by $v$ such that $x_{(k,q)} \leq v$ for $q \in \{1, \ldots, d\}$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the random draw of the training set $S$, the generalization risk $R(f)$ is bounded by

$$ R(f) \leq R_e(f) + \frac{2d\tau^2 \rho^2 v^2 n^2}{m\lambda \alpha} + \left( \frac{(4m - 2)d\tau^2 \rho^2 v^2 n^2}{m\lambda \alpha} + \gamma \right) \sqrt{\frac{\ln \delta^{-1}}{2m}} $$

(9)

1. Simple structure $\rightarrow$ low overfitting risk & high empirical risk
2. Complex structure $\rightarrow$ high overfitting risk & low empirical risk
3. Need a balanced complexity of structures

Some intuition in the proof (as in the full version paper):
1) The decomposition can improve **stability**
2) Better stability leads to better **generalization** (less overfitting)
Theoretical analysis: Learning speed

Proposition 5 (Convergence rates vs. structure regularization) With the aforementioned assumptions, let the SGD training have a learning rate defined as \( \eta = \frac{c \epsilon \beta \alpha^2}{q \kappa^2 n^2} \), where \( \epsilon > 0 \) is a convergence tolerance value and \( \beta \in (0, 1] \). Let \( t \) be an integer satisfying

\[
 t \geq \frac{q \kappa^2 n^2 \log (qa_0/\epsilon)}{\epsilon \beta c^2 \alpha^2} \tag{15}
\]

where \( n \) and \( \alpha \in [1, n] \) is like before, and \( a_0 \) is the initial distance which depends on the initialization of the weights \( w_0 \) and the minimizer \( w^* \), i.e., \( a_0 = ||w_0 - w^*||^2 \). Then, after \( t \) updates of \( w \) it converges to \( \mathbb{E}[g(w_t) - g(w^*)] \leq \epsilon \).

- SR also with faster speed
  (a by-product of simpler structures)

- ✔️ using structure regularization can quadratically accelerate the convergence rate
- Complex structures (high complexity)

- Simple structures (low complexity)
We propose **structure regularization (SR)** to find good complexity

- Simply split the structures!
- Can (almost) be seen as a preprocessing step of the training data
Will the split causes feature loss? – loss of long distance features?

No loss of any (long distance) features

→ We can first extract features, then split the structures
→ Or, by simply copying observations to mini-samples, i.e., the split is only on tag-structures, like this:
Is structure regularization also required for test data?

No, no use of SR for testing data (in current implementation & experiments)

→ Like other regularization methods, SR is only for the training

→ i.e., No SR on the test stage (no decomposition of test samples)!
Structure regularization

Structure & weight regularization

\[ R_{\alpha, \lambda}(G_S) \triangleq R_{\alpha}(G_S) + N_{\lambda}(G_S) \]

Algorithm 1 Training with structure regularization

1: **Input**: model weights \( w \), training set \( S \), structure regularization strength \( \alpha \)
2: **repeat**
3: \( S' \leftarrow \emptyset \)
4: **for** \( i = 1 \rightarrow m \) **do**
5: \( \text{Randomly decompose } z_i \in S \text{ into mini-samples } N_{\alpha}(z_i) = \{z_{i,1}, \ldots, z_{i,\alpha}\} \)
6: \( S' \leftarrow S' \cup N_{\alpha}(z_i) \)
7: **end for**
8: **for** \( i = 1 \rightarrow |S'| \) **do**
9: \( \text{Sample } z' \text{ uniformly at random from } S', \text{ with gradient } \nabla g_{z'}(w) \)
10: \( w \leftarrow w - \eta \nabla g_{z'}(w) \)
11: **end for**
12: **until** Convergence
13: **return** \( w \)

The implementation is very simple
Some advantages

- If the original obj. function is convex, can still keep the convexity of the objective function

- No conflict with the weight regularization
  - E.g., L2, and/or L1 regularization

- General purpose and model-independent (because act like a preprocessing step)
  - E.g., can be used for different types of models, including CRFs, perceptrons, & neural networks
State-of-the-art scores on competitive tasks

Experiments-1: accuracy

![Graphs showing accuracy and F-score for POS-Tagging, Bio-NER, Word-Seg, and Act-Recog tasks using CRF and Perc models.]
Experiments-2: Learning speed

- Also with faster speed
  (a by-product of simpler structures)
Question: Is structure complexity matters in structured prediction?

Theoretical analysis to the question

1) Yes it matters  
2) High complexity of structures $\rightarrow$ high overfitting risk 
3) Low complexity $\rightarrow$ high empirical risk 
4) We need to find an optimal complexity of structures

Proposed a solution

- Split the original structure to find the optimal complexity
- Better accuracies in real tasks, & faster (a by-product)

This work is published at NIPS 2014:
Thanks for your attention!

Plz email xusun@pku.edu.cn if any question.

Source code is available upon request.