Structure Regularization for Structured Prediction

Xu SUN

xusun@pku.edu.cn

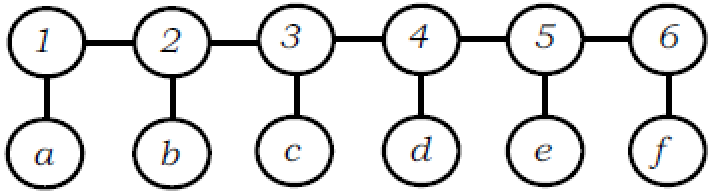
Peking University

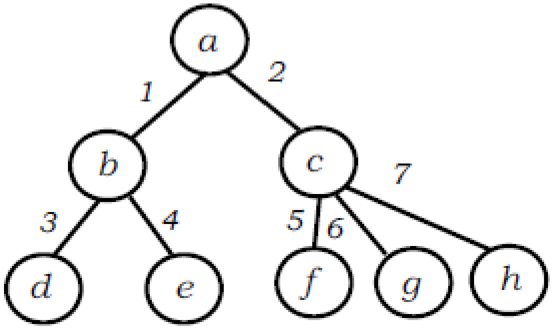
Structured prediction

- **Structured prediction methods are useful for many areas**
  - Natural language processing (NLP)
  - Vision recognition
  - Signal processing
  - Bioinformatics
  - Speech recognition
  - Etc.

Structured prediction

- For example, many natural language processing (NLP) tasks are **structured prediction** tasks
  - Parsing
  - SMT
  - POS tagging
  - Word segmentation
  - Named entity recognition
  - Chunking





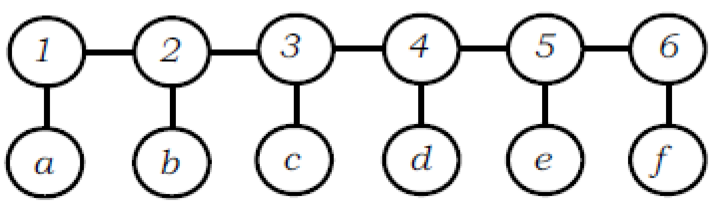
Basic question

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

He	PRP
reckons	VBZ
the	DT
current	JJ
account	NN
deficit	NN
will	MD
narrow	VB
to	TO
only	RB
#	#
1.8	CD
billion	CD
in	IN
September	NNP
.	.

Or, how about this scale?

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



Basic question

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

He	PRP
reckons	VBZ
the	DT
current	JJ
account	NN
deficit	NN
will	MD
narrow	VB
to	TO
only	RB
#	#
1.8	CD
billion	CD
in	IN
September	NNP
.	.

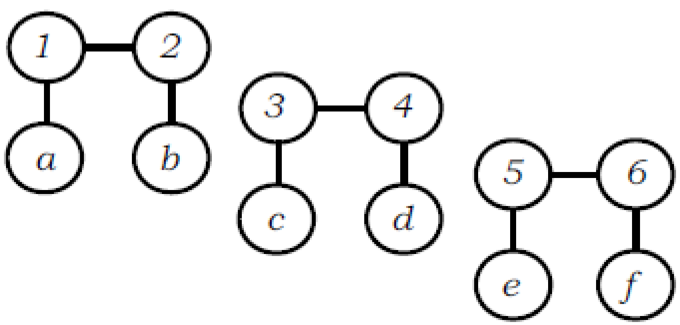
Basic question

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

He	PRP
reckons	VBZ
the	DT
current	JJ
account	NN
deficit	NN
will	MD
narrow	VB
to	TO
only	RB
#	#
1.8	CD
billion	CD
in	IN
September	NNP
.	.

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?

Basic question

- **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 λ, and the penalized function has a minimizer f:*

$$f = \operatorname*{argmin}_{g \in \mathcal{F}} R_{\alpha,\lambda}(g) = \operatorname*{argmin}_{g \in \mathcal{F}} \Big(\frac{1}{mn} \sum_{j=1}^{m\alpha} \mathcal{L}_\tau(g, \boldsymbol{z}'_j) + \frac{\lambda}{2} ||g||_2^2 \Big) \tag{8}$$

Assume the point-wise loss ℓ_τ is convex and differentiable, and is bounded by $\ell_\tau(f, \boldsymbol{z}, k) \leq \gamma$. Assume $f(\boldsymbol{x}, k)$ is ρ-admissible. Let a local feature value be bounded by v such that $\boldsymbol{x}_{(k,q)} \leq v$ 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

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

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 λ, and the penalized function has a minimizer f:*

$$f = \operatorname*{argmin}_{g \in \mathcal{F}} R_{\alpha,\lambda}(g) = \operatorname*{argmin}_{g \in \mathcal{F}} \Big(\frac{1}{mn} \sum_{j=1}^{m\alpha} \mathcal{L}_\tau(g, \boldsymbol{z}'_j) + \frac{\lambda}{2} ||g||_2^2 \Big) \tag{8}$$

Assume the point-wise loss ℓ_τ is convex and differentiable, and is bounded by $\ell_\tau(f, \boldsymbol{z}, k) \leq \gamma$. Assume $f(\boldsymbol{x}, k)$ is ρ-admissible. Let a local feature value be bounded by v such that $\boldsymbol{x}_{(k,q)} \leq v$ 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

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

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

→ 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 λ, and the penalized function has a minimizer f:*

$$f = \underset{g \in \mathcal{F}}{\operatorname{argmin}} R_{\alpha,\lambda}(g) = \underset{g \in \mathcal{F}}{\operatorname{argmin}} \Big(\frac{1}{mn} \sum_{j=1}^{m\alpha} \mathcal{L}_\tau(g, \boldsymbol{z}'_j) + \frac{\lambda}{2} ||g||_2^2 \Big) \quad (8)$$

Assume the point-wise loss ℓ_τ is convex and differentiable, and is bounded by $\ell_\tau(f, \boldsymbol{z}, k) \leq \gamma$. Assume $f(\boldsymbol{x}, k)$ is ρ-admissible. Let a local feature value be bounded by v such that $\boldsymbol{x}_{(k,q)} \leq v$ 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

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

Strength of structure regularization (strength of decomposition)

→ 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 λ, and the penalized function has a minimizer f:*

$$f = \operatorname*{argmin}_{g \in \mathcal{F}} R_{\alpha,\lambda}(g) = \operatorname*{argmin}_{g \in \mathcal{F}} \Big(\frac{1}{mn} \sum_{j=1}^{m\alpha} \mathcal{L}_\tau(g, \boldsymbol{z}'_j) + \frac{\lambda}{2} ||g||_2^2 \Big) \tag{8}$$

Assume the point-wise loss ℓ_τ is convex and differentiable, and is bounded by $\ell_\tau(f, \boldsymbol{z}, k) \leq \gamma$. Assume $f(\boldsymbol{x}, k)$ is ρ-admissible. Let a local feature value be bounded by v such that $\boldsymbol{x}_{(k,q)} \leq v$ 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

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

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 λ, and the penalized function has a minimizer f:*

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

Assume the point-wise loss ℓ_τ is convex and differentiable, and is bounded by $\ell_\tau(f, \boldsymbol{z}, k) \leq \gamma$. Assume $f(\boldsymbol{x}, k)$ is ρ-admissible. Let a local feature value be bounded by v such that $\boldsymbol{x}_{(k,q)} \leq v$ 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

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

✓Conclusions from our analysis:

1. Complex structure → low empirical risk & high overfitting risk
2. Simple structure → 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 λ, and the penalized function has a minimizer f:*

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

Assume the point-wise loss ℓ_τ is convex and differentiable, and is bounded by $\ell_\tau(f, \boldsymbol{z}, k) \leq \gamma$. Assume $f(\boldsymbol{x}, k)$ is ρ-admissible. Let a local feature value be bounded by v such that $\boldsymbol{x}_{(k,q)} \leq v$ 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

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

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

  Proper structure → good accuracy on training + not easy to overfit → 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 λ, and the penalized function has a minimizer f:*

$$f = \underset{g \in \mathcal{F}}{\operatorname{argmin}} R_{\alpha,\lambda}(g) = \underset{g \in \mathcal{F}}{\operatorname{argmin}} \Big(\frac{1}{mn} \sum_{j=1}^{m\alpha} \mathcal{L}_\tau(g, \boldsymbol{z}'_j) + \frac{\lambda}{2} ||g||_2^2 \Big) \tag{8}$$

Assume the point-wise loss ℓ_τ is convex and differentiable, and is bounded by $\ell_\tau(f, \boldsymbol{z}, k) \leq \gamma$. Assume $f(\boldsymbol{x}, k)$ is ρ-admissible. Let a local feature value be bounded by v such that $\boldsymbol{x}_{(k,q)} \leq v$ 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

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

1. Simple structure → low overfitting risk & high empirical risk
2. Complex structure → 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 a integer satisfying*

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

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 $\boldsymbol{w}_0$ and the minimizer $\boldsymbol{w}^*$, i.e., $a_0 = ||\boldsymbol{w}_0 - \boldsymbol{w}^*||^2$. Then, after t updates of $\boldsymbol{w}$ it converges to $\mathbb{E}[g(\boldsymbol{w}_t) - g(\boldsymbol{w}^*)] \leq \epsilon$.

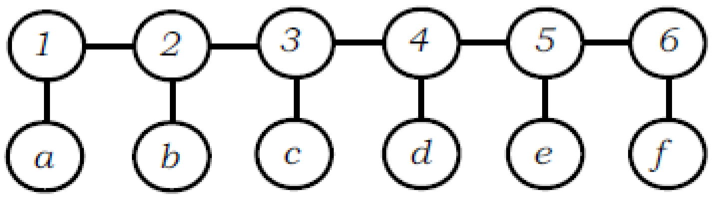
▫ SR also with faster speed

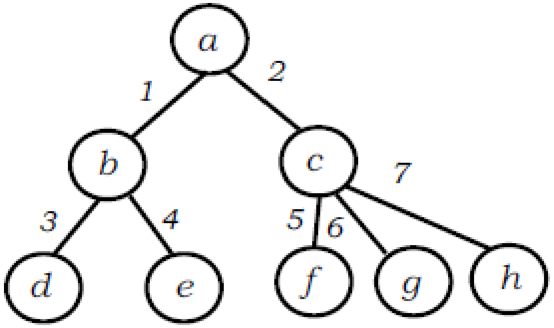
(a by-product of simpler structures)

✓ **using structure regularization can quadratically accelerate the convergence rate**

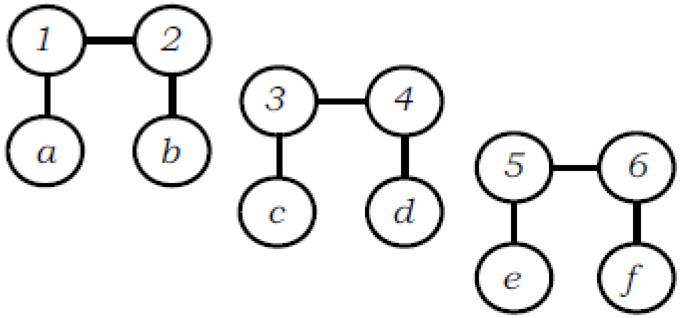
Illustration

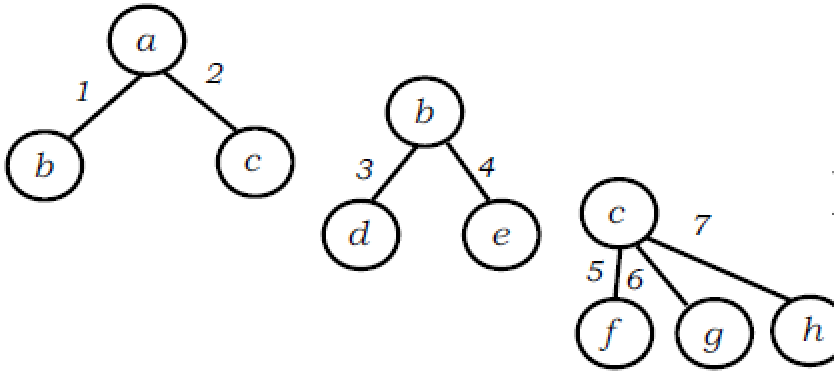
- Complex structures (high complexity)





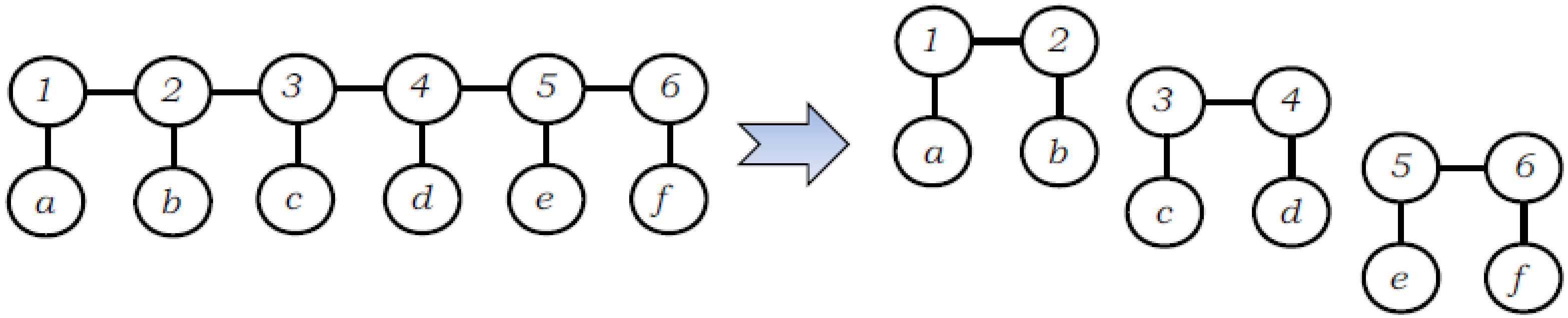
- Simple structures (low complexity)

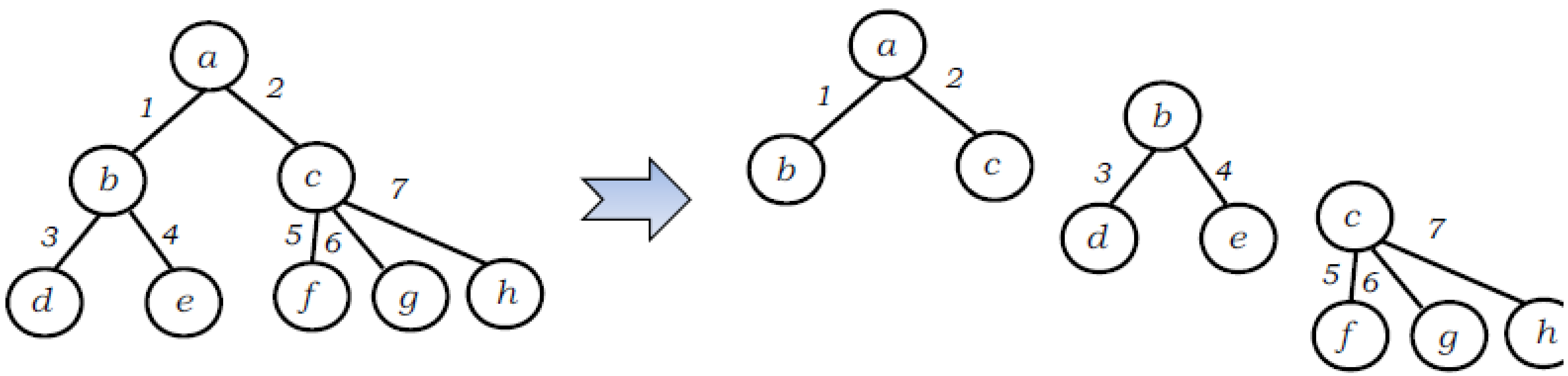




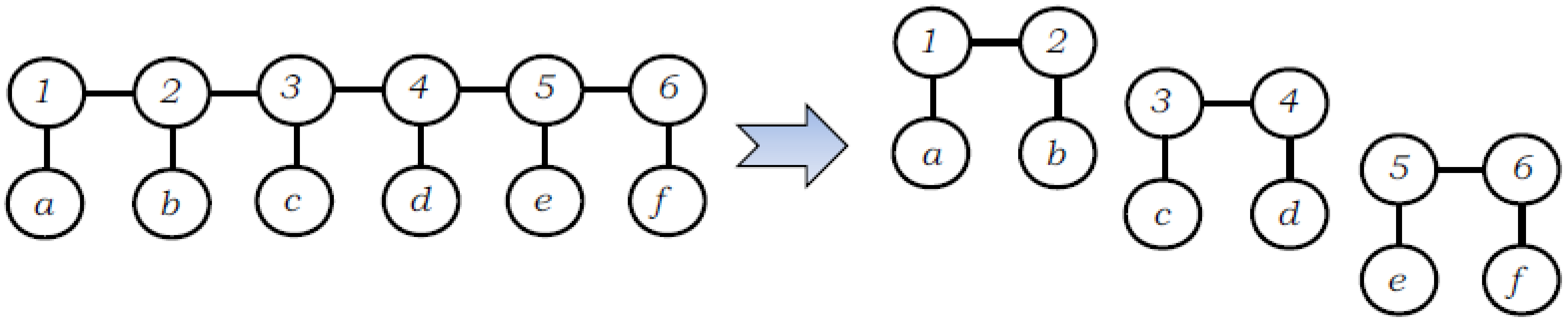
Structure regularization

- **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





Structure regularization

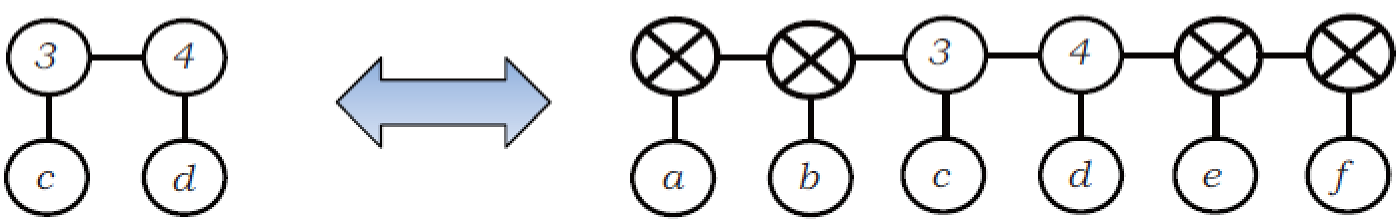


- **Will the split causes feature loss? – loss of long distance features?**

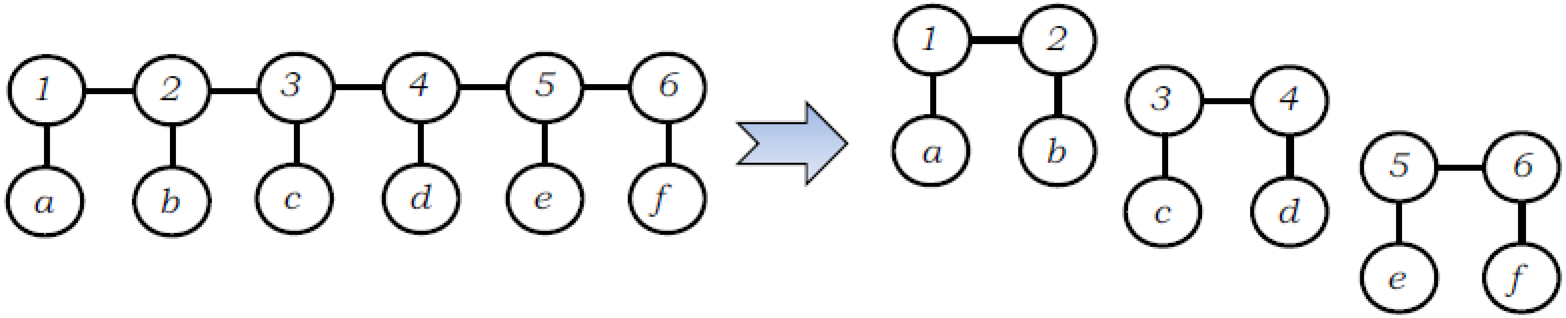
No loss of any (long distance) features

→ 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:



Structure regularization



- 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 & weight regularization

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

Algorithm 1 Training with structure regularization

Input: model weights $\boldsymbol{w}$, training set S, structure regularization strength α
repeat
  $S' \leftarrow \emptyset$
  for $i = 1 \to m$ **do**
    Randomly decompose $\boldsymbol{z}_i \in S$ into mini-samples $N_\alpha(\boldsymbol{z}_i) = \{\boldsymbol{z}_{(i,1)}, \ldots, \boldsymbol{z}_{(i,\alpha)}\}$
    $S' \leftarrow S' \cup N_\alpha(\boldsymbol{z}_i)$
  end for
  for $i = 1 \to |S'|$ **do**
    Sample $\boldsymbol{z}'$ uniformly at random from S', with gradient $\nabla g_{\boldsymbol{z}'}(\boldsymbol{w})$
    $\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \nabla g_{\boldsymbol{z}'}(\boldsymbol{w})$
  end for
until Convergence
return $\boldsymbol{w}$

The implementation is very simpl

Some advantages

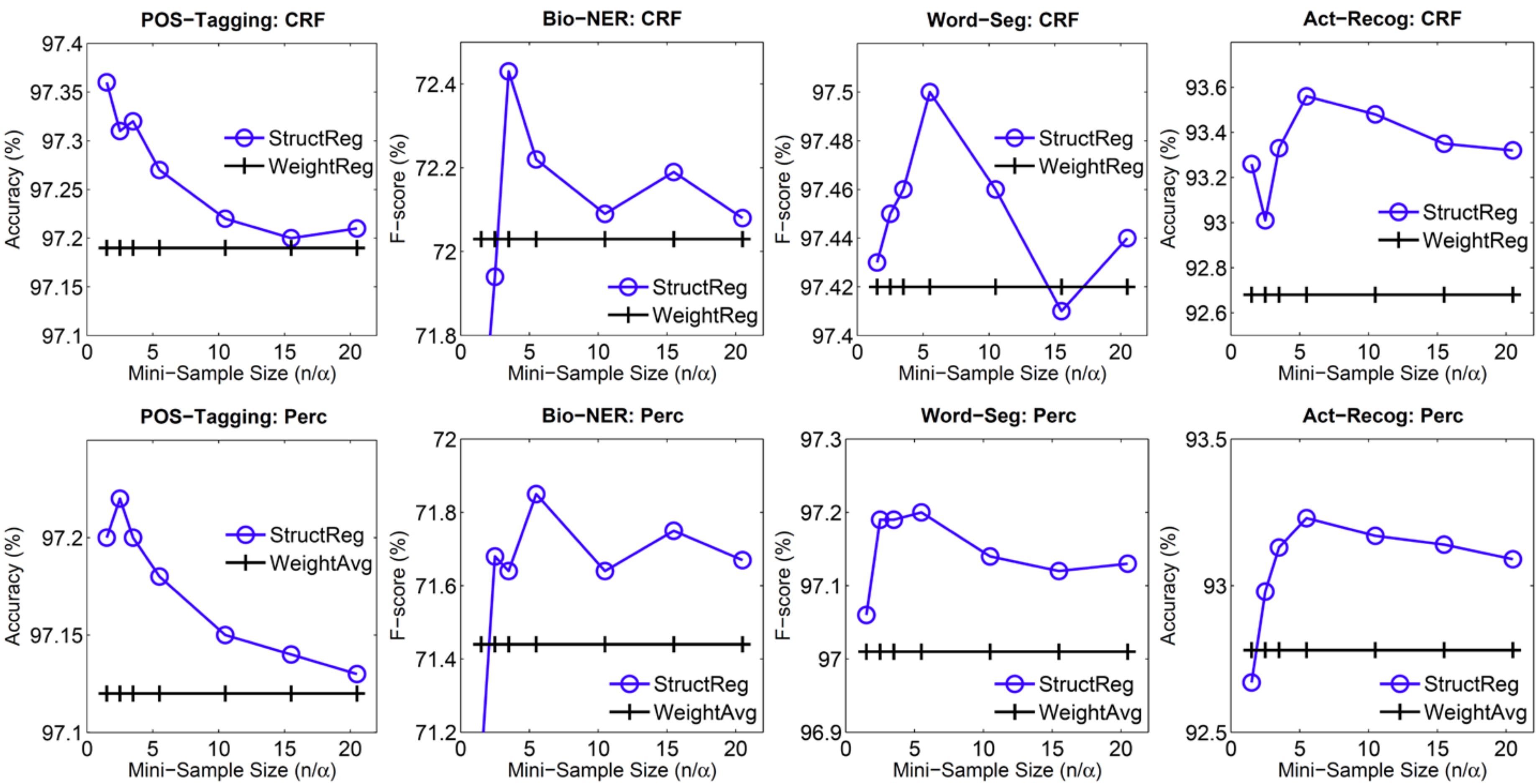
- If the original obj. function is convex, can still keep the convexity of the objective function

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

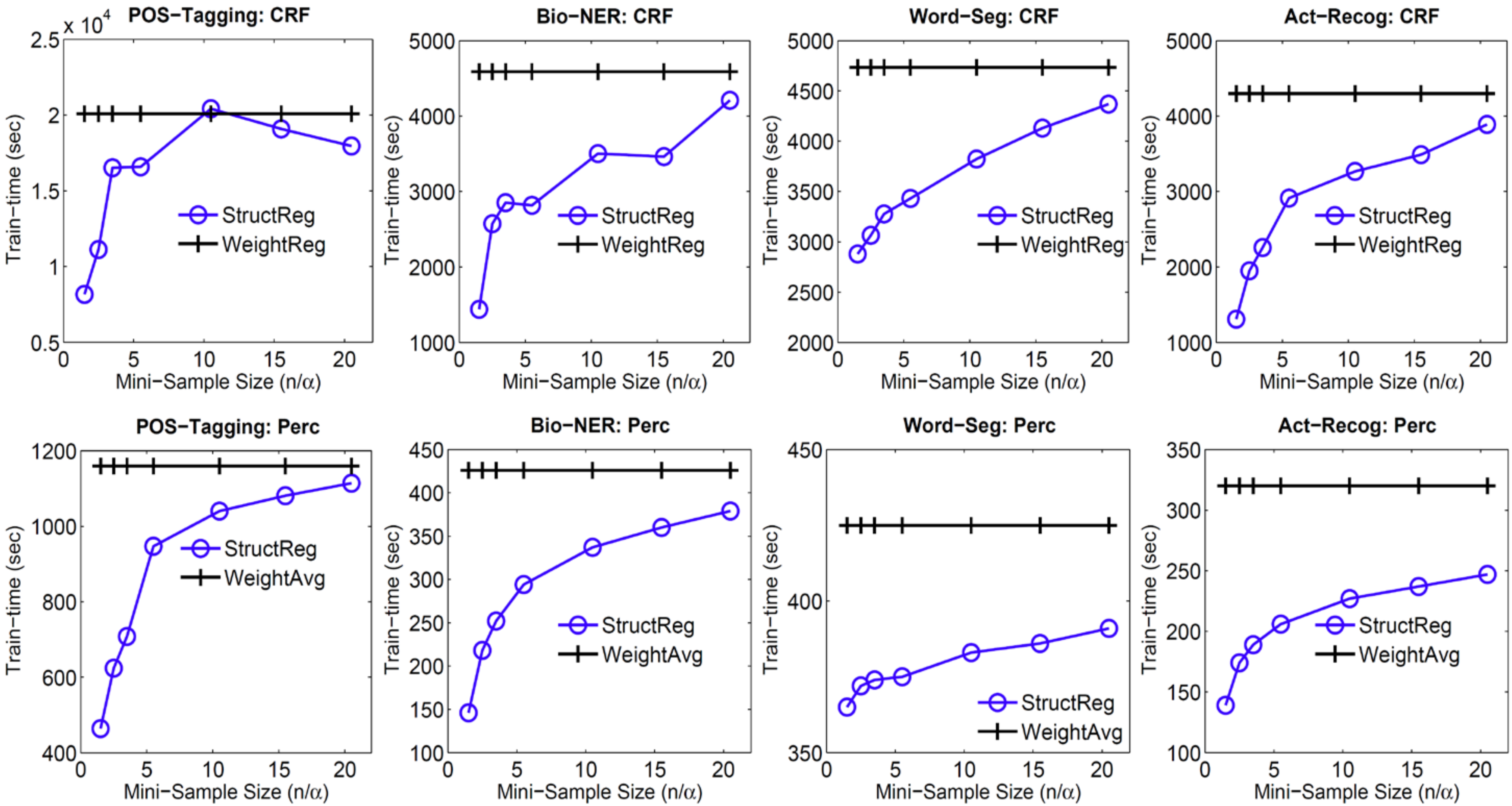
- General purpose and model-independent (because act like a preprocessing step)
  - E.g., can be used for different types of models, including CRFs, perceptrons, & neural networks

Experiments-1: accuracy

State-of-the-art scores on competitive tasks



Experiments-2 : Learning speed



- **Also with faster speed**

(a by-product of simpler structures)

Conclusions

- Question: Is structure complexity matters in structured prediction?
  - Theoretical analysis to the question
    - 1) Yes it matters
    - 2) High complexity of structures → high overfitting risk
    - 3) Low complexity → 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:

Xu Sun. Structure Regularization for Structured Prediction. In Advances in Neural Information Processing Systems (NIPS). 2402-2410. 2014

Thanks for your attention！

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

Source code is available upon request.