Inexact Search is “Good Enough”

Advanced Machine Learning for NLP
Jordan Boyd-Graber
MATHEMATICAL TREATMENT
Preliminaries: algorithm, separability

- Structured perceptron maintains set of “wrong features”
  \[ \Delta \Phi(x, y, z) \equiv \Phi(x, y) - \Phi(x, z) \]  
  (1)

- Structured perceptron updates weights with
  \[ \tilde{w} \leftarrow \tilde{w} + \Delta \Phi(x, y, z) \]  
  (2)

- Dataset \( D \) is linearly separable under features \( \Phi \) with margin \( \delta \) if
  \[ \tilde{u} \cdot \Delta \Phi(x, y, z) \geq \delta \quad \forall x, y, z \in D \]  
  (3)

  given some oracle unit vector \( u \).
Violations vs. Errors

• It may be difficult to find the highest scoring hypothesis
• It’s okay as long as inference finds a violation

\[ \vec{w} \cdot \Delta \Phi(x, y, z) \leq 0 \]  \hspace{1cm} (4)

• This means that \( y \) might not be answer algorithm gives (i.e., wrong)
Limited number of mistakes

• Define diameter $R$ as

$$R = \max_{(x,y,z)} ||\Delta \Phi(x,y,z)||$$

(5)
Limited number of mistakes

- Define diameter $R$ as

$$ R = \max_{(x, y, z)} \| \Delta \Phi (x, y, z) \| $$

\hspace{1cm} (5)

- Weight vector $\hat{w}$ grows with each error
- We can prove that $\| \hat{w} \|$ can’t get too big
- And thus, algorithm can only run for limited number of iterations $k$ where it updates weights
- Indeed, we’ll bound it from two directions

$$ k^2 \delta^2 \leq \| w^{(k+1)} \|^2 \leq kR^2 $$

\hspace{1cm} (6)
Lower Bound

\[ k^2 \delta^2 \leq \| w^{(k+1)} \|^2 \]
Lower Bound

\[ k^2 \delta^2 \leq \| w^{(k+1)} \|^2 \]

\[ \tilde{w}^{(k+1)} = w^{(k)} + \Delta \tilde{\Phi}(x, y, z) \]

Update equation
Lower Bound

\[ k^2 \delta^2 \leq \| w^{(k+1)} \|^2 \]

\[ \hat{w}^{(k+1)} = w^{(k)} + \Delta \hat{\Phi}(x, y, z) \quad (7) \]

\[ \hat{u} \cdot \hat{w}^{(k+1)} = \hat{u} \cdot w^{(k)} + \hat{u} \cdot \Delta \hat{\Phi}(x, y, z) \quad (8) \]

Multiply both sides by \( \hat{u} \)
Lower Bound

\[ k^2 \delta^2 \leq \| w^{(k+1)} \|^2 \]

\[ \hat{w}^{(k+1)} = w^{(k)} + \Delta \Phi(x, y, z) \tag{7} \]

\[ \hat{u} \cdot \hat{w}^{(k+1)} = \hat{u} \cdot w^{(k)} + \Delta \Phi(x, y, z) \tag{8} \]

\[ \hat{u} \cdot \hat{w}^{(k+1)} \geq \hat{u} \cdot w^{(k)} + \delta \tag{9} \]

Definition of margin
Lower Bound

\[ k^2 \delta^2 \leq \|w^{(k+1)}\|^2 \]

\[ \tilde{w}^{(k+1)} = w^{(k)} + \Delta \tilde{\Phi}(x, y, z) \quad (7) \]

\[ \tilde{u} \cdot \tilde{w}^{(k+1)} = \tilde{u} \cdot w^{(k)} + \tilde{u} \cdot \Delta \tilde{\Phi}(x, y, z) \quad (8) \]

\[ \tilde{u} \cdot \tilde{w}^{(k+1)} \geq \tilde{u} \cdot w^{(k)} + \delta \quad (9) \]

By induction, \( \tilde{u} \cdot \tilde{w}^{(k+1)} \geq k \delta \) (Base case: \( \tilde{w}^0 = \tilde{0} \))
Lower Bound

\[ k^2 \delta^2 \leq \| \mathbf{w}^{(k+1)} \|^2 \]

\[ \hat{u} \cdot \hat{w}^{(k+1)} \geq \hat{u} \cdot w^{(k)} + \delta \]  \hspace{1cm} (7)

By induction, \( \hat{u} \cdot \hat{w}^{(k+1)} \geq k \delta \) (Base case: \( \hat{w}^0 = \vec{0} \))

\[ \| \hat{u} \| \| \hat{w}^{(k+1)} \| \geq \hat{u} \cdot \hat{w} \geq k \delta \]  \hspace{1cm} (8)

For any vectors, \( \| \hat{a} \| \| \hat{b} \| \geq a \cdot b \)
Lower Bound

\[ k^2 \delta^2 \leq \| w^{(k+1)} \|^2 \]

\[ \tilde{u} \cdot \tilde{w}^{(k+1)} \geq \tilde{u} \cdot w^{(k)} + \delta \]  

(7)

By induction, \( \tilde{u} \cdot \tilde{w}^{(k+1)} \geq k\delta \) (Base case: \( \tilde{w}^0 = \tilde{0} \))

\[ \| \tilde{u} \| \| \tilde{w}^{(k+1)} \| \geq \tilde{u} \cdot \tilde{w} \geq k\delta \]  

(8)

\[ \| \tilde{w}^{(k+1)} \| \geq k\delta \]  

(9)

\( \tilde{u} \) is a unit vector
Lower Bound

\[ k^2 \delta^2 \leq \|w^{(k+1)}\|^2 \]

\[ \hat{u} \cdot \hat{w}^{(k+1)} \geq \hat{u} \cdot w^{(k)} + \delta \]  \hspace{1cm} (7)

By induction, \( \hat{u} \cdot \hat{w}^{(k+1)} \geq k\delta \) (Base case: \( \hat{w}^0 = \vec{0} \))

\[ \|\hat{u}\| \|\hat{w}^{(k+1)}\| \geq \hat{u} \cdot \hat{w} \geq k\delta \]  \hspace{1cm} (8)

\[ \|\hat{w}^{(k+1)}\| \geq k\delta \]  \hspace{1cm} (9)

\[ \|\hat{w}^{(k+1)}\|^2 \geq k^2 \delta^2 \]  \hspace{1cm} (10)

Square both sides, and we’re done!
Upper Bound

\[ \| \hat{w}^{(k+1)} \|^2 \leq kR^2 \]  \hspace{1cm} (11)
Upper Bound

\[ \| \tilde{w}^{(k+1)} \|^2 \leq kR^2 \] (11)

\[ \| \tilde{w}^{(k+1)} \|^2 = \| \tilde{w}^{(k)} + \Delta \tilde{\Phi}(x, y, z) \|^2 \] (12)

Update rule
Upper Bound

\[ \| \tilde{w}^{(k+1)} \|^2 \leq kR^2 \quad (11) \]

\[ \| \tilde{w}^{(k+1)} \|^2 = \| \tilde{w}^{(k)} + \Delta \tilde{\Phi}(x, y, z) \|^2 \quad (12) \]

\[ \| \tilde{w}^{(k+1)} \|^2 = \| \tilde{w}^{(k)} \|^2 + \| \Delta \tilde{\Phi}(x, y, z) \|^2 + 2 \tilde{w}^{(k)} \cdot \Delta \tilde{\Phi}(x, y, z) \quad (13) \]

Law of cosines
Upper Bound

\[ \| \hat{\mathbf{w}}^{(k+1)} \|^2 \leq kR^2 \] (11)

\[ \| \hat{\mathbf{w}}^{(k+1)} \|^2 = \| \hat{\mathbf{w}}^{(k)} + \Delta \Phi(x, y, z) \|^2 \] (12)

\[ \| \hat{\mathbf{w}}^{(k+1)} \|^2 = \| \hat{\mathbf{w}}^{(k)} \|^2 + \| \Delta \Phi(x, y, z) \|^2 + 2 \mathbf{w}^{(k)} \cdot \Delta \Phi(x, y, z) \] (13)

\[ \| \hat{\mathbf{w}}^{(k+1)} \|^2 \leq \| \hat{\mathbf{w}}^{(k)} \|^2 + R^2 + 2 \mathbf{w}^{(k)} \cdot \Delta \Phi(x, y, z) \] (14)

Definition of diameter
Upper Bound

\[ \|\mathbf{\hat{w}}^{(k+1)}\|^2 \leq kR^2 \]  \hspace{1cm} (11)

\[ \|\mathbf{\hat{w}}^{(k+1)}\|^2 = \|\mathbf{\hat{w}}^{(k)}\|^2 + \|\Delta \Phi(x, y, z)\|^2 + 2\mathbf{w}^{(k)} \cdot \Delta \Phi(x, y, z) \]  \hspace{1cm} (12)

\[ \|\mathbf{\hat{w}}^{(k+1)}\|^2 \leq \|\mathbf{\hat{w}}^{(k)}\|^2 + R^2 + 2\mathbf{w}^{(k)} \cdot \Delta \Phi(x, y, z) \]  \hspace{1cm} (13)

\[ \|\mathbf{\hat{w}}^{(k+1)}\|^2 \leq \|\mathbf{\hat{w}}^{(k)}\|^2 + R^2 + 0 \]  \hspace{1cm} (14)

\[ \|\mathbf{\hat{w}}^{(k+1)}\|^2 \leq \|\mathbf{\hat{w}}^{(k)}\|^2 + R^2 + 0 \]  \hspace{1cm} (15)

If violation
Upper Bound

\[ \| \mathbf{\hat{w}}(k+1) \| ^2 \leq kR^2 \]  \hspace{1cm} (11)

\[ \| \mathbf{\hat{w}}(k+1) \| ^2 = \| \mathbf{\hat{w}}(k) + \Delta \hat{\Phi}(x, y, z) \| ^2 \]  \hspace{1cm} (12)

\[ \| \mathbf{\hat{w}}(k+1) \| ^2 = \| \mathbf{\hat{w}}(k) \| ^2 + \| \Delta \hat{\Phi}(x, y, z) \| ^2 + 2w(k) \cdot \Delta \hat{\Phi}(x, y, z) \]  \hspace{1cm} (13)

\[ \| \mathbf{\hat{w}}(k+1) \| ^2 \leq \| \mathbf{\hat{w}}(k) \| ^2 + R^2 + 2w(k) \cdot \Delta \hat{\Phi}(x, y, z) \]  \hspace{1cm} (14)

\[ \| \mathbf{\hat{w}}(k+1) \| ^2 \leq \| \mathbf{\hat{w}}(k) \| ^2 + R^2 + 0 \]  \hspace{1cm} (15)

\[ \| \mathbf{\hat{w}}(k+1) \| ^2 \leq kR^2 \]  \hspace{1cm} (16)

Induction!
Putting it together

- Sandwich:

\[ k^2 \delta^2 \leq \|w^{(k+1)}\|^2 \leq kR^2 \]  \hspace{1cm} (17)

- Solve for \( k \):

\[ k \leq \frac{R^2}{\delta^2} \]  \hspace{1cm} (18)

- What does this mean?

  - Limited number of errors (updates)
    - Larger diameter increases errors (worst possible mistake)
    - Larger margin decreases errors (bigger separation from wrong answer)
  
  - Finding the largest violation wrong answer is best (but any violation okay)
Putting it together

- Sandwich:
  \[ k^2 \delta^2 \leq \|w^{(k+1)}\|^2 \leq kR^2 \]  
  (17)

- Solve for \( k \):
  \[ k \leq \frac{R^2}{\delta^2} \]  
  (18)
Putting it together

- Sandwich:
  \[ k^2 \delta^2 \leq \|w^{(k+1)}\|^2 \leq kR^2 \]  \hspace{1cm} (17)

- Solve for \( k \):
  \[ k \leq \frac{R^2}{\delta^2} \]  \hspace{1cm} (18)

- What does this mean?

  - Limited number of errors (updates)
    - Larger diameter increases errors (worst possible mistake)
    - Larger margin decreases errors (bigger separation from wrong answer)
  
  - Finding the largest violation wrong answer is best (but any violation okay)
Putting it together

• Sandwich:

\[ k^2 \delta^2 \leq ||w^{(k+1)}||^2 \leq kR^2 \]  

(17)

• Solve for \( k \):

\[ k \leq \frac{R^2}{\delta^2} \]  

(18)

• What does this mean?

• Limited number of errors (updates)
  ▪ Larger diameter increases errors (worst possible mistake)
  ▪ Larger margin decreases errors (bigger separation from wrong answer)

• Finding the largest violation wrong answer is best (but any violation okay)
In Practice

Harder the search space, the more max violation helps

![Graphs showing performance across different scenarios.

- Tagging accuracy (`b=1`):
  - Increase with training time.

- Incremental parsing accuracy (`b=8`):
  - Initial increase, stabilization over time.

- Unlabeled Attachment Score (UAS) for parsing:
  - Shows similar trends across different methods.

- BLEU score for machine translation:
  - Max-violation method consistently outperforms others.

Boyd-Graber