Structured Prediction

Machine Learning: Jordan Boyd-Graber
University of Maryland

INEXACT SEARCH IS “GOOD ENOUGH”
Preliminaries: algorithm, separability

- Structured perceptron maintains set of “wrong features”

\[ \Delta \Phi(x, y, z) \equiv \Phi(x, y) - \Phi(x, z) \quad (1) \]

- Structured perceptron updates weights with

\[ \tilde{w} \leftarrow \tilde{w} + \Delta \Phi(x, y, z) \quad (2) \]

- Dataset $D$ is linearly separable under features $\Phi$ with margin $\delta$ if

\[ \tilde{u} \cdot \Delta \Phi(x, y, z) \geq \delta \quad \forall x, y, z \in D \quad (3) \]

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

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

(4)

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

- Define diameter $R$ as

$$R = \max_{(x,y,z)} ||\Delta \tilde{\phi}(x,y,z)||$$  \hspace{1cm} (5)
Limited number of mistakes

- Define diameter $R$ as

$$R = \max_{(x,y,z)} \| \Delta \tilde{\Phi}(x, y, z) \|$$  \hspace{1cm} (5)

- Weight vector $\tilde{w}$ grows with each error
- We can prove that $\| \tilde{w} \|$ 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

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

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

Update equation
Lower Bound

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

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

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

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

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

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

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

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

Definition of margin
Lower Bound

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

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

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

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

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

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

\[ \tilde{u} \cdot \tilde{w}^{(k+1)} \geq \tilde{u} \cdot w^{(k)} + \delta \]  \hfill (7)

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

\[ ||\tilde{u}|| ||\tilde{w}^{(k+1)}|| \geq \tilde{u} \cdot \tilde{w} \geq k\delta \]  \hfill (8)

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

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

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

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

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

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

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

(12)
Upper Bound

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

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

Update rule
Upper Bound

Upper Bound

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

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

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

Law of cosines
Upper Bound

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

\[
\begin{align*}
\| \hat{w}^{(k+1)} \|^2 &= \| \hat{w}^{(k)} + \Delta \Phi(x, y, z) \|^2 \\
\| \hat{w}^{(k+1)} \|^2 &= \| \hat{w}^{(k)} \|^2 + \| \Delta \Phi(x, y, z) \|^2 + 2w^{(k)} \cdot \Delta \Phi(x, y, z)
\end{align*}
\] (12) (13)

Definition of diameter
Upper Bound

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

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

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

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

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

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

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

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

If violation, \( z \) is highest scoring candidate (so must be negative)
Upper Bound

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Induction!
Putting it together

- Sandwich:

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

- \[ \text{Solve for } k \text{:} \]

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

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

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

Harder the search space, the more max violation helps

- Tagging: $b=1$
- Incremental parsing: $b=8$
- Bottom-up parsing
- Machine translation