Generalization Error Bounds for Learning to Rank

Ambuj Tewari and Sougata Chaudhuri

Department of Statistics, and
Department of EECS,
University of Michigan, Ann Arbor

June 28, 2015
Learning to Rank at Query Level

Input space - Lists of \( m \) documents pertaining to queries. Formally: \( X \in \mathbb{R}^{m \times d} \)

Supervision space - Relevance vectors of length \( m \). Formally: \( Y \in \{0, 1, \ldots, K\}^m \)

Rank documents by sorting scores corresponding to a scoring function. For \( X \in X \), linear scoring function:

\[
    f_w(X) = Xw \in \mathbb{R}^m
\]
Input space - Lists of $m$ documents pertaining to queries.

Formally: $\mathcal{X} \in \mathbb{R}^{m \times d}$
Learning to Rank at Query Level

- Input space - Lists of \( m \) documents pertaining to queries.
  - Formally: \( X \in \mathbb{R}^{m \times d} \)

- Supervision space - Relevance vectors of length \( m \).
  - Formally: \( Y \in \{0, 1, \ldots, K\}^m \).

Rank documents by sorting scores corresponding to a scoring function.
For \( X \in X \), linear scoring function:
\[
 f_w(X) = Xw \in \mathbb{R}^m.
\]
Input space- Lists of $m$ documents pertaining to queries.
- Formally: $\mathcal{X} \in \mathbb{R}^{m \times d}$

Supervision space- Relevance vectors of length $m$.
- Formally: $\mathcal{Y} \in \{0, 1, \ldots, K\}^m$.

Rank documents by sorting scores corresponding to a scoring function.
Input space- Lists of $m$ documents pertaining to queries.
- Formally: $X \in \mathbb{R}^{m \times d}$

Supervision space- Relevance vectors of length $m$.
- Formally: $Y \in \{0, 1, \ldots, K\}^m$.

Rank documents by sorting scores corresponding to a scoring function.

For $X \in \mathcal{X}$, linear scoring function: $f_w(X) = Xw \in \mathbb{R}^m$. 

Ambuj Tewari and Sougata Chaudhuri
ICML 2015
Scoring function learnt from training data.

Training data: \((X_i, R_i) \overset{i.i.d.}{\sim} \mathcal{D}(\mathcal{X} \times \mathcal{Y})\)

Performance of function judged by target measures like NDCG, AP.

Computationally difficult to optimize the measures during training time.

Hence, development of a number of ranking surrogates.
Ranking Surrogates

Scoring function learnt from training data.

Training data: \((X_i, R_i) \sim \mathbb{D}(X \times Y)\)

Performance of function judged by target measures like NDCG, AP.
Scoring function learnt from training data.

- Training data: \((X_i, R_i) \sim \mathcal{D}(\mathcal{X} \times \mathcal{Y})\)

- Performance of function judged by target measures like NDCG, AP.

- Computationally difficult to optimize the measures during training time.
Scoring function learnt from training data.

- Training data: \((X_i, R_i) \sim_i d \mathcal{D}(\mathcal{X} \times \mathcal{Y})\)

- Performance of function judged by target measures like NDCG, AP.

- Computationally difficult to optimize the measures during training time.

- Hence, development of a number of ranking surrogates.
ERM algorithms learn a function by minimizing surrogate loss on training data.
ERM algorithms learn a function by minimizing surrogate loss on training data.

**Generalization Error**: What is the expected surrogate loss for the learnt function?
ERM algorithms learn a function by minimizing surrogate loss on training data.

**Generalization Error**: What is the expected surrogate loss for the learnt function?

**Calibration**: How does expected surrogate loss relate to expected *target measures* based losses?
ERM algorithms learn a function by minimizing surrogate loss on training data.

**Generalization Error**: What is the expected surrogate loss for the learnt function?

**Calibration**: How does expected surrogate loss relate to expected *target measures* based losses?

We address question on *generalization error*. 
Let $\phi$ be a ranking surrogate.

$\phi(s^w, R) \mapsto \mathbb{R}$, for $s^w = Xw \in \mathbb{R}^m$, $R \in \{0, \ldots, K\}^m$. 
Let $\phi$ be a ranking surrogate. 

$\phi(s^w, R) \mapsto \mathbb{R}$, for $s^w = Xw \in \mathbb{R}^m$, $R \in \{0, \ldots, K\}^m$.

Uniform generalization error (over parameter class $\mathcal{F}$):

$\mathbb{E}[\phi(s^w, R)] \leq \frac{1}{n} \sum_{i=1}^{n} \phi(s_i^w, R_i) + \text{Complexity}, \quad \forall \ w \in \mathcal{F}.$
Let $\phi$ be a ranking surrogate.

$\phi(s^w, R) \mapsto \mathbb{R}$, for $s^w = Xw \in \mathbb{R}^m$, $R \in \{0, \ldots, K\}^m$.

Uniform generalization error (over parameter class $\mathcal{F}$):

$$\mathbb{E}[\phi(s^w, R)] \leq \frac{1}{n} \sum_{i=1}^{n} \phi(s^w_i, R_i) + \text{Complexity}, \quad \forall w \in \mathcal{F}.$$  

Complexity term may depend on properties of $\phi$, sample size $n$, length of document list $m$ etc.
Since $\phi(s, R)$ is defined on vector valued predictions ($s \in \mathbb{R}^m$), Lipschitz property dependent on norm.
Since $\phi(s, R)$ is defined on vector valued predictions ($s \in \mathbb{R}^m$), Lipschitz property dependent on norm.

$\phi(s, R)$ is $L_2$ Lipschitz w.r.t $s$ in $\ell_2$ norm if
$$|\phi(s_1, R) - \phi(s_2, R)| \leq L_2 \|s_1 - s_2\|_2.$$
Since $\phi(s, R)$ is defined on vector valued predictions ($s \in \mathbb{R}^m$), Lipschitz property dependent on norm.

- $\phi(s, R)$ is $L_2$ Lipschitz w.r.t $s$ in $\ell_2$ norm if
  $$|\phi(s_1, R) - \phi(s_2, R)| \leq L_2 \|s_1 - s_2\|_2.$$

- $\phi(s, R)$ is $L_1$ Lipschitz w.r.t $s$ in $\ell_\infty$ norm if
  $$|\phi(s_1, R) - \phi(s_2, R)| \leq L_2 1 \|s_1 - s_2\|_\infty.$$
Since $\phi(s, R)$ is defined on vector valued predictions ($s \in \mathbb{R}^m$), Lipschitz property dependent on norm.

- $\phi(s, R)$ is $L_2$ Lipschitz w.r.t $s$ in $\ell_2$ norm if 
  \[ |\phi(s_1, R) - \phi(s_2, R)| \leq L_2 \|s_1 - s_2\|_2. \]

- $\phi(s, R)$ is $L_1$ Lipschitz w.r.t $s$ in $\ell_\infty$ norm if 
  \[ |\phi(s_1, R) - \phi(s_2, R)| \leq L_2 1 \|s_1 - s_2\|_\infty. \]

- $L_1 \leq \sqrt{m} L_2$. 

Ambuj Tewari and Sougata Chaudhuri

ICML 2015
Existing Result

Let $\|w\|_2 \leq W_2, RX$ be bound on $\ell_2$ norm of feature vectors.

Best known complexity for Lipschitz surrogates in $\ell_2$ norm:

$$O\left( L_2 W_2 R X \sqrt{mn} \right).$$

Proof technique was intrinsic to $\ell_2$ Lipschitz surrogates and necessitated $m$ dependence.
Let $\|w\|_2 \leq W_2, \ R_X$ be bound on $\ell_2$ norm of feature vectors.
Let $\|w\|_2 \leq W_2$, $R_X$ be bound on $\ell_2$ norm of feature vectors.

Best known complexity for Lipschitz surrogates in $\ell_2$ norm:

$O(L_2 W_2 R_X \sqrt{\frac{m}{n}})$.
Let $\|w\|_2 \leq W_2$, $R_X$ be bound on $\ell_2$ norm of feature vectors.

Best known complexity for Lipschitz surrogates in $\ell_2$ norm: $O(L_2 W_2 R_X \sqrt{\frac{m}{n}})$.

Proof technique was intrinsic to $\ell_2$ Lipschitz surrogates and necessitated $m$ dependence.
Complexity term depends on richness of class of scoring functions.
Necessary Dependence on $m$?

- Complexity term depends on richness of class of scoring functions.
- Linear scoring functions parameterized by $d$ dimensional vector ($w \in \mathbb{R}^d$), *independent of* $m$. 
Necessary Dependence on $m$?

- Complexity term depends on richness of class of scoring functions.
- Linear scoring functions parameterized by $d$ dimensional vector ($w \in \mathbb{R}^d$), *independent of* $m$.
- Should complexity term be independent of $m$?
- Complexity term depends on richness of class of scoring functions.
- Linear scoring functions parameterized by $d$ dimensional vector ($w \in \mathbb{R}^d$), independent of $m$.
- Should complexity term be independent of $m$?
- What role does Lipschitz norm play in complexity?
Our Contributions
Examples- ListNet and SmoothDCG

Listnet (convex) and SmoothDCG@1 (non-convex) are popular ranking surrogates. Both are $\ell_\infty$ Lipschitz with constants independent of $m$.

Previous generalization error bounds for both the surrogates had $m$ dependent complexity.
Listnet (convex) and SmoothDCG@1 (non-convex) are popular ranking surrogates.
Listnet (convex) and SmoothDCG@1 (non-convex) are popular ranking surrogates.

Both are $\ell_\infty$ Lipschitz with constants independent of $m$. 
Listnet (convex) and SmoothDCG@1 (non-convex) are popular ranking surrogates.

Both are $\ell_\infty$ Lipschitz with constants independent of $m$.

Previous generalization error bounds for both the surrogates had $m$ dependent complexity.
φ is $\ell_\infty$ Lipschitz (constant $L_1$), functional parameter $\|w\|_2 \leq W_2$, $R_X$ bound on $\ell_2$ norm of feature vectors.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Non-convexity</th>
<th>Complexity</th>
<th>Constants in $O(\cdot)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>OGD</td>
<td>No</td>
<td>$O(L_1 W_2 R_X \sqrt{\frac{1}{n}})$</td>
<td>smallest</td>
</tr>
<tr>
<td>RERM</td>
<td>No</td>
<td>$O(L_1 W_2 R_X \sqrt{\frac{1}{n}})$</td>
<td>small</td>
</tr>
<tr>
<td>ERM</td>
<td>Yes</td>
<td>$O(L_1 W_2 R_X \sqrt{\frac{1}{n}})$</td>
<td>several log factors</td>
</tr>
</tbody>
</table>
Analysis of Methods

OGD: Specific to convex surrogates and Online Gradient Descent algorithm.

RERM: Specific to convex surrogates, requires $\ell_2$ regularization function.

ERM: Applies to all Lipschitz surrogates, for all Empirical Risk Minimization algorithms.
OGD: Specific to convex surrogates and Online Gradient Descent algorithm.
Analysis of Methods

- **OGD**: Specific to convex surrogates and Online Gradient Descent algorithm.

- **RERM**: Specific to convex surrogates, requires $\ell_2$ regularization function.
OGD: Specific to convex surrogates and Online Gradient Descent algorithm.

RERM: Specific to convex surrogates, requires $\ell_2$ regularization function.

ERM: Applies to all Lipschitz surrogates, for all Empirical Risk Minimization algorithms.
Learning to rank problems can involve high dimensional features.
Learning to rank problems can involve high dimensional features.

Appropriately, let \( \{ w \in \mathbb{R}^d : \| w \|_1 \leq W \} \), \( \bar{R}_X \) be bound on \( \ell_\infty \) norm of feature vectors.
Learning to rank problems can involve high dimensional features.

Appropriately, let \( \{ w \in \mathbb{R}^d : \| w \|_1 \leq W_1 \} \), \( \bar{R}_X \) be bound on \( \ell_\infty \) norm of feature vectors.

Generalization error complexity: \( O(L_1 W_1 \bar{R}_X \sqrt{\frac{\log(d)}{n}}) \).
Learning to rank problems can involve high dimensional features.

Appropriately, let \( \{ w \in \mathbb{R}^d : \|w\|_1 \leq W_1 \} \), \( \bar{R}_X \) be bound on \( \ell_\infty \) norm of feature vectors.

Generalization error complexity: \( O(L_1 W_1 \bar{R}_X \sqrt{\frac{\log(d)}{n}}) \).

Complexity *nearly independent* of \( d \).
Let $\phi$ be a smooth surrogate w.r.t $\ell_\infty$ norm with constant $H_\phi$(definition in paper).
Let $\phi$ be a smooth surrogate w.r.t $\ell_\infty$ norm with constant $H_\phi$ (definition in paper).

Let $L_\phi (w^*) = \min_w \mathbb{E}[\phi(s^w, R)]$ and $C$ be constant depending on $H_\phi$. 

Generalization error complexity: $O(\sqrt{L_\phi (w^*) C n} + C n)$. 

Rate interpolates between $O(1/n)$ ($L_\phi (w^*) = 0$) and $O(\sqrt{1/n})$ ($L_\phi (w^*) > 0$).
Result for Smooth Surrogates

- Let $\phi$ be a smooth surrogate w.r.t $\ell_\infty$ norm with constant $H_\phi$ (definition in paper).
- Let $L_\phi(w^*) = \min_w \mathbb{E}[\phi(s^w, R)]$ and $C$ be constant depending on $H_\phi$.
- Generalization error complexity: $O(\sqrt{\frac{L_\phi(w^*)C}{n}} + \frac{C}{n})$. 

Ambuj Tewari and Sougata Chaudhuri
Let $\phi$ be a smooth surrogate w.r.t $\ell_\infty$ norm with constant $H_\phi$ (definition in paper).

Let $L_\phi(w^*) = \min_w \mathbb{E}[\phi(s^w, R)]$ and $C$ be constant depending on $H_\phi$.

Generalization error complexity: $O\left(\sqrt{\frac{L_\phi(w^*) C}{n}} + \frac{C}{n}\right)$.

Rate interpolates between $O\left(\frac{1}{n}\right)$ ($L_\phi(w^*) = 0$) and $O\left(\sqrt{\frac{1}{n}}\right)$ ($L_\phi(w^*) > 0$).