SELECTIVE SAMPLING WITH ALMOST OPTIMAL GUARANTEES FOR LEARNING TO RANK FROM PAIRWISE PREFERENCES

NIPS’11 Workshop
Nir Ailon, Ron Begleiter, Esther Ezra

Not to be confused with [Ailon, NIPS’11], which we improve
Agenda:

- Ranking from pairwise preferences
- Motivation
- State-of-the-art
- Near optimal selective sampling method
- Epsilon-smooth approximations
- Rank-SVM and other relaxations
To make the long story short -

Problem: Reducing the number of preferences :: Ranking from pairwise pref.

Idea: epsilon-smooth approximations :: powerful structural property

Problem+Idea: Near optimal. solution :: worst-case, info. theory, e.g., sorting
Consider a finite set of elements \( V \)

and a pairwise preference function \( W \) over \( V \times V \)

\[
W(u, v) = \begin{cases} 
1, & \text{if } u \text{ preferred over } v \\
0, & \text{when } v \text{ is preferred}
\end{cases}
\]

\[
W(u, v) + W(v, u) = 1
\]

Note: \( W \) is not a ranking. Can define preferences cycles: \( u < v < s < u \)
Given alternatives \( V \); and pairwise-preferences \( W \) find an order \( \pi \) over \( V \) with minimal disagreements with \( W \):

\[
d(\pi, W) = \frac{1}{2} \sum_{(u,v), u \neq v} \pi(u,v) W(v,u)
\]

Note: extends the well known Kendall-tau-distance
Ranking from pairwise preferences :: Toy example

\[ \pi(V) \text{ is} \]

\[ W \text{ is} \]

\[ W(V, \pi) = 0 \]

i.e. prefers over

Mind the preference cycle:

Least to most like ranking (left-right), with Kendall-tau cost = "2"

Can we have a "0"-cost rank??
Facts

When "W" is known this a.k.a. "Minimum Feedback Arc-Set in Tournaments" : It is NP-Hard [Alon'06]

There exists a Polynomial Time Approximation Scheme (PTAS) [Kenyon-Mathieu & Schudy'07]

The PTAS might need quadratic labels (know W)

Recent query efficient method by [Ailon'11]; Follows

[Kenyon-Mathieu & Schudy'07] Ideas. Our-work improves this by: (1) tighter bound; (2) applicable to Rank-SVM.

Next: -- We present a novel idea and analyze it --
epsilon-smooth approximations :: Notation

Consider

An instance space \( \mathcal{X} \) of pairs \((u, v) \in V \times V\)

A set \( \mathcal{F} \) of permutations -- func. from \( \mathcal{X} \) to \( \{0, 1\} \)

with \( \pi((u, v)) = 1 - \pi((v, u)) \) and \( \pi((u, w)) \leq \pi((u, v)) + \pi((v, w)) \)

A pseudo metric \( d(\pi, \sigma) = \frac{1}{2} \sum_{u, v \in V} 1_{\pi((u, v)) \neq \sigma((u, v))} \)

Goal - find \( \pi \) in \( \mathcal{F} \) with min. cost: \( \arg\min_{\mathcal{F}} d(\pi, W) \)
epsilon-smooth approximations :: Iterative algorithm

Algorithm 1.1 Finding an \((1 + O(\epsilon))\)-competitive estimation

**Input:** an initial solution \(\hat{\pi}_0 \in \mathcal{F}\), an estimation parameter \(\epsilon \in (0, 1/5)\), and number of iterations \(k\)

1. \(i \leftarrow 0\)
2. **repeat**
   1. \(\hat{\pi}_{i+1} \leftarrow \arg\min_{\sigma \in \mathcal{F}} \tilde{\Delta}_{\hat{\pi}_i}(\sigma)\), where \(\tilde{\Delta}_{\hat{\pi}_i}(\sigma)\) is an empirical risk of a “biased sample”
   2. \(i \leftarrow i + 1\)
3. **until** \(i\) equals \(k\)
4. **return** \(\hat{\sigma}_k\)

will be defined & clarified later
epsilon-smooth approximations :: Selective sampling

Assume we posses an arbitrary permutation $\pi$

Get a random sample $S \subseteq \mathcal{X}$ of $O(n \operatorname{poly} \log n, \varepsilon^{-1}))$ pairs

The *trick* :: Sample with bias

$$p_\pi(u, v) := \min \{1, \operatorname{poly} \log n, \varepsilon^{-1})/\delta_\pi(u, v)\}$$

$\delta_\pi(u, v)$ is the distance between position of $u$ and $v$ in the ordering induced by $\pi$

For example:: $\pi$ is $\{\text{A, B, C, D, E}\}$

$\delta_\pi(\{\text{A, B, C, D, E}\}) = 3$
Selective sampling :: The eps-smooth estimator

Define \( C_{u,v}(\sigma) = 1_{\sigma(u,v) \neq W(u,v)} \) the cost of a pair

The estimator is \( \hat{\Delta}_\pi(\sigma) = \frac{1}{2} \sum_{u,v \in S} p_\pi(u,v)^{-1} (C_{u,v}(\sigma) - C_{u,v}(\pi)) \)

Recall that \( d(\pi, \sigma) = \frac{1}{2} \sum_{u,v \in V} 1_{\pi(u,v) \neq \sigma(u,v)} \)

Lower weights to closer pairs
Relative regret w.r.t. single pair
The following turns out to be a powerful structural property:

\[ \pi \text{ is arbitrary} \quad \text{define:} \quad \Delta_{\pi}(\sigma) = d(\sigma, W) - d(\pi, W) \]

For \( \varepsilon < 1 \), \( \hat{\Delta}_{\pi}(\sigma) \) is \( \varepsilon \)-smooth approx. of \( \Delta_{\pi} \) if

for all \( \sigma \in \mathcal{F} \)

\[ |\hat{\Delta}_{\pi}(\sigma) - \Delta_{\pi}(\sigma)| \leq \varepsilon d(\pi, \sigma) \]
epsilon-smooth approximations :: Algorithm analysis

Algorithm 1.1 Finding an $(1 + O(\epsilon))$-competitive estimation

Input: an initial solution $\pi_0 \in \mathcal{F}$, an estimation parameter $\epsilon \in (0, 1/5)$, and number of iterations $k$

1. $i \leftarrow 0$
2. repeat
   1. $\hat{\pi}_{i+1} \leftarrow \text{argmin}_{\pi \in \mathcal{F}} \hat{\Delta}_{\pi, i}(\sigma)$, where $\hat{\Delta}_{\pi, i}(\sigma)$ is an $\epsilon$-smooth approximation of $\Delta_{\pi, i}$
   2. $i \leftarrow i + 1$
3. until $i$ equals $k$
4. return $\hat{\sigma}_k$

Reminder: Alg.

Theorem Let $\epsilon \in (0, 1/2)$. Let $\text{OPT}$ denote $\min_{\pi \in \mathcal{F}} d(\pi, W)$, and $\hat{\pi}_0$ be an arbitrary function. Then the following holds for $\hat{\pi}_i$ obtained in (previous slide) Algorithm for any $1 \leq i \leq k$:

$$d(\hat{\pi}_i, W) \leq (1 + O(\epsilon)) \text{OPT} + O(\epsilon^i)d(\hat{\pi}_0, W).$$

Proof idea: Using triangle inequality over $|\hat{\Delta}_{\pi}(\sigma) - \Delta_{\pi}(\sigma)| \leq \epsilon d(\pi, \sigma)$

along with simple manipulation we get $d(\hat{\pi}_k, W) \leq \left(\frac{1}{1 - \epsilon}\right)^2 + \left(\frac{1}{1 - \epsilon}\right)\sum_{i=1}^{k} 2\epsilon^i d(\hat{\pi}_{k-i}, W)$

sum of geometric series

$$\frac{2(1 - \epsilon^k)}{(1 - \epsilon)^2} \left(\text{OPT} + d(\hat{\pi}_0, W)\right)$$
\[ \hat{\Delta}_\pi(\sigma) = \frac{1}{2} \sum_{u,v \in V} p_\pi(u,v)^{-1} (C_{u,v}(\sigma) - C_{u,v}(\pi)) \] is eps-smooth

**Theorem**

With high probability \( \hat{\Delta}_\pi \) is eps-smooth estimator of \( \Delta_\pi \).

We compute \( \hat{\Delta}_\pi \) with \( O(n \text{ poly}(\log n, \varepsilon^{-1})) \) pairs (w.h.p.)

**Concluding**

Defined an Active-Learning algorithm achieving \( (1 + O(\varepsilon)) \) approx.

using only \( O(n \text{ poly}(\log n, \varepsilon^{-1})) \) pairs
SVM-Rank & Convex relaxations

Why?
Solving algo. step: \( \hat{\pi}_{i+1} \leftarrow \arg\min_{\sigma \in \mathcal{F}} \Delta \hat{\pi}_i (\sigma) \) is not clear

How?
use relaxed cost \( \tilde{d}(\pi, W) = \int_X \tilde{L} (\pi(u, v), W(u, v)) \, d\mu \)
where for all pairs \( L(a, b) \leq \tilde{L}(a, b) \)
Note: Similar theorem holds

Example: Subsampled SVM
Describing alternative \( u \) with feature vector \( u \in \mathbb{R}^d \)
and taking hinge-loss \( \tilde{L}(x, y) = \max\{0, 1 - xy\} \)
Conclusions & Future work

Done?
Near optimal query efficient & OPT-competitive solution
Worst case analysis; agnostic setting; improves recent [Ailon (NIPS'11)]

Future work:
Applying our idea to other settings:
For example - [Jamieson & Nowak (NIPS'11)] : avg.-case analysis, Embedding in $\mathbb{R}^d$
different hypothesis-class
From theory to practice: Empirical assessment
Applying eps-smooth on other classes of problems