Online and Stochastic Gradient Methods for Non-decomposable Loss Functions

Purushottam Kar*, Harikrishna Narasimhan†, and Prateek Jain*

*Microsoft Research, INDIA
†Indian Institute of Science, INDIA

Microsoft Research

The Goal

Scalable training algorithms for large scale optimization tasks with non-decomposable performance measures

Why "hinge loss" isn't Enough

Machine learning applications in sensitive domains:

- Medicine, biometrics, bioinformatics
- Essential: fine grained control over classification characteristics
  - Mild to severe label imbalance
  - Asymmetric misclassification penalties
- Pointwise performance measures (like hinge loss) deficient

Non-decomposable performance measures

The good news:

- Perform a holistic evaluation of classifiers over entire data
- Offer a high degree of control over prediction profile
  - Specific interest in top ranked results: prec@k
  - Sensitivity to FPR - partial AUC
  - Class imbalanced situations - F-measure

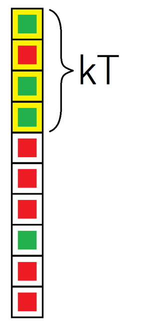
The not-so-good news:

- Frequently non-decomposable/non-additive
- Precludes application of a large body of work
  - Optimization theory - online, stochastic methods
  - Learning theory - OTB, generalization bounds
  - Prior work: mostly indirect/cutting plane-based *batch* methods

Examples of Non-decomposable Loss Functions

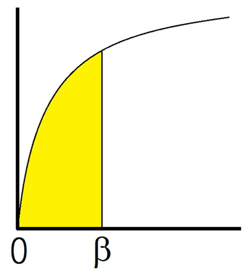
Data: $\mathbf{x}_1, \ldots, \mathbf{x}_T \in \mathbb{R}^d$, $\mathbf{y} = y_1 \ldots y_T \in \{-1, +1\}$
Surrogates: widely used *structSVM*-based formulations [Joachims05]

1. Prec@k: precision at the top k fraction of the ranked list



$$\ell_{\text{Prec}_{@k}}(\mathbf{w}) = \max_{\substack{\bar{\mathbf{y}} \in \{-1,+1\}^T \\ \sum_i \bar{y}_i = (2k-1)T}} \sum_i (\bar{y}_i - y_i) \cdot \mathbf{x}_i^\top \mathbf{w} - \sum_i \bar{y}_i y_i$$

2. pAUC(β): area under the ROC curve restricted to FPR $\in [0, \beta]$



$$\ell_{\text{pAUC}(\beta)}(\mathbf{w}) = \sum_{y_i > 0} \sum_{y_j < 0} \mathcal{T}^-_{\beta,T}(\mathbf{x}_j; \mathbf{w}) \cdot h((\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{w})$$

In other words, highly non-decomposable and holistic

Question I

Low regret algorithms and OTBC for non-decomposable loss functions.

A Novel Online Learning Framework

Challenges with the state-of-the-art:

- The very framework is unsuitable for non-decomposable functions
- Notions of instantaneous penalty, regret absent

An Extended Online Framework

A non-decomposable function $\ell_{\mathcal{P}} : (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_t) \times \mathbf{w} \mapsto \mathbb{R}_+$
Instantaneous Penalty: $\mathcal{L}_t(\mathbf{w}) := \ell_{\mathcal{P}}(\mathbf{z}_{1:t}; \mathbf{w}) - \ell_{\mathcal{P}}(\mathbf{z}_{1:t-1}; \mathbf{w})$
Regret: $\mathfrak{R}(T) := \sum \mathcal{L}_t(\mathbf{w}_t) - \arg\min_{\mathbf{w} \in \mathcal{W}} \ell_{\mathcal{P}}(\mathbf{z}_{1:T}; \mathbf{w})$

Properties of the framework:

- For additive losses, AUC, recovers existing frameworks
- Efficient vanishing regret online algorithms, OTB bounds

Low Regret Online Learning and OTB

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w} \in \mathcal{W}} \sum_{\tau=1}^{t} \mathcal{L}_\tau(\mathbf{w}) + \frac{\eta_t}{2} \|\mathbf{w}\|^2 \qquad \text{(FTRL)}$$

Theorem: If $\ell_{\mathcal{P}}(\cdot)$ is *stable* i.e. $\mathcal{L}_t(\cdot)$ is G_t-Lipschitz, then

$$\mathfrak{R}(T) \leq \|\mathcal{W}\| \sqrt{\sum_t G_t^2}$$

Proof Idea: Forward regret bound for non-convex losses + stability. Note that the updates are efficient since $\sum_{\tau=1}^{t} \mathcal{L}_\tau(\mathbf{w}) = \ell_{\mathcal{P}}(\mathbf{z}_{1:t}; \mathbf{w})$ which is convex

Stability Bounds: $\ell_{\text{Prec}_{@k}}$ and $\ell_{\text{pAUC}(\beta)}$ are $\mathcal{O}(1)$-stable

Proofs use a novel Structural Lemma on ranked lists
Sorted lists of inner products are Lipchitz at every position

OTB Bounds: Decompose stream into batches - $\mathbf{Z}_1, \ldots \mathbf{Z}_{T/s}$ and let $\mathcal{L}_t(\mathbf{w}) = \ell_{\mathcal{P}}(\mathbf{Z}_{1:t}; \mathbf{w}) - \ell_{\mathcal{P}}(\mathbf{Z}_{1:t-1}; \mathbf{w})$ with regret bound $\mathfrak{R}(T, s)$.
Theorem: If $\mathbf{w}_1, \ldots, \mathbf{w}_{T/s}$ is the online ensemble, $\bar{\mathbf{w}} = \frac{1}{T/s} \cdot \sum_t \mathbf{w}_t$, then for any $\epsilon \in (0, 0.5]$, $\mathbf{w}^* \in \mathcal{W}$, w.h.p.

$$\mathcal{R}_{\mathcal{P}}(\bar{\mathbf{w}}) \leq \mathcal{R}_{\mathcal{P}}(\mathbf{w}^*) + T^{-1} \cdot \mathfrak{R}(T, s) + e^{-s\epsilon^2} + \sqrt{s/T}$$

- For Prec$_{@k}$, pAUC(β), we show $\mathfrak{R}(T, \sqrt{T}) \leq T^{3/4}$ so $s = \sqrt{T}$ suffices
- Two-stage proof technique based on martingale bounds, UC

Question II

Scalable stochastic gradient methods for non-decomposable functions.

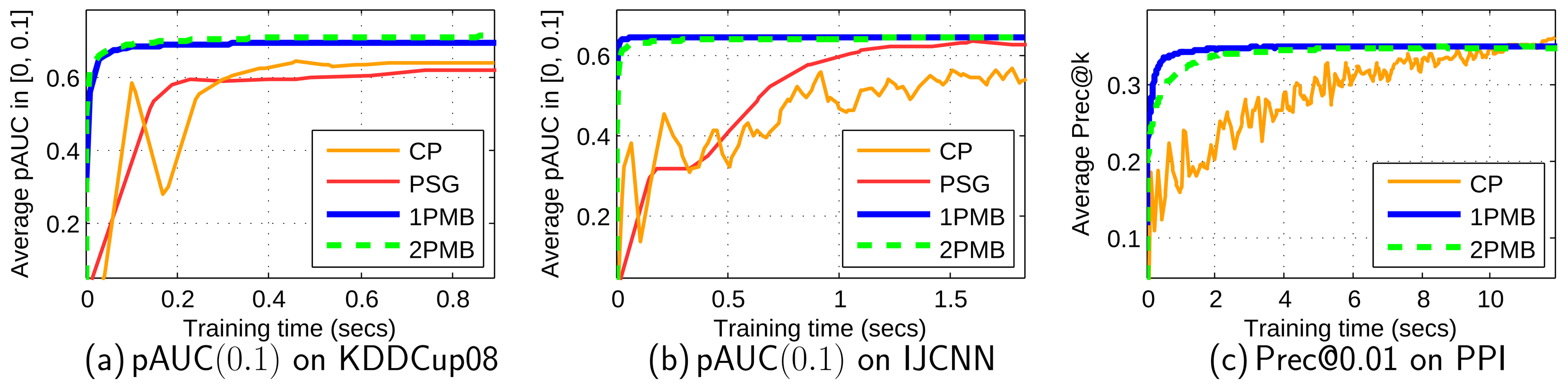
Stochastic gradients for non-additive functions

- Obtaining unbiased gradient estimates not cheap
  - Obtain cheap gradients with small bias instead

Algorithm 1 (1PMB) - Single Pass with Mini Batches

Input: Buffer B of size s, step length scale η
1. while stream not exhausted do
2. Collect s points $\mathbf{z}_1^e, \ldots \mathbf{z}_s^e$ in B
3. $\mathbf{w}_{e+1} \leftarrow \Pi_{\mathcal{W}} \left[\mathbf{w}_e - \frac{\eta}{\sqrt{e}} \cdot \nabla_{\mathbf{w}} \ell_{\mathcal{P}}(\mathbf{z}_{1:s}^e; \mathbf{w}_e) \right]$
4. end while
5. return $\bar{\mathbf{w}} = \frac{1}{T/s} \sum_e \mathbf{w}_e$

- Orders of magnitude faster than cutting plane/PSG methods
- Range of loss functions Prec$_{@k}$, pAUC(β), F-measure



(a) pAUC(0.1) on KDDCup08 (b) pAUC(0.1) on IJCNN (c) Prec@0.01 on PPI

UC Bounds for Non-decomposable losses

Non-decomposable functions exhibiting UC

A function $\ell(\mathbf{z}_1, \ldots; \mathbf{w})$ is $\alpha(s)$-UC if, for $\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_s$ randomly sampled from $\mathbf{z}_1, \ldots, \mathbf{z}_T$, we have w.h.p., $\sup_{\mathbf{w} \in \mathcal{W}} |\ell(\hat{\mathbf{z}}_{1:s}; \mathbf{w}) - \ell(\mathbf{z}_{1:T}; \mathbf{w})| \leq \alpha(s)$

Common proof techniques not applicable for non-decomposable functions

Novel UC Proofs: pAUC(β), Prec$_{@k}$, F-measure are $\mathcal{O}\left(\frac{1}{\sqrt{s}}\right)$-UC

Application: convergence bounds for **1PMB** method
Theorem: Suppose the stream is randomly ordered and $\ell_{\mathcal{P}}$ is $\alpha(s, \delta)$-UC

$$\ell_{\mathcal{P}}(\mathbf{z}_{1:T}; \bar{\mathbf{w}}) \leq \ell_{\mathcal{P}}(\mathbf{z}_{1:T}; \mathbf{w}^*) + 2\alpha(s, s\delta/T) + \sqrt{s/T}$$

Proof Idea: Regret bound, Hoeffding's lemma for randperms

28th Annual Conference on Neural Information Processing Systems (**NIPS 2014**)

Full Paper : http://tinyurl.com/llqws8z