Trade-offs in Large-Scale Distributed Tuplewise Estimation and Learning

Robin Vogel[1,2], Aurélien Bellet[3], Stéphan Clémençon[1], Ons Jelassi[1] and Guillaume Papa[1]. [1] Télécom Paris, [2] IDEMIA, [3] Inria

Motivation

Frameworks for cluster computing:
· ease the deployment of distributed algorithms,
· add *restrictions* on the type of operations available,
· behave well for standard averages.

The statistical learning literature:
· proves guarantees for statistical methods,
· generally ignores cluster-computing *restrictions*,
· handles computational aspects in a stylized manner,

U-statistics:
· are averages over all tuples of data points.
· arise in many practical problems.
· are well studied in a *centralized setting*,

In a *distributed setting*,
U-statistics need **lots of network communication.**

Contributions

Methods and analyses in a distributed setting for:
· statistical estimation of U-statistics,
· learning with U-statistics,
with good trade-off between **accuracy** & **scalability**.

Illustrative experiments.

Preliminaries

Define two independent i.i.d. samples, with $m \ll n$:

$$\mathcal{D}_n = \{X_k\}_{k=1}^n \subset \mathcal{X} \text{ and } \mathcal{Q}_m = \{Z_l\}_{l=1}^m \subset \mathcal{Z}.$$

Given $h : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$, we estimate $U(h)$, where

$$U(h) = \mathbb{E}[h(X_1, Z_1)].$$

A **two-sample U-statistic** writes:

$$U_{\mathbf{n}}(h) = \frac{1}{nm} \sum_{k=1}^{n} \sum_{l=1}^{m} h(X_k, Z_l), \qquad (1)$$

and is a MVUE of $U(h)$ that sums nm terms.

An **incomplete** two-sample U-statistic writes:

$$\widetilde{U}_B(H) = \frac{1}{B} \sum_{k,l \in \mathcal{D}_B} h(X_k, Z_l), \qquad (2)$$

with $\mathcal{D}_B$ a set of B pairs selected by uniform SWR.

In [1], $\widetilde{U}_B(H)$ is argued to be a statistically efficient approximation of $U_{\mathbf{n}}(h)$.

In a **distributed setting**, with N workers, denote by:
· $\mathcal{R}_i^{\mathcal{X}}$ the instances of $\mathcal{D}_n$ held by worker i,
· $\mathcal{R}_i^{\mathcal{Z}}$ the instances of $\mathcal{Q}_m$ held by worker i,
and $n_i = |\mathcal{R}_i^{\mathcal{X}}|$ and $m_i = |\mathcal{R}_i^{\mathcal{Z}}|$ for all $1 \le i \le N$.

The **full estimator on cluster** $\mathcal{R}_i$ writes:

$$U_{\mathcal{R}_i}(h) = \frac{1}{n_i m_i} \sum_{k \in \mathcal{R}_i^{\mathcal{X}}} \sum_{l \in \mathcal{R}_i^{\mathcal{Z}}} h(X_k, Z_l). \qquad (3)$$

The **incomplete estimator on cluster** $\mathcal{R}_i$ writes:

$$U_{B,\mathcal{R}_i}(h) = \frac{1}{B} \sum_{k,l \in \mathcal{R}_{i,B}} h(X_k, Z_l), \qquad (4)$$

with $\mathcal{R}_{i,B}$ sampled in $\mathcal{R}_i^{\mathcal{X}} \times \mathcal{R}_i^{\mathcal{Z}}$ as $\mathcal{D}_B$ in eq. (2).

Different strategies exist for distributing $\mathcal{D}_n, \mathcal{Q}_m$, here we took proportional ($n_i = n_j, \forall i \neq j$) SWOR.

Candidate estimators

Naive estimators: Use only *intra-cluster* pairs.
· Average full U-statistics $U_{\mathcal{R}_i}$ for each cluster:

$$U_{\mathbf{n},N} = \frac{1}{N} \sum_{i=1}^{N} U_{\mathcal{R}_i}.$$

· Average incomplete U-stats $U_{B,\mathcal{R}_i}$ for each cluster:

$$\widetilde{U}_{\mathbf{n},N,B} = \frac{1}{N} \sum_{i=1}^{N} \widetilde{U}_{B,\mathcal{R}_i}.$$

Proposed estimators: based on **redistributing data**.
· Average T estimators $U_{\mathbf{n},N}$ on T redistributions:

$$\widehat{U}_{\mathbf{n},N,T} = \frac{1}{T} \sum_{t=1}^{T} U_{\mathbf{n},N}^t.$$

· Average $\widetilde{U}_{\mathbf{n},N,B}$'s on T redistributions:

$$\widetilde{U}_{\mathbf{n},N,B,T} = \frac{1}{T} \sum_{t=1}^{T} \widetilde{U}_{\mathbf{n},N,B}^t.$$

Analysis

All estimators are **unbiased**, we study their **variance**.

Hoeffding's second decomposition [2] writes:

$$h(x,z) = h_0(x,z) + h_1(x) + h_2(z) - U(h),$$

with $h_1(x) = \mathbb{E}[h(x, Z_1)]$, $h_2(z) = \mathbb{E}[h(X_1, z)]$,
and $h_0(x,z) = h(x,z) - h_1(x) - h_2(z) + U(h)$.

Introduce $\sigma_1^2 = \mathrm{Var}(h_1(X_1))$, $\sigma_2^2 = \mathrm{Var}(h_2(Z_1))$, and $\sigma_0^2 = \mathrm{Var}(h_0(X_1, Z_1))$, it implies:

$$\mathrm{Var}(U_{\mathbf{n}}(h)) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m} + \frac{\sigma_0^2}{nm}.$$

The **naive estimators' variances** grow in N:

$$\mathrm{Var}(U_{\mathbf{n},N}(h)) = \mathrm{Var}(U_{\mathbf{n}}(h)) + (N-1)\frac{\sigma_0^2}{nm},$$

$$\mathrm{Var}(\widetilde{U}_{\mathbf{n},N,B}(h)) = \left(1 - \frac{1}{B}\right) \mathrm{Var}(U_{\mathbf{n},N}(h)) + \frac{\sigma^2}{NB},$$

while the **proposed estimators'** grow in N/T:

$$\mathrm{Var}(\widehat{U}_{\mathbf{n},N,T}(h)) = \mathrm{Var}(U_{\mathbf{n}}(h)) + (N-1)\frac{\sigma_0^2}{nmT},$$

$$\mathrm{Var}(\widetilde{U}_{\mathbf{n},N,B,T}(h)) = \mathrm{Var}(\widehat{U}_{\mathbf{n},N,T}(h)) + \frac{\sigma^2}{NTB} - \frac{1}{TB}\mathrm{Var}(U_{\mathbf{n},N}(h)).$$

The variance of gradient estimations impacts **SGD**. One idea is to redistribute data every n_r iterations.

References

[1] Stephan Clémençon, Igor Colin, and Aurélien Bellet. Scaling-up Empirical Risk Minimization: Optimization of Incomplete U-statistics. *JMLR*, 2016.

[2] Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 1948.

[3] Shebuti Rayana. ODDS library, 2016.

Experiments

The *shuttle* dataset is classic in outlier detection [3]. It verifies $n \approx 45,000$ and $m \approx 3,500$.

We optimize its *Area Under the ROC Curve* (**AUC**), by minimizing the U-statistic with kernel:

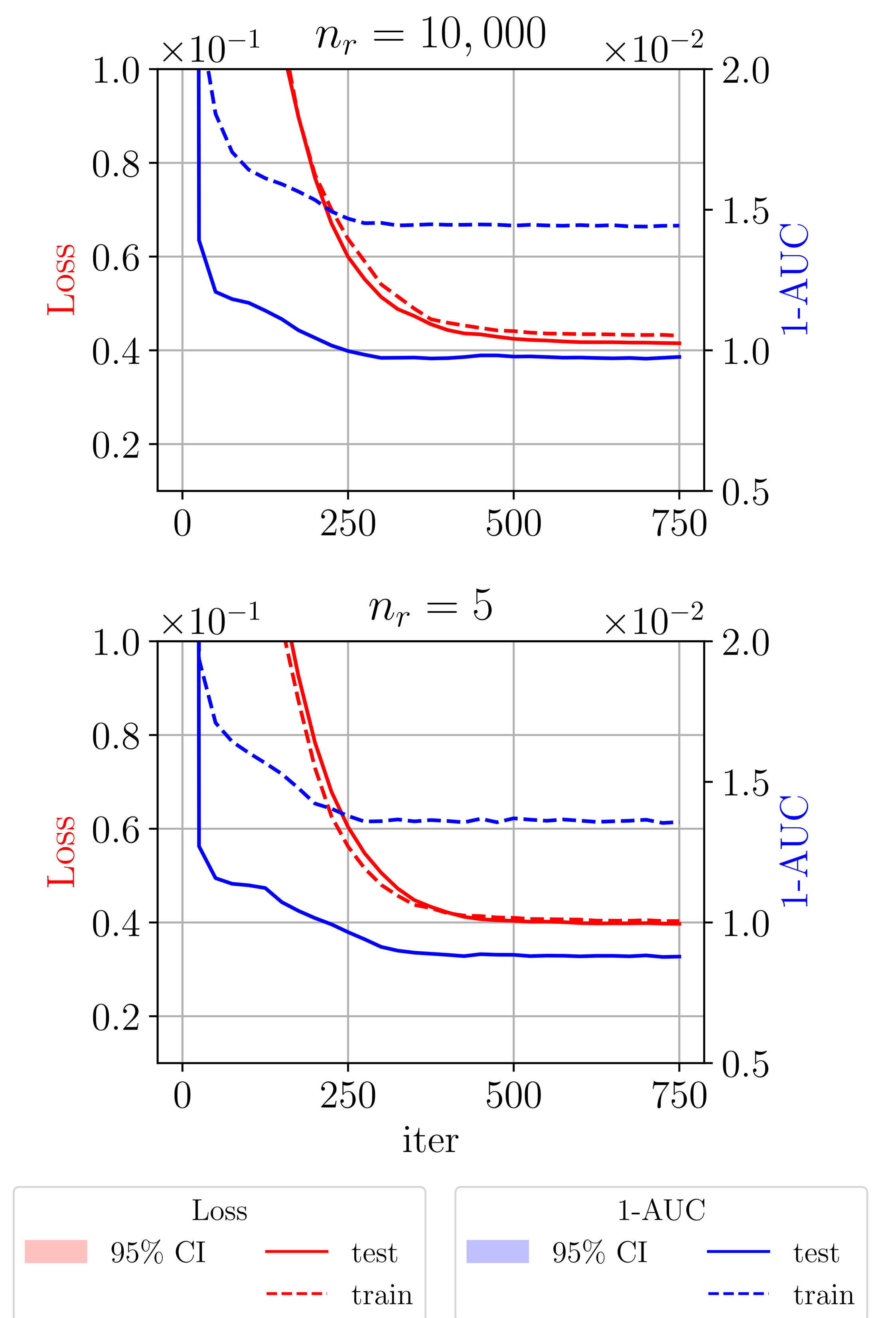
$$h_{w,b}(x,z) = \max(0, 1 + s_{w,b}(x) - s_{w,b}(z)),$$

where $s_{w,b}(x) = w^\top x + b$, with weight decay on w.

A full U-stat on 20% of the data tracks our test AUC. The rest of the data is split over $N = 100$ workers. The train AUC is tracked on a fixed sample of pairs.

We use GD with learning rate 0.01, momentum 0.9 and gradient estimates akin to $\widetilde{U}_{\mathbf{n},N,B}$ with $B = 100$.

Lines below are medians at each iter over 100 runs.



Variance-Time Tradeoff

For $n = 100,000$, $m = 200$ and $N = 100$, we plot the variance as a function of the number of evaluated pairs.