U-STATISTICS IN MACHINE LEARNING

LARGE-SCALE MINIMIZATION AND DECENTRALIZED ESTIMATION

Aurélien Bellet
MAGNET Team, INRIA Lille

Joint work with:
Stéphane Clémençon, Igor Colin, Guillaume Papa and Joseph Salmon (Télécom ParisTech)
1. Introduction: U-Statistics

2. Large-Scale Empirical Risk Minimization

3. Decentralized Estimation

4. Conclusion & Perspectives
INTRODUCTION: U-STATISTICS
• Let $\mu$ some (unknown) distribution on space $\mathcal{X}$

• Let $X_1, \ldots, X_n$ drawn i.i.d. from $\mu$

• **Univariate statistic**: estimate $\mathbb{E}_{X \sim \mu}[H(X)]$ with $\frac{1}{n} \sum_{i=1}^{n} H(X_i)$
  - $H : \mathcal{X} \to \mathbb{R}$
  - Example (sample mean): $\frac{1}{n} \sum_{i=1}^{n} X_i$

• **Pairwise statistic**: estimate $\mathbb{E}_{X_1, X_2 \sim \mu}[H(X_1, X_2)]$ with

  \[
  \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} H(X_i, X_j)
  \]

  - $H : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric
  - Example 1 (sample variance): $H(X, X') = (X - X')^2 / 2$
  - Example 2 (average distance): $H(X, X') = \|X - X'\|$
• **U-statistic of degree** $d$ with kernel $H$ [Hoeffding, 1948]:

$$U_n(H) = \frac{1}{\binom{n}{d}} \sum_{1 \leq i_1 < \ldots < i_d \leq n} H(X_{i_1}, \ldots, X_{i_d})$$

• $H : \mathcal{X}^d \to \mathbb{R}$ symmetric
• Note: can be generalized to multi-sample setting

• $U_n$ has **minimum variance** among all unbiased estimators of

$$U(H) = \mathbb{E}_{X_1, \ldots, X_d \sim \mu} [H(X_1, \ldots, X_d)]$$

• But for $d \geq 2$, not a sum of independent terms!

• Need specific tools to bound $|U_n(H) - U(H)|$
  • Decoupling: see for instance [de la Peña and Giné, 1999]
LARGE-SCALE MINIMIZATION  

NIPS ’15 + JMLR ’16
A standard paradigm in machine learning

- \( \mathcal{G} \): class of learning rules (e.g., linear classifiers)
- \( H_g : \mathcal{X}^d \rightarrow \mathbb{R} \): loss function associated with \( g \in \mathcal{G} \)
- True risk of rule \( g \in \mathcal{G} \): \( U(H_g) = \mathbb{E}_{X_1, \ldots, X_d \sim \mu} [H_g(X_1, \ldots, X_d)] \)
- Empirical risk of \( g \in \mathcal{G} \): \( U_n(H_g) = \frac{1}{n^d} \sum_{1 \leq i_1 < \ldots < i_d \leq n} H_g(X_{i_1}, \ldots, X_{i_d}) \)
- Empirical Risk Minimization (ERM): choose rule \( \hat{g} \in \arg \min_{g \in \mathcal{G}} U_n(H_g) \)
• Find a partition $\mathcal{P}$ of space $\mathcal{X}$

• Within-cluster point scatter [Clémençon, 2011]

$$W_n(\mathcal{P}) = \frac{2}{n(n-1)} \sum_{i<j} D(X_i, X_j) \cdot \mathbb{1}\{\exists C \in \mathcal{P} \text{ s.t. } X_i, X_j \in C^2\}$$
• Labeled data: \((X_i, Y_i) \in \mathcal{X} \times \{1, \ldots, C\}\)

• Learn distance measure adapted to the task [Bellet et al., 2015]

• Distance function \(D : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+\)

• Triplet-based criterion

\[
T_n(D) = \frac{6}{n(n-1)(n-2)} \sum_{i<j<k} \mathbb{I} \{D(X_i, X_j) > D(X_i, X_k), \ Y_i = Y_j \neq Y_k\}
\]
EXAMPLE: LEARNING TO RANK

- Labeled data: \((X_i, Y_i) \in \mathcal{X} \times \{-1, 1\}\)
- Learn to rank items (e.g., relevant vs irrelevant)
- Scoring function \(s : \mathcal{X} \to \mathbb{R}\)
- Area Under the ROC Curve (AUC) [Zhao et al., 2011]

\[
AUC_n(s) = \frac{1}{|X^+||X^-|} \sum_{X_i^+ \in X^+} \sum_{X_j^- \in X^-} \mathbb{I} \left\{ s(X_j^-) < s(X_i^+) \right\}
\]

where \(X^+ = \{X_i : Y_i = 1\}\) and \(X^- = \{X_i : Y_i = -1\}\)
- Generalizes to multi-partite ranking [Clémençon et al., 2013]
• Let \( \hat{g} \in \arg\min_{g \in G} U_n(H_g) \) the empirical risk minimizer

• Under suitable assumptions [Clémençon et al., 2008]

\[
U(H_{\hat{g}}) - \inf_{g \in G} U(H_g) = O_{\mathbb{P}}(1/\sqrt{n})
\]

• How to find \( \hat{g} \) efficiently? \( U_n(H_g) \) has \( O(n^d) \) terms!
  • Big data problem even for relatively small datasets

• We will exploit the dependence structure of \( U_n \)
Main idea: approximate $U_n$ by an incomplete $U$-statistic

$$\tilde{U}_B(H_g) = \frac{1}{B} \sum_{I \in \mathcal{D}_B} H_g(X_{I_1}, \ldots, X_{I_d})$$

where $\mathcal{D}_B$ is a set of cardinality $B$ drawn by sampling with replacement from the set of $d$-tuples

This is different from a $U$-statistic based on a subsample

Naive sampling (complete $U$-statistic)
Pair sampling (incomplete $U$-statistic)
Theorem ([Clémençon et al., 2016])

Let $\mathcal{H} = \{\mathcal{H}_g : g \in \mathcal{G}\}$ be a VC major class of functions with VC dimension $V < +\infty$ and uniformly bounded by $M_\mathcal{H} < +\infty$. For all $\eta > 0$, we have $\forall n, \forall B \geq 1$,

$$\Pr\left\{ \sup_{H \in \mathcal{H}} |\tilde{U}_B(H) - U_n(H)| > \eta \right\} \leq 2 \left( 1 + \left( \frac{n}{d} \right)^V \right) \times e^{-B\eta^2/M_\mathcal{H}^2}$$

- Prob. of large deviation decreases exponentially fast with $B$

- Main ingredients of the proof
  - Write $\tilde{U}_B(H) - U_n(H)$ as an average of $B$ independent variables
  - Sauer’s lemma
  - Union bound and Hoeffding’s inequality
Corollary ([Clémençon et al., 2016])

Let $\tilde{g}$ be an empirical risk minimizer of $\tilde{U}_B$ over $\mathcal{H}$, and $\delta > 0$. Under the previous assumptions, with probability at least $1 - \delta$, we have:

$$U(H_{\tilde{g}}) - \inf_{g \in G} U(H_g) \leq O\left(\sqrt{\frac{V \log(n) + \log(2/\delta)}{n}} + \sqrt{\frac{V \log\left(\binom{n}{d}\right) + \log(4/\delta)}{B}}\right)$$

- Choosing $B = O(n)$ preserves the $O_P(1/\sqrt{n})$ learning rate!

- In contrast: complete $U$-statistic with $O(n)$ terms leads to much slower rate of $O_P\left(\sqrt{1/n^{\frac{1}{d}}}\right)$

- Other results (not covered here): fast rates, model selection
• \( \Theta \subset \mathbb{R}^q \) parameter space

• \( H : \mathcal{X}^d \times \Theta \to \mathbb{R} \) strongly convex and smooth in 2nd argument

• Reformulation of true risk

\[
L(\theta) \overset{\text{def}}{=} U(H(\cdot; \theta))
\]

• Reformulation of empirical risk

\[
\hat{L}_n(\theta) \overset{\text{def}}{=} U_n(H(\cdot; \theta))
\]

• Reformulation of ERM problem

\[
\min_{\theta \in \Theta} \hat{L}_n(\theta)
\]
GRADIENT DESCENT

• Initialize $\theta_0 \in \Theta$ and follow the iterations

$$\theta_{t+1} = \theta_t - \eta_t \nabla_{\theta} \hat{L}_n(\theta_t), \quad \eta_t \geq 0$$

• Gradient of $\hat{L}_n(\theta)$ is

$$\nabla_{\theta} \hat{L}_n(\theta) = \frac{1}{n \choose d} \sum_{1 \leq i_1 < \ldots < i_d \leq n} \nabla_{\theta} H(X_{i_1}, \ldots, X_{i_d}; \theta)$$

• Each gradient involves summing over $n \choose d$ terms!

• Stochastic Gradient Descent (SGD): approximate gradient at each step using a random mini-batch of terms
Use incomplete $U$-statistic with $B$ terms to estimate the gradient

**Theorem ([Papa et al., 2015])**

Let $\mathcal{H} = \{H(\cdot; \theta) : \theta \in \Theta\}$ be a VC major class of functions with VC dimension $V < +\infty$ and uniformly bounded by $M_{\mathcal{H}} < +\infty$. Let $\theta^* = \arg \min_{\theta \in \Theta} L(\theta)$. Under appropriate conditions on the step size, we have for $\forall n$:

$$
\mathbb{E}[|L(\theta_t) - L(\theta^*)|] \leq O \left( \frac{1}{Bt} + M_{\mathcal{H}} \sqrt{\frac{V \log(n)}{n}} \right)
$$

- Decomposition into optimization and generalization errors
- Set $B = \binom{n'}{d}$. Alternative: use complete $U$-statistic of size $n'$
  - Both estimates consist of $B$ terms
  - But $B$ is replaced by $n'$ in the bound!
• Pairwise metric learning

\[ R_n(M) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} [y_{ij}(b - (X_i - X_j)^T M(X_i - X_j))]_+ \]

• \( M \): \( q \times q \) PSD matrix
• \( y_{ij} = 1 \) if \( y_i = y_j \), \(-1\) otherwise
• \([u]_+ = \max(0, 1 - u)\): hinge loss

• MNIST dataset
  • \( n = 60,000 \rightarrow 2 \times 10^9 \) pairs
Approximate risk by complete or incomplete $U$-statistic
SGD: Approximate gradient with complete or incomplete $U$-statistic

![Graph showing risk estimate on test set for iterations with $B=10$ and $B=55$.]
DECENTRALIZED ESTIMATION IN NETWORKS

- Estimation of statistics from data distributed over network graph
- Want asynchronous algorithm + limited communication/storage
- Applications: telecommunication, sensor networks, IoT
• Data points $X_1, \ldots, X_n \in \mathcal{X}$

• Network represented as a connected graph $G = (V, E)$
  - Nodes $V = \{1, \ldots, n\}$
  - Node $i$ holds point $X_i$
  - $(i, j) \in E$: $i$ and $j$ can exchange information directly

• **Goal:** estimate pairwise statistic

\[
\hat{U}_n(H) = \frac{1}{n^2} \sum_{i,j=1}^{n} H(X_i, X_j)
\]

• $\hat{U}_n(H)$ is a degree 2 $U$-statistic (up to normalization factor)
• **Synchronous algorithm**
  - *Global clock* ticking at the times of a rate 1 Poisson process
  - Each time the clock ticks, all nodes activate

• **Asynchronous algorithm**
  - Each node has a *local clock*
  - Each time a node’s clock ticks, it activates
  - For modeling purposes: equivalent to a single Poisson clock ticking at rate $n$ with random selection of node to activate
GOSSIP ALGORITHMS FOR STANDARD AVERAGING

- Gossip algorithms [Shah, 2009]: one edge activated at a time
- Canonical problem: estimate sample mean $\frac{1}{n} \sum_{i=1}^{n} X_i$
- Simple gossip algorithm [Boyd et al., 2006]
  - At each iteration, draw $(i, j) \in E$, $i$ and $j$ average their estimates
  - Geometric convergence
  - Natively asynchronous
- Naive extension to pairwise statistics $\rightarrow$ massive data transfer
GOSTA IN A NUTSHELL

- Each node stores an auxiliary observation and an estimate

\[
\begin{align*}
X_i & \quad Y_i & \quad Z_i \\
\text{original observation} & \quad \text{U-statistic estimator} & \quad \text{auxiliary observation}
\end{align*}
\]

- An iteration combines averaging and data propagation

\[
\begin{align*}
\text{Time } t: & & \quad \text{local memory} & \quad \text{local memory} \\
X_i & \quad Y_i & \quad Z_i & \quad X_j & \quad Y_j & \quad Z_j
\end{align*}
\]

\[
\begin{align*}
\text{mix estimates:} & \quad Z_i & \leftarrow & \frac{Z_i + Z_j}{2} & \quad Z_j & \leftarrow & \frac{Z_i + Z_j}{2}
\end{align*}
\]

\[
\begin{align*}
\text{update:} & \quad Z_i & \leftarrow & (1 - \alpha_t) Z_i + \alpha_t \text{H}(X_i, Y_i) & \quad Z_j & \leftarrow & (1 - \alpha_t) Z_j + \alpha_t \text{H}(X_j, Y_j)
\end{align*}
\]

\[
\begin{align*}
\text{swap auxiliary data:} & \quad Y_i & \leftarrow & Y_j & \quad Y_j & \leftarrow & Y_i
\end{align*}
\]
· Need a global clock

**Algorithm 1 GoSta-sync**

**Require:** Each node $k$ holds $X_k$
Each node $k$ initializes $Y_k = X_k$ and $Z_k = 0$

for $t = 1, 2, \ldots$ do

for $p = 1, \ldots, n$ do

Set $Z_p \leftarrow \frac{t-1}{t} Z_p + \frac{1}{t} H(X_p, Y_p)$

end for

Draw $(i, j)$ uniformly at random from $E$

Set $Z_i, Z_j \leftarrow \frac{1}{2} (Z_i + Z_j)$
Swap auxiliary observations: $Y_i \leftrightarrow Y_j$

end for
GOSTA: ASYNCHRONOUS VERSION

- No **global clock**: only selected nodes are active
- Each node $i$ stores an **unbiased estimate** $m_i$ of current iteration
  - Probability $p_i = 2d_i/|E|$ that $i$ awakes at a given iteration
  - When $i$ awakes, it updates $m_i \leftarrow m_i + 1/p_i$

---

**Algorithm 2 GoSta-async**

**Require:** Each node $k$ holds $X_k$ and $p_k$

Each node $k$ initializes $Y_k = X_k$, $Z_k = 0$ and $m_k = 0$

**for** $t = 1, 2, \ldots$ **do**

1. Draw $(i, j)$ uniformly at random from $E$
2. Set $m_i \leftarrow m_i + 1/p_i$ and $m_j \leftarrow m_j + 1/p_j$
3. Set $Z_i, Z_j \leftarrow \frac{1}{2}(Z_i + Z_j)$
4. Set $Z_i \leftarrow (1 - \frac{1}{p_im_i})Z_i + \frac{1}{p_im_i}H(X_i, Y_i)$
5. Set $Z_j \leftarrow (1 - \frac{1}{p_jm_j})Z_j + \frac{1}{p_jm_j}H(X_j, Y_j)$

Swap auxiliary observations: $Y_i \leftrightarrow Y_j$

**end for**
**CONVERGENCE ANALYSIS**

<table>
<thead>
<tr>
<th>Theorem ([Colin et al., 2015])</th>
</tr>
</thead>
<tbody>
<tr>
<td>If $G = (V, E)$ is connected and non-bipartite, then for any $t &gt; 0$:</td>
</tr>
<tr>
<td>$\left| E[Z(t)] - \hat{U}_n(H)1_n \right| \leq \frac{1}{ct} \left| \bar{h} - \hat{U}_n(H)1_n \right| + \left( \frac{2}{ct} + e^{-ct} \right) \left| H - \bar{h}1_n^T \right|$,</td>
</tr>
<tr>
<td>where $c = c(G) := \beta_{n-1}/</td>
</tr>
</tbody>
</table>

- **Data-dependent** terms: quantify difficulty of estimation problem
  - Dispersion measure between the values to be averaged

- **Network-dependent** terms: quantify how well things propagate
  - Graphs with larger spectral gap $\rightarrow$ better connectivity [Chung, 1997]
• Two estimation problems
  • *Within-cluster point scatter* on Wine quality dataset \((n = 1,599)\)
  • *Area Under the ROC Curve* on SMVguide3 dataset \((n = 1,260)\)

• Three types of graphs

![Diagram of three types of graphs: 2D-grid, Watts-Strogatz, Complete](image-url)
NUMERICAL SIMULATIONS

Comparison to U2-Gossip [Pelckmans and Suykens, 2009]

- U2-Gossip: propagates two observations, no averaging
- Only synchronous, worst theoretical guarantees
Comparison to U2-Gossip [Pelckmans and Suykens, 2009]

- GoSta scales better with $n$
- GoSta-sync and GoSta-async have similar performance
Comparison to “Master Node” baseline

- Baseline has access to master node connected to all nodes
- Our algorithm compensates well for lack of central node
CONCLUSION & PERSPECTIVES
CONCLUSION & PERSPECTIVES

Wrapping up

• $U$-statistics involved in many estimation and learning problems
• Sampling / stochastic optimization schemes to scale-up ERM
• Gossip algorithms for decentralized estimation

Looking ahead

• Decentralized ERM (ICML 2016 paper)
• Privacy, robustness to malicious users (under progress)
• Adaptive communication schemes: learn who to talk to
Thank you for your attention! Questions?
*Metric Learning.*

*Randomized gossip algorithms.*

*Spectral Graph Theory, volume 92.*
American Mathematical Society.

to appear.

*Ranking data with ordinal labels: Optimality and pairwise aggregation.*

*On U-processes and clustering performance.*
In NIPS, pages 37–45.


