Kickoff IA – Chaire BiSCottE

(Bridging Statistical and Computational Efficiency in AI)

Gilles Blanchard

Université Paris-Saclay

9 sept. 2020

université PARIS-SACLAY

Participating doctoral candidates:

- Jean-Baptiste Fermanian (ENS Rennes)
- Karl Hajjar (Saclay)
- Hannah Marienwald (TU Berlin)
- El Mehdi Saad (Saclay)
- Olympio Hacquard (Saclay)
- Jérémie Capitao Miniconi (Saclay)

Collaborating Colleagues:

- Sylvain Arlot (IMO, Saclay)
- Frédéric Chazal (INRIA, Saclay)
- Lénaïc Chizat (CNRS,IMO, Saclay)
- Elisabeth Gassiat (IMO, Saclay)
- Christophe Giraud (IMO, Saclay)
- Rémi Gribonval (INRIA, Lyon)

High-level goals

- Project positioned in current trend of of statistical and computational tradeoffs
- **Label efficiency** – information theoretic sense
  - Example: Requesting only just enough data (online) as needed for the task at hand
  - Example: "Small data" problem – many learning tasks with few data
- **Computational resource efficiency**
  - Computation time
  - Memory
  - Example: early stopping of iterative approximation/optimization
- **Structural efficiency** – taking advantage of unknown structures in data
  - Example: data lies (close to) an unknown manifold
  - Example: finding efficient representations
- **Mainly theoretical orientation – interactions welcome**

Efficient variable selection

Work with El Mehdi Saad

- Start from the fundamental linear regression problem:

$$Y_i = \langle X_i, \beta_* \rangle + \varepsilon_i, \text{ with } (X_i, Y_i) \text{ i.i.d.}$$

- **Assume** $X_i \in \mathbb{R}^d$ but

$$|\mathrm{Supp}(\beta_*)| \ll d \qquad \mathrm{Supp}(\beta_*) := \left\{ i \leq d : \beta_*^{(i)} \neq 0 \right\}.$$

- Many variable selection methods, **Orthogonal Matching Pursuit** still very popular:
  0. $\bar{\beta} \leftarrow 0, S \leftarrow \emptyset$, all data ($i = 1, \ldots, n$) available
  1. [Residuals] $R_i \leftarrow Y_i - \langle X_i, \bar{\beta} \rangle, i = 1, \ldots, n$
  2. [Selection] $S \leftarrow S \cup \underset{s \in [d] \setminus S}{\mathrm{Arg\,Max}} \widehat{E}(RX^{(s)})$
  3. [OLS] $\bar{\beta} \leftarrow \underset{\mathrm{Supp}(\beta) \subseteq S}{\mathrm{Arg\,Min}} \|Y - \langle X, \beta \rangle\|_n^2$
  4. Go to point 1.
- Statistical reliability studied by Zhang (JMLR 2009):
  minimum data size n (under appropriate assumptions) for selection consistency

Efficient variable selection
Online OMP

- Complexity of batch OMP (for k selection steps): $\mathcal{O}(knd)$
  and n depends on some a priori assumptions (RIP, smallest coefficient magnitude)

- Approach:
  - query data **only as needed** for reliable selection at each step (bandit arm style)
  - approximate OLS as needed by ASGD

- Study sample & computational complexity under:
  - **Data Base model** (arbitrary (data,coordinate) queries with unit cost)
  - **Data Stream model** (asked for partially observed new sample, can't query backwards)

Efficient multiple-mean estimation

Work with Hannah Marienwald, Jean-Baptiste Fermanian

- Independent samples $X_\bullet^{(b)}, b = 1, \dots, B$ on $\mathbb{R}^d$:

$$\begin{cases} X_\bullet^{(b)} := (X_i^{(b)})_{1 \le i \le N_b} \overset{i.i.d.}{\sim} \mathbb{P}_b, \\ (X_\bullet^{(1)}, \dots, X_\bullet^{(B)}) \text{ independent,} \end{cases}$$

- Goal is to estimate means $\mu_b := \mathbb{E}_{X \sim \mathbb{P}_b}[X] \in \mathbb{R}^d, \ b = 1, \dots, d$.

- **Question:** can we exploit unknown structure in the true means (clustering, manifold...) to improve over naive estimation $\widehat{\mu}_b^{\mathrm{NE}} := N_b^{-1} \sum_{i=1}^{N_b} X_i^{(b)}$?
  $\rightarrow$ **Structural efficiency and small data problem**

- **Relation to AI/machine learning?**
  - large databases of that form (e.g. medical records, online activity of many users)
  - relation to Kernel Mean Embedding (KME): estimation of $\Phi(P) = \mathbb{E}_{X \sim P}[\Phi(X)]$ where Φ is some kernel feature map
  - improving KME estimations has many applications (Muandet et al., ICML 2014)
  - improving multiple mean estimation also analyzed in ML (Feldman et al. NIPS 2012, JMLR 2014)

Multiple-mean estimation by local averaging

- Assume standard Gaussian distributions and equal sample sizes $N_b = N$
- Focus on estimating μ_0. Naive estimator $\widehat{\mu}_0^{\mathrm{NE}}$ has $\mathrm{MSE}(\mu_0) = d/N =: \sigma^2$
- Assume we know that $\Delta_i^2 = \|\mu_i - \mu_0\|^2 \leq \delta^2$ for "neighbor tasks" 1,...,K
- Consider simple neighbor averaging:

$$\widetilde{\mu}_0 := \frac{1}{K+1} \sum_{i=0}^{K} \widehat{\mu}_i^{\mathrm{NE}} \quad \text{then} \quad \mathrm{MSE}(\widetilde{\mu}_0) \leq \frac{\sigma^2}{K+1} + \delta^2.$$

- Gain if we can detect "neighboring tasks" s.t. $\Delta_i^2 \leq \delta \ll \sigma^2$.
- **Is it a pipe dream?** No, can detect $\Delta_i^2 \lesssim \sigma^2/\sqrt{d}$!
- **Blessing of dimensionality** phenomenon.

Thank you

(Do not hesitate to reach out!)