Localized Complexities for Transductive Learning

Ilya Tolstikhin\textsuperscript{1}, Gilles Blanchard\textsuperscript{2}, Marius Kloft\textsuperscript{3}

\textsuperscript{1}Russian Academy of Sciences
\textsuperscript{2}University of Potsdam
\textsuperscript{3}Humboldt University of Berlin

COLT 2014
Contents:

1. New concentration inequalities for sampling without replacement
2. Application to transductive learning
Concentration inequalities

- Function of many random variables $Q = g(X_1, \ldots, X_n)$
- We want to control random fluctuations of $Q$ around $\mathbb{E}[Q]$

We aim to show high-probability upper bounds on:

$$Q - \mathbb{E}[Q] \quad \text{and/or} \quad \mathbb{E}[Q] - Q.$$

Independent random variables
The case when $X_1, \ldots, X_n$ are independent has been very well studied and many useful results are available [Boucheron et al., 2013].
Concentration inequalities: independent random variables

Consider independent random variables $X_1, \ldots, X_n$ bounded in $[0, 1]$:

$$S_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$ 

Then for any $\delta \in (0, 1]$ with probability greater than $1 - \delta$:

**Hoeffding’s inequality**

$$|S_n - \mathbb{E}[S_n]| \leq \sqrt{\frac{\log(2/\delta)}{2n}};$$

**Bernstein’s inequality**:

$$|S_n - \mathbb{E}[S_n]| \leq \sqrt{\frac{2\sigma^2 \log(2/\delta)}{n}} + \frac{2 \log(2/\delta)}{3n},$$

where $\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \text{Var}[X_i]$.

**Message**: small variance leads to better convergence rates.
Concentration inequalities: independent random variables

Now consider i.i.d. sequence of r.v.'s $X_1, \ldots, X_n$, taking values in $\mathcal{X}$. Let $\mathcal{F}$ be a countable class of bounded functions $f : \mathcal{X} \to [-1, 1]$ such that $\mathbb{E}[f(X_1)] = 0$. Consider the supremum of empirical process:

$$Q_n = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

Then for any $\delta \in (0, 1]$ with probability greater than $1 - \delta$:

**McDiarmid’s inequality:**

$$|Q_n - \mathbb{E}[Q_n]| \leq \sqrt{\frac{2 \log(2/\delta)}{n}};$$

**Talagrand’s inequality (version due to [O. Bousquet, 2002]):**

$$Q_n - \mathbb{E}[Q_n] \leq \sqrt{\frac{2v \log(1/\delta)}{n}} + \frac{2 \log(1/\delta)}{3n},$$

where $v = \sup_{f \in \mathcal{F}} \text{Var}[f(X_1)] + 2\mathbb{E}[Q_n]$. 

Sampling without replacement

Now let $Z_1, \ldots, Z_n$ be sampled uniformly without replacement from given finite set $C = \{c_1, \ldots, c_N\}$ for $N \geq n$.

Note: $Z_1, \ldots, Z_n$ are not independent

Motivation:

- Cross-validation procedures;
- Transductive learning;
- Randomized sequential algorithms (SGD, . . .);
- Matrix completion;
- Low-rank matrix factorization (Collaborative filtering, . . .);
- . . .
Sampling without replacement: previous results

\[ S_n = \frac{1}{n} \sum_{i=1}^{n} Z_i. \]

[Hoeffding, 1963]:
Hoeffding's and Bernstein's inequalities also hold for this setting.

[Serfling, 1974]:
Moreover, for all \( \delta \in (0, 1] \) with prob. greater than \( 1 - \delta \):

\[ |S_n - \mathbb{E}[S_n]| \leq \sqrt{\left( \frac{N - n + 1}{N} \right) \frac{\log(2/\delta)}{2n}}. \]

[Bardenet and Maillard, 2013]:
Bernstein's inequality can be tightened in the same manner.

**Message:** things are more concentrated when random variables are sampled without replacement!
Sampling without replacement: previous results

\[ Q_n = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(Z_i). \]

[El-Yaniv and Pechyony, 2009; Cortes et al., 2009]:

for any \( \delta \in (0, 1] \) with prob. greater than \( 1 - \delta \):

\[
|Q_n - \mathbb{E}[Q_n]| \leq \sqrt{\left( \frac{N - n}{N - 1/2} \right) \frac{1}{\Delta(n, N)} \frac{2 \log(2/\delta)}{n}},
\]

where \( \Delta(n, N) = 1 - \frac{1}{2 \max\{n, N-n\}} \approx 1. \)

This inequality is a (tighter) version of McDiarmid’s inequality.

**Problem:** there is no version of Talagrand’s concentration inequality for sampling without replacement!
Our results

Let $X_1, \ldots, X_n$ and $Z_1, \ldots, Z_n$ be sampled with and without replacement respectively from $C = \{c_1, \ldots, c_N\}$. Consider:

$$Q_{iid}^n = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(X_i), \quad Q_n = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(Z_i), \quad \sigma_F^2 = \sup_{f \in \mathcal{F}} \text{Var}[f(X_1)].$$

**Theorem**

*For any $\delta \in (0, 1]$ with probability greater than $1 - \delta$:

$$|Q_n - \mathbb{E}[Q_n]| \leq 2\sqrt{\frac{2\sigma_F^2 \log(2/\delta)}{n}} \left(\frac{N}{n}\right).$$

**Theorem**

*For any $\delta \in (0, 1]$ with probability greater than $1 - \delta$:

$$Q_n - \mathbb{E}[Q_{iid}^n] \leq \sqrt{\frac{2(\sigma_F^2 + 2\mathbb{E}[Q_{iid}^n]) \log(1/\delta)}{n}} + \frac{\log(1/\delta)}{3n}.$$
Our results

Let $X_1, \ldots, X_n$ and $Z_1, \ldots, Z_n$ be sampled with and without replacement respectively from $C = \{c_1, \ldots, c_N\}$. Consider:

$$Q^{iid}_n = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(X_i), \quad Q_n = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(Z_i), \quad \sigma^2_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \text{Var}[f(X_1)].$$

**Theorem**

For any $\delta \in (0, 1]$ with probability greater than $1 - \delta$:

$$|Q_n - \mathbb{E}[Q_n]| \leq 2\sqrt{\frac{2\sigma^2_{\mathcal{F}} \log(2/\delta)}{n}} \left(\frac{N}{n}\right).$$

**Theorem**

For any $\delta \in (0, 1]$ with probability greater than $1 - \delta$:

$$Q_n - \mathbb{E}[Q^{iid}_n] \leq \sqrt{\frac{2(\sigma^2_{\mathcal{F}} + 2\mathbb{E}[Q^{iid}_n]) \log(1/\delta)}{n}} + \frac{\log(1/\delta)}{3n}.$$
Our results: discussion

\[ |Q_n - \mathbb{E}[Q_n]| \leq \sqrt{\frac{2 \log(2/\delta)}{n} \left( \frac{N - n}{N - 1/2} \right)}; \]  
\[ (\text{Old}) \]

\[ |Q_n - \mathbb{E}[Q_n]| \leq 2 \sqrt{\frac{2 \sigma_F^2 \log(2/\delta)}{n} \left( \frac{N}{n} \right)}; \]  
\[ (\text{New 1}) \]

\[ Q_n - \mathbb{E}[Q_{iid}^n] \leq \sqrt{\frac{2(\sigma_F^2 + 2\mathbb{E}[Q_{iid}^n]) \log(1/\delta)}{n}} + \frac{\log(1/\delta)}{3n}. \]  
\[ (\text{New 2}) \]

\begin{itemize}
  \item (Old) does not account for the variance (Hoeffding-type)
  \item If \( n = o(N) \) (Old) and (New 2) can outperform (New 1)
  \item If \( n = \Omega(N) \) (New 1) outperforms (Old) for \( \sigma_F^2 \leq 1/16 \)
  \item Comparison between (New 2) and (Old) depends on \( \sigma_F^2 \) and \( \mathbb{E}[Q_{iid}^n] \)
  \item \[ 0 \leq \mathbb{E}[Q_{iid}^n] - \mathbb{E}[Q_n] \leq 2n^3/N \]
\end{itemize}

Summary: (New 2) stays informative in all regimes of \( N \) and \( n \); (New 1) can give better results (at least for \( n = \Omega(N) \)).
Our results: discussion

\begin{align*}
|Q_n - \mathbb{E}[Q_n]| & \leq \sqrt{\frac{2 \log(2/\delta)}{n}} \left( \frac{N - n}{N - 1/2} \right); \\
(\text{Old}) \\
|Q_n - \mathbb{E}[Q_n]| & \leq 2 \sqrt{\frac{2 \sigma_F^2 \log(2/\delta)}{n}} \left( \frac{N}{n} \right); \\
(\text{New 1}) \\
Q_n - \mathbb{E}[Q_{n}^{iid}] & \leq \sqrt{\frac{2(\sigma_F^2 + 2 \mathbb{E}[Q_{n}^{iid}]) \log(1/\delta)}{n}} + \frac{\log(1/\delta)}{3n}. \\
(\text{New 2})
\end{align*}

- (Old) does not account for the variance (Hoeffding-type)
- If \( n = o(N) \) (Old) and (New 2) can outperform (New 1)
- If \( n = \Omega(N) \) (New 1) outperforms (Old) for \( \sigma_F^2 \leq 1/16 \)
- Comparison between (New 2) and (Old) depends on \( \sigma_F^2 \) and \( \mathbb{E}[Q_{n}^{iid}] \)
- \( 0 \leq \mathbb{E}[Q^{iid}] - \mathbb{E}[Q_n] \leq 2n^3/N \)

Summary: (New 2) stays informative in all regimes of \( N \) and \( n \); (New 1) can give better results (at least for \( n = \Omega(N) \)).
Contents:

1. New Concentration inequalities for sampling without replacement
2. Application to transductive learning
Transductive learning: setting and notations

Deterministic agnostic setting

Finite instance space $X_n = \{X_1, \ldots, X_N\} \subset \mathcal{X}$ and output space $\mathcal{Y}$

Class $\mathcal{H}$ of predictors $h : \mathcal{X} \rightarrow \mathcal{Y}$

Labelling function $\varphi : \mathcal{X} \rightarrow \mathcal{Y}$ (not necessarily in $\mathcal{H}$)

1. Sample $n \leq N$ inputs $X_n \subseteq X_N$ uniformly without replacement
2. Obtain outputs $Y_n$ for $X_n$ by applying function $\varphi : \mathcal{X} \rightarrow \mathcal{Y}$
3. Reveal training set $S_n = (X_n, Y_n)$ and $u = N - n$ test inputs $X_u$
Transductive learning: setting and notations

**Goal of the learner:** based on $S_n$ and $X_u$ find a predictor in hypothesis class $\mathcal{H}$ with minimal test error:

$$L_u(h) = \frac{1}{u} \sum_{X \in X_u} \ell(h(X), \varphi(X))$$

for *arbitrary bounded* loss function $\ell: \mathcal{Y} \times \mathcal{Y} \to [0, 1]$.

- $L_N(h)$ and $\hat{L}_n(h)$ are losses on $X_N$ and $X_n$ respectively;
- $\hat{h}_n$, $h_u^*$ and $h_N^*$ minimize $\hat{L}_n(h)$, $L_u(h)$ and $L_N(h)$ respectively;
- Excess loss

$$\mathcal{E}(\hat{h}_n) = L_u(\hat{h}_n) - L_u(h_u^*).$$

**Our goal:** obtain tight high-probability upper bounds on $\mathcal{E}(\hat{h}_n)$. 
Transductive learning: previous results

- [Vapnik, 1982; Blum and Langford, 2003] present an implicit bounds for binary loss function;

- [Cortes and Mohri, 2006] obtain bounds of order $\sqrt{\hat{L}_n(\hat{h}_n) \frac{\log(n+u)}{n}}$ for regression with quadratic loss;

- [Blum and Langford, 2003; Derbeko et al., 2004] PAC-Bayesian bounds for transductive learning which crucially depend on prior;

- [El-Yaniv and Pechyony, 2006; Cortes et al., 2009] Bounds of order $n^{-1/2}$ for binary and quadratic loss functions based on algorithmic stability;


- [Blum and Langford, 2003] provide bounds of the order $n^{-1}$ in the realizable setting (when $\varphi \in \mathcal{H}$) and binary loss function;

Message: all bounds have the “slow” rate $O(n^{-1/2})$ under general assumptions
Localized complexities and fast rates in inductive setting

**Inductive setting** assumes that $S_n \sim \text{i.i.d.}$ from unknown $P$ on $\mathcal{X} \times \mathcal{Y}$.

Classic VC-approach deals with uniform deviations:

$$\sup_{h \in \mathcal{H}} L_N(h) - \hat{L}_n(h)$$

and provides bounds of the slow rate of $O(n^{-1/2})$.

**Localized approach**
[Massart, 2000; Bartlett et al., 2005; Koltchinskii, 2006]

this is overpessimistic and we should study local fluctuations:

$$\sup_{h \in \mathcal{H}'} L_N(h) - \hat{L}_n(h),$$

where $\mathcal{H}' \subseteq \mathcal{H}$ contains functions with small variances. This often leads to the fast rates of $o(n^{-1/2})$ (e.g. Tsybakov’s low noise conditions, etc.).

**Important:** Localized approach is based on the Talagrand’s inequality.
Localized complexities and fast rates in inductive setting

**Inductive setting** assumes that $S_n \sim \text{i.i.d.}$ from unknown $P$ on $\mathcal{X} \times \mathcal{Y}$.

Classic VC-approach deals with uniform deviations:

$$
\sup_{h \in \mathcal{H}} L_N(h) - \hat{L}_n(h)
$$

and provides bounds of the slow rate of $O(n^{-1/2})$.

**Localized approach**

[Massart, 2000; Bartlett et al., 2005; Koltchinskii, 2006]

this is overpessimistic and we should study local fluctuations:

$$
\sup_{h \in \mathcal{H}'} L_N(h) - \hat{L}_n(h),
$$

where $\mathcal{H}' \subseteq \mathcal{H}$ contains functions with small variances. This often leads to the fast rates of $o(n^{-1/2})$ (e.g. Tsybakov’s low noise conditions, etc.).

**Important**: Localized approach is based on the Talagrand’s inequality.
Our results

Let \( \hat{L}_{n}^{iid}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell_{h}(Z_i) \), where \( Z_1, \ldots, Z_n \sim \text{i.i.d.} \) from \( X_N \).

Consider local neighbourhood of \( h_N^* \) in \( \mathcal{H} \):

\[
\mathcal{H}(r) = \left\{ h \in \mathcal{H} : \mathbb{E} \left[ (\ell_{h}(X) - \ell_{h_N^*}(X))^2 \right] \leq r \right\}.
\]

Theorem

Assume that there is a constant \( B > 0 \) such that for every \( h \in \mathcal{H} \):

\[
\mathbb{E} \left[ (\ell_{h}(X) - \ell_{h_N^*}(X))^2 \right] \leq B \cdot (L_N(h) - L_N(h_N^*)).
\]

Assume that there is a sub-root function \( \psi_n(r) \), such that:

\[
B \cdot \mathbb{E} \left[ \sup_{h \in \mathcal{H}(r)} L_N(h) - \hat{L}_{n}^{iid}(h) - (L_N(h_N^*) - \hat{L}_{n}^{iid}(h_N^*)) \right] \leq \psi_n(r).
\]

Let \( r_n^* \) be a fixed point of \( \psi_n(r) \). Then with prob. greater than \( 1 - \delta \):

\[
L_N(\hat{h}_n) - L_N(h_N^*) \leq 901 \frac{r_n^*}{B} + (16 + 25B) \frac{\log(1/\delta)}{3n} = \Delta_n(\delta).
\]
Our results

Theorem

Under assumptions of the previous theorem with prob. greater than \(1 - \delta\):

\[
L_u(\hat{h}_n) - L_u(h_u^*) \leq N \left( \frac{\Delta_n(\delta)}{u} + \frac{\Delta_u(\delta)}{n} \right), \quad \Delta_n(\delta) \sim r_n^* + n^{-1}.
\]

\[
\mathbb{E} \left[ (\ell_h(X) - \ell_{h_N^*}(X))^2 \right] \leq B \cdot (L_N(h) - L_N(h_N^*)).
\]

This condition is satisfied for many problems including:

- quadratic loss and uniformly bounded convex class \(\mathcal{H}\);
- binary loss and a class \(\mathcal{H}\) with finite VC-dimension if \(\varphi \in \mathcal{H}\).

For many interesting situations \(r_n^*\) is of the order \(o(n^{-1/2})\):

- [Massart, 2000] binary loss and VC-classes: \(r_n^* \sim \frac{\text{VC}(\mathcal{H}) \log n}{n}\).
Thank you for attention!

Many open questions:

- Can we “close the gap” in concentration inequalities?

- Can we obtain the tighter version of Talagrand’s inequality? (In the way Serfling’s bound tightens Hoeffding’s inequality)

- Local transductive Rademacher complexities.

- Other applications: non-asymptotic analysis of cross-validation, …

- Can we obtain transductive bounds useful in practice?

- …