New concentration inequalities for sampling without replacement and an application to transductive learning

Ilya Tolstikhin*
joint work with Gilles Blanchard and Marius Kloft

*Russian Academy of Sciences
iliya.tolstikhin@gmail.com

May 2014

Contents:

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

Concentration inequalities

- Function of many variables $g\colon \mathcal{X}^n \to \mathbb{R}$
- Sequence $X_1, \ldots, X_n$ of random variables taking values in $\mathcal{X}$
- We want to control random fluctuations of $Q = g(X_1, \ldots, X_n)$ around its expected value $\mathbb{E}[Q]$

We want to upper bound:

$$\mathbb{P}\{Q \geq \mathbb{E}[Q] + t\} \quad \text{and} \quad \mathbb{P}\{Q \leq \mathbb{E}[Q] - t\}$$

for $t > 0$. Or, equivalently, we want high-probability upper bounds on:

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

Independent random variables

The case when $X_1, \ldots, X_n$ are independent is very well studied and many useful results are available [Boucheron et al., 2013].

Concentration inequalities

- Function of many variables $g\colon \mathcal{X}^n \to \mathbb{R}$
- Sequence $X_1, \ldots, X_n$ of random variables taking values in $\mathcal{X}$
- We want to control random fluctuations of $Q = g(X_1, \ldots, X_n)$ around its expected value $\mathbb{E}[Q]$

We want to upper bound:

$$\mathbb{P}\{Q \geq \mathbb{E}[Q] + t\} \quad \text{and} \quad \mathbb{P}\{Q \leq \mathbb{E}[Q] - t\}$$

for $t > 0$. Or, equivalently, we want high-probability upper bounds on:

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

Independent random variables

The case when $X_1, \ldots, X_n$ are independent is very well studied and many useful results are available [Boucheron et al., 2013].

Concentration inequalities

- Function of many variables $g\colon \mathcal{X}^n \to \mathbb{R}$
- Sequence $X_1, \ldots, X_n$ of random variables taking values in $\mathcal{X}$
- We want to control random fluctuations of $Q = g(X_1, \ldots, X_n)$ around its expected value $\mathbb{E}[Q]$

We want to upper bound:

$$\mathbb{P}\{Q \geq \mathbb{E}[Q] + t\} \quad \text{and} \quad \mathbb{P}\{Q \leq \mathbb{E}[Q] - t\}$$

for $t > 0$. Or, equivalently, we want high-probability upper bounds on:

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

Independent random variables

The case when $X_1, \ldots, X_n$ are independent is very well studied and many useful results are available [Boucheron et al., 2013].

Concentration inequalities: independent random variables

Consider the normalized sum of independent and bounded random variables $X_1, \ldots, X_n$, such that $X_i \in [0,1]$, $i = 1, \ldots, n$, almost surely:

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

Hoeffding's inequality : for any $\delta \in [0,1]$ with prob. greater than $1 - \delta$:

$$Q - \mathbb{E}[Q] \leq \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Bernstein's inequality : for any $\delta \in [0,1]$ with prob. greater than $1 - \delta$:

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

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

Message: small variance leads to better convergence rates.

Concentration inequalities: independent random variables

Consider the normalized sum of independent and bounded random variables $X_1, \ldots, X_n$, such that $X_i \in [0,1]$, $i = 1, \ldots, n$, almost surely:

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

Hoeffding's inequality : for any $\delta \in [0,1]$ with prob. greater than $1 - \delta$:

$$Q - \mathbb{E}[Q] \leq \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Bernstein's inequality : for any $\delta \in [0,1]$ with prob. greater than $1 - \delta$:

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

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

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:

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

McDiarmid's inequality: for any $\delta \in [0, 1]$ with prob. greater than $1 - \delta$:

$$Q - \mathbb{E}[Q] \leq \sqrt{\frac{2 \log(1/\delta)}{n}}$$

Generalization of Hoeffding's inequality!

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

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

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

Generalization of Bernstein inequality!

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:

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

McDiarmid's inequality: for any $\delta \in [0, 1]$ with prob. greater than $1 - \delta$:

$$Q - \mathbb{E}[Q] \leq \sqrt{\frac{2 \log(1/\delta)}{n}}$$

Generalization of Hoeffding's inequality!

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

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

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

Generalization of Bernstein inequality!

Sampling without replacement

In many interesting situations the assumption that $X_1, \ldots, X_n$ are independent breaks.

Commonly used settings of non-i.i.d. random variables include martingale sequences, mixing processes, sampling without replacement...

Sampling without replacement setting:
$Z_1, \ldots, Z_n$ are sampled uniformly without replacement from some given finite set $\mathcal{C} = \{c_1, \ldots, c_N\}$ for $N \geq n$.

We want to obtain concentration inequalities for $Q = g(Z_1, \ldots, Z_n)$.

Motivation:

- Cross-validation procedures in machine learning
- Transductive setting of statistical learning theory
- ...

Sampling without replacement

In many interesting situations the assumption that $X_1, \ldots, X_n$ are independent breaks.

Commonly used settings of non-i.i.d. random variables include martingale sequences, mixing processes, sampling without replacement...

Sampling without replacement setting:
$Z_1, \ldots, Z_n$ are sampled uniformly without replacement from some given finite set $\mathcal{C} = \{c_1, \ldots, c_N\}$ for $N \geq n$.

We want to obtain concentration inequalities for $Q = g(Z_1, \ldots, Z_n)$.

Motivation:

- Cross-validation procedures in machine learning
- Transductive setting of statistical learning theory
- ...

Sampling without replacement

In many interesting situations the assumption that $X_1, \ldots, X_n$ are independent breaks.

Commonly used settings of non-i.i.d. random variables include martingale sequences, mixing processes, sampling without replacement...

Sampling without replacement setting:
$Z_1, \ldots, Z_n$ are sampled uniformly without replacement from some given finite set $\mathcal{C} = \{c_1, \ldots, c_N\}$ for $N \geq n$.

We want to obtain concentration inequalities for $Q = g(Z_1, \ldots, Z_n)$.

Motivation:

- Cross-validation procedures in machine learning
- Transductive setting of statistical learning theory
- ...

Sampling without replacement: stronger concentration!

Let $c_1, \ldots, c_N$ be bounded in $[0,1]$, $Z_1, \ldots, Z_n$ be sampled uniformly without replacement from $\{c_1, \ldots, c_N\}$ and consider the normalized sum:

$$Q = \frac{1}{n} \sum_{i=1}^{n} Z_i.$$

[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 $1-\delta$:

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

[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: stronger concentration!

Let $Z_1, \ldots, Z_n$ be sampled uniformly without replacement from $\mathcal{C} = \{c_1, \ldots, c_N\}$. Let $\mathcal{F}$ be a countable class of bounded functions $f : \mathcal{C} \to [-1, 1]$, such that $\mathbb{E}[f(Z_1)] = 0$. Consider the supremum of empirical process under sampling without replacement:

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

[El-Yaniv and Pechyony, 2009; Cortes et al., 2009]:
for any $\delta \in [0, 1]$ with prob. greater than $1 - \delta$:

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

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

This inequality is a (tighter) variant of McDiarmid's inequality.

Problem: there is no version of Talagrand's concentration inequality for sampling without replacement.

Sampling without replacement: stronger concentration!

Let $Z_1, \ldots, Z_n$ be sampled uniformly without replacement from $\mathcal{C} = \{c_1, \ldots, c_N\}$. Let $\mathcal{F}$ be a countable class of bounded functions $f : \mathcal{C} \to [-1, 1]$, such that $\mathbb{E}[f(Z_1)] = 0$. Consider the supremum of empirical process under sampling without replacement:

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

[El-Yaniv and Pechyony, 2009; Cortes et al., 2009]:
for any $\delta \in [0, 1]$ with prob. greater than $1 - \delta$:

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

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

This inequality is a (tighter) variant of McDiarmid's inequality.

Problem: there is no version of Talagrand's concentration inequality for sampling without replacement.

Sampling without replacement: stronger concentration!

Let $Z_1, \ldots, Z_n$ be sampled uniformly without replacement from $\mathcal{C} = \{c_1, \ldots, c_N\}$. Let $\mathcal{F}$ be a countable class of bounded functions $f : \mathcal{C} \to [-1, 1]$, such that $\mathbb{E}[f(Z_1)] = 0$. Consider the supremum of empirical process under sampling without replacement:

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

[El-Yaniv and Pechyony, 2009; Cortes et al., 2009]:
for any $\delta \in [0, 1]$ with prob. greater than $1 - \delta$:

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

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

This inequality is a (tighter) variant of McDiarmid's inequality.

Problem: there is no version of Talagrand's concentration inequality for sampling without replacement.

Main contributions

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

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

Theorem (Tolstikhin et al., 2014)

For any $\delta \in [0, 1]$ with prob. greater than $1 - \delta$:

$$Q - \mathbb{E}[Q] \leq 2\sqrt{\frac{2\sigma^2 \log(1/\delta)}{n} \left(\frac{N}{n}\right)}.$$

Theorem (Tolstikhin et al., 2014)

For any $\delta \in [0, 1]$ with prob. greater than $1 - \delta$:

$$Q - \mathbb{E}[Q'] \leq \sqrt{\frac{2(\sigma^2 + 2\mathbb{E}[Q']) \log(1/\delta)}{n}} + \frac{\log(1/\delta)}{3n}.$$

Main contributions

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

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

Theorem (Tolstikhin et al., 2014)

For any $\delta \in [0, 1]$ with prob. greater than $1 - \delta$:

$$Q - \mathbb{E}[Q] \leq 2\sqrt{\frac{2\sigma^2 \log(1/\delta)}{n} \left(\frac{N}{n}\right)}.$$

Theorem (Tolstikhin et al., 2014)

For any $\delta \in [0, 1]$ with prob. greater than $1 - \delta$:

$$Q - \mathbb{E}[Q'] \leq \sqrt{\frac{2(\sigma^2 + 2\mathbb{E}[Q']) \log(1/\delta)}{n}} + \frac{\log(1/\delta)}{3n}.$$

Main contributions: discussion

$$Q - \mathbb{E}[Q] \leq \sqrt{\frac{2\log(1/\delta)}{n}\left(\frac{N-m}{N-1/2}\right)}; \tag{EP}$$

$$Q - \mathbb{E}[Q] \leq 2\sqrt{\frac{2\sigma^2\log(1/\delta)}{n}\left(\frac{N}{n}\right)}; \tag{1}$$

$$Q - \mathbb{E}[Q'] \leq \sqrt{\frac{2(\sigma^2 + 2\mathbb{E}[Q'])\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{3n}. \tag{2}$$

- (EP) does not account for the variance (Hoeffding-type)
- (EP) outperforms (1) if $n = o(N)$ (which roughly means i.i.d.)
- (1) outperforms (EP) if $n = \Omega(N)$ and $\sigma^2 \leq 1/16$.
- Upper bound of (2) equals the upper bound of Bousquet's inequality
- Comparison between (2) and (EP) depends on σ^2
- $\mathbb{E}[Q]$ is always smaller than $\mathbb{E}[Q']$ and $\mathbb{E}[Q'] - \mathbb{E}[Q] \leq 2m^3/N$

Summarizing: All the results for Q' also hold for Q. But (1) can sometimes give even better results.

Main contributions: discussion

$$Q - \mathbb{E}[Q] \leq \sqrt{\frac{2\log(1/\delta)}{n}\left(\frac{N-m}{N-1/2}\right)}; \tag{EP}$$

$$Q - \mathbb{E}[Q] \leq 2\sqrt{\frac{2\sigma^2\log(1/\delta)}{n}\left(\frac{N}{n}\right)}; \tag{1}$$

$$Q - \mathbb{E}[Q'] \leq \sqrt{\frac{2(\sigma^2 + 2\mathbb{E}[Q'])\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{3n}. \tag{2}$$

- (EP) does not account for the variance (Hoeffding-type)
- (EP) outperforms (1) if $n = o(N)$ (which roughly means i.i.d.)
- (1) outperforms (EP) if $n = \Omega(N)$ and $\sigma^2 \leq 1/16$.
- Upper bound of (2) equals the upper bound of Bousquet's inequality
- Comparison between (2) and (EP) depends on σ^2
- $\mathbb{E}[Q]$ is always smaller than $\mathbb{E}[Q']$ and $\mathbb{E}[Q'] - \mathbb{E}[Q] \leq 2m^3/N$

Summarizing: All the results for Q' also hold for Q. But (1) can sometimes give even better results.

Main contributions: discussion

$$Q - \mathbb{E}[Q] \leq \sqrt{\frac{2\log(1/\delta)}{n}\left(\frac{N-m}{N-1/2}\right)}; \tag{EP}$$

$$Q - \mathbb{E}[Q] \leq 2\sqrt{\frac{2\sigma^2\log(1/\delta)}{n}\left(\frac{N}{n}\right)}; \tag{1}$$

$$Q - \mathbb{E}[Q'] \leq \sqrt{\frac{2(\sigma^2 + 2\mathbb{E}[Q'])\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{3n}. \tag{2}$$

- (EP) does not account for the variance (Hoeffding-type)
- (EP) outperforms (1) if $n = o(N)$ (which roughly means i.i.d.)
- (1) outperforms (EP) if $n = \Omega(N)$ and $\sigma^2 \leq 1/16$.
- Upper bound of (2) equals the upper bound of Bousquet's inequality
- Comparison between (2) and (EP) depends on σ^2
- $\mathbb{E}[Q]$ is always smaller than $\mathbb{E}[Q']$ and $\mathbb{E}[Q'] - \mathbb{E}[Q] \leq 2m^3/N$

Summarizing: All the results for Q' also hold for Q. But (1) can sometimes give even better results.

Contents:

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

Transductive learning: setting and notations

- Given set of N input points $\boldsymbol{X}_N = \{X_1, \ldots, X_N\} \subseteq \mathcal{X}$ sample $n \leq N$ objects $\boldsymbol{X}_n \subseteq \boldsymbol{X}_N$ uniformly without replacement;
- Obtain answers $\boldsymbol{Y}_n = \{Y_1, \ldots, Y_n\}$ for $\boldsymbol{X}_n$ by sampling for each input $X \in \boldsymbol{X}_n$ an output $Y \in \mathcal{Y}$ from unknown distribution $P(Y|X)$
- Denote the training set $S_n = (\boldsymbol{X}_n, \boldsymbol{Y}_n)$ and unlabelled test set $\boldsymbol{X}_u = \boldsymbol{X}_N \setminus \boldsymbol{X}_u$, where $u = N - n$;
- Consider fixed hypothesis class $\mathcal{H}$ of predictors $h\colon \mathcal{X} \to \mathcal{Y}$.
- Deterministic agnostic setting: there exists $\varphi\colon \mathcal{X} \to \mathcal{Y}$ such that:

$$P\big(Y = \varphi(x) | X = x\big) = 1,$$

  but φ does not necessary belong to $\mathcal{H}$.

Main goal: find a predictor in $\mathcal{H}$ based on both S_n and $\boldsymbol{X}_u$ with minimal error on test set.

Transductive learning: excess risk bounds

For loss function $\ell\colon \mathcal{Y} \times \mathcal{Y} \to [0,1]$ denote $\ell_h(X) = \ell\big(h(X), \varphi(X)\big)$ and

$$\hat{L}_n(h) = \frac{1}{n} \sum_{X \in \boldsymbol{X}_n} \ell_h(X), \quad L_u(h) = \frac{1}{u} \sum_{X \in \boldsymbol{X}_u} \ell_h(X).$$

We want to minimize $L_u(h)$ on $\mathcal{H}$ but instead we minimize $\hat{L}_n(h)$:

$$\hat{h}_n = \arg\min_{h \in \mathcal{H}} \hat{L}_n(h), \quad h_u^* = \arg\min_{h \in \mathcal{H}} L_u(h).$$

We also define risk on the whole sample $\boldsymbol{X}_N$ and its minimizer:

$$L_N(h) = \frac{1}{N} \sum_{X \in \boldsymbol{X}_N} \ell_h(X) \quad \text{and} \quad h_N^* = \arg\min_{h \in \mathcal{H}} L_N(h).$$

Excess risk of predictor $h \in \mathcal{H}$ is defined as:

$$\mathcal{E}_u(h) = L_u(h) - L_u(h_u^*) = L_u(h) - \inf_{g \in \mathcal{H}} L_u(g).$$

Excess risk $\mathcal{E}_u(\hat{h}_n)$ is a random quantity depending on training sample S_n and we want to obtain tight high-probability upper bounds on it.

Localized complexities and fast rates in inductive setting

More familiar inductive setting assumes that training examples are sampled i.i.d. from unknown distribution 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] show that 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.).

Localized approach is based on the Talagrand's inequality

Transductive learning: previous results

- [Vapnik, 1982; Blum and Langford, 2003] present an implicit bounds for binary loss function;
- [Blum and Langford, 2003] provide bounds of the order $\frac{1}{\min\{n,u\}}$ in the realizable setting (when $\varphi \in \mathcal{H}$) and binary loss function;
- [Cortes and Mohri, 2006] obtain bounds of order $\sqrt{\hat{L}_n(\hat{h}_n)\frac{\log N}{\min\{n,u\}}}$ 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 $\min\{n,u\}^{-1/2}$ for binary and quadratic loss functions based on algorithmic stability;
- [El-Yaniv and Pechyony, 2009] Bounds of order $\min\{n,u\}^{-1/2}$ for bounded loss functions based on global Rademacher complexities.

Problem: there are no excess risk bounds of fast rates $o(\min\{n,u\}^{-1/2})$ that hold without too restrictive assumptions

Transductive learning: previous results

- [Vapnik, 1982; Blum and Langford, 2003] present an implicit bounds for binary loss function;
- [Blum and Langford, 2003] provide bounds of the order $\frac{1}{\min\{n,u\}}$ in the realizable setting (when $\varphi \in \mathcal{H}$) and binary loss function;
- [Cortes and Mohri, 2006] obtain bounds of order $\sqrt{\hat{L}_n(\hat{h}_n)\frac{\log N}{\min\{n,u\}}}$ 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 $\min\{n,u\}^{-1/2}$ for binary and quadratic loss functions based on algorithmic stability;
- [El-Yaniv and Pechyony, 2009] Bounds of order $\min\{n,u\}^{-1/2}$ for bounded loss functions based on global Rademacher complexities.

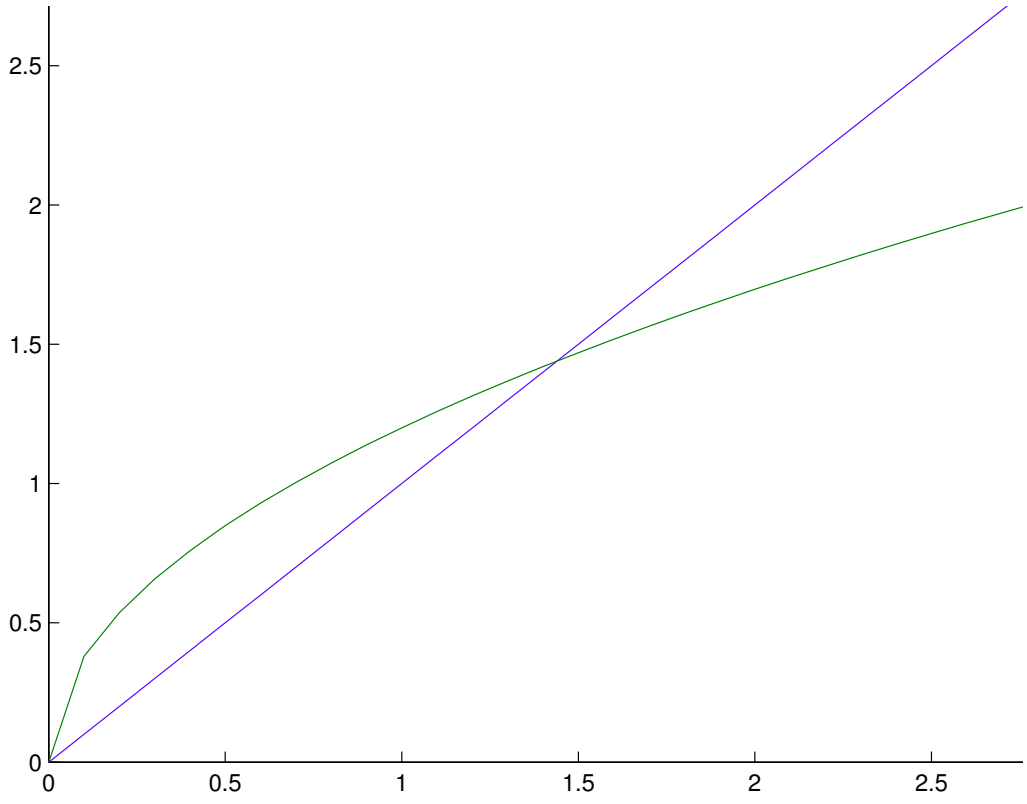
Problem: there are no excess risk bounds of fast rates $o(\min\{n,u\}^{-1/2})$ that hold without too restrictive assumptions

Transductive learning: main contributions

The function $\psi\colon [0,+\infty) \to [0,+\infty)$ is sub-root if

- it is nondecreasing,
- it is nonnegative,
- $r \to \psi(r)/\sqrt{r}$ is nonincreasing for $r > 0$.

Any sub-root function has unique fixed point.



Transductive learning: main contributions

Define

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

Let $\tilde{L}_n(h) = \frac{1}{n} \sum_{i=1}^{n} \ell_h(Z_i)$, where $Z_1, \ldots, Z_n$ are sampled uniformly with replacement from $\boldsymbol{X}_N$.

Theorem (Tolstikhin et al., 2014)

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

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

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

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

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

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

Transductive learning: main contributions

Define

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

Let $\tilde{L}_n(h) = \frac{1}{n} \sum_{i=1}^n \ell_h(Z_i)$, where $Z_1, \ldots, Z_n$ are sampled uniformly with replacement from $\boldsymbol{X}_N$.

Theorem (Tolstikhin et al., 2014)

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

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

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

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

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

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

Transductive learning: main contributions

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

Condition (*) is satisfied for:

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

Message: for many interesting situations r^*_n is of the order $o(n^{-1/2})$.

We also have the following excess risk bound:

Theorem (Tolstikhin et al., 2014)

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

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

$$\frac{N}{u}\left(901\frac{r^*_n}{B} + (16 + 25B)\frac{\log(1/\delta)}{3n}\right) + \frac{N}{n}\left(901\frac{r^*_u}{B} + (16 + 25B)\frac{\log(1/\delta)}{3u}\right).$$

Sanity check: consider the case $u = N - n \to N$ which brings us to the inductive setting. If $r^*_n = O(n^{-1})$ then $\mathcal{E}_u(\hat{h}_n) = O(n^{-1})$ as expected.

Transductive learning: main contributions

$$\mathbb{E}\left[\left(\ell_h(X) - \ell_{h^*_N}(X)\right)^2\right] \le B \cdot \left(L_N(h) - L_N(h^*_N)\right) \qquad (*)$$

Condition (*) is satisfied for:

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

Message: for many interesting situations r^*_n is of the order $o(n^{-1/2})$.

We also have the following excess risk bound:

Theorem (Tolstikhin et al., 2014)

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

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

$$\frac{N}{u}\left(901\frac{r^*_n}{B} + (16+25B)\frac{\log(1/\delta)}{3n}\right) + \frac{N}{n}\left(901\frac{r^*_u}{B} + (16+25B)\frac{\log(1/\delta)}{3u}\right).$$

Sanity check: consider the case $u = N - n \to N$ which brings us to the inductive setting. If $r^*_n = O(n^{-1})$ then $\mathcal{E}_u(\hat{h}_n) = O(n^{-1})$ as expected.

Transductive learning: main contributions

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

Condition (*) is satisfied for:

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

Message: for many interesting situations r^*_n is of the order $o(n^{-1/2})$.

We also have the following excess risk bound:

Theorem (Tolstikhin et al., 2014)

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

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

$$\frac{N}{u}\left(901\frac{r^*_n}{B} + (16 + 25B)\frac{\log(1/\delta)}{3n}\right) + \frac{N}{n}\left(901\frac{r^*_u}{B} + (16 + 25B)\frac{\log(1/\delta)}{3u}\right).$$

Sanity check: consider the case $u = N - n \to N$ which brings us to the inductive setting. If $r^*_n = O(n^{-1})$ then $\mathcal{E}_u(\hat{h}_n) = O(n^{-1})$ as expected.

Transductive learning: kernel classes

Consider:

- positive semidefinite kernel $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, such that $k(X, X) \leq 1$;
- $\mathcal{C}_k$ is the RKHS associated to k;
- hypothesis class $\mathcal{H} = \{f \in \mathcal{C}_k\colon \|f\|_k \leq 1\}$;
- K_N is a Gram matrix: $(K_N)_{ij} = \frac{1}{N} k(X_i, X_j)$ for $X_i, X_j \in \boldsymbol{X}_N$;
- $\lambda_{1,N} \geq \cdots \geq \lambda_{N,N}$ are ordered eigenvalues of K_N.

Theorem (Mendelson, 2002)

If loss function ℓ is L-Lipschitz in its first argument and there is $B > 0$:

$$\forall h \in \mathcal{H}: \quad \mathbb{E}\left[\left(\ell_h(X) - \ell_{h^*_N}(X)\right)^2\right] \leq B \cdot \left(L_N(h) - L_N(h^*_N)\right),$$

then

$$r^*_n \leq c_L \min_{0 \leq \theta \leq n} \left(\frac{\theta}{n} + \sqrt{\frac{1}{n} \sum_{i \geq \theta} \lambda_{i,N}} \right). \tag{1}$$

Message: r. h. s. of (1) is at most of the order $O(n^{-1/2})$ but can be much smaller if the decay is fast.

Thank you for attention!

Many open questions:

- Can we "close the gap" in our 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
- Can we obtain transductive bounds useful on practice?
- Other applications: non-asymptotic analysis of cross-validation
- ...