

Localized Complexities for Transductive Learning

Ilya Tolstikhin¹, Gilles Blanchard², Marius Kloft³

¹Russian Academy of Sciences, ²University of Potsdam, ³Humboldt University of Berlin

Short Summary

Concentration inequalities

- For **independent** random variables see overview in [BLM13]:

	1st order	2nd order
Sum	Hoeffding's	Bennett's Bernstein's
Supremum of empirical process	McDiarmid's	Talagrand's Bousquet's

- For random variables **sampled without replacement (SWOR)**:

	1st order	2nd order
Sum	Serfling's	[BM13]
Supremum of empirical process	[EYP09] [CMPR09]	(New results)

Application:

Local complexities (and fast rates) in statistical learning

- Inductive setting: [Mas00, BBM05, Kol06] and others.
- Transductive setting: **(New results)**

Concentration Inequalities: known results

Given:

random variables $\mathbf{X}_1^n = \{X_1, \dots, X_n\} \subset \mathcal{X}$

function $g: \mathcal{X}^n \rightarrow \mathbb{R}$

random variable $Q = g(X_1, \dots, X_n)$

Find: high-probability (over X_1, \dots, X_n) upper bounds on:

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

Sums of random variables

For random variables X_1, \dots, X_n bounded in $[0, 1]$ consider:

$$S_n = \frac{1}{n} \sum_{i=1}^n X_i, \quad \sigma^2 = \frac{1}{n} \sum_{i=1}^n \text{Var}[X_i]$$

$S_n - \mathbb{E}[S_n]$ and $\mathbb{E}[S_n] - S_n$ are upper bounded with prob. $\geq 1 - \delta$ by:

	\mathbf{X}_1^n are independent	\mathbf{X}_1^n are SWOR from $\{c_1, \dots, c_N\}$
1st order	Hoeffding $\sqrt{\frac{\ln(1/\delta)}{2n}}$	Serfling $\sqrt{\frac{\ln(1/\delta)}{2n} \left(\frac{N-n+1}{N} \right)}$
2nd order	Bernstein $\sqrt{\frac{2\sigma^2 \ln \frac{1}{\delta}}{n} + \frac{2 \ln(1/\delta)}{3n}}$	[BM13] $\sqrt{\frac{2\sigma^2 \ln \frac{1}{\delta}}{n} \left(\frac{N-n+1}{N} \right) + O(n^{-1})}$

Factor $\left(\frac{N-n+1}{N} \right)$ is a SWOR-specific improvement.

Suprema of empirical processes

For **identically distributed** random variables X_1, \dots, X_n and countable class \mathcal{F} of functions $f: \mathcal{X} \rightarrow [-1, 1]$, such that $\mathbb{E}[f(X_1)] = 0$, consider:

$$Q_n = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(X_i), \quad v_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \text{Var}[f(X_1)] + 2\mathbb{E}[Q_n]$$

$Q_n - \mathbb{E}[Q_n]$ and $\mathbb{E}[Q_n] - Q_n$ are upper bounded with prob. $\geq 1 - \delta$ by:

	\mathbf{X}_1^n are independent	\mathbf{X}_1^n are SWOR from $\{c_1, \dots, c_N\}$
1st order	McDiarmid $\sqrt{\frac{2 \ln(1/\delta)}{n}}$	[EYP09, CMPR09] $\sqrt{\frac{2 \ln(1/\delta)}{n} \left(\frac{N-n}{N-1/2} \right)}$ (EP)
2nd order	Talagrand $\sqrt{\frac{2v_{\mathcal{F}} \ln \frac{1}{\delta}}{n} + \frac{2 \ln(1/\delta)}{3n}}$ (only for $Q_n - \mathbb{E}[Q_n]$)	(New results): $2 \sqrt{\frac{2\sigma_{\mathcal{F}}^2 \log(1/\delta)}{n} \left(\frac{N}{n} \right)}$ $\sqrt{\frac{2(\sigma_{\mathcal{F}}^2 + 2\mathbb{E}[Q_n^{iid}]) \log \frac{1}{\delta}}{n} + O(n^{-1})}$

Factor $\left(\frac{N-n}{N-1/2} \right)$ is a SWOR-specific improvement.

Motivation

Applications of sampling without replacement

- Cross-validation procedures
- Transductive learning
- Low rank matrix factorization (collaborative filtering, ...)
- Randomized sequential algorithms (SGD, ...)

Motivation of our work

McDiarmid-type concentration inequality **(EP)** for SWOR does not account for the variance. There are situations when variance $\sim O(n^{-1/2})$.

New concentration inequalities for SWOR

Arbitrary finite set $\mathcal{C} = \{c_1, \dots, c_N\}$

Z_1, \dots, Z_n sampled uniformly **with replacement** from \mathcal{C}

X_1, \dots, X_n sampled uniformly **without replacement** from \mathcal{C}

Countable class \mathcal{F} of functions $f: \mathcal{C} \rightarrow [-1, 1]$, such that $\mathbb{E}[f(X_1)] = 0$

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

Theorem 1 ((Sub-Gaussian concentr. ineq. for SWOR)).

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

$$Q_n - \mathbb{E}[Q_n] \leq 2 \sqrt{\frac{2\sigma_{\mathcal{F}}^2 \log(1/\delta)}{n}} \cdot \left(\frac{N}{n} \right). \quad \text{(New1)}$$

The same bound also holds for $\mathbb{E}[Q_n] - Q_n$.

Theorem 2 ((Talagrand-type concentr. ineq. for SWOR)).

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

$$Q_n - \mathbb{E}[Q_n^{iid}] \leq \sqrt{\frac{2(\sigma_{\mathcal{F}}^2 + 2\mathbb{E}[Q_n^{iid}]) \log(1/\delta)}{n}} + \frac{\log(1/\delta)}{3n}. \quad \text{(New2)}$$

Comparison with State-of-the-art

- (EP)** does not account for $\sigma_{\mathcal{F}}^2$ while **(New1)** and **(New2)** do.
- Large N and $n = o(N)$: **(EP)** can outperform **(New1)**.
- Large N and $n = \Omega(N)$: **(New1)** outperforms **(EP)** for $\sigma_{\mathcal{F}}^2 < 1/16$.
- Shortcomings of the new results:

(New1): factor of N/n which can be large;

(New2): concentration not around $\mathbb{E}[Q_n]$ but around $\mathbb{E}[Q_n^{iid}]$.

Lemma.

$$0 \leq \mathbb{E}[Q_n^{iid}] - \mathbb{E}[Q_n] \leq \frac{n^3}{N}.$$

- However, often $\mathbb{E}[Q_n^{iid}] \sim O(n^{-1/2})$ (e.g. finite \mathcal{F} or VC-class \mathcal{F})

Summary:

(New1) stays informative in all regimes of n and N .

(New2) can be preferable for $n = \Omega(N)$.

Both **(New1)** and **(New2)** account for the variance.

Proof Idea

Chernoff's bounding method: for $\lambda > 0$

$$\mathbb{P}\{\xi \geq t\} = \mathbb{P}\{e^{\lambda \xi} \geq e^{\lambda t}\} \leq \frac{\mathbb{E}[e^{\lambda \xi}]}{e^{\lambda t}} \quad (1)$$

Theorem 3 ([Bob04], simplified version). Consider fixed finite set $\mathcal{C} = \{c_1, \dots, c_N\}$, a set $[\mathcal{C}]^n$ of all disjoint partitions $\mathcal{C} = \mathcal{C}' \cup \mathcal{C}''$ where $|\mathcal{C}'| = n$, and a function $g: [\mathcal{C}]^n \rightarrow \mathbb{R}$.

If Euclidian norm of the discrete gradient $\|\nabla g(\mathcal{C}' \cup \mathcal{C}'')\|_2^2$ is bounded by Σ^2 and $(U \cup V)$ is sampled uniformly from $[\mathcal{C}]^n$ then the following upper bound on the MGF for $Q = g(U \cup V)$ holds:

$$\mathbb{E}\left[e^{\lambda(Q - \mathbb{E}[Q])}\right] \leq \exp\left(\frac{\Sigma^2 \lambda^2}{N+2}\right), \quad \lambda \in \mathbb{R}. \quad (2)$$

Theorem 4 ([Hoe63]). Let V_1, \dots, V_n and U_1, \dots, U_n be sampled uniformly from a finite set $\{v_1, \dots, v_N\} \subset \mathbb{R}^d$ without and with replacements respectively. For any convex function $F: \mathbb{R}^d \rightarrow \mathbb{R}$:

$$\mathbb{E}[F(\sum_{i=1}^n V_i)] \leq \mathbb{E}[F(\sum_{i=1}^n U_i)]. \quad (3)$$

Proof sketch of Theorem 1: Use (1) for $\xi = Q_n - \mathbb{E}[Q_n^{iid}]$. Notice that $F: \mathbf{x} \rightarrow \exp(\sup_{i=1, \dots, d} x_i)$ for $\mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$ is convex. Apply (3) and upper bound MGF using Talagrand's inequality.

Proof sketch of Theorem 2: Note that Q_n is a function over partitions of \mathcal{C} . Apply (1) and upper bound MGF using (2).

Application: Transductive Learning

Deterministic agnostic setting

Finite instance space $\mathbf{X}_n = \{X_1, \dots, X_N\} \subset \mathcal{X}$ 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})

- Sample $n \leq N$ inputs $\mathbf{X}_n \subset \mathbf{X}_N$ uniformly **without replacement**
- Obtain outputs \mathbf{Y}_n for \mathbf{X}_n by applying function $\varphi: \mathcal{X} \rightarrow \mathcal{Y}$
- Reveal training set $S_n = (\mathbf{X}_n, \mathbf{Y}_n)$ and $u = N - n$ test inputs \mathbf{X}_u

Goal of the learner: based on S_n and \mathbf{X}_u find a predictor in hypothesis class $\mathcal{H} = \{h: \mathcal{X} \rightarrow \mathcal{Y}\}$ with minimal test error:

$$L_u(h) = \frac{1}{u} \sum_{X \in \mathbf{X}_u} \underbrace{\ell(h(X), \varphi(X))}_{\ell_h(X)}$$

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

Transductive learning: Notations

$L_N(h)$ and $\hat{L}_n(h)$ are errors over \mathbf{X}_N and \mathbf{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

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

measures how well ERM \hat{h}_n approximates optimal h_u^* on \mathbf{X}_u .

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

State-of-the-art transductive bounds

Algo. stability: $\sim 1/\sqrt{n}$

$$L_2\text{-loss: } \sim \sqrt{\hat{L}_n(\hat{h}_n) \frac{\log(n+u)}{n}}$$

Global Rademacher: $\sim 1/\sqrt{n}$

Binary loss and $\varphi \in \mathcal{H}$: $\sim 1/n$

Summary: Without restrictive assumptions all bounds $\sim 1/\sqrt{n}$

Localized transductive bounds

Let $\hat{L}_n^{iid}(h) = \frac{1}{n} \sum_{i=1}^n \ell_h(Z_i)$, where $Z_1, \dots, Z_n \sim \text{i.i.d.}$ from \mathbf{X}_N .

Consider the 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 5 ((Localized transductive excess risk bound)). Assume that there is a constant $B > 0$ such that for every $h \in \mathcal{H}$:

$$\mathbb{E} \left[(\ell_h(X) - \ell_{h_N^*}(X))^2 \right] \leq B \cdot (L_N(h) - L_N(h_N^*)). \quad \text{(C1)}$$

Assume 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 $1 - \delta$:

$$L_N(\hat{h}_n) - L_N(h_N^*) \leq 901 \frac{r_n^*}{B} + (16+25B) \frac{\log(1/\delta)}{3n} = \Delta_n(\delta). \quad \text{(New3)}$$

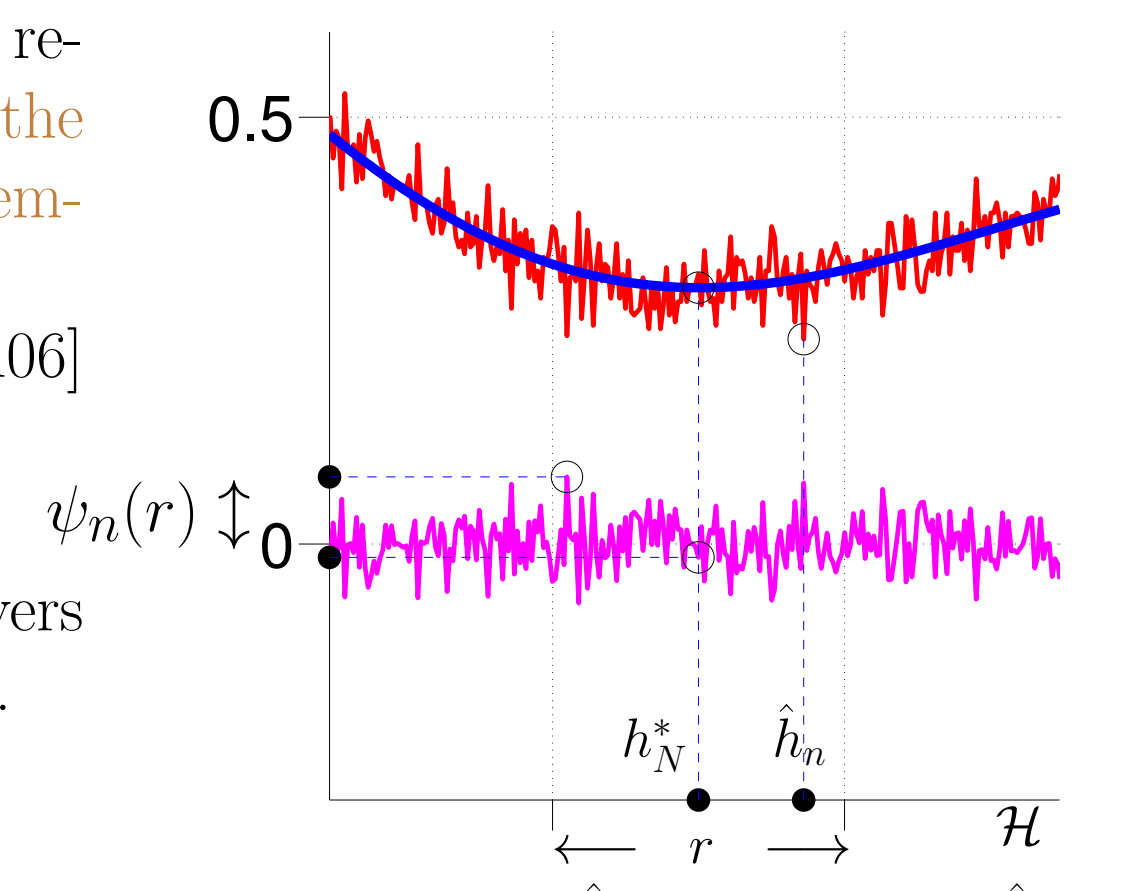
Also

$$\mathcal{E}_u(\hat{h}_n) \leq N \left(\frac{\Delta_n(\delta)}{u} + \frac{\Delta_u(\delta)}{n} \right). \quad \text{(New4)}$$

Discussion:

"Order of the excess risk is related to the fixed point r_n^* of the modulus of continuity of the empirical process". [Mas00, BBM05, Kol06]

Message: Theorem 5 recovers it in the transductive setting.



Condition **(C1)** is satisfied for:

- L_2 -loss and uniformly bounded convex class \mathcal{H} with $B = 1$;
- binary loss and VC-class \mathcal{H} if $\varphi \in \mathcal{H}$ with $B = 8$.

Important: Fixed point r_n^* can be of the order $\sim o(n^{-1/2})$:

- For binary loss and VC-classes: $r_n^* \sim \frac{\text{VC}(\mathcal{H}) \log n}{n}$ [Mas00].
- For Lipschitz losses and balls in RKHS [Men03].

Future Work

- Close the gap in concentration inequalities **(New1)** and **(New2)**.
- SWOR-specific improvements for Talagrand's inequality?
- Local transductive Rademacher complexities.
- Other applications: Cross-Validation, ...

References

- [BBM05] P. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497-1537, 2005.
- [BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.
- [BM13] Rémi Bardenet and Odalric-Ambrym Maillard. Concentration inequalities for sampling without replacement. <http://arxiv.org/abs/1309.4029>, 2013.
- [Bob04] S. Bobkov. Concentration of normalized sums and a central limit theorem for noncorrelated random variables. *Annals of Probability*, 32, 2004.
- [CMPR09] C. Cortes, M. Mohri, D. Pechony, and A. Rastogi. Stability analysis and learning bounds for transductive regression algorithms. <http://arxiv.org/abs/0904.0814>, 2009.
- [EYP09] Ran El-Yaniv and Dmitry Pechony. Transductive rademacher complexity and its applications. *Journal of Artificial Intelligence Research*, 2009.
- [Hoe63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13-30, 1963.
- [Kol06] V. Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593-2656, 2006.
- [Mas00] Pascal Massart. Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math.*, 9(6):245-303, 2000.
- [Men03] Shahar Mendelson. On the performance of kernel classes. *J. Mach. Learn. Res.*, 4:759-771, December 2003.