Sharp Bernstein inequality and applications to Machine Learning

Liming Wu

HIT and UCA

31 July 2023
in Conference on Markov Processes and Related Fields. Tianjin University

Outline

1. Introduction: empirical risk principle (ERP) in ML.
2. Probabilistic problems coming from the ERP: the curse of dimension (CoD)
3. Talagrand's concentration inequality: overcoming the curse of dimension (CoD)?
4. Best known bias estimate in terms of VC dimension
5. Bernstein's concentration inequality: old and some new results
6. Applications: we can verify quickly the non-efficiency of a learning machine.

1. Introduction: empirical risk principle (ERP) in ML.

We have two random variables (r.v.): X valued in a domain D of $\mathbb{R}^d$, real-valued r.v. Y; X is thought as cause, Y is the effect. The joint law $\mu(\cdot) = \mathbb{P}(Z \in \cdot)$ of $Z = (X, Y)$ is unknown. We want to know what is the "best" way to describe the dependence of Y upon X. To this purpose we dispose of a great sample of data

$$Z_1 = (X_1, Y_1), \cdots, Z_n = (X_n, Y_n)$$

assumed to be the independent copies of $Z = (X, Y)$.
Learning machines furnish a special class of functions

$$\mathcal{F} = \{f(x, \theta); \theta \in \Theta\}$$

to approximatively learn the the dependence of Y upon X, where $\Theta \subset \mathbb{R}^N$ is a domain of $\mathbb{R}^N$, N being the number of training parameters which is often very huge ($N \asymp 10^{11}$ for ChatGPT).

1. Introduction: ERP in ML (cont')

To describe what means the "best way", we are given a risk or loss function

$$Q(z, \theta) = (y - f(x, \theta))^2 \text{ or } |z - f(x, \theta)| \ \text{ or other forms,}$$

where $z = (x, y) \in D \times \mathbb{R}$. One main purpose of learning machines is to minimise the empirical risk function

$$R_{E,n}(\theta) = \frac{1}{n} \sum_{k=1}^{n} Q(Z_k, \theta) \tag{1}$$

among all $\theta \in \Theta$, i.e. to find the minimisers of

$$\arg\min_{\theta \in \Theta} R_{E,n}(\theta) = \{\hat{\theta}_n \in \Theta \mid R_{E,n}(\hat{\theta}_n) \leqslant R_{E,n}(\theta),\ \forall \theta \in \Theta\}. \tag{2}$$

(that is called "training the parameters" in machine learning).

1. Introduction: ERP in ML (cont')

When $Q(z, \theta) = (y - f(x, \theta))^2$, the theoretical risk of the learning machine for a given θ is

$$R(\theta) = \mathbb{E}(Y - f(X, \theta))^2 = \mathbb{E}(Y - f_0(X))^2 + \mathbb{E}(f_0(X) - f(X, \theta))^2$$

where $f_0(x) = \mathbb{E}(Y|X = x)$ is the conditional expectation, known as the non-linear regression function. Then the theoretical minimal risk of the learning machine is

$$\inf_{\theta \in \Theta} R(\theta) = \mathbb{E}(Y - f_0(X))^2 + \inf_{\theta \in \Theta} \mathbb{E}(f_0(X) - f(X, \theta))^2. \quad (3)$$

The first term at the right hand side (r.h.s.) can not be diminished by any learning machine (because of the "random" dependence assumption of Y upon X), and the least-square error

$$\inf_{\theta \in \Theta} \mathbb{E}(f_0(X) - f(X, \theta))^2.$$

qualifies the (theoretical optimal) efficiency of the learning machine.

Empirical Risk Principle (**ERP** in short), laid by

V.N. Vapnik: *The Nature of Statistical Learning Theory*, Second Edition. Springer 1999.

as a basic (starting) principle for statistical learning theory, means roughly

$$\begin{aligned} p_+(n,\varepsilon) &:= \mathbb{P}\left(\inf_{\theta\in\Theta} R_{E,n}(\theta) < \inf_{\theta\in\Theta} R(\theta) - \varepsilon\right) \\ p_-(n,\varepsilon) &:= \mathbb{P}\left(\inf_{\theta\in\Theta} R_{E,n}(\theta) > \inf_{\theta\in\Theta} R(\theta) + \varepsilon\right) \end{aligned} \tag{4}$$

go both to zero for any $\varepsilon > 0$. That is a consequence of the Glivenko-Cantelli theorem about the (uniform) law of large number in empirical processes.

When $|Q(z,\theta)| \leqslant M$ is bounded, a necessary and sufficient condition for the Glivenko-Cantelli theorem in terms of the VC entropy number is known ([14, §2.3.4, Theorem 2.3]).

On the other hand, if $Q(z,\theta)$ is continuous in θ and Θ is compact, ERP holds.

2. Probabilistic problems comed from the ERP

1. The first error probability $p_+(n, \varepsilon)$ gives an upper bound of the theoretical minimal risk:

$$\inf_{\theta \in \Theta} R(\theta) \leqslant \inf_{\theta \in \Theta} R_{E,n}(\theta) + \varepsilon$$

   with probability $1 - p_+(n, \varepsilon)$ (the so called confidence level),

2. whereas the second error probability $p_-(n, \varepsilon)$ gives a lower bound of the theoretical minimal risk:

$$\inf_{\theta \in \Theta} R(\theta) \geqslant \inf_{\theta \in \Theta} R_{E,n}(\theta) - \varepsilon$$

   with probability $1 - p_-(n, \varepsilon)$.

In other words, $p_+(n, \varepsilon)$ quantifies how good a leaning machine is; $p_-(n, \varepsilon)$ quantifies the non-efficiency of a leaning machine.

2. Probabilistic problems comed from the ERP: the curse of dimension (CoD)

Estimating the above two error probabilities is then a fundamental question in machine learning.
At first the classical limit theorems such as Donsker's central limit theorem (or invariance principle, see [9], [10]), the large and moderate deviation principles (W. [17] (94); R. Wang et al. [16](10)), which are only asymptotic (when $n \to +\infty$), **can not** be applied directly, because the disposed sample size n can not be much bigger than the number N of parameters, and the dimension d of the input vector X is often very high (256×256 pixels for a picture for example).

2. Probabilistic problems comed from the ERP: CoD (cont'd)

Recent progresses in high dimensional probability show that the error probabilities depend often on the dimension d and the number N of parameters, see Fournier and Guillin [5] (2015) for the dimension dependence:

$$W_1(L_n, \mu) \asymp \frac{1}{n^{1/d}}, \ d \geqslant 3;$$

and the recent book in preparation [15] (2020) by Vershynin for an account of art for the dependence on N, d.
See the works of F.Y. Wang and his collaborators for the Wasserstein distance between the empirical distribution and its stationary distribution of diffusions.

Conclusion : Wasserstein distance is too sensible to the dimension d, it gives rise to the CoD.

2. Probabilistic problems comed from the ERP: overcoming the curse of dimension (CoD)?

The whole book of Vapnik is to show that $p_+(n, \varepsilon) \to 0$ with an explicit concentration inequality in terms of VC dimension or VC entropy number.

His results together with recent developments in approximation theory of the neural network:

1. Approximation theory: for deterministic dependence $Y = f(X)$ and for neural network,

$$R_{min} := \inf_{\theta \in \Theta} \mathbb{E}|f(X) - f(X, \theta)| \to 0$$

   if the neural network is sufficiently wide or depth: the number of units is large enough.
   See E (ICM2020) .
2. For 1-layer neural network, $p_+(\varepsilon)$ may be small, even for large N and d, but not so great (Vapnik [14]).

Those 2 demands are contradictory!

No word about $p_-(n, \varepsilon)$ in Vapnik [14] !

3. Talagrand's concentration inequality: overcoming the curse of dimension (CoD)?

Talagrand ([11, 12, 13], 94AOP, 95IHES, 96Inv.Math.) investigated in depth the concentration phenomena on product measure spaces and renewed the theory of empirical processes. Massart [8] (AOP00) found explicit constants in Talagrand's concentration inequality, by refining the log-Sobolev inequality approach of Ledoux.

Theorem 1

Given

1. *a sequence of i.i.d.r.v.* $(\xi_k)_{k\geqslant 1}$ *valued in some Polish space* S *equipped with the Borel* σ*-field, of common law* μ*;*
2. *an at most countable class* $\mathcal{H}$ *of bounded measurable functions* h *on* S *such that* $|h| \leqslant b$*;*

3. Talagrand's concentration inequality: overcoming the curse of dimension (CoD)?

let

$$Z = \sup_{h \in \mathcal{H}} |\frac{1}{n} \sum_{k=1}^{n} (h(\xi_k) - \mu(h))| \ (\mu(h) := \int_S h d\mu = \mathbb{E} h(\xi_1))$$

and

$$\sigma^2(\mathcal{H}) = \sup_{h \in \mathcal{H}} \mathrm{Var}_\mu(h). \tag{5}$$

Then for any $\varepsilon > 0$,

$$\mathbb{P}\left(Z > (1+\varepsilon)\mathbb{E}Z + \sigma(\mathcal{H})\sqrt{\frac{8x}{n}} + \kappa(\varepsilon)\frac{bx}{n} \right) \leqslant e^{-x}, \ \forall x > 0 \tag{6}$$

where $\kappa(\varepsilon) = 2.5 + \frac{32}{\varepsilon}$.

3. Talagrand's concentration inequality: overcoming the curse of dimension (CoD)?

Applying it to $\mathcal{H} = \{Q(z,\theta);\ \theta \in \Theta\}$ we get

$$\begin{aligned}\max\{p_+(\varepsilon), p_-(\varepsilon)\} &\leqslant \mathbb{P}\left(|\inf_{\theta\in\Theta} R_{E,n}(\theta) - \inf_{\theta\in\Theta} R(\theta)| > \varepsilon\right) \\ &\leqslant \mathbb{P}\left(\sup_{\theta\in\Theta} |R_{E,n}(\theta) - R(\theta)| > \varepsilon\right) \\ &\leqslant e^{-x},\ x > 0\end{aligned}$$

where

$$\varepsilon = (1+\delta)\mathbb{E}\sup_{\theta\in\Theta}|R_{E,n}(\theta) - R(\theta)| + \sigma(\mathcal{H})\sqrt{\frac{8x}{n}} + \kappa(\delta)\frac{bx}{n} \quad (7)$$

for an arbitrary $\delta > 0$. Except the bias

$$\mathbb{E}\sup_{\theta\in\Theta}|R_{E,n}(\theta) - R(\theta)|,$$

the dependence on N, d are only through the maximal variance $\sigma^2(\mathcal{H})$!

4. Best known bias estimate in terms of VC dimension

Theorem 2

(Vershynin [15, Theorem 8.3.23]) *Assume that* $a \leqslant h \leqslant b$ *for all* $h \in \mathcal{H}$ *and the VC dimension* $\mathrm{vc}(\mathcal{H})$ *of* $\mathcal{H}$ *is finite (the so called VC class). Then*

$$\mathbb{E} \sup_{\mathcal{H}} (L_n(h), \mu(h)) \leqslant K \sqrt{\frac{\mathrm{vc}(\mathcal{H})}{n}} (b - a). \tag{8}$$

where $K > 0$ *is an absolute constant, and*

$$L_n = \frac{1}{n} \sum_{k=1}^{n} \delta_{Z_k}.$$

4. Best known bias estimate in terms of VC dimension (cont'd)

Conclusion: with probability $\alpha \in (0,5,1)$, for $x = -\log(1-\alpha)$,

$$|\inf_\theta R(\theta) - \inf_\theta R_{E,n}(\theta)| \leqslant \left(K\sqrt{\frac{\mathrm{vc}(\mathcal{H})}{n}} + \sigma(\mathcal{H})\sqrt{\frac{8x}{n}} + \kappa(\delta)\frac{x}{2n} \right)(b-a) \tag{9}$$

In other words, if the sample size $n \gg \mathrm{vc}(\mathcal{H})$, the empirical minimal risk $\inf_\theta R_{E,n}(\theta)$ of the learning machine attains the theoretical minimal risk $\inf_\theta R_E(\theta)$.
The dimension-dependence on d, N is transformed into that on the VC dimension $\mathrm{vc}(\mathcal{H})$ of the learning machine.

5. Sharp Bernstein's concentration inequality

Purposes of this talk:

1. A dimension-free estimate for $p_{-}(n, \varepsilon)$
2. removing the boundedness assumption.

5.1. Bernstein's concentration inequality: some known results

Theorem 3

(Gozlan-Léonard [7]) *Given the constants* $c_B > 0, M \geqslant 0$ *and a* μ*-exponentially integrable function* h *on* $S(= D \times \mathbb{R})$*, i.e.*

$$\exists \delta > 0 : \int_S e^{\delta |f|} d\mu < +\infty, \tag{10}$$

the following properties are equivalent:

(1) *The log-Laplace transform of* $h(Z)$ *satisfies*

$$\Lambda(\lambda) := \log \mathbb{E} e^{\lambda[h(Z) - \mu(h)]} \leqslant \frac{c_B \lambda^2}{2(1 - \lambda M)}, \ \lambda \in (0, 1/M); \tag{11}$$

5.1. Bernstein's concentration inequality: some known results

(2) *for any* $r > 0$ *and* $n \geqslant 1$,

$$\mathbb{P}\left(L_n(f) - \mu(f) > r\right) \leqslant \exp\left(-n\frac{2r^2}{c_B\left(\sqrt{1+\frac{2Mr}{c_B}}+1\right)^2}\right); \tag{12}$$

(3) *for any* $x > 0$ *and* $n \geqslant 1$,

$$\mathbb{P}\left(L_n(f) - \mu(f) > \sqrt{\frac{2c_B x}{n}} + M\frac{x}{n}\right) \leqslant e^{-x}; \tag{13}$$

(4) *the following transport-entropy inequality holds:*

$$\nu(f) - \mu(f) \leqslant \sqrt{2c_B H(\nu|\mu)} + MH(\nu|\mu), \tag{14}$$

for all $\nu \in M_1(S)$ *such that* $\nu \ll \mu$ *and* $\nu(|f|) < +\infty$.

5.1. Bernstein's concentration inequality: some known results

In particular, when (11) *holds, then the following Bernstein's concentration inequality holds:*

$$\mathbb{P}\left(L_n(f) - \mu(f) > r\right) \leqslant \exp\left(-\frac{nr^2}{2(c_B + Mr)}\right), \ r > 0. \qquad (15)$$

Remarks:

(1) By the order 2 limit expansion of Taylor-Young at $\lambda = 0+$ in (11), we see that the Bernstein concentration constant c_B satisfies

$$c_B \geqslant \mathrm{Var}_\mu(f) = \mu(f^2) - (\mu(f))^2. \qquad (16)$$

5.1. Bernstein's concentration inequality: some known results

(2) Two sided (for both $\pm f$) Bernstein's concentration inequality holds for some c_B, M if and only if $f(Z)$ is exponentially integrable:

$$\|f\|_{\psi_1} := \inf\{C > 0; \int_S (e^{|f|/C} - 1) d\mu \leqslant 1\} < +\infty$$

(Orlicz norm in $L^{\psi_1}(\mu)$, $\psi_1(x) := e^{|x|} - 1$).

Theorem 4

(Classical) *If f is upper bounded and μ-square integrable,* (11) *holds with*

$$c_B = \mathrm{Var}_\mu(f),\ M = \frac{1}{3}\|(f - \mu(f))^+\|_\infty. \tag{17}$$

5.1. Bernstein's concentration inequality: some known results

We recall its proof.

Proof.

We may assume that $\mu(f) = 0$ and $f \leqslant 1$. By the inequality

$$e^x \leqslant 1 + x + \frac{x^2}{2} \cdot \frac{1}{1 - x^+/3}, \ \ x < 3$$

we have for all $\lambda \in (0, 3)$,

$$\mathbb{E}e^{\lambda f} \leqslant 1 + \lambda\mu(f) + \frac{\lambda^2 \mu(f^2)}{2(1 - \lambda/3)} \leqslant \exp\left(\frac{\lambda^2 \mu(f^2)}{2(1 - \lambda/3)}\right).$$

That is (11) with c_B, M given in (17). □

5.1. Bernstein's concentration inequality: some known results

Theorem 5

(Bolley-Villani 04, Gozlan-Léonard 07) If f is μ-exponentially integrable, then the transport-entropy inequality (14) *holds with*

$$c_B = 2\|f\|_{\psi_1}^2,\ M = \|f\|_{\psi_1}. \tag{18}$$

5.2. Bernstein's concentration inequality: new results

Theorem 6

(under Gaussian integrabity) *If f is of Gaussian integrability:*

$$\exists\delta > 0: \ \mathbb{E}e^{\delta f(X)^2} = \int_S e^{\delta f^2(x)}\mu(dx) < +\infty$$

then for any $\varepsilon \in (0, \delta)$, (13) *holds with*

$$c_B = \mathrm{Var}_\mu(f) + \frac{1}{3}L(\varepsilon), \ M = \sqrt{\frac{2}{3\varepsilon}}, \tag{19}$$

where

$$L(\varepsilon) = \frac{1}{\varepsilon}\log\int_S e^{\varepsilon(\tilde{f}^2 - \mu(\tilde{f}^2))}d\mu, \ \tilde{f} := f - \mu(f) \tag{20}$$

satisfies for all $0 < \varepsilon < \frac{1}{\|\tilde{f}^2 - \mu(\tilde{f}^2)\|_{\psi_1}}$,

$$L(\varepsilon) \leqslant \frac{\varepsilon \|\tilde{f}^2 - \mu(\tilde{f}^2)\|_{\psi_1}^2}{1 - \varepsilon \cdot \|\tilde{f}^2 - \mu(\tilde{f}^2)\|_{\psi_1}}. \tag{21}$$

Moreover for all $n \geqslant 1, x > 0$ *such that* $0 < \frac{x}{n} \leqslant \frac{1}{2}$, *we have*

$$\mathbb{P}\left(L_n(f) - \mu(f) > \sqrt{2\mathrm{Var}_\mu(f)\frac{x}{n}} + \sqrt{\frac{\sqrt{2}\|\tilde{f}^2 - \mu(\tilde{f}^2)\|_{\psi_1}}{3}} \left(\frac{x}{n}\right)^{3/4} \right) \leqslant e^{-x}. \tag{22}$$

5.2. Bernstein's concentration inequality: new results

Theorem 7

(under one-sided exponential integrabity) *If* $f \in L^2(S,\mu)$ *and the positive part* $f^+(x) = \max\{f(x), 0\}$ *is exponentially integrable, i.e.*

$$\exists \delta > 0: \ \mathbb{E}e^{\delta f^+(X)} = \int_S e^{\delta f^+(x)} \mu(dx) < +\infty,$$

then for any $L > 0$, *setting* $\varepsilon(L) := \|f - f \wedge L\|_{\psi_1}$, *the Bernstein's concentration inequality* (13) *holds with*

$$\begin{cases} c_B & = \mathrm{Var}_\mu(f) + 2\varepsilon(L)\sqrt{2\mathrm{Var}_\mu(f)} + 2\varepsilon(L)^2 \\ & \leqslant (1+\varepsilon(L))\mathrm{Var}_\mu(f) + 2(\varepsilon(L) + \varepsilon^2(L)) \\ M & = \frac{L}{3} + \varepsilon(L). \end{cases} \tag{23}$$

5.2. Bernstein's concentration inequality: ideas of proof

1. some technique from large deviations
2. Transport-entropy inequality, refining the arguments of Bolley-Villani [1] for our purpose.
3. Based on the known results recalled before

Comments on other concentrationn inequalities

1. Hoeffding's gaussian concentration inequality (corresponding to $M = 0$ in Bernstein's inequality) is equivalent to the Gaussian integrability (Djellout et al. AOP04)
2. Bernstein's inequality is not sharp for large deviations: finer estimate in this range was found by Fan-Grama-Liu [3, 4]
3. Classical asymptotic edge-expansion in moderate deviations: Cramèr, Bahadur-Rao, ...
4. Comparison with the Gaussian distribution...
5. For continuous-time symmetric Markov processes satisfying the log-Sobolev or transport inequalities, this was proved by Gao et al. [6] (SIAM 14).

6. Applications: we can verify quickly the non-efficiency of a learning machine.

Theorem 8

Assume the Gaussian integrability of the loss function $Q(z, \theta)$. For all $n \geqslant 1$ and $0 < x < \frac{n}{2}$,

$$p_-(n, \varepsilon) := \mathbb{P}\left(\inf_{\theta \in \Theta} R_{E,n}(\theta) > \inf_{\theta \in \Theta} R(\theta) + \varepsilon(n, x) \right) \leqslant e^{-x},$$

$$\varepsilon(n, x) = \sqrt{\frac{2\sigma^2(\Theta)x}{n}} + \sqrt{\frac{\sqrt{2} C_{GI}(\Theta)}{3}} \left(\frac{x}{n} \right)^{3/4}$$

$$\sigma^2(\Theta) = \sup_{\theta} \mathrm{Var}(Q(\cdot, \theta));$$

$$C_{GI}(\Theta) = \sup_{\theta \in \Theta} \|(Q(\cdot, \theta) - R(\theta))^2 - \mu((Q(\cdot, \theta) - R(\theta))^2)\|_{\psi_1}$$

Similar result holds under the exponential inequality of $Q^+(\cdot, \theta)$.

6. Applications: we can verify quickly the non-efficiency of a learning machine.

To verify if a learning machine does not work, given a confidence level $\alpha \in (0.5, 1)$, one takes

$$Y = f(X) \text{ and } Y_i = f(X_i),\ 1 \leqslant i \leqslant n$$

for some n so that

$$\varepsilon\left(n, \log \frac{1}{1-\alpha}\right) \leqslant \varepsilon_0$$

roughly $n \succeq \frac{\sigma(\Theta)}{\varepsilon_0^2} \log \frac{1}{1-\alpha}$, which is dimension-free. Then by Theorem 8, the minimal error of the learning machine

$$\inf_{\theta \in \Theta} R(\theta) \geqslant \inf_{\theta \in \Theta} R_{E,n}(\theta) - \varepsilon_0$$

with probability α.

Conclusion: if $\inf_{\theta \in \Theta} R_{E,n}(\theta)$ is not small with a sample size n given above (dimension-free), then the learning machine is not good most probably.

End Slide

Thanks for your attention

- F. Bolley and C. Villani, Weighted Csiszar-Kullback-Pinsker inequality and applications to transportation inequalities. *Ann. Fac. Sci. Toulouse Math. (6) 14 (2005), no. 3, 331-352.*
- H. Djellout, A. Guillin and L. Wu. Transportation cost-information inequalities for random dynamical systems and diffusions. *Ann. Probab.* 32 (3B) (2004) 2702-2732.
- X. Fan, I. Grama, Q.S. Liu. Hoeffding's inequality for supermartingales. *Stoch. Proc. Appl. 122 (2012), 3545-3559.*
- X. Fan, I. Grama, Q.S. Liu. Sharp large deviation results for sums of independent random variables. *Sciences in China: Mathematics (2015), Vol. 58, No. 9, 1939-1958.*
- N. Fournier, A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Relat. Fields 162, 707-738 (2015)*

- F.Q. Gao; A. Guillin and L. Wu. Bernstein-type concentration inequalities for symmetric Markov processes. *SIAM. Theory Probab. Appl. 58, 3 (2014), 358-382.*
- N. Gozlan and C. Léonard. A large deviation approach to some transportation cost inequalities. *Probability Theory and Related Fields*, 1-2:235–283, 2007.
- P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Probab. 28 (2000), no. 2, 863-884.*
- M. Ledoux, M. Talagrand, *Probability in Banach spaces. Isoperimetry and processes.* Ergebnisse der Mathematik und ihrer Grenzgebiete (3), 23. Springer-Verlag, Berlin, 1991
- A. Van der Vaart, J.A. Wellner. *Weak convergence of empirical processes.* Springer-Verlag, 1996
- M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Ann. Probab. 22 (1994), no.1, 28-76.*

- M. Talagrand, Concentration of measure and isoperimetric inequalities in product spaces. *Publications mathématiques de l'I.H.E.S., tome 81 , p. 73-205*(1995).
- M. Talagrand. New concentration inequalities in product spaces. *Invent. Math. (1996), No. 126, 505-563.*
- V.N. Vapnik: *The Nature of Statistical Learning Theory*, Second Edition. Springer 1999.
- R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science.* University of California, Irvine June 9, 2020. https://www.math.uci.edu/ rvershyn
- R. Wang, X.Y. Wang and L. Wu, Sanov's theorem in the Wasserstein metric: a necessary and sufficient condition. *Stat. Proba. Letters 80 (2010), 505-512.*
- L. Wu. Large deviations, moderate deviations and LIL for empirical processes, *Ann. Probab. vol. 22, No. 1, p17-27.*