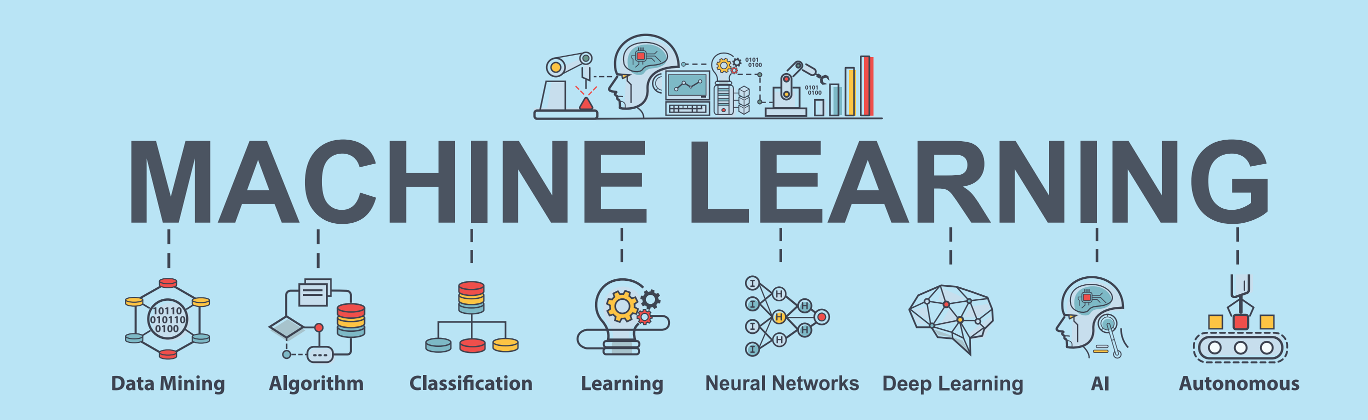
From Decoupling and Self-Normalization to Machine Learning



Victor H. de la Pena

Introduction

We now live in the age of the digital revolution, where machine learning and artificial intelligence have transformed the way data is generated, processed, and analyzed to solve complex research problems. One of the goals in the development of algorithms in machine learning is to extract as much information from the data as possible that cannot be achieved through traditional techniques. Decoupling and self-normalization are statistical tools that handle complex problems that are beyond the spheres of classical statistical methods. Self-normalization can be traced back to the seminal work of Gosset [9], which is considered a breakthrough in science. Notably, his Student t-statistic allowed statistical inference about the value of the mean of a (Gaussian) distribution without knowledge of the actual value of the variance, provided one has a random sample from the target population. Remarkably, the self-normalization approach provides techniques to extend the t-statistic to the case of non-Gaussian distributions and nonindependent variables.

Then what do decoupling, self-normalization, and machine learning have in common? They enable the development of algorithms with minimal dependence or parametric assumptions. In essence, decoupling and self-normalization are areas that grew out of the need to extend martingale methods to high and infinite dimensions and complex nonlinear dependence structures. Decoupling provides tools for treating dependent variables as if they were independent. In particular, it provides a natural tool for developing sharp exponential inequalities for self-normalized martingales. Prototypical examples of self-normalized processes are the t-statistic in dependent random variables as well as (self-normalized) extensions of the law of the iterated logarithm of Kolmogorov. As will be described in the last section, these results have been used to establish important machine-learning tools, the development of efficient algorithms for the stochastic multiarmed bandit problem, and for the so-called learning to rank (LTR) model.

Complete Decoupling

Let $\{d_i\ i = 1, \dots, n\}$ be a sequence of dependent random variables with $E|d_i| < \infty$. Let $\{y_i\ i = 1, \dots, n\}$ be a sequence of independent variables where for each i, d_i and y_i have the same marginal distributions. Since $Ed_i = Ey_i$, linearity of expectations provides the first "complete de-

Victor H. de la Pena is a professor of statistics at Columbia University. His email address is `vp@stat.columbia.edu`.

Communicated by Notices *Associate Editor Noah Simon.*

For permission to reprint this article, please contact: `reprint-permission@ams.org`.

DOI: https://doi.org/10.1090/noti1978

coupling" equality:

$$E\sum_{i=1}^{n} d_i = E\sum_{i=1}^{n} y_i. \tag{1}$$

As a concrete example for constructing the sequence $\{y_i\}$, let $\{d_i^{(j)}, i = 1, \dots, n; j = 1, \dots, n\}$ be independent copies of $\{d_i\}$ and take $y_i = d_i^{(i)}$. Then, it is easy to see that $\{y_i\}$ is a sequence of independent random variables, since each row in the array is independent of the others and

$$E\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} Ed_i = \sum_{i=1}^{n} Ed_i^{(i)} = \sum_{i=1}^{n} Ey_i = E\sum_{i=1}^{n} y_i.$$

In complete decoupling, one compares $Ef(\sum d_i)$ to $Ef(\sum y_i)$ for more general functions than $f(x) = x$. In statistics, Hoeffding's [11] inequality for comparing sampling without replacement to sampling with replacement provides an early example. More recent work involves extensions to the case of negatively associated variables of Shao [19], as well as to the case of sequences of nonnegative random variables with arbitrary dependence structures. Applications of complete decoupling include, among others, tools for the optimization of stochastic processes such as the scheduling of dependent computer servers (see Makarychev and Sviridenko [17]).

Let the population C consist of N values $c_2, c_2, \dots, c_N$. Let $d_1, d_2, \dots, d_n$ denote a random sample without replacement from C, and let $y_1, y_2, \dots, y_n$ denote a random sample with replacement from C. The random variables $y_1, \dots, y_n$ are independent and identically distributed. Moreover, for all i, d_i and y_i have the same marginal distributions. Hoeffding [11] developed the following widely used complete decoupling inequality: For every continuous convex function Φ,

$$E\Phi\left(\sum_{i=1}^{n} d_i\right) \le E\Phi\left(\sum_{i=1}^{n} y_i\right). \tag{2}$$

Shao [19] extended it to the case of negatively associated random variables.

In what follows we present a (sharp) complete decoupling inequality for sums of nonnegative dependent random variables that provides a reverse Hoeffding's inequality for L_p moments. The price one pays is a constant.

Theorem. *Let $\pi(1)$ be a Poisson random variable with mean 1. Assume the d_i's are a sequence of arbitrarily dependent nonnegative random variables. Let $y_i, \dots, y_n$ be independent random variables with y_i having the same distributions as d_i for all i. Then, for $p \ge 1$,*

$$E\left(\sum_{i}^{n} y_i\right)^p \le E(\pi(1)^p)E\left(\sum_{i=1}^{n} d_i\right)^p. \tag{3}$$

In particular, when $p = 1$ we obtain the original complete decoupling equality in (1),

$$E\sum_{i}^{n} y_i = E\pi(1)E\sum_{i=1}^{n} d_i = E\sum_{i=1}^{n} d_i.$$

The above result was introduced by de la Pena [4] in the case of more general functions, including powers with different constants. The sharp constants were first obtained by Makarychev and Sviridenko [17].

Conditionally Independent (Tangent) Decoupling

The theory of martingale inequalities has been central in the development of modern probability theory. Recently it has been expanded widely through the introduction of the theory of conditionally independent (tangent) decoupling. This approach to decoupling can be traced back to a result of Burkholder and McConell included in Burkholder [2] that represents a step in extending the theory of martingales to Banach spaces.

Let $\{d_i\}$ and $\{e_i\}$ be two sequences of random variables adapted to the σ-fields $\{F_i\}$. Then $\{d_i\}$ and $\{e_i\}$ are said to be tangent with respect to $\{\mathcal{F}_i\}$ if, for all i,

$$\mathcal{L}(d_i|\mathcal{F}_{i-1}) = \mathcal{L}(e_i|\mathcal{F}_{i-1}), \tag{4}$$

where $\mathcal{L}(d_i|\mathcal{F}_{i-1})$ denotes the conditional probability law of d_i given $\mathcal{F}_{i-1}$.

Let $d_1, \dots, d_n$ be an arbitrary sequence of dependent random variables adapted to an increasing sequence of σ-fields $\{\mathcal{F}_i\}$. Then, as shown in de la Pena and Gine [6], one can construct a sequence $e_1, \dots, e_n$ of random variables that is conditionally independent given $\mathcal{G} = \mathcal{F}_n$. The construction proceeds as follows: First we take e_1 and d_1 to be two independent copies of the same random mechanism. Having constructed $d_1, \dots, d_{i-1}; e_1, \dots, e_{i-1}$, the ith pair of variables d_i and e_i comes from i.i.d. copies of the same random mechanism given $\mathcal{F}_{i-1}$. It is easy to see that using this construction and taking

$$\mathcal{F}_i' = \mathcal{F}_i \vee \sigma(e_1, \dots, e_i),$$

the sequences $\{d_i\}$, $\{e_i\}$ satisfy

$$\mathcal{L}(d_i|\mathcal{F}_{i-1}') = \mathcal{L}(e_i|\mathcal{F}_{i-1}') = \mathcal{L}(e_i|\mathcal{G})$$

and the sequence $e_1, \dots, e_n$ is conditionally independent given $\mathcal{G} = \mathcal{F}_n$.

A sequence $\{e_i\}$ of random variables satisfying the above conditions is said to be a *decoupled* $\mathcal{F}_i'$-tangent version of $\{d_i\}$.

As in the case of complete decoupling, linearity of expectations provides the canonical example of a decoupling "equality." In conditionally independent decoupling one replaces dependent random variables with decoupled (conditionally independent) random variables. If $E|d_i| <$

∞ for all i, then

$$E\sum_{i=1}^{n} d_i = E\sum_{i=1}^{n} e_i. \tag{5}$$

To see this note that

$$\begin{aligned} E\sum_{i=1}^{n} d_i &= \sum_{i=1}^{n} Ed_i = \sum_{i=1}^{n} E(E(d_i|\mathcal{F}'_{i-1})) \\ &= \sum_{i=1}^{n} E(E(e_i|\mathcal{F}'_{i-1})) = \sum_{i=1}^{n} E(E(e_i|\mathcal{G})) \\ &= E\left(E\left(\sum_{i=1}^{n} e_i|\mathcal{G}\right)\right) = E\sum_{i=1}^{n} e_i. \end{aligned}$$

The first general decoupling inequality for tangent sequences was obtained by Zinn [20] and extended by Hitczenko [10].

A turning point in the theory of decoupling for tangent sequences has been a 1986 result of Kwapien and Woyczynski (see Kwapien and Woyczinsky [14] for the exact reference). It is shown in that paper for the first time, and in a precise manner, that one can always obtain a decoupled tangent sequence to any adapted sequence hence, making general decoupling inequalities widely applicable.

Developments of the theory and applications are found in hard copy in Kwapien and Woyczynski [14] and in de la Pena and Gine [6].

Decoupling and Self-Normalization

Next, we present a sharp decoupling inequality with constraints from de la Pena [5]. This result will be used later to obtain a sharp extension of Bernstein's inequality for independent random variables to the case of self-normalized martingales.

Let $\{d_i\}$ and $\{e_i\}$ be two tangent sequences with $\{e_i\}$ decoupled. Then for all $g > 0$ adapted to $\sigma(\{d_i\})$,

$$Eg\exp\{\lambda\sum_{i=1}^{n} d_i\} \le \sqrt{Eg^2\exp\{2\lambda\sum_{i=1}^{n} e_i\}}. \tag{6}$$

As a first application we use (6) in the context of the sampling schemes discussed above.

Example (conditionally independent sampling). In this example we show how to decouple a sample without replacement and show how the decoupled sequence relates to sampling without replacement and sampling with replacement. (In survey sampling we treat draws without replacement as if they were independent, though they are actually weakly coupled.) As before, consider drawing samples of size n from a population C that consists of N values. Let $d_1, \dots, d_n$ denote a sample without replacement and let $y_1, \dots, y_n$ denote a sample with replacement. A conditionally independent sample can be constructed as follows. At the ith stage of a simple random sampling without replacement, both d_i and e_i are obtained sampling uniformly from $\{c_1, \dots, c_N\}$, excluding $\{d_1, \dots, d_{i-1}\}$. This may be attained by selectively returning items to C. More precisely, at the ith stage first draw e_i and return it to the population. Then draw d_i and put its value aside. It is easy to see that the above procedure will make $\{e_i\}$ tangent to $\{d_i\}$ with $\mathcal{F}_n = \sigma(d_1, \dots, d_n; e_1, \dots, e_n)$. Moreover, the sequence $\{e_i\}$ is conditionally independent given $\mathcal{G} = \sigma(d_1, \dots, d_N)$.

A use of the exponential decoupling inequality (with $g = 1$) found in (6) gives

$$E\exp\{\lambda\sum_{i=1}^{n} d_i\} \le \sqrt{E\exp\{2\lambda\sum_{i=1}^{n} e_i\}}. \tag{7}$$

In some sense, conditionally independent sampling can be viewed as a sampling scheme in between sampling with replacement and sampling without replacement.

Next we state Bernstein's inequality for sums of independent random variables and its self-normalized counterpart in the case of self-normalized martingales.

Bernstein's inequality. Let $\{x_i\}$ be a sequence of independent random variables. Assume that $E(x_j) = 0$ and $E(x_j^2) = \sigma_j^2 < \infty$ and set $v_n^2 = \sum_{j=1}^{n} \sigma_j^2$. Furthermore, assume that there exists a constant $0 < c < \infty$ such that, almost surely, $E(|x_j|^k) \le (k!/2)\sigma_j^2 c^{k-2}$ for all $k > 2$. Then for all $x > 0$,

$$P\left(\sum_{i=1}^{n} x_i > x\right) \le \exp\left(-\frac{x^2}{2(v_n^2 + cx)}\right). \tag{8}$$

The following is a self-normalized inequality from de la Pena [5]:

Self-normalized Bernstein's inequality. Let $\{d_i, \mathcal{F}_i\}$ be a martingale difference sequence. Assume that $E(d_j|\mathcal{F}_{j-1}) = 0$ and $E(d_j^2|\mathcal{F}_{j-1}) = \sigma_j^2 < \infty$ (satisfied by subexponential random variables) and set $V_n^2 = \sum_{j=1}^{n} \sigma_j^2$. Furthermore, assume that there exists a constant $0 < c < \infty$ such that, almost surely, $E(|d_j|^k|\mathcal{F}_{j-1}) \le (k!/2)\sigma_j^2 c^{k-2}$ for all $k > 2$. Then for all $x, y > 0$,

$$P\left(\frac{\sum_{i=1}^{n} d_i}{V_n^2} > x, \frac{1}{V_n^2} \le y\right) \le \exp\left(-\frac{x^2}{2y(1 + cx)}\right). \tag{9}$$

Remark. The inequality is sharp, since when $V_n^2 = v_n^2$ (nonrandom) the two inequalities are equivalent. The key steps in obtaining this result involve the use of Markov's inequality followed by the decoupling inequality presented in (6). We then (conditionally) apply the standard results for sums of independent random variables to complete the proof.

Self-Normalization

This area goes back to the seminal work of Gosset [9] ("Student"), in which he introduced the t-statistic and the t-distribution. For more than a century, the t-statistic has evolved into much more general Studentized statistics and self-normalized processes. Let $X_1, X_2, \dots, X_n$ be a sequence of i.i.d. normal random variables. Gosset considered the problem of statistical inference on the mean μ when the standard deviation σ of the underlying distribution is unknown. Let $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$, $\hat{\sigma}_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n-1}$ be the sample mean and the sample variance, respectively. In his 1908 paper Gosset derived the distribution of the t-statistic $T_n = \sqrt{n}\frac{\sum_{i=1}^n (X_i - \mu)}{\hat{\sigma}_n}$ for normal X_i; this is the t-distribution with $n-1$ degrees of freedom. The t-distribution converges to the standard normal distribution, and in fact T_n has a limiting standard normal distribution even when the X_i are nonnormal.

It is noted that for "fat-tailed" distributions with infinite second or even first absolute moments, it has been found that the t-test of $\mu = \mu_0$ is robust against nonnormality in terms of the type I error probability. Furthermore, directly plugging in the true variance will actually result in a substantially worse statistic (that is extremely anticonservative in the case of, e.g., tests involving Cauchy random variables). Without loss of generality, consider testing the hypothesis $\mu_0 = 0$ so that

$$T_n = \frac{\sqrt{n}\bar{X}_n}{\hat{\sigma}_n} = \frac{S_n}{V_n}\{\frac{n-1}{n-(S_n/V_n)^2}\}^{1/2}, \tag{10}$$

where $S_n = \sum_{i=1}^n X_i$, $V_n^2 = \sum_{i=1}^n X_i^2$.

In view of the above equality, if T_n or S_n/V_n has limiting distribution, then so does the other, and it is well known that they coincide. In de la Pena et al. [8], a more complete historical perspective of the general theory is provided, as well as numerous applications in statistics.

Pseudo-maximization (Method of Mixtures)

The method of pseudo-maximization (also known as the method of mixtures) was used in de la Pena et al. [7] and is based on the following assumption: Let (A, B) be an arbitrarily dependent vector of random variables, with $B > 0$. Assume that $-\infty < \lambda < \infty$. Then, the pair is said to satisfy the canonical assumption if

$$E\exp(\lambda A - \lambda^2 B^2/2) \le 1. \tag{11}$$

The appendix provides a wide class of processes that satisfy this assumption. These include martingales, randomly stopped processes, and sums of conditionally symmetric random variables. For some applications the range of λ can be smaller, e.g., $0 < \lambda < \lambda_0$. Note that if the integrand in $E\exp(\lambda A - \lambda^2 B/2)$ can be maximized over λ inside the expectation (as can be done if A/B^2 is nonrandom), taking $\lambda = A/B^2$ would yield $E\exp(\frac{A^2}{2B^2}) \le 1$. This in turn would give the optimal Chebyshev-type bound

$$P\left(\frac{A}{B} > x\right) \le \exp\left(\frac{-x^2}{2}\right). \tag{12}$$

However, since A/B^2 cannot (in general) be taken to be nonrandom, we need to find an alternative method for dealing with this maximization. One approach for attaining a similar effect involves integrating over a probability measure F and using Fubini's theorem to interchange the order of integration with respect to P and F:

$$\begin{aligned} &\int E\exp(\lambda A - \lambda^2 B^2/2)F(d\lambda) \\ &\quad = E\int \exp(\lambda A - \lambda^2 B^2/2)F(d\lambda) \\ &\quad \le \int F(d\lambda) \le 1. \end{aligned} \tag{13}$$

To be effective for all possible pairs (A, B), the F chosen would need to be as uniform as possible so as to include the maximum value of $E\exp(\lambda A - \lambda^2 B^2/2)$ regardless of where it might occur. Thus some mass is certain to be assigned to and near the random value $\lambda = A/B^2$ that maximizes $\exp(\lambda A - \lambda^2 B^2/2)$. Since all uniform measures are multiples of Lebesgue measure (which is infinite), we construct a finite measure (or a sequence of finite measures) that tapers off to zero as $\lambda \to \infty$ as slowly as we can manage. One can obtain different results by changing the measure F.

For example, as shown in de la Pena et al. [7], by integrating over a Gaussian measure in (13) one can develop the following self-normalized exponential inequality:

$$P(|A|/\sqrt{B^2 + (EB)^2} \ge x) \le \sqrt{2}\exp(-x^2/4) \tag{14}$$

for all $x > 0$.

We remark that we almost get the optimal Chebyshev bound, the ideal inequality (see (12)) up to constants, and the term $(EB)^2$.

Under the following refinement of the canonical assumption we obtain an LIL bound. Assume that

$$\begin{gathered} \{\exp(\lambda A_t - \lambda^2 B_t^2/2), t \ge 0\} \\ \text{is a super martingale with mean} \le 1. \end{gathered} \tag{15}$$

Then,

$$\limsup_{t\to\infty} \frac{A_t}{\sqrt{2B_t^2 \log\log B_t^2}} \le 1, \tag{16}$$

on the set $\{\lim_{t\to\infty} B_t^2 = \infty\}$.

As formalized in Lemma A.3, for conditionally symmetric increments, d_i, we can use this result to get

$$\limsup_{n\to\infty} \frac{\sum_{i=1}^n d_i}{\sqrt{2\sum_{i=1}^n d_i^2 \log\log \sum_{i=1}^n d_i^2}} \le 1, \tag{17}$$

on the set $\{\lim_{n\to\infty} \sum_{i=1}^{n} d_i^2 = \infty\}$, which is a sharp extension of Kolmogorov's LIL without moments assumptions. In particular, the result is valid for i.i.d. centered Cauchy random variables.

Applications to Machine Learning

An area of important application in machine learning is the stochastic multiarmed bandit problem. Following the formulation by Kaufmann et al. [12], the model involves sequentially sampling from a set of k probability distributions, called arms, each having an unknown mean μ_k. At time t, an arm is selected according to a sampling strategy that depends on the history of past arm selections and samples, and then a sample X_t is drawn from the associated distribution. A key objective is to adjust the sampling strategy in order to maximize the expected value of the rewards gathered up to a specified time horizon T. In their paper Kaufmann et al. provide improved sequential stopping rules that have guaranteed error probabilities and shorter average running times. In a related paper Kaufmann and Koolen [13] establish asymptotic optimality of a class of sequential tests generalizing the track-and-stop method to problems beyond best arm identification. The approach they take involves the use of the method of mixtures and a self-normalized law of the iterated logarithm (see (16) above).

In addition in machine learning, self-normalization techniques are being used in the development of efficient algorithms for the so-called learning to rank (LTR). The primary objective of LTR, which is used in web search and recommender systems (Liu [16]), is to optimally select a subset from a large set of documents that maximizes the satisfaction of the user. It is well known that some of the commonly proposed approaches have limited applications, including lack of convergence in certain situations.

In a recent paper, Lattimore et al. [15] introduced an algorithm, called TopRank, with several desirable features, such as performance superior to many of the competing algorithms. A major step in the development of the proposed algorithm is use of self-normalization principles. For example, in their paper, the proof of their Lemma 6 is based on this principle in the construction of the bounds on relevant quantities.

The theory for self-normalized sums has also been applied in the formulation of regression models for high-dimensional data. Notably, Belloni et al. [1] used the theory to achieve Gaussian-like results under weak conditions for a self-tuning and square-root function of the lasso method that works well with unknown scale, heteroscedasticity, and nonnormality of the error terms.

Further Applications

The tools developed have been successfully applied in diverse areas such as extension of martingale results to infinite dimensions, including Banach spaces; self-normalized martingales; stochastic integration; empirical processes, including U-statistics and U-processes; density estimation; sequential analysis; survival analysis; efficient Monte Carlo methods; matrix completion; large deviations; robust estimation; likelihood; and Bayesian inference. See Kwapien and Woyczynski [14], de la Pena and Gine [6], and de la Pena et al. [8] for details.

More recent applications of conditionally independent decoupling include Candes and Recht [3], which deals with exact matrix completion via convex optimization. There has been a recent surge of interest in applying and developing the methods presented in this survey. In particular, Rahklyn et al. [18] uses conditional independent decoupling techniques to study sequential complexities and exponential inequalities for martingales in Banach spaces. Makarychev and Sviridenko [17] uses complete decoupling inequalities to develop stochastic optimization tools for energy efficient routing load balancing in parallel machines.

Concluding Remarks

As can be seen from the broad range of results and applications, it is worth looking at problems using the decoupling and self-normalization perspectives. In fact, there are still multiple open problems in these areas, including the development of new sharp algorithms in the area of machine learning.

Appendix

Lemma A.1. *Let W_t be a standard Brownian motion. Assume that T is a stopping time such that $T < \infty$ a.s. Then for all $-\infty < \lambda < \infty$,*

$$E \exp\{\lambda W_T - \lambda^2 T/2\} \leq 1. \tag{18}$$

Lemma A.2. *Let M_t be a continuous, square-integrable martingale, with $M_0 = 0$. Then, for all $-\infty < \lambda < \infty$,*

$$\exp\{\lambda M_t - \lambda^2 < M >_t /2\} \leq 1. \tag{19}$$

If M_t is only assumed to be a continuous local martingale, the inequality is also valid (by application of Fatou's lemma).

Lemma A.3. *Let $\{d_i\}$ be a sequence of variables adapted to an increasing sequence of σ-fields $\{\mathcal{F}_i\}$. Assume that the d_i's are conditionally symmetric (i.e., $\mathcal{L}(d_i|\mathcal{F}_{i-1}) = \mathcal{L}(-d_i|\mathcal{F}_{i-1})$). Then,*

$$E \exp\{\lambda \sum_{i=1}^{n} d_i - \lambda^2 \sum_{i=1}^{n} d_i^2/2\} \leq 1 \tag{20}$$

for all $-\infty < \lambda < \infty$.

References

[1] Belloni A, Chernozhukov V, Wang L. Pivotal estimation via square-root Lasso in nonparametric regression, *Ann. Statist.*, (42, no. 2): 757–788, 2014. MR3210986

[2] Burkholder DL. A geometric condition that implies the existence of certain singular integrals of Banach-space-valued functions. *Conference on harmonic analysis in honor of Antoni Zygmund, Vol. I, II (Chicago Ill., 1981)*. Wadsworth. Math. Ser., Wadsworth, Belmont, CA; 1983: 270–286. MR730072

[3] Candes EJ, Recht B. Exact matrix completion via convex optimization, *Found. Comput. Math.*, (9, no. 6): 717–772, 2009. MR2565240

[4] de la Pena VH. Bounds on the expectation of functions of martingales and sums of positive RVs in terms of norms of sums of independent random variables, *Proc. Amer. Math. Soc.*, (108, no. 1): 233–239, 1990. MR990432

[5] de la Pena VH. A general class of exponential inequalities for martingales and ratios, *Ann. Probab.*, (27, no. 1): 537–564, 1999. MR1681153

[6] de la Pena VH, Gine E. *Decoupling: From Dependence to Independence*, Springer, New York, 1999. MR1666908

[7] de la Pena VH, Klass MJ, Lai TL. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws, *Ann. Probab.*, (32, no. 3A): 1902–1933, 2004. MR2073181

[8] de la Pena VH, Lai TL, Shao Q-M. *Self-Normalized Processes: Limit Theory and Statistical Applications*, Springer, 2009. MR2488094

[9] Gosset (Student) WS. The probable error of the mean, *Biometrika*, (6): 1–25, 1908.

[10] Hitczenko P. Comparison of moments for tangent sequences of random variables, *Probab. Theory Related Fields*, (78, no. 2): 223–230, 1988. MR945110

[11] Hoeffding W. Probability inequalities for sums of bounded random variables, *J. Amer. Statist. Assoc.*, (58, no. 301): 13–30, 1963. MR0144363

[12] Kaufmann E, Cappe O, Garivier A. On the complexity of best-arm identification in multi-armed bandit models, *J. Mach. Learn. Res.*, (17): 1–42, 2016. MR3482921

[13] Kaufmann E, Koolen W. Mixtures martingales revisited with applications to sequential tests and confidence intervals, arXiv:1811.11419v1, 2018.

[14] Kwapien S, Woyczynski W. *Random Series and Stochastic Integrals: Single and Multiple*, Birkhauser, Boston, 1992. MR1167198

[15] Lattimore T, Kveton B, Li S, Szpervary Z. TopRank: A practical algorithm for online stochastic ranking, arXiv:1806.02248v2, Neural Information Processing Systems (NIPS), Montreal, Canada, 2018.

[16] Liu T. *Learning to Rank for Information Retrieval*, Springer, 2011.

[17] Makarychev K, Sviridenko M. Solving optimization problems with diseconomies of scale via decoupling, *J. ACM*, (65, no. 6): Article 42, 2018. MR3882265

[18] Rakhlin A, Sridharan K, Tewari A. Sequential complexities and uniform martingale laws of large numbers, *Probab. Theory Related Fields*, (161, issue 1-2): 111–153, 2015. MR3304748

[19] Shao Q-M. A comparison theorem on moment inequalities between negatively associated and independent random variables, *J. Theoret. Probab.*, (13, no. 2): 343–356, 2000. MR1777538

[20] Zinn J. Comparison of martingale difference sequences. In: *Probability in Banach Spaces V*, Lecture Notes in Math., (1153). Springer; 1985: 453–457. MR821997

Victor H. de la Pena

Credits

Opening image is courtesy of Getty.

Photo of the author is courtesy of the author.