How ~~to~~ I read research papers

Aaditya Ramdas

Dept. of Statistics and Data Science
Machine Learning Dept.
Carnegie Mellon University

Thanks to: Sarah Tan, Pablo Samuel Castro

https://cs.stanford.edu/~rishig/courses/ref/paper-reading-overview.pdf
https://cs.stanford.edu/~rishig/courses/ref/paper-reading-technical.pdf

This talk

To give you an idea **how reading papers evolved** for me (and might evolve for you) as you go from being an inexperienced researcher to a mature researcher.

To show you some **tools that I use** to keep track of papers that I have read, so that I do not forget key points.

<u>Types of papers:</u>
theoretically inclined papers in statistical machine learning or mathematical statistics or applied probability

Know why you are reading a research paper

Good reasons:

a) "Read directly from the masters" — a lot can be lost or omitted by someone else summarizing or paraphrasing a classic paper.

b) The authors had a new insights on an old problem

Bad reasons:

a) **"I am new to the area, I wanted to read the original proof."**

Often, the original authors' proof was complicated and has been far simplified in later works.

If you are new to an area, the simpler proofs may be a better first read.

b) **"I am citing this paper, so I should read it fully."**

Sometimes, we are only interested in porting a very particular lemma or result from a paper, and one does *not* need to read it fully. One definitely needs to verify the correctness of the claim being made about the other paper, or the correctness of the result being borrowed.

Typically, the few *closest* papers to yours do need to understood fully, but not every cited paper.

Start of PhD

Looks like
an important
paper. Let me read it
from start to end

The Annals of Statistics
2009, Vol. 37, No. 5A, 2178–2201
DOI: 10.1214/08-AOS646

HIGH-DIMENSIONAL VARIABLE SELECTION

BY LARRY WASSERMAN AND KATHRYN ROEDER[1]

Carnegie Mellon University

This paper explores the following question: what kind of statistical guarantees can be given when doing variable selection in high-dimensional models? In particular, we look at the error rates and power of some multi-stage regression methods. In the first stage we fit a set of candidate models. In the second stage we select one model by cross-validation. In the third stage we use hypothesis testing to eliminate some variables. We refer to the first two stages as "screening" and the last stage as "cleaning." We consider three screening methods: the lasso, marginal regression, and forward stepwise regression. Our method gives consistent variable selection under certain conditions.

Today

I never
read a paper
from start to end on my
first opening (or ever)

Common (wrong) belief: papers should be read linearly

No! How exactly I read a paper depends on

a. My goal (paper reviewer vs. finding related work for your own paper vs. curiosity-driven daily reading vs. trying to get into a new field)

b. How well I know the topic (and how well I want to know it)

c. How much time I have right now (more than 10mins, less than 2hrs)

If so, then why are papers written linearly in a somewhat standard high-level ordering?

To help you in your non-linear search!?
You can jump forward to what you're looking for.

Papers are not novels: nonlinear order is the norm

a. Can often skip entire sections

b. May need to read other paragraphs or subsections multiple times

c. Sometimes the reading needs to split across days

What are you looking for?

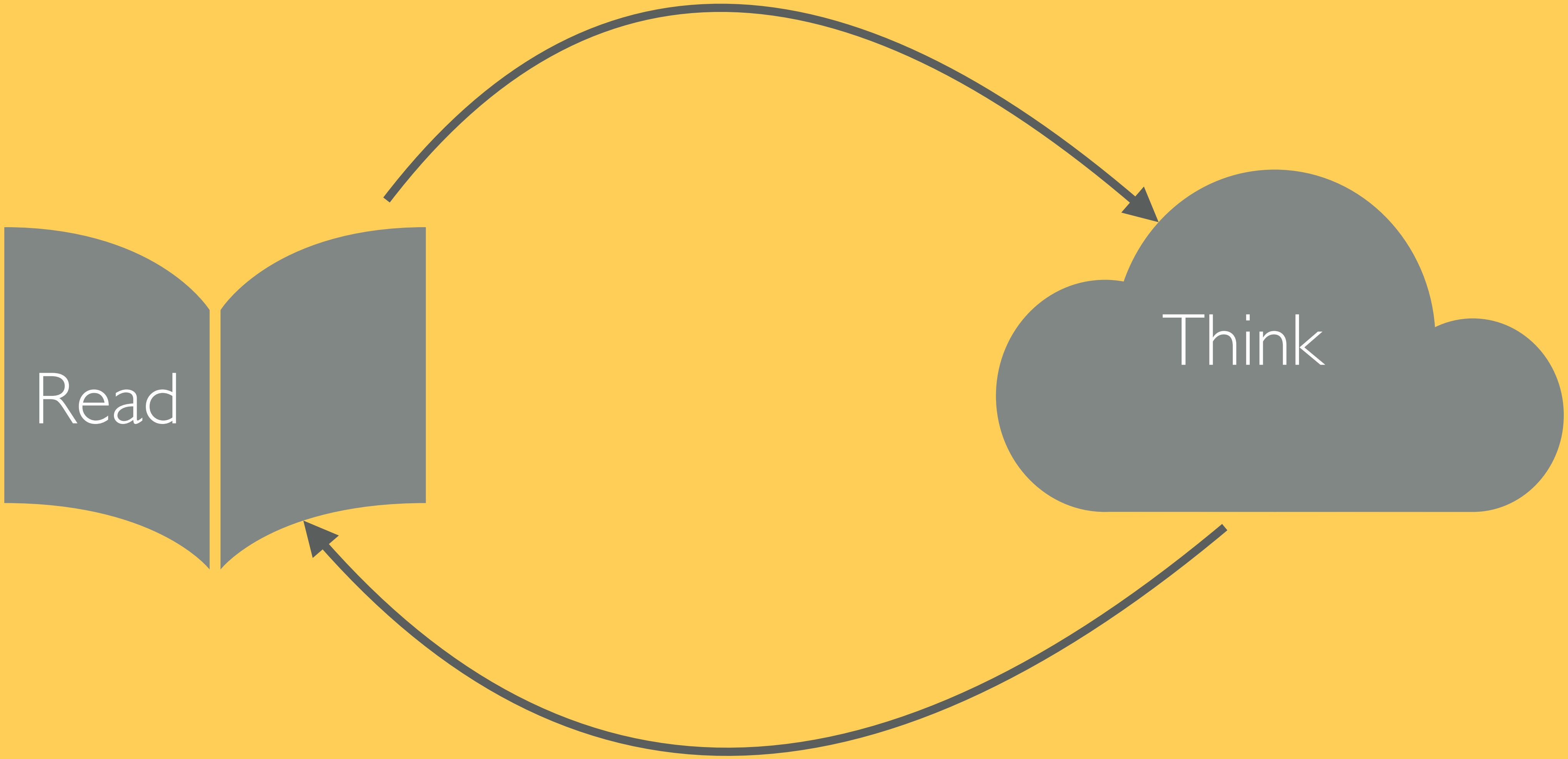
a. Do you just want to know the problem being solved?

b. Maybe you want to understand the main claim(s) being made?

c. What was the major past hurdle and how was it overcome?

d. Is there a nifty, cute proof technique I can borrow?

The principle of iterative refinement



First pass: jigsaw puzzle theme (5-30mins)

a. What is the problem being solved?
   [problem context]

b. Why is it interesting and nontrivial?
   [be critical about assumptions, but not too much]

c. What is the main claim being made?
   [at least in English, preferably in Math]

<u>**Sources**</u>: abstract/intro, problem definition, main theorem, discussion.

For papers with interesting titles or abstracts, 75% end at first pass.
(one per day?)

Second pass: scuba diving (30mins-2hrs)

a. What was the main technical hurdle faced by past work?
   How does this paper overcome it?

b. What is the simplest nontrivial baseline? According to what metric is the new method better?

c. What's still open and why does their insight not apply there?

d. Does their insight apply to other unconsidered problems?

e. What are the caveats and takeaways?

Sources: examples, special cases, key lemmas/propositions, proof outlines

For papers with interesting titles or abstracts, 20% end at second pass.
(one per week?)

Third pass: the swamp (multiple days/weeks)

a. How did they prove their lemmas, propositions, theorems?

b. Can I reprove (in spirit) their result from scratch?
[High bar! Read for concepts, even if they are technical, skip algebra or symbolic manipulation.]

c. If I cannot, what piece of intuition am I missing? Does an additional assumption make it easier?

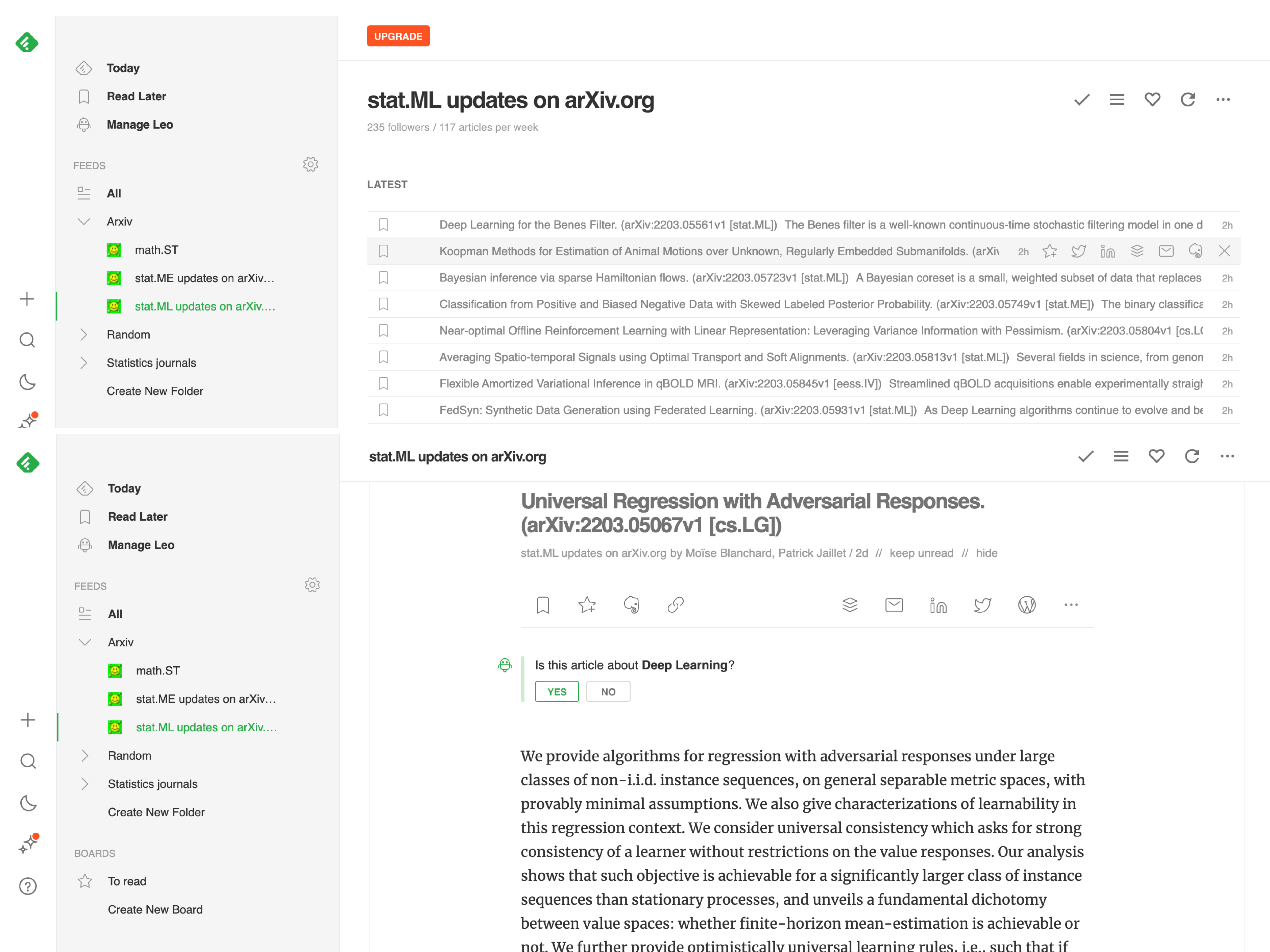
d. Can I simplify their proof using the tools I know, or prove their main result in a very different way, once I get their intuition?
[often easier than reproducing their proof, can help you avoid reading a tedious proof :)]

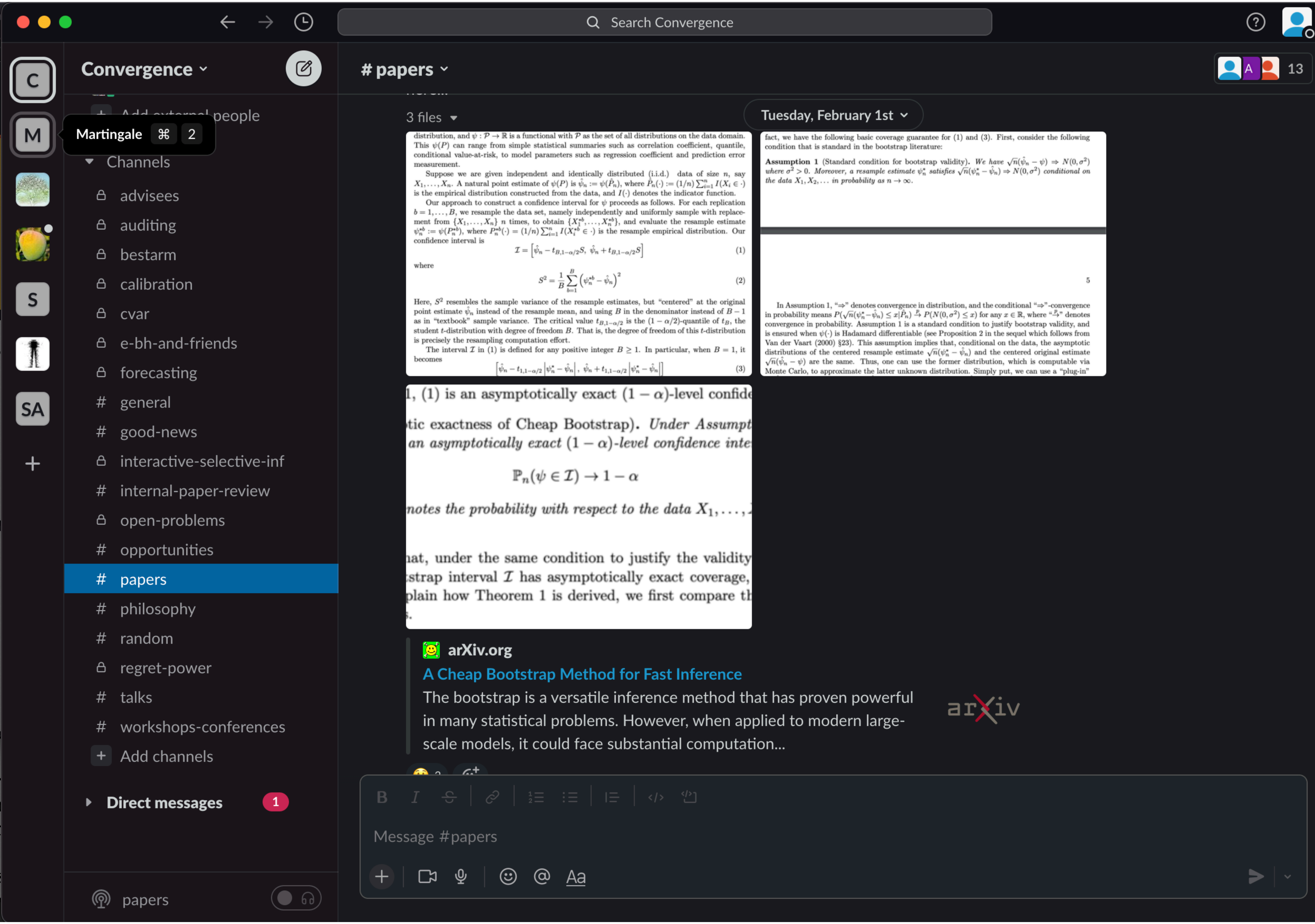
Sources: appendices, proof details, corollaries, remarks, related work

For papers with interesting titles or abstracts, 5% reach a third pass.
(one per month?)

How I organize my reading

Feedly
Chrome tabs
PDF markings
Slack


UPGRADE
Today
Read Later
Manage Leo
FEEDS
All
Arxiv
math.ST
stat.ME updates on arXiv…
stat.ML updates on arXiv.…
Random
Statistics journals
Create New Folder
stat.ML updates on arXiv.org
235 followers / 117 articles per week
LATEST
Deep Learning for the Benes Filter. (arXiv:2203.05561v1 [stat.ML]) The Benes filter is a well-known continuous-time stochastic filtering model in one d 2h
Koopman Methods for Estimation of Animal Motions over Unknown, Regularly Embedded Submanifolds. (arXiv 2h
Bayesian inference via sparse Hamiltonian flows. (arXiv:2203.05723v1 [stat.ML]) A Bayesian coreset is a small, weighted subset of data that replaces 2h
Classification from Positive and Biased Negative Data with Skewed Labeled Posterior Probability. (arXiv:2203.05749v1 [stat.ME]) The binary classifica 2h
Near-optimal Offline Reinforcement Learning with Linear Representation: Leveraging Variance Information with Pessimism. (arXiv:2203.05804v1 [cs.L(2h
Averaging Spatio-temporal Signals using Optimal Transport and Soft Alignments. (arXiv:2203.05813v1 [stat.ML]) Several fields in science, from genon 2h
Flexible Amortized Variational Inference in qBOLD MRI. (arXiv:2203.05845v1 [eess.IV]) Streamlined qBOLD acquisitions enable experimentally straigl 2h
FedSyn: Synthetic Data Generation using Federated Learning. (arXiv:2203.05931v1 [stat.ML]) As Deep Learning algorithms continue to evolve and be 2h
Today
Read Later
Manage Leo
FEEDS
All
Arxiv
math.ST
stat.ME updates on arXiv…
stat.ML updates on arXiv.…
Random
Statistics journals
Create New Folder
BOARDS
To read
Create New Board
stat.ML updates on arXiv.org
Universal Regression with Adversarial Responses. (arXiv:2203.05067v1 [cs.LG])
stat.ML updates on arXiv.org by Moïse Blanchard, Patrick Jaillet / 2d // keep unread // hide
Is this article about Deep Learning?
YES
NO
We provide algorithms for regression with adversarial responses under large classes of non-i.i.d. instance sequences, on general separable metric spaces, with provably minimal assumptions. We also give characterizations of learnability in this regression context. We consider universal consistency which asks for strong consistency of a learner without restrictions on the value responses. Our analysis shows that such objective is achievable for a significantly larger class of instance sequences than stationary processes, and unveils a fundamental dichotomy between value spaces: whether finite-horizon mean-estimation is achievable or not. We further provide optimistically universal learning rules, i.e., such that if



Search Convergence
Convergence
papers
13
Add external people
Martingale ⌘ 2
Channels
advisees
auditing
bestarm
calibration
cvar
e-bh-and-friends
forecasting
general
good-news
interactive-selective-inf
internal-paper-review
open-problems
opportunities
papers
philosophy
random
regret-power
talks
workshops-conferences
Add channels
Direct messages 1
papers
3 files
Tuesday, February 1st
distribution, and $\psi : \mathcal{P} \to \mathbb{R}$ is a functional with $\mathcal{P}$ as the set of all distributions on the data domain. This $\psi(P)$ can range from simple statistical summaries such as correlation coefficient, quantile, conditional value-at-risk, to model parameters such as regression coefficient and prediction error measurement.
Suppose we are given independent and identically distributed (i.i.d.) data of size n, say $X_1, \ldots, X_n$. A natural point estimate of $\psi(P)$ is $\hat{\psi}_n := \psi(\hat{P}_n)$, where $\hat{P}_n(\cdot) := (1/n)\sum_{i=1}^n I(X_i \in \cdot)$ is the empirical distribution constructed from the data, and $I(\cdot)$ denotes the indicator function.
Our approach to construct a confidence interval for ψ proceeds as follows. For each replication $b = 1, \ldots, B$, we resample the data set, namely independently and uniformly sample with replacement from $\{X_1, \ldots, X_n\}$ n times, to obtain $\{X_1^{*b}, \ldots, X_n^{*b}\}$, and evaluate the resample estimate $\psi_n^{*b} := \psi(P_n^{*b})$, where $P_n^{*b}(\cdot) = (1/n)\sum_{i=1}^n I(X_i^{*b} \in \cdot)$ is the resample empirical distribution. Our confidence interval is
$$\mathcal{I} = \left[\hat{\psi}_n - t_{B,1-\alpha/2}S,\ \hat{\psi}_n + t_{B,1-\alpha/2}S\right] \quad (1)$$
where
$$S^2 = \frac{1}{B}\sum_{b=1}^{B}\left(\psi_n^{*b} - \hat{\psi}_n\right)^2 \quad (2)$$
Here, S^2 resembles the sample variance of the resample estimates, but "centered" at the original point estimate $\hat{\psi}_n$ instead of the resample mean, and using B in the denominator instead of $B-1$ as in "textbook" sample variance. The critical value $t_{B,1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of t_B, the student t-distribution with degree of freedom B. That is, the degree of freedom of this t-distribution is precisely the resampling computation effort.
The interval $\mathcal{I}$ in (1) is defined for any positive integer $B \geq 1$. In particular, when $B = 1$, it becomes
$$\left[\hat{\psi}_n - t_{1,1-\alpha/2}\left|\psi_n^* - \hat{\psi}_n\right|,\ \hat{\psi}_n + t_{1,1-\alpha/2}\left|\psi_n^* - \hat{\psi}_n\right|\right] \quad (3)$$
fact, we have the following basic coverage guarantee for (1) and (3). First, consider the following condition that is standard in the bootstrap literature:
Assumption 1 (Standard condition for bootstrap validity). *We have $\sqrt{n}(\hat{\psi}_n - \psi) \Rightarrow N(0, \sigma^2)$ where $\sigma^2 > 0$. Moreover, a resample estimate ψ_n^* satisfies $\sqrt{n}(\psi_n^* - \hat{\psi}_n) \Rightarrow N(0, \sigma^2)$ conditional on the data $X_1, X_2, \ldots$ in probability as $n \to \infty$.*
5
In Assumption 1, "$\Rightarrow$" denotes convergence in distribution, and the conditional "$\Rightarrow$"-convergence in probability means $P(\sqrt{n}(\psi_n^* - \hat{\psi}_n) \leq x | \hat{P}_n) \xrightarrow{p} P(N(0, \sigma^2) \leq x)$ for any $x \in \mathbb{R}$, where "$\xrightarrow{p}$" denotes convergence in probability. Assumption 1 is a standard condition to justify bootstrap validity, and is ensured when $\psi(\cdot)$ is Hadamard differentiable (see Proposition 2 in the sequel which follows from Van der Vaart (2000) §23). This assumption implies that, conditional on the data, the asymptotic distributions of the centered resample estimate $\sqrt{n}(\psi_n^* - \hat{\psi}_n)$ and the centered original estimate $\sqrt{n}(\hat{\psi}_n - \psi)$ are the same. Thus, one can use the former distribution, which is computable via Monte Carlo, to approximate the latter unknown distribution. Simply put, we can use a "plug-in"
1, (1) is an asymptotically exact $(1-\alpha)$-level confide
tic exactness of Cheap Bootstrap). *Under Assumpt*
an asymptotically exact $(1-\alpha)$-level confidence inte
$$\mathbb{P}_n(\psi \in \mathcal{I}) \to 1 - \alpha$$
notes the probability with respect to the data $X_1, \ldots,$
hat, under the same condition to justify the validity
strap interval $\mathcal{I}$ has asymptotically exact coverage,
plain how Theorem 1 is derived, we first compare th
arXiv.org
A Cheap Bootstrap Method for Fast Inference
The bootstrap is a versatile inference method that has proven powerful in many statistical problems. However, when applied to modern large-scale models, it could face substantial computation...
arXiv
Message #papers
Aa

On the other hand, the multichart CUSUM will give a reasonably good approximation to the best possible performance at the intermediate points $\theta \neq \theta_i$, and, therefore, may be considered as a reasonable candidate for practical applications. The same asymptotic performance can be obtained by using a multichart S–R detection test.

Yet another possible (and asymptotically efficient) solution can be constructed based on the maximal invariant sequence $Y_n = X_n - X_1, \quad n \geq 2$. Specifically, we conjecture that building likelihood ratios for Y_n and applying the corresponding invariant S–R test N_A will allow one to obtain an asymptotically optimal solution (as $\alpha \to 0$) with respect to the average detection delay $\mathbb{E}_k(T - k \mid T \geq k)$ uniformly for every $k \geq 2$ in the class of invariant detection procedures $\Delta_\alpha = \{T : \sup_k \mathbb{P}_\infty (T < k + m \mid T \geq k) \leq \alpha\}$ that confines the supremum local PFA. In fact, because the invariant S–R statistic R_n is a non-negative submartingale with mean $\mathbb{E}_\infty R_n = n$, it follows that $\mathbb{P}_\infty(N_A < k + m \mid N_A \geq k) \leq m/A$. Choose $m_\alpha = O(|\log \alpha|)$ and $A = A_\alpha$ as a solution of the equation $m_\alpha / A_\alpha = \alpha$. Generalizing an argument in Tartakovsky (2005) may lead to the desired asymptotic optimality result. This problem will be addressed elsewhere. Note also that the global ARL2FA metric may not be a good choice for the FAR, because the sequence $\{Y_n\}_{n \geq 2}$ is not i.i.d.

5. DETECTION OF A CHANGE OCCURRING AT A FAR HORIZON

Universally Consistent Online Learning with Arbitrarily Dependent Responses

Steve Hanneke STEVE.HANNEKE@GMAIL.COM
Purdue University

Abstract

This work provides an online learning rule that is universally consistent under processes on (X, Y) pairs, under conditions only on the X process. As a special case, the conditions admit all processes on (X, Y) such that the process on X is stationary. This generalizes past results which required stationarity for the joint process on (X, Y), and additionally required this process to be ergodic. In particular, this means that ergodicity is superfluous for the purpose of universally consistent online learning.

Keywords: statistical learning theory, universal consistency, nonparametric estimation, stochastic processes, stationary processes, non-ergodic processes, online learning

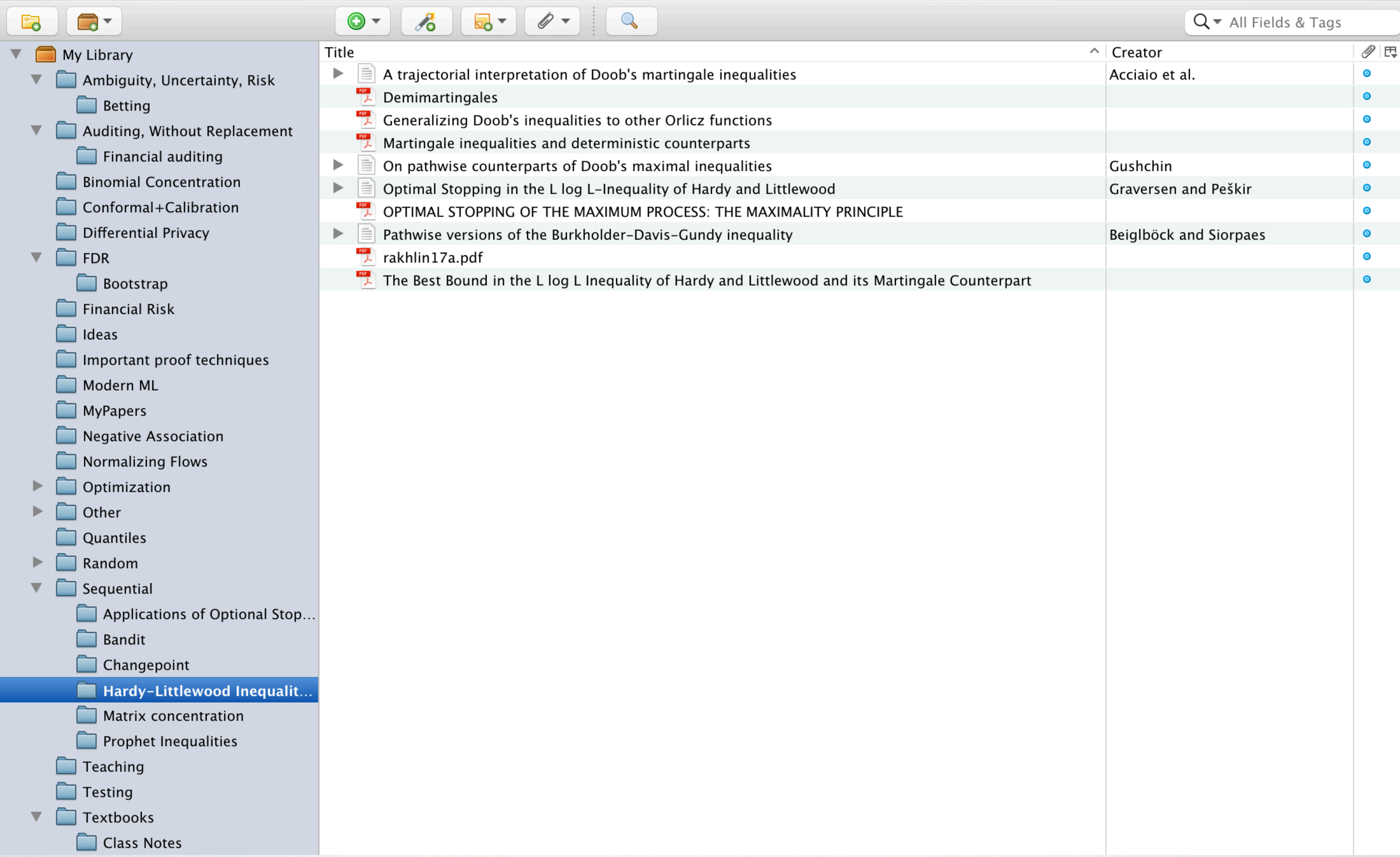
1. Introduction

The task of achieving low expected *regret* in online learning is a classic topic in learning theory. Specifically, we consider a sequential setting, where at each time t, a learner observes a point X_t, makes a *prediction* $\hat{Y}_t$, and then observes a true *response* Y_t: that is, $\hat{Y}_t = f_t(X_{1:(t-1)}, Y_{1:(t-1)}, X_t)$ for some function f_t (possibly randomized). We are then interested in the rate of growth of the long-run cumulative *loss* of the learner: i.e., $\sum_{t=1}^{T} \ell(\hat{Y}_t, Y_t)$, for a given loss function ℓ. However, as it may sometimes be impossible to achieve low cumulative loss in an absolute sense, we are often interested in understanding the *excess* loss compared to some particular *fixed* predictor f_0: i.e., $\sum_{t=1}^{T} \ell(\hat{Y}_t, Y_t) - \sum_{t=1}^{T} \ell(f_0(X_t), Y_t)$, known as the *regret* (relative to f_0).

Several different formulations of the subject have been proposed, leading to different algorithmic approaches and theoretical analyses of regret. For instance, there is a rich theory of online learning with *arbitrary* sequences $\{(X_t, Y_t)\}_{t=1}^{\infty}$, but where the reference function f_0 is restricted to belong to some particular function class $\mathcal{F}$ (see e.g., Cesa-Bianchi and Lugosi, 2006; Ben-David, Pál, and Shalev-Shwartz, 2009; Rakhlin, Sridharan, and Tewari, 2015).

Getting into a new area (BFS)

Zotero
Google Scholar
Author homepages
PhD theses
Survey articles
Blogs
Course notes
Recorded lectures


All Fields & Tags
My Library
Ambiguity, Uncertainty, Risk
Betting
Auditing, Without Replacement
Financial auditing
Binomial Concentration
Conformal+Calibration
Differential Privacy
FDR
Bootstrap
Financial Risk
Ideas
Important proof techniques
Modern ML
MyPapers
Negative Association
Normalizing Flows
Optimization
Other
Quantiles
Random
Sequential
Applications of Optional Stop…
Bandit
Changepoint
Hardy-Littlewood Inequalit…
Matrix concentration
Prophet Inequalities
Teaching
Testing
Textbooks
Class Notes
Title
Creator
A trajectorial interpretation of Doob's martingale inequalities
Acciaio et al.
Demimartingales
Generalizing Doob's inequalities to other Orlicz functions
Martingale inequalities and deterministic counterparts
On pathwise counterparts of Doob's maximal inequalities
Gushchin
Optimal Stopping in the L log L-Inequality of Hardy and Littlewood
Graversen and Peškir
OPTIMAL STOPPING OF THE MAXIMUM PROCESS: THE MAXIMALITY PRINCIPLE
Pathwise versions of the Burkholder-Davis-Gundy inequality
Beiglböck and Siorpaes
rakhlin17a.pdf
The Best Bound in the L log L Inequality of Hardy and Littlewood and its Martingale Counterpart

Deep dive into who cited a specific paper

[PDF] History of the efficient market hypothesis — [PDF] psu.edu

M Sewell - Rn, 2011 - Citeseer

A market is said to be efficient with respect to an information set if the price 'fully reflects' that information set, ie if the price would be unaffected by revealing the information set to all …

Save Cite Cited by 315 Related articles All 8 versions

Competitive on-line statistics — [PDF] wiley.com

V Vovk - International Statistical Review, 2001 - Wiley Online Library

A radically new approach to statistical modelling, which combines mathematical techniques of Bayesian statistics with the philosophy of the theory of competitive on-line algorithms, has …

Save Cite Cited by 312 Related articles All 16 versions

[PDF] Information, divergence and risk for binary experiments — [PDF] jmlr.org

M Reid, R Williamson - 2011 - jmlr.org

We unify f-divergences, Bregman divergences, surrogate regret bounds, proper scoring rules, cost curves, ROC-curves and statistical information. We do this by systematically …

Save Cite Cited by 178 Related articles All 23 versions

[BOOK] Forecasting volatility in the financial markets

S Satchell, J Knight - 2011 - books.google.com

Forecasting Volatility in the Financial Markets, Third Edition assumes that the reader has a firm grounding in the key principles and methods of understanding volatility measurement …

Save Cite Cited by 176 Related articles All 6 versions

Deep dive into specific authors

Google Scholar

Leonid A. Levin

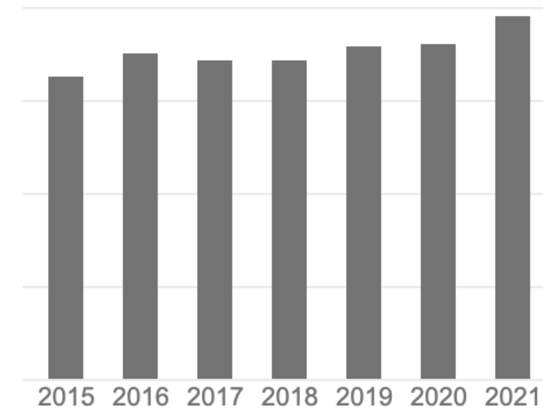
Unknown affiliation
Verified email at bu.edu - Homepage

FOLLOW

TITLE	CITED BY	YEAR
Universal sequential search problems LA Levin Problemy Peredachi Informatsii 9 (3), 115-116	1788 *	1973
A pseudorandom generator from any one-way function J Håstad, R Impagliazzo, LA Levin, M Luby SIAM J. Comput. 28 (4), 1364-1396	1757 *	1999
A hard-core predicate for all one-way functions O Goldreich, LA Levin Proceedings of the twenty-first annual ACM symposium on Theory of computing …	1436	1989
The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms AK Zvonkin, LA Levin Russian Mathematical Surveys 25, 83	1003 *	1970
Pseudo-random generation from one-way functions R Impagliazzo, LA Levin, M Luby Proceedings of the twenty-first annual ACM symposium on Theory of computing …	951	1989
Checking computations in polylogarithmic time L Babai, L Fortnow, LA Levin, M Szegedy Proceedings of the twenty-third annual ACM symposium on Theory of computing …	749	1991

Cited by

	All
Citations	12271
h-index	30
i10-index	40



Public access

0 articles

not available

Based on funding mandates

Co-authors

Parting thoughts

a. All papers have typos/mistakes. They're usually fixable (>95%), not fundamental errors (<5%).

b. Deep understanding can take weeks or months, even for experts. I still re-read fundamental papers in my area, and learn new things from them.

c. There is always a pyramid of understanding for important papers: lots of people understand things at a high level, and very few people outside the authors may understand the intricacies. Thus understanding a technical paper = an almost unique superpower!

Takeaway messages

1. What and how you read depends on goals and time constraints.
2. Ask the right questions for the goals.
3. Refine your understanding iteratively (non-linear reading).
4. Reading proofs is often about knowing what to gloss over.
5. Use Feedly/Scholar/Slack/Zotero to organize reading.