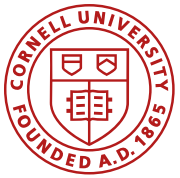
Invited talk for INFORMS Applied Probability Society

An Adversarial Analysis of Thompson Sampling for Full-information Online Learning: from Finite to Infinite Action Spaces

Joint work with Jeff Negrea

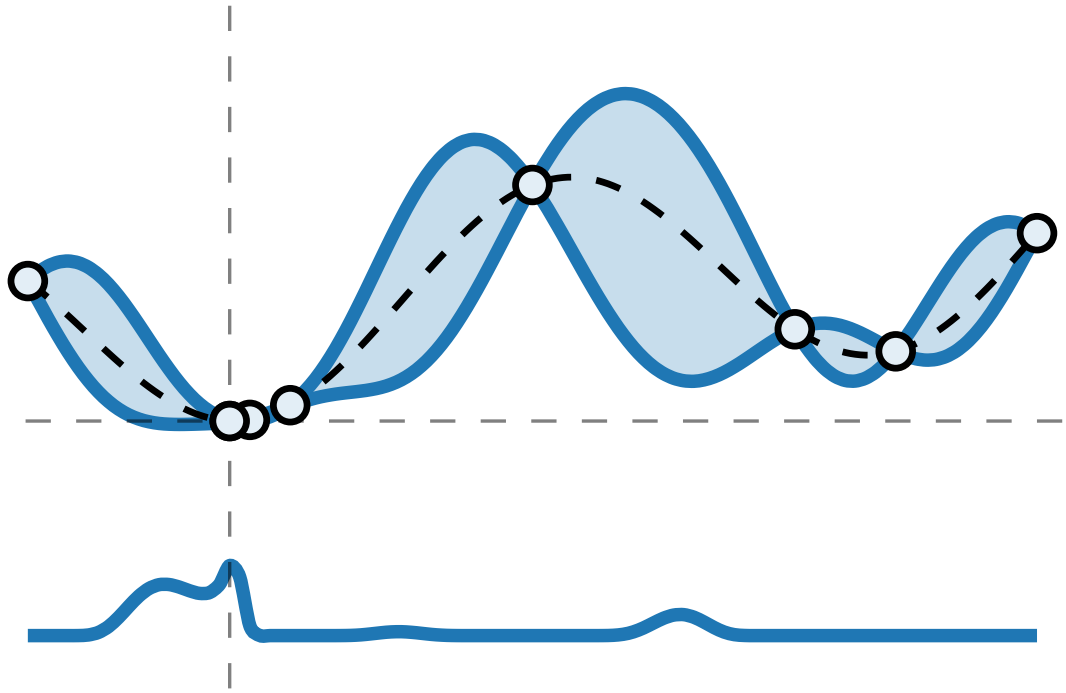


Cornell University

Alexander Terenin

HTTPS://AVT.IM/ · @AVT_IM

Probabilistic Decision-making



This talk: adversarial analogs of problems like this

The Online Learning Game

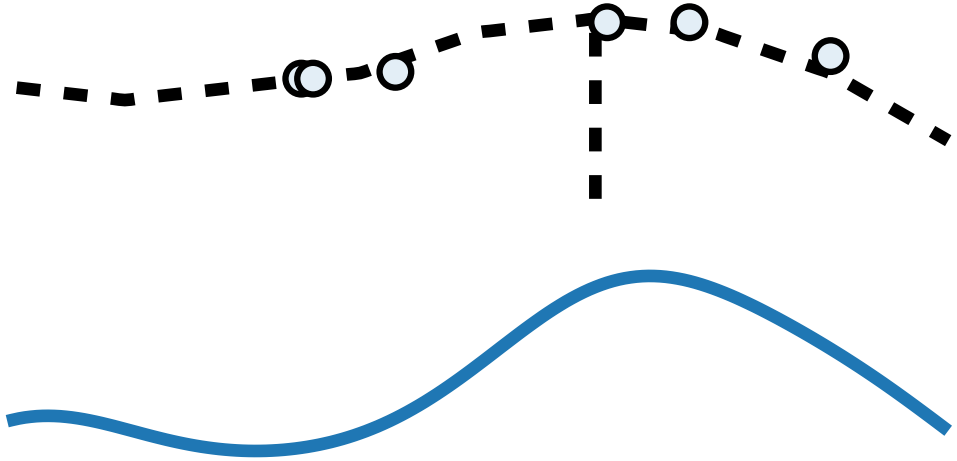
At each time $t = 1, .., T$:

1. Learner picks a random *action* $x_t \sim p_t \in \mathcal{M}_1(X)$.
2. Adversary responds with a *reward function* $y_t : X \to \mathbb{R}$, chosen adaptively from a given function class.

Regret:

$$R(p, y) = \mathop{\mathbb{E}}_{x_t \sim p_t} \underbrace{\sup_{x \in X} \sum_{t=1}^{T} y_t(x)}_{\substack{\text{single best action} \\ \text{in hindsight}}} - \underbrace{\sum_{t=1}^{T} y_t(x_t)}_{\substack{\text{total reward of} \\ \text{learner's actions}}}$$

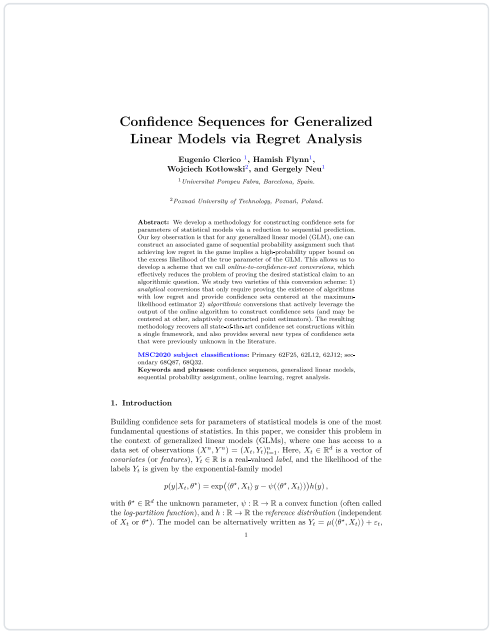
Non-discrete Online Learning



Adversarial problem: seems to have nothing to do with Bayes' Rule?

No-regret Online Learning Algorithms

Given various no-regret online learning oracles, one can obtain:



Confidence Sequences for Generalized Linear Models via Regret Analysis

Eugenio Clerico [1], Hamish Flynn[1], Wojciech Kotłowski[2], and Gergely Neu[1]

[1] Universitat Pompeu Fabra, Barcelona, Spain.

[2] Poznań University of Technology, Poznań, Poland.

Abstract: We develop a methodology for constructing confidence sets for parameters of statistical models via a reduction to sequential prediction. Our key observation is that for any generalized linear model (GLM), one can construct an associated game of sequential probability assignment such that achieving low regret in the game implies a high-probability upper bound on the excess likelihood of the true parameter of the GLM. This allows us to develop a scheme that we call *online-to-confidence-set conversions*, which effectively reduces the problem of proving the desired statistical claim to an algorithmic question. We study two varieties of this conversion scheme: 1) *analytical conversions* that only require proving the existence of algorithms with low regret and provide confidence sets centered at the maximum-likelihood estimator 2) *algorithmic conversions* that actively leverage the output of the online algorithm to construct confidence sets (and may be centered at other, adaptively constructed point estimators). The resulting methodology recovers all state-of-the-art confidence set constructions within a single framework, and also provides several new types of confidence sets that were previously unknown in the literature.

MSC2020 subject classifications: Primary 62F25, 62L12, 62J12; secondary 68Q87, 68Q32.

Keywords and phrases: confidence sequences, generalized linear models, sequential probability assignment, online learning, regret analysis.

1. Introduction

Building confidence sets for parameters of statistical models is one of the most fundamental questions of statistics. In this paper, we consider this problem in the context of generalized linear models (GLMs), where one has access to a data set of observations $(X^n, Y^n) = (X_t, Y_t)_{t=1}^n$. Here, $X_t \in \mathbb{R}^d$ is a vector of *covariates* (or *features*), $Y_t \in \mathbb{R}$ is a real-valued label, and the likelihood of the labels Y_t is given by the exponential-family model

$$p(y|X_t, \theta^*) = \exp(\langle \theta^*, X_t \rangle \, y - \psi(\langle \theta^*, X_t \rangle)) h(y) \,,$$

with $\theta^* \in \mathbb{R}^d$ the unknown parameter, $\psi : \mathbb{R} \to \mathbb{R}$ a convex function (often called the *log-partition function*), and $h : \mathbb{R} \to \mathbb{R}$ the *reference distribution* (independent of X_t or θ^*). The model can be alternatively written as $Y_t = \mu(\langle \theta^*, X_t \rangle) + \varepsilon_t$,

Confidence sets

Online-to-PAC Conversions: Generalization Bounds via Regret Analysis

Gábor Lugosi
ICREA, Universitat Pompeu Fabra, and Barcelona School of Economics, Barcelona, Spain

Gergely Neu
Universitat Pompeu Fabra, Barcelona, Spain

Abstract

We present a new framework for deriving bounds on the generalization bound of statistical learning algorithms from the perspective of online learning. Specifically, we construct an online learning game called the "generalization game", where an online learner is trying to compete with a fixed statistical learning algorithm in predicting the sequence of generalization gaps on a training set of i.i.d. data points. We establish a connection between the online and statistical learning setting by showing that the existence of an online learning algorithm with bounded regret in this game implies a bound on the generalization error of the statistical learning algorithm, up to a martingale concentration term that is independent of the complexity of the statistical learning method. This technique allows us to recover several standard generalization bounds including a range of PAC-Bayesian and information-theoretic guarantees, as well as generalizations thereof.

Keywords: statistical learning, generalization error, online learning, regret analysis

1. Introduction

We study the standard model of statistical learning. We are given a training sample of n i.i.d. data points $S_n = (Z_1, \ldots, Z_n)$ drawn from a distribution μ over a measurable *instance space* $\mathcal{Z}$. A *learning algorithm* maps the training sample to an output $W_n = \mathcal{A}(S_n)$ taking values in a measurable set $\mathcal{W}$ (called the *hypothesis class*) in a potentially randomized way. In other words, a randomized learning algorithm assigns, to any n-tuple of samples from $\mathcal{Z}$, a probability distribution over $\mathcal{W}$ and draws a sample from that distribution, independently of S_n. The resulting random element is denoted by W_n. More precisely, denoting the set of distributions over hypotheses by $\Delta_{\mathcal{W}}$, a learning algorithm can thus be formally written as a mapping $\mathcal{A} : \mathcal{Z}^n \to \Delta_{\mathcal{W}}$, and $W_n = \mathcal{A}(S_n)$.

We study the performance of the learning algorithm measured by a *loss function* $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}_+$. Two key objects of interest are the *risk* (or *test error*) $\mathbb{E}[\ell(w, Z')]$ and the *empirical risk* (or *training error*) $L(w, S_n) = \frac{1}{n} \sum_{i=1}^n \ell(w, Z_i)$ of a hypothesis $w \in \mathcal{W}$, where the random element Z' has the same distribution as Z_i and is independent of S_n. The ultimate goal in statistical learning is to find algorithms with small *excess risk*

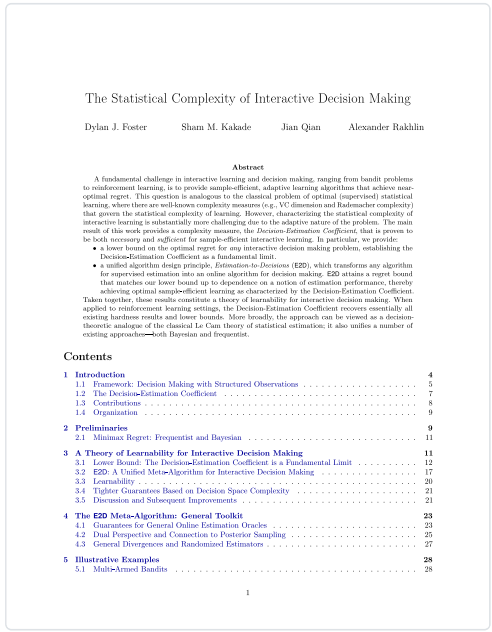
$$\mathcal{E}(W_n) = \mathbb{E}\left[\ell(W_n, Z') | W_n\right] - \inf_{w \in \mathcal{W}} \mathbb{E}\left[\ell(w, Z')\right] ,$$

which is typically decomposed into the task of minimizing the empirical risk $L(\cdot, S_n)$, and showing that the gap between the true risk and the empirical risk is small. This gap is commonly called the *generalization error* of the algorithm, and is defined formally as

$$\mathrm{gen}(W_n, S_n) = \mathbb{E}\left[\ell(W_n, Z') | W_n\right] - L(W_n, S_n) \,.$$

In other words, the generalization error measures the extent of *overfitting* occurring during training. As such, understanding (and more specifically, upper bounding) the generalization error has been in the center of focus of statistical learning theory ever since its inception. Over the past half century, numerous approaches have

Generalization bounds



The Statistical Complexity of Interactive Decision Making

Dylan J. Foster Sham M. Kakade Jian Qian Alexander Rakhlin

Abstract

A fundamental challenge in interactive learning and decision making, ranging from bandit problems to reinforcement learning, is to provide sample-efficient, adaptive learning algorithms that achieve near-optimal regret. This question is analogous to the classical problem of optimal (supervised) statistical learning, where there are well-known complexity measures (e.g., VC dimension and Rademacher complexity) that govern the statistical complexity of learning. However, characterizing the statistical complexity of interactive learning is substantially more challenging due to the adaptive nature of the problem. The main result of this work provides a complexity measure, the *Decision-Estimation Coefficient*, that is proven to be both necessary and *sufficient* for sample-efficient interactive learning. In particular, we provide:

- a lower bound on the optimal regret for any interactive decision making problem, establishing the Decision-Estimation Coefficient as a fundamental limit.
- a unified algorithm design principle, *Estimation-to-Decisions* (E2D), which transforms any algorithm for supervised estimation into an online algorithm for decision making. E2D attains a regret bound that matches our lower bound up to dependence on a notion of estimation performance, thereby achieving optimal sample-efficient learning as characterized by the Decision-Estimation Coefficient.

Taken together, these results constitute a theory of learnability for interactive decision making. When applied to reinforcement learning settings, the Decision-Estimation Coefficient recovers essentially all existing hardness results and lower bounds. More broadly, the approach can be viewed as a decision-theoretic analogue of the classical Le Cam theory of statistical estimation; it also unifies a number of existing approaches—both Bayesian and frequentist.

Contents

Sample-efficient reinforcement learning

No-Regret Learning Dynamics for Extensive-Form Correlated Equilibrium

Andrea Celli[*] Politecnico di Milano
Alberto Marchesi[*] Politecnico di Milano
Gabriele Farina[*] Carnegie Mellon University
Nicola Gatti Politecnico di Milano

Abstract

The existence of simple, uncoupled no-regret dynamics that converge to correlated equilibria in normal-form games is a celebrated result in the theory of multi-agent systems. Specifically, it has been known for more than 20 years that when all players seek to minimize their *internal* regret in a repeated normal-form game, the empirical frequency of play converges to a normal-form correlated equilibrium. Extensive-form (that is, tree-form) games generalize normal-form games by modeling both sequential and simultaneous moves, as well as private information. Because of the sequential nature and presence of partial information in the game, extensive-form correlation has significantly different properties than the normal-form counterpart, many of which are still open research directions. Extensive-form correlated equilibrium (EFCE) has been proposed as the natural extensive-form counterpart to normal-form correlated equilibrium. However, it was currently unknown whether EFCE emerges as the result of uncoupled agent dynamics. In this paper, we give the first uncoupled no-regret dynamics that converge to the set of EFCEs in n-player general-sum extensive-form games with perfect recall. First, we introduce a notion of *trigger regret* in extensive-form games, which extends that of internal regret in normal-form games. When each player has low trigger regret, the empirical frequency of play is close to an EFCE. Then, we give an efficient no-trigger-regret algorithm. Our algorithm decomposes trigger regret into local subproblems at each decision point for the player, and constructs a global strategy of the player from the local solutions at each decision point.

1 Introduction

The *Nash equilibrium* (NE) [37] is the most common notion of rationality in game theory, and its computation in two-player, zero-sum games has been the flagship computational challenge in the area at the interplay between computer science and game theory (see, *e.g.*, the landmark results in heads-up no-limit poker by Brown and Sandholm [5] and Moravčík et al. [34]). The assumption underpinning NE is that the interaction among players is fully *decentralized*. Therefore, an NE is a distribution on the *uncorrelated* strategy space (*i.e.*, a product of independent distributions, one per player). A competing notion of rationality is the *correlated equilibrium* (CE) proposed by Aumann [2]. A *correlated strategy* is a general distribution over joint action profiles and it is customarily modeled via a trusted external *mediator* that draws an action profile from this distribution, and

[*] Equal contribution.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Equilibrium computation algorithms

Online Learning: Discrete Action Spaces

Typical algorithm: mirror descent or follow-the-regularized-leader

$$p_t = \operatorname*{arg\,max}_{p \in \mathcal{M}_1(X)} \mathbb{E}_{x \sim p} \sum_{t=1}^{T} y_t(x) - \Gamma(p)$$

Parameterized by convex regularizer $\Gamma : X \to \mathbb{R}$

- Not obvious how to pick Γ in non-discrete settings
- Standard choice of KL ignores adversary's smoothness class
- Unclear how to perform numerics in general

Adversarial problem: seemingly nothing to do with Bayesian learning?

Bayesian Algorithms for Adversarial Learning

Idea: think of this game in a Bayesian way

1. Place a prior $q^{(\gamma)}$ on the *adversary's future reward functions* $\gamma_1, .., \gamma_T \in \mathcal{M}_1(Y)$
2. Condition on observed rewards $\gamma_1 = y_1, .., \gamma_{t-1} = y_{t-1}$
3. Draw a sample from the posterior, and play

$$x_t = \arg\max_{x \in X} \sum_{\tau=1}^{t-1} y_\tau(x) + \sum_{\tau=t}^{T} \gamma_\tau(x)$$

Algorithm: Thompson sampling

Bayesian Algorithms for Online Learning

Thompson sampling: FTPL with very specific learning rates

- Why should this be a good idea?

Result (Gravin, Peres, and Sivan, 2016). If $X = \{1, 2, 3\}$, the (infinite-horizon discounted) online learning game's Nash equilibrium is Thompson sampling with respect to a certain optimal prior.

Conjecture: Thompson sampling with strong priors is minimax-strong

This work: prove the finite and simplest infinite-dimensional case

Idea: Analyze Bayesian Algorithm in a Bayesian Way

Regret decomposition:

$$\underbrace{R(p,y)}_{\text{regret}} = \underbrace{R(p,q^{(\gamma)})}_{\text{prior regret}} + \underbrace{E_{q^{(\gamma)}}(p,y)}_{\text{excess regret}}$$

where

$$E_{q^{(\gamma)}}(p,y) = \sum_{t=1}^{T} \Gamma^*_{t+1}(y_{1:t}) - \Gamma^*_t(y_{1:t-1}) - \langle y_t | p_t \rangle + \mathbb{E}\left\langle \gamma_t | p_t^{(\gamma)} \right\rangle$$

and $\Gamma^*_t(f) = \sup_{x \in X} f(x) + \gamma_{t:T}(x)$.

Strong Priors for Adversarial Feedback

Strong prior over adversary:

1. Equalizing: $\mathbb{E}\,\gamma(x) = \mathbb{E}\,\gamma(x')$ for all x, x'
2. Certifies a sharp regret lower bound

$$R(\cdot, q^{(\gamma)}) \leq \min_{p} \max_{q} R(p, q)$$

Practical approximation: Gaussian process with matching smoothness

- Matérn priors: widely-used in Bayesian optimization today
- Straightforward and well-understood numerics

A Bregman Divergence Bound on Excess Regret

For a strong prior:

$$E_{q^{(\gamma)}}(p, y) \leq \sum_{t=1}^{T} \underbrace{D_{\Gamma_t^*}(y_{1:t} \mid\mid y_{1:t-1})}_{\substack{\text{Bregman divergence} \\ \text{induced by } \Gamma_t^*}}$$

Gaussian adversary: rates

- Finite X with ℓ^∞ adversary: $R(p, q) \leq 2\sqrt{T \log N}$
- $X = [0, 1]$, BL adversary: $R(p, q) \leq \mathcal{O}\left(\beta\sqrt{Td\log(1 + \sqrt{d}\frac{\lambda}{\beta})}\right)$

Hessian Bounds for Gaussian Adversaries

Taylor form of the Bregman divergence:

$$D_{\Gamma_t^*}(y_{1:t} \mid\mid y_{1:t-1}) = \frac{1}{2}\int_0^1 \underbrace{\partial^2_{y_t,y_t}\Gamma_t^*(y_{1:t-1} + \alpha y_t)}_{\text{Gâteaux Hessian}} \mathrm{d}\alpha$$

$$\partial^2_{u,v}\Gamma_t^*(f) = \frac{1}{\sqrt{T-t+1}}\,\mathbb{E}\,u(\underbrace{x^*_{f+\gamma_{t:T}}}_{\substack{\text{perturbed}\\ \text{maximizer}}})\,\langle \gamma_{t:T} | \underbrace{\mathcal{K}^{-1}}_{\substack{\text{covariance}\\ \text{operator}}} v\rangle$$

Classical finite-dimensional argument: $\|\Gamma_t^*(f)\|_{L_{\infty,1}} \leq 2\,\mathrm{tr}\,\Gamma_t^*(f)$

- Only works if $\mathcal{K}$ is the identity matrix
- Essentially all obvious workarounds give vacuous rates

Probabilistic Hessian Bounds

Idea: condition on maximizer and work with truncated normals

- Algebraic properties from discrete case have probabilistic analogs

Theorem. Suppose adversary satisfies $\|y\|_\infty \leq \beta$ and prior is IID over time with constant variance $k(x,x) = \sigma^2$. Suppose that

$$\sup_{y \in Y} y(x) - y(x') \frac{k(x,x')}{k(x',x')} \leq C_{Y,k} \left(1 - \frac{k(x,x')}{k(x',x')}\right).$$

Then we have

$$D_{\Gamma_t^*}(y_{1:t} \mid\mid y_{1:t-1}) \leq \frac{\beta^2 + \beta C_{Y,k}}{2\sigma^2 \sqrt{T-t+1}} \mathbb{E} \sup_{x \in X} \gamma_{t:T}(x).$$

Bounded Lipschitz Adversary

Bounded Lipschitz (β, λ) adversary: Matérn kernel with length scale κ

$$\sup_{y \in Y} y(x) - y(x') \frac{k(x, x')}{k(x', x')} \leq \underbrace{\frac{\beta(\lambda + \frac{1}{\kappa})}{\frac{1}{\kappa}\left(1 - e^{-\frac{2}{\lambda\kappa + 1}}\right)}}_{C_{Y,k}} \left(1 - \frac{k(x, x')}{k(x', x')}\right)$$

Thompson sampling regret:

$$R(p, q) \leq \beta \left(32 + \frac{32}{1 - \frac{1}{e}}\right) \sqrt{T d \log\left(1 + \sqrt{d}\frac{\lambda}{\beta}\right)}$$

Takeaways

Algorithmic design principle:

To explore by random actions, don't be too predictable, and match smoothness

Non-discrete online learning:

- Thompson sampling: perturbation based algorithm
- Bayesian viewpoint: match smoothness using Gaussian priors
- Analysis: probabilistic argument for non-discrete Hessian bounds

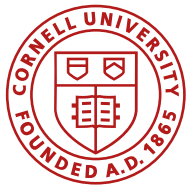
Future work:

- More general smoothness classes
- Learning in games
- Bandit feedback

Thank you!

https://avt.im/ · @avt_im

A. Terenin and J. Negrea. An Adversarial Analysis of Thompson Sampling for Full-information Online Learning: from Finite to Infinite Action Spaces. *arXiv:2502.14790*, 2025.



Cornell University