Robust Learning from Batches

The Best Things in Life are (Almost) Free

Alon Orlitsky

UC San Diego

Based on Joint Work with Ayush Jain

Overview

Fundamental learning tasks

Data corruption

Robust learning

Learning from batches

Density estimation

Discrete and Continuous

Classification

Learn (almost) as well as from genuine data

The Golden Age of Machine Learning

The Golden Age of Machine Learning

Many important applications accurately learnable with modest resources

- Amount of data
- Computation time

Build on theoretical advances in fundamental learning paradigms

- Hypothesis testing
- Density estimation $\longleftarrow$
- Classification $\longleftarrow$
- Regression
- Clustering
- Reinforcement learning
- .... Deep neural networks

Results

- As data $\nearrow$ error $\searrow$
- Polynomial-time algorithms

Density Estimation

Known distribution (density) class

Unknown distribution in class

Generates samples

Estimate distribution

Best estimate?

Discrete Distributions

Support set

Wolog $[k] = \{0, \ldots, k-1\}$

Set of distributions

Δ_k = { Distributions over $[k]$ }

Distribution

$p = (p_0, \ldots, p_{k-1}) \qquad p_i \geq 0 \qquad \sum p_i = 1$

s Samples

$X^s = X_1, \ldots, X_s \sim p$

Independent $\quad p(X^s = x^s) = \prod p_{x_i}$

Distribution Estimation

Unknown $p \in \Delta_k \quad \rightarrow \quad X^s \quad \rightarrow$ estimate of p

Estimator

$q^{\mathsf{est}} : [k]^s \rightarrow \Delta_k$

Estimate $q^{\mathsf{est}}(X^s)$

Distance measure

L_1 distance: $||p - q||_1 \stackrel{\text{def}}{=} \sum_{i=0}^{k-1} |p_i - q_i|$

Total-variation distance: $||p - q||_{\mathsf{TV}} \stackrel{\text{def}}{=} \sum_{i:p_i > q_i} (p_i - q_i) = \frac{1}{2} ||p - q||_1$

Distance of q^{est} from p when observing X^s

$d(q^{\mathsf{est}}(X^s), p)$

Particular q^{est}, X^s, p

Fundamental difficulty of whole estimation task

Min-Max Expected Loss

Remove X^s – Expectation

$$L_s(q^{\textsf{est}}, p) \stackrel{\text{def}}{=} \mathbb{E}_{X^s \sim p}\ d(q^{\textsf{est}}(X^s), p)$$

Remove p – Worst

$$L_s(q^{\textsf{est}}) \stackrel{\text{def}}{=} \max_{p \in \Delta_k} L_s(q^{\textsf{est}}, p)$$

Remove $q^{\textsf{est}}$ - Best

$$\begin{aligned} L_{k,s} &\stackrel{\text{def}}{=} \min_{q^{\textsf{est}}} L_s(q^{\textsf{est}}) \\ &= \min_{q^{\textsf{est}}}\ \max_{p \in \Delta_k}\ \mathbb{E}_{X^s \sim p}\ d(q^{\textsf{est}}(X^s), p) \end{aligned}$$

Expected Loss of the best estimator for worst distribution

Min-max Expected loss

$L_{k,s} = ?$

Binary and Larger Alphabets

$L_{2,s} \quad \to \quad L_{k,s}$

$X^s = X_1, \ldots, X_s \sim$ Ber(p) independently Estimate p

N – # 1's in X^s

$N \sim$ Bin(p, s) $\mathbb{E}(N) = sp$ $\sigma(N) = \sqrt{sp(1-p)}$

$q^{\mathsf{emp}}(X^s) \stackrel{\text{def}}{=} \frac{N}{s}$ $\mathbb{E}(\frac{N}{s}) = p$ $\sigma(\frac{N}{s}) = \sqrt{\frac{p(1-p)}{s}}$

$\mathbb{E}\,||(1-\frac{N}{s}, \frac{N}{s}) - (1-p, p)||_{\mathsf{TV}} = \mathbb{E}\,|\frac{N}{s} - p| = \Theta(\sigma(\frac{N}{s})) = \Theta(\sqrt{\frac{p(1-p)}{s}})$

$L_{2,s} = \Theta(\sqrt{\frac{1}{s}})$

$L_{k,s} = \sqrt{\frac{k-1}{2\pi s}} + o(\frac{1}{\sqrt{s}})$ [Kamath, O, Pichapati, Suresh 2015]

As $s \nearrow \infty$, $L_{k,s} \searrow 0$ ☺

Statistical limit

Trouble in Paradise

Corrupt Data

With big data come big problems

Samples oft corrupt

- Inadvertent
- Faulty
- Biased
- Malicious
- Adversarial, based on p and other samples

Identities of corrupt samples unknown

Can p still be learned accurately?

Robust Statistics

Early

Tukey, Huber, Donoho, 70's

Books

Robust Statistics; Huber, 1981

Robust Statistics and Influence Functions; Hampel et al, 1986

Robust Statistics and Outlier Detection; Rousseeuw and Leroy, 2003

Robust Estimation and Hypothesis Testing; Wilcox, 2011

Robust Statistics: Theory and Methods; Maronna, 2018

...

Recent

Efficient mean estimation in high dimensions

[Diakonikolas, Kamath, Kane, Li, Moitra, Stewart 2016]

[Lai, Rao, Vempala 2016]

[Charikar, Steinhart, Valiant 2017]

Common Model [Huber]

Parameter $\beta < 1/2$

Fraction $\leq \beta$ of samples corrupt

- Different distribution, biased, arbitrary, adversarial

Remaining $\geq 1 - \beta$ fraction of samples genuine $\sim p$

- Generalizations to roughly p

Typically estimate distribution parameters

- Median of s samples estimates univariate Gaussian mean to $\mathcal{O}(\frac{\sigma}{\sqrt{n}} \vee \beta\sigma)$
- $\vee$ – $\max$

If $\beta > 1/2$, cannot know which distribution genuine

Little Secret

Even as $s \nearrow \infty$, error does not $\searrow 0$

Hard limit on performance in the presence of corrupt data

- Hypothesis testing
- Density estimation $\longleftarrow$
- Classification

The Source of All Evil

$k = 2$, β fraction of adversarial samples

Two possible distributions: Ber$(\frac{1}{2} - \frac{\beta}{2}) = (\frac{1}{2} + \frac{\beta}{2}, \frac{1}{2} - \frac{\beta}{2})$, Ber$(\frac{1}{2} + \frac{\beta}{2})$

Ber$(\frac{1}{2} - \frac{\beta}{2})$: # 1's in genuine smpls $\approx s(1-\beta)(\frac{1}{2} - \frac{\beta}{2}) = s(\frac{1}{2} - \frac{\beta}{2} - \frac{\beta}{2} + \frac{\beta^2}{2}) \gtrsim s(\frac{1}{2} - \beta)$

Adversary can add βs 1's, force $\frac{s}{2}$ 0's and 1's, Similarly for Ber$(\frac{1}{2} + \frac{\beta}{2})$

Same overall samples for both Ber$(\frac{1}{2} - \frac{\beta}{2})$ and Ber$(\frac{1}{2} + \frac{\beta}{2})$

Underlying distribution cannot be determined better than random

$||$Ber$(\frac{1}{2} - \frac{\beta}{2})$ − Ber$(\frac{1}{2} + \frac{\beta}{2})||_{TV} = \beta$

Triangle inequality

Any estimated distribution is at distance $\geq \frac{\beta}{2}$ from one of two distributions

While $L_{2,s} \approx \frac{1}{\sqrt{2\pi s}} \searrow 0$, $L_{2,s,\beta} \geq \frac{\beta}{2}$ for all s

Implications

Accuracy does not improve with sample size s

E.g., for $\beta = 0.2$, with however many samples, $L_{2,0.2,s} \geq 0.1$

Propagates to all learning problems!

End is near?

There is hope

Batches

Many applications: samples collected from multiple sources

Each provides a batch of samples

- Sensor networks
- Recommendation systems
- Natural language processing
- Crowd sourcing
- Federated learning

Can batched data be used for robust learning?

Faulty Batches

Most batches genuine

Often some are not

- Faulty sensors
- Biased feedback
- Wrongly attributed texts
- Malicious sources - may even falsify data based on other samples

In some applications significant fraction of batches unreliable

- CNN: 5% of active Facebook accounts are fake
- Harvard: 20% of Yelp reviews fake
- BBC: Fake Amazon reviews cost £5
- ClickCease: 14% of ad campaign clicks fraudulent
- Finance: Analysis is only as clean as your data

Model [Qiao and Valiant '17]

k alphabet size

m batches

n samples each

Good batches: i.i.d. samples from p

Can be relaxed

Adversarial batches: arbitrary, may depend on p, even on good batches

β — upper bound on fraction of adversarial batches

$L_{k,m,n,\beta} = ?$

Why Batches Help

Binary alphabet, $k = 2$

Three batches, $m = 3$

$\beta = 1/3 \quad \rightarrow \quad$ 2 batches genuine $\sim p$, 1 batch adversarial

As batch size $n \to \infty$, genuine batches $\to p$ as $\mathcal{O}(1/\sqrt{n})$

Find distribution q within $\mathcal{O}(1/\sqrt{n})$ from two batches

Exists, e.g. p

q within $\mathcal{O}(1/\sqrt{n})$ from p

Adversarial Lower Bound [Donoho and Liu 1988]

General framework, follows from minimum distance functionals

$\{f_\theta : \theta \in \Theta\}$ parametric distribution family

s samples, $1 - \beta$ fraction $\sim f_\theta$ for some $\theta \in \Theta$, rest adversarial

Estimate θ in distance measure d

General lower bound

For any s and β, $\quad L_{s,\beta} \geq \frac{1}{2} \max\{d(\theta, \theta') : \|f_\theta - f_{\theta'}\|_{\mathsf{TV}} \leq \beta\}$

Proof similar to binary example

$\|f - f'\|_{\mathsf{TV}} \leq \beta \quad \rightarrow \quad \exists g$, g' such that $(1-\beta)f + \beta g = (1-\beta)f' + \beta g'$

Adversary can make overall distributions appear same

Prior Work

Lower bound

Each batch can be viewed as $\sim \mathsf{Mul}(p, n)$

Falls in general adversarial framework [Donoho and Liu 1988]

[Qiao and Valiant '17] applied adversarial lower bound to $f_p = \mathsf{Mul}(p, n)$

For k, m, n, and $\beta < 1/2$,

$$L_{k,m,n,\beta} \geq \frac{1}{2} \max\{||p - p'||_{\mathsf{TV}} : ||\mathsf{Mul}(p, n) - \mathsf{Mul}(p', n)||_{\mathsf{TV}} \leq \beta\}$$
$$\geq \frac{\beta}{2\sqrt{2n}}$$

As in binary example, applies to $k = 2$, $m \to \infty$ batches

Adversarial lower bound

Upper bound

[Qiao and Valiant '17] derived an estimator $q^{\mathsf{Q}V}$

- For $\beta \leq 1/900$ $\quad L_{k,m,n,\beta}(q^{\mathsf{Q}V}) = \mathcal{O}\left(\frac{\beta}{\sqrt{n}} \vee \sqrt{\frac{k+n}{mn}}\right)$
- $q^{\mathsf{Q}V}$ runs in time exponential in k

[Jain, O '19] Polynomial-time estimator with near optimal complexity + line

[Chen, Li, Moitra '19] Quasi polynomial time and sample complexity

Near Optimal Learning in Polynomial Time

Loss lower bounds: Statistical $\Omega\Big(\sqrt{\frac{k}{m \cdot n}}\Big)$ $\quad m \cdot n = s$

Adversarial $\Omega\Big(\frac{\beta}{\sqrt{n}}\Big)$

Estimator $q^{\mathsf{n}ew}$

Polynomial-time estimator, for all $\beta \le 0.49$, k, n, m

$$L_{k,n,m,\beta}(q^{\mathsf{n}ew}) \le \mathcal{O}\Big(\frac{\beta}{\sqrt{n}} \cdot \sqrt{\log \frac{1}{\beta}} \vee \sqrt{\frac{k}{m \cdot n}}\Big)$$

Works for all $\beta \le 0.49$

Achieves both lower bounds: Statistical to constant factor

Adversarial to small $\sqrt{\log 1/\beta}$ factor

No tradeoff

Polynomial time

First to allow implementation and simulations

Robustness is (Almost) Free

k - alphabet size n - batch size m - # batches β - adversarial fraction

$$\Omega\left(\frac{\beta}{\sqrt{n}} \vee \sqrt{\frac{k}{m \cdot n}}\right) \le L_{k,n,m,\beta} \le \mathcal{O}\left(\frac{\beta}{\sqrt{n}} \cdot \sqrt{\log \frac{1}{\beta}} \vee \sqrt{\frac{k}{m \cdot n}}\right)$$

Statistical lower bound

- $\sqrt{k/(mn)}$ – Even for genuine samples

Adversarial lower bound

- Individual samples - $\beta/2$
- Batches - $\beta/\sqrt{n}$

If desired error

- Below adversarial lower bound – cannot
- Above lower bound $(\times\sqrt{\log(1/\beta)})$ – can achieve statistical lower bound
- $\beta = 0.1,\ n = 1000 \quad \rightarrow \quad \beta\sqrt{\log(1/\beta)}/\sqrt{n} \approx 0.005$
- Robustness (almost) free

Experiments

p random distribution in Δ_k

Different adversarial distributions with varied TV distances from p

Show results for worst adversary

Compare algorithm's performance to two estimators

Naive empirical estimator

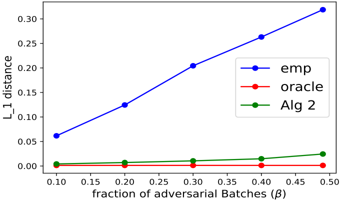
- Does not utilize batch structure
- Estimates p as empirical distribution of all samples
- May incur loss $\geq \beta/2$

Oracle

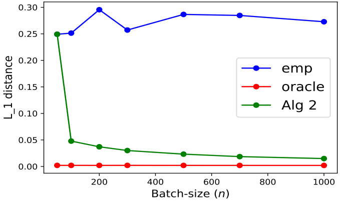
- Knows identity of adversarial and good batches
- Estimate as empirical distribution of good batches
- Not affected by adversarial batches
- Achieves statistical lower bound $\Theta\left(\sqrt{\frac{k}{m \cdot n(1-\beta)}}\right)$

Results

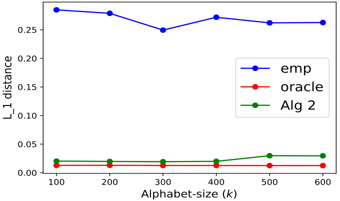
(a)+(b): m chosen so # good samples $m \cdot n \cdot (1 - \beta)$ is large constant so statistical limit stays same, m large so adversarial bound dominates.



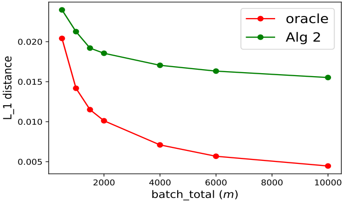
(a) $n = 1000,\ k = 200$



(b) $\beta = 0.4,\ k = 200$



(c) $\beta = 0.4,\ n = 1000,\ m = \frac{k}{\beta^2}$



(d) k=200,n=1000, β=0.4

Algorithm in a Nutshell

Filtering algorithm: Removes suspected bad batches

Empirical frequency on remaining batches

Binary distributions: Median of batch means approximates p to $1/\sqrt{n}$

Remove batches with mean ≥ $\sqrt{\log(1/\beta)/n}$ away from median

General k: TV distance is highest probability difference over all 2^k subsets

Small k: Remove outlier batches for all 2^k subsets

Large k: Intractable, every batch may be an outlier for some subset

Want subsets with fat tails, as contain more outlier (adversarial) batches

Statistical measure capturing effect of outliers on a subset probability

Easy to find a subset among 2^k where measure is approximately highest

Recursively remove outliers till measure small for all subsets

Time complexity linear in mn, small polynomial in k

Statistical Measure: Binomial Distributions

Consider $X_1, X_2, ..., X_m \sim \text{Bin}(n, p)$

$\mathbb{E}[X_i] = np$ and $\text{Var}(X_i) = np(1-p)$

Let $\hat{p} = \frac{1}{mn}\sum_i X_i$, as $m \to \infty$, $\hat{p} \to p$

Two variance estimates

- First moment: $\mathbb{V}^1 = n\hat{p}(1-\hat{p})$
- Second moment: $\mathbb{V}^2 = \frac{1}{m}\sum_i (X_i - n\hat{p})^2$

If all samples geniune then $\mathbb{V}^2 - \mathbb{V}^1 \to 0$ as $m \to \infty$

If βm samples are corrupt so that $|\hat{p} - p|$ is large then tail is fat

Fat tails increase second-moment $\mathbb{V}^2$ more than first-moment $\mathbb{V}^1$

A large value of $\mathbb{V}^2 - \mathbb{V}^1$ indicates fat tail

Idea generalizes to Multinomial

Statistical Measure: Multinomial Distributions

Consider $X_1, X_2, ..., X_m \sim$ Mul(n, p)

$\mathbb{E}[X_i] = np$ and Cov$(X_i) = n($Diag$(p) - pp^\top)$

Let $\hat{p} = \frac{1}{mn}\sum_i X_i$, as $m \to \infty$, $\hat{p} \to p$

Two covariance estimates

First moment: $\mathbb{V}^1 = n($Diag$(\hat{p}) - \hat{p}\hat{p}^\top)$

Second moment: $\mathbb{V}^2 = \frac{1}{m}\sum_i (X_i - n\hat{p})(X_i - n\hat{p})^\top$

$TV(p, \hat{p}) = \max_S |p(S) - \hat{p}(S)| = \max_{u \in \{0,1\}^k} |p \cdot u - \hat{p} \cdot u|$

For small TV, ensure no binary vector u corresponds to a fat tail

Identify fat tails by finding $\arg\max_{u \in \{0,1\}^k} |u^\top(\mathbb{V}^2 - \mathbb{V}^1)u|$, NP-hard

SDP approximation (Alon and Naor, 2004) finds u^* for which the above quantity $\geq$ half the maximum

Show it suffices

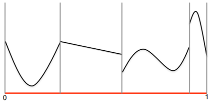
Piecewise Polynomial Distributions

t-piece degree-d distribution – t pieces, each a degree-d polynomial

$\mathcal{P}_{t,d}$ – all t-piece degree-d distributions

$\mathcal{P}_{t,0}$ histograms

$\mathcal{P}_{t,1}$ piecewise-linear distributions

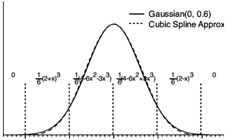


4-piece degree-3 distribution

Approximate any piecewise continuous distribution with large enough t, d

Approximate many staple distribution families with very low t and d

Gaussians, mixtures, log-concave, low-modal



Gaussian approximation by 4-piece degree-3 distribution

Robust Learning Piecewise-Polynomial Distributions

Loss lower bounds: Statistical $2 \cdot d(p, \mathcal{P}_{t,d}) + \Omega\left(\sqrt{\frac{t(d+1)}{m \cdot n}}\right)$

Adversarial $2 \cdot d(p, \mathcal{P}_{t,d}) + \Omega\left(\frac{\beta}{\sqrt{n}}\right)$

Estimator q^{new}

Polynomial-time estimator, for all $\beta <$ universal constant, $\alpha \approx 3$, t, d

$$L(q^{\mathsf{new}}) \le \alpha \cdot d(p, \mathcal{P}_{t,d}) + \mathcal{O}\left(\frac{\beta}{\sqrt{n}} \cdot \sqrt{\log \frac{1}{\beta}}\right) \vee \tilde{\mathcal{O}}\left(\sqrt{\frac{t(d+1)}{mn}}\right)$$

Use our algorithm to find p' close to p in $\mathcal{A}_k$ distance for $k = t(d+1)$

Use algorithm in [ADLS 17] for this p'

Same comments as before

First robust estimation algorithm for continuous distributions from batches

Allows first simulations

Related Work

For structured discrete distributions

[Chen, Li, Moitra '19] quasi-polynomial time

[Chen, Li, Moitra '20] polynomial time, suboptimal in # batches

Experiments

Compared to same two estimators

Naive empirical estimator: Does not utilize batch structure

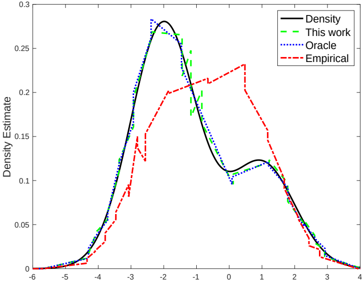
Oracle: Knows identity of good batches, uses them alone

$\beta = 0.4$, $n = 500$, $m(1 - \beta) = 62$

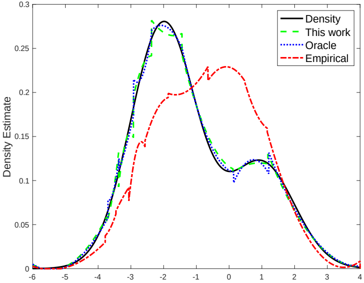
Gaussian Mixtures

Genuine distribution: $0.7\mathcal{N}(-2,1) + 0.3\mathcal{N}(1,1)$

Adversarial distribution: $\mathcal{N}(0,1)$



(a) Piecewise linear polynomials

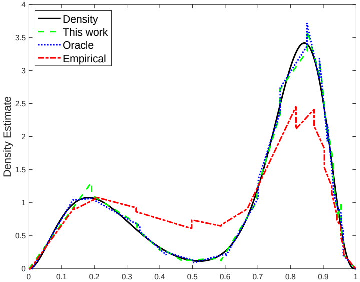


(b) Piecewise degree-2 polynomials

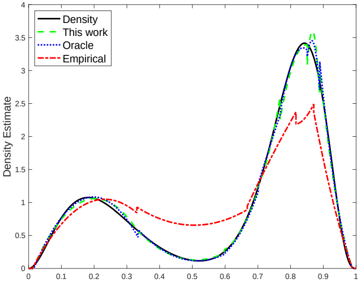
Beta Mixtures

Genuine distribution: $0.7\,\text{Beta}(17,4) + 0.3\,\text{Beta}(3,10)$,

Adversarial distribution: $\text{Beta}(2,2)$



(a) Piecewise linear polynomials



(b) Piecewise degree-2 polynomials

Classification

Hypothesis class $\mathcal{H}$ – family of Boolean functions $\Omega \to \{0,1\}$

Finite VC dimension $V_{\mathcal{H}}$

Excess loss of classifier: Classifier error probability - $\min_{h\in\mathcal{H}} h$ error probability

h^{ERM} empirical risk minimizer – $h \in \mathcal{H}$ with lowest empirical error

With s genuine samples, h^{ERM} achieves excess loss $\mathcal{O}\left(\sqrt{\frac{V_{\mathcal{H}}}{s}}\right)$

Min-max optimal over all classifiers

Robust Classification

With β fraction adversarial samples, exc. loss of h^{ERM} may be $\geq\Omega(\beta)$

Excess loss lower bounds: Statistical $\Omega\Big(\sqrt{\frac{V_{\mathcal{H}}}{m\cdot n}}\Big)$

Adversarial $\Omega\big(\frac{\beta}{\sqrt{n}}\big)$

Estimator h^{new}

For all $\beta < 1/2$, $\mathcal{H}$, n, m, p, expected excess loss of h^{new}

$$\leq \mathcal{O}\Big(\frac{\beta}{\sqrt{n}} \cdot \sqrt{\log \frac{1}{\beta}}\Big) \vee \tilde{\mathcal{O}}\Big(\sqrt{\frac{V_{\mathcal{H}}}{mn}}\Big)$$

Key idea: Learn p robustly in distance defined by $\mathcal{H}$, then ERM

Achieves adversarial batch lower bound to small $\sqrt{\log 1/\beta}$ factor

Achieves statistical lower bound to log factors of $\frac{n}{\beta}$

Robustness (almost) free

Review

Robust learning – some samples corrupt

- β fraction of samples corrupt
- Hard limit on accuracy, for any number of samples

Robust learning from batches

- Arises in many natural applications – sensors, recommendations, NLP
- β fraction of batches corrupt even adversarial

Firsts (robust learning from batches)

- Computationally efficient algorithm + essentially optimal
- Simulations for discrete distributions + positive results
- Estimation of continuous distributions + near optimal + efficient for $\mathcal{P}_{t,d}$
- Classification + near optimal + efficient for interval classification

The best things in life are (almost) free

- Up to adversarial bound achieve same accuracy as for genuine samples

Thank You!

References

Optimal Robust Learning of Discrete Distributions from Batches

Ayush Jain and O, ICML 2019

A General Method for Robust Learning from Batches

Ayush Jain and O, NeurIPS 2020

Robust Learning from Batches: The Best Things in Life are (Nearly) Free

Ayush Jain and O, ICML 2021