Beyond Theory:
Realizing Robustness

Gautam Kamath
MIT → This workshop → Simons Institute → University of Waterloo
Workshop on Computational Efficiency and High-Dimensional Robust Statistics
August 17, 2018
Huber’s (1997) Call to Arms

“It is one thing to design a theoretical algorithm whose purpose is to prove [large fractions of corruptions can be tolerated] and quite another thing to design a practical version that can be used not merely on small, but also on medium sized regression problems, with a 2000 by 50 matrix or so. This last requirement would seem to exclude all of the recently proposed [techniques].”
Classic Motivations

• Model misspecification
  • Nature doesn’t actually sample from Gaussians

• Dirty datasets
  • Some measurement/recording errors
  • Data from multiple inconsistent sources
Modern Motivations

• Data poisoning and adversarial attacks
• Current machine learning systems are surprisingly brittle

Figure from [Goodfellow Shlens Szegedy ‘14]
Modern Motivations

From [Gu Dolan-Gavitt Garg ‘17]
Modern Motivations

From [Athalye Engstrom Ilyas Kwok, ICML ‘18]
Unsupervised Learning

Being Robust (in High Dimensions) Can Be Practical

[Diakonikolas K Kane Moitra Li Stewart, ICML ‘17]
Does the filter “work”?

- 90% Gaussian data, 10% adversarial noise
- Isotropic Gaussian
  - Estimate mean
  - Estimate covariance
- Skewed Gaussian
  - Estimate covariance
Synthetic Experiments, Unknown Mean

Code: https://github.com/hoonose/robust-filter
Synthetic Experiments, Unknown Covariance

Code: https://github.com/hoonose/robust-filter
Exploratory Data Analysis

Being Robust (in High Dimensions) Can Be Practical

[Diakonikolas K Kane Moitra Li Stewart, ICML ‘17]
Robust PCA

• Our setting: incomparable with Robust PCA setting of Candes et al.
Gene Expression PCA Contains Europe

• Genes Mirror Geography in Europe. [Novembre et al.], Nature ‘08

Code: https://github.com/hoonose/robust-filter
Naively, Corruptions Destroy Europe

• Genes Mirror Geography in Europe. [Novembre et al.], Nature ‘08

Code: https://github.com/hoonose/robust-filter
Europe is RANSACked

- Genes Mirror Geography in Europe. [Novembre et al.], Nature ‘08

Code: https://github.com/hoonose/robust-filter
Robust PCA SDPs couldn’t save them...

- Genes Mirror Geography in Europe. [Novembre et al.], Nature ‘08

Code: https://github.com/hoonose/robust-filter
Our Algorithms Fix Europe!

- Genes Mirror Geography in Europe. [Novembre et al.], Nature ‘08

Code: https://github.com/hoonose/robust-filter
Supervised Learning

Sever: A Robust Meta-Algorithm for Stochastic Optimization
[Diakonikolas K Kane Li Steinhardt Stewart ‘18]
Beyond Robust Statistics

• Can we optimize more complicated objectives with corruptions?
  • Distribution $D$ over $(X, y)$ pairs
  • Loss function $\ell(X, y, w)$
• Given an $\varepsilon$-corrupted set of samples from $D$, find $w$ that minimizes
  
  $$f(w) = \mathbb{E}_{(X,y) \sim D}[\ell(X, y, w)].$$

• Examples:
  • Linear regression: $\ell(X, y, w) = (y - \langle w, x \rangle)^2$
  • SVMs: $\ell(X, y, w) = \max(0, 1 - y \langle w, x \rangle)$
  • GLMs
Stochastic Optimization

• Gradient descent:
\[ w_{t+1} \leftarrow w_t - \eta_t \cdot \frac{1}{n} \sum \nabla \ell(X_i, y_i, w_t) \]

• Want to follow \(-\nabla f(w_t)\)
• The empirical gradient works in the vanilla setting:
  • \( \nabla f(w_t) = \mathbb{E}[\nabla \ell(X, y, w)] = \frac{1}{n} \sum \nabla \ell(X_i, y_i, w_t) \)
• But what if some \((X_i, y_i)\) are corrupted?
• How do we robustly estimate \(\mathbb{E}[\nabla \ell(X, y, w)]\)?
Sever: Robust Stochastic Optimization

• How do we robustly estimate $\mathbb{E}[\nabla \ell(X, y, w)]$?
• Modified gradient descent:

$$w_{t+1} \leftarrow w_t - \eta_t \cdot g_t$$

• $g_t$ is a robust estimate of $\nabla f(w_t)$
  • Obtained via robust estimators from earlier
  • Bounded moments of $X$ often suffice to bound moments of $\nabla \ell(X, y, w)$

Same idea used in [Prasad Suggala Balakrishnan Ravikumar ‘18]
Sever

• If $\text{Cov}[\nabla \ell(X, y, w)] \ll \sigma^2 I$, then Sever locates an $O(\sigma \sqrt{\varepsilon})$-approximate critical point
  • Based on a “second-moment” filter

• If $\ell(X, y, w)$ is convex, can approximate optimal parameters:
  $$f(\hat{w}) - \argmin_w f(w) \leq O(\sigma \sqrt{\varepsilon})$$

• Specific sample complexity results for linear regression, SVM, logistic regression
Making it practical

- Problem: Gradient descent is fast, filter is (comparatively) slow
- Solution: Run filter once GD has converged to “sever” outliers
- Same(ish) theoretical guarantees, much faster in practice
- Even simpler: removing some hyperparameters
- Project onto top singular vector of gradients, remove $\frac{\epsilon n}{k}$ most extreme points, repeat $k$ times
Attacks

• How do we know a defense works?
  • Generate effective attacks
• Data poisoning attacks of [Steinhardt Koh Liang, NIPS’17]
• If the attacker knew the defender’s strategy, what should he do?
  • Generally a hard problem...
• If defender’s strategy is “fixed” (not data dependent), can generate nearly optimal attacks using no-regret learning
• With “simple” data-dependent defenses, effective heuristic methods
• Forthcoming work bypasses more defenses [Koh Steinhardt Liang, ‘??]
Experiments

- Ridge regression and Support Vector Machines (SVMs)
- Synthetic and real datasets
  - Drug discovery (regression) and Enron spam (classification)
- Generated large suite of attacks for a range of $\varepsilon$ (from 0.5% to 10%)
- Comparison: other baselines which attempt to remove “large” points
  - Large norm, loss, norm of gradient, or distance of gradient from mean
Ridge Regression

Code: coming soon...
SVMs, synthetic data

Code: coming soon...
SVMs, Enron dataset

Code: coming soon...
Conclusions

• Robustness is real and better than ever!
• Useful for data analysis in unsupervised and supervised settings
• Next steps: practical tools for more real-world settings