1F4-3

学習の安定化と最適なノイズ抑制

Stabilization of learning algorithms and optimal suppression of noise

マシュー ホーランド　池田和司
Matthew J. Holland　Kazushi Ikeda

奈良先端科学技術大学院大学
Nara Institute of Science and Technology (NAIST)

The chief topic of discussion is the theory and practice of systematically robustifying learning algorithms such that they realize provably stable performance over wide classes of data distributions. One concrete approach of interest is to focus on the impact that highly reliable estimates of algorithm objective functions has on off-sample predictive performance. We show that one may introduce nearly mechanical modifications to many popular algorithms which results in a distinct robustness property. As well, the notion of "how much information to throw away" optimally is introduced and basic results are given.

1. Background

In many statistical estimation tasks, a common paradigm for designing reliable procedures is to use a combination of heuristics and rigorous formal guarantees. The theoretical results typically are very appealing, with the caveat that they only hold under rather restrictive assumptions on the underlying data distribution. Using such algorithms in practice provides additional insight, and we build a picture of the conditions under which a given routine succeeds, and where it tends to break down or become unstable in its performance. Ideally, one would hope that formal and empirical insights would align, such that even with highly incomplete *a priori* information on the data, learning algorithms perform "as we expect."

Our talk focuses primarily on the theory and practice surrounding the problem of how to systematically robustify wide classes of learning algorithms. The present reality is that many popular routines in common use are in fact very sensitive to contaminated data or extreme observations when sample sizes are small. In pursuing algorithm design such that our routines perform (with high probability) as we expect them to, it is clear that the fundamental goal is a form of robustness. This important notion has been present in the statistics literature since the 1960s and there is a rich body of literature [Hampel et al., 1986, Staudte and Sheather, 1990, Huber and Ronchetti, 2009]. In the machine learning community as well, recent work such as strategic data sub-sampling techniques under partial corruption [McWilliams et al., 2014], engineering objective functions resistant to extreme observations [Candès et al., 2011], truncating the processed observations [Audibert and Catoni, 2011], removing noisy or irrelevant features through regularized objectives designed to suit the underlying distribution [Soltanolkotabi et al., 2014], as well as a prominent workshop at NIPS 2010 on "robust statistical learning," including talks from eminent researchers such as Peter Bickel and Emmanuel Candès on the topic.

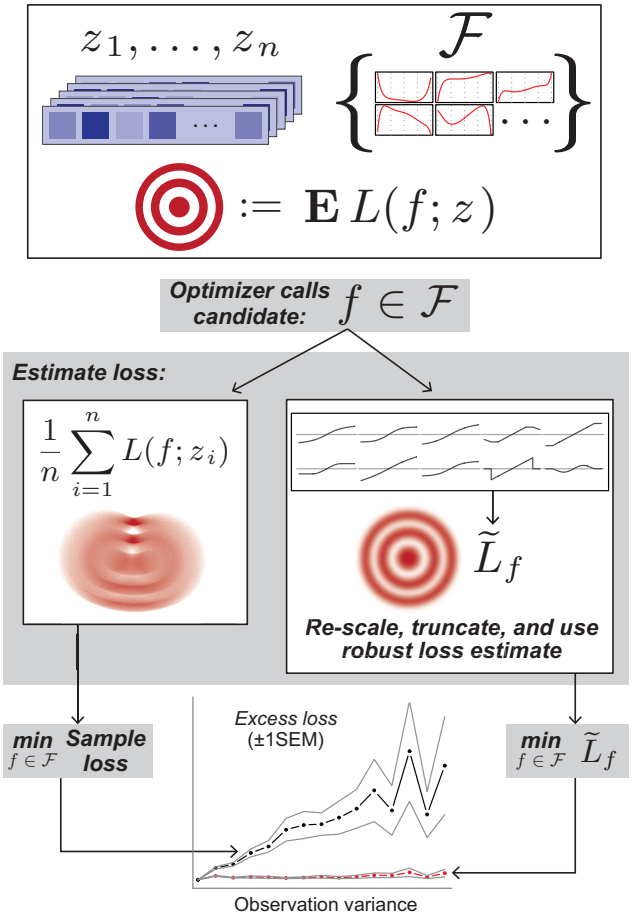

図 1: Schematic of basic robust target algorithm. The test data is from a series of experiments using the linear model with standard Normal inputs and log-Normal noise, and comparing OLS with a robust alternative.

To raise more precise questions and make more substan-

連絡先: M.J. Holland, NAIST 情報科学研究科
数理情報学研究室, 奈良県生駒市高山町 8916-5,
matthew-h@is.naist.jp

tial statements, let us formalize a concrete common machine learning task. Assume independent random observations $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ with common distribution P_z, of the form $\boldsymbol{z} = (\boldsymbol{x}, y) \in \mathbb{R}^{d+1}$. Denote the sample compactly by $\boldsymbol{z}_{(n)}$. In the regression task, we seek a correspondence $\boldsymbol{x} \mapsto \widehat{f}(\boldsymbol{x})$ with low prediction error off-sample, such that $\widehat{f}(\boldsymbol{x}) \approx y$ is a sharp approximation, with high confidence. A *learning algorithm* is any procedure which returns $\widehat{f}$ given input $\boldsymbol{z}_{(n)}$. To talk about performance, one may introduce a loss $L(f; \boldsymbol{z}) \geq 0$ to quantify predictive accuracy, say $L(f; \boldsymbol{z}) := (f(\boldsymbol{x}) - y)^2$. A natural benchmark is $L^* = \inf \mathbf{E}\, L(f; \boldsymbol{z})$, with the infimum is taken over f in some algorithm-dependent *model*, or function class. There are many routes of analysis, but put roughly, "good" performance means that $\mathbf{E}\, L(\widehat{f}; \boldsymbol{z}) - L^* \approx 0$ is a very sharp approximation for n not too large. When comparing rival algorithms $\widehat{f}$ and $\widehat{f}'$, the key questions to ask are as follows:

1. For fixed P_z, which routine should perform best?
2. How sensitive is performance to assumptions on P_z?

In the remainder of this document we introduce outlining some of the ideas underlying our work, making reference to classical and modern results of import.

2. Robust procedures in the machine learning era

Historically, the notion of creating a "robust" procedure was, thanks to Hampel's theorem, roughly a matter of designing $\widehat{f}$ such that for any small $\varepsilon > 0$ one has a constant δ where

$$\|P_0 - P\| < \delta \implies \|\widehat{f}(P_0) - \widehat{f}(P)\| < \varepsilon,$$

and this can be done for each P_0 in some model $\mathcal{P}$ (a class of distributions). This is just a continuity property for the map of data distributions to the procedure's final output, here a function $\widehat{f}$. This is a very useful notion when we have some good, known, parametric model $\mathcal{P} = \mathcal{P}(\Theta)$ and want to guard against small "corruptions," or mixtures where $\boldsymbol{z} \sim P_z$ and $P_z = (1 - \eta)P + \eta Q$ for small η, $P \in \mathcal{P}(\Theta)$, and some disruptive distribution Q not too far from the known class (cf. [Huber and Ronchetti, 2009]). Note the key assumption:

A good a priori parametric model $\mathcal{P}(\Theta)$ is available.

This setup is excellent in the context of 1970s statistics, since estimation tasks were typically small in dimension, sample sizes were assumed large (relative to dimension), and carried out by professional statisticians. At present we face different issues, since many users without formal training are making use of machine learning algorithms for high-dimensional statistical estimation tasks, which are expected to simply "work," with little in the way of assumption checking.

In this more modern setting, it might be more productive to consider robustness in a different sense. For example, we might say procedure A is more robust than procedure B if at the same confidence level, the same performance can be guaranteed under weaker assumptions on P_z. In the regression task, this might amount to showing that for small $\delta \in (0, 1)$ we have

$$\mathbf{E}\, L(\widehat{f}; \boldsymbol{z}) - L^* < \varepsilon(n, \delta)$$

with high probability, with the conditions on P_z being the weaker the better, all else equal.

Given this context, we are interested in the theory and practice of taking a given algorithm $\widehat{f}$, and systematically "robustifying" it by modifying key sub-routines in a methodical way such that near-optimal performance is guaranteed for a provably larger class of data distributions. One natural idea is to take algorithms utilizing sample loss optimizers, namely those $\widehat{f} = \arg\min_f \sum_{i=1}^n L(f; \boldsymbol{z}_i)/n$, and re-code them to simply

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \widetilde{L}_f \approx \mathbf{E}\, L(f; \boldsymbol{z}), \quad P_z \in \mathcal{P}$$

where $\mathcal{P}$ is some large, non-parametric class where the approximation quality satisfies some pre-specified threshold (e.g., $\varepsilon(n, \delta) = O(1/\sqrt{(n)})$ is about as good as it gets orderwise). The basic idea is represented in Fig. 1. The task then is how to construct the estimate $\widetilde{L}_f$ given $L(\cdot; \boldsymbol{z})$, $\mathcal{F}$, and $\boldsymbol{z}_{(n)}$, in such a way that the optimization can be implemented and formal guarantees may be made. Defining an M-estimator (with scaling parameter s) of the form

$$\widetilde{L}_f \in \left\{ \theta : \sum_{i=1}^n \psi\left(\frac{(L(f; \boldsymbol{z}_i) - \theta)}{s} \right) = 0 \right\}$$

is a common starting point, where ψ is usually a smooth function on $\mathbb{R}$, and may be monotonic or re-descending in the limit. A fascinating result [Catoni, 2012] says that if we have access to $\sigma_f^2 := \mathrm{var} L(f; \boldsymbol{z})$, then there exists a (computable) class of increasing functions ψ such that

$$|\widetilde{L}_f - \mathbf{E}\, L(f; \boldsymbol{z})| \leq O(\sigma \sqrt{\log(\delta^{-1}/n)})$$

with probability no less than $1 - 2\delta$. Of course assuming σ to be known is artificial and restrictive, but doing so renders the scale estimation task trivial. As a result of this, very recent work from [Brownlees et al., 2015] has shown connections with this sort of re-coding and learning performance. For example, they show that with known variance bound $\sup_{f \in \mathcal{F}} \mathrm{var} L(f; \boldsymbol{z})$, and linear model $\mathcal{F} = \{f : f(\boldsymbol{x}) = \boldsymbol{x}^T \beta, \beta \in \mathbb{R}^d\}$, an argument using key multiplier inequalities [van der Vaart and Wellner, 1996] and chaining techniques [Pollard, 1990] allows one to verify that the minimizer of $\widetilde{L}_f$ for the ℓ_2 loss L is such that $\mathbf{E}\, L(\widehat{f}; \boldsymbol{z}) - L^*$ scales with $O(\sqrt{\log(\delta^{-1})/n})$ plus a term scaling with a metric entropy model complexity parameter [Dudley, 1967]. The key fact is that as long as P_z is such that the uniform bound is finite, the results hold, making for a significant weakening of assumptions on the noise distribution.

One then naturally asks, outside this idealized setting, how should we estimate s? This essentially as "how much

information to truncate" in a routine that adapts to P_z. Clearly some tradeoffs must be made; today we look at what sort of a price must be paid, discuss some basic results and routines for tackling the more general problem.

参考文献

[Audibert and Catoni, 2011] Audibert, J.-Y. and Catoni, O. (2011). Robust linear least squares regression. *Annals of Statistics*, 39(5):2766–2794.

[Brownlees et al., 2015] Brownlees, C., Joly, E., and Lugosi, G. (2015). Empirical risk minimization for heavy-tailed losses. *Annals of Statistics*, 43(6):2507–2536.

[Candès et al., 2011] Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM*, 58(3).

[Catoni, 2012] Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 48, pages 1148–1185. Institut Henri Poincaré.

[Dudley, 1967] Dudley, R. M. (1967). The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1(3):290–330.

[Hampel et al., 1986] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons.

[Huber and Ronchetti, 2009] Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. John Wiley & Sons, 2nd edition.

[McWilliams et al., 2014] McWilliams, B., Krummenacher, G., Lucic, M., and Buhmann, J. M. (2014). Fast and robust least squares estimation in corrupted linear models. In *Advances in Neural Information Processing Systems 27*, pages 415–423.

[Pollard, 1990] Pollard, D. (1990). Empirical processes: theory and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, volume 2.

[Soltanolkotabi et al., 2014] Soltanolkotabi, M., Elhamifar, E., and Candès, E. J. (2014). Robust subspace clustering. *Annals of Statistics*, 42(2):669–699.

[Staudte and Sheather, 1990] Staudte, R. G. and Sheather, S. J. (1990). *Robust Estimation and Testing*. John Wiley & Sons.

[van der Vaart and Wellner, 1996] van der Vaart, A. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.