Reading Report for Theoretical Topics of Machine Learning

Xiang Li
1801110058
School of Mathematical Science

June 21, 2020

1 Introduction

The book I choose is [7] that tries to specify the link between statistics and optimization and conduct statistic inference through the lens of convex optimization. The most concerned topics includes sparse recovery and statistical hypothesis test.

The former is quite popular earlier in the decade due to its wide application in compressed sensing, signal de-noising, statistical model selection and so on. The main idea of sparse recovery lies in that a suitably high dimensional sparse signal can be inferred from very few linear observations. The great development of the sparse recovery theory drives fruitful applications in the general field of information processing, including communications channel estimation, dictionary leaning, data compression, optical imaging, machine learning etc [8]. It also inspires researchers to develop extensions to the recovery of low-rank matrices and higher order tensors from incomplete linear information [6].

The later lies in the core of statistic decision framework: how to make decisions with minimum risk [5]. The process of finding the optimal test via minimizing risk inevitably involves optimization. By statistical hypothesis tests, we can interpret data by assuming a specific structure our outcome and use statistical methods to confirm or reject the assumption. Whenever we want to make claims about the distribution of data or whether one set of results are different from another set of results in applied machine learning, we must rely on statistical hypothesis tests. Unlike the simple cases introduced in [9], [7] attempt to handle multiple hypotheses testing and sequential hypothesis testing. What's more, the testing machinery can help the estimation problem, which extends the boundary of statistical hypothesis testing.

2 Sparse Recovery

The problem and ℓ_1 minimization. One of the basic problems in Signal Processing is the problem of recovering a *signal* $x \in \mathbb{R}^n$ from its *noisy observations*

$$y = Ax + \eta \tag{1}$$

of the affine image of the signal under a given *sensing mapping* $x \mapsto Ax : \mathbb{R}^n \to \mathbb{R}^m$, η is the *observation error* and A is called *sensing matrix.* It is almost impossible to recover a dense x when $m < n$. Therefore, make the case of $m < n$ meaningful is to add to the observations (1) some a priori information on the signal: x is sparse in the sense that it has only s non-zero entries. The approach of Compressed Sensing [4] differs from previous works in that no assumption on the location of non-zero entries is made. By Compressed Sensing, we can only compute $m < n$ other linear forms of the signal and then use signal's sparsity in a known to us basis (i.e., how we construct A) in order to recover the signal reasonably well from these m observations. Mathematically speaking, the problem is formulated as:

$$\min_z\{\|z\|_0 := \sum_{i=1}^n 1_{\{z_i \neq 0\}} : y = Az\}. \tag{2}$$

Such a combinatorial optimization problem is NP-hard, and we almost could do nothing but process it by "brute force". Instead, researchers propose to replace the $\|\cdot\|_0$ with it convex surrogate $\|\cdot\|_1$; with this approximation, (2) converts into the l_1 minimization problem:

$$\min_z\{\|z\|_1 := \sum_{i=1}^n |z_i| : y = Az\}. \tag{3}$$

which are efficiently solvable by convex solvers, like celebrated FISTA [1].

Perfect ℓ_1 minimization. The remains question is how the ℓ_1 relaxation correlates with the original problem, which would make several constraints on A. The minimal requirement on sensing matrix A is to guarantee the correct recovery of **exactly** s-sparse signals in the **noiseless** case. We say A is *s-good*, if whenever in y in (2) is of the form $y = Ax$ with s-sparse x, x is the unique optimal solution to (2). A necessary and sufficient for A to be s-good is the so-called *Nullspace property*:

$$\exists \kappa \in (0, \frac{1}{2}),\ \|w\|_{1,s} := \max_{|I|=s} \|w_I\|_1 \leq \kappa \|w\|_1,\ \forall w \in \text{Ker } A. \tag{4}$$

Imperfect ℓ_1 minimization. However, in reality, the two assumptions typically are violated: (i) imperfect sparsity, i.e., we only have nearly s-sparse signal x; and (ii) noisy observations. Fortunately, we can quantify the Nullspace property to allow for instructive error analysis. In particular, the Nullspace property holds if and only if for a properly selected constant C one has

$$\exists \kappa \in (0, \frac{1}{2}),\ \|w\|_{1,s} \leq C\|Aw\|_2 + \kappa\|w\|_1,\ \forall w \tag{5}$$

For our purposes, it is convenient to present the condition (5) in a more systematic form:

Definition 2.1 ($Q_q(s,\kappa)$ conditions)**.** *Given a $m \times n$ matrix A, sparsity level $s < n$ and $\kappa \in (0, 1/2)$, we say that $m \times N$ matrix H and a norm $\|\cdot\|$ on $\mathbb{R}^N$ satisfy condition $Q_q(s,\kappa)$ with $1 \leq q \leq \infty$, if*

$$\|w\|_{q,s} \leq s^{\frac{1}{q}}\|H^T Aw\| + \kappa s^{\frac{1}{q}-1}\|w\|_1,\ \forall w \in \mathbb{R}^n. \tag{6}$$

Here H is named as the *contrast matrix.* We define the regular ℓ_1 recovery of x via observation y as the solution of the following convex optimization problem:

$$\hat{x}_{\text{reg}}(y) \in \arg\min_u\{\|u\|_1 : \|H^T(Au - y)\| \leq \rho, \} \tag{7}$$

and assume that $(H, \|\cdot\|)$ satisfies $Q_q(s,\kappa)$ condition associated with A. Then for all $x \in \mathbb{R}^n$ and $\eta \in \Xi_\rho := \{\eta : \|H^T\eta\| \leq \rho\}$, it follows that

$$\|\widehat{x}_{\text{reg}}(Ax+\eta) - x\|_p \leq \frac{4(2s)^{\frac{1}{p}}}{1-2\kappa}\left[\rho + \frac{\|x-x^s\|_1}{2s}\right], 1 \leq p \leq q.$$

In practice, we often design H such that $(H, \|\cdots\|)$ obeys Q-condition with small enough ρ. If η is some random vector, ρ can be replaced by certain quantile bound.

Signal recovery from Gaussian observations. When the noise conforms to Gaussian distribution, we have more general results. Given positive definite $m \times m$ matrix Γ, $m \times n$ matrix A, $\nu \times n$ matrix B, and indirect noisy observation $\omega = Ax + \xi$ with $\xi \sim \mathcal{N}(0, \Gamma)$ of unknown "signal" x known to belong to a given symmetric convex compact subset $\mathcal{X} \in \mathbb{R}^n$, we want to recover the image $Bx \in \mathbb{R}^\nu$. We focus first on the case where the quality of a candidate recovery $\omega \mapsto \widehat{x}(\omega)$ is quantified by its worst-case, over $x \in \mathcal{X}$, expected $\|\cdot\|_2^2$ -error, that is, by the risk

$$\text{Risk}[\widehat{x}(\cdot)|\mathcal{X}] = \sup_{x\in\mathcal{X}} \sqrt{\mathbb{E}_{\xi\sim\mathcal{N}(0,\Gamma)}\left\{\|\widehat{x}(Ax+\xi) - Bx\|_2^2\right\}}.$$

Actually, we can establish a rather general result on near-optimality of properly built linear estimates (that is, $\widehat{x}(\omega) = H^T\omega$) as compared to all possible estimates by imposing some restrictions on $\mathcal{X}$. For example, $\mathcal{X}$ is a high-dimensional $\|\cdot\|_1$-ball or the intersection of multiple ellipsoids/elliptic cylinders. In the latter case, the risk of properly selected linear estimate $\widehat{x}_{H_*}$ with both H_* and the risk efficiently computable, satisfies the bound

$$\text{Risk}\left[\widehat{x}_{H_*}|\mathcal{X}\right] \leq O(1)\sqrt{\ln(K+1)}\,\text{Risk}_{\text{opt}}[\mathcal{X}] \text{ with } \text{Risk}_{\text{opt}}[\mathcal{X}] := \inf_{\widehat{x}(\cdot)} \text{Risk}[\widehat{x}|\mathcal{X}].$$

3 Statistical Hypothesis Test

Problem setup. Hypothesis Testing is one of the most basic problems of Statistics. Informally, when given an (or several) observation $\omega \in \Omega$, which is a realization of random variable with unknown (at least partially) probability distribution, and several hypotheses on the actual distribution of the observed variable $H_1, H_2, \cdots, H_L$ with H_l stating that the true probability distribution $P \in \mathcal{P}_l$, we want find a measurable function $\mathcal{T} : \Omega \mapsto \{1, 2, \cdots, L\}$ that helps specify which one of the hypotheses is true. If multiple observations are available (say K samples), we can replace Ω with the product space $\underbrace{\Omega \times \ldots \times \Omega}_{K}$. The criterion to compare different $\mathcal{T}$ is various kinds of risk: (i) *partial risk* are the worst-case probability for $\mathcal{T}$ to reject l-th hypothesis, defined as,

$$\text{Risk}_\ell\left(\mathcal{T}|H_1, \ldots, H_L\right) = \sup_{P\in\mathcal{P}_\ell} \text{Prob}_{\omega\sim P}\{\omega : \mathcal{T}(\omega) \neq \{\ell\}\}, \ell = 1, \ldots, L; \tag{8}$$

(ii) *risk* is the the maximum of all L partial risks:

$$\text{Risk}\left(\mathcal{T}|H_1, \ldots, H_L\right) = \max_{1\leq\ell\leq L} \text{Risk}_\ell\left(\mathcal{T}|H_1, \ldots, H_L\right). \tag{9}$$

Hypothesis testing via Euclidean separation. When we focus on the decision of distribution mean, the hypothesis problem can be reduced quite a lot: (i) $\omega = x + \xi$ where ξ is random with distribution from $\mathcal{P}$; (ii) the l-th hypothesis says $H_l : x \in X_l$ where X_l is some convex set. When we consider only two hypothesis H_1 and H_2 and assume that $X_1 \cap X_2 = \emptyset$, the optimal test is the hyperplane that separates the two non-intersect convex sets with maximum margins, whose existence is guaranteed by the *Hyperplane Separation Theorem* [3].

When multiple samples are available, a popular strategy to extend a test function derived from one observation is *majority voting*. Formally, given K-repeated observation $\omega^K = (\omega_1, \cdots, \omega_K)$ and the decision given by a test $\mathcal{T}$ at hand, the majority test $\mathcal{T}_K^{\text{maj}}$ is given by the following procedure: if at least $K/2$ rejects H_1, then we accept H_2 and vice versa. One can show that the majority test is **near optimal** in terms of the number of observations to reach a sufficient small risk.

Multiple hypothesis testing via pairwise tests. When considering multiple hypothesis testings, we can't obtain a low-risk simple test if some H_l's resemble each other. To solve the problem, we allow ambiguity in similar hypothesis and redefine the risk as $\mathcal{C}$-risk:

$$\text{Risk}_\ell^{\mathcal{C}}\left(\mathcal{T}|H_1, \ldots, H_L\right) = \sup_{P \in \mathcal{P}_\ell} \text{Prob}_{\omega \sim P}\left\{[\ell \notin \mathcal{T}(\omega)] \text{ or } \left[\exists \ell' \in \mathcal{T}(\omega) : \left(\ell, \ell'\right) \notin \mathcal{C}\right]\right\}. \tag{10}$$

Here $\mathcal{C}$ is defined as some set of pairs (ℓ, ℓ') with $1 \le \ell, \ell' \le L$ with $(\ell, \ell') \in \mathcal{C}$ meaning H_ℓ and $H_{\ell'}$ are close to each other.

Given a set of simple tests $\mathcal{T}_{\{\ell,\ell'\}}$ deciding on H_ℓ v.s. $H_{\ell'}$ via observation ω, we can assemble them into a test $\mathcal{T}$ that decides on $H_1, \cdots, H_L$ up to closeness $\mathcal{C}$. The construction follows: we build a $L \times L$ matrix $T(\omega)$ as (i) if $(\ell, \ell') \in \mathcal{C}$, $T_{\ell,\ell'}(\cdot) = T_{\ell',\ell}(\cdot) = 0$; (ii) if $(\ell, \ell') \notin \mathcal{C}$, we set

$$T_{\ell\ell'}(\omega) = \begin{cases} 1, & \mathcal{T}_{\{\ell,\ell'\}}(\omega) = \{\ell\} \\ -1, & \mathcal{T}_{\{\ell,\ell'\}}(\omega) = \{\ell'\} \end{cases}.$$

Since $\mathcal{C}$ is symmetric, we have $T_{\ell\ell'}(\omega) \equiv -T_{\ell'\ell}(\omega), 1 \le \ell, \ell' \le L$. And (iii) we accept exactly those of the hypotheses H_ℓ for which ℓ-th row in $T(\omega) = [T_{\ell\ell'}(\omega)]$ is nonnegative. We can find the optimal test by tuning the risk of all pairwise $\mathcal{T}_{\{\ell,\ell'\}}$ to minimize the risk of $\mathcal{T}$ via convex optimization. It is also straightforward to combine the strategy with the majority test, when multiple observations and multiples hypothesis test are taken into consideration.

Detector-based test. If no geometry structure is available, we can make use of *detector* to define a statistical test in a systematic manner. For a simple test, $H_1 : P \in \mathcal{P}_1$ v.s. $H_2 : P \in \mathcal{P}_2$, the detector $\phi(\omega) : \Omega \mapsto \mathbb{R}$ is a real-valued function with risk defined:

$$\text{Risk}_-\left[\phi|\mathcal{P}_1\right] = \sup_{P \in \mathcal{P}_1} \int_\Omega \exp\{-\phi(\omega)\} P(d\omega), \ \text{Risk}_+\left[\phi|\mathcal{P}_2\right] = \sup_{P \in \mathcal{P}_2} \int_\Omega \exp\{\phi(\omega)\} P(d\omega)$$
$$\text{Risk}\left[\phi|\mathcal{P}_1, \mathcal{P}_2\right] = \max\left[\text{Risk}_-\left[\phi|\mathcal{P}_1\right], \text{Risk}_+\left[\phi|\mathcal{P}_2\right]\right]$$

Given a detector ϕ, we can associate with it simple test $\mathcal{T}_\phi(\omega) = 1_{\phi(\omega) \ge 0}$ via observation $\omega \sim P$, on the hypotheses $H_1 : P \in \mathcal{P}_1$ v.s. $H_2 : P \in \mathcal{P}_2$. The risk of $\mathcal{T}_\phi(\omega)$ satisfies

$$\text{Risk}_1\left(\mathcal{T}_\phi|H_1, H_2\right) \le \text{Risk}_-\left[\phi|\mathcal{P}_1\right], \ \text{Risk}_2\left(\mathcal{T}_\phi|H_1, H_2\right) \le \text{Risk}_+\left[\phi|\mathcal{P}_2\right]$$

Detector can be easily extended to cases where multiple observations are available: $\phi^{(K)}\left(\omega^K\right) = \sum_{k=1}^{K}\phi_k\left(\omega_k\right) : \Omega^K \to \mathbb{R}$.

To find the optimal $\mathcal{T}_\phi(\omega)$ that has a minimum risk, we we arrive at the following design problem:

$$\mathrm{Opt} = \min_{\phi:\Omega\to\mathbb{R}} \max\left[\underbrace{\sup_{P\in\mathcal{P}_1}\int_\Omega e^{-\phi(\omega)}P(d\omega)}_{F[\phi]}, \underbrace{\sup_{P\in\mathcal{P}_2}\int_\Omega e^{\phi(\omega)}P(d\omega)}_{G[\phi]}\right] \tag{11}$$

If we parametrize $\mathcal{P}_\ell$ with $\mathcal{P}_\ell = \{p_\mu : \mu \in M_\ell\}$ where $p_\mu(\cdot)$ belongs to a given "parametric" family of probability densities. Then (11) can be rewritten equivalently as a minmax optimization problem:

$$\ln(\mathrm{Opt}) = \min_{\phi:\Omega\to\mathbb{R}} \max_{\mu\in M_1,\nu\in M_2} \underbrace{\frac{1}{2}\left[\ln\left(\int_\Omega e^{-\phi(\omega)}p_\mu(\omega)\Pi(d\omega)\right) + \ln\left(\int_\Omega e^{\phi(\omega)}p_\nu(\omega)\Pi(d\omega)\right)\right]}_{\Phi(\phi;\mu,\nu)}. \tag{12}$$

Under some regularity, the optimal ϕ^*, μ^* and ν^* exist, serving as the saddle point of problem (12) and satisfying $\phi^*(\omega) = \frac{1}{2}\ln\left(p_{\mu^*}(\omega)/p_{\nu^*}(\omega)\right)$.

Estimating functions via hypothesis testing. The hypothesis testing techniques could help estimate properly structured scalar functionals in a simple observation scheme, which is a tuple $\mathcal{O} = ((\Omega,\Pi), \{p_\mu(\cdot) : \mu \in \mathcal{M}\}, \mathcal{F})$. Given random observation $\omega^K = (\omega_1, \ldots, \omega_K)$ with $\omega \overset{\text{i.i.d.}}{\sim} p_{A_j(x)}$ with $j \leq I$ and $x \in X_j$, where affine mappings $A_j(\cdot) : \mathbb{R}^n \to \mathbb{R}^M$ such that $A_j(x) \in \mathcal{M}$ whenever $x \in X_j, 1 \leq j \leq I$ and X_j is a convex set, we want to recover a linear function $g^T x$ on $\mathbb{R}^n$. It should be stressed that we do not know neither j nor x underlying our observation. Given reliability tolerance $\epsilon \in (0,1)$, we quantify the performance of a candidate $\widehat{g}(\cdot) : \Omega \to \mathbb{R}$ by the concept of (ρ,ϵ)-*reliable*, that says,

$$\forall\left(j \leq I, x \in X_j\right) : \mathrm{Prob}_{\omega\sim p_{A_j(x)}}\left\{\left|\widehat{g}(\omega) - g^T x\right| > \rho\right\} \leq \epsilon.$$

We define ϵ-risk of the estimate as

$$\mathrm{Risk}_\epsilon[\widehat{g}] = \inf\{\rho : \widehat{g} \text{ is } (\rho,\epsilon)\text{-reliable}\}. \tag{13}$$

Based on detector-based test, it can be proved that in this situation the estimate

$$\widehat{g}\left(\omega^K\right) = \sum_k \phi\left(\omega_k\right) + \kappa \tag{14}$$

with properly selected $\phi \in \mathcal{F}$ and $\kappa \in \mathbb{R}$ is near-optimal in the sense that when $K = \Omega(\bar{K})$, we have $\mathrm{Risk}_\epsilon[\widehat{g}] \leq \mathrm{Risk}_\epsilon^*(K) := \inf_{\widehat{g}(\cdot)} \mathrm{Risk}_\epsilon[\widehat{g}]$.

4 Discussion

Optimization facilitates error analysis. In this book, I learned how to mingle optimization with statistical inference. In sparsity recovery, the first (or zero-th) order stationary condition derived from the convex problem (5) is used to connect the solution and the true parameter, which provides ways to analyze the properties of the solution. For example, in non-parametric regression [11], the difference between the estimated function and the ground truth function can be bounded by a empirical process via using the zero-th stationary condition of optimization. By this way, the error analysis is reduced to the problem of uniform bounding a empirical process, for which we have abundant tools to offer, like entropy arguments and non-parametric statistics (see Chapter 19 in [10]). Besides, the asymptotic error analysis of LASSO also makes use of the idea [2].

Optimization improves decision performance. This book provides a general approach to hypothesis testing. The main "building block" of the proposed construction is a detector-based simple test for a pair of hypotheses in the situation where each particular hypothesis states that the vector of parameters identifying the distribution of observations belongs to a convex compact set associated with the hypothesis. When it comes to multiple hypothesis testing, we can build a test based on many pairwise simple tests, and then tune the Type I error and Type II error of each simple test in order to minimize the risk of final composite test. Such generic approach involves optimization naturally and smartly.

References

[1] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[2] Peter J Bickel, Ya'acov Ritov, Alexandre B Tsybakov, et al. Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics*, 37(4):1705–1732, 2009.

[3] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization.* Cambridge university press, 2004.

[4] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[5] Thomas S Ferguson. *Mathematical statistics: A decision theoretic approach*, volume 1. Academic press, 2014.

[6] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations.* CRC press, 2015.

[7] Anatoli Juditsky and Arkadi Nemirovski. *Statistical Inference via Convex Optimization*, volume 69. Princeton University Press, 2020.

[8] Elaine Crespo Marques, Nilson Maciel, Lirida Naviner, Hao Cai, and Jun Yang. A review of sparse recovery algorithms. *IEEE Access*, 7:1300–1322, 2018.

[9] Jun Shao. *Mathematical statistics.* Springer Science & Business Media, 2006.

[10] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[11] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.