Statistical Inference with Local Optima

Yen-Chi Chen

Department of Statistics
University of Washington

March 16, 2019

Estimator from optimization

- Many estimators can be written in the form of optimizing an objective function.
- For one famous example, the MLE (maximum likelihood estimator) is defined to be

$$\widehat{\theta}_{MLE} = \text{argmax}_{\theta \in \Theta} L_n(\theta),$$

where Θ is the parameter space and

$$L_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} L(\theta | X_i)$$

is the log-likelihood function and $X_1, \cdots, X_n$ are IID from an unknown distribution function P_0.
- The objective function is the log-likelihood function.

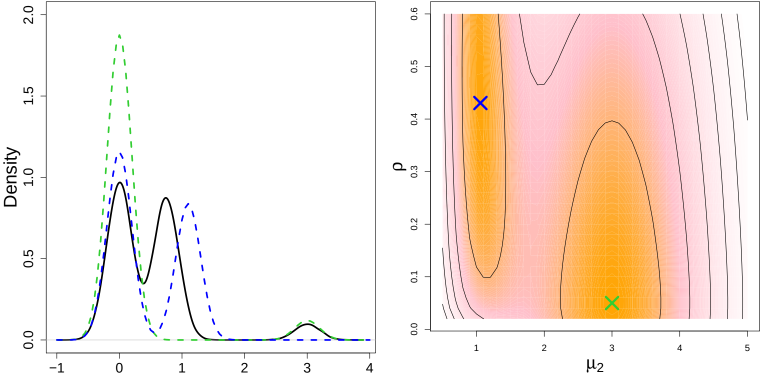
M-Estimator and its theory

- When the estimator is constructed by maximizing an objective function, it is often called an M-estimator.
- There are many well-known theory about the M-estimator such as consistency, convergence rate, and asymptotic normality.
- See, e.g., van der Vaart's *Asymptotic Statistics*.

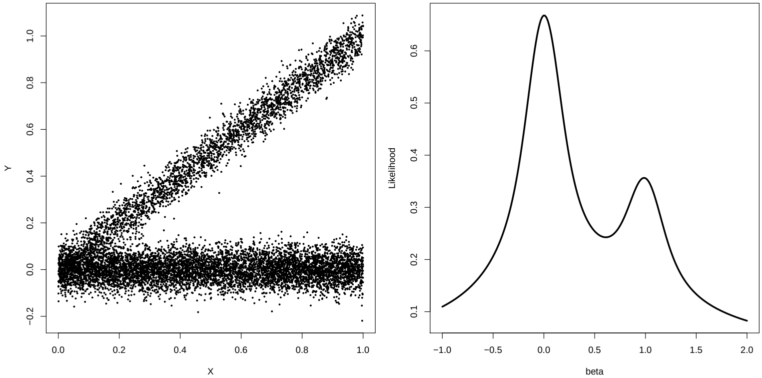
Challenge of the M-estimator and MLE

- M-estimator and MLE are nice and beautiful but they may not be tractable in practice.
- In many cases, the MLE does not have a closed-form so we have to use numerical approach to compute it.
- What's worse, in certain cases, the objective function (log-likelihood function) is not convex and may have multiple local modes.
- There is no simple way to find the MLE.
- A common case is the mixture model (Titteringtonet al., 1985; Redner and Walker, 1984).

An example of non-convex log-likelihood function



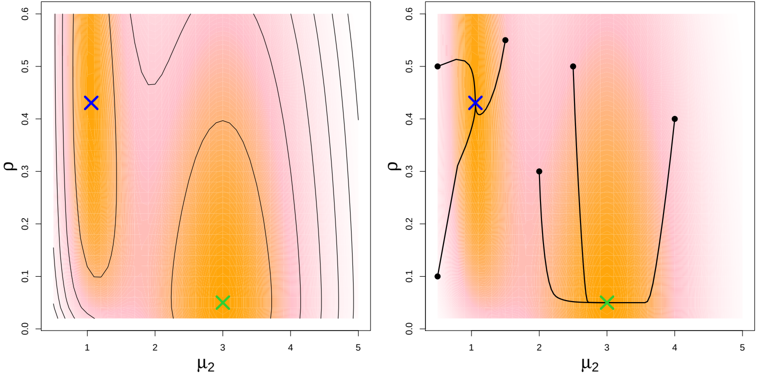
An example of non-convex log-likelihood function



Non-convex log-likelihood function

- When the log-likelihood function is non-convex, here is what people do in practice (see, e.g., McLachlan and Peel, 2004; Jin et al., 2016).
- We randomly choose an initial starting point of the parameter, denoted as θ_0.
- We apply EM algorithm or a gradient ascent algorithm with the initial point being θ_0 until the algorithm converges. We record the log-likelihood value at the convergent point.
- Repeat the above two steps many times, pick the convergent point with the highest log-likelihood value as the 'MLE'.
- Report the 'MLE' and use asymptotic theory of the MLE to construct a confidence interval.

Non-convex log-likelihood function: illustration



Optimizing a non-convex log-likelihood function

Formally, the above procedure can be written as follows.

1. Choose θ_0 from a distribution Π defined over Θ.
2. Define the gradient flow $\widehat{\gamma}_\theta : \mathbb{R} \mapsto \Theta$ such that
$$\widehat{\gamma}_\theta(0) = \theta, \quad \widehat{\gamma}'_\theta(t) = \nabla L_n(\widehat{\gamma}_\theta(t)).$$
Let the destination of the gradient flow starting at θ_0 be
$$\widehat{\gamma}_{\theta_0}(\infty) = \lim_{t\to\infty} \widehat{\gamma}_{\theta_0}(t).$$
This is the convergent point we have in the gradient ascent algorithm.
3. Repeat the above procedure M times, leading to M destinations
$$\widehat{\gamma}_{\theta_0^{(1)}}(\infty), \cdots, \widehat{\gamma}_{\theta_0^{(M)}}(\infty)$$
4. Define the estimator
$$\widehat{\theta}_{n,M} = \widehat{\gamma}_{\theta_0^{(J^*)}}(\infty)$$
$$J^* = \text{argmax}_{j=1,\cdots,M} L_n(\widehat{\gamma}_{\theta_0^{(j)}}(\infty)).$$

Questions we want to address

- The estimator $\widehat{\theta}_{n,M}$ may not be the MLE $\widehat{\theta}_{MLE}$.
- Thus, the inference may not be correct if we are pretending the estimator is the MLE.
- Our goal is to understand how bad the estimator $\widehat{\theta}_{n,M}$ can be when M is fixed and n is allowed to increase to infinity.

The population log-likelihood function - 1

- The log-likelihood function $L_n(\theta)$ converges to the population log-likelihood function

$$L(\theta) = \mathbb{E}(L(\theta|X_1))$$

  due to the law of large number.
- Our gradient ascent algorithm with $L_n(\theta)$ can be viewed as a sample version of the population gradient ascent flow $\gamma_\theta(t)$:

$$\gamma_\theta(0) = \theta, \quad \gamma'_\theta(t) = \nabla L(\gamma_\theta(t)).$$

- Let $\gamma_\theta(\infty)$ be the destination of the population gradient flow starting at θ.

The population log-likelihood function - 2

- Let $\theta_{MLE} = \text{argmax}_{\theta \in \Theta} L(\theta)$ be the population MLE.
- When the log-likelihood function is a Morse function, the population MLE is a mode of the log-likelihood function so it will be the destination of some gradient flows.
- We define the basin of attraction of θ_{MLE} as

$$\mathcal{A}_{MLE} = \{\theta : \gamma_\theta(\infty) = \theta_{MLE}\}.$$

- With the above notation, the probability

$$\Pi(\mathcal{A}_{MLE}) = P(Y \in \mathcal{A}_{MLE}),$$

where Y is a random variable from the distribution Π describes the chance of an initial parameter falls within the right basin of attraction.

The population log-likelihood function - 3

- Thus, if we draw M points from Π and apply the gradient ascent algorithm, the obtained maximum θ_M has a probability of

$$1 - (1 - \Pi(\mathcal{A}_{MLE}))^M$$

being the same as θ_{MLE}!

- Thus, the same argument applies to the sample MLE case. Let

$$\widehat{\mathcal{A}}_{MLE} = \{\theta : \widehat{\gamma}_\theta(\infty) = \widehat{\theta}_{MLE}\}$$

be the basin of attraction of the sample MLE with the sample gradient ascent flow.

- Then

$$P(\widehat{\theta}_{n,M} = \widehat{\theta}_{MLE} | X_1, \cdots, X_n) = 1 - (1 - \Pi(\widehat{\mathcal{A}}_{MLE}))^M$$

Chance to recover MLE

Theorem

Under regularity conditions,

$$\mathrm{Haus}(\widehat{\mathcal{A}}_{MLE}, \mathcal{A}_{MLE}) = O\left(\sup_{\theta} \|\nabla L_n(\theta) - \nabla L(\theta)\|_{\max}\right).$$

- Therefore, as $n \to \infty$ and M being fixed,

$$\begin{aligned} P(\widehat{\theta}_{n,M} = \widehat{\theta}_{MLE} | X_1, \cdots, X_n) &= 1 - (1 - \Pi(\widehat{\mathcal{A}}_{MLE}))^M \\ &= 1 - (1 - \Pi(\mathcal{A}_{MLE}))^M + O_P\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

Revising the confidence statement

- The above result shows that we need to modify our statement about the 'confidence' in constructing a confidence interval.
- Let $C_{n,M,\alpha}$ be a confidence interval from the asymptotic normality of $\widehat{\theta}_{MLE}$ but centered at $\widehat{\theta}_{n,M}$, then

$$P(\theta_{MLE} \in C_{n,M,\alpha}) = 1 - \alpha - (1 - \Pi(\mathcal{A}_{MLE}))^M + O\left(\frac{1}{\sqrt{n}}\right).$$

- $(1 - \Pi(\mathcal{A}_{MLE}))^M$ is the coverage deficiency due to the finite number of initializations.

Bootstrap confidence interval

- One may want to use the bootstrap to construct a confidence interval.
- But here comes the question: how should we initialize the starting point of gradient ascent algorithm in each bootstrap sample?
- If we want to obtain the same result as the previous confidence interval, we only need to initialize it once and use the same initial point $\widehat{\theta}_{n,M}$–the original estimator.
- Let $C^*_{n,M,\alpha}$ be the bootstrap confidence interval. Then

$$P(\theta_{MLE} \in C^*_{n,M,\alpha}) = 1 - \alpha - (1 - \Pi(\mathcal{A}_{MLE}))^M + O\left(\frac{1}{\sqrt{n}}\right).$$

Confidence intervals from inverting a test - 1

- Another common approach to constructing a confidence interval is via inverting a hypothesis testing procedure.
- There are three common approaches: the likelihood ratio test (LRT), the score test, and the Wald test.
- In the classical settings (when the log-likelihood function is convex), these tests are asymptotically equivalent.
- However, when the log-likelihood function has multiple local modes, they can be very different.

Confidence intervals from inverting a test - 2

- The LRT:

$$C_{LRT,\alpha} = \left\{ \theta : 2n(L_n(\widehat{\theta}_{n,M}) - L_n(\theta)) \geq \chi^2_{d,1-\alpha} \right\},$$

where $\chi^2_{d,1-\alpha}$ is the $1-\alpha$ quantile of a χ^2 distribution with d degrees of freedom.

- The score test:

$$C_{S,\alpha} = \left\{ \theta : n\nabla L_n(\theta)^T I_n^{-1}(\theta) \nabla L_n(\theta) \leq \chi^2_{d,1-\alpha} \right\},$$

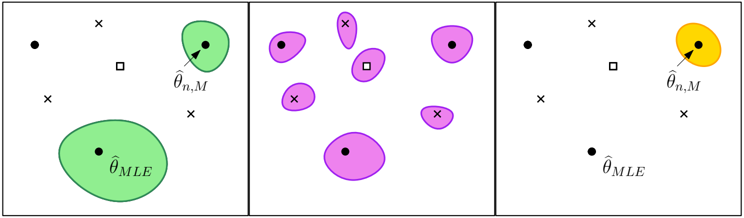
where $I_n(\theta)$ is the Fisher's information matrix.

- The Wald test:

$$C_{Wald,\alpha} = \left\{ \theta : (\widehat{\theta}_{n,M} - \theta)^T \widehat{\text{Cov}}(\widehat{\theta}_{n,M})(\widehat{\theta}_{n,M} - \theta) \leq \chi^2_{d,1-\alpha} \right\},$$

where $\widehat{\text{Cov}}(\widehat{\theta}_{n,M})$ is an estimate of the covariance matrix of $\widehat{\theta}_{n,M}$.

Confidence intervals from inverting a test - 3



- Left: the LRT; middle: the score test; right: the Wald test.
- The LRT and score tests always have the right coverage.
- The Wald test has the similar coverage as the usual confidence interval.

Applications of this frameworks

- Although we worked on the gradient ascent algorithm, a similar result can be obtained for the EM algorithm.
- Also, we can perform the same analysis for nonparametric bump hunting problem where the parameter of interest is the global mode of the density function.

Comparing initialization approaches

- Using the proposed framework, we can compare different approaches for generating the initial points.
- An initialization approach can be viewed as a distribution Π.
- Let Π_1 and Π_2 be two initialization methods.
- We can argue that the first method is better than the second method if

$$\Pi_1(\mathcal{A}_{MLE}) > \Pi_2(\mathcal{A}_{MLE}).$$

Reproducibility

- Because the estimator $\widehat{\theta}_{n,M}$ is computed with several random initializations, the reproducibility may be challenging.
- Another group with identical data and identical method may not leads to the same estimator due to the randomness of initializations.
- However, here is a simple way to test reproducibility if we keep track of the likelihood values of every destination in our initializations.
- The likelihood values of destinations of gradient flows will be IID points from a discrete distribution.
- If we have this information, another team can do a two-sample test to see if their observed likelihood values are from the same distribution as ours.

Discussion

- When our estimator is derived from optimizing a non-convex function, we need to be very cautious about our inference.
- The conventional confidence interval will not have the nominal coverage.
- Also, when inverting a test to a confidence interval, the LRT, score, and Wald tests may give you different answers.
- Many open questions left: generalizations to stochastic gradient ascent methods, bounding the coverage deficiency, controlling the algorithmic errors.

Thank you!

Paper reference: `https://arxiv.org/abs/1807.04431` (Statistical Inference with Local Optima).

References

1. Chen, Yen-Chi. “Statistical inference with local optima.” arXiv preprint arXiv:1807.04431 (2018).
2. Van der Vaart, Aad W. Asymptotic statistics. Vol. 3. Cambridge university press, 2000.
3. Titterington, D. Michael, Adrian FM Smith, and Udi E. Makov. Statistical analysis of finite mixture distributions. Wiley,, 1985.
4. Redner, Richard A., and Homer F. Walker. “Mixture densities, maximum likelihood and the EM algorithm.” SIAM review 26, no. 2 (1984): 195-239.
5. McLachlan, Geoffrey J., Sharon X. Lee, and Suren I. Rathnayake. “Finite mixture models.” Annual Review of Statistics and Its Application 0 (2000).
6. Jin, Chi, Yuchen Zhang, Sivaraman Balakrishnan, Martin J. Wainwright, and Michael I. Jordan. “Local maxima in the likelihood of gaussian mixture models: Structural results and algorithmic consequences.” In Advances in Neural Information Processing Systems, pp. 4116-4124. 2016.