Report：Statistical inference using SGD

Sitao Min

M-estimators

In statistics, **M-estimators** are a broad class of estimators, which are obtained as the minima of sums of functions of the data. In 1964, Peter J. Hube[1] proposed generalizing maximum likelihood estimation to the minimization of $\sum_{i=1}^{n} \rho(x_i,\ \theta)$, where ρ is a function with certain properties. The solutions $\hat{\theta} = arg\ min(\sum_{i=1}^{n} \rho(x_i,\ \theta))$ are called **M-estimators** ("M" for "maximum likelihood-type"). In frequentist inference (**Question1, what is frequrntist inference?**), we can write this as a minimizer of population risk (A **population** is any large collection of objects or individuals, such as Americans, students, or trees about which information is desired)

$$\theta^* = argmin \mathbb{E}_P[f(\theta; X)] = argmin \int_x f(\theta; x) dP(x) \quad (1)$$

As we don't know distribution P in real practice, we usually estimate θ^* by sloving empirical risk minimizaton problem, use estimate $\hat{\theta}$:

$$\hat{\theta} = argmin \frac{1}{n} \sum_{i=1}^{n} f(\theta; X_i) \quad (2)$$

Because the solution of M-estimation problems satisfies asymptotic normality, which means the distribution of $\sqrt{n}(\hat{\theta} - \theta^*)$ converges weakly to a normal distibution $\mathcal{N}(0, H^{*-1} G^* H^{*-1})$, in which $H^* = \mathbb{E}[\nabla^2 f(\theta^*; X)]$ and $G^* = \mathbb{E}[\nabla f(\theta^*; X) \cdot \nabla f(\theta^*; X)^T]$ (Theorem 5.21 in [1]) . We can use some statistical techniques such as confidence interval to obtain information of $\hat{\theta}$ (A **confidence interval** (**CI**) is a type of interval estimate(of a population parameter that is computed from the observed data).

So the key problem in M-estimator problem is to efficiently estimate $H^{*-1} G^* H^{*-1}$. And in this paper, **the author use SGD to efficiently estimatethis key item**.

Statistical inference using SGD

The main idea of this method is to proceed *t* consecutive SGD interation and use the average of this *t* consecutive SGD iterations as empirical minimum $\hat{\theta}$

Statistical Inference Using SGD Steps

Each SGD iteraton updating rule: $\theta_{t+1} = \theta_t - \eta g_s(\theta_t)$

1. Burn in first SGD iterates $\theta_{-b}, \theta_{-b+1}, \ldots, \theta_0$
2. For each “segment” of $t + d$ iterates, we use the first t iterates to compute $\bar{\theta}^{(i)} = \frac{1}{n} \sum_{j=1}^{t} \theta_j^{(i)}$ and discard the last d iterates, where i indicates the $i - th$ segment. And we proceed R segments.
3. The final empirical minimum $\hat{\theta} \approx \frac{1}{R} \sum_{i=1}^{R} \bar{\theta}_t^{(i)}$ [2].

4. *Statistical inference*: $\theta^{(i)} = \hat{\theta} + \sqrt{\frac{Ks \cdot t}{n}}(\bar{\theta}_t^{(i)} - \hat{\theta})$ and using variance of $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(R)}$ for statistical inference

---

Theoretical guarantees

Theorem1

Assume that $||\theta_1 - \hat{\theta}||_2^2 = O(\eta)$; then for sufficiently small step size $\eta > 0$, the average SGD sequence, $\bar{\theta}_t$ satisfies:

$$||t\mathbb{E}[(\bar{\theta}_t - \hat{\theta})(\bar{\theta}_t - \hat{\theta})^T] - H^{-1}GH^{-1}||_2 \leq \sqrt{\eta} + \sqrt{\frac{1}{t\eta} + t\eta^2} \quad (3)$$

(Question2:cannot understand what this conclusion means? Does it mean sequence of SGD aproximately equal to HGH?)

Reference

[1] A.W. van der Vaart. Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

[2] S´ ebastien Bubeck. Convex optimization: Algorithms and complexity. Found. Trends Mach. Learn., 8(3-4):231–357, November 2015.