

非凸な目的関数に対する確率的最適化

総合研究大学院大学 川島 孝行

統計数理研究所 藤澤 洋徳

1 はじめに

機械学習や統計、信号処理などにあらわれる多くの最適化問題は損失関数 $f(\theta)$ と正則化項 $h(\theta)$ の和を目的関数とした次のような形をしている。

$$\operatorname{argmin}_{\theta} f(\theta) + h(\theta). \quad (1)$$

例えば、Lasso (Tibshirani, 1996) は $f(\theta) = \frac{1}{n} \sum_{i=1}^n (y_i - x_i^T \theta)^2$, $h(\theta) = \lambda \|\theta\|_1$ とすることで、(1) の枠組みにはいる。上記の (1) の最適化を行う手法は多く提案されている (ADMM: Boyd et al 2010, Proximal Gradient Descent: Parikh and Boyd 2013)。特に、大規模なデータを対象とした場合には、計算効率の良さから Stochastic Gradient Descent (Bottou 2010) が広く使われている。このため、近年、Stochastic Gradient Descent の収束性に関する理論解析が精力的に研究されている (Duchi et al 2010, Xiao 2010, Defazio et al 2014, Ghadimi et al 2016, Ghadimi and Lan 2016, Allen-Zhu 2017)。しかしながら、現在までに損失関数 $f(\theta)$ と正則化項 $h(\theta)$ がどちらも非凸な場合の収束性に関する結果は得られておらず、特に正則化項は凸の場合にしか収束性が保証されていない。推定量の理論的な性質の観点から、凸性に基づく正則化項では推定量にバイアスが生じることが広く知られており (SCAD: Fan and Li 2001, MCP: Zhang 2010)、非凸な正則化項への拡張が望まれる。そこで、本発表では正則化項も非凸な場合についての Stochastic Gradient Descent の収束性に関する理論解析について発表を行う。

2 Stochastic Gradient Descent

上記の (1) に対する Stochastic Gradient Descent の更新式は以下のようにあらわされる。

$$\theta^{(t+1)} = \operatorname{argmin}_{\theta} \left\langle \frac{1}{m_t} \sum_{i=1}^{m_t} G_{X_{t,i}}(\theta^{(t)}), \theta \right\rangle + h(\theta) + \frac{1}{2\eta_t} \|\theta - \theta^{(t)}\|_2^2, \quad (2)$$

ただし、 $E_{X_{t,i}} [G_{X_{t,i}}(\theta^{(t)})] = \nabla f(\theta^{(t)})$, m_t と η_t はそれぞれ t 回目の反復におけるミニバッチサイズ、ステップサイズである。正則化項 $h(\theta)$ が凸であれば、更新式 (2) は凸関数の最小化になっており、凸性に基づいた距離射影や劣勾配を用いて解を特徴づけることが可能となる。しかしながら、一般に正則化項 $h(\theta)$ が非凸の場合には、その方法は使えない。そこで、正則化項 $h(\theta)$ が非凸のなかでも準凸と呼ばれるクラスの関数であると考え、最適性の必要条件などをもとに、凸である場合と同じように収束性を得ることが可能であることがわかった。また、準凸な正則化項 $h(\theta)$ は次のように、多くの重要な非凸な正則化項を含む。

$h(\theta)$: SCAD, MCP, Log-sum penalty, Capped- l_1 penalty, l_0 penalty.

発表当日は、より詳細な理論的性質およびステップサイズ、ミニバッチサイズの選択方法、数値実験についても発表を行う予定である。