確率的勾配降下法による期待識別誤差の線形収束性

二反田 篤史 †‡ 鈴木 大慈 †‡　† 東京大学 ‡ 理研 AIP

本研究では再生核ヒルベルト空間を仮説集合とする二値識別問題に対して，確率的勾配降下法が強低ノイズ条件下で線形収束する事を示す．二値識別問題では期待識別誤差を最小化する可測関数であるベイズ識別器を求める事が究極の目標だが，識別誤差は $\{0,1\}$-値関数であり直接的な最適化が困難である．そこで，ロジスティック損失のような一致性を備えた凸損失関数で近似し，代わりにこの損失関数が定める期待損失関数の最小化を試みる．このような近似は損失関数の一致性により期待識別誤差の最適値からのギャップが，期待損失関数値とその最適値とのギャップで上から抑えられる事によって正当化される．確率的勾配降下法はスケーラビリティ，実装容易性，収束性能の高さから期待損失最小化問題を含め大規模機械学習問題で有用な最適化手法であり，その収束性は活発に研究されている．これら一致性と期待損失関数の収束解析を通じて，期待識別誤差に対する収束率はサンプル数について劣線形である事が一般的に示される．しかしながら，損失関数の一致性を経由した収束率は最適とは限らず，更に高速な期待識別誤差の収束が適切な条件下で達成されるかは興味深い問題である．経験損失最小解に関してもこのような性質が成り立つが [1, 2] では，ラベルについて強低ノイズ条件を課す事で，期待識別誤差の収束率は劣線形より圧倒的な高速な線形収束となる事が示された．但し，強低ノイズ条件の他に必要とされる仮定及び問題設定により，その適用対象はやや限定的であった．一方，確率的勾配降下法についても二乗損失を用いた場合に強低ノイズ条件下で線形収束する事が [3] によって近年示されたが，二乗損失関数は外れ値に敏感であり識別問題において好まれないという問題がある．そこで本研究では，強低ノイズ条件下での確率的勾配降下法による期待識別誤差の線形収束性を，ロジスティック損失や指数損失を含む識別問題にとってより適切な関数クラスに対して示す．また，数値実験において実際に識別誤差が損失関数値に比べ高速に収束する事を紹介する．

主結果　以下，簡単のためガウシアンカーネルが定める再生核ヒルベルト空間 $\mathcal{H}_k$ 上のロジスティック回帰と，それに対する平均化確率的勾配降下法に限定した結果を紹介する．平均化確率的勾配降下法は $g_1 \in \mathcal{H}_k$ を初期点とした確率的勾配降下法が生成する反復点列 $\{g_t\}_{t=1}^T$ の重み付き平均 $\overline{g}_{T+1}$ を返す．具体的には γ をハイパーパラメータ，λ を正則化係数，$G_\lambda(g_t, z_t)$ を $\mathcal{H}_k$ での正則化項付き損失関数の確率的勾配とすると，以下の様に定義される．

$$g_{t+1} \leftarrow g_t - \frac{2}{\lambda(\gamma+t)} G_\lambda(g_t, z_t), \quad \overline{g}_{T+1} \leftarrow \sum_{t=1}^{T+1} \frac{2(\gamma+t-1)}{(2\gamma+T)(T+1)} g_t.$$

期待識別誤差とそのベイズ最適値を $\mathcal{R}(g)$ 及び $\mathcal{R}_*$ とする．特徴ベクトルと二値ラベル $\{-1,1\}$ の確率変数を X, Y，その確率分布を $\rho(X,Y)$ として，X についての周辺分布及び条件付き確率を $\rho(X)$, $\rho(Y|X)$ で表す．

定理. 期待損失最小解が $\mathcal{H}_k$ に含まれ，強低ノイズ条件 $|\rho(1|X) - \frac{1}{2}| > \delta$ $(\delta \in (0, 1/2))$ が $\rho(X)$ について至る所で成り立つとする．$\eta_1 = \frac{2}{\lambda(\gamma+1)}$ として，十分小さな正則化係数 $\lambda > 0$ に対し $\gamma > 0$ は $\eta_1 \leq \min\{1/L, 1/2\lambda\}$ を満たすように定め，更に $\|g_1\|_{\mathcal{H}_k} \leq \left(2\eta_1 + \frac{1}{\lambda}\right)/4$ とする．この時，十分大きな T に対し次が成り立つ．

$$\mathbb{E}[\mathcal{R}(g_{T+1}) - \mathcal{R}_*] \leq 2\exp\left(-\frac{\delta^2\lambda^2(2\gamma+T)}{36}\right).$$

参考文献

[1] Vladimir Koltchinskii and Olexandra Beznosova. Exponential convergence rates in classification. In *International Conference on Computational Learning Theory*, pages 295–307, 2005.

[2] Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633, 2007.

[3] Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Exponential convergence of testing error for stochastic gradient methods. *arXiv preprint arXiv:1712.04755*, 2017.