Generalization Error Analysis of Stochastic Gradient Descent on Classification Problems under Low Noise Condition

（低ノイズ条件下の識別問題における確率的勾配降下法の汎化誤差解析）

数理情報学専攻 48-186219　八嶋 晋吾

指導教員　鈴木 大慈 准教授

1　背景

本研究では，仮説空間を再生核ヒルベルト空間 (RKHS) とするカーネル法を用いて，二値識別問題を解く問題設定を考える．入力空間を $\mathcal{X}$，ラベル空間を $\mathcal{Y} = \{-1, 1\}$ とし，訓練サンプルはある分布 ρ から生成されるとする．このとき，二値識別問題は以下で定義される**期待識別誤差** $\mathcal{R}$ を最小化する識別器 $g : \mathcal{X} \to \mathbb{R}$ を求める問題として定式化できる:

$$\mathcal{R}(g) \stackrel{\mathrm{def}}{=} \mathbb{E}_{(X,Y)\sim\rho}\left[\mathbb{1}(\mathrm{sgn}(g(X)) \neq Y)\right].$$

しかし，0-1 損失は非凸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 ℓ でそれを近似し，代わりに ℓ が定める**期待損失関数** $\mathcal{L}$ を最小化することで期待識別誤差を小さくすることを試みる:

$$\mathcal{L}(g) \stackrel{\mathrm{def}}{=} \mathbb{E}_{(X,Y)\sim\rho}[\ell(g(X), Y)].$$

すると，勾配降下法といった一般的な凸最適化手法により効率的に識別機を学習することができる．

一方，実用上においてカーネル法はサンプルサイズに対して計算量がスケールしないという問題を持ち，大規模な問題では近似して計算量を減らす方法がよく用いられる．そのアプローチは最適化における近似と仮説空間における近似の二つに大きく分けられ，前者として勾配を一つのサンプルのみから計算する確率的勾配降下法 (SGD) [7]，後者として再生核ヒルベルト空間を確率的なサンプリングにより低次元に近似する Random Feature [6] が，計算効率性と実装の簡便さから広く用いられている．

これらの手法の理論的な性質については多くの既存研究があるが，その多くは最適化の目的関数，つまり期待損失関数の収束の観点からの性能解析であった．本研究では二値識別問題における真の目的である期待識別誤差の収束の観点から，SGD と Random Feature の有効性について議論する．

2　先行研究

凸な損失関数に対し，SGD が達成する最適レートはイテレーション数 T に対し $O(1/\sqrt{T})$ であることが知られている [1]．さらに，一般には識別誤差の収束レートは損失関数の収束レートよりも速くなることはない [2] が，一方で速い識別誤差の収束を実現するような条件についても古くから考えられてきた．

定義 1 (低ノイズ条件 [8])**.** サンプル分布 ρ に対し，ある α で

$$\mathbb{P}\left[|\rho(Y = 1|X) - 1/2| \leq \delta\right] \lesssim \delta^{\alpha}.$$

が成り立つとき，**弱低ノイズ条件**を満たすという．また，ある $\delta \in (0, 1/2)$ で

$$|\rho(Y = 1|x) - 1/2| > \delta \quad (\rho_{\mathcal{X}}\text{-}a.s.)$$

が成り立つとき，**強低ノイズ条件**を満たすという．

この条件はラベル確率が $1/2$ に近いようなサンプルがどれくらい含まれるかを表したもので，弱低ノイズ条件の α が大きいほど簡単な問題となり，$\alpha = \infty$ の場合が強低ノイズ条件に対応する．近年の研究で，強低ノイズ条件のもとで SGD が識別誤差の線形収束（指数的収束）を達成することが示されている [5, 4]．本研究ではそれらの研究を元に，以下の 2 つの貢献を行った．

1. 弱低ノイズ条件下でも，SGD により損失関数よりも速い識別誤差の収束が達成されることを示した．
2. 強低ノイズ条件下では，SGD に加えて Random Feature での近似が有効であることを理論的・実験的に示した．

3　扱うアルゴリズムの概略

3.1　確率的勾配降下法 (SGD) [7]

カーネル関数 k の RKHS $\mathcal{H}$ 上の期待損失最小化問題

$$\min_{g\in\mathcal{H}} \left\{ \mathcal{L}(g) + \frac{\lambda}{2}\|g\|_{\mathcal{H}}^2 \right\} \tag{1}$$

を SGD で最適化する手順は以下のようになる．学習率 η_t, 重み θ_t は適当に定めるパラメータである．

Algorithm 1 SGD on RKHS

1: initialize $g_1 \in \mathcal{H}$
2: **for** $t = 1, \dots, T$ **do**
3: sample $(x_t, y_t) \sim \rho$
4: $g_{t+1} \leftarrow g_t - \eta_t \left(\nabla \ell(g_t(x_t), y_t) k(x_t, \cdot) + \lambda g_t\right)$
5: **end for**
6: **return** $\overline{g}_{T+1} = \sum_{t=1}^{T+1} \theta_t g_t$

3.2 Random Feature [6]

カーネル関数 k が以下のように展開されるとする:

$$k(x, y) = \int_\Omega \varphi(x, \omega)\varphi(y, \omega) d\tau(\omega).$$

このような展開は特に平行移動不変なカーネル (e.g., ガウシアン) で成り立つ．この積分をモンテカルロ近似することにより M 次元空間 $\mathcal{H}_M$ に仮説空間を制限するのが Random Feature である:

$$k_M(x, y) = \frac{1}{M} \sum_{i=1}^{M} \varphi(x, \omega_i)\varphi(y, \omega_i), \quad \omega_i \overset{i.i.d.}{\sim} \tau.$$

4 弱低ノイズ条件下での SGD の速い収束

分布 ρ が弱低ノイズ条件を満たす場合，SGD による期待識別誤差 $\mathcal{R}$ の収束について以下の結果を得た．

定理 2. 弱低ノイズ条件がパラメータ α で成り立つとする．いくつかの仮定のもと，適当に定めた $\lambda, \eta_t, \theta_t$ に対し，ある定数 $C > 0$ が存在し次が成り立つ:

$$\mathbb{E}\left[\mathcal{R}(\overline{g}_{T+1}) - \mathcal{R}(g_*)\right] \leq C T^{-\frac{(\alpha+1)\kappa}{2+2\kappa}}.$$

ただし，κ は $\mathcal{H}$ の複雑さを表すパラメータ，g_* は最適な識別器である．

この結果より，ラベルノイズが小さい (α が大きい) サンプルの場合，識別誤差の収束は損失関数の最適レートよりも速くなることがわかる．

5 強低ノイズ条件下での SGD と Random Feature の有効性

以降簡単のため，具体的に k をガウシアンカーネル，ℓ をロジスティック損失とした場合の結果を紹介する．まず，最小化問題 (1) の $\mathcal{H}, \mathcal{H}_M$ 上での解 $g_\lambda, g_{M,\lambda}$ について次のバウンドを得た．

定理 3. feature のサンプリングに関して確率 $1 - 2\delta'$ 以上で以下が成り立つ:

$$\|g_\lambda - g_{M,\lambda}\|_{L^\infty} \lesssim \left(1 + \frac{1}{\delta'}\right)^{\frac{1}{4}} \left(\frac{1}{\lambda}\right)^{\frac{3}{4}} \left(\frac{1}{M} \log \frac{1}{\delta'}\right)^{\frac{1}{8}}.$$

この結果は既存研究 [3] の拡張になっている．この結果のもと，次の強低ノイズ条件下での期待識別誤差の線形収束性が示される．

定理 4. 適当に定めた $\lambda, \eta_t, \theta_t$ のもと，

$$M \gtrsim \left(\frac{\left(1 + \frac{1}{\delta'}\right) \|g_*\|_{\mathcal{H}}^4}{\lambda^3 \log^4 \frac{1+2\delta}{1-2\delta}}\right)^2 \log \frac{1}{\delta'}.$$

とすると，パラメータ δ の強低ノイズ条件といくつかの仮定のもと，確率 $1 - 2\delta'$ 以上で次が成り立つ：パラメータに依存する T が存在し，$t > T$ で

$$\mathbb{E}\left[\mathcal{R}(\overline{g}_{t+1}) - \mathcal{R}(g_*)\right] \leq 2 \exp\left(-\frac{\lambda^2(2\gamma + t)}{2^{12} \cdot 9} \log^2 \frac{1 + 2\delta}{1 - 2\delta}\right).$$

この結果より，識別誤差の収束に必要な feature 数 M はサンプルサイズ T によらないことがわかる．これは損失関数の収束を考えた場合は見られない現象であり，Random Feature が強低ノイズ条件下では実際に計算を効率化できていることを示唆している．さらに数値実験によりこの事実を確かめることができた．

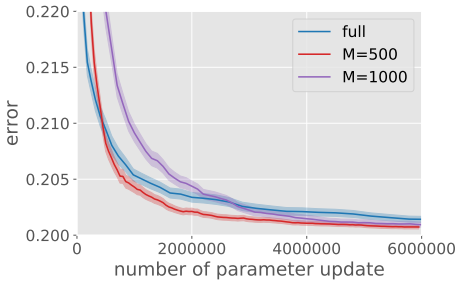

図 1. 近似なしと Random Feature の計算量の比較

参考文献

[1] Alekh Agarwal, Martin J Wainwright, Peter L Bartlett, and Pradeep K Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems*, pp. 1–9, 2009.

[2] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, Vol. 101, No. 473, pp. 138–156, 2006.

[3] Corinna Cortes, Mehryar Mohri, and Ameet Talwalkar. On the impact of kernel approximation on learning accuracy. In *International Conference on Artificial Intelligence and Statistics*, pp. 113–120, 2010.

[4] Atsushi Nitanda and Taiji Suzuki. Stochastic gradient descent with exponential convergence rates of expected classification errors. In *International Conference on Artificial Intelligence and Statistics*, pp. 1417–1426, 2019.

[5] Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Exponential convergence of testing error for stochastic gradient methods. In *Conference on Learning Theory*, pp. 250–296, 2018.

[6] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pp. 1177–1184, 2008.

[7] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

[8] Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, Vol. 32, No. 1, pp. 135–166, 2004.