Scientific Research Publishing

Journal of Computer and Communications, 2017, 5, 28-33
http://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

A Study on Differential Private Online Learning

Weilin Nie, Cheng Wang

Huizhou University, Huizhou, China
Email: niewl@hzu.edu.cn, wangch@hzu.edu.cn

How to cite this paper: Nie, W.L. and Wang, C. (2017) A Study on Differential Private Online Learning. *Journal of Computer and Communications*, **5**, 28-33.
https://doi.org/10.4236/jcc.2017.52004

Received: December 18, 2016
Accepted: February 18, 2017
Published: February 21, 2017

Abstract

Online learning algorithms are very attractive, in which iterations are applied efficiently instead of solving some optimization problems. In this paper, online learning with protecting privacy is considered. A perturbation term is added into the classical online algorithms to obtain the differential privacy property. Firstly the distribution for the perturbation term is deduced, and then an error analysis for the new algorithms is performed, which shows the convergence and learning rate. From the error analysis, a choice for the parameters for differential privacy can be found theoretically.

Keywords

Online Learning, Differential Privacy, Output Perturbation, Error Decomposition, Learning Rate

1. Introduction

Online learning is widely used recently in computer sciences, due to its efficiency in calculation and well theoretical results. Compared with the classical batch learning in learning theory, online algorithms update the output only according to the last sample point. So such algorithms are very effective to handle the practical problems and have been studied in [1] [2] [3] [4] [5] and etc. However, as the technologic development of data analysis, there are risks for applying such algorithms on a big data set. A commonly used notion for measuring the risk is differential privacy [6]. Little references on this topic can be found except for [7]. There the authors conducted an analysis for online convex programming. Choice for the parameters of differential privacy and utilities analysis are presented for algorithms such as implicit gradient descent and generalized infinitesimal gradient descent. In this paper, the line of work begins with [8] is considered which can be thought as a kernel online learning algorithm.

2. Fundamental Principles

Our setting for online learning is introduced as follow. Let the input space X be a compact metric space, and output $Y \in [-M, M]$ for some $M > 0$ as a regression problem. Denote $Z := X \times Y$ as the sample space. Assume there is a probability measure ρ on Z, which can be decomposed to marginal distribution ρ_X on X and conditional distribution $\rho(y|x)$ on Y at $x \in X$. Then the regression function is defined by

$$f_\rho = \int_Y y \mathrm{d}\rho(y|x) \tag{1}$$

which is indeed the conditional expectation of y given x. The regression function minimizes the least square generalization error (see [9] for more details)

$$\mathcal{E}(f) := \int_Z (f(x) - y)^2 \mathrm{d}\rho \tag{2}$$

So learning algorithms always aim to approximate the regression function based on samples $\{z_t = (x_t, y_t)\}_{t=0,1,2,\cdots}$, which are drawn independently from distribution ρ. Let $K : X \times X \to R$ be a Mercer Kernel, and $\mathcal{H}_K$ is the induced reproducing kernel Hilbert space (RKHS, [10]), *i.e.*, the completion of $\operatorname{span}\{K_x : x \in X\}$ where $K_x(x') = K(x, x')$ for any $x, x' \in X$ with respect to the inner product $\langle K_x, K_{x'} \rangle_K = K(x, x')$. The corresponding norm in $\mathcal{H}_K$ is denoted as $\|\cdot\|_K$. Now our online learning algorithm as

$$f_{t+1} = f_t - \eta_t \left[(f_t(x_t) - y_t) K_{x_t} + \lambda_t f_t \right], \quad t = 0, 1, 2, \cdots \tag{3}$$

with $f_0 = 0$. Here $\eta_t > 0$ is the step size and $\lambda_t > 0$ is the regularization parameter.

When applying this online algorithm on private data set, it may leak some sensitive information. To deal with this privacy problem, Dwork *et al.* introduced differential privacy in [11]. Which can be described as follow. For the sample space Z introduced above, the Hamming distance between two sample sets $\{z_1, z_2\} \in Z^m$ is

$$\mathrm{d}(z_1, z_2) = \#\{i = 1, \cdots, m : z_{1,i} \neq z_{2,i}\} \tag{4}$$

Definition 1 *A random algorithm* $A : Z^m \to \operatorname{Range}(A)$ *is* ϵ*-differential private if for every two data sets* z_1, z_2 *satisfying* $\mathrm{d}(z_1, z_2) = 1$, *and every set* $\mathcal{O} \in \operatorname{Range}(A(z_1)) \cap \operatorname{Range}(A(z_2))$, *there holds*

$$\Pr\{A(z_1) \in \mathcal{O}\} \le e^{\epsilon} \cdot \Pr\{A(z_2) \in \mathcal{O}\} \tag{5}$$

To endow our online algorithm the differential privacy property, a perturbation term is added into the output of (3), that is,

$$f_{t,\mathcal{A}} = f_t + b_t \tag{6}$$

where b_t takes value in R with distribution to be determined in following analysis.

Differential private online learning has already been studied in [7], there the authors consider a differentially private online convex programming problem.

Here our algorithm is different, which is based on the Mercer kernels. Our purpose in this paper is to firstly provide the explicit density function for b and then conduct an error analysis for (6), which reveals the learning rate.

3. Differential Privacy Analysis

In this section, a detail analysis for the perturbation term b_t in algorithm (6) will be conducted. Firstly recall the useful definition of sensitivity and lemma proposed in [11].

Definition 2 *denote* Δf_t *as the maximum infinite norm of difference between the outputs when changing the last sample point in* z. *Let* $z=\{(x_i,y_i)\}_{i=0}^{t}$ *and* $z=\{(x_1,y_1),(x_2,y_2),\cdots,(x_{t-1},y_{t-1}),(x_{\bar{t}},y_{\bar{t}})\}$, f_t *and* $f_{\bar{t}}$ *derived from* (3) *accordingly, it is clear that*

$$\Delta f_t := \sup_{z,\bar{z}} \| f_t - f_{\bar{t}} \|_\infty \tag{7}$$

Then a similar result to [11] is:

Lemma 1 *Assume* Δf_t *is bounded by* $C_t > 0$, *and* b_t *has density function proportion to* $\exp\left\{-\frac{\epsilon|b|}{C_t}\right\}$, *then algorithm* (6) *provides* ϵ *-differential privacy.*

Proof. For all possible output function r, and $z,\bar{z}$ differ in last element, then

$$\Pr\{f_{t,\mathcal{A}} = r\} = \Pr_{b_t}\{b_t = r - f_t\} \propto \exp\left(-\frac{\epsilon|r-f_t|}{C_t}\right) \tag{8}$$

and

$$\Pr\{f_{\bar{t},\mathcal{A}} = r\} = \Pr_{b_t}\{b_t = r - f_{\bar{t}}\} \propto \exp\left(-\frac{\epsilon|r-f_{\bar{t}}|}{C_t}\right) \tag{9}$$

So by triangle inequality,

$$\Pr\{f_{t,\mathcal{A}} = r\} \le \Pr\{f_{\bar{t},\mathcal{A}} = r\} \times e^{\frac{\epsilon|f_t - f_{\bar{t}}|}{C_t}} \le e^{\epsilon} \Pr\{f_{\bar{t},\mathcal{A}} = r\} \tag{10}$$

Then the lemma is proved by a union bound. □

It is obvious that if finding the upper bound for Δf_t, the distribution for b_t can be derived. Set $\eta_t = 1/(t+t_0)^{\theta}$ and $\lambda_t = 1/(t+t_0)^{1-\theta}$ for some $t_0 > 0$ and $0<\theta<1$. Moreover, denote $\kappa = \sup_{x,x'\in X} K(x,x')$ (as K is Mercer Kernel on compact metric space X). The next lemma is taken from [1] to bound $\|f_t\|_K$.

Lemma 2 *If* $t_0^{\theta} \ge \kappa^2 + 1$, *then for all* $t \in \mathbb{N}$, *there holds*

$$\|f_t\|_K \le \frac{\kappa M}{\lambda_t} \tag{11}$$

Now the main result for differential privacy for algorithm (6) follows.

Theorem 1 *When choosing* $\eta_t = 1/(t+t_0)^{\theta}$ *and* $\lambda_t = 1/(t+t_0)^{1-\theta}$ *for some* $\frac{1}{2} < \theta < 1$ *and* $t_0^{\theta} \ge \kappa^2 + 1$, *let the density function of* b_t *is* $\frac{1}{\alpha}\exp\left\{\frac{\epsilon|b|}{C_t}\right\}$ *with* $\alpha = 2C_t/\epsilon$ *and*

$$C_t = \frac{2\left(\kappa^2+1\right)\kappa M}{\left(t-1+t_0\right)^{2\theta-1}} \tag{12}$$

then the algorithm (6) provides ϵ -differential privacy.

Proof. From (3) there holds

$$f_t = f_{t-1} - \eta_{t-1}\left[\left(f_{t-1}\left(x_{t-1}\right) - y_{t-1}\right)K_{x_{t-1}} + \lambda_{t-1} f_{t-1}\right] \tag{13}$$

and

$$f_{\bar{t}} = f_{t-1} - \eta_{t-1}\left[\left(f_{t-1}\left(x_{\bar{t}-1}\right) - y_{\bar{t}-1}\right)K_{x_{\bar{t}-1}} + \lambda_{t-1} f_{t-1}\right] \tag{14}$$

Then

$$f_t - f_{\bar{t}} = \eta_{t-1}\left[\left(f_{t-1}\left(x_{\bar{t}-1}\right) - y_{\bar{t}-1}\right)K_{x_{\bar{t}-1}} - \left(f_{t-1}\left(x_{t-1}\right) - y_{t-1}\right)K_{x_{t-1}}\right] \tag{15}$$

From the above lemma $\left\| f_{t-1} \right\|_K \le \frac{\kappa M}{\lambda_{t-1}}$ for all t. By the reproducing property that $f\left(x\right) = \left\langle f, K_x \right\rangle_K \le \left\| f \right\|_K \left\| K_x \right\|_K \le \kappa \left\| f \right\|_K$ (see [9]),

$$\left\| f_t - f_{\bar{t}} \right\|_K \le 2\eta_{t-1}\left(\frac{\kappa^2 M}{\lambda_{t-1}} + M\right)\kappa \tag{16}$$

Therefore

$$\Delta f_t = \sup_{z,\bar{z}} \left\| f_t - f_{\bar{t}} \right\|_\infty \le \frac{2\kappa^2\left(\kappa^2+1\right)M}{\left(t-1+t_0\right)^{2\theta-1}} \tag{17}$$

Set B_t to be the right hand side in lemma 1 then the theorem is proved. □

4. Error Analysis

In this section, $f_\rho \in \mathcal{H}_K$ is assumed for simplicity. It will be shown that f_t obtained from (6) still converge to regression function f_ρ by choosing appropriate parameter ϵ under the choice of η_t and λ_t as in the theorem in the last section. To this end, an error decomposition is needed. Denote operators $L_t : \mathcal{H}_K \to \mathcal{H}_K$ as $L_t\left(f\right) = f\left(x_t\right)K_{x_t}$ for $t = 0,1,2,\cdots$, and I as the identity operator. It is easy to verify that $\left\| L_t \right\| \le \kappa^2$. Notice that $f_\rho \in \mathcal{H}_K$, the following decomposition can be deduced:

$$f_{t+1} - f_\rho = \left(I - \eta_t\lambda_t I - \eta_t L_t\right)\left(f_t - f_\rho\right) + \eta_t\left[y_t K_{x_t} - \left(\lambda_t I + L_t\right)f_\rho\right] \tag{18}$$

$$= A_t\left(f_t - f_\rho\right) + B_t = A_t\left[A_{t-1}\left(f_{t-1} - f_\rho\right) + B_{t-1}\right] + B_t = \cdots \tag{19}$$

$$= A_t A_{t-1} \cdots A_0\left(f_0 - f_\rho\right) + \left[B_t + A_t B_{t-1} + \cdots + A_t A_{t-1} \cdots A_1 B_0\right] \tag{20}$$

Here $A_t = I - \eta_t\lambda_t I - \eta_t L_t$ and $B_t = \eta_t\left[y_t K_{x_t} - \left(\lambda_t I + L_t\right)f_\rho\right]$. In the following the first term is called initial error and second one is sample error. The initial error is easy to bound from the analysis above. Since t_0 is such that $t_0^\theta \ge \kappa^2 + 1$, A_t is a positive operator with $\left\| A_t \right\| \le 1 - \eta_t\lambda_t = \left(t + t_0 - 1\right)/\left(t + t_0\right)$.

$$\begin{aligned}\left\| A_t A_{t-1} \cdots A_0\left(f_0 - f_\rho\right) \right\|_K &\le \left\| A_t \right\| \left\| A_{t-1} \right\| \cdots \left\| A_0 \right\| \cdot \left\| f_0 - f_\rho \right\|_K \\ &\le \Pi_{j=0}^t \frac{j + t_0 - 1}{j + t_0} \left\| f_0 - f_\rho \right\|_K \\ &= \frac{t_0}{t + t_0} \left\| f_\rho \right\|_K .\end{aligned}$$

For the sample error, it is more difficult and the Pinelis-Bernstein inequality [4] will be applied.

Lemma 3 *Let* ξ_i *be a martingale difference sequence in a Hilbert space. Suppose that almost surely* $\|\xi_i\| \le B$ *and* $\sum_{i=1}^{t} \mathbb{E}_{i-1} \|\xi_i\|^2 \le \sigma_t^2$ *for some constants* $B, \sigma_t > 0, t = 1, 2, \cdots$. *Then for any* $0 < \delta < 1$, *with probability at least* $1-\delta$, *there holds*

$$\left\| \sum_{i=1}^{t} \xi_i \right\| \le 2\left(\frac{B}{3} + \sigma_t\right) \ln\left(\frac{2}{\delta}\right) \tag{21}$$

Now the error bounds for sample error can be derived. Notice that $\|B_t\|_K \le \eta_t \left[M\kappa + \left(\kappa^2 + \lambda_t\right) \|f_\rho\|_K \right] \le \eta_t \left[M\kappa + \left(\kappa^2 + 1\right) \|f_\rho\|_K \right]$. Set $\xi_i = A_t A_{t-1} \cdots A_i B_{i-1}, i = 1, 2, \cdots, t$, then

$$\|\xi_i\|_K \le \|A_t\| \|A_{t-1}\| \cdots \|A_i\| \|B_{i-1}\|_K \le \frac{i + t_0 - 1}{t + t_0} \eta_{i-1} \left[M\kappa + \left(\kappa^2 + 1\right) \|f_\rho\|_K \right] \tag{22}$$

$$= \frac{1}{t + t_0} \frac{1}{\lambda_{i-1}} \left[M\kappa + \left(\kappa^2 + 1\right) \|f_\rho\|_K \right] \le \frac{1}{t + t_0} \frac{1}{\lambda_t} \left[M\kappa + \left(\kappa^2 + 1\right) \|f_\rho\|_K \right] \tag{23}$$

$$= \eta_t \left[M\kappa + \left(\kappa^2 + 1\right) \|f_\rho\|_K \right]. \tag{24}$$

And $\sum_{i=1}^{t} \mathbb{E}_{i-1} \|X_i\|_K^2 \le t\eta_t^2 \left[M\kappa + \left(\kappa^2 + 1\right) \|f_\rho\|_K \right]$. So for any $0 < \delta < 1$, with probability at least $1-\delta$,

$$\begin{aligned} \left\| \sum_{i=1}^{t} \xi_i \right\|_K &\le \frac{8}{3} \sqrt{t} \eta_t \left[M\kappa + \left(\kappa^2 + 1\right) \|f_\rho\|_K \right] \ln\left(\frac{2}{\delta}\right) \\ &\le \frac{8}{3t^{\theta - 1/2}} \left[M\kappa + \left(\kappa^2 + 1\right) \|f_\rho\|_K \right] \ln\left(\frac{2}{\delta}\right) \end{aligned} \tag{25}$$

Note that $\|B_t\|_K \le \eta_t \left[M\kappa + \left(\kappa^2 + 1\right) \|f_\rho\|_K \right]$, hence

$$\|B_t + A_t B_{t-1} + \cdots + A_t A_{t-1} \cdots A_1 B_0\|_K \le \frac{11}{3t^{\theta - 1/2}} \left[M\kappa + \left(\kappa^2 + 1\right) \|f_\rho\|_K \right] \ln\left(\frac{2}{\delta}\right) \tag{26}$$

Combining the initial error, sample error bounds and applying Markov inequality for the fact that $\mathbb{E}|b_{t+1}| = C_{t+1}/\epsilon$, the total error estimation is obtained.

Theorem 2 *Choose* η_t, λ_t *and* b_t *as in the theorem in the last section, with confidence* $1 - \delta (0 < \delta < 1)$, *there holds*

$$\left\| f_{t+1,\mathcal{A}} - f_\rho \right\|_K \le C_\epsilon \frac{1}{t^{\theta - 1/2}} \frac{4}{\delta} \tag{27}$$

where constant $C_\epsilon = 4\left(\kappa^2 + 1\right)\kappa M / \epsilon + t_0 \|f_\rho\|_K + \frac{11}{3}\left(\kappa M + \left(\kappa^2 + 1\right) \|f_\rho\|_K\right)$.

5. Conclusion

In this paper, analysis is performed for the differential privacy (Theorem 1) and generalization property (Theorem 2) for the online differential private learning algorithm (6). Under the choice of parameters in our theorems, the algorithm (6) can provide ϵ-differential privacy and keep learning rate close to $1/2$, for any $\epsilon > 0$. However, this error bound is not satisfactory enough. It might be an interesting problem to promote the error bound from $2/\delta$ to $\ln(2/\delta)$ in our future work.

Fund

This work is supported by NSFC (Nos. 11326096, 11401247), NSF of Guangdong Province in China (No. 2015A030313674), Foundation for Distinguished Young Talents in Higher Education of Guangdong, China (No. 2013LYM_0089), National Social Science Fund in China (No. 15BTJ024), Planning Fund Project of Humanities and Social Science Research in Chinese Ministry of Education (No. 14YJAZH040) and Doctor Grants of Huizhou University (No. C511.0206).

References

[1] Ying, Y. and Zhou, D.X. (2006) Online Regularized Classification Algorithms. *IEEE Transactions on Information Theory*, **11**, 4775-4788. https://doi.org/10.1109/TIT.2006.883632

[2] Smale, S. and Zhou, D.X. (2009) Online Learning with Markov Sampling. *Analysis and Applications*, **7**, 87-113. https://doi.org/10.1142/S0219530509001293

[3] Guo, Z.C. and Wang, C. (2011) Online Regression with Unbounded Sampling. *International Journal of Computer Mathematics*, **88**, 2936-2941. https://doi.org/10.1080/00207160.2011.587510

[4] Tarrés, P. and Yao, Y. (2014) Online Learning as Stochastic Approximations of Regularization Paths. *IEEE Transactions on Information Theory*, **60**, 5716-5735.

[5] Ying, Y. and Zhou, D.X. (2015) Online Pairwise Learning Algorithms with Kernels. *Computer Science*, **10**, 441-449.

[6] Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006) Calibrating Noise to Sensitivity in Private Data Analysis. *Theory of Cryptography*, Springer, 265-284. https://doi.org/10.1007/11681878_14

[7] Jain, P., Kothari, P. and Thakurta, A. (2012) Differential Private Online Learning. *JMLR*: *Workshop and Conference Proceedings*, **2012**, 1-34.

[8] Smale, S. and Yao, Y. (2006) Online Learning Algorithms. *Foundations of Computational Mathematics*, **6**, 145-170. https://doi.org/10.1007/s10208-004-0160-z

[9] Cucker, F. and Smale, S. (2002) On the Mathematical Foundations of Learning. *Bulletin of the American Mathematical Society*, **39**, 1-49. https://doi.org/10.1090/S0273-0979-01-00923-5

[10] Aronszajn, N. (1950) Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, **68**, 337-404. https://doi.org/10.1090/S0002-9947-1950-0051437-7

[11] Dwork, C. (2006) Differential Privacy. *ICALP*, Springer, 1-12. https://doi.org/10.1007/11787006_1