ANZIAM J. 58(2017), 231–237
doi:10.1017/S1446181116000390

A STUDY ON THE ERROR OF DISTRIBUTED ALGORITHMS FOR BIG DATA CLASSIFICATION WITH SVM

CHENG WANG[1] and FEILONG CAO✉[1]

(Received 11 July, 2016; accepted 15 October, 2016; first published online 7 March 2017)

Abstract

The error of a distributed algorithm for big data classification with a support vector machine (SVM) is analysed in this paper. First, the given big data sets are divided into small subsets, on which the classical SVM with Gaussian kernels is used. Then, the classification error of the SVM for each subset is analysed based on the Tsybakov exponent, geometric noise, and width of the Gaussian kernels. Finally, the whole error of the distributed algorithm is estimated in terms of the error of each subset.

2010 *Mathematics subject classification*: 68T05.

Keywords and phrases: distributed algorithm, big data, support vector machine, Tsybakov exponent, geometric noise exponent.

1. Introduction

Steinwart and Scovel [7] studied the classification error of support vector machines (SVMs) based on the Tsybakov exponent, geometric noise, the varying regularization parameter, and Gaussian kernels. We discuss the classification error of a distributed algorithm for big data under their framework. For convenience, we first introduce some concepts and notation (see [3, 7, 8] for details).

Let $X \subset \mathbb{R}^d$ be the input space. To represent the two classes, the output space is written as $Y = \{-1, 1\}$. Clearly, classification algorithms produce binary classifiers, such as $C : X \to Y$, the prediction power of which can be measured with its classification error defined by

$$\mathcal{R}(C) = P(C(x) \neq y) = \int_X P(y \neq C(x)|x)\, d\rho_X,$$

where ρ is a probability distribution on $Z = X \times Y$, and ρ_x is the marginal distribution of ρ on X. The so-called Bayes rule [3] is the classifier minimizing $\mathcal{R}(C)$, and is given

[1]Applied Mathematics Department of China Jiliang University, China; e-mail: wangc628@cjlu.edu.cn, feilongcao@gmail.com.

by

$$f_c(x) = \begin{cases} 1 & \text{if } \rho(y=1|x) \geq \rho(y=-1|x), \\ -1 & \text{otherwise.} \end{cases}$$

So the excess classification error, $\mathcal{R}(C) - \mathcal{R}(f_c)$ of a classifier C can be used to measure the performance of the classifier C. We consider classifiers C_f induced by a real-valued function $f : X \to \mathbb{R}$, which is defined by $C_f = \text{sign}(f)(x)$, where $\text{sign}(f)$ is the sign function.

Xiang and Zhou [8] and Cucker and Zhou [3] presented a general convex loss function. In this paper, we consider an SVM using hinge loss ℓ [7], defined by $\ell(t) = (1-t)_+ = \max\{0, 1-t\}$. In the literature [2, 3], the Tikhonov regularization scheme with loss ℓ, Gaussian kernel K^σ and a training sample $T = \{(x_i, y_i)\}_{i=1}^n \in Z^n$ is defined as the solution of the following minimization problem:

$$\hat{f}_{T,\lambda} = \arg\min_{f \in \mathcal{H}_\sigma} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i f(x_i)) + \lambda \|f\|^2_{\mathcal{H}_\sigma} \right\}, \tag{1.1}$$

where $\lambda > 0$ is called the *regularization parameter*.

We define $\mathcal{E}^\ell(f) = \int_Z \ell(yf(x))\, d\rho$ and $f^\ell_{\sigma,\lambda} = \arg\min_{f \in \mathcal{H}_\sigma}\{\mathcal{E}^\ell(f) + \lambda\|f\|^2_{\mathcal{H}_\sigma}\}$. This function is called a *regular function*, and was used by De Vito et al. [4]. We also define

$$f^\ell_\rho(x) = \arg\min_{t \in \mathbb{R}} \int_Y \ell(yt)\, d\rho(y|x) = \arg\min_{t \in \mathbb{R}} \{\ell(t)P(y=1|x) + \ell(-t)P(y=-1|x)\},$$

for $x \in X$, almost everywhere.

Barlett et al. [1] and Cucker and Zhou [3] proved that for any measurable $f : X \to \mathbb{R}$, the inequality

$$\mathcal{R}(\text{sign}(f)) - \mathcal{R}(f_c) \leq \mathcal{E}^\ell(f) - \mathcal{E}^\ell(f_c) \tag{1.2}$$

holds. Although the theoretical properties of (1.1) have been extensively investigated, the computation of (1.1) is complicated for big data with size N.

In this work, we study the so-called distributed algorithm for big data. Recently, more researchers have worked on to this approach; for instance, McDonald et al. [6] used perceptron-based algorithms, while Kleiner et al. [5] applied a bootstrap approach. The aim of this paper is to study the binary classification error of the distributed algorithm with varying λ and σ, based on the Tsybakov noise exponent [7] and geometric exponent [7] under an SVM.

We provide some preliminaries in Section 2. The main results and associated proofs are presented in Section 3. Conclusions are presented in Section 4.

2. Preliminaries

We first describe the distributed algorithm [9, 10]. We assume that N samples $(x_1, y_1), \ldots, (x_N, y_N)$ are given, which are independently and identically distributed (i.i.d.) according to the distribution ρ on $Z = X \times Y$. Instead of solving (1.1) on all N samples, we take the following three steps.

https://doi.org/10.1017/S1446181116000390 Published online by Cambridge University Press

(I) Divide the set of samples $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ randomly and evenly into m disjoint subsets $S_1, \ldots, S_m \subset Z$, where each S_i has $n = N/m$ samples.

(II) For each $i = 1, 2, \ldots, m$, compute the local estimates

$$\hat{f}_{i,\lambda} = \arg\min_{f \in \mathcal{H}_\sigma} \left\{ \frac{1}{n} \sum_{(x,y) \in S_i} \ell(yf(x)) + \lambda \|f\|^2_{\mathcal{H}_\sigma} \right\}.$$

(III) Take athe verage of the local estimates and output $\bar{f} = (1/m) \sum_{i=1}^m \hat{f}_{i,\lambda}$.

Our aim is to estimate the error $\mathcal{R}(\text{sign}(\bar{f})) - \mathcal{R}(f_c)$. However, using (1.2), we only need to estimate $\mathcal{E}^\ell(\bar{f}) - \mathcal{E}^\ell(f_c)$. In Section 3, some results to bound $\mathcal{E}^\ell(\bar{f}) - \mathcal{E}^\ell(f_c)$ and $\mathcal{R}(\text{sign}(\bar{f})) - \mathcal{R}(f_c)$ will be presented. When solving each $\hat{f}_{i,\lambda}$, we take λ and σ as in Lemma 3.2.

First we present some concepts and lemmas (see [7] for details) on the Tsybakov noise exponent q and geometric noise exponent $\alpha > 0$ of the probability ρ. We denote $\eta(x) = \rho(y = 1|x)$ in the following.

Definition 2.1. Let $0 \le q \le \infty$ and ρ be a probability measure on $X \times Y$. We say that ρ has Tsybakov noise exponent q, if there exists a constant $C > 0$ such that

$$\rho_X(x \in X \mid |2\eta(x) - 1| \le t) \le Ct^q$$

for all sufficiently small $t > 0$.

Definition 2.2. Let $X \subset \mathbb{R}^d$ be compact and ρ be a probability measure on $X \times Y$. We say that ρ has geometric noise exponent $\alpha > 0$, if there exists a constant $C > 0$ such that

$$\int_X |2\eta(x) - 1| \exp\left(-\frac{\tau_x^2}{t}\right) \rho_X(dx) \le Ct^{\alpha d/2} \tag{2.1}$$

for all $t > 0$, where τ_x is the distance of x to the decision boundary. We say that ρ has geometric noise exponent $\alpha = \infty$, if it has geometric noise exponent α' for all $\alpha' > 0$.

Given a reproducing kernel Hilbert space $\mathcal{H}$ over X, we define the approximation error function [7] with respect to $\mathcal{H}$ and P by

$$a(\lambda) = \inf_{f \in \mathcal{H}} \{\lambda \|f\|^2_{\mathcal{H}} + \mathcal{E}^\ell(f) - \mathcal{E}^\ell(f_\rho^\ell)\}, \quad \lambda \ge 0.$$

The following two lemmas [7] are important for obtaining our results in the next section.

Lemma 2.3. *Let $\sigma > 0$, X be a closed unit ball in the Euclidean space $\mathbb{R}^d$, and $a_\sigma(.)$ be the approximation error function with respect to $\mathcal{H}_\sigma(X)$. Furthermore, let ρ be a distribution on $X \times Y$ that has geometric noise exponent $0 < \alpha < \infty$ with constant C in (2.1). Then there is a constant $c_d > 0$ depending only on the dimension d such that, for all $\lambda > 0$, we have*

$$a_\sigma(\lambda) \le c_d(\sigma^d \lambda + C(4d)^{\alpha d/2} \sigma^{-\alpha d}).$$

https://doi.org/10.1017/S1446181116000390 Published online by Cambridge University Press

LEMMA 2.4. *Let $\mathcal{F}$ be a set of bounded measurable functions on X. Using the above notation, we define*

$$\mathcal{G} = \{[\ell(yf(x)) + \lambda\|f\|^2_{\mathcal{H}_\sigma}] - [\ell(yf^\ell_{\sigma,\lambda}(x)) + \lambda\|f^\ell_{\sigma,\lambda}\|^2_{\mathcal{H}_\sigma}] \mid f \in \mathcal{F}\}.$$

Suppose that there are constants $c > 0$, $0 < \alpha \le 1$, $\delta > 0$ and $B > 0$ with $\mathbb{E}_\rho g^2 \le c(\mathbb{E}_\rho g)^\alpha + \delta$, and $\|g\|_\infty \le B$ for all $g \in \mathcal{G}$. Furthermore, assume that the covering number for $B^{-1}\mathcal{G}$ satisfies the condition of Steinwart and Scovel [7]. Then

$$P(T \in Z^n \mid [\mathcal{E}^\ell(f_T) + \lambda\|f_T\|^2] - [\mathcal{E}^l(f^\ell_{\sigma,\lambda}) + \lambda\|f^\ell_{\sigma,\lambda}\|^2] \le c_p\epsilon(n, a, B, c, \delta, x)) \ge 1 - e^{-x},$$

where $\epsilon(n, a, B, c, \delta, x)$ is as in [7].

3. Main results

LEMMA 3.1. *We have*

$$\mathcal{E}^\ell(\bar{f}) - \mathcal{E}^\ell(f_c) \le \frac{1}{m}\sum_{i=1}^m (\mathcal{E}^\ell(\hat{f}_{i,\lambda}) - \mathcal{E}^\ell(f_c)).$$

PROOF. The convexity of ℓ yields

$$\mathcal{E}^\ell(\bar{f}) = \int_Z \ell(y\bar{f}(x))\,d\rho \le \int_Z \frac{1}{m}\sum_{i=1}^m \ell(y\hat{f}_{i,\lambda}(x))\,d\rho = \frac{1}{m}\sum_{i=1}^m \int_Z \ell(y\hat{f}_{i,\lambda}(x))\,d\rho$$
$$= \frac{1}{m}\sum_{i=1}^m \mathcal{E}^\ell(\hat{f}_{i,\lambda}),$$

therefore, $\mathcal{E}^\ell(\bar{f}) - \mathcal{E}^\ell(f_c) \le (1/m)\sum_{i=1}^m (\mathcal{E}^\ell(\hat{f}_{i,\lambda}) - \mathcal{E}^\ell(f_c))$. □

Now in order to find a boundary for $\mathcal{E}^\ell(\bar{f}) - \mathcal{E}^\ell(f_c)$, we only need to estimate $\mathcal{E}^\ell(\hat{f}_{i,\lambda}) - \mathcal{E}^\ell(f_c)$ for each i. In fact, the results for each i are the same, because the $\hat{f}_{i,\lambda}$ $(i = 1, 2, \ldots, m)$ are i.i.d. distributed, and share the same properties. To estimate $\mathcal{E}^\ell(\hat{f}_{i,\lambda}) - \mathcal{E}^\ell(f_c)$, we first consider $\mathcal{E}^\ell(\hat{f}_{i,\lambda}) - \mathcal{E}^\ell(f^\ell_\rho)$ by taking the approach of Steinwart and Scovel [7] and make some modifications.

LEMMA 3.2. *Let X be a closed unit ball in the Euclidean space $\mathbb{R}^d$, and ρ be a distribution on $X \times Y$ with Tsybakov noise exponent $0 \le q \le \infty$ and geometric noise exponent $0 < \alpha < \infty$. Also, let us assume that for some $0 < \gamma \le 2$ and $0 < p < 2$,*

$$\sup_{T \in Z^n} \ln \mathcal{N}(B_{H_\sigma}, \epsilon, L_2(T_X)) \le c\sigma^{\gamma d}\epsilon^{-p}$$

holds for all $\epsilon > 0$, $\sigma \ne 1$. Given $0 \le \zeta < 1/5$, we define

$$\lambda_n = n^{-4(\alpha+1)(q+1)/[(2\alpha+1)(2q+pq+4)+4\gamma(q+1)(1-\zeta)]},$$
$$\sigma_n = \lambda_n^{-1/(\alpha+1)d}.$$

Then for any $\epsilon > 0$, there is a constant $C > 0$ such that for all $x \ge 1$ and all $n \ge 1$, with probability not less than $1 - 2e^{-x}$, we have

$$\mathcal{E}^\ell(\hat{f}_{i,\lambda_n}) - \mathcal{E}^\ell(f^\ell_\rho) \le Cx^2 n^{-((4\alpha(q+1))/((2\alpha+1)(2q+pq+4)+4\gamma(q+1)))\cdot(1/(1-\zeta))+20\zeta+\epsilon}.$$

https://doi.org/10.1017/S1446181116000390 Published online by Cambridge University Press

Proof. Iteratively using the result of Steinwart and Scovel [7, Lemma 7.2], we find a constant $C \geq 1$ such that for $Q = \{1/2(\alpha+1)\} + 4\zeta + \epsilon$, and all $x \geq 1$, $n \geq 1$, we have

$$P(\|\hat{f}_{i,\lambda_n}\| \leq Cx\lambda^{-Q}) \geq 1 - e^{-x}.$$

Let $\tilde{f}_{i,\lambda_n}$ be the minimizer of $(1/n)\sum_{i=1}^n \ell(y_i f(x_i)) + \lambda\|f\|^2_{\mathcal{H}_\sigma}$ on $Cx\lambda^{(Q-1)/2}B_{H_{\sigma_n}}$. Then $P(\tilde{f}_{i,\lambda_n} = \hat{f}_{i,\lambda_n}) \geq 1 - e^{-x}$.

By the result of Steinwart and Scovel [7, Proposition 6.8], we may choose B, a, c, δ in Lemma 2.4 such that

$$B \sim x\lambda_n^{-Q}, \quad a \sim \lambda_n^{-\gamma/(\alpha+1)}, \quad c \sim x^{(q+2)/(q+1)}\lambda_n^{-Q(q+2)/(q+1)},$$
$$\delta \sim x^{(q+2)/(q+1)}\lambda_n^{[\alpha q - Q(q+2)(\alpha+1)]/(\alpha+1)(q+1)}.$$

Then simple calculations show that

$$\epsilon(n,a,B,c,\delta,x) \leq x^2\lambda_n^{[\alpha/(\alpha+1)]-[\{2Q(\alpha+1)-1\}/2(\alpha+1)]-[\zeta\cdot(2\alpha+1)(2q+pq+4)+4\gamma(q+1)/2(\alpha+1)(2q+pq+4)]}.$$

Now from Lemma 2.4, we have

$$\begin{aligned}
&\lambda_n\|\tilde{f}_{i,\lambda_n}\|^2 + \mathcal{E}^\ell(\tilde{f}_{i,\lambda_n}) - \mathcal{E}^\ell(f_\rho^\ell) \\
&\leq \lambda_n\|f^\ell_{\sigma_n,\lambda_n}\|^2 + \mathcal{E}^\ell(f^\ell_{\sigma_n,\lambda_n}) - \mathcal{E}^\ell(f_\rho^\ell) + \tilde{C}_1x^2\lambda_n^{[\alpha/(\alpha+1)]-[\{2Q(\alpha+1)-1\}/2(\alpha+1)]-4\zeta} \\
&\leq \tilde{C}_2\lambda_n^{\alpha/(\alpha+1)} + \tilde{C}_1x^2\lambda_n^{[\alpha/(\alpha+1)]-[\{2Q(\alpha+1)-1\}/2(\alpha+1)]-4\zeta},
\end{aligned}$$

where the last inequality is due to Lemma 2.3.

Since $P(\tilde{f}_{i,\lambda_n} = \hat{f}_{i,\lambda_n}) \geq 1 - e^{-x}$, we have

$$\lambda_n\|\hat{f}_{i,\lambda_n}\|^2 + \mathcal{E}^\ell(\hat{f}_{i,\lambda_n}) - \mathcal{E}^\ell(f_\rho^\ell) \leq \tilde{C}_2\lambda_n^{\alpha/(\alpha+1)} + \tilde{C}_1x^2\lambda_n^{[\alpha/(\alpha+1)]-[\{2Q(\alpha+1)-1\}/2(\alpha+1)]-4\zeta}$$

with probability not less than $1 - 2e^{-x}$, which leads to

$$\mathcal{E}^\ell(\hat{f}_{i,\lambda_n}) - \mathcal{E}^\ell(f_\rho^\ell) \leq \tilde{C}_2\lambda_n^{\alpha/(\alpha+1)} + \tilde{C}_1x^2\lambda_n^{[\alpha/(\alpha+1)]-[\{2Q(\alpha+1)-1\}/2(\alpha+1)]-4\zeta}.$$

Due to the definition of Q, we have

$$\begin{aligned}
\lambda_n^{[\alpha/(\alpha+1)]-[\{2Q(\alpha+1)-1\}/2(\alpha+1)]-4\zeta} &= \lambda_n^{[\alpha/(\alpha+1)]-4\zeta-\epsilon-4\zeta} \\
&\leq n^{[-4\alpha(q+1)/(2\alpha+1)(2q+pq+4)+4\gamma(q+1)(1-\zeta)]+20\zeta+3\epsilon},
\end{aligned}$$

and the assertion is obtained. □

Theorem 3.3. *Under the conditions of Lemma 3.2, with probability not less than* $1 - e^{-y}$ *(for any* $y > 1$*), we have*

$$\mathcal{E}^\ell(\bar{f}) - \mathcal{E}^\ell(f_c) \leq C(y + \ln(2m))^2 n^{[-4\alpha(q+1)/(2\alpha+1)(2q+pq+4)+4\gamma(q+1)(1-\zeta)]+20\zeta+3\epsilon}.$$

Proof. First, from the definition of f_ρ^ℓ, we have

$$\mathcal{E}^\ell(\hat{f}_{i,\lambda_n}) - \mathcal{E}^\ell(f_c) \leq \mathcal{E}^\ell(\hat{f}_{i,\lambda_n}) - \mathcal{E}^\ell(f_\rho^\ell),$$

https://doi.org/10.1017/S1446181116000390 Published online by Cambridge University Press

so from Lemma 3.2, with probability at least $1 - 2e^{-x}$, we have

$$\mathcal{E}^\ell(\hat{f}_{i,\lambda_n}) - \mathcal{E}^\ell(f_c) \le Cx^2 n^{[-4\alpha(q+1)/(2\alpha+1)(2q+pq+4)+4\gamma(q+1)(1-\zeta)]+20\zeta+\epsilon}.$$

Hence, using Lemma 3.1, we get

$$\begin{aligned}
&P\{\mathcal{E}^\ell(\bar{f}) - \mathcal{E}^\ell(f_c) \le Cx^2 n^{[-4\alpha(q+1)/(2\alpha+1)(2q+pq+4)+4\gamma(q+1)(1-\zeta)]+20\zeta+\epsilon}\}\\
&\ge P\left\{\frac{1}{m}\sum_{i=1}^{m}(\mathcal{E}^\ell(\hat{f}_{i,\lambda}) - \mathcal{E}^\ell(f_c)) \le Cx^2 n^{[-4\alpha(q+1)/(2\alpha+1)(2q+pq+4)+4\gamma(q+1)(1-\zeta)]+20\zeta+\epsilon}\right\}\\
&\ge P\left\{\bigcap_{i=1}^{m}(\mathcal{E}^\ell(\hat{f}_{i,\lambda}) - \mathcal{E}^\ell(f_c) \le Cx^2 n^{[-4\alpha(q+1)/(2\alpha+1)(2q+pq+4)+4\gamma(q+1)(1-\zeta)]+20\zeta+\epsilon})\right\}\\
&\ge 1 - 2me^{-x}.
\end{aligned}$$

Replacing x by $y + \ln(2m)$, we obtain that

$$\mathcal{E}^\ell(\bar{f}) - \mathcal{E}^\ell(f_c) \le C(y + \ln(2m))^2 n^{[-4\alpha(q+1)/(2\alpha+1)(2q+pq+4)+4\gamma(q+1)(1-\zeta)]+20\zeta+\epsilon}$$

holds with probability at least $1 - e^{-y}$. □

Considering (1.2), with the same calculation as in [7], we obtain the main result.

THEOREM 3.4. *Let X be a closed unit ball in $\mathbb{R}^d$, and P be a distribution on $X \times Y$ with Tsybakov noise exponent $q \in [0, \infty]$ and geometric noise exponent $\alpha \in (0, \infty)$. We define*

$$\lambda_n = \begin{cases} n^{-(\alpha+1)/(2\alpha+1)} & \text{if } \alpha \le \dfrac{q+2}{2q},\\ n^{-[2(\alpha+1)(q+1)]/[2\alpha(q+2)+3q+4]} & \text{otherwise,}\end{cases}$$

and $\sigma_n = \lambda_n^{-1/(\alpha+1)d}$. Then for all $\epsilon > 0$, there exists $C > 0$ such that for all $y > 1$, with probability at least $1 - e^{-y}$, the algorithm satisfies

$$\mathcal{R}(\mathrm{sign}(f)) - \mathcal{R}(f_c) \le \begin{cases} C(y + \ln(2m))^2 n^{-[\alpha/(2\alpha+1)]+\epsilon} & \text{if } \alpha \le \dfrac{q+2}{2q},\\ C(y + \ln(2m))^2 n^{-[2(\alpha+1)(q+1)/\{2\alpha(q+2)+3q+4\}]+\epsilon} & \text{otherwise.}\end{cases}$$

4. Conclusion

Problems in big data analysis have recently become a hot topic. In this paper, we apply a distributed algorithm to big data classification. The classification error bound is discussed, based on Tsybakov noise and geometrical noise exponents. We use a Tikhonov regularization scheme with hinge loss to obtain estimators. Other schemes (such as back-propagation network (BPN)) and general convex loss functions can also be considered. In our work, we have taken the ordinary average of local estimators. A weighted average may lead to a better boundary for classification error, an issue we intend to address in future work.

https://doi.org/10.1017/S1446181116000390 Published online by Cambridge University Press

Acknowledgements

This work is supported by Zhejiang Provincial Natural Science Foundation of China (no. LY14A010026) and National Natural Science Foundation of China (no. 61672477).

References

[1] P. L. Bartlett, M. I. Jordan and J. D. McAuliffe, "Convexity, classification, and risk bounds", *J. Amer. Statist. Assoc.* **101** (2006) 138–156; doi:10.1198/016214505000000907.

[2] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines* (Cambridge University Press, Cambridge, 2000) http://dl.acm.org/citation.cfm?id=345662.

[3] F. Cucker and D. X. Zhou, *Learning theory: an approximation theory viewpoint* (Cambridge University Press, Cambridge, 2007); doi:10.1017/CBO9780511618796.

[4] E. DeVito, A. Caponnetto and L. Rosasco, "Model selection for regularized least-squares algorithm in learning theory", *Found. Comput. Math.* **5** (2006) 59–85; doi:10.1007/s10208-004-0134-1.

[5] A. Kleiner, A. Talwalkar, P. Sarkar and M. Jordan, "The big data bootstrap", in: *Proc. 29th Inter. Conf. Mach. Lear.* (Omnipress, Madison, WI, USA, 2012) 1759–1766; https://arxiv.org/ftp/arxiv/papers/1206/1206.6415.pdf.

[6] R. McDonald, K. Hall and G. Mann, "Distributed training strategies for the structured perceptron", in: *Proc. HLT '10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (Association for Computational Linguistics Stroudsburg, PA, USA, 2010) 456–464; http://aclweb.org/anthology/N/N10/N10-1069.pdf.

[7] I. Steinwart and C. Scovel, "Fast rates for support vector machines using Gaussian kernels", *Ann. Statist.* **35** (2007) 575–607; doi:10.1214/009053606000001226.

[8] D. H. Xiang and D. X. Zhou, "Classification with Gaussians and convex loss", *J. Mach. Learn. Res.* **10** (2009) 1147–1468; doi:10.1007/s11425-010-4043-2.

[9] Y. Zhang, J. C. Duchi and M. J. Wainwright, "Communication-efficient algorithms for statistical optimization", *J. Mach. Learn. Res.* **14** (2013) 3321–3363; doi:10.1109/CDC.2012.6426691.

[10] Y. Zhang, J. Duchi and M. Wainwright, "Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates", *J. Mach. Learn. Res.* **30** (2013) 592–617; https://arxiv.org/pdf/1305.5029v2.pdf.