Nie and Wang *Journal of Inequalities and Applications* (2017) 2017:9
DOI 10.1186/s13660-016-1280-0

Journal of Inequalities and Applications
a SpringerOpen Journal

RESEARCH Open Access

Perturbation of convex risk minimization and its application in differential private learning algorithms

Weilin Nie and Cheng Wang*

*Correspondence:
wangch@hzu.edu.cn
Huizhou University, Huizhou, P.R. China

Abstract

Convex risk minimization is a commonly used setting in learning theory. In this paper, we firstly give a perturbation analysis for such algorithms, and then we apply this result to differential private learning algorithms. Our analysis needs the objective functions to be strongly convex. This leads to an extension of our previous analysis to the non-differentiable loss functions, when constructing differential private algorithms. Finally, an error analysis is then provided to show the selection for the parameters.

Keywords: differential privacy; convex risk minimization; perturbation; concentration inequality; error decomposition

1 Introduction

In learning theory, convex optimization is one of the powerful tools in analysis and algorithm designs, which is especially used for empirical risk minimization (ERM) (Vapnik 1998 [1]). When running on a sensitive data set, algorithms may leak private information. This has motivated the notion of differential privacy (Dwork *et al.* 2006, 2016 [2, 3]).

For the sample space Z, denote the Hamming distance between two sample sets $\{\mathbf{z_1}, \mathbf{z_2}\} \in Z^m$ as

$$d(\mathbf{z_1}, \mathbf{z_2}) = \#\{i = 1, \dots, m : z_{1,i} \neq z_{2,i}\},$$

i.e., there is only one element that is different. Then ϵ-differential privacy is defined as follows.

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

$$\Pr\{A(\mathbf{z_1}) \in \mathcal{O}\} \leq e^{\epsilon} \cdot \Pr\{A(\mathbf{z_2}) \in \mathcal{O}\}.$$

Throughout the paper, we assume $\epsilon < 1$ for meaningful privacy guaranties. The relaxation (ϵ, δ)-differential privacy is also interesting and has been studied in some recent literature. However, it is out of our scope and we will just focus on the ϵ-differential privacy

throughout the paper. Extension of our results to (ϵ, δ)-differential privacy or concentrated differential privacy [3] may be studied in future work.

A mechanism obtains differential privacy usually by adding a perturbation term to an original definite output (Dwork *et al.* 2006 [4]), *i.e.*, the so-called Laplacian mechanism. McSherry and Talwar 2007 [5] proposed the exponential mechanism, which chooses an output based on its utility function. Indeed, the two mechanisms are related, and both of them are dependent with some kinds of sensitivity of the original definite output. We refer to Dwork 2008 [6] and Ji *et al.* 2014 [7] for a general idea of the differential private algorithms and applications.

A line of work, beginning with Chaudhuri *et al.* 2011 [8], introduced the output perturbation and objective perturbation algorithm to obtain differential privacy for the ERM algorithms. This is following [9–13], etc. However, most of the literature needs a differentiable loss function, sometimes a double-differentiable condition is required (see [8] for detail analysis). This limits the application for the algorithms, such as ERM algorithms with hinge loss (SVM) or pinball loss ([14]), and it motivates our work.

On the other hand, sensitivity in a differential private algorithm, which can be considered as the perturbation for the ERM algorithms, or the stability, has been studied in Bousquet and Elisseeff 2002 [15] and Shalev-Shwartz *et al.* 2010 [16] in the classical learning theory setting. More recently, the relationship between the stability and differential privacy has been revealed in Wang *et al.* 2015 [17].

The main contribution of this paper is to present a different perturbation analysis for the ERM algorithms, in which the condition is just in having convex loss functions and strongly convex regularization terms. Thus the output perturbation mechanisms can still be valid directly in SVM or other non-differentiable loss cases. Besides, an error analysis is conducted, from which we find a choice for the parameter ϵ to balance the privacy and generalization ability.

2 Perturbation analysis for ERM algorithms

In this section we consider the general regularized ERM algorithms. Let X be a compact metric space, and output $Y \subset \mathbb{R}$, where $|y| \leq M$ for some $M > 0$. (We refer to Cucker and Smale 2002 [18] and Cucker and Zhou 2007 [19] for more details as regards this learning theory setting.) A function $f_{\mathbf{z},\mathcal{A}} : X \to Y$ is obtained via some algorithm $\mathcal{A}$ based on the sample $\mathbf{z} = \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m$, which is drawn according to a distribution function ρ on the sample space $Z := X \times Y$. Furthermore, we assume there is a marginal distribution ρ_X on X and a conditional distribution $\rho(y|x)$ on Y given some x.

Firstly we introduce our notations which will be used in the following statements and analysis. Let the loss function $L(f(x), y)$ be positive and convex for the first variable. Denote

$$\mathcal{E}(f) = \int_Z L\big(f(x), y\big)\, d\rho,$$

$$\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m} \sum_{i=1}^{m} L\big(f(x_i), y_i\big).$$

Without loss of generality, we set $\bar{\mathbf{z}} = \{z_1, z_2, \ldots, z_{m-1}, \bar{z}_m\}$, which replaces the last element of $\mathbf{z}$, and $\mathbf{z}^-= \{z_1, z_2, \ldots, z_{m-1}\}$ as a sample set deleting the last element of $\mathbf{z}$. Then similar

notations can be given:

$$\mathcal{E}_{\bar{\mathbf{z}}}(f) = \frac{1}{m}\left(\sum_{i=1}^{m-1} L\big(f(x_i), y_i\big) + L\big(f(\bar{x}_m), \bar{y}_m\big)\right),$$

$$\mathcal{E}_{\mathbf{z}_-}(f) = \frac{1}{m-1}\sum_{i=1}^{m-1} L\big(f(x_i), y_i\big).$$

Denote $(\mathcal{H}_K, \|\cdot\|_K)$ as the reproducing kernel Hilbert space (RKHS) on X, *i.e.*, $\mathcal{H}_K := \overline{\operatorname{span}\{K(x,\cdot), x \in X\}}$, where $K : X \times X \to \mathbb{R}$ is a Mercer kernel. Let $K_x(y) = K(x,y)$ for any $x, y \in X$, and $\kappa = \sup_{x,y\in X}\sqrt{K(x,y)}$. Then the reproducing property tells us that $f(x) = \langle f, K_x\rangle_K$. Now a typical regularized ERM algorithm can be stated as

$$f_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}_K} \frac{1}{m}\sum_{i=1}^{m} L\big(f(x_i), y_i\big) + \lambda\Omega(f). \tag{1}$$

Here $\lambda > 0$ is the regularization parameter and $\Omega(f)$ is a γ-strongly ($\gamma > 0$) convex function with respect to the K norm, *i.e.*, for any $f_1, f_2 \in \mathcal{H}_K$ and $t \in [0,1]$,

$$\Omega\big(tf_1 + (1-t)f_2\big) \le t\Omega(f_1) + (1-t)\Omega(f_2) - \frac{\gamma}{2}t(1-t)\|f_1 - f_2\|_K^2.$$

This definition of being strongly convex is taken from Sridharan 2008 [20], where the authors derived some kind of uniform convergence under the strongly convex assumption. It has been widely used in the subsequent literature such as [8, 12, 16, 17], etc. By denoting

$$f_{\bar{\mathbf{z}}} = \arg\min_{f\in\mathcal{H}_K} \mathcal{E}_{\bar{\mathbf{z}}}(f) + \lambda\Omega(f),$$

$$f_{\mathbf{z}_-} = \arg\min_{f\in\mathcal{H}_K} \mathcal{E}_{\mathbf{z}_-}(f) + \lambda\Omega(f),$$

we have the following result.

Theorem 1 *Let $f_{\mathbf{z}}$ and $f_{\bar{\mathbf{z}}}$ be defined as above. Ω is γ-strongly convex and L is convex w.r.t. the first variable. Assume there is a $B > 0$ such that $\lambda\Omega(f_S) \le B$ and $|L(f_S(x), y)| \le B$ for any $\mathbf{S} \in Z^m$, $m \in \mathbb{N}$ and $(x,y) \in Z$. Then we have*

$$\|f_{\mathbf{z}} - f_{\bar{\mathbf{z}}}\|_K \le \sqrt{\frac{16B}{\lambda\gamma m}}.$$

Proof We will prove the result in three steps.

(1) For any $S \in Z^m$ and f_S from (1),

$$\big|\mathcal{E}_{\mathbf{z}}(f_S) - \mathcal{E}_{\bar{\mathbf{z}}}(f_S)\big| \le \frac{2B}{m}.$$

It is obvious from the definition above that

$$\big|\mathcal{E}_{\mathbf{z}}(f_{\mathbf{S}}) - \mathcal{E}_{\bar{\mathbf{z}}}(f_{\mathbf{S}})\big| \le \frac{1}{m}\big|L\big(f_{\mathbf{S}}(x_m), y_m\big) - L\big(f_{\mathbf{S}}(\bar{x}_m), \bar{y}_m\big)\big| \le \frac{2B}{m}.$$

(2) The minimization of the two objective functions are close, *i.e.*,

$$\left|\left(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})+\lambda\Omega(f_{\mathbf{z}})\right)-\left(\mathcal{E}_{\bar{\mathbf{z}}}(f_{\bar{\mathbf{z}}})+\lambda\Omega(f_{\bar{\mathbf{z}}})\right)\right|\leq\frac{2B}{m}.$$

From the notations above, we have

$$\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}_-})+\lambda\Omega(f_{\mathbf{z}_-})\geq\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})+\lambda\Omega(f_{\mathbf{z}}),$$

i.e.,

$$\begin{aligned}&\sum_{i=1}^{m}L\left(f_{\mathbf{z}_-}(x_i),y_i\right)+\lambda m\Omega(f_{\mathbf{z}_-})\\&\quad\geq\sum_{i=1}^{m}L\left(f_{\mathbf{z}}(x_i),y_i\right)+\lambda m\Omega(f_{\mathbf{z}})\\&\quad\geq\sum_{i=1}^{m-1}L\left(f_{\mathbf{z}}(x_i),y_i\right)+\lambda(m-1)\Omega(f_{\mathbf{z}})\geq\sum_{i=1}^{m-1}L\left(f_{\mathbf{z}_-}(x_i),y_i\right)+\lambda(m-1)\Omega(f_{\mathbf{z}_-}).\end{aligned}$$

A similar analysis for $f_{\bar{\mathbf{z}}}$ can be given as follows:

$$\begin{aligned}&\sum_{i=1}^{m-1}L\left(f_{\mathbf{z}_-}(x_i),y_i\right)+L\left(f_{\mathbf{z}_-}(\bar{x}_m),\bar{y}_m\right)+\lambda m\Omega(f_{\mathbf{z}_-})\\&\quad\geq\sum_{i=1}^{m-1}L\left(f_{\bar{\mathbf{z}}}(x_i),y_i\right)+L\left(f_{\bar{\mathbf{z}}}(\bar{x}_m),\bar{y}_m\right)+\lambda m\Omega(f_{\bar{\mathbf{z}}})\\&\quad\geq\sum_{i=1}^{m-1}L\left(f_{\bar{\mathbf{z}}}(x_i),y_i\right)+\lambda(m-1)\Omega(f_{\bar{\mathbf{z}}})\geq\sum_{i=1}^{m-1}L\left(f_{\mathbf{z}_-}(x_i),y_i\right)+\lambda(m-1)\Omega(f_{\mathbf{z}_-}).\end{aligned}$$

Note that $\sum_{i=1}^{m}L(f_{\mathbf{z}}(x_i),y_i)+\lambda m\Omega(f_{\mathbf{z}})$ is indeed $m(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})+\lambda\Omega(f_{\mathbf{z}}))$, and the two lower bounds above is the same, we have

$$\begin{aligned}&\left|m\left[\left(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})+\lambda\Omega(f_{\mathbf{z}})\right)-\left(\mathcal{E}_{\bar{\mathbf{z}}}(f_{\mathbf{z}})+\lambda\Omega(f_{\bar{\mathbf{z}}})\right)\right]\right|\\&\quad\leq\max\left\{L\left(f_{\mathbf{z}_-}(x_m),y_m\right)+\lambda\Omega(f_{\mathbf{z}_-}),L\left(f_{\mathbf{z}_-}(\bar{x}_m),\bar{y}_m\right)+\lambda\Omega(f_{\mathbf{z}_-})\right\}.\end{aligned}$$

We can deduce that

$$\left|\left(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})+\lambda\Omega(f_{\mathbf{z}})\right)-\left(\mathcal{E}_{\bar{\mathbf{z}}}(f_{\bar{\mathbf{z}}})+\lambda\Omega(f_{\bar{\mathbf{z}}})\right)\right|\leq\frac{2B}{m}.$$

(3) Now we can prove our main result. Since Ω is γ-strongly convex, and $L(f(x),y)$ is convex w.r.t. the first argument, which leads to the convexity of $\mathcal{E}_{\mathbf{z}}(f)$, for any $0<t<1$, it follows that

$$\begin{aligned}&\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})+\lambda\Omega(f_{\mathbf{z}})\\&\quad\leq\mathcal{E}_{\mathbf{z}}\left(tf_{\mathbf{z}}+(1-t)f_{\bar{\mathbf{z}}}\right)+\lambda\Omega\left(tf_{\mathbf{z}}+(1-t)f_{\bar{\mathbf{z}}}\right)\end{aligned}$$

$$
\begin{aligned}
&\leq t\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + (1-t)\mathcal{E}_{\mathbf{z}}(f_{\bar{\mathbf{z}}}) + \lambda\left[t\Omega(f_{\mathbf{z}}) + (1-t)\Omega(f_{\bar{\mathbf{z}}}) - \frac{\gamma}{2}t(1-t)\|f_{\mathbf{z}} - f_{\bar{\mathbf{z}}}\|_K^2\right] \\
&= t\big(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\Omega(f_{\mathbf{z}})\big) + (1-t)\big(\mathcal{E}_{\mathbf{z}}(f_{\bar{\mathbf{z}}}) + \lambda\Omega(f_{\bar{\mathbf{z}}})\big) - \frac{\lambda\gamma}{2}t(1-t)\|f_{\mathbf{z}} - f_{\bar{\mathbf{z}}}\|_K^2 \\
&\overset{(1)}{\leq} t\big(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\Omega(f_{\mathbf{z}})\big) + (1-t)\left(\mathcal{E}_{\bar{\mathbf{z}}}(f_{\bar{\mathbf{z}}}) + \lambda\Omega(f_{\bar{\mathbf{z}}}) + \frac{2B}{m}\right) - \frac{\lambda\gamma}{2}t(1-t)\|f_{\mathbf{z}} - f_{\bar{\mathbf{z}}}\|_K^2 \\
&\overset{(2)}{\leq} t\big(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\Omega(f_{\mathbf{z}})\big) + (1-t)\left(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\Omega(f_{\mathbf{z}}) + \frac{4B}{m}\right) - \frac{\lambda\gamma}{2}t(1-t)\|f_{\mathbf{z}} - f_{\bar{\mathbf{z}}}\|_K^2 \\
&= \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\Omega(f_{\mathbf{z}}) + \frac{4(1-t)B}{m} - \frac{\lambda\gamma}{2}t(1-t)\|f_{\mathbf{z}} - f_{\bar{\mathbf{z}}}\|_K^2.
\end{aligned}
$$

Therefore,

$$
\frac{\lambda\gamma t}{2}\|f_{\mathbf{z}} - f_{\bar{\mathbf{z}}}\|_K^2 \leq \frac{4B}{m}.
$$

Simply taking $t = \frac{1}{2}$ we have

$$
\|f_{\mathbf{z}} - f_{\bar{\mathbf{z}}}\|_K \leq \sqrt{\frac{16B}{\lambda\gamma m}},
$$

which proves our result. □

Now let us make a brief remark about this result. In our theorem, only convexity for the loss function and γ-strongly convexity for Ω are assumed. The assumption $\lambda\Omega(f_{\mathbf{S}}) \leq B$ is trivial for algorithms such as general SVM or coefficient regularization [21], since $\mathcal{E}_{\mathbf{S}}(f_{\mathbf{S}}) + \lambda\Omega(f_{\mathbf{S}})$ is the minimum value. The advantage of this result is that most of our learning algorithms satisfy this condition, especially including hinge loss for SVM and pinball loss for quantile regression. Perturbation, or stability analysis has already been performed in [15, 16]. There the authors proposed quite a few stability definitions, which is mainly used for classical generalization analysis. References [10, 22] also studied the differential private learning algorithms with different kernels and Lipschitz losses, with a regularization term of square norm. A similar result to theirs with our notations is as follows.

Theorem 2 *Let $f_{\mathbf{z}}$, $f_{\bar{\mathbf{z}}}$, $f_{\mathbf{z}-}$ be defined as above. Assume $|L(t_1, y) - L(t_2, y)| \leq C_L|t_1 - t_2|$ for any t_1, t_2, y and some $C_L > 0$, then we have*

$$
\|f_{\mathbf{z}} - f_{\bar{\mathbf{z}}}\|_K \leq \frac{2\kappa C_L}{\lambda\gamma m}.
$$

Proof From the convexity of the loss function and regularization term, we have, for any $f \in \mathcal{H}_K$ and $0 < t < 1$,

$$
\begin{aligned}
\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\Omega(f_{\mathbf{z}}) &\leq \mathcal{E}_{\mathbf{z}}\big(tf_{\mathbf{z}} + (1-t)f\big) + \lambda\Omega\big(tf_{\mathbf{z}} + (1-t)f\big) \\
&\leq t\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + (1-t)\mathcal{E}_{\mathbf{z}}(f) + \lambda\left[t\Omega(f_{\mathbf{z}}) + (1-t)\Omega(f) - \frac{\gamma}{2}t(1-t)\|f - f_{\mathbf{z}}\|_K^2\right].
\end{aligned}
$$

This leads to

$$(1-t)\big(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})+\lambda\Omega(f_{\mathbf{z}})\big)\le(1-t)\big(\mathcal{E}_{\mathbf{z}}(f)+\lambda\Omega(f)\big)-\frac{\lambda\gamma}{2}t(1-t)\|f-f_{\mathbf{z}}\|_K^2,$$

i.e.,

$$\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})+\lambda\Omega(f_{\mathbf{z}})\le\mathcal{E}_{\mathbf{z}}(f)+\lambda\Omega(f)-\frac{\lambda\gamma}{2}t\|f-f_{\mathbf{z}}\|_K^2.$$

Let t tend to 1, we have

$$\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})+\lambda\Omega(f_{\mathbf{z}})\le\mathcal{E}_{\mathbf{z}}(f)+\lambda\Omega(f)-\frac{\lambda\gamma}{2}\|f-f_{\mathbf{z}}\|_K^2$$

for any $f\in\mathcal{H}_K$. Similarly, we also have

$$\mathcal{E}_{\bar{\mathbf{z}}}(f_{\bar{\mathbf{z}}})+\lambda\Omega(f_{\bar{\mathbf{z}}})\le\mathcal{E}_{\mathbf{z}}(f)+\lambda\Omega(f)-\frac{\lambda\gamma}{2}\|f-f_{\bar{\mathbf{z}}}\|_K^2$$

for any $f\in\mathcal{H}_K$. Therefore,

$$\begin{aligned}&\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})+\lambda\Omega(f_{\mathbf{z}})\le\mathcal{E}_{\mathbf{z}}(f_{\bar{\mathbf{z}}})+\lambda\Omega(f_{\bar{\mathbf{z}}})-\frac{\lambda\gamma}{2}\|f_{\bar{\mathbf{z}}}-f_{\mathbf{z}}\|_K^2,\\&\mathcal{E}_{\bar{\mathbf{z}}}(f_{\bar{\mathbf{z}}})+\lambda\Omega(f_{\bar{\mathbf{z}}})\le\mathcal{E}_{\bar{\mathbf{z}}}(f_{\bar{\mathbf{z}}})+\lambda\Omega(f_{\bar{\mathbf{z}}})-\frac{\lambda\gamma}{2}\|f_{\mathbf{z}}-f_{\bar{\mathbf{z}}}\|_K^2.\end{aligned}$$

By adding the two equations we have

$$\begin{aligned}\lambda\gamma\|f_{\bar{\mathbf{z}}}-f_{\mathbf{z}}\|_K^2&\le\big(\mathcal{E}_{\bar{\mathbf{z}}}(f_{\mathbf{z}})-\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})\big)+\big(\mathcal{E}_{\bar{\mathbf{z}}}(f_{\mathbf{z}})\mathcal{E}_{\bar{\mathbf{z}}}(f_{\bar{\mathbf{z}}})\big)\\&=\frac{1}{m}\big(L\big(f_{\mathbf{z}}(\bar{x}_m),\bar{y}_m\big)-L\big(f_{\mathbf{z}}(x_m),y_m\big)\big)+\frac{1}{m}\big(L\big(f_{\bar{\mathbf{z}}}(x_m),y_m\big)-L\big(f_{\bar{\mathbf{z}}}(\bar{x}_m),\bar{y}_m\big)\big)\\&\le\frac{2C_L}{m}\|f_{\mathbf{z}}-f_{\bar{\mathbf{z}}}\|_\infty.\end{aligned}$$

From the fact that $\|f\|_\infty=\sup_{x\in X}|f(x)|\le\sup_{x\in X}\langle f,K_x\rangle_K\le\kappa\|f\|_K$ for any $f\in\mathcal{H}_K$ we have

$$\|f_{\bar{\mathbf{z}}}-f_{\mathbf{z}}\|_K\le\frac{2\kappa C_L}{\lambda\gamma m},$$

and the theorem is proved. □

Though the condition for the latter result is stronger than the first one, we will still apply this to the analysis below, as the bound is sharper and most of the loss functions satisfy the Lipschitz condition above.

3 Differential private learning algorithms

In this section, we will describe the general differential private learning algorithms based on an output perturbation method. Perturbation ERM algorithms give a random output by adding a random perturbation term on the above deterministic output. That is,

$$f_{\mathcal{A},\mathbf{z}}=f_{\mathbf{z}}+b,\tag{2}$$

where $f_{\mathbf{z}}$ is derived from (1). To determine the distribution of b, we firstly recall the sensitivity, introduced in Dwork 2006 [2], in our settings.

Definition 2 We denote Δf as the maximum infinite norm of difference between the outputs when changing one sample point in $\mathbf{z}$. Let $\mathbf{z}$ and $\bar{\mathbf{z}}$ be defined as in the previous section, and $f_{\mathbf{z}}$ and $f_{\bar{\mathbf{z}}}$ be derived from (1) accordingly, we can see that

$$\Delta f := \sup_{\mathbf{z},\bar{\mathbf{z}}} \|f_{\mathbf{z}} - f_{\bar{\mathbf{z}}}\|_\infty.$$

Then a similar result to [2] is the following.

Lemma 1 *Assume Δf is bounded by $B_\Delta > 0$, and b has a density function proportional to* $\exp\{-\frac{\epsilon|b|}{B_\Delta}\}$, *then algorithm* (2) *provides ϵ-differential privacy.*

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

$$\Pr\{f_{\mathbf{z},\mathcal{A}} = r\} = \Pr_b\{b = r - f_{\mathbf{z}}\} \propto \exp\left(-\frac{\epsilon|r - f_{\mathbf{z}}|}{B_\Delta}\right)$$

and

$$\Pr\{f_{\bar{\mathbf{z}},\mathcal{A}} = r\} = \Pr_b\{b = r - f_{\bar{\mathbf{z}}}\} \propto \exp\left(-\frac{\epsilon|r - f_{\bar{\mathbf{z}}}|}{B_\Delta}\right).$$

So by the triangle inequality,

$$\Pr\{f_{\mathbf{z},\mathcal{A}} = r\} \le \Pr\{f_{\bar{\mathbf{z}},\mathcal{A}} = r\} \times e^{\frac{\epsilon|f_{\mathbf{z}} - f_{\bar{\mathbf{z}}}|}{B_\Delta}} \le e^\epsilon \Pr\{f_{\bar{\mathbf{z}},\mathcal{A}} = r\}.$$

Then the lemma is proved by a union bound. □

Combining this with the result in the previous section, we can choose the noise term b as follows.

Proposition 1 *Assume the conditions in Theorem* 1 *hold, and b takes value in* $(-\infty, +\infty)$, *we choose the density of b to be* $\frac{1}{\alpha}\exp(-\frac{\lambda\gamma m\epsilon|b|}{\kappa^2 C_L})$, *where* $\alpha = \frac{2\kappa^2 C_L}{\lambda\gamma m\epsilon}$, *then the algorithm* (2) *provides ϵ-differential privacy.*

Proof Since from the previous section we have

$$\|f_{\mathbf{z}} - f_{\bar{\mathbf{z}}}\|_K \le \frac{2\kappa C_L}{\lambda\gamma m}$$

for any $\mathbf{z}$ and $\bar{\mathbf{z}}$ differing in the last sample point. Then from the reproducing property,

$$\Delta f_{\mathbf{z}} = \sup_{\mathbf{z},\bar{\mathbf{z}}} \|f_{\mathbf{z}} - f_{\bar{\mathbf{z}}}\|_\infty \le \frac{2\kappa^2 C_L}{\lambda\gamma m}.$$

The proposition is proved by substitute $B_\Delta = \frac{2\kappa^2 C_L}{\lambda\gamma m}$ in the last lemma. □

4 Error analysis

In this section, we conduct the error analysis for the general differential private ERM algorithm (2). We denote

$$f_\rho = \arg\min_f \mathcal{E}(f) = \arg\min_f \int_Z L\big(f(x), y\big)\, d\rho$$

as our goal function. In the following in this section, we always assume the Lipshitz continuous condition for the loss function, *i.e.* $|L(t_1, y) - L(t_2, y)| \le C_L|t_1 - t_2|$ for any t_1, t_2, y and some $C_L > 0$. Now let us introduce our error decomposition,

$$\begin{aligned}
\mathcal{E}(f_{\mathbf{z},\mathcal{A}}) - \mathcal{E}(f_\rho) &\le \mathcal{E}(f_{\mathbf{z},\mathcal{A}}) - \mathcal{E}(f_\rho) + \lambda\Omega(f_{\mathbf{z}}) \\
&\le \mathcal{E}(f_{\mathbf{z},\mathcal{A}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\mathcal{A}}) + \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\mathcal{A}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\Omega(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \\
&\le \mathcal{E}(f_{\mathbf{z},\mathcal{A}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\mathcal{A}}) + \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\mathcal{A}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \mathcal{E}_{\mathbf{z}}(f_\lambda) + \lambda\Omega(f_\lambda) - \mathcal{E}(f_\rho) \\
&\le \mathcal{R}_1 + \mathcal{R}_2 + \mathcal{S} + D(\lambda), \qquad (3)
\end{aligned}$$

where f_λ is a function in $\mathcal{H}_K$ to be determined and

$$\begin{aligned}
&\mathcal{R}_1 = \mathcal{E}(f_{\mathbf{z},\mathcal{A}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\mathcal{A}}), \qquad \mathcal{R}_2 = \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\mathcal{A}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}), \\
&\mathcal{S} = \mathcal{E}_{\mathbf{z}}(f_\lambda) - \mathcal{E}(f_\lambda), \qquad D(\lambda) = \mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho) + \lambda\Omega(f_\lambda).
\end{aligned}$$

Here $\mathcal{R}_1$ and $\mathcal{R}_2$ involve the function $f_{\mathbf{z},\mathcal{A}}$ from random algorithm (2) so we call them random errors. $\mathcal{S}$ and $D(\lambda)$ are similar to the classical ones in the literature in learning theory and are called sample error and approximation error. In the following we will study these errors, respectively.

4.1 Concentration inequality and error bounds for random errors

To bound the first random error, we need a concentration inequality. Dwork *et al.* 2015 [23] have proposed such an inequality under their differential private setting. Soon Bassily *et al.* 2015 [13] gave a different proof for the concentration inequality, which enlightens our error analysis.

Theorem 3 *If an algorithm $\mathcal{A}$ provides ϵ-differential privacy, and outputs a positive function $g_{\mathbf{z},\mathcal{A}} : Z \to \mathbb{R}$ with bounded expectation $\mathbb{E}_{\mathbf{z},\mathcal{A}} \frac{1}{m}\sum_{i=1}^m g_{\mathbf{z},\mathcal{A}}(z_i) \le G$ some $G > 0$, where the expectation is taken over the sample and the output of the random algorithm. Then*

$$\mathbb{E}_{\mathbf{z},\mathcal{A}}\left(\frac{1}{m}\sum_{i=1}^m g_{\mathbf{z},\mathcal{A}}(z_i) - \int_Z g_{\mathbf{z},\mathcal{A}}(z)\, d\rho\right) \le 2G\epsilon$$

and

$$\mathbb{E}_{\mathbf{z},\mathcal{A}}\left(\int_Z g_{\mathbf{z},\mathcal{A}}(z)\, d\rho - \frac{1}{m}\sum_{i=1}^m g_{\mathbf{z},\mathcal{A}}(z_i)\right) \le 2G\epsilon.$$

Proof Denote the sample sets $\mathbf{w}_j = \{z_1, z_2, \ldots, z_{j-1}, z'_j, z_{j+1}, \ldots, z_m\}$ for $j \in \{1, 2, \ldots, m\}$. We observe that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z},\mathcal{A}}\left(\frac{1}{m}\sum_{i=1}^{m} g_{\mathbf{z},\mathcal{A}}(z_i)\right)\\
&= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{\mathbf{z}}\mathbb{E}_{\mathcal{A}}\left(g_{\mathbf{z},\mathcal{A}}(z_i)\right)\\
&= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{\mathbf{z}}\mathbb{E}_{z'_i}\int_0^{+\infty} \Pr_{\mathcal{A}}\left\{g_{\mathbf{z},\mathcal{A}}(z_i) \geq t\right\} dt \leq \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{\mathbf{z}}\mathbb{E}_{z'_i}\int_0^{+\infty} e^{\epsilon} \Pr_{\mathcal{A}}\left\{g_{\mathbf{w}_i,\mathcal{A}}(z_i) \geq t\right\} dt\\
&= e^{\epsilon}\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{\mathbf{w}_i}\mathbb{E}_{z_i}\mathbb{E}_{\mathcal{A}}\left(g_{\mathbf{w}_i,\mathcal{A}}(z_i)\right) = e^{\epsilon}\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{\mathbf{w}_i,\mathcal{A}}\mathbb{E}_{z_i}\left(g_{\mathbf{w}_i,\mathcal{A}}(z_i)\right)\\
&= e^{\epsilon}\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{\mathbf{w}_i,\mathcal{A}}\int_Z g_{\mathbf{w}_i,\mathcal{A}}(z)\,d\rho = e^{\epsilon}\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{\mathbf{z},\mathcal{A}}\int_Z g_{\mathbf{z},\mathcal{A}}(z)\,d\rho\\
&= e^{\epsilon}\mathbb{E}_{\mathbf{z},\mathcal{A}}\int_Z g_{\mathbf{z},\mathcal{A}}(z)\,d\rho.
\end{aligned}$$

Then

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z},\mathcal{A}}\left(\frac{1}{m}\sum_{i=1}^{m} g_{\mathbf{z},\mathcal{A}}(z_i) - \int_Z g_{\mathbf{z},\mathcal{A}}(z)\,d\rho\right)\\
&\leq \left(1 - e^{-\epsilon}\right)\mathbb{E}_{\mathbf{z},\mathcal{A}}\left(\frac{1}{m}\sum_{i=1}^{m} g_{\mathbf{z},\mathcal{A}}(z_i)\right) \leq 2G\epsilon.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z},\mathcal{A}}\int_Z g_{\mathbf{z},\mathcal{A}}(z)\,d\rho\\
&= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{\mathbf{z}}\mathbb{E}_{\mathcal{A}}\int_Z g_{\mathbf{z},\mathcal{A}}(z)\,d\rho\\
&= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{\mathbf{w}_i}\mathbb{E}_{\mathcal{A}}\int_Z g_{\mathbf{w}_i,\mathcal{A}}(z)\,d\rho = \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{\mathbf{w}_i}\mathbb{E}_{\mathcal{A}}\int_Z g_{\mathbf{w}_i,\mathcal{A}}(z_i)\,d\rho(z_i)\\
&= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{\mathbf{w}_i}\mathbb{E}_{z_i}\mathbb{E}_{\mathcal{A}}\left(g_{\mathbf{w}_i,\mathcal{A}}(z_i)\right) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{\mathbf{z}}\mathbb{E}_{z'_i}\int_0^{+\infty} \Pr_{\mathcal{A}}\left\{g_{\mathbf{w}_i,\mathcal{A}}(z_i) \geq t\right\} dt\\
&\leq \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{\mathbf{z}}\mathbb{E}_{z'_i}e^{\epsilon}\int_0^{+\infty} \Pr_{\mathcal{A}}\left\{g_{\mathbf{z},\mathcal{A}}(z_i) \geq t\right\} dt\\
&= e^{\epsilon}\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{\mathbf{z}}\mathbb{E}_{\mathcal{A}}\left(g_{\mathbf{z},\mathcal{A}}(z_i)\right) = e^{\epsilon}\mathbb{E}_{\mathbf{z},\mathcal{A}}\frac{1}{m}\sum_{i=1}^{m} g_{\mathbf{z},\mathcal{A}}(z_i).
\end{aligned}$$

This leads to

$$\mathbb{E}_{\mathbf{z},\mathcal{A}}\left(\int_Z g_{\mathbf{z},\mathcal{A}}(z)\,d\rho - \frac{1}{m}\sum_{i=1}^m g_{\mathbf{z},\mathcal{A}}(z_i)\right)$$
$$= \left(e^{\epsilon}-1\right)\mathbb{E}_{\mathbf{z},\mathcal{A}}\frac{1}{m}\sum_{i=1}^m g_{\mathbf{z},\mathcal{A}}(z_j) \le 2G\epsilon.$$

These verify our results. □

Remark 1 In [23] and [13], the authors restrict the function to take values in $[0,1]$ or $\{0,1\}$ for their special use, our result here extends the result to the function taking values in $\mathbb{R}^+$. This makes our following error analysis implementable.

Since y is bounded by $M > 0$ throughout our paper, it is reasonable to assume that $\mathcal{E}_{\mathbf{z}}(0) = \frac{1}{m}\sum_{i=1}^m L(0,y_i) \le B_0$ for some $B_0 > 0$ depending just on M. Then we apply this concentration inequality to the random error $\mathcal{R}_1$.

Proposition 2 *Let $f_{\mathbf{z},\mathcal{A}}$ be obtained from algorithm* (2). *Assume $\mathcal{E}_{\mathbf{z}}(0) \le B_0$ for some constant $B_0 > 0$. We have*

$$\mathbb{E}_{\mathbf{z},\mathcal{A}}\mathcal{R}_1 = \mathbb{E}_{\mathbf{z},\mathcal{A}}\left(\mathcal{E}(f_{\mathbf{z},\mathcal{A}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\mathcal{A}})\right) \le 2\tilde{B}\epsilon + 2\epsilon\mathbb{E}_{\mathbf{z},\mathcal{A}}\mathcal{R}_2,$$

where $\tilde{B} = 2(B_0 + \lambda\Omega(0))$ is a constant independent of m.

Proof Let $g_{\mathbf{z},\mathcal{A}}(z) = L(f_{\mathbf{z},\mathcal{A}}(x), y)$, which is always positive. Note that

$$\mathbb{E}_{\mathbf{z},\mathcal{A}}\left(\frac{1}{m}\sum_{i=1}^m g_{\mathbf{z},\mathcal{A}}(z_i)\right) = \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\mathbf{z},\mathcal{A}}L\left(f_{\mathbf{z},\mathcal{A}}(x_i), y_i\right) = \mathbb{E}_{\mathbf{z},\mathcal{A}}\mathcal{R}_2 + \mathbb{E}_{\mathbf{z},\mathcal{A}}\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})$$

and

$$\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) \le \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\Omega(f_{\mathbf{z}}) \le \mathcal{E}_{\mathbf{z}}(0) + \lambda\Omega(0) \le B_0 + \lambda\Omega(0),$$

we have

$$\mathbb{E}_{\mathbf{z},\mathcal{A}}\left(\frac{1}{m}\sum_{i=1}^m g_{\mathbf{z},\mathcal{A}}(z_i)\right) \le \mathbb{E}_{\mathbf{z},\mathcal{A}}\mathcal{R}_2 + B_0 + \lambda\Omega(0).$$

By applying the concentration inequality for the given $g_{\mathbf{z},\mathcal{A}}$ we can prove the result with constant $\tilde{B} = 2(B_0 + \lambda\Omega(0))$. □

For the random error $\mathcal{R}_2$, we have the following estimation.

Proposition 3 *For the function $f_{\mathbf{z},\mathcal{A}}$ obtained from algorithm* (2), *we have*

$$\mathbb{E}_{\mathbf{z},\mathcal{A}}\mathcal{R}_2 = \mathbb{E}_{\mathbf{z},\mathcal{A}}\left(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\mathcal{A}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})\right) \le \frac{\kappa^2 C_L}{\lambda\gamma m\epsilon}.$$

Proof Note that

$$\left|L\left(f_{\mathbf{z},\mathcal{A}}(x_i),y_i\right)-L\left(f_{\mathbf{z}}(x_i),y_i\right)\right|\le C_L\left|f_{\mathbf{z},\mathcal{A}}(x_i)-f_{\mathbf{z}}(x_i)\right|=C_L|b|.$$

Therefore,

$$\begin{aligned}\mathbb{E}_{\mathbf{z},\mathcal{A}}\mathcal{R}_2&=\mathbb{E}_{\mathbf{z},\mathcal{A}}\left(\frac{1}{m}\sum_{i=1}^{m}\left[L\left(f_{\mathbf{z},\mathcal{A}}(x_i),y_i\right)-L\left(f_{\mathbf{z}}(x_i),y_i\right)\right]\right)\\&\le\mathbb{E}_{\mathbf{z},\mathcal{A}}C_L|b|=C_L\mathbb{E}_b|b|=\frac{\kappa^2C_L}{\lambda\gamma m\epsilon}.\end{aligned}$$

This verifies our bound. □

4.2 Error estimate for the other error terms

For the sample error and approximation error, we choose f_λ to be some function in $\mathcal{H}_K$ close to f_ρ, which satisfies $|L(f_\lambda(x),y)|\le B_\rho$ for some $B_\rho>0$. Explicit expressions of f_λ and B_ρ will be presented in the next section, with respect to different algorithms. To bound the sample error, we should recall the Hoeffding inequality [24].

Lemma 2 *Let ξ be a random variable on a probability space Z satisfying $|\xi(z)-\mathbb{E}\xi|\le\Xi$ for some $\Xi>0$ for almost all $z\in Z$. Denote $\sigma^2=\sigma^2(\xi)$, then, for any $t>0$,*

$$\Pr\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i)-\mathbb{E}\xi\ge t\right|\right\}\le 2\exp\left\{-\frac{mt^2}{2\Xi^2}\right\}.$$

Now we have the following proposition.

Proposition 4 *Let $L(f_\lambda(x),y)\le B_\rho$ for any $(x,y)\in Z$, we have*

$$\mathbb{E}_{\mathbf{z},\mathcal{A}}\mathcal{S}\le\frac{2\sqrt{2\pi}B_\rho}{\sqrt{m}}.$$

Proof Since

$$\mathcal{S}=\int_Z L\left(f_\lambda(x),y\right)d\rho-\frac{1}{m}\sum_{i=1}^{m}L\left(f_\lambda(x_i),y_i\right),$$

we apply the Hoeffding inequality to $\xi(z)=-L(f_\lambda(x),y)$. Note that $|\xi-\mathbb{E}\xi|\le 2B_\rho$ and

$$\Pr_{\mathbf{z}}\left\{\left|\int_Z L\left(f_\lambda(x),y\right)d\rho-\frac{1}{m}\sum_{i=1}^{m}L\left(f_\lambda(x_i),y_i\right)\right|\ge\varepsilon\right\}\le 2\exp\left\{-\frac{m\varepsilon^2}{8B_\rho^2}\right\}.$$

Therefore

$$\begin{aligned}\mathbb{E}_{\mathbf{z},\mathcal{A}}\mathcal{S}\le\mathbb{E}_{\mathbf{z}}|\mathcal{S}|&=\int_0^{+\infty}\Pr_{\mathbf{z}}\left\{|\mathcal{S}|\ge t\right\}dt\\&\le\int_0^{+\infty}2\exp\left\{-\frac{mt^2}{8B_\rho^2}\right\}dt\le\frac{2\sqrt{2\pi}B_\rho}{\sqrt{m}},\end{aligned}$$

and the proposition is proved. □

Let us turn to the approximation error $D(\lambda)$. It is difficult to give the upper bound for the abstract approximation error. So we use the natural assumption on $D(\lambda)$, which is

$$D(\lambda) \leq c_\beta \lambda^\beta, \tag{4}$$

for some $0 < \beta < 1$ and $c_\beta > 0$. This assumption is trivial in concrete algorithms; see [25–27], etc.

4.3 Total error bound

Now we can deduce our total error by combining all the error bounds above.

Theorem 4 *Let $f_{\mathbf{z},\mathcal{A}}$ defined as* (2), *f_ρ defined as above. Assume $\mathcal{E}_\mathbf{z}(0) \leq B_0$, $|L(f_\lambda(x), y)| \leq B_\rho$, and* (4) *hold. By choosing $\epsilon = 1/\sqrt{\lambda m}$ and $\lambda = m^{-1/(2\beta+1)}$ we have*

$$\mathbb{E}_{\mathbf{z},\mathcal{A}}\big(\mathcal{E}(f_{\mathbf{z},\mathcal{A}}) - \mathcal{E}(f_\rho)\big) \leq \left(2B_0 + 2\Omega(0) + \frac{3\kappa^2 C_L}{\gamma} + c_\beta\right)\left(\frac{1}{m}\right)^{\frac{\beta}{2\beta+1}}.$$

Proof By substituting the upper bounds above in the error decomposition (3), we have

$$\mathbb{E}_{\mathbf{z},\mathcal{A}}\big(\mathcal{E}(f_{\mathbf{z},\mathcal{A}}) - \mathcal{E}(f_\rho)\big) \leq 2\big(B_0 + \lambda\Omega(0)\big)\epsilon + \frac{(1+2\epsilon)\kappa^2 C_L}{\lambda\gamma m\epsilon} + \frac{2\sqrt{2\pi}B_\rho}{\sqrt{m}} + c_\beta\lambda^\beta.$$

Take $\epsilon = 1/\sqrt{\lambda m}$ and $\lambda = m^{-1/(2\beta+1)}$ for balance, then the result is proved. □

Here we present a general convergence result for the general differential private ERM learning algorithms. In this theorem, we provide a choice for the parameters ϵ and λ, under some conditions above, which leads to a learning rate $m^{-\beta/(2\beta+1)}$ with fixed B and γ. However, in an explicit algorithm B and γ may depend on λ and the learning rate will vary accordingly. We cannot go further without a specific description of the algorithms, which will be studied in the next section.

5 Applications

In this section, we will apply our results to several frequently used learning algorithms. First of all, let us take a look at the assumptions as regards f_ρ. Denote the integral operator L_K as $L_K f(t) = \int_X f(x)K(x,t)\,d\rho_X(x)$. It is well known that [18] $\|L_K\| \leq \kappa^2$. Then $f_\rho \in L_K^r(L^2_{\rho_X})$ for some $r > 0$ is often used in learning theory literature. When $r = 1/2$, it is the same as $f_\rho \in \mathcal{H}_K$ [18]. It is natural if we consider $L(\pi(f(x)), y) \leq L(f(x), y)$ for any function f and $(x, y) \in Z$, which means $\pi(f(x))$ is more close than $f(x)$ to y in some sense, as $|y| \leq M$. Here

$$\pi\big(f(x)\big) = \begin{cases} M, & f(x) > M, \\ f(x), & -M \leq f(x) \leq M, \\ -M, & f(x) < -M. \end{cases}$$

Then $\int_Z (\pi(f_\rho(x)), y)\,d\rho \leq \int_Z (f_\rho(x), y)\,d\rho$, *i.e.*, $|f_\rho(x)| \leq M$ always holds. So without loss of generality, we also assume $\|f_\rho\|_\infty \leq M$.

5.1 Differential private least squares regularization

Our first example is the differential private least squares regularization algorithm,

$$f_{\mathbf{z}}^{ls} = \arg\min_{f\in\mathcal{H}_K} \frac{1}{m}\sum_{i=1}^{m}\big(f(x_i)-y_i\big)^2 + \lambda\|f\|_K^2,$$

and perturbation

$$f_{\mathbf{z},\mathcal{A}}^{ls} = f_{\mathbf{z}}^{ls} + b_{ls}.$$

Such an algorithm has been studied in our previous work [28]. Now we will try to apply the above analysis. Firstly we can verify that $\Omega(f) = \|f\|_K^2$ is 2-strongly convex, *i.e.*, $\gamma = 2$ in our settings. Since $\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}^{ls}) + \lambda\|f_{\mathbf{z}}^{ls}\|_K^2 \le \mathcal{E}_{\mathbf{z}}(0) + 0 \le M^2$ with $|y| \le M$, we have $\|f_{\mathbf{z}}^{ls}\|_K \le \frac{M}{\sqrt{\lambda}}$, which leads to $\|f_{\mathbf{z}}^{ls}\|_\infty \le \frac{\kappa M}{\sqrt{\lambda}}$ for any $\mathbf{z}\in Z^m$. Therefore though the least square loss is not Lipschitz continuous, it satisfies

$$\begin{aligned}
&\big|L\big(f_{S_1}^{ls}(x),y\big) - L\big(f_{S_2}^{ls}(x),y\big)\big| \\
&\quad = \big|\big(f_{S_1}^{ls}(x)-y\big)^2 - \big(f_{S_2}^{ls}(x)-y\big)^2\big| \\
&\quad \le \big|f_{S_1}^{ls}(x) + f_{S_2}^{ls}(x) - 2y\big|\cdot\big|f_{S_1}^{ls}(x) - f_{S_2}^{ls}(x)\big| \le \frac{2M(\kappa+1)}{\sqrt{\lambda}}\cdot\big|f_{S_1}^{ls}(x) - f_{S_2}^{ls}(x)\big|
\end{aligned}$$

for any $S_1, S_2 \in Z^m$. So we set $C_L = \frac{2M(\kappa+1)}{\sqrt{\lambda}}$ in Proposition 1. Then b_{ls} has a density function $\frac{1}{\alpha}\exp\{-\frac{2|b|}{\alpha}\}$ with $\alpha = \frac{2M\kappa^2(\kappa+1)}{\lambda^{3/2}m\epsilon}$, which makes the algorithm provide ϵ-differential privacy.

A generalization analysis for this algorithm can also be found in [28]. What we shall mention here is that direct use of our error bound in the previous section leads to an unsatisfactory learning rate, since C_L tends to ∞ when $m\to\infty$. However, note that

$$\big(f_{\mathbf{z},\mathcal{A}}^{ls}(x_i)-y_i\big)^2 - \big(f_{ls}(x_i)-y_i\big)^2 = 2b\big(f_{\mathbf{z}}^{ls}(x_i)-y_i\big) + b^2$$

for any $i = 1,2,\ldots,m$, then

$$\mathbb{E}_{\mathbf{z},\mathcal{A}}\big(\mathcal{E}_{\mathbf{z}}\big(f_{\mathbf{z},\mathcal{A}}^{ls}\big) - \mathcal{E}_{\mathbf{z}}\big(f_{\mathbf{z}}^{ls}\big)\big) = \mathbb{E}_b b^2 = \frac{2M^2\kappa^4(\kappa+1)^2}{\lambda^3 m^2\epsilon^2}.$$

When $f_\rho^{ls} \in L_K^r(L^2_{\rho_X})$, let $f_\lambda = (L_K+\lambda I)^{-1}L_K f_\rho$, we have $B_\rho = 4M^2$, and (4) holds with $\beta = \min\{1,2r\}$ in Theorem 4 [29]. Then by choosing $\epsilon = 1/(\lambda m^{\frac{2}{3}})$ and $\lambda = (1/m)^{\frac{2}{3(\beta+1)}}$, we can derive an error bound in the form of

$$\mathbb{E}_{\mathbf{z},\mathcal{A}}\big(\mathcal{E}\big(f_{\mathbf{z},\mathcal{A}}^{ls}\big) - \mathcal{E}\big(f_\rho^{ls}\big)\big) \le \tilde{C}(1/m)^{\frac{2\beta}{3(\beta+1)}}$$

for some $\tilde{C}$ independent with m, from the total error bound in the last section. We omit the detail complex analysis here.

5.2 Differential private SVM

The second example is differential private SVM. We describe the SVM algorithm as in [19], *i.e.*, when $Y = \{-1, +1\}$,

$$f_{\mathbf{z}}^h = \arg\min_{f\in\mathcal{H}_K} \frac{1}{m}\sum_{i=1}^{m}\big(1-y_i f(x_i)\big)_+ + \lambda\|f\|_K^2,$$

and perturbation

$$f_{\mathbf{z},\mathcal{A}}^h = f_{\mathbf{z}}^h + b_h,$$

where the hinge loss $L_h(f(x),y) = (1-yf(x))_+ = \max\{0, 1-yf(x)\}$ is used in the ERM setting. Then the output classifier is $\operatorname{sgn}(f_{\mathbf{z},\mathcal{A}}^h)$.

Firstly we consider the differential privacy of this algorithm. Note that $|a_+ - b_+| \le |a-b|$ for any $a, b \in \mathbb{R}$, by the discussion, we have

$$\big|L\big(f_1(x),y\big) - L\big(f_2(x),y\big)\big| = \big|\big(1-yf_1(x)\big)_+ - \big(1-yf_2(x)\big)_+\big| \le \big|f_1(x)-f_2(x)\big|.$$

Then $C_L = 1$ and $\gamma = 2$ in Proposition 1. Therefore b_h here has a density function $1/\alpha \exp\{-\frac{2|b|}{\alpha}\}$ with $\alpha = \frac{\kappa^2}{\lambda m\epsilon}$. In this case, we have, for any possible output set $\mathcal{O}$,

$$\Pr\big\{f_{\mathbf{z},\mathcal{A}}^h \in \mathcal{O}\big\} \le e^{\epsilon}\Pr\big\{f_{\bar{\mathbf{z}},\mathcal{A}}^h \in \mathcal{O}\big\},$$

where $\bar{\mathbf{z}}$ differs from $\mathbf{z}$ in one element. Then, for any possible classifier g defined on X,

$$\Pr_{\mathcal{A}}\big\{\operatorname{sgn}\big(f_{\mathbf{z},\mathcal{A}}^h\big) = g\big\} \le e^{\epsilon}\Pr_{\mathcal{A}}\big\{\operatorname{sgn}\big(f_{\bar{\mathbf{z}},\mathcal{A}}^h\big) = g\big\}.$$

This verifies the ϵ-differential privacy of the algorithm.

Now let us turn to the error analysis. When hinge loss is applied in the ERM setting, Theorem 14 of [30] reveals the comparison theorem, that is, denote $R(f) = \Pr(y \ne f(x)) = \int_X \Pr(y \ne f(x)|x)\, d\rho_X$, then

$$R(f) - R(f_c) \le \sqrt{2\big(\mathcal{E}(f) - \mathcal{E}\big(f_\rho^h\big)\big)}$$

for any measurable function f. Here

$$f_\rho^h = \arg\min_f \int_Z \big(1-yf(x)\big)_+ d\rho.$$

Assume $f_\rho^h \in L_K^r(L^2_{\rho_X})$, for some $r > 0$ and f_c is the Bayes classifier, *i.e.*,

$$f_c(x) = \begin{cases} 1, & \Pr(y=1|x) \ge \Pr(y=-1|x), \\ -1, & \Pr(y=1|x) < \Pr(y=-1|x). \end{cases}$$

Then

$$\mathbb{E}_{\mathbf{z},\mathcal{A}}\big(R\big(f_{\mathbf{z},\mathcal{A}}^h\big) - R(f_c)\big) \le \sqrt{2\mathbb{E}_{\mathbf{z},\mathcal{A}}\big(\mathcal{E}\big(f_{\mathbf{z},\mathcal{A}}^h\big) - \mathcal{E}\big(f_\rho^h\big)\big)}.$$

Still we choose stepping-stone function $f_\lambda = (L_K + \lambda I)^{-1} L_K f_\rho^h$, which leads to $\|f_\lambda\|_\infty \le M$ and $B_\rho = (M+1)^2$. Reference [31] shows that $D(\lambda) \le \lambda^{\min\{r,1\}}$, so we can follow the choice for ϵ and λ in Theorem 4 with $\beta = \min\{r,1\}$ to get the learning rate as

$$\mathbb{E}_{\mathbf{z},\mathcal{A}}\left(R\left(f_{\mathbf{z},\mathcal{A}}^h\right) - R(f_c)\right) \le \tilde{C}\left(\frac{1}{m}\right)^{\frac{\beta}{2(2\beta+1)}},$$

where $\tilde{C}$ is a constant independent of m.

6 Results and conclusions

In this paper, we present two results in the analysis of the differential private convex risk minimization algorithms.

The first one is the perturbation results for general convex risk minimization algorithms. We studied two cases of the general algorithms. The second one is applied in the following analysis, as it leads to a sharper upper bound of the error between two outputs differ in 1 sample point. However, the first one is more relaxed, without Lipschitz continuity of the loss function. Based on such perturbation results we obtain a choice for the random terms of the differential private algorithms, *i.e.*, Proposition 1. This gives us a theoretical and practical construction of differential private algorithms.

An error analysis is the second contribution of this paper. The analysis relies on the concentration inequality in the setting of differential privacy. After conducting a different error decomposition using the above concentration inequality, we provide an upper bound or learning rate of the expected generalization error. In this result we find a selection of the parameter of differential privacy ϵ and the regularization parameter λ, both of which depend on the sample size m. Since smaller ϵ always means more effective privacy protection, this indicates that generalized algorithms must not be too much privacy protected.

In [8], the authors proposed that the learning rate can be $\frac{1}{2}$ under the strong assumption on the loss function and with regularization term $\frac{1}{2}\|f\|_K^2$. However, the differential private parameter ϵ is fixed there. In this paper we obtain a learning rate $\frac{1}{3}$ with weak conditions on the loss function and $r \ge \frac{1}{2}$ when choosing appropriate parameters ϵ and λ. As we pointed out above, ϵ should not be too small to derive convergent algorithms. In fact, for a fixed ϵ, we as well can deduce a learning rate of $\frac{1}{2}$ (with a slightly different form); see [28] for a detailed analysis.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
All authors contributed equally to the writing of this paper. All authors read and approved the final manuscript.

Acknowledgements
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). The authors would like to thank the associated editors and anonymous referees for their valuable comments and suggestions, which have helped to improve the paper.

Received: 6 October 2016 Accepted: 23 December 2016 Published online: 05 January 2017

References

1. Vapnik, V: Statistical Learning Theory. Wiley, New York (1998)
2. Dwork, C: Differential privacy. In: ICALP, pp. 1-12. Springer, Berlin (2006)
3. Dwork, C, Rothblum, GN: Concentrated differential privacy. arXiv:1603.01887
4. Dwork, C, McSherry, F, Nissim, K, Smith, A: Calibrating noise to sensitivity in private data analysis. In: Theory of Cryptography, pp. 265-284. Springer, Berlin (2006)
5. McSherry, F, Talwar, K: Mechanism design via differential privacy. In: Proceedings of the 48th Annual Symposium on Foundations of Computer Science, pp. 94-103 (2007)
6. Dwork, C: Differential privacy: a survey of results. In: Theory and Applications of Models of Computation. Lecture Notes in Computer Science, vol. 4978, pp. 1-19 (2008)
7. Ji, ZL, Lipton, ZC, Elkan, C: Differential privacy and machine learning: a survey and review (2014). arXiv:1412.7584
8. Chaudhuri, K, Monteleoni, C, Sarwate, AD: Differentially private empirical risk minimization. J. Mach. Learn. Res. **12**, 1069-1109 (2011)
9. Kifer, D, Smith, A, Thakurta, A: Private convex empirical risk minimization and high-dimensional regression. In: Conference on Learning Theory, pp. 25.1-25.40
10. Jain, P, Thakurta, AG: Differentially private learning with kernels. In: ICML (2013)
11. Jain, P, Thakurta, AG: (Near) dimension independent risk bounds for differentially private learning. In: ICML (2014)
12. Bassily, R, Smith, A, Thakurta, A: Differential private empirical risk minimization: efficient algorithms and tight error bounds. In: FOCS. IEEE (2014)
13. Bassily, R, Nissim, K, Smith, A, Steinke, T, Stemmer, U, Ullman, J: Algorithmic stability for adaptive data analysis (2015). arXiv:1511.02513
14. Steinwart, I, Christmann, A: Estimating conditional quantiles with the help of the pinball loss. Bernoulli **17**(1), 211-225 (2008)
15. Bousquet, O, Elisseeff, A: Stability and generalization. J. Mach. Learn. Res. **2**, 499-526 (2002)
16. Shalev-Shwartz, S, Shamir, O, Screbro, N, Scridharan, K: Learnability, stability and uniform convergence. J. Mach. Learn. Res. **11**, 2635-2670 (2010)
17. Wang, Y-X, Lei, J, Fienberg, SE: Learning with differential privacy: stability, learnability and the sufficiency and necessity of ERM principle. arXiv:1502.06309
18. Cucker, F, Smale, S: On the mathematical foundations of learning. Bull. Am. Math. Soc. **39**, 1-49 (2002)
19. Cucker, F, Zhou, DX: Learning Theory: An Approximation Theory Viewpoint. Cambridge University Press, Cambridge (2007)
20. Sridharan, K, Srebro, N, Shalev-Shwartz, S: Fast rates for regularized objectives. In: Advances in Neural Information Processing Systems 22, pp. 1545-1552 (2008)
21. Wu, Q, Zhou, DX: Learning with sample dependent hypothesis space. Comput. Math. Appl. **56**, 2896-2907 (2008)
22. Rubinstein, BIP, Bartlett, PL, Huang, L, Taft, N: Learning in a large function space: privacy-preserving mechanisms for SVM learning. J. Priv. Confid. **4**(1), 65-100 (2012)
23. Dwork, C, Feldman, V, Hardt, M, Pitassi, T, Reingold, O, Roth, A: Preserving statistical validity in adaptive data analysis. In: ACM Symposium on the Theory of Computing (STOC). ACM (2015)
24. Hoeffding, W: Probability inequalities for sums of bounded random variables. J. Am. Stat. Assoc. **58**(301), 13-30 (1963)
25. Wang, C, Zhou, DX: Optimal learning rates for least squares regularized regression with unbounded sampling. J. Complex. **27**, 55-67 (2011)
26. Shi, L: Learning theory estimates for coefficient-based regularized regression. Appl. Comput. Harmon. Anal. **34**, 252-265 (2013)
27. Xiang, DH: Conditional quantiles with varying Gaussians. Adv. Comput. Math. **38**, 723-735 (2013)
28. Nie, WL, Wang, C: Error analysis and variable selection for differential private learning algorithm. Preprint (2016)
29. Smale, S, Zhou, DX: Learning theory estimates via integral operators and their applications. Constr. Approx. **26**, 153-172 (2007)
30. Chen, DR, Wu, Q, Ying, Y, Zhou, DX: Support vector machine soft margin classifiers: error analysis. J. Mach. Learn. Res. **5**, 1143-1175 (2004)
31. Xiang, DH, Hu, T, Zhou, DX: Approximation analysis of learning algorithms for support vector regression and quantile regression. J. Appl. Math. **2012**, Article ID 902139 (2012). doi:10.1155/2012/902139