1 Appendix

Proposition 3.1 If the loss function ϕ is L-Lipschitz and all data points lie in a unit ball, then the L_2 sensitivity of the algorithm that computes $\nabla G_k(\mathbf{w}_t)$ from D_k is at most $2L$.

Proof. Let D_k, D'_k be any two data sets differing in a single element. Without loss of generality, we can assume that D_k and D'_k differ only in their last element, with $D_k = \{(\mathbf{x}_j^k, y_j^k)\}_{j=1}^{N_k}$ and $D'_k = \{(\mathbf{x}_1^k, y_1^k), \ldots, (\mathbf{x}_{N_k-1}^k, y_{N_k-1}^k), (\mathbf{x}_{N_k}'^k, y_{N_k}'^k)\}$. Let $\nabla G'_k(\mathbf{w}_t)$ denote the gradient at $\mathbf{w}_t$ of the cumulative loss on D'_k.

By definition, we have for datasets D_k and D'_k,

$$\nabla G_k(\mathbf{w}_t) = \sum_{j=1}^{N_k} \phi'(y_j^k \mathbf{w}_t^\top \mathbf{x}_j^k)(y_j^k \mathbf{x}_j^k)$$

and

$$\nabla G'_k(\mathbf{w}_t) = \sum_{j=1}^{N_k-1} \phi'(y_j^k \mathbf{w}_t^\top \mathbf{x}_j^k)(y_j^k \mathbf{x}_j^k) + \phi'(y_j'^k \mathbf{w}_t^\top \mathbf{x}_j'^k)(y_j'^k \mathbf{x}_j'^k)$$

Then we have

$$\begin{aligned}
&\|\nabla G_k(\mathbf{w}_t) - \nabla G'_k(\mathbf{w}_t)\|_2 \\
&= \left\|\phi'(y_{N_k}^k \mathbf{w}_t^\top \mathbf{x}_{N_k}^k)(y_{N_k}^k \mathbf{x}_{N_k}^k) - \phi'(y_{N_k}'^k \mathbf{w}_t^\top \mathbf{x}_{N_k}'^k)(y_{N_k}'^k \mathbf{x}_{N_k}'^k)\right\|_2 \\
&\leq \left|\phi'(y_{N_k}^k \mathbf{w}_t^\top \mathbf{x}_{N_k}^k)\right| \left\|(y_{N_k}^k \mathbf{x}_{N_k}^k)\right\|_2 + \left|\phi'(y_{N_k}'^k \mathbf{w}_t^\top \mathbf{x}_{N_k}'^k)\right| \left\|(y_{N_k}'^k \mathbf{x}_{N_k}'^k)\right\|_2 \\
&\leq 2L\,,
\end{aligned}$$

where the first inequality follows from triangle inequality and the last inequality follows from the assumptions that ϕ is L-Lipschitz (which gives $|\phi'(z)| \leq L\ \forall z$), $y_j^k \in \{\pm 1\}$, and $\|\mathbf{x}_j^k\|_2 \leq 1$. □ □

Theorem 4.1 If ϕ is convex and doubly differentiable with $\phi'(z) \leq 1$ and $\phi''(z) \leq c\ \forall z$, then Algorithm 4 is (ϵ, δ)-differentially private.

Proof. The proof is similar in overall structure to the proof of Theorem 9 of Chaudhuri et al. [4]; We provide the details here for completeness.

To prove (ϵ, δ) differential privacy, we need the ratio $\frac{\mathbf{Pr}(\mathbf{w}_{\text{gop}}|D)}{\mathbf{Pr}(\mathbf{w}_{\text{gop}}|D')}$ of the densities of $\mathbf{w}_{\text{gop}}$ under two neighboring data sets D and D'. As the objective function that Algorithm 4 minimizes is convex, it has a unique minimizer and hence one can show that there exists a bijection between the noise that is added to the objective and the output of the Algorithm. This allows us to write the ratios of the densities of getting any fixed vector as output for adjacent data sets in terms of the ratios of the noises added to achieve that particular output vector. Specifically, if $\mathbf{w}_{\text{gop}}$ is the output of the algorithm, the following holds:

$$\frac{\mathbf{Pr}(\mathbf{w}_{\text{gop}}|D)}{\mathbf{Pr}(\mathbf{w}_{\text{gop}}|D')} = \frac{\mathbf{Pr}(\boldsymbol{\eta}_D|D)}{\mathbf{Pr}(\boldsymbol{\eta}_{D'}|D')} \cdot \frac{|det(\mathbf{J}(\mathbf{w}_{\text{gop}} \to \boldsymbol{\eta}_D|D))|^{-1}}{|det(\mathbf{J}(\mathbf{w}_{\text{gop}} \to \boldsymbol{\eta}_{D'}|D'))|^{-1}}$$

where $\mathbf{J}(\mathbf{w}_{\text{gop}} \to \boldsymbol{\eta}_D|D)$ represents the Jacobian matrix of the mapping from $\mathbf{w}_{\text{gop}}$ to $\boldsymbol{\eta}_D$ whose (j,k)-th entry is given by

$$\mathbf{J}(\mathbf{w}_{\text{gop}} \to \boldsymbol{\eta}_D|D)_{jk} = \frac{\partial \eta_D^{(j)}}{\partial w_{\text{gop}}^{(k)}}$$

We will first bound the ratio of the determinants.

Note that the mapping from $\mathbf{w}_{\text{gop}}$ to $\boldsymbol{\eta}_D$ is got by setting the derivative of the perturbed objective function involving the dataset D to zero. Thus we have for dataset D, the following mapping.

$$\boldsymbol{\eta}_D = -N(\lambda + \Delta)\mathbf{w}_{\text{gop}} + \sum_{i=1}^{N} \phi'(y_i \mathbf{w}_{\text{gop}}^\top \mathbf{x}_i)(y_i \mathbf{x}_i)\,.$$

Under this mapping, the (j,k)-th entry of the Jacobian $\mathbf{J}(\mathbf{w}_{\text{gop}} \to \boldsymbol{\eta}_D|D)$ is given by

$$\mathbf{J}(\mathbf{w}_{\text{gop}} \to \boldsymbol{\eta}_D|D)_{jk} = -N(\lambda + \Delta)\mathbf{I}(j = k) - \sum_{i=1}^{N} y_i^2 \phi''(y_i \mathbf{w}_{\text{gop}}^\top \mathbf{x}_i) x_i^{(j)} x_i^{(k)}$$

Define

$$\mathbf{A} := N(\lambda + \Delta)\mathbf{I}_d + \sum_{i=1}^{N} y_i^2 \phi''(y_i \mathbf{w}_{\text{gop}}^\top \mathbf{x}_i)\mathbf{x}_i \mathbf{x}_i^\top$$

and

$$\mathbf{E} := y_N'^2 \phi''(y'_N \mathbf{w}_{\text{gop}}^\top \mathbf{x}'_N)\mathbf{x}'_N \mathbf{x}'^\top_N - y_N^2 \phi''(y_N \mathbf{w}_{\text{gop}}^\top \mathbf{x}_N)\mathbf{x}_N \mathbf{x}_N^\top$$

Thus we have, the ratio of the determinants of the Jacobians equal to

$$\frac{|det(\mathbf{J}(\mathbf{w}_{\text{gop}} \to \boldsymbol{\eta}_D|D))|^{-1}}{|det(\mathbf{J}(\mathbf{w}_{\text{gop}} \to \boldsymbol{\eta}_{D'}|D'))|^{-1}} = \frac{|det(\mathbf{A} + \mathbf{E})|}{|det(\mathbf{A})|}$$

As the matrix $\mathbf{E}$ is of rank atmost 2, by Lemma 2 of [4], this determinant ratio is equal to

$$\frac{|det(\mathbf{A} + \mathbf{E})|}{|det(\mathbf{A})|} = |1 + \theta_1(\mathbf{A}^{-1}\mathbf{E}) + \theta_2(\mathbf{A}^{-1}\mathbf{E}) + \theta_1(\mathbf{A}^{-1}\mathbf{E})\theta_2(\mathbf{A}^{-1}\mathbf{E})|$$

where θ_1 and θ_2 are the largest and second largest eigenvalues of the matrix $\mathbf{A}^{-1}\mathbf{E}$ in absolute value.

As ϕ is doubly differentiable and convex, the second term in the definition of $\mathbf{A}$ is a positive semi definite matrix. Thus it follows that the eigenvalues of $\mathbf{A}$ are greater than $N(\lambda+\Delta)$ which is the eigenvalue of the first term in the definition of $\mathbf{A}$. Hence we have

$$|\theta_l(\mathbf{A}^{-1}\mathbf{E})| \leq |\frac{\theta_l(\mathbf{E})}{N(\lambda+\Delta)}| \;\; l=1,2$$

Now applying Lemma 3 of [4] gives us the following relations for the eigenvalues of the matrix $\mathbf{E}$

$$|\theta_1(\mathbf{E})| + |\theta_2(\mathbf{E})| \leq 2c$$

and

$$|\theta_1(\mathbf{E})|.|\theta_2(\mathbf{E})| \leq c^2$$

Substituting these in the bound above, we get

$$\frac{|det(\mathbf{A}+\mathbf{E})|}{|det(\mathbf{A})|} \leq \left(1+\frac{c}{N(\lambda+\Delta)}\right)^2$$

It is easy to verify that the choice of Δ in Algorithm 4 implies that the above quantity is bounded by $e^{(\epsilon-\tilde{\epsilon})}$. Thus we have established the following:

$$\frac{|det(\mathbf{J}(\mathbf{w}_{\text{gop}} \to \boldsymbol{\eta}_D|D))|^{-1}}{|det(\mathbf{J}(\mathbf{w}_{\text{gop}} \to \boldsymbol{\eta}_{D'}|D'))|^{-1}} \leq e^{(\epsilon-\tilde{\epsilon})}$$

Now if we can choose a variance for the Gaussian noise such that the ratio $\frac{\mathbf{Pr}(\boldsymbol{\eta}_D|D)}{\mathbf{Pr}(\boldsymbol{\eta}_{D'}|D')}$ is bounded by $e^{\tilde{\epsilon}}$, then we have ϵ-differential privacy overall. Unfortunately, unlike the noise considered in [4], such a bound is not possible with the Gaussian noise we use. Instead, for any given data set D, we will show that the output space $\mathbb{R}^d$ can be divided into two sets Ω_1, Ω_2 such that $\mathbf{Pr}(\mathcal{A}(D) \in \Omega_1) \leq \delta$, and for all data sets D' that differ from D in one element, we get $\frac{\mathbf{Pr}(\mathcal{A}(D)=\mathbf{w})}{\mathbf{Pr}(\mathcal{A}(D')=\mathbf{w})} \leq e^{\epsilon}$ for all $\mathbf{w} \in \Omega_2$; this will allow us to show (ϵ, δ) differential privacy for the Gaussian objective perturbation algorithm.

To prove a (high probability) bound on the ratio of densities, we proceed in the usual way. Let the noise that is added to data sets D and D' be $\boldsymbol{\eta}_D$ and $\boldsymbol{\eta}_{D'}$. For now we will assume each component of the noise vectors is chosen independently from $\mathcal{N}(0,\sigma^2)$ without committing to the exact value of σ; this will be selected later. Then we need to show

$$\frac{\mathbf{Pr}(\boldsymbol{\eta}_D)}{\mathbf{Pr}(\boldsymbol{\eta}_{D'})} \leq e^{\tilde{\epsilon}}\,.$$

Using the fact that the noise vectors are Gaussian, we will consider when the following inequality holds:

$$\exp\left(-\frac{1}{2\sigma^2}(\|\boldsymbol{\eta}_D\|^2 - \|\boldsymbol{\eta}_{D'}\|^2)\right) \leq e^{\tilde{\epsilon}}\,.$$

Equivalently we can consider when the following holds:

$$\|\boldsymbol{\eta}_{D'}\|^2 - \|\boldsymbol{\eta}_D\|^2 \leq 2\sigma^2\tilde{\epsilon}\,.$$

Consider the LHS of the above expression. This is equal to

$$\|\boldsymbol{\eta}_{D'} + \boldsymbol{\eta}_D - \boldsymbol{\eta}_D\|^2 - \|\boldsymbol{\eta}_D\|^2$$

$$= \|\boldsymbol{\eta}_{D'} - \boldsymbol{\eta}_D\|^2 + \|\boldsymbol{\eta}_D\|^2 + 2(\boldsymbol{\eta}_{D'} - \boldsymbol{\eta}_D)^\top \boldsymbol{\eta}_D - \|\boldsymbol{\eta}_D\|^2$$

which by Cauchy-Schwartz becomes

$$\leq \|\boldsymbol{\eta}_{D'} - \boldsymbol{\eta}_D\|^2 + 2\|\boldsymbol{\eta}_{D'} - \boldsymbol{\eta}_D\|\,\|\boldsymbol{\eta}_D\|$$

To obtain $\mathbf{w}_{\text{gop}}$ as output for both D and D', we know from before that we must have

$$\boldsymbol{\eta}_D = -N(\lambda+\Delta)\mathbf{w}_{\text{gop}} - \sum_{i=1}^{N} \phi'(y_i \mathbf{w}_{\text{gop}}^\top \mathbf{x}_i)(y_i\mathbf{x}_i)$$

and

$$\begin{aligned}\boldsymbol{\eta}_{D'} = &-N(\lambda+\Delta)\mathbf{w}_{\text{gop}} \\ &- \sum_{i=1}^{N-1} \phi'(y_i \mathbf{w}_{\text{gop}}^\top \mathbf{x}_i)(y_i\mathbf{x}_i) \\ &-\phi'(y'_N \mathbf{w}^\top \mathbf{x}'_N)(y'_N\mathbf{x}'_N)\end{aligned}$$

Now by using the fact that $|\phi'(z)| \leq 1\;\forall z$ and $\|\mathbf{x}_i\| \leq 1\;\forall i$, we have

$$\|\boldsymbol{\eta}_{D'} - \boldsymbol{\eta}_D\| \leq 2\,.$$

After applying this bound to the previous step and a few steps of algebra, we find that we need to consider when the following holds:

$$\|\boldsymbol{\eta}_D\| \leq \frac{\sigma^2\tilde{\epsilon} - 2}{2} \tag{1}$$

The above equation cannot be satisfied for all values that the random variable $\|\eta_D\|$ can take. However, we can always choose a variance σ such that

$$\mathbf{Pr}\left(\|\boldsymbol{\eta}_D\| \leq \frac{\sigma^2\tilde{\epsilon} - 2}{2}\right) \geq 1-\delta\,.$$

In particular, let $\sigma^* = \sigma^*(d,\tilde{\epsilon},\delta)$ be the value that satisfies the above relation with equality. Note that this corresponds to choosing σ^* such that

$$\mathbf{Pr}\left(U \leq \left(\frac{\sigma^2\tilde{\epsilon} - 2}{2\sigma}\right)^2\right) = 1-\delta\,.$$

where U is a χ^2 random variable with d degrees of freedom. Now recall that the noise $\boldsymbol{\eta}_D$ that must be added to the objective with data set D to get $\mathbf{w}$ as the output satisfies the following

$$\boldsymbol{\eta}_D(\mathbf{w}) = -N(\lambda+\Delta)\mathbf{w}_{\text{gop}} + \sum_{i=1}^{N} \phi'(y_i \mathbf{w}_{\text{gop}}^\top \mathbf{x}_i)(y_i\mathbf{x}_i)\,.$$

For a given a data set D, consider Ω_1, Ω_2 defined as below:

$$\Omega_1 = \left\{ \mathbf{w} \in \mathbb{R}^d \;\middle|\; \|\boldsymbol{\eta}_D(\mathbf{w})\| > \frac{\tilde{\epsilon}\sigma^{*2} - 2}{2} \right\}$$

$$\Omega_2 = \mathbb{R}^d \setminus \Omega_1 \,.$$

Clearly, if $\mathcal{A}(D) \in \Omega_2$, then the corresponding Gaussian noise $\boldsymbol{\eta}_D$ that was generated to perturb the objective satisfies the property $\|\boldsymbol{\eta}_D\| \leq \frac{\tilde{\epsilon}\sigma^{*2}-2}{2}$. By choice of σ^*, this happens with probability $1-\delta$ as required. Putting these arguments together with the bound on the Jacobian proves that Algorithm 4 is (ϵ, δ)-differentially private. □

Theorem 4.2 If on each iteration t the third party receives the *sum* of noisy gradients from all the parties in a cryptographically secure manner, then Algorithm 1 (with σ^* chosen as described in Procedure 3) is (ϵ, δ)-differentially private.

Proof. Note that the noise vector $\boldsymbol{\eta}^k$ generated once by each party P_k is sampled according to a multivariate Gaussian, each component of which is drawn according to $\mathcal{N}(0, \frac{\sigma^{*2}}{K})$. Thus the overall noise added to the objective minimized by the third party is also a multivariate Gaussian, with each component drawn from $\mathcal{N}(0, \sigma^{*2})$. As Algorithm 1 can be viewed as Gaussian objective perturbation,by Theorem 4.1 , it follows directly that Algorithm 1 is (ϵ, δ)-differentially private. □

Theorem 5.1 If ϕ is such that $\phi'(z) \leq 1 \, \forall z$, $\phi''(z) \leq c \, \forall z$, all the data instances $\mathbf{x}_i$ lie in a unit ball and $\Delta = 0$, then with probability at least $1 - \delta'$ (over the draw of $D \sim \mathcal{Q}^N$ and randomization in the algorithm), the excess empirical regularized risk of the perturbed classifier $\mathbf{w}_{\text{gop}}$ learned by Algorithm 4 over the minimizer $\mathbf{w}^*$ of $J(\mathbf{w})$ is bounded as

$$J(\mathbf{w}_{\text{gop}}) \leq J(\mathbf{w}^*) + \frac{(c+\lambda)}{2N^2\lambda^2}\widehat{T}$$

where $\widehat{T} := \widehat{T}(d, \tilde{\epsilon}, \delta, \delta')$ satisfies the equation

$$\mathbf{Pr}\left(U \leq \frac{\widehat{T}}{\sigma^{*2}} \right) = 1 - \delta'$$

where U is a χ^2 random variable with d degrees of freedom and $\sigma^* = \sigma^*(d, \tilde{\epsilon}, \delta)$ is as chosen in Procedure [3].

Proof. By Taylor series expansion of J, we have

$$J(\mathbf{w}_{\text{gop}}) = J(\mathbf{w}^*) + (\mathbf{w}_{\text{gop}} - \mathbf{w}^*)^\top \nabla J(\mathbf{w}^*) + \frac{1}{2}(\mathbf{w}_{\text{gop}} - \mathbf{w}^*)^T \nabla^2 \left(J(\mathbf{w})\right) (\mathbf{w}_{\text{gop}} - \mathbf{w}^*)$$

for some $\mathbf{w} \in \mathbb{R}^d$. By definition $\nabla J(\mathbf{w}^*) = 0$.

By Cauchy-Schwartz, we have

$$|J(\mathbf{w}_{\text{gop}}) - J(\mathbf{w}^*)| \leq \frac{1}{2}\|\mathbf{w}_{\text{gop}} - \mathbf{w}^*\|^2 \|\nabla^2 J(\mathbf{w})\| \quad (2)$$

where the norm with respect the $\nabla^2 J(\mathbf{w})$ is the matrix L_2 norm. But

$$\nabla^2 J(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N} \phi''(y_i \mathbf{w}^\top \mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top + \lambda \mathbf{I}_d$$

which implies

$$\|\nabla^2 J(\mathbf{w})\| = \|\frac{1}{N}\sum_{i=1}^{N} \phi''(y_i \mathbf{w}^\top \mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top + \lambda \mathbf{I}_d\|$$

Applying triangle inequality for matrix norms, we have

$$\|\nabla^2 J(\mathbf{w})\| \leq \|\frac{1}{N}\sum_{i=1}^{N} \phi''(y_i \mathbf{w}^\top \mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top\| + \|\lambda \mathbf{I}_d\|$$

Since $|\phi''(z)| \leq c$, we have,

$$\|\nabla^2 J(\mathbf{w})\| \leq c\|\frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^\top\| + \lambda\|\mathbf{I}_d\|$$

Again applying triangle inequality, we get

$$\|\nabla^2 J(\mathbf{w})\| \leq c\frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i\mathbf{x}_i^\top\| + \lambda\|\mathbf{I}_d\|$$

Since $\mathbf{x}_i$'s lie in the unit ball, we have that $\|\mathbf{x}_i\mathbf{x}_i^\top\| \leq 1$. Also $\|\mathbf{I}_d\| = 1$. Thus we get,

$$\|\nabla^2 J(\mathbf{w})\| \leq c + \lambda$$

Substituting this in Equation 2 gives

$$J(\mathbf{w}_{\text{gop}}) \leq J(\mathbf{w}^*) + \frac{(c+\lambda)}{2}\|\mathbf{w}_{\text{gop}} - \mathbf{w}^*\|^2$$

Using Lemma 7 of [4], we can show that for objective perturbation, if $\Delta = 0$, then we have

$$\|\mathbf{w}_{\text{gop}} - \mathbf{w}^*\|^2 \leq \frac{\|\boldsymbol{\eta}\|^2}{N^2\lambda^2}$$

Thus we have

$$J(\mathbf{w}_{\text{gop}}) \leq J(\mathbf{w}^*) + \frac{(c+\lambda)}{2}\frac{\|\boldsymbol{\eta}\|^2}{N^2\lambda^2}$$

We now can bound $\|\boldsymbol{\eta}\|^2$ using $\frac{\|\boldsymbol{\eta}\|^2}{\sigma^{*2}}$ which is χ^2 distributed by equating the cumulative distribution function to $(1-\delta')$ to get the statement of the theorem. □

Proposition 5.2 The following bound holds for parameters as in Theorem [5.1] and Algorithm 4

$$\widehat{T} = O(\frac{d^2 \log(\frac{1}{\delta}) \log(\frac{1}{\delta'})}{\tilde{\epsilon}^2}) \ \forall \ \delta, \delta' \leq \frac{1}{e}$$

Proof. Let U be a χ^2 distributed random variable with d degrees of freedom. Then the value of $\widehat{T} = \widehat{T}(d, \tilde{\epsilon}, \delta, \delta')$ is such that it satisfies the following equation.

$$\mathbf{Pr}(U \leq \frac{T}{\sigma^{*2}}) = 1 - \delta' \tag{3}$$

where $\sigma^* = \sigma^*(d, \tilde{\epsilon}, \delta)$ is chosen as in Procedure [3].

To see how the value of $\widehat{T}$ grows with the parameters it depends on, we use the following exponential tail bound for a χ^2 random variable from corollary of Lemma 1 in [9]:

If U is χ^2 distributed with D degrees of freedom, for any positive x,

$$\mathbf{Pr}(U \geq D + 2\sqrt{Dx} + 2x) \leq e^{-x} \tag{4}$$

Substituting $x = \log(\frac{1}{\delta'})$ and $D = d$ in the above equation yields us

$$\mathbf{Pr}(U \geq d + 2\sqrt{d \log(\frac{1}{\delta'})} + 2\log(\frac{1}{\delta'})) \leq \delta'$$

Comparing the above two tail inequalities for U, we obtain that

$$\widehat{T} \leq \sigma^{*2}(d + 2\sqrt{d \log(\frac{1}{\delta'})} + 2\log(\frac{1}{\delta'})) \tag{5}$$

As mentioned before, σ^* is chosen such that the following holds

$$\mathbf{Pr}(U \geq \frac{(\sigma^{*2}\tilde{\epsilon} - 2)^2}{4\sigma^{*2}}) = \delta$$

Using the same tail bound as before, but substituting $x = \log(\frac{1}{\delta})$, we have

$$\mathbf{Pr}(U \geq d + 2\sqrt{d \log(\frac{1}{\delta})} + 2\log(\frac{1}{\delta})) \leq \delta$$

It then follows by comparing the above two equations that

$$\frac{(\sigma^{*2}\tilde{\epsilon} - 2)^2}{4\sigma^{*2}} \leq d + 2\sqrt{d \log(\frac{1}{\delta})} + 2\log(\frac{1}{\delta}))$$

Expanding the LHS gives

$$\frac{\sigma^{*2}\tilde{\epsilon}^2}{4} + \frac{1}{\sigma^{*2}} - \tilde{\epsilon} \leq d + 2\sqrt{d \log(\frac{1}{\delta})} + 2\log(\frac{1}{\delta}))$$

Dropping the second term from the bound, we get

$$\sigma^{*2} \leq \frac{4}{\tilde{\epsilon}^2} R_\delta + \frac{4}{\tilde{\epsilon}}$$

where R_θ is defined as $(d + 2\sqrt{d \log(\frac{1}{\theta})} + 2\log(\frac{1}{\theta}))$ for $0 \leq \theta \leq 1$

As $\frac{1}{\tilde{\epsilon}^2}$ would eventually dominate the above bound for σ^{*2}, it implies that there exists some constant C such that

$$\sigma^{*2} \leq C \frac{1}{\tilde{\epsilon}^2} R_\delta \tag{6}$$

Substituting the bound got here in Equation 5, we get

$$\widehat{T} \leq \frac{C}{\tilde{\epsilon}^2} R_\delta R_{\delta'} \tag{7}$$

Now consider the quantity R_θ for some θ

$$R_\theta = d + 2\sqrt{d \log(\frac{1}{\theta})} + 2\log(\frac{1}{\theta}))$$

$$R_\theta \leq d + 2\sqrt{d \log(\frac{1}{\theta})} + 2d\log(\frac{1}{\theta}))$$

Now if $\theta \leq \frac{1}{e}$, we have $\log(\frac{1}{\theta}) \geq 1$. Thus in this case, we have

$$R_\theta \leq d\log(\frac{1}{\theta}) + 2\sqrt{d \log(\frac{1}{\theta})} + 2d\log(\frac{1}{\theta}))$$

The above can now be written as

$$R_\theta \leq C' d \log(\frac{1}{\theta})$$

for some C'

Substituting this in Equation 7 for δ and δ', it follows that

$$\widehat{T} \leq \frac{C''}{\tilde{\epsilon}^2} d^2 \log(\frac{1}{\delta}) \log(\frac{1}{\delta'}) \tag{8}$$

for some C''.

□

Theorem 5.3 Let ϕ be convex and doubly differentiable with $\phi'(z) \leq 1$ and $\phi''(z) \leq c \ \forall z$. Then there exists a constant κ such that for any fixed weight vector $\mathbf{w}_0$, any $\delta' > 0$, if

$$N > \kappa \max\left(\frac{\|\mathbf{w}_0\|^2 \log(\frac{1}{\delta'})}{\tau^2}, \frac{c\|\mathbf{w}_0\|^2}{\tau\epsilon}, \frac{\|\mathbf{w}_0\|\widehat{T}^{\frac{1}{2}}}{\tau}\right)$$

where $\widehat{T} = \widehat{T}(d, \tilde{\epsilon}, \delta, \delta')$ is as in Theorem [5.1], then with probability at least $1 - 2\delta'$ (over $D \sim \mathcal{Q}^N$ and randomization in the algorithm), the output $\mathbf{w}_{\text{gop}}$ of Algorithm 4 satisfies

$$\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{Q}}\left[\phi(y\mathbf{w}_{\text{gop}}^\top\mathbf{x})\right] - \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{Q}}\left[\phi(y\mathbf{w}_0^\top\mathbf{x})\right] \leq \tau\,.$$

Proof. The proof of this theorem is similar to the proof of Theorem 18 in [4].

We provide the proof here for completeness.

Let $\mathbf{w}_{\text{e}}$ be the minimizer of the regularized expected risk $J_{\text{e}}(\mathbf{w})$ and $\mathbf{w}^*$ be the minimizer of the regularized empirical error objective $J(\mathbf{w})$ for the dataset D:

$$\mathbf{w}_{\text{e}} = \operatorname*{argmin}_{\mathbf{w}} J_{\text{e}}(\mathbf{w})$$

where

$$J_{\text{e}}(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{Q}}\left[\phi(y\mathbf{w}^\top\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{w}\|^2\right]$$

$$\mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} J(\mathbf{w}) := \operatorname*{argmin}_{\mathbf{w}}\left[\sum_{i=1}^{N}\phi(y_i\mathbf{w}^\top\mathbf{x}_i) + \frac{\lambda}{2}\|\mathbf{w}\|^2\right]$$

Let $L(\mathbf{w})$ denote the expected risk associated with output vector $\mathbf{w}$.

$$L(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{Q}}\left[\phi(y\mathbf{w}^\top\mathbf{x})\right]$$

We are then interested in the quantity $L(\mathbf{w}_{\text{gop}}) - L(\mathbf{w}_0)$ which can be rewritten as follows:

$$\begin{aligned} L(\mathbf{w}_{\text{gop}}) - L(\mathbf{w}_0) &= (J_{\text{e}}(\mathbf{w}_{\text{gop}}) - J_{\text{e}}(\mathbf{w}_{\text{e}})) \\ &+ (J_{\text{e}}(\mathbf{w}_{\text{e}}) - J_{\text{e}}(\mathbf{w}_0)) \\ &+ \frac{\lambda}{2}\|\mathbf{w}_0\|^2 - \frac{\lambda}{2}\|\mathbf{w}_{\text{gop}}\|^2 \end{aligned}$$

Dropping the last term which is non-negative, we get

$$\begin{aligned} L(\mathbf{w}_{\text{gop}}) - L(\mathbf{w}_0) &\leq (J_{\text{e}}(\mathbf{w}_{\text{gop}}) - J_{\text{e}}(\mathbf{w}_{\text{e}})) \\ &+ (J_{\text{e}}(\mathbf{w}_{\text{e}}) - J_{\text{e}}(\mathbf{w}_0)) \\ &+ \frac{\lambda}{2}\|\mathbf{w}_0\|^2 \end{aligned}$$

Also as $\mathbf{w}_{\text{e}}$ minimizes J_{e}, the second term on the right hand side is negative and can be removed from the bound. Thus we have

$$L(\mathbf{w}_{\text{gop}}) - L(\mathbf{w}_0) \leq (J_{\text{e}}(\mathbf{w}_{\text{gop}}) - J_{\text{e}}(\mathbf{w}_{\text{e}})) + \frac{\lambda}{2}\|\mathbf{w}_0\|^2$$

From [6], we have the following result:

With probability $1 - \delta'$,

$$\begin{aligned} J_{\text{e}}(\mathbf{w}_{\text{gop}}) - J_{\text{e}}(\mathbf{w}_{\text{e}}) \leq 2(J_{\text{emp}}(\mathbf{w}_{\text{gop}}) - J_{\text{emp}}(\mathbf{w}_{\text{emp}})) \\ + O(\frac{\log(\frac{1}{\delta'})}{N\lambda}) \end{aligned}$$

where $\mathbf{w}_{\text{emp}}$ is the non-regularized empirical risk minimizer given by

$$\mathbf{w}_{\text{emp}} = \operatorname*{argmin}_{\mathbf{w}} J_{\text{emp}}(\mathbf{w}) := \operatorname*{argmin}_{\mathbf{w}} \sum_{i=1}^{N}\phi(y_i\mathbf{w}^\top\mathbf{x}_i)$$

Thus applying the above bound to the previous equation, we have with probability $1 - \delta'$,

$$\begin{aligned} L(\mathbf{w}_{\text{gop}}) - L(\mathbf{w}_0) \leq 2(J_{\text{emp}}(\mathbf{w}_{\text{gop}}) - J_{\text{emp}}(\mathbf{w}_{\text{emp}})) \quad (9) \\ + O(\frac{\log(\frac{1}{\delta'})}{N\lambda}) + \frac{\lambda}{2}\|\mathbf{w}_0\|^2 \quad (10) \end{aligned}$$

Using Lemma 6 of [4] ,we can get the following bound for the the quantity $J_{\text{emp}}(\mathbf{w}_{\text{gop}}) - J_{\text{emp}}(\mathbf{w}_{\text{emp}})$ when $\Delta = 0$

$$J_{\text{emp}}(\mathbf{w}_{\text{gop}}) - J_{\text{emp}}(\mathbf{w}_{\text{emp}}) \leq \frac{\|\boldsymbol{\eta}\|^2}{N^2\lambda} \quad (11)$$

It can be shown that if sufficiently large number of samples are taken and λ chosen appropriately, then the condition for setting $\Delta = 0$ in Algorithm 4 is satisfied. Precisely, notice that when $N \geq \frac{4c\|\mathbf{w}_0\|^2}{\tau\epsilon}$ and $\lambda \geq \frac{\tau}{\|\mathbf{w}_0\|^2}$, we have $N\lambda \geq \frac{4c}{\epsilon}$. From definition of $\tilde{\epsilon}$, we have

For these values of N and λ, we have $\tilde{\epsilon} \geq 0$ and hence $\Delta = 0$. Thus we can substitute the bound in Equation 11 in Equation 9 to get the following.

With probability $1 - \delta'$,

$$L(\mathbf{w}_{\text{gop}}) - L(\mathbf{w}_0) \leq 2\frac{\|\boldsymbol{\eta}\|^2}{N^2\lambda} + O(\frac{\log(\frac{1}{\delta'})}{N\lambda}) + \frac{\lambda}{2}\|\mathbf{w}_0\|^2 \quad (12)$$

Substituting $\lambda = \frac{\tau}{\|\mathbf{w}_0\|^2}$ in the above equation, we get with probability $1 - \delta'$,

$$L(\mathbf{w}_{\text{gop}}) - L(\mathbf{w}_0) \leq \frac{\|\boldsymbol{\eta}\|^2\|\mathbf{w}_0\|^2}{N^2\tau} + O(\frac{\log(\frac{1}{\delta'})\|\mathbf{w_0}\|^2}{N\tau}) + \frac{\tau}{2} \quad (13)$$

We know that if $\widehat{T}$ is chosen as in Theorem [5.1], with probability $1 - \delta'$, $\|\boldsymbol{\eta}\|^2 \leq \widehat{T}$. Applying this to the above equation, we now get with probability $1 - 2\delta'$,

$$L(\mathbf{w}_{\text{gop}}) - L(\mathbf{w}_0) \leq \frac{\widehat{T}\|\mathbf{w}_0\|^2}{N^2\tau} + O(\frac{\log(\frac{1}{\delta'})\|\mathbf{w_0}\|^2}{N\tau}) + \frac{\tau}{2} \quad (14)$$

As we want the bound on the righthand side to be less than or equal to τ, we equate the first two terms on the righthand

side to $\frac{\tau}{2}$ separately to get bounds on N. The first term yields

$$N = \frac{2\sqrt{\widehat{T}}\|\mathbf{w}_0\|}{\tau}$$

and the second term yields

$$N = 2\frac{\log(\frac{1}{\delta'})\|\mathbf{w}_0\|^2)}{\tau^2}$$

Ignoring constants in the above two equations and taking the maximum value of N and combining with the constraint got for N imposed by the requirement that $\Delta = 0$, we get the statement of the theorem.

□