A Appendix

A.1 DP-SGD CNN for MNIST

In this experiment, we use CNN for about 26k parameters and 52k parameters by widening layers to train on MNIST using DP-SGD. The result is similar comparing with the introduction that models with more parameters perform worse.

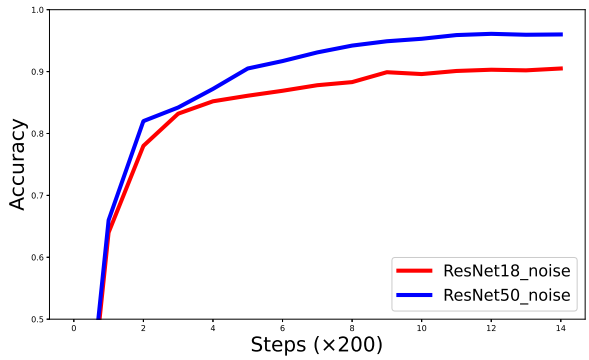


Figure 1: Accuracy for CNN using 52k parameters after DP-SGD perform 5% lower than 26k parameters.

This also shows that both width and depth of neural network would influence accuracy, leading to problem towards dimension.

A.2 Proof of Theorem 6

Before proof, we agree that character with hat is observation in this proof and truth value without hat.

Proof. First, since for normal distribution, if we have $x \sim \mathcal{N}(a, b)$ and $y \sim \mathcal{N}(c, d)$, assume x and y are independent, then

$$x + y \sim \mathcal{N}(a + c, b + d)$$

Thus data with perturbation can regard as a new data set. We will show result with new $\mathbf{\Sigma} \triangleq \mathbf{\Sigma} + \sigma^2 * \mathbf{I}_p$. Also, we define ϵ_{ij} is the bias for data i and feature j from true means, S_j is average estimation variance for feature j and $\hat{\epsilon}_{kj}$ is the average bias for class k and feature j.

For estimation $\hat{\Sigma}$, we have following inequality:

$$P\left(\max_{j=1,\cdots,p}\left|S_j^2 - \sigma_j^2\right| > \varepsilon\right) \leq \sum_{j=1}^{p} P\left(\left|S_j^2 - \sigma_j^2\right| > \varepsilon\right) \leq \sum_{j=1}^{p} P\left(\left|\sum_{i=1}^{n}\left(\epsilon_{ij}^2 - \sigma_j^2\right)\right| > n\varepsilon\right) \equiv I_1. \tag{1}$$

It follows from Bernstein's inequality that

$$P\left(\left|\sum_{i=1}^{n_k}\left(\epsilon_{kij}^2 - \sigma_{kj}^2\right)\right| > n\varepsilon\right) \leq 2\exp\left\{-\frac{c}{2}\frac{n^2\varepsilon^2}{\sum_{j=1}^{p}\sigma_j^2}\right\},$$

where c is the parameter for Bernstein's inequality.

Since $\log p = o(n)$, when $p \to \infty$, $n \to \infty$. So $I_1 = o_p(1)$. Thus $P\left(\max_{j=1,\cdots,p}\left|S_j^2 - \sigma_j^2\right| > \varepsilon\right) \xrightarrow{P} 0$. So $\hat{\mathbf{\Sigma}} = (1 + o_p(1))\mathbf{\Sigma}$.

Then we back to definition of classification error, since we assume Σ is already a diagonal matrix, after simplification, it can be written in form $W(\hat{\delta}, \theta) = 1 - \Phi(\Psi)$ where

$$\Psi \geq \frac{(\mu_1 - \hat{\mu}) \prime \hat{\mathbf{\Sigma}}^{-1} (\hat{\mu}_1 - \hat{\mu}_2)}{\sqrt{(\hat{\mu}_1 - \hat{\mu}_2) \prime \mathbf{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_2)}} (1 + o_p(1)),$$

Since $\hat{\mathbf{\Sigma}} = (1 + o_p(1))\mathbf{\Sigma}$, $\hat{\mathbf{\Sigma}}^{-1} = (1 + o_p(1))\mathbf{\Sigma}^{-1}$.

For numerator, we have

$$(\mu_1 - \hat{\mu}) \prime \hat{\mathbf{\Sigma}}^{-1} (\hat{\mu}_1 - \hat{\mu}_2) = \frac{1}{2} \alpha \prime \hat{\mathbf{\Sigma}}^{-1} \alpha - \frac{1}{2} (1 + o_P(1)) \sum \frac{\hat{\epsilon}_{1j}^2}{\sigma_j^2} + \frac{1}{2} (1 + o_P(1)) \sum \frac{\hat{\epsilon}_{2j}^2}{\sigma_j^2}.$$

The third term is in the same form with fourth, so they vanish.

For denominator, it is complicated, but in the same way.

$$\begin{aligned}
(\hat{\mu}_1 - \hat{\mu}_2) \prime \mathbf{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_2) &= \alpha \prime \mathbf{\Sigma}^{-1} \alpha + 2 \sum \alpha_j \frac{\hat{\epsilon}_{1j} \hat{\epsilon}_{2j}}{\sigma_j^2} + \sum \frac{(\hat{\epsilon}_{1j} - \hat{\epsilon}_{2j})^2}{\sigma_j^2} \\
&= \alpha \prime \mathbf{\Sigma}^{-1} \alpha + 2 \sum \frac{\alpha_j}{\sigma_j^2} (\hat{\epsilon}_{1j} - \hat{\epsilon}_{2j}) + \sum \frac{(\hat{\epsilon}_{1j} - \hat{\epsilon}_{2j})^2}{\sigma_j^2}
\end{aligned}$$

The third term is with distribution $\hat{\epsilon}_{1j} - \hat{\epsilon}_{2j} \sim \mathcal{N}(0, (4/n)\sigma_j^2)$. In term, it need to divide σ_j^2, so it converges to $4/n$.

$$\sum \frac{(\hat{\epsilon}_{1j} - \hat{\epsilon}_{2j})^2}{\sigma_j^2} \xrightarrow{P} \frac{4p}{n}$$

Then the second term is the same. $\frac{\alpha_j}{\sigma_j^2} (\hat{\epsilon}_{1j} - \hat{\epsilon}_{2j}) \sim N(0, (4/n)\alpha_j \sigma_j^{-2} \alpha_j)$. Then as variance is $o_p(1)$, so the whole term is in the order of $o_p(1)\alpha \prime \hat{\mathbf{\Sigma}}^{-1} \alpha$.

Finally, together above result, we can complete our proof. □

A.3 LEMMA 2

Let $n = n_1 + n_2$. Assume that there exist $0 < c_1 \leq c_2 < 1$ such that $c_1 \leq n_1/n_2 \leq c_2$. Let $\tilde{T}_j = T_j - \frac{\mu_{j1} - \mu_{j2}}{\sqrt{s_{1j}^2/n_1 + S_{1j}^2/n_2}}$. Then for any $x \equiv x(n_1, n_2)$ satisfying $x \to \infty$ and $x = o\left(n^{1/2}\right)$,

$$\log P\left(\tilde{T}_j \geq x\right) \sim -x^2/2, \quad \text{as} \quad n_1, n_2 \to \infty$$

If in addition, if we have $E|Y_{1ij}|^3 < \infty$ and $E|Y_{2ij}|^3 < \infty$, then

$$\frac{P\left(\tilde{T}_j \geq x\right)}{1 - \Phi(x)} = 1 + O(1)(1 + x)^3 n^{-1/2} d^3, \quad \text{for} \quad 0 \leq x \leq n^{1/6}/d$$

where $d = \left(E|X_{1ij}|^3 + E|X_{2ij}|^3\right) / \left(\text{var}(X_{1ij}) + \text{var}(X_{2ij})\right)^{3/2}$ and $O(1)$ is a finite constant depending only on c_1 and c_2.

A.4 PROOF OF THEOREM 8

Proof. First, since we consider Gaussian distribution, so lemma 2 is always tenable in below proof, we will use it directly.

Second, take into two parts. a) First, we check probability $P(max_{j>s}|D_j| > x)$. For any probability, it is clear that

$$P\left(\max_{j>s} |D_j| > x\right) \leq \sum_{j=s+1}^{p} P(|D_j| \geq x).$$

With lemma 2 and the max variance bounded after normalization, we can infer that

$$P(\max_{j>s}|D_j| > v\frac{2}{\sqrt{n}}x) \leq (1-\Phi(x))\left(1+C(1+x)^3n^{-1/2}d^3\right)$$

with $d = \left(E\left|Y_{1ij}\right|^3 + E\left|Y_{2ij}\right|^3\right) / \left(\sigma_{1j}^2+\sigma_{2j}^2\right)^{3/2}$.
Since $|T_j|$ obtain following inequality:

$$|T_j| = \frac{|D_j|}{\sqrt{s_{1j}^2/n_1 + s_{2j}^2/n_2}} \geq \frac{|D_j|}{v}\sqrt{\frac{n_1n_2}{n}} \geq \frac{\sqrt{n}}{2}\frac{|D_j|}{v}.$$

Also, with normal distribution

$$1-\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-x^2/2}dx < \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-xy/2}dy,$$

we can give that

$$1-\Phi(x) \leq \frac{1}{\sqrt{2\pi}}\frac{1}{x}e^{-x^2/2}$$

This together with the symmetry of D_j gives

$$P(|D_j| > v\frac{2}{\sqrt{n}}x) \leq 2\frac{1}{\sqrt{2\pi}}\frac{1}{x}e^{-x^2/2}\left(1+C(1+x)^3n^{-1/2}d^3\right).$$

Combining the above inequality, we have

$$\sum_{j>s}P(|D_j| > v\frac{2}{\sqrt{n}}x) \leq (p-s)\frac{2}{\sqrt{2\pi}}\frac{1}{x}e^{-x^2/2}\left(1+C(1+x)^3n^{-1/2}d^3\right).$$

Since $\log(p-s) = o\left(n^\gamma\right)$ with $0<\gamma<1/3$, if we let $x = cn^{\gamma/2}$, that is $y = cvn^{(\gamma-1)/2}$, then

$$\sum_{j>s}P\left(|D_j| \geq y\right) = n^{\gamma-1/2}.$$

So we can draw that

$$\sum_{j>s}P\left(|D_j| \geq y\right) \to 0.$$

This equality yields

$$P\left(\max_{j>s}|D_j| \geq y\right) \to 0.$$

b) Then we consider $P(\min_{j\leq s}|D_j| \leq y)$. Notice that when $j \leq s$, $\alpha_j = \mu_{1j} - \mu_{2j} \neq 0$. So also with lemma 2, we define $\tilde{D}_j = D_j - \alpha_j$, it is same like a)

$$P(\max_{j\leq s}|\tilde{D}_j| > v\frac{2}{\sqrt{n}}x) \to 0. \tag{2}$$

For addition, there is an inequality

$$y > \min_{j\leq s}|D_j| = \min_{j\leq s}|\tilde{D}_j + \alpha_j| \geq \min_{j\leq s}|\alpha_j| - \max_{j\leq s}|\tilde{D}_j|.$$

So in probability

$$P\left(\min_{j\leq s}|D_j| \leq y\right) \leq P\left(\max_{j>s}\left|\tilde{D}_j\right| \geq \min_{j\leq s}|\alpha_j| - y\right).$$

Then with all assumption above and some $\beta_n \to \infty$

$$\min_{j\leq s}|\alpha_j| - y = vn^{-\gamma}\beta_n - 2cvn^{(\gamma-1)/2} \geq y.$$

Together with (2), b) is established. Combination two parts complete the theorem. □

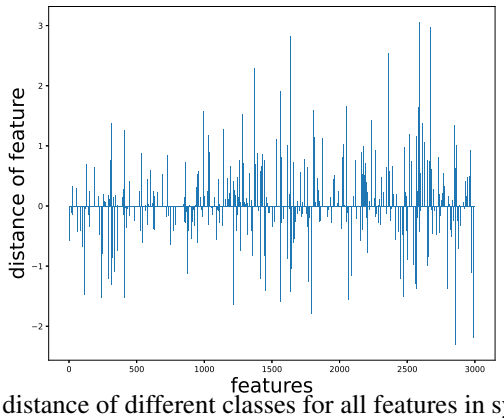


Figure 2: distance of different classes for all features in synthetic data

Algorithm 1: DP Feature Release Algorithm with k classes

Input: [[$\mathbf{X}_{11}$],...,[$\mathbf{X}_{1n_1}$]] to [[$\mathbf{X}_{k1}$],...,[$\mathbf{X}_{kn_k}$]]
Calculate average of features: $\hat{\mu}_1 = [a_{11}, ..., a_{1p}]$ to $\hat{\mu}_k = [a_{k1}, ..., a_{kp}]$
Calculate max distance of features: $D_j = \max_{c,q \leq k} |\hat{\mu}_{qj} - \hat{\mu}_{cj}|$
Rank features with distance: X_r = [[$x_{1[1]}$,...,$x_{1[p]}$],...,[$x_{n[1]}$,...,$x_{n[p]}$]]
Cut the first m features: X_c = [[$x_{1[1]}$,...,$x_{1[m]}$],...,[$x_{n[1]}$,...,$x_{n[m]}$]]
Calculate the maximum norm in X_c: $N_{max} \triangleq \max_{i \leq n, X_i \in X_c} \|X_i\|_1$
Generate noise: $n \times m$ matrix ε with i.i.d. $\varepsilon_{ij} \sim \mathcal{N}(0, 2N_{max} \ln(1/\delta)/\epsilon)$
Add noise to feature: $\hat{X} = X_c + \varepsilon$
Output: feature with noise $\hat{X}$, *Label*

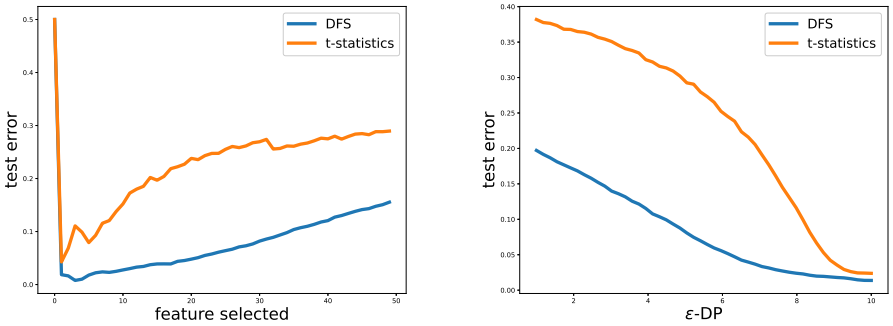
A.5 Means of μ_1 in toy experiment

Fig. 2 is a bar figure of our $\boldsymbol{\mu}_1$ in toy experiment. We can see most of the features are sparse.

A.6 Multiple class criterion

For Fisher's classifier, we consider in binary classification. But our approach can be generalized to multiple classification. We list changed algorithm in CIFAR-10 part.

A.7 Fisher classifier for CIFAR-10



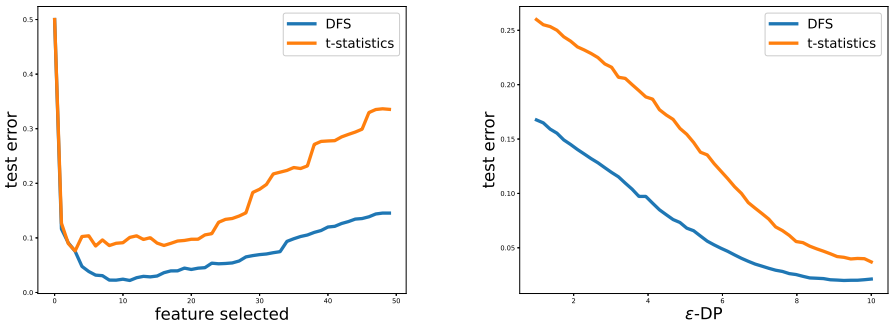
(a) fixed $\varepsilon = 6$ with p increasing

(b) fixed $p = 25$ with ε increasing

Figure 3: Results for CIFAR-10

For Fisher Classifier on CIFAR-10, left Fig. 3a shows our curve is lower and smoother which means robustness with dimension increasing. Right Fig.3b proves when ε is tiny, noise is large, DFS can perform over t-testing for more than 0.2 in test error.

A.8 Experiment for MNIST



(a) fixed $\varepsilon = 6$ with p increasing

(b) fixed $p = 25$ with ε increasing

Figure 4: Results for MNIST

Left Fig.4a shows robustness similar to CIFAR-10 since curve is lower and smoother. Right Fig.4b shows t-testing is susceptible to DP noise, even $\varepsilon = 10$ would cause error increasing.