

## A BIAS-VARIANCE BOUND ANALYSIS

In this section, we prove the bias-variance bound used in FairDP.

**Theorem 1.** *Suppose the function  $f$  is  $L$ -Lipschitz smooth, for the  $t$ -th iteration of differentially private SGD with learning rate  $\mu^t$ , batch size  $n$ , clipping threshold parameter  $C$  and noise scale  $\sigma$ , we have the clipping bias bounded by*

$$2 \left(1 + \frac{1}{\mu^t L}\right) \|\nabla f(\theta^t)\| \cdot \left(\frac{1}{n} \sum_{\|g^t(x_i)\| > C} (\|g^t(x_i)\| - C)\right) + \left(\frac{1}{n} \sum_{\|g^t(x_i)\| > C} (\|g^t(x_i)\| - C)\right)^2,$$

and the noise-addition variance

$$\frac{1}{n^2} \cdot \sigma^2 C^2 |\mathbf{I}|.$$

*Proof.* The update is of the form:

$$\theta^{t+1} = \theta^t - \mu^t \cdot \frac{1}{n} \left( \sum_i \frac{g^t(x_i)}{\max(1, \frac{\|g^t(x_i)\|_2}{C})} + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I}) \right).$$

Then

$$\begin{aligned} \mathbb{E}f(\theta^{t+1}) &\leq f(\theta^t) + \mathbb{E}[\langle \nabla f(\theta^t), \theta^{t+1} - \theta^t \rangle] + \frac{L}{2} \mathbb{E}[\|\theta^{t+1} - \theta^t\|^2] \\ &= f(\theta^t) - \mu^t \cdot \mathbb{E} \left[ \left\langle \nabla f(\theta^t), \frac{1}{n} \left( \sum_i \frac{g^t(x_i)}{\max(1, \frac{\|g^t(x_i)\|_2}{C})} + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I}) \right) \right\rangle \right] \\ &\quad + \frac{\mu^{t2} L}{2} \cdot \mathbb{E} \left[ \left\| \frac{1}{n} \left( \sum_i \frac{g^t(x_i)}{\max(1, \frac{\|g^t(x_i)\|_2}{C})} + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I}) \right) \right\|^2 \right] \\ &\leq f(\theta^t) - \mu^t \cdot \mathbb{E} \left[ \left\langle \nabla f(\theta^t), \frac{1}{n} \left( \sum_i \frac{g^t(x_i)}{\max(1, \frac{\|g^t(x_i)\|_2}{C})} + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I}) \right) \right\rangle \right] \\ &\quad + \frac{\mu^{t2} L}{2} \cdot \mathbb{E} \left[ \left\| \frac{1}{n} \sum_i \frac{g^t(x_i)}{\max(1, \frac{\|g^t(x_i)\|_2}{C})} \right\|^2 \right] + \frac{\mu^{t2} L}{2n^2} \cdot \mathbb{E}[\|\mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\|^2] \\ &\quad + \frac{\mu^{t2} L}{n} \cdot \mathbb{E} \left[ \left\langle \frac{1}{n} \sum_i \frac{g^t(x_i)}{\max(1, \frac{\|g^t(x_i)\|_2}{C})}, \mathcal{N}(0, \sigma^2 C^2 \mathbf{I}) \right\rangle \right]. \end{aligned}$$

**Lemma 1.** *The Gamma Distribution has the scaling property. That is, if  $X \sim \Gamma(\alpha, \beta)$ , then  $Y = cX$  follows  $\Gamma(\alpha, c\beta)$ , where  $c$  is a positive and real constant.*

*Proof.* Let the random variable  $X$  has the  $\Gamma(\alpha, \beta)$  with probability density function

$$f_X(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}},$$

where  $x > 0$ .

The transformation  $Y = cX$  is with inverse  $X = \frac{Y}{c}$  and  $\frac{dX}{dY} = \frac{1}{c}$ .

Therefore, the probability density function of  $Y$  is

$$\begin{aligned} f_Y(y) &= f_X\left(\frac{y}{c}\right) \frac{dX}{dY} \\ &= \frac{1}{\beta^\alpha \Gamma(\alpha)} \left(\frac{y}{c}\right)^{\alpha-1} e^{-\frac{y}{c\beta}} \frac{1}{c} \\ &= \frac{1}{(c\beta)^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-\frac{y}{c\beta}}, \end{aligned}$$

which is the probability density function of  $\Gamma(\alpha, c\beta)$ .  $\square$

If  $x_i \sim \mathcal{N}(0, \sigma^2 C^2)$ , then

$$\begin{aligned}\frac{x_i}{\sigma C} &\sim \mathcal{N}(0, 1) \\ \frac{x_i^2}{\sigma^2 C^2} &\sim \mathcal{X}^2(1) \sim \Gamma(\frac{1}{2}, 2) \\ x_i^2 &\sim \sigma^2 C^2 \Gamma(\frac{1}{2}, 2) \sim \Gamma(\frac{1}{2}, 2\sigma^2 C^2) \\ \sum_i x_i^2 &\sim \sum_i \Gamma(\frac{1}{2}, 2\sigma^2 C^2) \sim \Gamma(\frac{|\mathbf{I}|}{2}, 2\sigma^2 C^2),\end{aligned}$$

where the gamma distribution uses scale parameterization.

From the fact that  $\mathbb{E}[\mathcal{N}(0, \sigma^2 C^2 \mathbf{I})] = 0$  and  $\mathbb{E}[\|\mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\|^2] = \sigma^2 C^2 |\mathbf{I}|$ , we have

$$\begin{aligned}\mathbb{E}f(\theta^{t+1}) &\leq f(\theta^t) - \mu^t \cdot \mathbb{E} \left\langle \nabla f(\theta^t), \frac{1}{n} \sum_i \frac{g^t(x_i)}{\max(1, \frac{\|g^t(x_i)\|_2}{C})} \right\rangle \\ &\quad + \frac{\mu^{t^2} L}{2} \cdot \mathbb{E} \left\| \frac{1}{n} \sum_i \frac{g^t(x_i)}{\max(1, \frac{\|g^t(x_i)\|_2}{C})} \right\|^2 + \frac{\mu^{t^2} L}{2n^2} \cdot \sigma^2 C^2 |\mathbf{I}|.\end{aligned}$$

For the ease of exposition, we assume that  $g^t(x_i)$  are arranged in ascending order based on their  $L^2$ -norm. For any  $C$ ,

$$\begin{aligned}\frac{1}{n} \sum_i \frac{g^t(x_i)}{\max(1, \frac{\|g^t(x_i)\|_2}{C})} &= \frac{1}{n} \sum_i \begin{cases} g^t(x_i), & \text{if } \|g^t(x_i)\| \leq C \\ g^t(x_i) \cdot \frac{C}{\|g^t(x_i)\|}, & \text{otherwise} \end{cases} \\ &= \frac{\sum_{i=1}^{\mathbb{I}_{\|g^t(x_i)\| \leq C}} g^t(x_i)}{n} + \frac{\sum_{i=\mathbb{I}_{\|g^t(x_i)\| > C}^n g^t(x_i) \cdot \frac{C}{\|g^t(x_i)\|}}{n} \\ &= \nabla f(\theta^t) - \frac{1}{n} \sum_{\mathbb{I}_{\|g^t(x_i)\| > C}} g^t(x_i) \left( 1 - \frac{C}{\|g^t(x_i)\|} \right).\end{aligned}$$

Therefore, from the above inequality, we have

$$\begin{aligned}
\mathbb{E}f(\theta^{t+1}) &\leq f(\theta^t) - \mu^t \cdot \mathbb{E} \left[ \left\langle \nabla f(\theta^t), \nabla f(\theta^t) - \frac{1}{n} \sum_{\mathbb{I}_{g^t(x_i) > C}}^n g^t(x_i) \left(1 - \frac{C}{\|g^t(x_i)\|}\right) \right\rangle \right] \\
&\quad + \frac{\mu^{t^2}L}{2} \cdot \mathbb{E} \left[ \left\| \nabla f(\theta^t) - \frac{1}{n} \sum_{\mathbb{I}_{g^t(x_i) > C}}^n g^t(x_i) \left(1 - \frac{C}{\|g^t(x_i)\|}\right) \right\|^2 \right] + \frac{\mu^{t^2}L}{2n^2} \cdot \sigma^2 C^2 |\mathbf{I}| \\
&\leq f(\theta^t) - \mu^t \|\nabla f(\theta^t)\|^2 + \mu^t \cdot \mathbb{E} \left[ \left\langle \nabla f(\theta^t), \frac{1}{n} \sum_{\mathbb{I}_{g^t(x_i) > C}}^n g^t(x_i) \left(1 - \frac{C}{\|g^t(x_i)\|}\right) \right\rangle \right] \\
&\quad + \frac{\mu^{t^2}L}{2} \|\nabla f(\theta^t)\|^2 + \frac{\mu^{t^2}L}{2} \left\| \frac{1}{n} \sum_{\mathbb{I}_{g^t(x_i) > C}}^n g^t(x_i) \left(1 - \frac{C}{\|g^t(x_i)\|}\right) \right\|^2 \\
&\quad - \mu^{t^2}L \cdot \mathbb{E} \left[ \left\langle \nabla f(\theta^t), \frac{1}{n} \sum_{\mathbb{I}_{g^t(x_i) > C}}^n g^t(x_i) \left(1 - \frac{C}{\|g^t(x_i)\|}\right) \right\rangle \right] + \frac{\mu^{t^2}L}{2n^2} \cdot \sigma^2 C^2 |\mathbf{I}| \\
&= f(\theta^t) - \mu^t \|\nabla f(\theta^t)\|^2 + \frac{\mu^{t^2}L}{2} \|\nabla f(\theta^t)\|^2 \\
&\quad + (\mu^t - \mu^{t^2}L) \cdot \mathbb{E} \left[ \left\langle \nabla f(\theta^t), \frac{1}{n} \sum_{\mathbb{I}_{g^t(x_i) > C}}^n g^t(x_i) \left(1 - \frac{C}{\|g^t(x_i)\|}\right) \right\rangle \right] \\
&\quad + \frac{\mu^{t^2}L}{2} \left\| \frac{1}{n} \sum_{\mathbb{I}_{g^t(x_i) > C}}^n g^t(x_i) \left(1 - \frac{C}{\|g^t(x_i)\|}\right) \right\|^2 + \frac{\mu^{t^2}L}{2n^2} \cdot \sigma^2 C^2 |\mathbf{I}|
\end{aligned}$$

Suppose the coefficients of  $\|\nabla f(\theta^t)\|^2 < 0$ , that is  $-\mu^t + \frac{\mu^{t^2}L}{2} < 0$ , then  $\mu^t < \frac{2}{L}$ .

By Cauchy–Schwarz inequality, we know

$$\begin{aligned}
\mathbb{E}f(\theta^{t+1}) &\leq f(\theta^t) - \mu^t \|\nabla f(\theta^t)\|^2 + \frac{\mu^{t^2}L}{2} \|\nabla f(\theta^t)\|^2 \\
&\quad + (\mu^t + \mu^{t^2}L) \cdot \mu^t \|\nabla f(\theta^t)\| \cdot \left\| \frac{1}{n} \sum_{\mathbb{I}_{g^t(x_i) > C}}^n g^t(x_i) \left(1 - \frac{C}{\|g^t(x_i)\|}\right) \right\| \\
&\quad + \frac{\mu^{t^2}L}{2} \left\| \frac{1}{n} \sum_{\mathbb{I}_{g^t(x_i) > C}}^n g^t(x_i) \left(1 - \frac{C}{\|g^t(x_i)\|}\right) \right\|^2 + \frac{\mu^{t^2}L}{2n^2} \cdot \sigma^2 C^2 |\mathbf{I}|.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\left\| \frac{1}{n} \sum_{\mathbb{I}_{g^t(x_i) > C}}^n g^t(x_i) \left(1 - \frac{C}{\|g^t(x_i)\|}\right) \right\| \\
&\leq \frac{1}{n} \sum_{\mathbb{I}_{g^t(x_i) > C}}^n \|g^t(x_i)\| \cdot \left(1 - \frac{C}{\|g^t(x_i)\|}\right) \\
&\leq \frac{1}{n} \sum_{\mathbb{I}_{g^t(x_i) > C}}^n (\|g^t(x_i)\| - C)
\end{aligned}$$

Here, the desired result is obtained. The clipping bias is bounded by

$$2 \left( 1 + \frac{1}{\mu^t L} \right) \|\nabla f(\theta^t)\| \cdot \left( \frac{1}{n} \sum_{\mathbb{I}_{g^t(x_i)} > C}^n (\|g^t(x_i)\| - C) \right) + \left( \frac{1}{n} \sum_{\mathbb{I}_{g^t(x_i)} > C}^n (\|g^t(x_i)\| - C) \right)^2.$$

The noise-addition variance is

$$\frac{1}{n^2} \cdot \sigma^2 C^2 |\mathbf{I}|.$$

□

## B DESCRIPTION OF DATASETS

- We use two census datasets, Adult (Dua & Graff, 2017) and Dutch (Kamiran & Calders, 2011). For both datasets, we consider “Sex” as the protected attribute and “Income” as decision. For unprotected attributes, we convert categorical attributes to one-hot vectors and normalize numerical attributes to [0,1] range. After preprocessing, we have 13 unprotected attributes (103 dimension) for Adult and 10 unprotected attributes (59 dimension) for Dutch. The instances with unknown values are removed from Adult dataset (train=30162, test=15060). The Dutch dataset is close to balanced with 30,273 males and 30,147 females. We split the Adult and Dutch dataset into 80% training data and 20% testing data.
- We use MNIST dataset (LeCun et al., 1998) and replicate the setting in (Bagdasaryan et al., 2019). The original MNIST dataset is a balanced dataset with 60,000 training samples and each class has about 6,000 samples. Class 8 has the most false negatives, hence we choose it as the artificially underrepresented group (reducing the number of training samples from 5,851 to 500) in the unbalanced MNIST dataset. We compare the underrepresented Class 8 with the well-represented Class 2 that shares the fewest false negatives with Class 8 and therefore can be considered independent. The testing dataset has 10,000 testing samples with about 1,000 for each class.

## C COMPARISON METHODS

- SGD computes a separate gradient for each training example and averages them per class on each batch.
- DPSGD (Abadi et al., 2016): The gradients of all the parameters are clipped before grouped together to compute the norm. (also known as “flat clipping”)
- DP-FedAvg (McMahan et al., 2018): Per-layer clipping  $C = \sqrt{\sum_{j=1}^m C_j^2}$ ,  $C_j = \frac{C}{\sqrt{m}}$ , where the model has  $m$  layers; when  $m = 1$ , DP-FedAvg degenerates to DPSGD.
- Opt-Q (abbreviation for Optimal-Quantile in this paper, adapted from (Amin et al., 2019), see proof in Appendix C.1, we further improve their performance by refining  $C$  into classes):  $\left(1 - \sqrt{\frac{2}{\pi}} \cdot \frac{\sigma}{e}\right)$ -quantile;
- Dpsgd-F (Xu et al., 2020):  $C^k = C^0 \times (1 + \frac{p^k}{\frac{q^k}{e}})$ , where  $\frac{p^k}{q^k}$  represents the fraction of instances in the class with gradients greater than hyper-parameter  $C^0$ .

### C.1 OPTIMAL QUANTILE (OPT-Q)

(Amin et al., 2019) tries to do influence limitation by applying the Laplace mechanism. We follow their example to do this by applying the Gaussian mechanism.

**Theorem 2.** *Opt-Q says that the limit we should impose on instance contributions is the  $\left(1 - \sqrt{\frac{2}{\pi}} \cdot \frac{\sigma}{e}\right)$ -quantile of the gradients themselves.*

*Proof.* Recall that the noise added to  $\nabla f(\theta^t)$  follows a Gaussian distribution with scale parameter  $\sigma C$ . We can decompose the expected error of the estimate  $g^t(x^t)$  into a variance term (due to the noise) and a bias term (due to the contribution limit):

$$\begin{aligned} & \mathbb{E}|g - \nabla f(\theta)| \\ & \leq \mathbb{E}|g - g_C| + |g_C - \nabla f(\theta)| \\ & = \sqrt{\frac{2}{\pi}} \cdot \frac{\sigma C}{e} + \sum_{i=1}^n \max(0, g(x_i) - C) \end{aligned}$$

where we use the fact that the mean of the folded Gaussian variable is equal to  $\sqrt{\frac{2}{\pi}} \cdot \frac{\sigma C}{e}$ . We can find the optimal  $C$  by noting that the bound is convex with sub-derivative

$$\sqrt{\frac{2}{\pi}} \cdot \frac{\sigma}{e} - |\{i : g(x_i) > C\}|,$$

thus the minimum is achieved when  $C$  is equal to the  $\sqrt{\frac{2}{\pi}} \cdot \frac{\sigma n}{e}$ -th largest gradient. It says that the limit we should impose on sample contributions is just the  $\left(1 - \sqrt{\frac{2}{\pi}} \cdot \frac{\sigma}{e}\right)$ -quantile of the gradients themselves.  $\square$

## D CODE APPENDIX

### D.1 COMPUTING INFRASTRUCTURE FOR RUNNING EXPERIMENTS

#### D.1.1 HARDWARE

**GPU:** NVIDIA GeForce RTX 2080Ti

**GPU model:** The peak memory usage of the neural network with 2 convolutional layers and 2 fully-connected layers for MNIST dataset is 1509MB.

**CPU:** Intel(R) Xeon(R) CPU E5-2678 v3 @ 2.50GHz

**CPU model:** The peak memory usage of the logistic regression for Adult and Dutch datasets is 285MB.

#### D.1.2 SOFTWARE LIBRARIES AND FRAMEWORKS

We use PyTorch 1.6.0 to implement all the methods. Our code depend on Opacus library<sup>5</sup> It enables vectorized per-sample gradient computation that is 10x faster than microbatching.

We use the privacy testing library in TensorFlow<sup>6</sup> to assess the privacy properties of SGD, DPSGD and FairDP under membership attack. The accumulated privacy budget  $\epsilon$  for each setting is computed using the privacy moments accounting method. We use Rényi DP only to estimate privacy loss. This does not change the DPSGD algorithm of (Abadi et al., 2016) but rather provides tighter bounds on privacy loss, allowing to reduce the amount of added noise. The TensorFlow Privacy tool<sup>7</sup> enables estimation of  $\epsilon$  given the input parameters (dataset size, number of epochs, batch size,  $\ell^2$  noise, delta) before starting the training.

### D.2 SEEDS TO ALLOW REPLICATION OF RESULTS / THE NUMBER OF ALGORITHM RUNS

The private learning methods (i.e., DPSGD, DP-FedAVG, Opt-Q, DPSGD-F, and Fair-DP) depend on randomness. In our experiments, we set seeds from 0 to 9 and run the algorithms 10 times repeatedly to compute each reported result.

<sup>5</sup><https://github.com/pytorch/opacus>

<sup>6</sup>[https://github.com/tensorflow/privacy/tree/master/tensorflow\\_privacy/privacy/membership\\_inference\\_attack](https://github.com/tensorflow/privacy/tree/master/tensorflow_privacy/privacy/membership_inference_attack)

<sup>7</sup>[https://github.com/tensorflow/privacy/blob/master/tensorflow\\_privacy/privacy/analysis/compute\\_dp\\_sgd\\_privacy.py](https://github.com/tensorflow/privacy/blob/master/tensorflow_privacy/privacy/analysis/compute_dp_sgd_privacy.py)

### D.3 HYPER-PARAMETERS FOR EACH MODEL

For the Adult and Dutch datasets, we use a logistic regression model with regularization parameter 0.01, learning rate  $\frac{1}{\sqrt{T}}$ , batch size 256, and training epochs 20.

For the MNIST dataset, we use a neural network with 2 convolutional layers and 2 linear layers with 431K parameters in total. We use learning rate 0.01, batch size 256 and 128 respectively, and the number of training epochs 60,  $\sigma = 1.0$ .

### D.4 SUMMARY OF PERFORMANCE

#### D.4.1 SUMMARY OF FAIRNESS

To draw an analogy between machine learning and the problem of resource allocation, one can think of the model as a resource that is meant to serve the classes. In this sense, it is natural to ask questions about the fairness of the model performance for classes.

We take “performance” to be the testing accuracy of applying the trained model on the test data for each class. Our definition can be seen as a relaxed version of accuracy parity (Zafar et al., 2017), in that we optimize for similar accuracy reduction but not necessarily identical performance for each class.

There are many ways to mathematically evaluate the uniformity of the performance. In this work, we use four indexes (Bureau, 2016), including *Atkinson Index*, *Gini index*, *Mean Log Deviation*, and *Theil Index*.

**Atkinson Index**  $= 1 - \frac{1}{\bar{y}} \left( \prod_{k=1}^{\ell} y_k^{\mathcal{G}_k} \right)^{1/\mathcal{G}}$ , where  $\bar{y} = \sum_{k=1}^{\ell} y_k \mathcal{G}_k$ .

**Gini**  $= \frac{1}{2\bar{y}\mathcal{G}^2} \sum_{k=1}^{\ell} \sum_{k'=1}^{\ell} \mathcal{G}_k \mathcal{G}_{k'} |y_k - y_{k'}|$ .

**Mean Log Deviation (MLD)**  $= \frac{1}{\mathcal{G}} \sum_{k=1}^{\ell} \mathcal{G}_k \ln \frac{\bar{y}}{y_i}$ .

**Theil Index:** The Theil *T* index  $T_T = T_1 = \frac{1}{\mathcal{G}} \sum_{k=1}^{\ell} \frac{y_i}{\bar{y}} \ln \left( \frac{y_i}{\bar{y}} \right)$ ; the Theil *L* index  $T_L = T_0 = \frac{1}{\mathcal{G}} \sum_{k=1}^{\ell} \ln \left( \frac{\bar{y}}{y_i} \right)$ . In this paper, we choose  $T_1$  to compensate the other indexes’ observation.

#### D.4.2 SUMMARY OF PRIVACY

The vulnerability score (or memorization potential) is measured via the Area Under the ROC-Curve (AUC) and  $\max |r_{\text{fp}} - r_{\text{tp}}|$  (advantage) of the attack classifier.

## E DUTCH

Figure 7, 8 and 9 show the sensitivity of compared private methods on learning rate, batch size and noise on Dutch dataset (supplementary to Section 4.3). Opt-Q and DPSGD have higher accuracy on Class “Female” at the cost of decline on Class “Male” that already has lower accuracy in non-private learning.

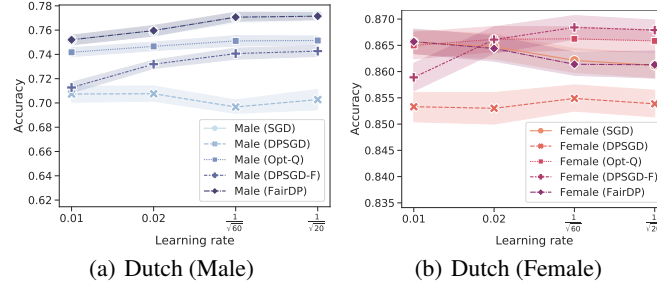


Figure 7: Effect of learning rate on training

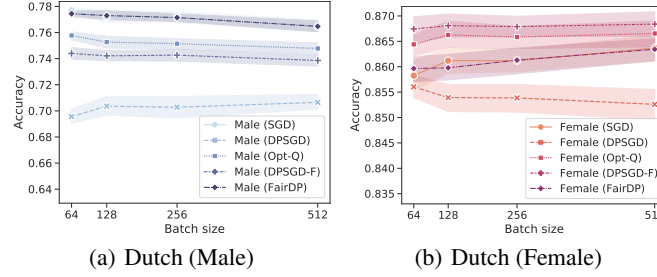


Figure 8: Effect of batch size on training

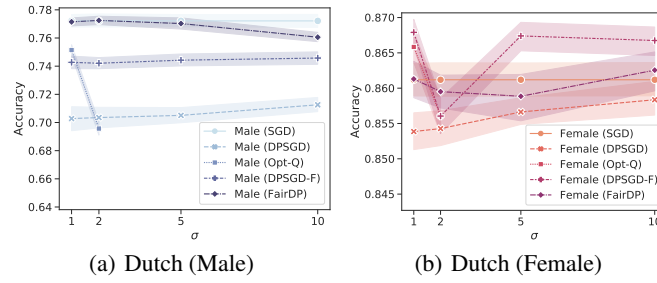


Figure 9: Effect of noise on training