Appendix

The Appendix is organized as follows:

- Appendix A provides more details on experimental setups for training, presents the effect of Monte Carlo estimation and runtime of attacks, and reports the results on backdoored test set.
- Appendix B provides proofs for our Theorem 2 and Lemma 1 related to model closeness.
- Appendix C gives proofs for our Theorem 3 related to the parameter smoothing.

A. Experimental Details

A.1. More Details on Experiment Setup for Training

We focus on multi-class logistic regression (one linear layer with softmax function and cross-entropy loss), which is a convex classification problem. We train the FL system following our CRFL framework with three datasets: Lending Club Loan Data (LOAN) (Kan, 2019), MNIST (LeCun & Cortes, 2010), and EMNIST (Cohen et al., 2017). The financial dataset LOAN is a tabular dataset that contains the current loan status (Current, Late, Fully Paid, etc.) and latest payment information, which can be used for loan status prediction. It consists of 1,808,534 data samples and we divide them by 51 US states, each of whom represents a client in FL, hence the data distribution is non-i.i.d. 80% of data samples are used for training and the rest is for testing. EMNIST is an extended MNIST dataset that contains 10 digits and 37 letters. In the two image datasets, we split the training data for FL clients in an i.i.d. manner. The data description and other parameter setups are summarized in Table 1. For these datasets, the local learning rate η_i is 0.001 for all clients. The server performs an adaptive norm clipping threshold ρ_t that increases by time so that the normal learning ability of the model can be preserved (described in Table 1), and sets the fixed training noise level $\sigma_t = 0.01$ ($t < T$). When the clipping threshold is not a fixed value, $L_{\mathcal{Z}}$ is calculated based on $\rho_{\mathsf{t_{adv}}}$ following Lemma 1 for our experiment,.

Regarding the attack setting, by default, we set $R = 1$, and if there are more adversarial clients, we use same parameters setups for all of them. For the pixel-pattern backdoor in MNIST and EMNIST, the attackers add the backdoor pattern (see Figure. 10 for an example) in images and swap the label of any sample with such patterns into the target label, which is "digit 0". Similarly, for the preprocessed[1] LOAN dataset, the attackers increase the value of the two features (i.e., num_tl_120dpd_2m, num_tl_90g_dpd_24m) as a backdoor pattern, and swap label to "Does not meet the credit policy. Status:Fully Paid". Since we adopt Lemma 1 for our experiments, we focus on the backdoor pattern $\|\delta_i\| = \|\delta_{i_x}\|$. The magnitude of backdoored pattern in every example is $\|\delta_i\| = 0.1$ on three datasets. Every attacker's batch is mixed with correctly labeled data and such backdoored data with poison ratio q_{B_i}/n_{B_i}.

We train the FL global model until convergence and then use our certification in Algorithm 2 for robustness evaluation.

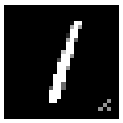


Figure 10. Backdoor pattern for image datasets

Dataset	Classes	#Training samples	Features	N	q_{B_i}/n_{B_i}	τ_i	γ_i	$\mathsf{t_{adv}}$	ρ_t
LOAN	9	1446827	91	20	40/800	143	10	6	0.025t+2
MNIST	10	60000	784	20	5/100	30	10	10	0.1t+2
EMNIST	47	697932	784	50	5/200	70	20	10	0.25t+4

Table 1. Dataset description and parameters

A.2. More Experimental Results on Clean Test Set

Effect of Monte Carlo estimation Recall that we use M and α when calculating the lower bound $\underline{p_A}$ and the upper bound $\overline{p_B}$. Figure 11 (left) shows that larger number M of noisy models used for certification can result in larger certified radius. Figure 11 (middle) presents that the certified radius is smaller when the error tolerance α is smaller but overall the certified accuracy is not very sensitive to α.

Effect of Attack Timing $\mathsf{t_{adv}}$ For Figure 11 (right), we use a strong attack (γ=100, R=2) and report the certified accuracy with different $\mathsf{t_{adv}}$. As described in Table 1, $\rho_{\mathsf{t_{adv}}}$ increases with $\mathsf{t_{adv}}$, and $L_{\mathcal{Z}}$ is calculated based on $\rho_{\mathsf{t_{adv}}}$. In order to control

[1]We preprocess LOAN by dropping the features which are not digital and cannot be one-hot encoded, and then normalizing the rest 90 features and so that the value of each feature is between 0 and 1.

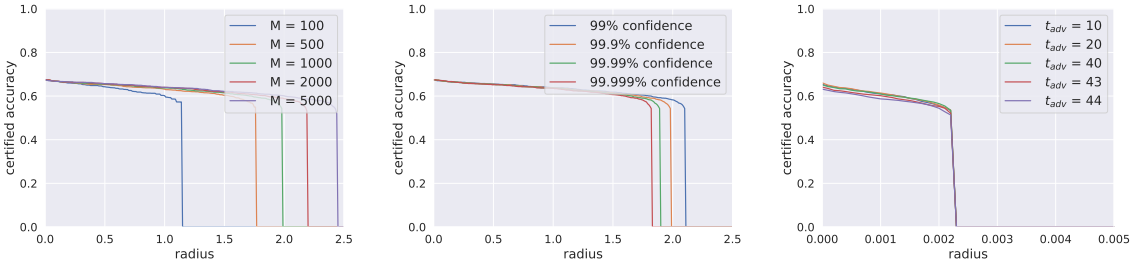
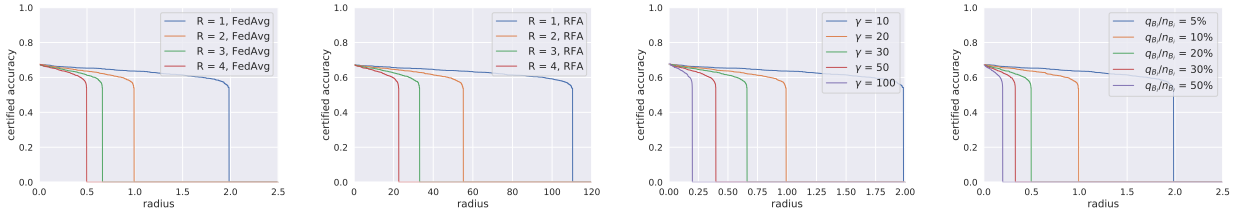


Figure 11. Left: Certified accuracy on MNIST with different number of smoothed models M for certification. Middle: Certified accuracy on MNIST with different error tolerance α for certification. Right: Certified accuracy with different $\mathsf{t}_{\mathsf{adv}}$ on MNIST.



(a) Different number of adversarial clients R

(b) Different R under robust aggregation RFA(Pillutla et al., 2019)

(c) Different scaling factor γ

(d) Different poison ratio q_{B_i}/n_{B_i}

Figure 12. Certified accuracy with different attack ability (a)(c)(d) and certified accuracy under robust aggregation RFA (Pillutla et al., 2019) (b) on MNIST backdoored test set.

variable, we use the same, loose L_z which is calculated based on ρ_{44} for all $\mathsf{t}_{\mathsf{adv}} = 10, 20, 40, 43, 44$. The results show that the certified radius is not sensitive to the attack timing $\mathsf{t}_{\mathsf{adv}}$ after training sufficient number of rounds with clean datasets after $\mathsf{t}_{\mathsf{adv}}$.

A.3. Experimental Results on Backdoored Test Set

In this section, we report the certified accuracy on the backdoored test set. For every test sample, the backdoor pattern is added to the input while the label is still correct. As shown in Figure 12 and 13, the results are similar to the results on the clean test set.

B. Proofs of Model Closeness

In this section, we will present preliminaries on f-divergence, define the problem of model closeness and then provide the detailed proofs for our Theorem 2 and Lemma 1 that are related to model closeness. Let us list the notations used in the paper and the Appendix in Table 2.

Throughout this paper, "benign training process" is the process that trains with clean dataset D for T rounds and outputs $\mathcal{M}(D)$; "backdoored training process" is the process that trains with poisoned dataset D' at round $\mathsf{t}_{\mathsf{adv}}$, trains with original clean dataset when $t \neq \mathsf{t}_{\mathsf{adv}}$, and outputs $\mathcal{M}(D')$.

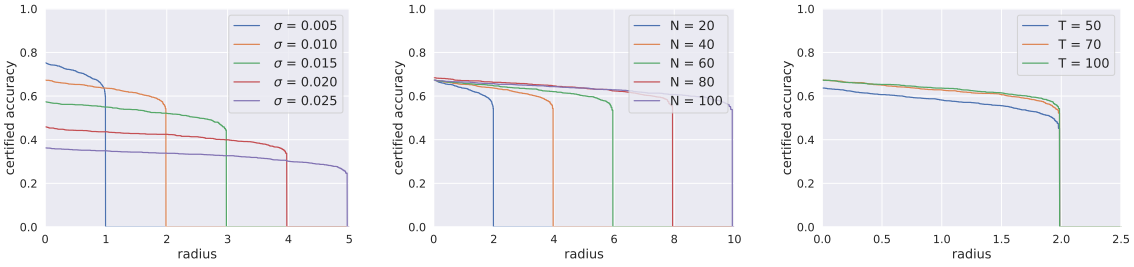
B.1. Preliminaries on f-divergence

Let $f : (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$, ν and ρ be two probability distributions. Then f-divergence is defined as

$$D_f(\nu||\rho) = E_{W\sim\rho}[f(\frac{\nu(W)}{\rho(W)})]. \tag{3}$$

Common f-divergence includes Total variation $f(x) = \frac{1}{2}\|x - 1\|$ and Kullback-Leibler (KL) divergence $f(x) = x \log x$.

Lemma 2. *For $m_1, m_2 \in \mathbb{R}^d$ and σ>0, let $\mathcal{N}_1$and $\mathcal{N}_2$ denote Gaussian distribution $\mathcal{N}_1(m_1, \sigma^2 I)$ and $\mathcal{N}_2(m_2, \sigma^2 I)$,*



(a) Different training time noise level σ (b) Different number of total clients N (c) Different training rounds T

Figure 13. Certified accuracy with different σ (a), N (b) and T (c) on MNIST backdoored test set.

Table 2. Table of notations.

Notation	Description
$\mathcal{M}(\cdot)$	the training protocol in Algorithm 2
$z_j^i := \{x_j^i, y_j^i\}$	j-th data sample at client i with input x_j^i and label y_j^i
$z'^i_j := \{x_j^i + \delta_{ix}, y_j^i + \delta_{iy}\}$	backdoored version of z_j^i where δ_{ix} is input backdoor pattern and δ_{iy} is label flipping effect
$D := \{S_1, S_2, \ldots, S_N\}$	Clean training dataset, the union of clean local dataset of N clients
$D' = D + \{\{\delta_i\}_{j=1}^{q_i}\}_{i=1}^{R}$.	poisoned training dataset in round $\mathsf{t}_{\mathsf{adv}}$ with R attackers and q_i poisoned samples in i-th attacker's local dataset
$\mathcal{M}(D)$	the clipped global model obtained from $\mathcal{M}$ using D
$\mathcal{M}(D')$	the clipped global model obtained from $\mathcal{M}$ that uses D' at round $\mathsf{t}_{\mathsf{adv}}$ and uses D at round $t \neq \mathsf{t}_{\mathsf{adv}}$
$g_i(w) = g_i(w; \xi^i)$	local gradients at client i w.r.t w with clean batch ξ^i
$g'_i(w) = g_i(w; \xi'^i)$	local gradients at client i w.r.t w with poisoned batch $\xi' i$
$\mathcal{B}^i \triangleq g'_i(w) - g_i(w)$	the difference between poisoned local gradient and benign local gradient w.r.t same model parameters w
w_s^i	client i's local model parameters at local iteration s
$w_t \leftarrow \widetilde{w}_{t-1} + \sum_{i=1}^{N} p_i(w^i_{t\tau_i} - \widetilde{w}_{t-1})$	aggregated global model at round t
$\mathrm{Clip}_{\rho_t}(w_t) \leftarrow w_t / \max(1, \frac{\|w_t\|}{\rho_t})$	clipped global model with model parameters norm threshold ρ_t at round t
$\widetilde{w}_t \leftarrow \mathrm{Clip}_{\rho_t}(w_t) + \epsilon_t$	global model at round t that is perturbed by noise ϵ_t
h_s	the smoothed classifier transferred from the base classifier h
$p_c = H_s^c(w; x_{test}) = \mathbb{P}_{W\sim\mu(w)}[h(W; x_{test}) = c]$	the probability (the majority votes) of class c for the given w and x_{test}
$h_s(w; x_{test}) = \arg\max_{c\in\mathcal{Y}} H_s^c(w; x_{test})$	the mostly probable label among all classes (the majority vote winner) for the given w and x_{test}

respectively. Then,

$$D_{KL}(\mathcal{N}_1||\mathcal{N}_2) = \frac{\|m_2 - m_1\|^2}{2\sigma^2}, \tag{4}$$

$$D_{TV}(\mathcal{N}_1||\mathcal{N}_2) = 2\Phi\left(\frac{\|m_2 - m_1\|}{\sigma}\right) - 1, \tag{5}$$

where Φ is the CDF of the Gaussian distribution.

The well-known data processing inequality (Polyanskiy & Wu, 2015) for the relative entropy states that, for any convex function f and any stochastic transformation (probability transition kernel), i.e., Markov Kernel K, we have

$$D_f(\nu K||\rho K) \leq D_f(\nu||\rho),$$

where νK denotes the push-forward of ν by K, i.e., $\nu K = \int \nu(dW)K(W)$. In other words, $D_f(\nu||\rho)$ decreases by post-processing. (Asoodeh & Calmon, 2020) extends it into machine learning and the operations in a Markov Kernel contain one step of Stochastic Gradient Descent (SGD).

To capture this effect, the quantity of the noisiness of a Markov operator (Raginsky, 2016) for f-divergence, i.e., contraction coefficient (Asoodeh & Calmon, 2020), is defined as

$$\eta_f(K) := \sup_{\nu,\rho; D_f(\nu||\rho)\neq 0} \frac{D_f(\nu K||\rho K)}{D_f(\nu||\rho)}. \tag{6}$$

Lemma 3 (Two-point characterization of Total variation (Dobrushin, 1956)). *The supremum in the definition of $\eta_{TV}(K)$ can be restricted to point mass:*

$$\eta_{TV}(K) := \sup_{y_1, y_2 \in y} D_{TV}(K(y_1)||K(y_2)) \tag{7}$$

Lemma 4 ($\eta_{TV}(K)$ Upper Bound (Makur, 2019)). *For any f-divergence, we have*

$$\eta_f(K) \leq \eta_{TV}(K) \tag{8}$$

B.2. Problem Definition

As described in Algorithm 1, due to the Gaussian noise perturbation mechanism, in each iteration the global model can be viewed as a random vector with the Gaussian smoothing measure μ. We use the f-divergence between $\mu(\mathcal{M}(D'))$ and $\mu(\mathcal{M}(D))$ as a statistical distance for measuring model closeness. According to the data post-processing inequality, when we interpret each round of CRFL as a probability transition kernel, i.e., a Markov Kernel, the contraction coefficient of Markov Kernel can help bound the divergence over multiple training rounds of FL.

Iteration as Markov Kernel We identify each iteration as a Markov Kernel. At iteration t, the central server produces the new model by $\widetilde{w}_t \leftarrow \text{Clip}_{\rho_t}(w_t) + \epsilon_t$ where w_t is the aggregated model. We denote $w_t = \Psi_t(\widetilde{w}_{t-1})$, and

$$\widetilde{w}_t \leftarrow \text{Clip}_{\rho_t}(\Psi_t(\widetilde{w}_{t-1})) + \epsilon_t, \tag{9}$$

where

$$\Psi_t(\widetilde{w}_{t-1}) \triangleq \widetilde{w}_{t-1} - \sum_{i=1}^{N} p_i \eta_i \sum_{s=(t-1)\tau_i+1}^{t\tau_i} g_i\left(w_{s-1}^i; \xi_{s-1}^i\right) \tag{10}$$

is the federated learning SGD process and the local model is initialized as $w_{(t-1)\tau_i}^i \leftarrow \widetilde{w}_{t-1}$. Therefore, iteration t can be realized by K_t, a Markov Kernel associated with the mapping $\widetilde{w}_{t-1} \to \text{Clip}_{\rho_t}(\Psi_t(\widetilde{w}_{t-1})) + \epsilon_t$. K_t receives $\widetilde{w}_{t-1}$ and then generates $\widetilde{w}_t$. Let μ_t denote the distribution of global model $\widetilde{w}_t$, and we have $\widetilde{w}_{t-1} \sim \mu_{t-1}$, then $\mu_t = \int \mu_{t-1}(dy) K_t(y)$.

Model Replacement Attack at $\mathsf{t}_{\mathsf{adv}}$ We define the backdoored federated learning SGD process Ψ'_t at round $t = \mathsf{t}_{\mathsf{adv}}$ as

$$\Psi'_t(\widetilde{w}_{t-1}) \triangleq \widetilde{w}_{t-1} - \sum_{i=1}^{R} p_i \gamma_i \eta_i \sum_{s=(t-1)\tau_i+1}^{t\tau_i} g_i\left(w'^i_{s-1}; \xi'^i_{s-1}\right) - \sum_{i=R+1}^{N} p_j \eta_j \sum_{s=(t-1)\tau_i+1}^{t\tau_i} g_j\left(w_{s-1}^i; \xi_{s-1}^i\right) \tag{11}$$

where the local model is initialized as $w'^i_{(t-1)\tau_i} \leftarrow \widetilde{w}_{t-1}$. Then we define the corresponding Markov Kernel K'_t associated with the mapping $\widetilde{w}_{t-1} \to \text{Clip}_{\rho_t}(\Psi'_t(\widetilde{w}_{t-1})) + \epsilon_t$. Through aggregation, the global model is influenced by adversarial clients. Let μ'_t denotes the distribution of backdoored global model $\widetilde{w}'_t$, and we have $\widetilde{w}_{t-1} \sim \mu_{t-1}$, then $\mu'_t = \int \mu_{t-1}(dy) K'_t(y)$.

After Model Replacement Attack After $\mathsf{t}_{\mathsf{adv}}$, all clients use the original clean datasets to update their local model. However, the global model in the backdoored training process already begins to differ from the one in the benign training process from round $\mathsf{t}_{\mathsf{adv}}$ so it is difficult to analysis it through distributed SGD. Therefore, we use Markov Kernel to quantify the poisoning effect. When $t > \mathsf{t}_{\mathsf{adv}}$, we have $\widetilde{w}'_{t-1} \sim \mu'_{t-1}$, then $\mu'_t = \int \mu'_{t-1}(dy) K_t(y)$. Because the clean datasets are used for both clean and backdoored training process when $t > \mathsf{t}_{\mathsf{adv}}$, the Markov Kernel K_t is the same. We define the contraction coefficient (Asoodeh & Calmon, 2020) as:

$$\eta_f(K_t) := \sup_{\substack{\mu_{t-1}, \mu'_{t-1}; \\ D_f(\mu_{t-1}\|\mu'_{t-1}) \neq 0}} \frac{D_f(\mu_{t-1}K_t \| \mu'_{t-1}K_t)}{D_f(\mu_{t-1}\|\mu'_{t-1})}. \tag{12}$$

Therefore, $\eta_f(K_t)$ can serve as the upper bound for the real $\frac{D_f(\mu_t\|\mu'_t)}{D_f(\mu_{t-1}\|\mu'_{t-1})}$. Then we write the model closeness $D_f(\mu_T\|\mu'_T)$

as:

$$D_f(\mu_T\|\mu'_T) = D_f(\mu_{\mathsf{t_{adv}}}\|\mu'_{\mathsf{t_{adv}}})\frac{D_f(\mu_{\mathsf{t_{adv}}+1}\|\mu'_{\mathsf{t_{adv}}+1})}{D_f(\mu_{\mathsf{t_{adv}}}\|\mu'_{\mathsf{t_{adv}}})}\cdots\frac{D_f(\mu_T\|\mu'_T)}{D_f(\mu_{T-1}\|\mu'_{T-1})}$$
$$\leq D_f(\mu_{\mathsf{t_{adv}}}\|\mu'_{\mathsf{t_{adv}}})\prod_{t=\mathsf{t_{adv}}+1}^{T}\eta_f(K_t). \tag{13}$$

We will compute $D_f(\mu_{\mathsf{t_{adv}}}\|\mu'_{\mathsf{t_{adv}}})$ and $\eta_f(K_t)$ respectively in the following sections.

B.3. Analysis for $t = \mathsf{t_{adv}}$

We would like to bound the divergence of the global model at round $\mathsf{t_{adv}}$ between the benign training process and the backdoor training process, i.e., $D_f(\mu_{\mathsf{t_{adv}}}\|\mu'_{\mathsf{t_{adv}}})$. We consider KL divergence. Based on the KL divergence for two Gaussian distributions in Lemma 2 and Assumption 3, we have

$$D_{KL}(\mu_{\mathsf{t_{adv}}}\|\mu'_{\mathsf{t_{adv}}}) = D_{KL}\left(\mathcal{N}\left(\mathrm{Clip}_{\rho_{\mathsf{t_{adv}}}}\left(w_{\mathsf{t_{adv}}}\right),\sigma^2_{\mathsf{t_{adv}}}\mathbf{I}\right)\|\mathcal{N}\left(\mathrm{Clip}_{\rho_{\mathsf{t_{adv}}}}\left(w'_{\mathsf{t_{adv}}}\right),\sigma^2_{\mathsf{t_{adv}}}\mathbf{I}\right)\right)$$
$$= \frac{\left\|\mathrm{Clip}_{\rho_{\mathsf{t_{adv}}}}(w_{\mathsf{t_{adv}}}) - \mathrm{Clip}_{\rho_{\mathsf{t_{adv}}}}\left(w'_{\mathsf{t_{adv}}}\right)\right\|^2}{2\sigma^2_{\mathsf{t_{adv}}}}$$
$$\leq \frac{\left\|w_{\mathsf{t_{adv}}} - w'_{\mathsf{t_{adv}}}\right\|^2}{2\sigma^2_{\mathsf{t_{adv}}}}. \tag{14}$$

Accumulated Effect in Local Iterations In order to bound $\left\|w_{\mathsf{t_{adv}}} - w'_{\mathsf{t_{adv}}}\right\|^2$, we look at the local iterations $s = (t-1)\tau_i + 1, (t-1)\tau_i + 2, \ldots, t\tau_i$ of adversarial client i for the benign training process and the backdoored training process. We use $\underline{s} = s - (\mathsf{t_{adv}} - 1)\tau_i, \underline{s} = 1, 2, \ldots, \tau_i$ for simplicity. We denote $\Delta^i_{\underline{s}} \triangleq w^i_{\underline{s}} - w'^i_{\underline{s}}$. Note that $\Delta^i_0 = 0$ because in the start of round $\mathsf{t_{adv}}$, the initial local model is the same benign global model $w^i_{(\mathsf{t_{adv}}-1)\tau_i} = w'^i_{(\mathsf{t_{adv}}-1)\tau_i} = \widetilde{w}_{\mathsf{t_{adv}}-1}$ for all clients $i \in [N]$ in both benign and backdoored training process. For simplicity, we will use $g_i(w), g'_i(w)$ instead of $g_i(w;\xi), g_i(w;\xi')$ in the rest of this section. We denote $\mathcal{B}^i \triangleq g'_i(w) - g_i(w)$.

Lemma 5. *Under Assumption 1 and the condition $\eta_i \leq \frac{1}{\beta}$, for $\underline{s} \in [1, \tau_i]$, we have*

$$\Delta^{i\,2}_{\underline{s}+1} \leq \Delta^{i\,2}_{\underline{s}} + 2\eta_i\left\|\mathcal{B}^i\right\|\Delta^i_{\underline{s}} + 2\eta_i^2\left\|\mathcal{B}^i\right\|^2. \tag{15}$$

We defer the proof to Section B.5. Lemma 5 states that the deviation at the current local iteration $\Delta^i_{\underline{s}}$ is added upon the deviation at the last iteration.

Lemma 6. *Based on Lemma 5, under Assumption 1 and the condition $\eta_i \leq \frac{1}{\beta}$, for $\underline{s} \in [1, \tau_i]$, we have*

$$\Delta^i_{\underline{s}} \leq 2\eta_i\left\|\mathcal{B}^i\right\|\underline{s}. \tag{16}$$

Proof. We prove it using induction argument (Zhang et al., 2017). Due to the fact $\Delta^i_0 = 0$, so $\Delta^i_1 \leq \sqrt{2\eta_i^2\left\|\mathcal{B}^i\right\|^2} \leq 2\eta_i\left\|\mathcal{B}^i\right\|$. Therefore, $\Delta^i_{\underline{s}} \leq 2\eta_i\left\|\mathcal{B}^i\right\|\underline{s}$ for $\underline{s} = 1$. Suppose the argument $\Delta^i_{\underline{s}} \leq 2\eta_i\left\|\mathcal{B}^i\right\|\underline{s}$ holds for some $\underline{s}$, then we verify $\underline{s}+1$,

$$\Delta^{i\,2}_{\underline{s}+1} \leq 4\eta_i^2\left\|\mathcal{B}^i\right\|^2\underline{s}^2 + 4\eta_i^2\left\|\mathcal{B}^i\right\|^2\underline{s} + 2\eta_i^2\left\|\mathcal{B}^i\right\|^2$$
$$= \eta_i^2\left\|\mathcal{B}^i\right\|^2(4\underline{s}^2 + 8\underline{s} + 4)$$
$$\leq 4\eta_i^2\left\|\mathcal{B}^i\right\|^2(\underline{s}+1)^2.$$

It turns out that $\Delta^i_{\underline{s}} \leq 2\eta_i\left\|\mathcal{B}^i\right\|\underline{s}$ also holds for $\underline{s}+1$. Thus, the argument is correct. □

Lemma 6 states that the deviation is accumulated over the local iterations. The larger number of local iterations τ_i, the larger deviation $\Delta^i_{\tau_i}$. Next, we provide the upper bound for $\|\mathcal{B}^i\|$.

Lemma 7. *Under the Assumption 2 on Lipschitz gradient w.r.t. data, when the adversarial clients have q_{Bi} backdoored samples out of a batch with size n_{Bi}, we have*

$$\|\mathcal{B}^i\| \leq \frac{q_{Bi}}{n_{Bi}} L_{\mathcal{Z}} \|\delta_i\|. \tag{17}$$

Proof.

$$\begin{aligned}
\left\|\mathcal{B}^i\right\| &= \|{g'}_i(w) - g_i(w)\| \\
&= \left\| \frac{1}{n_{B_i}} \left(\sum_{j=1}^{q_{B_i}} \nabla \ell(w; {z'}_j^i) + \sum_{j=q_{B_i}+1}^{n_{B_i}} \nabla \ell(w; z_j^i) \right) - \frac{1}{n_{B_i}} \sum_{j=1}^{n_{B_i}} \nabla \ell(w; z_j^i) \right\| \\
&= \left\| \frac{1}{n_{B_i}} \sum_{j=1}^{q_{B_i}} \left(\nabla \ell(w; {z'}_j^i) - \nabla \ell(w; z_j^i) \right) \right\| \\
&\leq \left\| \frac{1}{n_{B_i}} L_{\mathcal{Z}} \sum_{j=1}^{q_{B_i}} ({z'}_j^i - z_j^i) \right\| \\
&= \frac{q_{Bi}}{n_{Bi}} L_{\mathcal{Z}} \left\| \delta_i \right\|.
\end{aligned}$$

□

Scaling and Aggregation Let the scale factor be γ_i for i-th adversarial client, then the scaled malicious local update is $\gamma_i({w'}^i_{\mathsf{t}_{\mathsf{adv}}\tau_i} - \widetilde{w}_{\mathsf{t}_{\mathsf{adv}}-1})$. We assume in the benign setting (which is a virtual training process for analyzing, and we do not really train such model), this client also scales its clean local updates as $\gamma_i(w^i_{\mathsf{t}_{\mathsf{adv}}\tau_i} - \widetilde{w}_{\mathsf{t}_{\mathsf{adv}}-1})$, which can be expanded as $-\eta_i\gamma_i \sum_{s=(\mathsf{t}_{\mathsf{adv}}-1)\tau_i+1}^{\mathsf{t}_{\mathsf{adv}}\tau_i} g_i\left(w^i_{s-1}; \xi^i_{s-1}\right)$. This assumption does not hurt the global model performance in the virtual benign setting since the local learning objectives are benign so scaling the updates is equivalent to scale its local learning rate $\eta_i \leftarrow \eta_i\gamma_i$.

After aggregation, the deviation between global model parameters in benign and backdoored training process can be bounded. Note that the benign local model updates are cancelled out since they are the same in the two training process.

Lemma 8. *The deviation between the aggregated global model in the benign training process and the global model in the backdoored training process at round $\mathsf{t}_{\mathsf{adv}}$ is*

$$\|w_{\mathsf{t}_{\mathsf{adv}}} - w'_{\mathsf{t}_{\mathsf{adv}}}\|^2 = R \sum_{i=1}^{R} (\gamma_i p_i \Delta^i_{\tau_i})^2. \tag{18}$$

Proof.

$$\begin{aligned}
&\left\|w_{\mathsf{t}_{\mathsf{adv}}} - w'_{\mathsf{t}_{\mathsf{adv}}}\right\|^2 \\
&= \left\| \sum_{i=1}^{R} p_i\gamma_i (w^i_{\mathsf{t}_{\mathsf{adv}}\tau_i} - w_{t-1}) - \sum_{i=1}^{R} p_i\gamma_i ({w'}^i_{\mathsf{t}_{\mathsf{adv}}\tau_i} - w_{t-1}) \right\|^2 \\
&= \left\| \sum_{i=1}^{R} p_i\gamma_i \left(w^i_{\mathsf{t}_{\mathsf{adv}}\tau_i} - {w'}^i_{\mathsf{t}_{\mathsf{adv}}\tau_i} \right) \right\|^2 \\
&= \left\| \sum_{i=1}^{R} p_i\gamma_i \Delta^i_{\tau_i} \right\|^2 \\
&\leq R \sum_{i=1}^{R} \left(p_i\gamma_i \Delta^i_{\tau_i} \right)^2,
\end{aligned}$$

where we use the fact from linear algebra that $\| \sum_{i=1}^{R} a_i\|^2 \leq R \sum_{i=1}^{R} \|a_i\|^2$. □

Lemma 9. *Under Assumption 1, 2, 3 and the condition $\eta_i \leq \frac{1}{\beta}$, we have*

$$D_{KL}(\mu_{\mathsf{t}_{\mathsf{adv}}} \| \mu'_{\mathsf{t}_{\mathsf{adv}}}) \leq \frac{2R\sum_{i=1}^{R}\left(p_i\gamma_i\tau_i\eta_i\,\frac{q_{Bi}}{n_{Bi}}L_{\mathcal{Z}}\|\delta_i\|\right)^2}{\sigma^2_{\mathsf{t}_{\mathsf{adv}}}}. \tag{19}$$

Proof. Plugging Lemma 6 and Lemma 7 into Lemma 8, we have:

$$\left\|w_{\mathsf{t}_{\mathsf{adv}}} - w'_{\mathsf{t}_{\mathsf{adv}}}\right\|^2 \leq R\sum_{i=1}^{R}(2p_i\gamma_i\tau_i\eta_i\frac{q_{Bi}}{n_{Bi}}L_{\mathcal{Z}}\|\delta_i\|)^2. \tag{20}$$

Plugging Eq. 20 to. Eq. 14, it is clear that the divergence of noisy global model parameters between the benign and backdoor training process at round $\mathsf{t}_{\mathsf{adv}}$ is bounded. □

B.4. Analysis for $t > \mathsf{t}_{\mathsf{adv}}$

Now we focus on the contraction coefficient $\eta_f(K_t)$ when $t{>}\mathsf{t}_{\mathsf{adv}}$.

Lemma 10. *Based on Lemma 2 and 3, under Assumption 3, we have*

$$\eta_{TV}(K_t) \leq 2\Phi\left(\frac{\rho_t}{\sigma_t}\right) - 1. \tag{21}$$

Proof.

$$\begin{aligned}
&\eta_{TV}(K_t) := \sup_{w_1,w_2\in W} D_{TV}(K_t(w_1)\|K_t(w_2)) \\
&\leq \sup_{w_1,w_2\in W} D_{TV}\left(\mathcal{N}\Big(\mathrm{Clip}_{\rho_t}(\Psi(w_1)),\sigma_t^2\mathbf{I}\Big)\|\mathcal{N}\Big(\mathrm{Clip}_{\rho_t}(\Psi(w_2)),\sigma_t^2\mathbf{I}\Big)\right) \\
&= \sup_{w_3,w_4\in ball(\rho_t)} D_{TV}\left(\mathcal{N}\Big(w_3,\sigma_t^2\mathbf{I}\Big)\|\mathcal{N}\Big(w_4,\sigma_t^2\mathbf{I}\Big)\right) \\
&= \sup_{w_3,w_4\in ball(\rho_t)} 2\Phi\left(\frac{\|w_3-w_4\|}{2\sigma_t}\right) - 1 \\
&= 2\Phi\left(\frac{\rho_t}{\sigma_t}\right) - 1. && \triangleright \textit{ the norm of model parameters is bounded by } \rho_t
\end{aligned}$$

□

Finally, we obtain the divergence of global model in round T. We restate our Theorem 2 here.

Theorem 2. *When $\eta_i \leq \frac{1}{\beta}$ and Assumptions 1, 2, and 3 hold, the KL divergence between $\mu(\mathcal{M}(D))$ and $\mu(\mathcal{M}(D'))$ with $\mu(w) = \mathcal{N}(w, {\sigma_T}^2\mathbf{I})$ is bounded as:*

$$D_{KL}(\mu(\mathcal{M}(D))||\mu(\mathcal{M}(D'))) \leq \frac{2R\sum_{i=1}^{R}\left(p_i\gamma_i\tau_i\eta_i\frac{q_{Bi}}{n_{Bi}}L_{\mathcal{Z}}\|\delta_i\|\right)^2}{\sigma^2_{\mathsf{t}_{\mathsf{adv}}}}\prod_{t=\mathsf{t}_{\mathsf{adv}}+1}^{T}\left(2\Phi\left(\frac{\rho_t}{\sigma_t}\right)-1\right)$$

Proof.

$$
\begin{aligned}
&D_{KL}(\mu(\mathcal{M}(D))||\mu(\mathcal{M}(D'))) = D_{KL}(\mu_T||\mu'_T) \\
&\le D_{KL}(\mu_{\mathsf{t}_{\mathsf{adv}}}||\mu'_{\mathsf{t}_{\mathsf{adv}}}) \prod_{t=\mathsf{t}_{\mathsf{adv}}+1}^{T} \eta_{KL}(K_t) && \triangleright \textit{because of Eq. 13} \\
&\le D_{KL}(\mu_{\mathsf{t}_{\mathsf{adv}}}||\mu'_{\mathsf{t}_{\mathsf{adv}}}) \prod_{t=\mathsf{t}_{\mathsf{adv}}+1}^{T} \eta_{TV}(K_t) && \triangleright \textit{because of Lemma 4} \\
&\le \frac{2R\sum_{i=1}^{R}\left(p_i\gamma_i\tau_i\eta_i\left\|\mathcal{B}^i\right\|\right)^2}{\sigma^2_{\mathsf{t}_{\mathsf{adv}}}} \prod_{t=\mathsf{t}_{\mathsf{adv}}+1}^{T}\left(2\Phi\left(\frac{\rho_t}{\sigma_t}\right)-1\right) && \triangleright \textit{because of Lemma 9 and 10} \\
&\le \frac{2R\sum_{i=1}^{R}\left(p_i\gamma_i\tau_i\eta_i\frac{q_{B_i}}{n_{B_i}}L_{\mathcal{Z}}\|\delta_i\|\right)^2}{\sigma^2_{\mathsf{t}_{\mathsf{adv}}}} \prod_{t=\mathsf{t}_{\mathsf{adv}}+1}^{T}\left(2\Phi\left(\frac{\rho_t}{\sigma_t}\right)-1\right). && \triangleright \textit{because of Lemma 7}
\end{aligned}
$$

□

B.5. Proof of Lemma 5

We first introduce a new lemma, which will be used to prove Lemma 5.

Lemma 11. *Under Assumption 1 on convexity and smoothness, we have*

$$
\left\|g_i\left(w^i_{\underline{s}}\right)-{g'}_i\left(w'^i_{\underline{s}}\right)\right\|^2 \le 2\beta\left\langle\Delta^i_{\underline{s}}, g_i\left(w^i_{\underline{s}}\right)-g_i\left(w'^i_{\underline{s}}\right)\right\rangle + 2\left\|{g'}_i\left(w'^i_{\underline{s}}\right)-g_i\left(w'^i_{\underline{s}}\right)\right\|^2. \tag{22}
$$

Proof.

$$
\begin{aligned}
&\left\|g_i\left(w^i_{\underline{s}}\right)-{g'}_i\left(w'^i_{\underline{s}}\right)\right\|^2 \\
&= \left\|\left[g_i\left(w^i_{\underline{s}}\right)-g_i\left(w'^i_{\underline{s}}\right)\right]-\left[{g'}_i\left(w'^i_{\underline{s}}\right)-g_i\left(w'^i_{\underline{s}}\right)\right]\right\|^2 \\
&\le 2\left\|g_i\left(w^i_{\underline{s}}\right)-g_i\left(w'^i_{\underline{s}}\right)\right\|^2 + 2\left\|{g'}_i\left(w'^i_{\underline{s}}\right)-g_i\left(w'^i_{\underline{s}}\right)\right\|^2 \\
&\le 2\beta\left\langle\Delta^i_{\underline{s}}, g_i\left(w^i_{\underline{s}}\right)-g_i\left(w'^i_{\underline{s}}\right)\right\rangle + 2\left\|{g'}_i\left(w'^i_{\underline{s}}\right)-g_i\left(w'^i_{\underline{s}}\right)\right\|^2. && \triangleright \textit{because of Assumption 1}
\end{aligned}
$$

□

Next we provide the proof of Lemma 5.

Proof of Lemma 5. When $\eta_i \le \frac{1}{\beta}$,

$$
\begin{aligned}
&{\Delta^i_{\underline{s}+1}}^2 \triangleq \left\|w^i_{\underline{s}+1}-{w'}^i_{\underline{s}+1}\right\|^2 \\
&= \left\|(w^i_{\underline{s}}-{w'}^i_{\underline{s}})-\eta_i\left[g_i\left(w^i_{\underline{s}}\right)-{g'}_i\left(w'^i_{\underline{s}}\right)\right]\right\|^2 \\
&= {\Delta^i_{\underline{s}}}^2+\eta_i^2\left\|g_i\left(w^i_{\underline{s}}\right)-{g'}_i\left(w'^i_{\underline{s}}\right)\right\|^2 - 2\eta_i\left\langle w^i_{\underline{s}}-{w'}^i_{\underline{s}}, g_i\left(w^i_{\underline{s}}\right)-{g'}_i\left(w'^i_{\underline{s}}\right)\right\rangle \\
&= {\Delta^i_{\underline{s}}}^2+\eta_i^2\left\|g_i\left(w^i_{\underline{s}}\right)-{g'}_i\left(w'^i_{\underline{s}}\right)\right\|^2 + 2\eta_i\left\langle w^i_{\underline{s}}-{w'}^i_{\underline{s}}, {g'}_i\left(w'^i_{\underline{s}}\right)-g_i\left(w'^i_{\underline{s}}\right)\right\rangle - 2\eta_i\left\langle w^i_{\underline{s}}-{w'}^i_{\underline{s}}, g_i\left(w^i_{\underline{s}}\right)-g_i\left(w'^i_{\underline{s}}\right)\right\rangle \\
&\le {\Delta^i_{\underline{s}}}^2+2\eta_i^2\left\|{g'}_i\left(w'^i_{\underline{s}}\right)-g_i\left(w'^i_{\underline{s}}\right)\right\|^2 + 2\eta_i\left\langle w^i_{\underline{s}}-{w'}^i_{\underline{s}}, {g'}_i\left(w'^i_{\underline{s}}\right)-g_i\left(w'^i_{\underline{s}}\right)\right\rangle + (2\beta\eta_i^2-2\eta_i)\left\langle w^i_{\underline{s}}-{w'}^i_{\underline{s}}, g_i\left(w^i_{\underline{s}}\right)-g_i\left(w'^i_{\underline{s}}\right)\right\rangle \\
&\qquad \triangleright \textit{because of Lemma 11} \\
&\le {\Delta^i_{\underline{s}}}^2+2\eta_i^2\left\|{g'}_i\left(w'^i_{\underline{s}}\right)-g_i\left(w'^i_{\underline{s}}\right)\right\|^2 + 2\eta_i\left\langle w^i_{\underline{s}}-{w'}^i_{\underline{s}}, {g'}_i\left(w'^i_{\underline{s}}\right)-g_i\left(w'^i_{\underline{s}}\right)\right\rangle && \triangleright \textit{because of } \eta_i \le \tfrac{1}{\beta} \\
&\le {\Delta^i_{\underline{s}}}^2+2\eta_i^2\left\|{g'}_i\left(w'^i_{\underline{s}}\right)-g_i\left(w'^i_{\underline{s}}\right)\right\|^2 + 2\eta_i\Delta^i_{\underline{s}}\left\|{g'}_i\left(w'^i_{\underline{s}}\right)-g_i\left(w'^i_{\underline{s}}\right)\right\| && \triangleright \textit{because of } \langle a,b\rangle \le \|a\|\|b\| \\
&= {\Delta^i_{\underline{s}}}^2+2\eta_i\left\|\mathcal{B}^i\right\|\Delta^i_{\underline{s}}+2\eta_i^2\left\|\mathcal{B}^i\right\|^2. && \triangleright \textit{because of the definition } \mathcal{B}^i \triangleq {g'}_i(w)-g_i(w)
\end{aligned}
$$

□

B.6. Proof of Lemma 1

We first restate our Lemma 1 here and then provide the detailed proof.

Lemma 1. *Given the upper bound on model parameters norm, i.e., $\|w\| \leq \rho$, and two data samples z_1 and z_2 with $x_1 \neq x_2$ ($y_1 = y_2$), for multi-class logistic regression (i.e., one linear layer followed by a softmax function and trained by cross-entropy loss), its Lipschitz gradient constant w.r.t data is $L_{\mathcal{Z}} = \sqrt{2 + 2\rho + \rho^2}$. That is,*

$$\|\nabla \ell(w; z_1) - \nabla \ell(w; z_2)\| \leq \sqrt{2 + 2\rho + \rho^2}\|z_1 - z_2\|.$$

Proof. Given model parameters W of one linear layer, data samples $z = \{x, y\}$ and $z' = \{x', y\}$, we denote their loss as $\ell(W; z)$ and $\ell(W; z')$, where $x \in \mathbb{R}^{1\times d_x}$, $W \in \mathbb{R}^{d_x \times C}$. $Y \in \mathbb{R}^{1\times C}$ is a one-hot vector for C classes where $Y_i = \mathbb{1}\{i = y\}$. For x, we denote xW as the output of the linear layer, $P_i(x) = \mathrm{softmax}(xW)_i$ as the normalized probability for class i (the output of the softmax function). The cross-entropy loss is calculated as

$$\ell(x) = -\sum_i Y_i \log P_i(x) = -\sum_i Y_i \log \mathrm{softmax}(xW)_i. \tag{23}$$

We define $G \in \mathbb{R}^{d_x \times C}$ as the gradient for one sample:

$$G(x) = \nabla \ell(W; \{x, y\}) = \frac{d\ell}{dW}(x) = x^\top (P(x) - Y), \tag{24}$$

and we define G' as

$$G(x') = \nabla \ell(W; \{x', y\}) = \frac{d\ell}{dW}(x') = x'^\top (P(x') - Y). \tag{25}$$

According to the mean value theorem (Rudin, 1976), for a continuous vector-valued function $f : [a, b] \to \mathbb{R}^k$ differentiable on (a, b), there exist $c \in (a, b)$ such that

$$\frac{\|f(b) - f(a)\|}{b - a} \leq \|f'(c)\|. \tag{26}$$

Because x is normalized to $[0, 1]$ (a common dataset pre-processing method), when we define $G_l(t) = G(x' + t(x - x')), t \in [0, 1]$, based on the mean value theorem we have

$$\begin{aligned}
\|G(x) - G(x')\| &= \|G_l(1) - G_l(0)\| \\
&\leq \left\| \frac{dG_l}{dt}(t_0) \right\| (1 - 0) \\
&= \left\| \frac{dG}{dx}(\xi) \odot (x - x') \right\| \\
&\leq \left\| \frac{dG}{dx}(\xi) \right\| \|x - x'\|
\end{aligned}$$

where $\xi = x' + t_0(x - x'), t_0 \in [0, 1]$, $\frac{dG}{dx}(\xi)$ is a 3 dimension tenosr and $\odot$ is tensor product. We reduce the computation to 2 dimension matrix for simplification. Let G_i denote the ith colunm of matrix G (the gradient w.r.t W_i). Let $\mathbf{1}_i$ denote a row vector where i-th element is 1 and the others is 0. We have

$$
\begin{aligned}
&\|G(x) - G(x')\| \\
&\le \left\| \frac{dG}{dx}(\xi) \right\| \|x - x'\| \\
&= \sqrt{\sum_i^C \left\| \frac{dG_i}{dx}(\xi) \right\|^2} \|x - x'\| \\
&= \sqrt{\sum_i^C \left\| \frac{dx^\top (P_i - Y_i)}{dx}(\xi) \right\|^2} \|x - x'\| && \triangleright \textit{as } G_i(x) = x^\top (P_i(x) - Y_i) \\
&= \sqrt{\sum_i^C \left\| \frac{dx^\top}{dx}(\xi)(P_i - Y_i) + x^\top \frac{d(P_i - Y_i)}{dx}(\xi) \right\|^2} \|x - x'\| \\
&= \sqrt{\sum_i^C \|(P_i(\xi) - Y_i)I + x^\top (P_i(\xi)\mathbf{1}_i - P_i(\xi)P(\xi))W^\top\|^2} \|x - x'\| \\
&&& \triangleright \textit{as } \tfrac{d(P_i - Y_i)}{dx} = \tfrac{d\mathrm{softmax}(xW)_i}{dx} = (P_i\mathbf{1}_i - P_iP)W^\top \\
&\le \sqrt{\sum_i^C \|(P_i - Y_i)\|^2 + 2\|(P_i - Y_i)\|\|x^\top (P_i\mathbf{1}_i - P_iP)W^\top\| + \|x^\top (P_i\mathbf{1}_i - P_iP)W^\top\|^2} \|x - x'\| \\
&&& \triangleright \textit{denote } P_i \textit{ as } P_i(\xi) \textit{ for simplicity} \\
&\le \sqrt{\sum_i^C \|(P_i - Y_i)\| + 2||x^\top (P_i\mathbf{1}_i - P_iP)W^\top\| + \|x^\top (P_i\mathbf{1}_i - P_iP)W^\top\|^2} \|x - x'\|, && \triangleright \textit{as } \|(P_i - Y_i)\| \le 1 \\
&\le \sqrt{\sum_i^C \|(P_i - Y_i)\| + 2P_i\|x\|\|(\mathbf{1}_i - P)W^\top\| + P_i^2\|x\|^2\|(\mathbf{1}_i - P)W^\top\|^2} \|x - x'\| \\
&\le \sqrt{\sum_i^C \|(P_i - Y_i)\| + 2P_i\|W\| + P_i\|W\|^2} \|x - x'\|, && \triangleright \textit{as } \|x\| \le 1 \textit{ and } 0 \le P_i \le 1 \\
&\le \sqrt{\sum_i^C \|(P_i - Y_i)\| + 2P_i\rho + P_i^2\rho^2} \|x - x'\|, && \triangleright \textit{as } \|W\| \le \rho \\
&\le \sqrt{2 + 2\rho + \rho^2} \|x - x'\|.
\end{aligned}
$$

□

C. Proofs of Parameter Smoothing

In this section, we explain our parameter smoothing for general f-divergence, and give closed-form certification for KL divergence, which corresponds to the proofs for our Theorems 3.

C.1. General Framework for Robustness Certification

Consider a classifier $h : (\mathcal{W}, \mathcal{X}) \to \mathcal{Y}$. The output of the classifier depends on both the test input and its model parameters (i.e., model weights) of this classifier. In the testing phase, the model weight w is fixed, just like x_{test}, so it can be seen as an argument for the classifier h. For example, in a one-linear-layer model, $h(w; x_{test}) = \mathrm{softmax}(w \times x_{test})$, where $\times$ is the multiplication operation; in a one-conv-layer model, $h(w; x_{test}) = \mathrm{softmax}(w \circledast x_{test})$ where $\circledast$ is the convolution operation. In a model with multiple layers, the expression of model prediction $h(w; x_{test})$ also holds, where w consists of the weights from all layers. To our best knowledge, this is the first work to study *parameter* smoothing on w rather than input smoothing on x_{test}.

We want to verify the robustness of smoothed multi-class classifier. Recall that we smooth the classifier $h : (\mathcal{W}, \mathcal{X}) \to \mathcal{Y}$

with finite set of label $\mathcal{Y}$ using a smoothing measure $\mu : \mathcal{W} \mapsto \mathcal{P}(\mathcal{W})$. The resulting randomly smoothed classifier h_s is

$$h_s(w; x_{test}) = \arg\max_{c \in \mathcal{Y}} \mathbb{P}_{W \sim \mu(w)}[h(W; x_{test}) = c] \tag{27}$$

Our goal is to certify that the prediction $h_s(w; x_{test})$ is robust to model parameters perturbations of size at most ϵ measured by some distance function d, i.e.,

$$h_s(w'; x_{test}) = h_s(w; x_{test}) \; \forall w' \text{ such that } d(w, w') \leq \epsilon \tag{28}$$

We assume $\mathcal{W} \subseteq \mathbb{R}^d$ (a d dimensional model parameters space). Our framework involves a reference measure $\rho = \mu(w)$, the set of perturbed distributions $\mathcal{D}_{w,\epsilon} = \{\mu(w') : d(w, w') \leq \epsilon\}$, and a set of specifications $\phi : (\mathcal{W}, \mathcal{X}) \to \mathcal{Z} \subseteq \mathbb{R}$. Specifically, let $c = h_s(w; x_{test})$. Since we are working on the multi-class classification problem, for every pair of classes $\{c, c'\}$ where $c' \in \mathcal{Y} \setminus \{c\}$, we need a ϕ, which is a generic function over the model parameters space that we want to verify has robustness properties. Following (Dvijotham et al., 2020), for every $c' \in \mathcal{Y} \setminus \{c\}$, we define a specification $\phi_{c,c'} : (\mathcal{W}, \mathcal{X}) \mapsto \{-1, 0, +1\}$ as follows:

$$\phi_{c,c'}(w) = \begin{cases} +1 & \text{if } h(w; x_{test}) = c \\ -1 & \text{if } h(w; x_{test}) = c' \\ 0 & \text{otherwise} \end{cases} \tag{29}$$

where we denote $\phi_{c,c'}(w; x_{test})$ as $\phi_{c,c'}(w)$ for simplicity.

Proposition 1. *The smoothed classifier h_s is robustly certified, i.e., Eq. 28 holds, if and only if for every $c' \in \mathcal{Y} \setminus \{c\}$, $\phi_{c,c'}$ is robustly certified at $\mu(w)$ w.r.t $\mathcal{D}_{w,\epsilon}$. Verifying that a given specification ϕ is robustly certified is equivalent to checking if the optimal value of the following optimization problem is non-negative:*

$$OPT(\phi, \rho, \mathcal{D}_{w,\epsilon}) := \min_{\nu \in \mathcal{D}_{w,\epsilon}} \mathbb{E}_{W' \sim \nu}(\phi(W')) \tag{30}$$

Proof. Note that for any perturbed distribution $\nu \in \mathcal{D}_{w,\epsilon}$, according to the definition of expectation and Eq. 29, we have

$$\mathbb{E}_{W' \sim \nu}[\phi_{c,c'}(W')] = \mathbb{P}_{W' \sim \nu}[h(W'; x_{test}) = c] - \mathbb{P}_{W' \sim \nu}[h(W'; x_{test}) = c']. \tag{31}$$

Therefore, $\mathbb{E}_{W' \sim \nu}[\phi_{c,c'}(W')] \geq 0$ for all $c' \in \mathcal{Y} \setminus \{c\}$ is equivalent to $c = \arg\max_{y \in \mathcal{C}} \mathbb{P}_{W' \sim \nu}[h(W'; x_{test}) = y]$. For $\nu = \mu(w')$, this means that $h_s(w'; x_{test}) = c$. In other words, $\mathbb{E}_{W' \sim \nu}[\phi_{c,c'}(W')] \geq 0$ for all $c' \in \mathcal{Y} \setminus \{c\}$ and all $\nu = \mu(w') \subset \mathcal{D}_{w,\epsilon}$ if and only if $h_s(w'; x_{test}) = c$ for all w' such that $d(w, w') \leq \epsilon$, proving the required robustness certificate. □

Then we define the certification problem [2]:

Definition 1. *Given a reference distribution $\rho \in \mathcal{P}(\mathcal{W})$, probabilities p_A ,p_B that satisfy $p_A, p_B \geq 0$, $p_A + p_B \leq 1$, we define the class of specifications S:*

$$S = \{\phi : (\mathcal{W}, \mathcal{X}) \mapsto \{-1, 0, +1\} \text{ s.t. } \mathbb{P}_{W \sim \rho}[\phi(W) = +1] \geq p_A, \mathbb{P}_{W \sim \rho}[\phi(W) = -1] \leq p_B\} \tag{32}$$

Given the above definition of S, we can rewrite Proposition 1 as:

Proposition 2. *The smoothed classifier h_s is robustly certified, i.e., Eq. 28 holds, if and only if S is robustly certified at $\mu(w)$ w.r.t $\mathcal{D}_{w,\epsilon}$. Verifying that S is robustly certified is equivalent to checking if the condition $\mathbb{E}_{W' \sim \nu}[\phi(W')] \geq 0$ holds for all $\nu \in \mathcal{D}_{w,\epsilon}$ and $\phi \in S$.*

We need to provide guarantees that hold simultaneously over a whole class of specifications ($\phi_{c,c'}$ for all $c' \in \mathcal{Y} \setminus \{c\}$). In fact, p_A can be the seen as the "votes" for the top-one class c, and p_B can be seen as the "votes" for the runner-up class. We note that the function $f(\cdot)$ used in f-divergence is convex. As shown in (Dvijotham et al., 2020) (but for input smoothing), for perturbation sets $\mathcal{D}_{w,\epsilon} = \{\mu(w') : d(w, w') \leq \epsilon\} = \{\nu : D_f(\nu \| \mu(w)) \leq \epsilon\}$ specified by a f-divergence D_f bound ϵ, this certification task can be solved efficiently using convex optimization.

[2] It is called information-limited robust certification in (Dvijotham et al., 2020) for input smoothing.

Theorem 4. *Let D_f be f-divergence, ϵ be the divergence constraint, S, p_A, p_B be as in Definition 1. The smoothed classifier h_s is robustly certified at reference distribution ρ with respect to $\mathcal{D}_{w,\epsilon} = \{\nu : D_f(\nu\|\rho) \leq \epsilon\}$ if and only if the optimal value of the following convex optimization problem is non-negative:*

$$\max_{\lambda\geq 0,\kappa} \kappa - \lambda\epsilon - p_A f_\lambda^*(\kappa - 1) - p_B f_\lambda^*(\kappa + 1) - (1 - p_A - p_B) f_\lambda^*(\kappa) \geq 0 \tag{33}$$

Proof. We prove the theorem according to Proposition 2. Let $\rho(W)$ be the clean model parameters distribution, $\nu(W)$ be the perturbed model parameters distribution, $r(W) = \frac{\nu(W)}{\rho(W)}$ be likelihood ratio. We have

$$\begin{aligned} \mathbb{E}_{W\sim\nu}[\phi(W)] &= \mathbb{E}_{W\sim\rho}[r(W)\phi(W)], \\ D_f(\nu\|\rho) &= \mathbb{E}_{W\sim\rho}[f(r(W))], \\ \mathbb{E}_{W\sim\rho}[r(W)] &= 1. \end{aligned} \tag{34}$$

The third condition is obtained using the fact that ν is a probability measure. The optimization over ν, which is equivalent to optimizing over r, can be written as

$$\begin{aligned} &\min_{r\geq 0} \mathbb{E}_{W\sim\rho}[r(W)\phi(W)] \\ &s.t.\ \mathbb{E}_{W\sim\rho}[f(r(W))] \leq \epsilon, \mathbb{E}_{W\sim\rho}[r(W)] = 1 \end{aligned} \tag{35}$$

We solve the optimization using Lagrangian duality as follows. We first dualize the constraints on r (Dvijotham et al., 2020) to obtain

$$\begin{aligned} &\min_{r\geq 0} \mathbb{E}_{W\sim\rho}[r(W)\phi(W)] + \lambda(\mathbb{E}_{W\sim\rho}[f(r(W))] - \epsilon) + \kappa(1 - \mathbb{E}_{W\sim\rho}[r(W)]) \\ &= \min_{r\geq 0} \mathbb{E}_{W\sim\rho}[r(W)\phi(W) + \lambda f(r(W)) - \kappa r(W)] + \kappa - \lambda\epsilon \\ &= \kappa - \lambda\epsilon - \mathbb{E}_{W\sim\rho}[\max_{r\geq 0} \kappa r(W) - r(W)\phi(W) - \lambda f(r(W))] \\ &= \kappa - \lambda\epsilon - \mathbb{E}_{W\sim\rho}[\max_{r\geq 0} r(W)(\kappa - \phi(W)) - \lambda f(r(W))] \\ &= \kappa - \lambda\epsilon - \mathbb{E}_{W\sim\rho}[\max_{r\geq 0} r(W)(\kappa - \phi(W)) - f_\lambda(r(W))] \\ &\leq \kappa - \lambda\epsilon - \mathbb{E}_{W\sim\rho}[f_\lambda^*(\kappa - \phi(W))] \end{aligned} \tag{36}$$

where $f_\lambda^*(u) = \max_{v\geq 0}(uv - f_\lambda(v))$, $f_\lambda(v) = \lambda f(v)$. By strong duality, maximizing the final expression in Eq. 36 with respect to $\lambda \geq 0, \kappa$ achieves the optimal value in Eq. 35. If the optimal value is non-negative, the specification S is robustly certified.

$$\max_{\lambda\geq 0,\kappa} \kappa - \lambda\epsilon - \mathbb{E}_{W\sim\rho}[f_\lambda^*(\kappa - \phi(W))] \tag{37}$$

We can plug in p_A, p_B defined in Definition 1:

$$\max_{\lambda\geq 0,\kappa} \kappa - \lambda\epsilon - p_A f_\lambda^*(\kappa - 1) - p_B f_\lambda^*(\kappa + 1) - (1 - p_A - p_B) f_\lambda^*(\kappa) \tag{38}$$

where $p_A = \mathbb{P}_{W\sim\rho}[\phi(W) = +1]$, $p_B = \mathbb{P}_{W\sim\rho}[\phi(W) = -1]$, $1 - p_A - p_B = \mathbb{P}_{W\sim\rho}[\phi(W) = 0]$, □

Remark. Note that our differences from (Dvijotham et al., 2020) are in two aspects: (1) Our certification is with respect to the smoothing scheme on model parameters W; (2) We concretize the corresponding Theorem 2 in (Dvijotham et al., 2020) by the explicit constraints on p_A, p_B.

C.2. Closed-form Certificate for KL Divergence

We instantize Theorem 4 with KL divergence.

Lemma 12. *Let D_{KL} be the KL divergence, ϵ be the divergence constraint, S, p_A, p_B be as in Definition 1. The smoothed classifier h_s is robustly certified at reference distribution ρ with respect to $\mathcal{D}_{w,\epsilon} = \{\nu : D_{KL}(\nu\|\rho) \leq \epsilon\}$ if and only if:*

$$\epsilon \leq -\log\left(1 - (\sqrt{p_A} - \sqrt{p_B})^2\right) \tag{39}$$

Proof for Lemma 12. The function $f(u) = ulog(u)$ for KL divergence is a convex function with $f(1) = 0$, then we have

$$f_\lambda^*(u) = \max_{v\geq 0}(uv - \lambda f(v)) = \max_{v\geq 0}(uv - \lambda v\log(v)).$$

Setting the derivative with respect to v to 0 and solving for v, we obtain $v = \exp\left(\frac{u-\lambda}{\lambda}\right), \lambda{>}0$. So we have

$$f_\lambda^*(u) = \lambda\exp\left(\frac{u}{\lambda} - 1\right). \tag{40}$$

Suppose we have a bound on the KL divergence $D_f(\nu\|\rho) \leq \epsilon$, then we want that the optimal certificate is non-negative:

$$\max_{\lambda>0,\kappa}\left(\kappa - \lambda\epsilon - p_A\lambda\exp\left(\frac{\kappa-1}{\lambda} - 1\right) - p_B\lambda\exp\left(\frac{\kappa+1}{\lambda} - 1\right) - (1-p_A-p_B)\lambda\exp\left(\frac{\kappa}{\lambda} - 1\right)\right) \geq 0. \tag{41}$$

Setting $y = \kappa/\lambda, z = \frac{1}{\lambda}(z{>}0)$, we can rewrite Eq. 41 as:

$$\max_{z>0,y}\left(\frac{1}{z}\Big(y - \epsilon - p_A\exp(y-z-1) - p_B\exp(y+z-1) - (1-p_A-p_B)\exp(y-1)\Big)\right) \geq 0. \tag{42}$$

Because $\frac{1}{z}$ is positive, we divide both the LHS and RHS by $\frac{1}{z}$ and our goal can be rewritten as:

$$\max_{z>0,y}\left(y - \epsilon - p_A\exp(y-z-1) - p_B\exp(y+z-1) - (1-p_A-p_B)\exp(y-1)\right) \geq 0. \tag{43}$$

Setting the derivative of the LHS with respect to z to 0 and solving for z, we obtain

$$\begin{aligned} p_A\exp(y-z-1) - p_B\exp(y+z-1) &= 0 \\ z &= \log(\sqrt{\frac{p_A}{p_B}}). \end{aligned} \tag{44}$$

Thus the LHS of Eq. 43 reduces to

$$\max_y\left(y - \epsilon - \left(1 - (\sqrt{p_A} - \sqrt{p_B})^2\right)\exp(y-1)\right). \tag{45}$$

Setting the derivative with respect to y to 0 and solving for y, we obtain

$$\begin{aligned} 1 - \left(1 - (\sqrt{p_A} - \sqrt{p_B})^2\right)\exp(y-1) &= 0 \\ y &= 1 - \log\left(1 - (\sqrt{p_A} - \sqrt{p_B})^2\right). \end{aligned} \tag{46}$$

Now the LHS of Eq. 43 reduces to

$$-\log\left(1 - (\sqrt{p_A} - \sqrt{p_B})^2\right) - \epsilon. \tag{47}$$

For this number to be positive, we need

$$\epsilon \leq -\log\left(1 - (\sqrt{p_A} - \sqrt{p_B})^2\right). \tag{48}$$

Hence, proved. □

Remark. The challenges are: 1) we divide both the LHS and RHS of Eq. 42 by $\frac{1}{z}$ to obtain Eq. 43, otherwise the derivative of the LHS of Eq. 42 cannot be calculated directly. Moreover, setting $y = \kappa/\lambda, z = \frac{1}{\lambda}$ makes it much easier to solve the optimization problem. 2) (Dvijotham et al., 2020) does not directly provide proof for KL Divergence. They proves the certification for Renyi Divergence and then regard KL as a special case of Renyi Divergence.

Finally, we restate our Theorem 3 here.

Theorem 3. *Let h_s be defined as in Eq. 1. Suppose $c_A \in \mathcal{Y}$ and $\underline{p_A}, \overline{p_B} \in [0,1]$ satisfy*

$$H_s^{c_A}(w'; x_{test}) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c \neq c_A} H_s^c(w'; x_{test}),$$

then $h_s(w'; x_{test}) = h_s(w; x_{test}) = c_A$ for all w such that $D_{KL}(\mu(w), \mu(w')) \leq \epsilon$, where

$$\epsilon = -\log\left(1 - (\sqrt{\underline{p_A}} - \sqrt{\overline{p_B}})^2\right)$$

Proof. We use Lemma 12 to prove Theorem 3. In practice, since the server does not know the global model in the current FL system is poisoned or not, we assume the model is already backdoored and derive the condition when its prediction will be certifiably consistent with the prediction of the clean model. Therefore, the reference distribution $\rho = \mu(w')$ and $\nu = \mu(w)$. Moreover, $H_s^{c_A}(w'; x_{test}) \geq \underline{p_A}$ is equivalent to $\mathbb{P}_{W\sim\rho}[\phi(W) = +1] \geq p_A$, and $\max_{c \neq c_A} H_s^c(w'; x_{test}) \leq \overline{p_B}$ is equivalent to $\mathbb{P}_{W\sim\rho}[\phi(W) = -1] \leq p_B$. Rewriting Lemma 12 leads to Theorem 3. □