A. Proofs

Lemma 3. *Let $\sigma : \mathbb{R} \to \mathbb{R}$ and $\mu : \mathbb{R} \to \mathbb{R}$ be given functions. The following facts hold:*

(i) if $\sigma \in L^\infty(\mathbb{R})$ and $\mu \in BV(\mathbb{R})$ then $\mu * \sigma \in W^{1,\infty}(\mathbb{R})$;

(ii) if $\sigma \in L^\infty(\mathbb{R})$ and $\mu \in W^{1,1}(\mathbb{R})$ then the weak derivative of $\mu * \sigma$ satisfies

$$D(\mu * \sigma)(x) = (D\mu * \sigma)(x)$$

for almost all $x \in \mathbb{R}$;

(iii) if $\sigma \in BV(\mathbb{R})$ and $\mu \in W^{1,1}(\mathbb{R})$ then $\mu * \sigma \in C^1(\mathbb{R})$, its derivative is uniformly continuous and one has

$$\frac{d(\mu * \sigma)}{dx}(x) = (D\mu * \sigma)(x)$$

for all $x \in \mathbb{R}$.

Proof. We first show (i). It is immediate to check that

$$\|\mu * \sigma\|_\infty \leq \|\mu\|_1 \, \|\sigma\|_\infty < +\infty \,. \tag{22}$$

We choose $x_1 < x_2 \in \mathbb{R}$ and, thanks to Fubini's theorem and a change of variable, we obtain the following estimate:

$$\begin{aligned}
&|(\mu * \sigma)(x_1) - (\mu * \sigma)(x_2)| \\
&\quad = \left| \int_{y \in \mathbb{R}} \big(\mu(x_1 - y) - \mu(x_2 - y)\big)\sigma(y)\, dy \right| \\
&\quad \leq \int_{y \in \mathbb{R}} |D\mu|\big([x_1 - y, x_2 - y]\big)\, |\sigma(y)|\, dy \\
&\quad \leq \|\sigma\|_\infty \int_{y \in \mathbb{R}} \int_{z \in \mathbb{R}} \chi_{[x_1, x_2]}(z + y)\, d|D\mu|(z)\, dy \\
&\quad = \|\sigma\|_\infty \int_{z \in \mathbb{R}} \int_{y \in \mathbb{R}} \chi_{[x_1, x_2]}(z + y)\, dy\, d|D\mu|(z) \\
&\quad \leq \|\sigma\|_\infty \, |D\mu|(\mathbb{R})\, |x_1 - x_2| \,.
\end{aligned}$$

This shows that $\mu * \sigma$ is a Lipschitz function (with Lipschitz constant bounded above by $\|\sigma\|_\infty \, |D\mu|(\mathbb{R})$). Therefore the proof of (i) follows from the Sobolev characterisation of Lipschitz functions combined with (22). Let us prove (ii) by showing that $\mu * \sigma$ is weakly differentiable, thus providing a pointwise almost everywhere representation of its weak derivative. Let $\phi \in C_c^\infty(\mathbb{R})$ be a given test function. By using Fubini's Theorem, the definition of weak derivative, and the change of variable in the integration, we obtain

$$\begin{aligned}
&\int_{\mathbb{R}} (\mu * \sigma)(x)\, \frac{d\phi}{dx}(x)\, dx \\
&\quad = \int_{x \in \mathbb{R}} \int_{y \in \mathbb{R}} \mu(x - y)\sigma(y)\, dy\, \frac{d\phi}{dx}(x)\, dx \\
&\quad = \int_{y \in \mathbb{R}} \int_{x \in \mathbb{R}} \mu(x - y) \frac{d\phi}{dx}(x)\, dx\, \sigma(y)\, dy \\
&\quad = -\int_{x \in \mathbb{R}} \int_{y \in \mathbb{R}} D\mu(x - y)\, \sigma(y)\, dy\, \phi(x)\, dx \\
&\quad = -\int_{x \in \mathbb{R}} (D\mu * \sigma)(x)\, \phi(x)\, dx \,.
\end{aligned}$$

This shows (ii). We finally prove (iii). By (i) and (ii) we already know that $\mu * \sigma \in W^{1,\infty}(\mathbb{R})$ and its weak derivative satisfies $D(\mu * \sigma)(x) = (D\mu * \sigma)(x)$ for almost all $x \in \mathbb{R}$. The conclusion is achieved as soon as we show that $(D\mu * \sigma)(x)$ is a continuous function. We have for $x_1 < x_2 \in \mathbb{R}$

$$\begin{aligned}
&|(D\mu * \sigma)(x_1) - (D\mu * \sigma)(x_2)| \\
&\quad = \left| \int_{y \in \mathbb{R}} D\mu(x_1 - y)\sigma(y)\, dy + \right. \\
&\qquad \left. - \int_{y \in \mathbb{R}} D\mu(x_2 - y)\sigma(y)\, dy \right| \\
&\quad = \left| \int_{z \in \mathbb{R}} D\mu(z)\, \big(\sigma(x_1 - z) - \sigma(x_2 - z)\big)\, dz \right| \\
&\quad \leq \int_{z \in \mathbb{R}} \int_{t \in \mathbb{R}} \chi_{[x_1, x_2]}(t + z)\, d|D\sigma|(t)\, |D\mu(z)| dz \\
&\quad = \int_{t \in \mathbb{R}} \int_{z \in \mathbb{R}} \chi_{[x_1, x_2]}(t + z)\, |D\mu(z)| dz\, d|D\sigma|(t) \\
&\quad = \int_{t \in \mathbb{R}} \int_{z \in \mathbb{R}} \chi_{[x_1, x_2]}(t + z)\, |D\mu(z)| dz\, d|D\sigma|(t) \,.
\end{aligned}$$

Denote by κ the non-negative, finite Borel measure defined by $d\kappa = |D\mu(z)| dz$. Since κ is absolutely continuous with respect to the Lebesgue measure, for all $\epsilon > 0$ there exists $\delta > 0$ such that $|x_1 - x_2| < \delta$ implies $\kappa([x_1, x_2]) < \epsilon$. Therefore we get

$$\begin{aligned}
&|(D\mu * \sigma)(x_1) - (D\mu * \sigma)(x_2)| \\
&\qquad \leq \int_{t \in \mathbb{R}} \kappa([x_1 - t, x_2 - t])\, d|D\sigma|(t) \\
&\qquad \leq \epsilon\, |D\sigma|(\mathbb{R})
\end{aligned}$$

as soon as $|x_1 - x_2| \leq \delta$, which proves the uniform continuity of $D\mu * \sigma$ and concludes the proof. □

Proof of Proposition 1

Proof. The first claim (i) follows from the definition of convolution. The proofs of (ii) and (iii) follow from the application of Lemma 3, noticing that, by definition, a quantiser (1) satisfies $\sigma \in L^\infty(\mathbb{R})$ (in particular, $\|\sigma\|_\infty = \max_{q \in Q}\{|q|\}$). □

Proof of Theorem 2

Proof. First, we note that

$$\|\mathbf{x}_{\hat{\lambda}_\ell,\ell} - \mathbf{x}_\ell\| \coloneqq \left(\sum_{i=1}^{n_\ell} |x_{\hat{\lambda}_\ell,\ell,i} - x_{\ell,i}|^2\right)^{\frac{1}{2}} \leq \sqrt{n_\ell} \max_{i\in\{1,2,\dots,n_\ell\}} \{|x_{\hat{\lambda}_\ell,\ell,i} - x_{\ell,i}|\}\,,$$

where $x_{\hat{\lambda}_\ell,\ell,i}$ and $x_{\ell,i}$ denote the i-th components of the ℓ-th layer regularised and quantised features, respectively. We define

$$\bar{i} \coloneqq \operatorname*{arg\,max}_{i\in\{1,2,\dots,n_\ell\}} \{|x_{\hat{\lambda}_\ell,\ell,i} - x_{\ell,i}|\}\,.$$

Therefore, since n_ℓ is arbitrary but finite, a sufficient condition for (19) is

$$\frac{|x_{\hat{\lambda}_\ell,\ell,\bar{i}} - x_{\ell,\bar{i}}|}{r^\ell(\lambda)} \xrightarrow[\lambda\to 0]{} 0\,. \tag{23}$$

To simplify the notation, in the following we will omit the subscript index $\bar{i}$. First, we conveniently rewrite (23) according to the definition of limit:

$$\begin{aligned}&\forall\,\varepsilon>0\,,\ \exists\,\tilde{\lambda}>0\ :\\ &\quad |x_{\hat{\lambda}_\ell,\ell} - x_\ell| < \varepsilon r_\ell(\lambda)\,,\ \forall\, 0<\lambda<\tilde{\lambda}\,.\end{aligned} \tag{24}$$

Then, we argue by induction.

To prove the base step ($\ell = 1$) we need to consider two cases: $x_1 = 0$ and $x_1 = 1$. First, we suppose $x_1 = \sigma(S_{\mathbf{m}_1}(\mathbf{x}_0)) = 0$; this implies that $S_{\mathbf{m}_1}(\mathbf{x}_0) < 0$. Property (14) implies $x_{\hat{\lambda}_1,1} \geq x_1$, which implies $|x_{\hat{\lambda}_1,1} - x_1| = x_{\hat{\lambda}_1,1} = \sigma_{\lambda_1}(S_{\mathbf{m}_1}(\mathbf{x}_0))$. Then, we can apply $\sigma_{\lambda_1}^{-1}$ to both sides of the inequality in (24) obtaining the following condition:

$$\begin{aligned}&\forall\,\varepsilon>0\,,\ \exists\,\tilde{\lambda}>0\ :\\ &\quad S_{\mathbf{m}^1}(\mathbf{x}^0) < \sigma_{\lambda_1}^{-1}(\varepsilon r_1(\lambda))\,,\ \forall\, 0<\lambda<\tilde{\lambda}\,,\end{aligned}$$

whose validity is guaranteed by hypothesis (15). Now, we analyse the case $x_1 = 1$. Property (14) implies $x_{\hat{\lambda}_1,1} \leq x_1$, hence $|x_{\hat{\lambda}_1,1} - x_1| = 1 - x_{\hat{\lambda}_1,1} = 1 - \sigma_{\lambda_1}(S_{\mathbf{m}_1}(\mathbf{x}_0))$. In this case, condition (24) becomes

$$\begin{aligned}&\forall\,\varepsilon>0\,,\ \exists\,\tilde{\lambda}>0\ :\\ &\quad 1 - \sigma_{\lambda_1}(S_{\mathbf{m}_1}(\mathbf{x}_0)) < \varepsilon r_1(\lambda)\,,\ \forall\, 0<\lambda<\tilde{\lambda}\,.\end{aligned} \tag{25}$$

We have two sub-cases: $S_{\mathbf{m}_1}(\mathbf{x}_0) > 0$ and $S_{\mathbf{m}_1}(\mathbf{x}_0) = 0$. In the first sub-case, by rearranging terms and applying $\sigma_{\lambda_1}^{-1}$ to both sides of (25), we derive the condition

$$\begin{aligned}&\forall\,\varepsilon>0\,,\ \exists\,\tilde{\lambda}>0\ :\\ &\quad \sigma_{\lambda_1}^{-1}(1-\varepsilon r_1(\lambda)) < S_{\mathbf{m}_1}(\mathbf{x}_0)\,,\ \forall\, 0<\lambda<\tilde{\lambda}\,,\end{aligned}$$

which is granted by hypothesis (16). In the second sub-case, we can divide both sides of (25) by $r_1(\lambda)$ and obtain the condition

$$\begin{aligned}&\forall\,\varepsilon>0\,,\ \exists\,\tilde{\lambda}>0\ :\\ &\quad \frac{1-\sigma_{\lambda_1}(0)}{r_1(\lambda)} < \varepsilon\,,\ \forall\, 0<\lambda<\tilde{\lambda}\,,\end{aligned}$$

which holds by hypothesis (17).

We now proceed to the inductive step ($\ell > 1$). We have two possibilities for x_ℓ:

(A) $x_\ell = 0$;

(B) $x_\ell = 1$.

We start with case (A). We observe that

$$\begin{aligned} s_{\hat{\lambda}_{\ell-1},\ell} - s_\ell &= S_{\mathbf{m}_\ell}(\mathbf{x}_{\hat{\lambda}_{\ell-1},\ell-1}) - S_{\mathbf{m}_\ell}(\mathbf{x}_{\ell-1}) \\ &= S_{\mathbf{m}_\ell}(\mathbf{x}_{\hat{\lambda}_{\ell-1},\ell-1} - \mathbf{x}_{\ell-1}) \xrightarrow[\lambda\to 0]{} 0\,, \end{aligned} \tag{26}$$

since $S_{\mathbf{m}_\ell}$ is linear and $\|\mathbf{x}_{\hat{\lambda}_{\ell-1},\ell-1} - \mathbf{x}_{\ell-1}\| \xrightarrow[\lambda\to 0]{} 0$ by the inductive hypothesis. With reference to H_0^+, case (A) implies that $s_\ell < 0$. Together with (26), this implies that

$$\begin{aligned}&\exists\,\lambda^* = \lambda^*(s_\ell) > 0\ :\\ &\quad s_{\hat{\lambda}_{\ell-1},\ell} < -\frac{|s_\ell|}{2} < 0\,,\ \forall\, 0<\lambda<\lambda^*\,.\end{aligned}$$

Since $x_\ell = 0$ and $x_{\hat{\lambda}_\ell,\ell} = \sigma_{\lambda_\ell}(s_{\hat{\lambda}_{\ell-1},\ell}) \geq 0$, condition (24) can be rewritten as

$$\begin{aligned}&\forall\,\varepsilon>0\,,\ \exists\,\tilde{\lambda}>0\ :\\ &\quad \sigma_{\lambda_\ell}(s_{\hat{\lambda}_{\ell-1},\ell}) < \varepsilon r_\ell(\lambda)\,,\ \forall\, 0<\lambda<\tilde{\lambda}\,.\end{aligned}$$

Due to the monotonicity of σ_{λ_ℓ}, we have $\sigma_{\lambda_\ell}(s_{\hat{\lambda}_{\ell-1},\ell}) < \sigma_{\lambda_\ell}(-|s_\ell|/2)\,,\ \forall\, 0<\lambda<\lambda^*$. Therefore, a sufficient condition to guarantee the convergence is that

$$\begin{aligned}&\forall\,\varepsilon>0\,,\ \exists\, 0<\tilde{\lambda}\leq\lambda^*\ :\\ &\quad -\frac{|s_\ell|}{2} < \sigma_{\lambda_\ell}^{-1}(\varepsilon r_\ell(\lambda))\,,\ \forall\, 0<\lambda<\tilde{\lambda}\,.\end{aligned}$$

This condition is granted for every $s_\ell < 0$ by (15). We now move to case (B). This case ($x_\ell = 1$) might originate from two sub-cases:

(i) $s_\ell > 0$;

(ii) $s_\ell = 0$.

The proof of sub-case (i) is similar to the proof for case (A). Given $s_\ell > 0$, since $s_{\hat{\lambda}_{\ell-1},\ell} \xrightarrow[\lambda\to 0]{} s_\ell$ by the inductive hypothesis, we have that

$$\begin{aligned}&\exists\,\lambda^* = \lambda^*(s_\ell) > 0\ :\\ &\quad 0 < \frac{s_\ell}{2} < s_{\hat{\lambda}_{\ell-1},\ell}\,.\end{aligned}$$

Then, since $x_\ell = 1$ and $x_\ell = \sigma_{\lambda_\ell}(s_{\hat{\lambda}_{\ell-1},\ell}) \leq 1$, we can rewrite (24) as

$$\forall\, \varepsilon > 0\,,\ \exists\, \tilde{\lambda} > 0\ :\\ 1 - \sigma_{\lambda_\ell}(s_{\hat{\lambda}_{\ell-1},\ell}) < \varepsilon r_\ell(\lambda)\,,\ \forall\, 0 < \lambda < \tilde{\lambda}\,.$$

Since $\sigma_{\lambda_\ell}(s_{\hat{\lambda}_{\ell-1},\ell}) > \sigma_{\lambda_\ell}(s_\ell/2)$, a sufficient condition to get convergence is that

$$\forall\, \varepsilon > 0\,,\ \exists\, 0 < \tilde{\lambda} < \lambda^*\ :\\ \frac{s_\ell}{2} > \sigma_{\lambda_\ell}^{-1}(1 - \varepsilon r_\ell(\lambda))\,,\ \forall\, 0 < \lambda < \tilde{\lambda}\,.$$

This is guaranteed for every $s_\ell > 0$ by (15). Case (ii) is more delicate, since $s_{\hat{\lambda}_{\ell-1},\ell}$ can be positioned in two ways with respect to $s_\ell = 0$:

(a) $\lambda > 0$ is such that $s_{\hat{\lambda}_{\ell-1},\ell} \geq 0$;

(b) $\lambda > 0$ is such that $s_{\hat{\lambda}_{\ell-1},\ell} < 0$.

In case (a), it is sufficient to note that the monotonicity of σ_{λ_ℓ} implies $1 - \sigma_{\lambda_\ell}(s_{\hat{\lambda}_{\ell-1},\ell}) \leq 1 - \sigma_{\lambda_\ell}(0)$, since $s_{\hat{\lambda}_{\ell-1},\ell} \geq 0$. Then, condition (24) can be rewritten as

$$\forall\, \varepsilon > 0\,,\ \exists\, \tilde{\lambda} > 0\ :\\ \frac{1 - \sigma_{\lambda_\ell}(0)}{r_\ell(\lambda)} < \varepsilon\,,\ \forall\, 0 < \lambda < \tilde{\lambda}\,.$$

This is guaranteed by (17). To prove the last case (b), we first observe that

$$\begin{aligned} s_{\hat{\lambda}_{\ell-1},\ell} &= s_{\hat{\lambda}_{\ell-1},\ell} - s_\ell \\ &= \Big(\langle \mathbf{w}_\ell, \mathbf{x}_{\hat{\lambda}_{\ell-1},\ell-1}\rangle + b_\ell\Big) - (\langle \mathbf{w}_\ell, \mathbf{x}_{\ell-1}\rangle + b_\ell) \\ &= \langle \mathbf{w}_\ell, \mathbf{x}_{\hat{\lambda}_{\ell-1},\ell-1} - \mathbf{x}_{\ell-1}\rangle \end{aligned}$$

(which follows from $s_\ell = 0$) and apply the Cauchy-Schwartz inequality to obtain the following upper bound:

$$|s_{\hat{\lambda}_{\ell-1},\ell}| \leq \|\mathbf{w}_\ell\|\, \|\mathbf{x}_{\hat{\lambda}_{\ell-1},\ell-1} - \mathbf{x}_{\ell-1}\|\,. \tag{27}$$

Then, we rewrite (24) as

$$\forall\, \varepsilon > 0\,,\ \exists\, \tilde{\lambda} > 0\ :\\ -s_{\hat{\lambda}_{\ell-1},\ell} < -\sigma_{\lambda_\ell}^{-1}(1 - \varepsilon r_\ell(\lambda))\,,\ \forall\, 0 < \lambda < \tilde{\lambda}\,.$$

Observation (27) allows us to write a slightly stronger but sufficient condition for convergence:

$$\forall\, \varepsilon > 0\,,\ \exists\, \tilde{\lambda} > 0\ :\\ \|\mathbf{w}_\ell\|\, \|\mathbf{x}_{\hat{\lambda}_{\ell-1},\ell-1} - \mathbf{x}_{\ell-1}\| \leq -\sigma_{\lambda_\ell}^{-1}(1 - \varepsilon r_\ell(\lambda))\,,\\ \forall\, 0 < \lambda < \tilde{\lambda}\,,$$

where we used the fact that $-s_{\hat{\lambda}_{\ell-1},\ell} = |s_{\hat{\lambda}_{\ell-1},\ell}|$ (since $s_{\hat{\lambda}_{\ell-1},\ell} < 0$). The inner inequality can be rewritten as

$$\frac{\|\mathbf{x}_{\hat{\lambda}_{\ell-1},\ell-1} - \mathbf{x}_{\ell-1}\|}{-\sigma_{\lambda_\ell}^{-1}(1 - \varepsilon r_\ell(\lambda))} \leq \frac{1}{\|\mathbf{w}_\ell\|}\,; \tag{28}$$

since $\mathbf{w}_\ell$ is fixed but arbitrary (it is part of the parameter $\mathbf{m}_\ell$), the term on the right can be arbitrarily small, and therefore a sufficient condition to ensure (28) for λ small enough is

$$\forall\, \varepsilon > 0\,,\ \frac{\|\mathbf{x}_{\hat{\lambda}_{\ell-1},\ell-1} - \mathbf{x}_{\ell-1}\|}{-\sigma_{\lambda_\ell}^{-1}(1 - \varepsilon r_\ell(\lambda))} \xrightarrow[\lambda \to 0]{} 0\,. \tag{29}$$

By the inductive hypothesis (where we set $\varepsilon = 1$),

$$\exists\, \tilde{\lambda}_{\ell-1} > 0\ :\\ \|\mathbf{x}_{\hat{\lambda}_{\ell-1},\ell-1} - \mathbf{x}_{\ell-1}\| < r_{\ell-1}(\lambda)\,,\ \forall\, 0 < \lambda < \tilde{\lambda}_{\ell-1}\,.$$

Therefore, (18) enforces the convergence of the upper bound in the following inequality:

$$\frac{\|\mathbf{x}_{\hat{\lambda}_{\ell-1},\ell-1} - \mathbf{x}_{\ell-1}\|}{-\sigma_{\lambda_\ell}^{-1}(1 - \varepsilon r_\ell(\lambda))} \leq \frac{r_{\ell-1}(\lambda)}{-\sigma_{\lambda_\ell}^{-1}(1 - \varepsilon r_\ell(\lambda))}\,,$$

and therefore (29) follows. This completes the proof of the theorem. □

B. Other experiments

The experimental findings reported in Section 4 refer to a ternary VGG-like network solving CIFAR-10. To corroborate the validity of our findings, we performed additional experiments on two different scenarios, where we changed the data set, the network topology, and the quantisation policy.

Given the findings reported in Section 4, we used different noise types only when analysing static noise schedules. We constrained the noise type to uniform when analysing dynamic noise schedules.

As in the original CIFAR-10 experiments, we evaluated each hyper-parameter configuration using five-fold cross-validation on the training partitions of the chosen data sets.

B.1. SVHN

Street View House Numbers (SVHN) is an image classification data set [19]. It contains $\sim 99k$ RGB-encoded images representing decimal digits from house number plates. It consists of a training partition ($\sim 73k$ images) and a validation partition ($\sim 26k$ images).

We used the same VGG-like network from the CIFAR-10 experiments. Again, we quantised all the weights and features to be ternary, and we kept the weights of the last layer in floating-point format.

In each experimental unit, we trained the network for 500 epochs using mini-batches of 256 images, the cross-entropy loss function, and the ADAM optimiser with an initial learning rate of 10^{-3}, decreased to 10^{-4} after 400 epochs.

In agreement with the CIFAR-10 findings, Figures 4a 4b show that QNNs trained using static STE variants based on different noise types converge to the same accuracy. We note that the uniform noise type in combination with the random forward computation strategy seems to perform slightly worse during the earlier stages of training.

Figures 5a, 5b, 5c show that the quality of different decay interval strategies (as measured by the final accuracy of the trained networks) is better for those that are more coherent with the hypothesis of Theorem 2, namely the partition and same start strategies. Independently of the forward computation strategy, the same end decay interval strategy is still the worst amongst the tested ones.

B.2. GSC

Google Speech Commands (GSC) is a keyword spotting data set [25]. Keyword spotting requires mapping word utterances to the corresponding items in a given vocabulary. It is an elementary though important speech recognition task, having widespread applications to speech-based user interactions with embedded devices such as smartphones or smartwatches. GSC contains $\sim 106k$ one-second utterances of 35 different keywords recorded at $16kHz$, plus

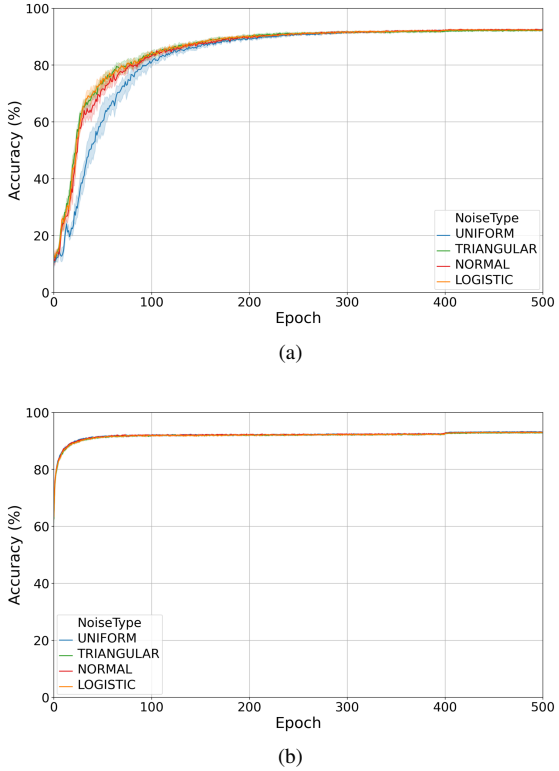


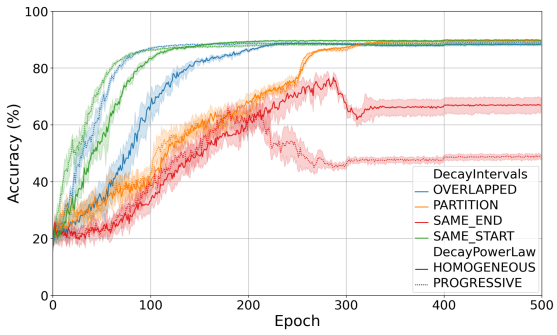
Figure 4. Performance of ANA on the SVHN data set using static noise schedules in combination with different forward computation strategies: random 4a, mode 4b. Each plot reports different noise types using different colours: uniform (blue), triangular (green), normal (red), logistic (yellow).

recordings of random background noise. There are different keyword spotting tasks associated with GSC; in our experiments, we focussed on the simplified 12-class classification problem.

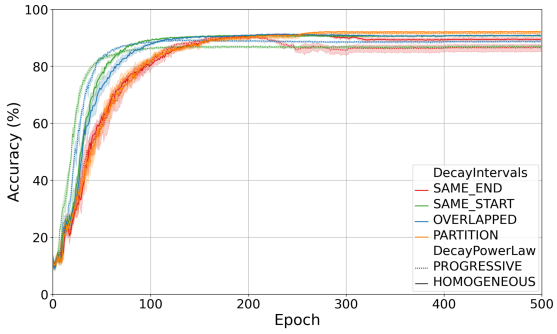
We used the DSCNN network topology [29], a fully-feedforward network topology consisting of eight convolutional layers (four blocks concatenating a depth-wise convolution with a point-wise one) and one fully-connected layer; therefore, $L = 9$. This time, we quantised weights aiming for the `INT4` (signed) data type, and features aiming for the `UINT4` (unsigned) data type. Coherently with literature practice, we kept the last layer in floating-point format.

In each experimental unit, we trained the network for 120 epochs using mini-batches of 256 pre-processed utterances, the cross-entropy loss function, and the ADAM optimiser with an initial learning rate of 10^{-3}, decreased to 10^{-4} after 100 epochs.

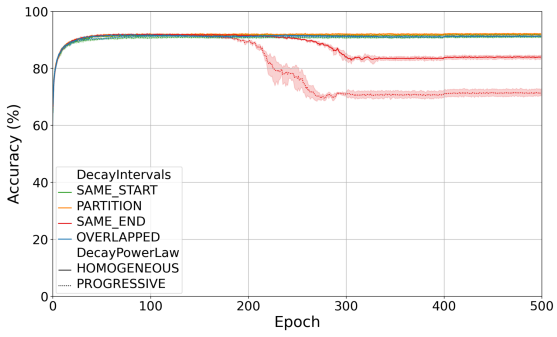
Figures 6a 6b show that QNNs trained using static STE variants based on different noise types still converge to approximately the same accuracy. However, in this scenario we can observe that the accuracy of QNNs trained using the



(a)



(b)

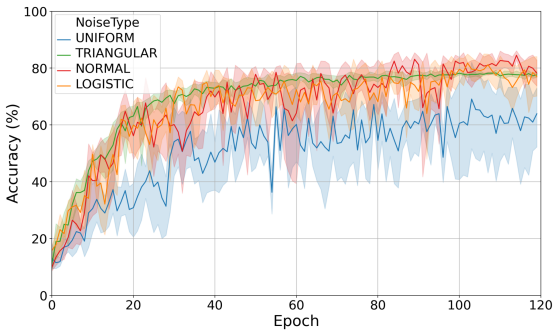


(c)

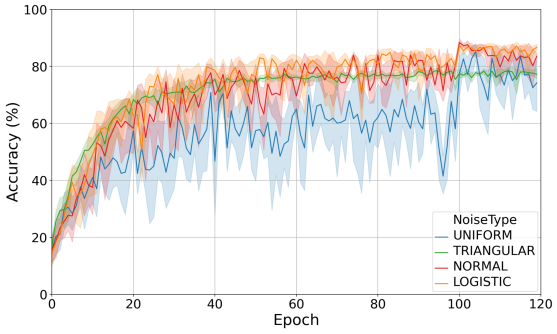
Figure 5. Performance of ANA on the SVHN data set using dynamic noise schedules (static means, dynamic variances) under uniform noise and different forward computation strategies: expectation 5a, random 5b, mode 5c. Each plot reports multiple schedules: decay intervals: same start (green), same end (red), partition (yellow), overlapped (blue); decay power law: homogeneous (continuous), progressive (dotted).

triangular noise type has lower variability, whereas that of QNNs trained using uniform noise has higher variability.

Figures 7a, 7b, 7c show that the quality of different decay interval strategies (as measured by the final accuracy of the trained networks) is better for those that are more coherent with the hypothesis of Theorem 2. In particular,



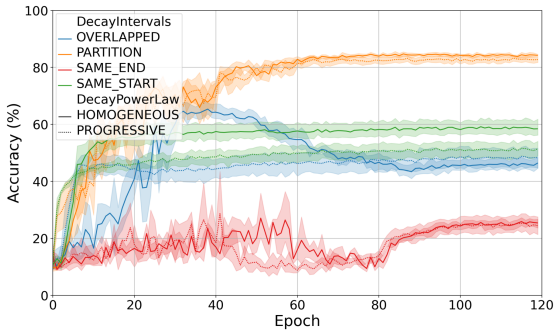
(a)



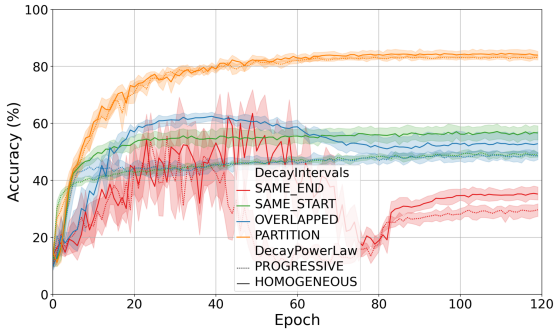
(b)

Figure 6. Performance of ANA on the GSC data set using static noise schedules in combination with different forward computation strategies: random 6a, mode 6b. Each plot reports different noise types using different colours: uniform (blue), triangular (green), normal (red), logistic (yellow).

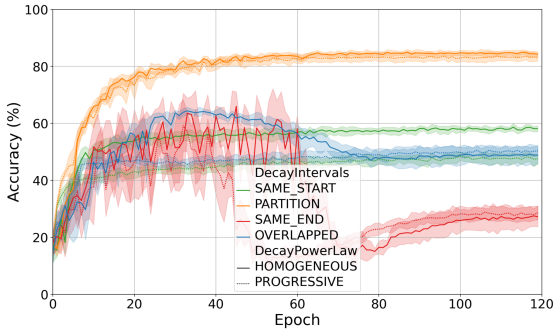
QNNs trained using the partition strategy can achieve approximately the same accuracy as networks trained using static noise schedules. Independently of the forward computation strategy, the same end decay interval strategy is still the worst amongst the tested ones.



(a)



(b)



(c)

Figure 7. Performance of ANA on the GSC data set using dynamic noise schedules (static means, dynamic variances) under uniform noise and different forward computation strategies: expectation 7a, random 7b, mode 7c. Each plot reports multiple schedules: decay intervals: same start (green), same end (red), partition (yellow), overlapped (blue); decay power law: homogeneous (continuous), progressive (dotted).