AdaSTE: An Adaptive Straight-Through Estimator to Train
Binary Neural Networks
Supplementary Material

Huu Le    Rasmus Kjær Høier    Che-Tsung Lin    Christopher Zach
Chalmers University of Technology, Gothenburg, Sweden
huul, hier, chetsung, zach@chalmers.se

1. Algorithmic Comparison between AdaSTE, ProxQuant and Mirror Descent

Algorithm 1 illustrates the differences between ProxQuant, Mirror Descent, and the proposed AdaSTE method to train binarized DNNs. For better clarity we display a full-batch gradient method, and we also omit the annealing aspect of ProxQuant and MD (i.e. we assume a fixed parameter absorbed into $\hat{s}$ and $\mathcal{E}$, respectively). (M38) refers to Equation 38 in the main text.

Mirror descent-based training and AdaSTE share the interpretation of $\theta(t)$ as the current latent weights, whereas $\theta(t)$ already tends to be binarized in ProxQuant. In the next section we show that AdaSTE can be considered as adaptive and time-varying variant of mirror descent.

Algorithm 1 ProxQuant/MD/AdaSTE training method.

1: Initialize $\theta(0)$, choose learning rates $\eta^t$, $t = 1, \ldots$
2: for $t = 1, \ldots$ do
3:     $w^* \leftarrow \theta(t)$
4:     $\hat{s} \leftarrow \hat{s}(\theta(t))$
5:     $w^* \leftarrow \hat{s}(\theta(t))$
6:     Run regular back-prop to determine $\ell'(w^*)$
7:     $\theta(t) \leftarrow \text{prox}_{\eta^t}(\theta(t) - \eta^t \ell'(w^*))$
8:     $\theta(t) \leftarrow \theta(t) - \eta^t \ell'(w^*)$
9:     Determine $\bar{\beta}(t)$ using (M38)
10: $\hat{w} \leftarrow \hat{s}(\theta(t) - \bar{\beta}(t) \odot \ell'(w^*))$
11: $\theta(t+1) \leftarrow \theta(t) - \eta^t (w^* - \hat{w}) \odot \bar{\beta}(t)$
12: end for

2. A Mirror Descent Interpretation of AdaSTE

In this section we establish a connection between AdaSTE and mirror descent with a data-adaptive and varying metric. Since the update in AdaSTE is applied element-wise, we focus on the update of $\theta_j$ (a scalar) in the following. For brevity of notation we drop the subscript $j$.

We consider using a “partial” chain rule as follows. Let the target forward mapping be the composition of $s_1$ and $s_2$, i.e. $s = s_2 \circ s_1$. Then the AdaSTE update step is abstractly given by

$$\theta(t+1) \leftarrow \theta(t) - \eta^t s_2'(s_1(\theta(t))) s_2'(s_1(\theta(t))) .$$  (1)

Observe that only one step of the chain rule is applied on $\ell$ as $s_1'$ is not used. We introduce an “intermediate” weight $u = s_1(\theta)$, and therefore $w = s_2(u) = s_2(s_1(\theta)) = s(\theta)$. Expressing the above update step in $u$ yields

$$s_2^{-1}(u(t+1)) \leftarrow s_2^{-1}(u(t)) - \eta^t (s_2(u(t))) s_2'(u(t)) ,$$  (2)

and identifying $s_2^{-1}$ with the mirror map $\nabla \Phi$ results eventually in

$$u(t+1) = \arg\min_u \frac{1}{\eta} D_\Phi(u\|u(t)) + \ell'(s_2(u)) s_2'(u(t)) = \arg\min_u \frac{1}{\eta} D_\Phi(u\|u(t)) + \frac{d}{du} \ell(s_2(u)) \big|_{u=u(t)} .$$  (3)

Now the question is whether there exist mappings $s_1$ and $s_2$ such that

$$s_2(s_1(\theta)) = s(\theta) \quad s_2'(u) = s'(s_2^{-1}(u) - h) ,$$  (4)

where will be chosen as $h = \beta \ell'$ in AdaSTE. The first relation yields

$$s_1(\theta) = s_2^{-1}(s(\theta)) \quad \text{and} \quad s_2^{-1}(u) = s^{-1}(s_2(u)) .$$  (5)

Hence, the second condition above is equivalent to

$$s_2'(u) = s'(s_2^{-1}(u) - h) = s'(s^{-1}(s_2(u)) - h) .$$

By expressing this relation in terms of $\theta$ we obtain

$$s_2'(s_1(\theta)) = s'(\theta - h) \quad \iff \quad s_2'(s_2^{-1}(s(\theta))) = s'(\theta - h)$$

$$\iff \frac{1}{s_2^{-1}(s(\theta))} = s'(\theta - h)$$

$$\iff \frac{1}{s_2^{-1}(w)} = \frac{1}{s'(s_2^{-1}(w) - h)} .$$


Consequently, $s_2^{-1}$ can be determined by solving
\[ s_2^{-1}(w) = \int_{w_0}^{w} \frac{1}{s(s^{-1}(w) - h)} \, dw. \] (6)

If $h = 0$, then $s_2^{-1} = s^{-1}$ (and therefore $s_1 = \text{id}$) is a valid solution. For $h \neq 0$, there is sometimes a closed-form expression for $s_2^{-1}$. We consider $s = \tanh$, i.e.
\[ s(\theta) = \frac{e^\theta - e^{-\theta}}{e^\theta + e^{-\theta}} = \frac{e^{2\theta} - 1}{e^{2\theta} + 1}, \quad s'(\theta) = \frac{4e^{2\theta}}{(e^{2\theta} + 1)^2}. \] (7)

With this choice we obtain (via a computer algebra system)
\[ (s_2^{-1})'(w) = \frac{1}{s'(s^{-1}(w) - h)} = \frac{e^{-2h}(e^{2h} - 1)w - e^{2h} - 1}{4(1 - w^2)} = \frac{((e^{2h} - 1)w - e^{2h} - 1)^2}{4(1 - w^2)}. \] (8)

Now the following relation holds,
\[ \int \frac{(aw + b)^2}{(1 - w^2)} \, dw = \frac{1}{2} \left( -2a^2w - (a + b)^2 \log(1 - w) + (a - b)^2 \log(1 + w) \right). \]

Plugging in the values $a = e^{2h} - 1$ and $b = -e^{2h} - 1$ (and therefore $a + b = -2e^{2h}$ and $a - b = 2e^{2h}$) results in
\[ s_2^{-1}(w) = \frac{1}{2} \left( (e^{2h} - 1)w - e^{2h} - 1 \right) \log(1 - w) + 4e^{2h} \log(1 + w) \]
\[ = \frac{1}{2} \left( e^{2h} \log(1 + w) - e^{2h} \log(1 - w) \right) - \frac{1}{4}(e^{2h} - 1)^2 w. \] (9)

As expected, for $h = 0$ we obtain $\tanh^{-1}$, and for $h \neq 0$ this mapping skewsn $\tanh^{-1}$. The important property is, that $s_2$ is strictly monotone since $s_2(s_1(\theta)) = s'(\theta - h) > 0$. We can recover $s_1$ via $s_1(x) = s_2^{-1}(s(\theta))$, but that seems to be a non-interpretable expression in this case.

3. AdaSTE: the case $\mu \alpha < 1$

As in the previous section we focus on one scalar weight $\theta_j / \omega_j$ and omit the subscript $j$ in the following. We know that the actual weight $w$ is obtained via
\[ w^* = \Pi_{[-1,1]} \left( \frac{\theta + \mu (1 + \alpha) \text{sgn}(\theta)}{1 + \mu} \right) \]
\[ \hat{w} = \Pi_{[-1,1]} \left( \frac{\hat{\theta} + \mu (1 + \alpha) \text{sgn}(\hat{\theta})}{1 + \mu} \right), \] (10)

where $\hat{\theta} = \theta - \beta \ell'$. We focus on $\theta < 0$, since the case $\theta > 0$ is symmetric. Hence,
\[ w^* = \begin{cases} -1 & \text{if } \theta < -1 + \mu \alpha \\ \frac{\theta - \mu (1 + \alpha)}{1 + \mu} & \text{if } \theta \in (-1 + \mu \alpha, 0) \end{cases} \] (11)
and
\[ \hat{w} = \begin{cases} -1 & \text{if } \hat{\theta} < -1 + \mu \alpha \\ \frac{\hat{\theta} - \mu (1 + \alpha)}{1 + \mu} & \text{if } \hat{\theta} \in (-1 + \mu \alpha, 0) \end{cases} \] (12)

We are now interested in values for $\beta > 0$ maximizing $|\hat{w} - w^*| / \beta$. We assume that $\mu \alpha < 1$, since the simpler setting $\mu \alpha \geq 1$ was discussed in the main paper.

Case $\ell' > 0$: We have $\hat{\theta} = \theta - \beta \ell' < \theta$ for all $\beta > 0$. Since $\hat{w}$ will be clamped at $-1$ for sufficiently large $\beta > 0$, the solution for $\beta$ satisfies
\[ \theta - \beta \ell' \in (-1 + \mu \alpha, 0). \] (13)

If $\theta \leq -1 + \mu \alpha$, then we have $w^* = \hat{w} = -1$ for all choices of $\beta$, and therefore $(\hat{w} - w^*) / \beta = 0$ regardless of $\beta$. Thus, we assume that $\theta > -1 + \mu \alpha$ and therefore $w^* > -1$. For $\beta$ constrained as above, we have
\[ \frac{\hat{w} - w^*}{\beta} = \frac{1}{\beta} \cdot \frac{\theta - \beta \ell' - \mu (1 + \alpha) - (\theta - \mu (1 + \alpha))}{1 + \mu} = \frac{1}{\beta} \cdot \frac{\beta \ell'}{1 + \mu} = \frac{\ell'}{1 + \mu}, \]

which is independent of the exact value of $\beta$ as long it is in the allowed range,
\[ \beta \in \frac{1}{\ell'} (\theta, \theta + 1 - \mu \alpha) \cap \mathbb{R}_{>0}. \] (14)

We can set $\beta$ as follows,
\[ \beta = \min \left\{ \beta_{\max}, \frac{\theta + 1 - \mu \alpha}{\ell'} \right\} \]

and the error signal is given by $(\hat{w} - w^*) / \beta = \ell' / (1 + \mu)$.

Case $\ell' < 0$: This means that $\hat{\theta} > \theta$ for $\beta > 0$. By inspecting the piecewise linear (and monotonically increasing) mapping $\theta \mapsto w^*$ we identify two relevant choices for $\beta$: $\beta_1$ as the smallest $\beta$ such that $\hat{w}$ is clamped at $+1$, and $\beta_0$ as the smallest $\beta$ such that $\hat{w}$ is positive. Note that $\hat{\theta}$ is clamped at $+1$ whenever $\hat{\theta} > 1 - \mu \alpha$. Therefore the defining constraints for $\beta_1$ and $\beta_0$ are given by
\[ \theta - \beta_1 \ell' = 1 - \mu \alpha \quad \theta - \beta_0 \ell' = 0^+, \]

where $\beta_0 \ell' < 0$.
i.e. $β_1 = (θ - 1 + μα)/ℓ'$ and $β_0 = θ/ℓ'$ (and $β_1 > β_0$ by construction). If $θ = 0^+$, then $\hat{w} = μ(1 + α)/(1 + μ)$. Consequently,

$$\frac{\hat{w}_1 - w^*}{β_1} = \frac{\ell'}{θ - 1 + μα} \left(1 - \max\left\{-1, \frac{θ - μ(1 + α)}{1 + μ}\right\}\right)$$

$$\frac{\hat{w}_0 - w^*}{β_0} = \frac{\ell'}{θ - 1 + μα} \left(1 + μ - \max\left\{-1, \frac{θ - μ(1 + α)}{1 + μ}\right\}\right).$$

If $θ \leq -1 + μα$ such that $w^* = -1$, then these expressions simplify to

$$\frac{\hat{w}_1 - w^*}{β_1} = \frac{2\ell'}{θ - 1 + μα} > 0$$

$$\frac{\hat{w}_0 - w^*}{β_0} = \frac{\ell'}{θ - 1 + μα} (1 + μ) > 0.$$

Now $(\hat{w}_1 - w^*)/β_1 > (\hat{w}_0 - w^*)/β_0$ iff

$$\frac{2\ell'}{θ - 1 + μα} > \frac{2(1 + 2μ + μα)}{(1 + μ)}$$

$$\iff \frac{2(1 + μ)}{(1 + μ)} > \frac{1 + 2μ}{θ - 1 + μα}$$

$$\iff 2(1 + μ)θ < (θ - 1 + μα)(1 + 2μ + μα)$$

$$\iff (1 - μα)(θ + 1 + 2μ + μα) < 0$$

$$\iff θ < -2 - 2μ - μα.$$

Visual inspection shows that $β_0$ is a good solution even when $β_1$ is the maximizer: $β_0$ does not maximize the slope $(\hat{w} - w^*)/β$, but its slope is close to the maximal one.

If $θ \in (-1 + μα, 0)$, then $w^* = (θ - μ(1 + α))/(1 + μ)$ and therefore

$$\frac{\hat{w}_1 - w^*}{β_1} = \frac{\ell'}{θ - 1 + μα} \left(1 - \frac{θ - μ(1 + α)}{1 + μ}\right)$$

$$= \frac{\ell'}{θ - 1 + μα} \left(1 + μ - θ + μ(1 + α)\right)$$

$$\frac{\hat{w}_0 - w^*}{β_0} = \frac{\ell'}{θ - 1 + μα} \left(1 + μ\right).$$

$(\hat{w}_1 - w^*)/β_1 > (\hat{w}_0 - w^*)/β_0$ iff (after dividing both sides by $1 + μ > 0$)

$$\frac{(1 + 2μ + μα - θ)\ell'}{θ - 1 + μα} > \frac{(2μ(1 + α) - θ)\ell'}{1 + μ}$$

$$\iff \frac{1 + 2μ + μα - θ}{θ - 1 + μα} < \frac{2μ(1 + α) - θ}{1 + μ}$$

$$\iff 1 + 2μ + μα - θ < (2μ(1 + α) - θ)(θ - 1 + μα)$$

$$\iff 2μ(1 - μα)(1 + α) < 0.$$

The l.h.s. is always positive under our assumptions, therefore $β_0 = θ/ℓ'$ is the maximizer in this case.

### 4. Proof of Proposition 1

**Proposition 1.** Let $E(w; θ) = G(w) - w^T θ$ and $w^* = \arg\min_w E(w; θ)$ be explicitly given as $w^* = s(θ)$. Then

$$\hat{w} = \tilde{s}(θ - β \odot ℓ'(w^*)).$$

**Proof.** We simply absorb the linear perturbation term into $θ$, yielding $\hat{θ} := θ - β \odot ℓ'(w^*)$, and therefore $\hat{w}$ solves

$$\hat{w} = \arg\min_w G(w) - w^T \hat{θ} = \arg\min_w E(w; \hat{θ}).$$

Hence, $\hat{w} = \tilde{s}(\hat{θ}) = \tilde{s}(θ - β \odot ℓ'(w^*))$ as claimed. □

### 5. Convergence analysis of AdaSTE

We use the following assumptions:

1. $ℓ$ is bounded from below and has a Lipschitz gradient with Lipschitz constant $L$.
2. $s$ is monotonically increasing and is $M$-Lipschitz continuous.

Both assumptions are often violated (since e.g. the standard cross-entropy loss is not bounded from below, and our choice for $s$ is not Lipschitz continuous). The respective convergence analysis of ProxQuant and mirror descent shares similar limitations.

The first assumption implies that

$$ℓ(w') ≥ ℓ(w) + \langle ∇ℓ(w), w' - w \rangle + \frac{L}{2} ∥w' - w∥^2.$$ (17)

Let $θ(t)$ be the latent weights in iteration $t$, and $w(t) := s(θ(t))$. We abbreviate $\nabla ℓ(w(t))$ as $g(t)$. Thus, $w^* = w(t)$ and $\hat{w}$ in iteration $t$ is given element-wise by

$$w_j(t) = s(θ_j(t) - β_j g_j(t)) = w_j(t) - α_j(t) β_j g_j(t)$$

for some $α_j(t) ≥ 0$ (due to the monotonicity of $s$). Moreover, using (M26) we identify $α_j(t)$ as (generalized) derivative of $s$ at a perturbation of $θ_j(t)$. Since $s$ is monotone and Lipschitz continuous, we deduce that $α_j(t) ≤ M$ (or $α_j(t) ≤ \min(1, M)$ in view of the gradient clipping described in Section 4.5 in the main text). Consequently,

$$θ_j(t+1) = θ_j(t) - \frac{η_j}{β_j} (w_j(t) - w_j(t)) = θ_j(t) - η_j α_j g_j(t).$$

Again, due to the monotonicity of $s$ we deduce that

$$(w_j(t+1) - w_j(t))^T g_j(t) = (s(θ_j(t+1)) - s(θ_j(t)))^T g_j(t) ≤ 0,$$ (20)
and \( w^{(t+1)} - w^{(t)} \) is thus a descent (but not necessarily the gradient) direction of \( \ell \) at \( w^{(t)} \). We have even something stronger:

\[
    w^{(t+1)}_j - w^{(t)}_j = -\eta \gamma_j g^{(t)}_j
\]

for \( \gamma_j \geq 0 \). In order to guarantee a reduction of \( \ell \) in each iteration, we require that each term in

\[
    \nabla \ell (w^{(t)})^T (w^{(t+1)} - w^{(t)}) + \frac{\ell}{2} \| w^{(t+1)} - w^{(t)} \|^2
    = \sum_i \left( g^{(t)}_j (w^{(t+1)}_j - w^{(t)}_j) + \frac{\ell}{2} (w^{(t+1)}_j - w^{(t)}_j)^2 \right)
\]

is non-positive. Recall that

\[
    g^{(t)}_j \cdot (w^{(t+1)}_j - w^{(t)}_j) \leq 0
\]

from above. Hence, each term can be written as

\[
    A^{(t)}_j := g^{(t)}_j (w^{(t+1)}_j - w^{(t)}_j) + \frac{\ell}{2} (w^{(t+1)}_j - w^{(t)}_j)^2
    = -|g^{(t)}_j| \cdot |w^{(t+1)}_j - w^{(t)}_j| + \frac{\ell}{2} (w^{(t+1)}_j - w^{(t)}_j)^2.
\]

Observe that \( A^{(t)}_j \) is a quadratic function in \( w^{(t+1)}_j \), and it is 0 for \( w^{(t+1)}_j = w^{(t)}_j \) and decreases monotonically until \( w^{(t+1)}_j = w^{(t)}_j - g^{(t)}_j / \ell \) (where \( A^{(t)}_j \) reaches its minimum value). Therefore we require that

\[
    |w^{(t+1)}_j - w^{(t)}_j| \leq \frac{1}{\ell} |g^{(t)}_j|.
\]

We employ the Lipschitz assumption on \( s \) and obtain,

\[
    |w^{(t+1)}_j - w^{(t)}_j| \leq M |\theta^{(t+1)}_j - \theta^{(t)}_j| = M \eta \alpha^{(t)}_j |g^{(t)}_j|.
\]

If we choose

\[
    \eta_t = \min_j \frac{1}{LM \alpha_j} = \frac{1}{LM} \cdot \frac{1}{\max_j \alpha^{(t)}_j},
\]

then

\[
    |w^{(t+1)}_j - w^{(t)}_j| \leq M \eta \alpha^{(t)}_j |g^{(t)}_j| = \frac{\alpha^{(t)}_j}{L \max_j \alpha^{(t)}_j} |g^{(t)}_j| \leq \frac{1}{\ell} |g^{(t)}_j|.
\]

as required. By recalling that \( \alpha^{(t)}_j \in [0, \min(1, M)] \) (using Lipschitz continuity of \( s \) and gradient clipping) we realize
that
\[
\max_j \alpha_j(t) \leq \min(1, M)
\]
\[
\Rightarrow \frac{1}{\max_j \alpha_j(t)} \geq \frac{1}{\min(1, M)} \geq 1,
\]
(28)
and \(\eta_t\) can in fact simply be chosen as
\[
\eta_t = \frac{1}{LM}.
\]
(29)

With this universal choice of \(\eta_t\) we read
\[
|w_j(t+1) - w_j(t)| \leq M\eta_t \alpha_j(t)|g_j(t)| = \frac{\alpha_j(t)}{T} |g_j(t)| \leq \frac{1}{T} |g_j(t)|
\]
due to \(\alpha_j(t) \in [0, 1]\). Thus, we obtain the first (and main) result: the sequence of objective values \((\ell(w(t)))_{t=1,...}\) is non-increasing.

The value of \(A_j(t)\) is given by
\[
A_j(t) = -|g_j(t)| \cdot |w_j(t+1) - w_j(t)| + \frac{1}{T} \left( w_j(t+1) - w_j(t) \right)^2
\]
\[
= -\frac{\alpha_j(t)}{L} |g_j(t)|^2 + \frac{(\alpha_j(t))^2}{2L} |g_j(t)|^2
\]
\[
= \frac{\alpha_j(t)(\alpha_j(t)-2)}{2L} |g_j(t)|^2 \leq 0.
\]
(30)

Summing over \(j\) and \(t\) yields (using the boundedness of \(\ell\) from below in the first relation)
\[
-\infty < \ell(w(T)) - \ell(w(0)) \leq \sum_{t=1}^T \sum_j A_j(t) \leq 0,
\]
(31)
which implies that \(A_j(T) \rightarrow 0\) for \(T \rightarrow \infty\). Thus, \(g_j(T) \rightarrow 0\) or \(\alpha_j(T) \rightarrow 0\). In the first case the target loss \(\ell\) is stationary w.r.t. \(w_j\), i.e. \(\frac{\partial}{\partial w_j} \ell(w(T)) \rightarrow 0\). In the second case \(\alpha_j(T) \rightarrow 0\) implies that \(s\) is behaving constant (as \(\hat{w}_j(T) \rightarrow w_j(T)\) and therefore the finite differences anchored at \(\theta(T)\) vanish). Hence, a solution \(\theta^{(\infty)} = \lim_{T \rightarrow \infty} \theta(T)\) is (component-wise) either stationary w.r.t. \(\ell\) or w.r.t. \(s\). If \(s\) is differentiable (in addition to being Lipschitz continuous), then this is analogous to the standard first-order optimality condition,
\[
\ell' \left( s(\theta^{(\infty)}) \right) \cdot s' \left( \theta^{(\infty)} \right) = 0.
\]
(32)

6. Imagenette Results and Mixup

In order to further justify if our model also works well on images at higher resolution, we conduct the same experiment on Imagenette dataset [3] which are sampled from Imagenet [1] without being downsamped and consists of 9469 training images and 3925 validation images. Besides, we also notice that mixup [2], a proven effective training trick, is also helpful in further boosting the classification accuracy. As can be seen in Table 1, it is quite obvious that our AdaSTE consistently outperforms BayesBiNN on both TinyImageNet and Imagenette datasets with and without mixup.

<table>
<thead>
<tr>
<th></th>
<th>TinyImageNet</th>
<th>Imagenette</th>
</tr>
</thead>
<tbody>
<tr>
<td>BayesBiNN</td>
<td>54.22</td>
<td>78.19</td>
</tr>
<tr>
<td>BayesBiNN (mixup)</td>
<td>55.84</td>
<td>79.59</td>
</tr>
<tr>
<td>AdaSTE</td>
<td>54.92</td>
<td>79.66</td>
</tr>
<tr>
<td>AdaSTE (mixup)</td>
<td>56.11</td>
<td>80.91</td>
</tr>
</tbody>
</table>

Table 1. Classification accuracy for different methods on Tiny ImageNet and Imagenette: Annealing is applied to our model with and without mixup.

7. Implementation Details

We implemented our AdaSTE algorithm in PyTorch, on top of the framework provided by BayesBiNN. In particular, we used SGD with momentum of 0.9 for all experiments.
- For CIFAR-10 and CIFAR-100 datasets, we used batch size of 128 with learning rate of \(10^{-5}\).
- For TinyImageNet, the chosen batch size was 100 with the learning rate of \(10^{-6}\).

The experimental results for BayesBiNN were produced with the following hyper parameters:
- Batch size: 128.
- Learning rate: 3 \times 10^{-4}.
- Momentum: 0.9.

8. CIFAR-100 Results

Similar to Fig. 3 and Fig. 4 in the main paper, in Fig. 1 and Fig. 2 (of this supplementary material), we also show the test accuracy and training loss versus number of epochs for the CIFAR-100 dataset with ResNet-18 and VGG-16 architectures. The same conclusion can also be drawn, where AdaSTE can quickly achieve very good performance, while it takes longer for other methods to yield high accuracy. This emphasizes the advantage of our method compared to existing approaches.

9. Training AdaSTE and BayesBiNN for a larger number of epochs

In Table 1 in the main paper, we report results obtained after training BayesBiNN and AdaSTE for 500 epochs. In
Fig. 3 (of this supplementary material), we further show the progress of BayesBiNN and AdaSTE after training for 700 epochs. As can be seen, the performance of both BayesBiNN and AdaSTE can still be improved, and BayesBiNN slowly approaches the performance of AdaSTE.

References

