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 $\vec{s}$ 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 $\eta^{(t)}$, $t = 1, \dots$
2: **for** $t = 1, \dots$ **do**
3: $w^* \leftarrow \theta^{(t)}$
4: $w^* \leftarrow \vec{s}(\theta^{(t)})$
5: $w^* \leftarrow \vec{s}(\theta^{(t)})$
6: Run regular back-prop to determine $\ell'(w^*)$
7: $\theta^{(t+1)} \leftarrow \text{prox}_{\eta^{(t)}\mathcal{E}}\left(\theta^{(t)} - \eta^{(t)}\ell'(w^*)\right)$
8: $\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta^{(t)}\ell'(w^*)$
9: Determine $\vec{\beta}^{(t)}$ using (M38)
10: $\hat{w} \leftarrow \vec{s}\big(\theta^{(t)} - \vec{\beta}^{(t)} \odot \ell'(w^*)\big)$
11: $\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta^{(t)}(w^* - \hat{w}) \oslash \vec{\beta}^{(t)}$
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 θ_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

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

Observe that only one step of the chain rule is applied on ℓ 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

$$s_1^{-1}(u^{(t+1)}) \leftarrow s_1^{-1}(u^{(t)}) - \eta\ell'(s_2(u^{(t)}))s_2'(u^{(t)}), \quad (2)$$

and identifying s_1^{-1} with the mirror map $\nabla\Phi$ results eventually in

$$\begin{aligned} u^{(t+1)} &= \arg\min_u \tfrac{1}{\eta} D_\Phi(u \| u^{(t)}) + \ell'(s_2(u^{(t)}))s_2'(u^{(t)}) \\ &= \arg\min_u \tfrac{1}{\eta} D_\Phi(u \| u^{(t)}) + \tfrac{d}{du}\ell(s_2(u))\big|_{u=u^{(t)}}. \end{aligned} \quad (3)$$

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

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

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

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

Hence, the second condition above is equivalent to

$$s_2'(u) = s'(s_1^{-1}(u) - h) = s'\big(s^{-1}(s_2(u)) - h\big).$$

By expressing this relation in terms of θ we obtain

$$\begin{aligned} & s_2'(s_1(\theta)) = s'(\theta - h) \iff s_2'(s_2^{-1}(s(\theta))) = s'(\theta - h) \\ \iff & \frac{1}{(s_2^{-1})'(s(\theta))} = s'(\theta - h) \\ \iff & (s_2^{-1})'(w) = \frac{1}{s'(s^{-1}(w) - h)}. \end{aligned}$$

Consequently, s_2^{-1} can be determined by solving

$$s_2^{-1}(w) = \int_{w_0}^{w} \frac{1}{s'(s^{-1}(\omega) - h)}\, d\omega. \tag{6}$$

If $h = 0$, then $s_2^{-1} = s^{-1}$ (and therefore $s_1 = \mathrm{id}$) 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.

$$s(\theta) = \frac{e^\theta - e^{-\theta}}{e^\theta + e^{-\theta}} = \frac{e^{2\theta} - 1}{e^{2\theta} + 1} \quad s'(\theta) = \frac{4e^{2\theta}}{(e^{2\theta}+1)^2}. \tag{7}$$

With this choice we obtain (via a computer algebra system)

$$\begin{aligned}(s_2^{-1})'(w) &= \frac{1}{s'(s^{-1}(w) - h)} \\ &= \frac{e^{-2h}\big((e^{2h}-1)w - e^{2h} - 1\big)^2}{4(1-w^2)} \\ &= \frac{\big((e^h - e^{-h})w - e^h - e^{-h}\big)^2}{4(1-w^2)}. \end{aligned} \tag{8}$$

Now the following relation holds,

$$\begin{aligned}&\int \frac{(aw+b)^2}{4(1-w^2)}\, dw \\ &\doteq \tfrac{1}{8}\left(-2a^2 w - (a+b)^2 \log(1-w) + (a-b)^2 \log(1+w)\right).\end{aligned}$$

Plugging in the values $a = e^h - e^{-h}$ and $b = -e^h - e^{-h}$ (and therefore $a + b = -2e^{-h}$ and $a - b = 2e^h$) results in

$$\begin{aligned}&s_2^{-1}(w) \\ &= \tfrac{1}{8}\left(-2(e^h - e^{-h})^2 w - 4e^{-2h}\log(1-w) + 4e^{2h}\log(1+w)\right) \\ &= \tfrac{1}{2}\left(e^{2h}\log(1+w) - e^{-2h}\log(1-w)\right) \\ &- \tfrac{1}{4}(e^h - e^{-h})^2 w. \end{aligned} \tag{9}$$

As expected, for $h = 0$ we obtain $\tanh^{-1}$, and for $h \neq 0$ this mapping skews $\tanh^{-1}$. 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 θ_j/w_j and omit the subscript j in the following. We know that the actual weight w is obtained via

$$\begin{aligned} w^* &= \Pi_{[-1,1]}\left(\frac{\theta + \mu(1+\alpha)\,\mathrm{sgn}(\theta)}{1+\mu}\right) \\ \hat{w} &= \Pi_{[-1,1]}\left(\frac{\tilde\theta + \mu(1+\alpha)\,\mathrm{sgn}(\tilde\theta)}{1+\mu}\right), \end{aligned} \tag{10}$$

where $\tilde\theta = \theta - \beta\ell'$. We focus on $\theta < 0$, since the case $\theta > 0$ is symmetric. Hence,

$$w^* = \begin{cases} -1 & \text{if } \theta \leq -1 + \mu\alpha \\ \frac{\theta - \mu(1+\alpha)}{1+\mu} & \text{if } \theta \in (-1+\mu\alpha, 0) \end{cases} \tag{11}$$

and

$$\hat{w} = \begin{cases} -1 & \text{if } \tilde\theta \leq -1 + \mu\alpha \\ \frac{\tilde\theta - \mu(1+\alpha)}{1+\mu} & \text{if } \tilde\theta \in (-1+\mu\alpha, 0) \end{cases}. \tag{12}$$

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 $\tilde\theta = \theta - \beta\ell' < \theta$ for all $\beta > 0$. Since $\hat{w}$ will be clamped at -1 for sufficiently large $\beta > 0$, the solution for β satisfies

$$\theta - \beta\ell' \in (-1 + \mu\alpha, 0). \tag{13}$$

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

$$\begin{aligned}\frac{\hat{w} - w^*}{\beta} &= \frac{1}{\beta} \cdot \frac{\theta - \beta\ell' - \mu(1+\alpha) - (\theta - \mu(1+\alpha))}{1+\mu} \\ &= \frac{1}{\beta} \cdot \frac{\beta\ell'}{1+\mu} = \frac{\ell'}{1+\mu},\end{aligned}$$

which is independent of the exact value of β as long it is in the allowed range,

$$\beta \in \tfrac{1}{\ell'}(\theta, \theta + 1 - \mu\alpha) \cap \mathbb{R}_{\geq 0}. \tag{14}$$

We can set β 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 $\tilde\theta > \theta$ for $\beta > 0$. By inspecting the piecewise linear (and monotonically increasing) mapping $\theta \mapsto w^*$ we identify two relevant choices for β: β_1 as the smallest β such that $\hat{w}$ is clamped at $+1$, and β_0 as the smallest β such that $\hat{w}$ is positive. Note that $\tilde\theta$ is clamped at $+1$ whenever $\tilde\theta > 1 - \mu\alpha$. Therefore the defining constraints for β_1 and β_0 are given by

$$\theta - \beta_1\ell' = 1 - \mu\alpha \qquad \theta - \beta_0\ell' = 0^+,$$

i.e. $\beta_1 = (\theta - 1 + \mu\alpha)/\ell'$ and $\beta_0 = \theta/\ell'$ (and $\beta_1 > \beta_0$ by construction). If $\tilde{\theta} = 0^+$, then $\hat{w} = \mu(1+\alpha)/(1+\mu)$. Consequently,

$$\frac{\hat{w}_1 - w^*}{\beta_1} = \frac{\ell'}{\theta - 1 + \mu\alpha}\left(1 - \max\left\{-1, \frac{\theta - \mu(1+\alpha)}{1+\mu}\right\}\right)$$
$$\frac{\hat{w}_0 - w^*}{\beta_0} = \frac{\ell'}{\theta}\left(\frac{\mu(1+\alpha)}{1+\mu} - \max\left\{-1, \frac{\theta - \mu(1+\alpha)}{1+\mu}\right\}\right).$$

If $\theta \leq -1 + \mu\alpha$ such that $w^* = -1$, then these expressions simplify to

$$\frac{\hat{w}_1 - w^*}{\beta_1} = \frac{2\ell'}{\theta - 1 + \mu\alpha} > 0$$
$$\frac{\hat{w}_0 - w^*}{\beta_0} = \frac{\ell'}{\theta} \cdot \frac{\mu + \mu\alpha + 1 + \mu}{1+\mu} = \frac{\ell'(1 + 2\mu + \mu\alpha)}{(1+\mu)\theta} > 0.$$

Now $(\hat{w}_1 - w^*)/\beta_1 > (\hat{w}_0 - w^*)/\beta_0$ iff

$$\begin{aligned}
&\frac{2\ell'}{\theta - 1 + \mu\alpha} > \frac{\ell'(1 + 2\mu + \mu\alpha)}{(1+\mu)\theta} \\
\iff& \frac{2}{\theta - 1 + \mu\alpha} < \frac{1 + 2\mu + \mu\alpha}{(1+\mu)\theta} \\
\iff& 2(1+\mu)\theta < (\theta - 1 + \mu\alpha)(1 + 2\mu + \mu\alpha) \\
\iff& (1 - \mu\alpha)(\theta + 1 + 2\mu + \mu\alpha) < 0 \\
\iff& \theta < -1 - 2\mu - \mu\alpha.
\end{aligned}$$

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

If $\theta \in (-1 + \mu\alpha, 0)$, then $w^* = (\theta - \mu(1+\alpha))/(1+\mu)$ and therefore

$$\begin{aligned}
\frac{\hat{w}_1 - w^*}{\beta_1} &= \frac{\ell'}{\theta - 1 + \mu\alpha}\left(1 - \frac{\theta - \mu(1+\alpha)}{1+\mu}\right) \\
&= \frac{\ell'}{\theta - 1 + \mu\alpha} \cdot \frac{1 + \mu - \theta + \mu(1+\alpha)}{1+\mu} \\
\frac{\hat{w}_0 - w^*}{\beta_0} &= \frac{\ell'}{\theta} \cdot \frac{\mu(1+\alpha) - \theta + \mu(1+\alpha)}{1+\mu}.
\end{aligned}$$

$(\hat{w}_1 - w^*)/\beta_1 > (\hat{w}_0 - w^*)/\beta_0$ iff (after dividing both sides by $1 + \mu > 0$)

$$\begin{aligned}
&\frac{(1 + 2\mu + \mu\alpha - \theta)\ell'}{\theta - 1 + \mu\alpha} > \frac{(2\mu(1+\alpha) - \theta)\ell'}{\theta} \\
\iff& \frac{1 + 2\mu + \mu\alpha - \theta}{\theta - 1 + \mu\alpha} < \frac{2\mu(1+\alpha) - \theta}{\theta} \\
\iff& (1 + 2\mu + \mu\alpha - \theta)\theta < (2\mu(1+\alpha) - \theta)(\theta - 1 + \mu\alpha) \\
\iff& 2\mu(1 - \mu\alpha)(1+\alpha) < 0
\end{aligned}$$

The l.h.s. is always positive under our assumptions, therefore $\beta_0 = \theta/\ell'$ is the maximizer in this case.

4. Proof of Proposition 1

Proposition 1. *Let $\mathcal{E}(w;\theta) = G(w) - w^\top\theta$ and $w^* = \arg\min_w \mathcal{E}(w;\theta)$ be explicitly given as $w^* = \vec{s}(\theta)$. Then*

$$\hat{w} = \vec{s}\big(\theta - \vec{\beta} \odot \ell'(w^*)\big). \tag{15}$$

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

$$\hat{w} = \arg\min_w G(w) - w^\top\tilde{\theta} = \arg\min_w \mathcal{E}(w;\tilde{\theta}). \tag{16}$$

Hence, $\hat{w} = \vec{s}(\tilde{\theta}) = \vec{s}(\theta - \vec{\beta} \odot \ell'(w^*))$ 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

$$\ell(w') \geq \ell(w) + \nabla\ell(w)^\top(w' - w) + \tfrac{L}{2}\|w' - w\|^2. \tag{17}$$

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

$$\hat{w}_j^{(t)} = s(\theta_j^{(t)} - \beta_j g_j^{(t)}) = w_j^{(t)} - \alpha_j^{(t)}\beta_j g_j^{(t)} \tag{18}$$

for some $\alpha_j^{(t)} \geq 0$ (due to the monontonicity of s). Moreover, using (M26) we identify $\alpha_j^{(t)}$ as (generalized) derivative of s at a perturbation of $\theta^{(t)}$. Since s is monotone and Lipschitz continuous, we deduce that $\alpha_j^{(t)} \leq M$ (or $\alpha_j^{(t)} \leq \min(1, M)$ in view of the gradient clipping described in Section 4.5 in the main text). Consequently,

$$\theta_j^{(t+1)} = \theta_j^{(t)} - \tfrac{\eta_t}{\beta_j}\left(w_j^{(t)} - \hat{w}_j^{(t)}\right) = \theta_j^{(t)} - \eta_t\alpha_j g_j^{(t)}. \tag{19}$$

Again, due to the monotonicity of s we deduce that

$$(w^{(t+1)} - w^{(t)})^\top g^{(t)} = (s(\theta^{(t+1)}) - s(\theta^{(t)}))^\top g^{(t)} \leq 0, \tag{20}$$

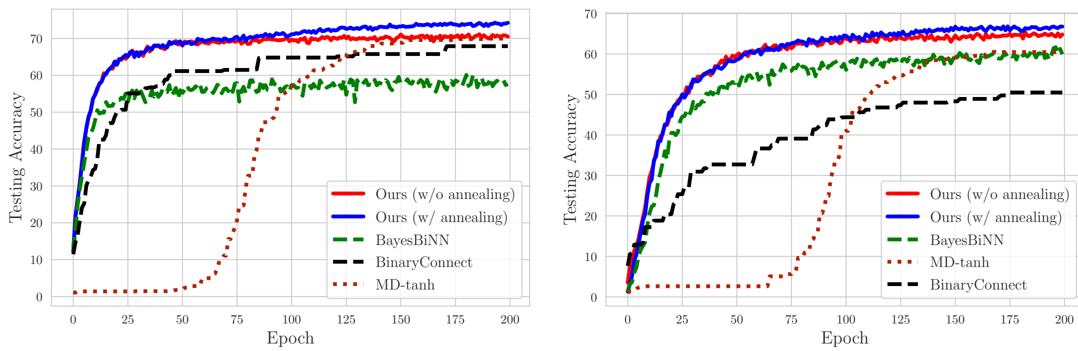

Figure 1. Testing accuracy achieved by the methods for the first 200 epochs with ResNet-18 (left) VGG16 (right) for CIFAR100 dataset.

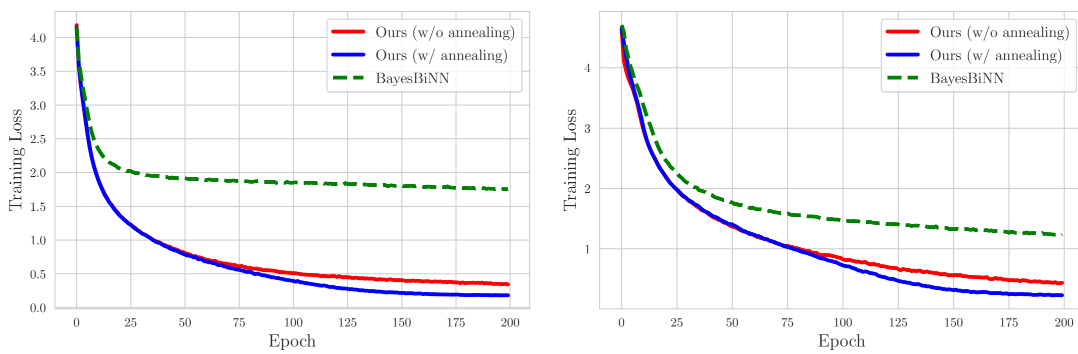

Figure 2. Training loss of the methods for the first 200 epochs with ResNet-18 (left) and VGG16 (right) on the CIFAR100 dataset.

and $w^{(t+1)} - w^{(t)}$ is thus a descent (but not necessarily the gradient) direction of ℓ at $w^{(t)}$. We have even something stronger:

$$w_j^{(t+1)} - w_j^{(t)} = -\eta_t \gamma_j g_j^{(t)} \tag{21}$$

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

$$\begin{aligned} &\nabla \ell(w^{(t)})^\top (w^{(t+1)} - w^{(t)}) + \tfrac{L}{2} \left\| w^{(t+1)} - w^{(t)} \right\|^2 \\ &= \sum_i \left(g_j^{(t)} (w_j^{(t+1)} - w_j^{(t)}) + \tfrac{L}{2} \left(w_j^{(t+1)} - w_j^{(t)} \right)^2 \right) \end{aligned} \tag{22}$$

is non-positive. Recall that

$$g_j^{(t)} \cdot (w_j^{(t+1)} - w_j^{(t)}) \leq 0 \tag{23}$$

from above. Hence, each term can be written as

$$\begin{aligned} A_j^{(t)} &:= g_j^{(t)} (w_j^{(t+1)} - w_j^{(t)}) + \tfrac{L}{2} \left(w_j^{(t+1)} - w_j^{(t)} \right)^2 \\ &= -\,|g_j^{(t)}| \cdot |w_j^{(t+1)} - w_j^{(t)}| + \tfrac{L}{2} \left(w_j^{(t+1)} - w_j^{(t)} \right)^2 . \end{aligned} \tag{24}$$

Observe that $A_j^{(t)}$ is a quadratic function in $w_j^{(t+1)}$, and it is 0 for $w_j^{(t+1)} = w_j^{(t)}$ and decreases monotonically until $w_j^{(t+1)} = w_j^{(t)} - g_j^{(t)}/L$ (where $A_j^{(t)}$ reaches its minimum value). Therefore we require that

$$|w_j^{(t+1)} - w_j^{(t)}| \leq \tfrac{1}{L} |g_j^{(t)}|. \tag{25}$$

We employ the Lipschitz assumption on s and obtain,

$$|w_j^{(t+1)} - w_j^{(t)}| \leq M |\theta_j^{(t+1)} - \theta_j^{(t)}| = M \eta_t \alpha_j^{(t)} |g_j^{(t)}|.$$

If we choose

$$\eta_t = \min_j \frac{1}{LM\alpha_j} = \frac{1}{LM} \cdot \frac{1}{\max_j \alpha_j^{(t)}}, \tag{26}$$

then

$$\begin{aligned} |w_j^{(t+1)} - w_j^{(t)}| \leq M \eta_t \alpha_j^{(t)} |g_j^{(t)}| &= \frac{\alpha_j^{(t)}}{L \max_{j'} \alpha_{j'}^{(t)}} |g_j^{(t)}| \\ &\leq \tfrac{1}{L} |g_j^{(t)}| \end{aligned} \tag{27}$$

as required. By recalling that $\alpha_j^{(t)} \in [0, \min(1, M)]$ (using Lipschitz continuity of s and gradient clipping) we realize that

$$\max_j \alpha_j^{(t)} \le \min(1, M)$$
$$\implies \frac{1}{\max_j \alpha_j^{(t)}} \ge \frac{1}{\min(1, M)} \ge 1, \tag{28}$$

and η_t can in fact simply be chosen as

$$\eta_t = \frac{1}{LM}. \tag{29}$$

With this universal choice of η_t we read

$$|w_j^{(t+1)} - w_j^{(t)}| \le M\eta_t \alpha_j^{(t)} |g_j^{(t)}| = \frac{\alpha_j^{(t)}}{L} |g_j^{(t)}| \le \frac{1}{L} |g_j^{(t)}|$$

due to $\alpha_j^{(t)} \in [0, 1]$. Thus, we obtain the first (and main) result: the sequence of objective values $(\ell(w^{(t)}))_{t=1,\dots}$ is non-increasing.

The value of $A_j^{(t)}$ is given by

$$\begin{aligned} A_j^{(t)} &= -|g_j^{(t)}| \cdot |w_j^{(t+1)} - w_j^{(t)}| + \tfrac{L}{2} \left(w_j^{(t+1)} - w_j^{(t)} \right)^2 \\ &= -\tfrac{\alpha_j^{(t)}}{L} |g_j^{(t)}|^2 + \tfrac{(\alpha_j^{(t)})^2}{2L} |g_j^{(t)}|^2 \\ &= \tfrac{\alpha_j^{(t)} \left(\alpha_j^{(t)} - 2 \right)}{2L} \cdot |g_j^{(t)}|^2 \le 0. \end{aligned} \tag{30}$$

Summing over j and t yields (using the boundedness of ℓ from below in the first relation)

$$-\infty < \ell(w^{(T)}) - \ell(w^{(0)}) \le \sum_{t=1}^{T} \sum_j A_j^{(t)} \le 0, \tag{31}$$

which implies that $A_j^{(T)} \to 0$ for $T \to \infty$. Thus, $g_j^{(T)} \to 0$ or $\alpha_j^{(T)} \to 0$. In the first case the target loss ℓ is stationary w.r.t. w_j, i.e. $\frac{\partial}{\partial w_j} \ell(w^{(T)}) \to 0$. In the second case $\alpha_j^{(T)} \to 0$ implies that s is behaving constant (as $\hat{w}_j^{(T)} \to w_j^{(T)}$ and therefore the finite differences anchored at $\theta^{(T)}$ vanish). Hence, a solution $\theta^{(\infty)} = \lim_{T\to\infty} \theta^{(t)}$ is (component-wise) either stationary w.r.t. ℓ 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,

$$\ell' \left(s(\theta^{(\infty)}) \right) \cdot s'(\theta^{(\infty)}) = \mathbf{0}. \tag{32}$$

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 downsampled 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 consistenly outperforms BayesBiNN on both TinyImageNet and Imagenette datasets with and without mixup.

	TinyImageNet ResNet-18	Imagenette ResNet-18
BayesBiNN	54.22	78.19
BayesBinn (mixup)	55.84	79.59
AdaSTE	54.92	79.66
AdaSTE (mixup)	56.11	80.91

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: 3×10^{-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

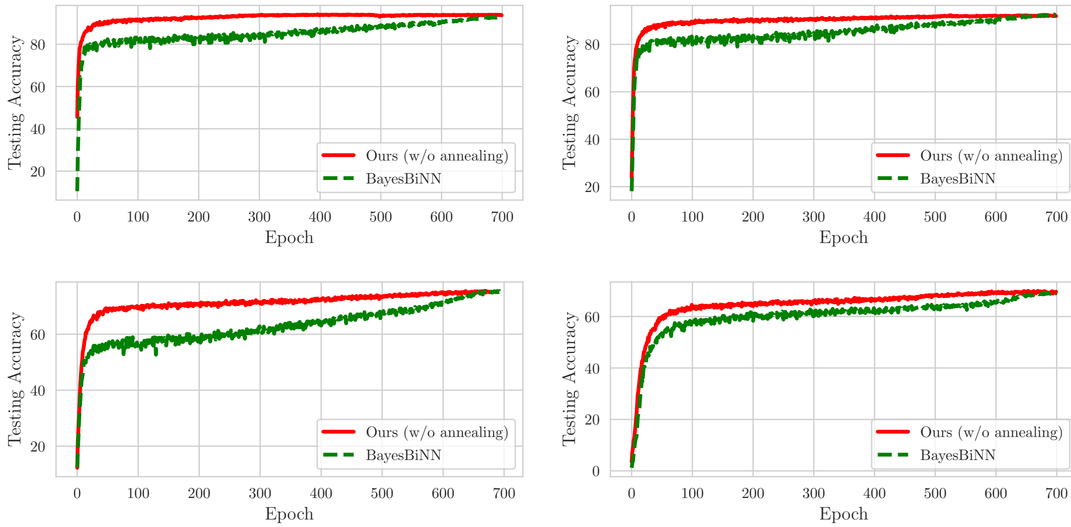

Figure 3. Testing accuracy achieved by the AdaSTE (no annealing) and BayesBiNN for 700 epochs. Top: CIFAR-10 with ResNet-18 (left) and VGG16 (right)

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

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 5

[2] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *CVPR*, pages 558–567, 2019. 5

[3] Jeremy Howard and Sylvain Gugger. Fastai: a layered api for deep learning. *Information*, 11(2):108, 2020. 5