Paper ID 1 APPENDIX

BAMSProd: A Step towards Generalizing the Adaptive Optimization Methods to Deep Binary Model

A Optimization setup

We follow the work [Reddi, Kale, and Kumar 2018] to employ a common definition of the online learning [Bianchi, Conconi, and Gentile 2002] with a sequence of observations for analyzing the deep learning optimizers. At t-th time step, the algorithm picks a model with parameter $w_t \in \mathcal{F}$, and a loss $f_t(w_t)$ is produced by the function $f_t(\cdot)$ in w_t. After T rounds, the regret of the algorithm is given by $R_T = \sum_{i=1}^{T} f_t(w_t) - \min_{w \in \mathcal{F}} \sum_{i=1}^{T} f_t(w^*)$, where w^* is used to denote the global optimal solution. In the following of the paper, we assume that the $\mathcal{F}$ has bounded diameter and $\|\nabla f_t(w_t)\|_\infty$ is bounded for all $t \in [T]$ and $w \in \mathcal{F}$. For the main problem, our objective is to devise an algorithm that ensures $R_T = o(T)$, which implies that the model is converged to the optimal on average with average regret $R_T/T \to 0$ as $T \to \infty$.

B Auxiliary Lemmas

Lemma 1. *For any $w_t \in \mathbb{R}^d$ and convex feasible set $\mathcal{F} \subset \mathbb{R}^d$, suppose $c_1 = \min_{\tilde{w}_t \in \mathcal{F}} \|w_t - \tilde{w}_t\|$ and $\tilde{w}_t = \alpha_t sign(w_t)$ s.t. $\alpha_t \in \mathbb{R}$, then we have $\alpha_t \geq 0;\ \forall t \in \mathbb{N}$.*

Proof. We provide the proof here for completeness. Since $c_1 = \min_{\tilde{w}_t \in \mathcal{F}} \|w_t - \tilde{w}_t\|$ and $\tilde{w}_t = \alpha_t sign(w_t)$, we have the following:

$$c_1 = \min_{\tilde{w}_t \in \mathcal{F}} \|w_t - \alpha_t sign(w_t)\|.$$

After rearranging, if given any $t \in \mathbb{N}$, we assume that exists the $\alpha_t < 0$, then we have

$$\alpha_t = \frac{1}{d} \sum_{i=0}^{d} \frac{w_{t,i} - c_1 \mathbb{I}}{sign(w_{t,i})} < 0.$$

Since the property of projection operator $\tilde{w}_t \in \mathcal{F}$ and convex feasible set $\mathcal{F}$, we have

$$\sum_{i=0}^{d} \left(|w_{t,i}| - \frac{c_1 \mathbb{I}}{sign(w_{t,i})} \right) \geq 0. \tag{1}$$

Hence the above inequality is false, and the inequality is true if only $\alpha_t \geq 0$.

We complete the proof of this lemma.

Lemma 2. *Suppose $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$ with $v_0 = \mathbf{0}$ and $0 \leq \beta_2 < 1$. Given $\|g_t\| \leq G_\infty$, we have*

$$\sum_{t=1}^{T} \|v_t\| \leq \sum_{t=1}^{T} G_\infty^2. \tag{2}$$

Proof. If $\beta_2 = 0$, the $v_t = g_t^2$, it satisfies our claim. Otherwise, for $0 < \beta_2 < 1$, we have

$$\begin{aligned} \|v_t\| &= \beta_2 \|v_{t-1}\| + (1 - \beta_2) \|g_t^2\| \\ &\leq \beta_2 \|v_{t-1}\| + (1 - \beta_2) \|G_\infty^2\|. \end{aligned}$$

The inequality follows from the gradient constraint $\|g_t\| \leq G_\infty$. In particular, since that $v_0 = \mathbf{0}$, we get

$$\|v_t\| \leq (1 - \beta_2) \sum_{i=1}^{t} \|g_i^2\| \beta_2^{t-i}.$$

Take the summation of above inequality with $t = [1, T]$, we have

$$\sum_{t=1}^{T}\|v_t\| \leq (1-\beta_2)\sum_{t=1}^{T}\sum_{i=1}^{t}\|g_i^2\|\beta_2^{t-i}$$
$$= (1-\beta_2)\sum_{i=1}^{T}\sum_{t=i}^{T}\|g_i^2\|\beta_2^{t-i}$$
$$\leq \sum_{t=1}^{T}\|g_t\|^2 \leq \sum_{t=1}^{T}G_\infty^2.$$

The second inequality follows from the constraint of

$$\sum_{i=0}^{N}\beta_2^i \leq \sum_{i=0}^{\infty}\beta_2^i = \frac{1}{1-\beta_2}.$$

for $0 < \beta_2 < 1$.

We complete the proof of this lemma.

Lemma 3. *For the parameter settings and conditions assumed in Theorem 4, we have*

$$\begin{aligned}\sum_{t=1}^{T}\frac{\eta_t}{2(1-\beta_{1t})}\|\hat{V}_t^{-1/4}m_t\|^2 + \sum_{t=1}^{T}\frac{\eta_t\beta_{1t}}{2(1-\beta_{1t})}\|\hat{V}_t^{-1/4}m_{t-1}\|^2 \\ \leq \frac{\eta\sqrt{1+\log T}}{(1-\beta_1)^2(1-\gamma)\sqrt{(1-\beta_2)}}\sum_{i=1}^{d}\|\alpha_{1:T,i}\|_2\varpi\end{aligned} \tag{3}$$

Proof. We start the proof with the following:

$$\begin{aligned}&\sum_{t=1}^{T}\frac{\eta_t}{2(1-\beta_{1t})}\|\hat{V}_t^{-1/4}m_t\|^2 + \sum_{t=1}^{T}\frac{\eta_t\beta_{1t}}{2(1-\beta_{1t})}\|\hat{V}_t^{-1/4}m_{t-1}\|^2 \\ &\leq \frac{1}{2(1-\beta_1)}\Big[\sum_{t=1}^{T}\eta_t\|\hat{V}_t^{-1/4}m_t\|^2 + \sum_{t=1}^{T}\eta_t\|\hat{V}_t^{-1/4}m_{t-1}\|^2\Big] \\ &\leq \frac{1}{(1-\beta_1)}\Big[\sum_{t=1}^{T-1}\eta_t\|\hat{V}_t^{-1/4}m_t\|^2 + \eta\sum_{i=1}^{d}\frac{(\sum_{j=1}^{T}\prod_{k=1}^{T-j}\beta_{1(T-k+1)}\alpha_{j,i}\varpi)^2}{\sqrt{T((1-\beta_2)\sum_{j=1}^{T}\beta_2^{T-j}\alpha_{j,i}^2\varpi)}}\Big] \\ &\leq \frac{1}{(1-\beta_1)}\Big[\sum_{t=1}^{T-1}\eta_t\|\hat{V}_t^{-1/4}m_t\|^2 + \eta\sum_{i=1}^{d}\frac{(\sum_{j=1}^{T}\prod_{k=1}^{T-j}\beta_{1(T-k+1)})(\sum_{j=1}^{T}\prod_{k=1}^{T-j}\beta_{1(T-k+1)}\alpha_{j,i}\varpi)^2}{\sqrt{T((1-\beta_2)\sum_{j=1}^{T}\beta_2^{T-j}\alpha_{j,i}^2\varpi)}}\Big] \\ &\leq \frac{1}{(1-\beta_1)}\Big[\sum_{t=1}^{T-1}\eta_t\|\hat{V}_t^{-1/4}m_t\|^2 + \frac{\eta}{(1-\beta_1)\sqrt{T(1-\beta_1)}}\sum_{i=1}^{d}\sum_{j=1}^{T}\frac{\beta_1^{T-j}}{\sqrt{\beta_2^{T-j}}}\alpha_{j,i}\Big]\end{aligned} \tag{4}$$

The first inequality follows from the fact of $\beta_{1t} \leq \beta 1 < 1$. The second inequality follows from the definition of $\alpha_{T,i}\hat{v}_{T,i}$, which is maximum of all $\alpha_{T,i}v_{T,i}$ until the current time step and the update rule of Algorithm 1. The third inequality follows from Cauchy-Schwarz inequality. The fourth inequality follows from the fact the $\beta_{1k} \leq \beta_1; \forall k \in [T]$ and the $\sum_{j=1}^{T}\beta_1^{T-j} \leq 1/(1-\beta_1)$. By using the upper bounds for all time steps, the quantity in Eq. 4 is further bounded as:

$$\sum_{t=1}^{T} \frac{\eta_t}{2(1-\beta_{1t})} \|\hat{V}_t^{-1/4} m_t\|^2 + \sum_{t=1}^{T} \frac{\eta_t \beta_{1t}}{2(1-\beta_{1t})} \|\hat{V}_t^{-1/4} m_{t-1}\|^2$$
$$\leq \sum_{t=1}^{T} \frac{\eta}{(1-\beta_1)^2 \sqrt{t(1-\beta_1)}} \sum_{i=1}^{d} \sum_{j=1}^{T} \frac{\beta_1^{T-j}}{\sqrt{\beta_2^{T-j}}} \alpha_{j,i}$$
$$\leq \frac{\eta}{(1-\beta_1)^2 (1-\beta_1/\sqrt{\beta_2}) \sqrt{(1-\beta_2)}} \sum_{i=1}^{d} \|\alpha_{1:T,i}\|_2 \sqrt{\sum_{t=1}^{T} \frac{1}{t}}$$
$$\leq \frac{\eta \sqrt{1+\log T}}{(1-\beta_1)^2 (1-\beta_1/\sqrt{\beta_2}) \sqrt{(1-\beta_2)}} \sum_{i=1}^{d} \|\alpha_{1:T,i}\|_2$$

The second inequality follows from the proof by Reddi et al. (Reddi, Kale, and Kumar (2018)) in their Appendix D. The last inequality is due to the bound on harmonic sum

$$\sum_{t=1}^{T} \frac{1}{t} \leq (1 + \log T)$$

We complete the proof of this lemma.

Lemma 4 [Reddi, Kale, and Kumar 2018]. *Suppose* $\mathcal{F} = [a, b]$ *for* $a, b \in \mathbb{R}$ *and*

$$y_{t+1} = \prod\nolimits_{\mathcal{F}} (y_t + \delta_t).$$

$\forall t \in [T], y_1 \in \mathcal{F}$ *and furthermore, there* $\exists i \in [T]$ *such that* $\delta_j \leq 0;\ \forall j \leq i$ *and* $\delta_j > 0;\ \forall j > 0$. *Then we have,*

$$y_{T+1} \geq \min\{b, y_1 + \sum_{j=1}^{T} \delta_j\}. \tag{5}$$

Proof. $y_{i+1} \geq y_1 + \sum_{j=1}^{i} \delta_j$ since $\delta_j \leq 0;\ \forall j \leq i$.

Furthermore, $y_{T+1} \geq \min b, y_{i+1} + \sum_{j=i+1}^{T} \delta_j$ since $\delta_j \geq 0;\ \forall j > i$.

Lemma 5 [Mcmahan & Streeter 2010] *For any* $Q \in S_+^d$ *and convex feasible set* $\mathcal{F} \subset \mathbb{R}^d$, *suppose* $u_1 = \min_{w \in \mathcal{F}} \|Q^{1/2}(w - z_1)\|$ *and* $u_2 = \min_{w \in \mathcal{F}} \|Q^{1/2}(w - z_2)\|$ *then we have*

$$\|Q^{1/2}(u_1 - u_2)\| \leq \|Q^{1/2}(z_1 - z_2)\|. \tag{6}$$

Proof. Since $u_1 = \min_{w \in \mathcal{F}} \|Q^{1/2}(w - z_1)\|$ and $u_2 = \min_{w \in \mathcal{F}} \|Q^{1/2}(w - z_2)\|$ and from the property of projection operator, we have the following:

$$\langle z_1 - u_1, Q(z_2 - z_1) \rangle \geq 0 \text{ and } \langle z_2 - u_2, Q(z_1 - z_2) \rangle \geq 0.$$

Combining the above inequalities, we have

$$\langle u_2 - u_1, Q(z_2 - z_1) \rangle \geq \langle z_2 - z_1, Q(z_2 - z_1) \rangle. \tag{7}$$

Also, observe the following:

$$\langle u_2 - u_1, Q(z_2 - z_1) \rangle \leq \frac{1}{2} [\langle u_2 - u_1, Q(u_2, -u_1) \rangle + \langle z_2 - z_1, Q(z_2 - z_1) \rangle].$$

The above inequality can be obtained from the fact that

$$\langle (u_2 - u_1) - (z_2 - z_1), \langle Q((u_2, -u_1) - Q(z_2 - z_1)) \rangle \geq 0.$$

as for any $Q \in S_+^d$.

Rearranging the terms. Combining the above inequality with Eq. 7, we have the required result.

C Proof of Theorem 1

Proof. We consider the setting where f_t are linear functions with the latent weights and $\mathcal{F} = [-1, 1]$. In details, we define the following function sequences:

$$f_t(w) = \begin{cases} -1 & \text{for } \tilde{w} = -1; \\ \tilde{w} & \text{for } -1 < \tilde{w} \leq 1 \text{ and } t \bmod C = 1; \\ -\tilde{w} & \text{for } -1 < \tilde{w} \leq 1 \text{ and } t \bmod C = 2; \\ 0 & \text{otherwise,} \end{cases}$$

where the $\tilde{w} = \alpha w_b$ is the latent weight, which approximate the w by minimizing the quantization errors $\min_{\tilde{w}_t \in \mathcal{F}} \|w_t - \tilde{w}_t\|$ then we have

$$f_t(w) = \begin{cases} -1 & \text{for } \alpha w_b = -1; \\ \alpha w_b & \text{for } -1 < \alpha w_b \leq 1 \text{ and } t \bmod C = 1; \\ -\alpha w_b & \text{for } -1 < \alpha w_b \leq 1 \text{ and } t \bmod C = 2; \\ 0 & \text{otherwise,} \end{cases}$$

For simplifying the notation, we drop the $\odot$ as the problem is one-dimensional, and we further drop indices representing coordinates from all quantizes in Adam optimization.

For this function sequence, it is not hard to see that the $\tilde{w} = -1$ provides the minimum regret. Without loss of generality, assume that the initial point is $w_1 = 1$. This can be assumed without any loss of generality because for any choice of initial point, we can always translate the coordinate system such that the initial point is $w_1 = 1$ in the new coordinate system and then choose the sequence of functions as above in the new coordinate system.

As the execution of Adam algorithm for this sequence of functions with

$$\beta_1 = 0, \frac{\beta_1^2}{\beta_2} < 1 \text{ and } \eta_t = \frac{\eta}{\sqrt{t}}.$$

and $\mathcal{F}$ has bounded L_∞ diameter. All conditions on parameters required for Adam are satisfied (Diederik and Jimmy (2015)).

For proving the said theorem, we claim that for any initial step size η, we have $\tilde{w}_t > 0;\ \forall t \in \mathbb{N}$, and furthermore, $\tilde{w}_{Ct+3} = 1;\ \forall C \in \mathbb{N} \cup \{0\}$. In details, we resort to the principle of mathematical induction. For every C steps, suppose for some $t \in \mathbb{N}$, we have $\tilde{w}_{Ct+1} \geq 0$. Our aim at proving that for $\tilde{w}_{Ct+i} > 0;\ \forall i \in \mathbb{N} \cap [2, C+1]$. It is not hard to see that the conditions holds if $\tilde{w}_{Ct+1} > 1$. Now we assume $\tilde{w}_{Ct+1} \leq 1$, as $w_b = sign(w)$ and $\tilde{w}_t > 0$, we observe that the gradients have:

$$\nabla f_i(w) = \begin{cases} \alpha_t & \text{for } -1 < \alpha \leq 1 \text{ and } i \bmod C = 1; \\ -\alpha_t & \text{for } -1 < \alpha \leq 1 \text{ and } i \bmod C = 2; \\ 0 & \text{otherwise,} \end{cases}$$

For the $(Ct+1)$-th update of Adam, we obtain

$$\hat{w}_{Ct+2} = \tilde{w}_{Ct+1} - \frac{\eta}{\sqrt{Ct+1}} \frac{\alpha_{Ct}}{\sqrt{\beta_2 v_{Ct} + \alpha_{Ct}^2 (1-\beta_2)}}$$

As $\beta_2 v_{Ct} \geq 0; \forall t \in \mathbb{N}$, we have the following

$$\begin{aligned} \frac{\eta}{\sqrt{Ct+1}} \frac{\alpha_{Ct}}{\sqrt{\beta_2 v_{Ct} + \alpha_{Ct}^2 (1-\beta_2)}} &\leq \frac{\eta}{\sqrt{Ct+1}} \frac{\alpha_{Ct}}{\sqrt{\alpha_{Ct}^2 (1-\beta_2)}} \\ &= \frac{\eta}{\sqrt{Ct+1}} \frac{1}{\sqrt{(1-\beta_2)}} < 1 \end{aligned}$$

The second inequality follows from the fact that $0 < \eta < \frac{1}{\sqrt{(Ct+1)(1-\beta_2)}}$. Therefore, we have $0 < \hat{w}_{Ct+2} < 1$ and hence $\tilde{w}_{Ct+2} = \hat{w}_{Ct+2}$.

To complete the proof, we need to prove that $\tilde{w}_{Ct+3} = 1$. For proving this claim, if $\hat{w}_{Ct+3} \geq 1$, it readily translates to $\tilde{w}_{Ct+3} = 1$ as $\tilde{w}_{Ct+3} = \prod_{\mathcal{F}}(\hat{w}_{Ct+3})$ and $\mathcal{F} = [-1, 1]$, where $\prod_{\mathcal{F}}$ is simple Euclidean projection ($\prod_{\mathcal{F}, \sqrt{V_t}} = \prod_{\mathcal{F}}$ in

one-dimension).

In particular, we need to consider the following case:

After $(Ct+2)$-th update, we have:

$$\begin{aligned}\hat{w}_{Ct+3} &= \hat{w}_{Ct+2} + \frac{\eta}{\sqrt{Ct+2}} \frac{\alpha_{Ct+1}}{\sqrt{\beta_2 v_{Ct+1} + \alpha_{Ct+1}^2(1-\beta_2)}} \\ &= \tilde{w}_{Ct+1} - \frac{\eta}{\sqrt{Ct+1}} \frac{\alpha_{Ct}}{\sqrt{\beta_2 v_{Ct} + \alpha_{Ct}^2(1-\beta_2)}} + \frac{\eta}{\sqrt{Ct+2}} \frac{\alpha_{Ct+1}}{\sqrt{\beta_2 v_{Ct+1} + \alpha_{Ct+1}^2(1-\beta_2)}}\end{aligned}$$

The third equality is due to $\hat{w}_{Ct+2} = \tilde{w}_{Ct+2}$. For proving $\hat{w}_{Ct+3} \geq 1$, we aim at proving:

$$\frac{\eta}{\sqrt{Ct+1}} \frac{\alpha_{Ct}}{\sqrt{\beta_2 v_{Ct} + \alpha_{Ct}^2(1-\beta_2)}} \leq \frac{\eta}{\sqrt{Ct+2}} \frac{\alpha_{Ct+1}}{\sqrt{\beta_2 v_{Ct+1} + \alpha_{Ct+1}^2(1-\beta_2)}}$$

Rearranging the terms, then we have

$$\begin{aligned}&\frac{\eta}{\sqrt{Ct+2}} \frac{\alpha_{Ct+1}}{\sqrt{\beta_2 v_{Ct+1} + \alpha_{Ct+1}^2(1-\beta_2)}} - \frac{\eta}{\sqrt{Ct+1}} \frac{\alpha_{Ct}}{\sqrt{\beta_2 v_{Ct} + \alpha_{Ct}^2(1-\beta_2)}} \geq \\ &\frac{\eta\beta_2}{\sqrt{Ct+2}\sqrt{\beta_2 v_{Ct} + (1-\beta_2)}\sqrt{\beta_2 v_{Ct+1} + (1-\beta_2)}} (\alpha_{Ct+1} v_{Ct} - \alpha_{Ct} v_{Ct+1}) \geq 0\end{aligned}$$

This last inequality is due to the following lower bound:

$$\begin{aligned}\alpha_{Ct+1} v_{Ct} - \alpha_{Ct} v_{Ct+1} &= \alpha_{Ct+1} v_{Ct} - \alpha_{Ct}(\beta_2 v_{Ct} + (1-\beta_2)\alpha_{Ct}^2) \\ &= (\alpha_{Ct+1} - \alpha_{Ct}\beta_2)\left[(1-\beta_2)(\sum_{i=1}^{t} \beta_2^{Ci-1}\alpha_{Ci-1}^2 + \sum_{i=1}^{t} \beta_2^{Ci-2}\alpha_{Ci-2}^2)\right] - (1-\beta_2)\alpha_{Ct}^3 \\ &\geq (\alpha_{Ct+1} - \alpha_{Ct}\beta_2)\left[(1-\beta_2)(\frac{\beta_2^{C-1}\alpha_{Ct}^2}{1-\beta_2^C} + \frac{\beta_2^{C-2}\alpha_{Ct}^2}{1-\beta_2^C})\right] - (1-\beta_2)\alpha_{Ct}^3 \\ &\geq 2(\alpha_{Ct+1} - \alpha_{Ct}\beta_2)\beta_2^{C-2}\alpha_{Ct}^2 - (1-\beta_2)\alpha_{Ct}^3 \geq 0\end{aligned}$$

The first and second equality follows from the definition of v_t and rearranging the terms. The first inequality follows from Lemma 1. The second inequality is due to the fact of $\beta_2 < 1$. The last inequality follows from the assumption if $C \in \mathbb{N}$ satisfies $\beta_2^{C-2} \leq \frac{2(\alpha_{t+1} - \alpha_t\beta_2)}{1-\beta_2}$.

Furthermore, since the gradient is equal to 0 when $\tilde{w}_i \geq 0$ and $i \bmod C \neq 1$ or 2, we have

$$\begin{aligned}\tilde{w}_{Ct+4} = \hat{w}_{Ct+3} = \tilde{w}_{Ct+3} \geq 0 \\ \tilde{w}_{Ct+5} = \hat{w}_{Ct+4} = \tilde{w}_{Ct+4} \geq 0 \\ \cdots \\ \tilde{w}_{Ct+C+1} = \hat{w}_{Ct+C} = \tilde{w}_{Ct+C} \geq 0\end{aligned}$$

Therefore, given $w_1 = 1$, it holds for all $t \in \mathbb{N}$ by the principle of mathematical induction. Thus, we have

$$\sum_{i=0}^{C} f_{kC+i}(w_{kC+i}) - \sum_{i=0}^{C} f_{kC+i}(-1) \geq 0 - (-C) = C$$

where $k \in \mathbb{N}$. Therefor, for every C steps, ADAM suffers a regret of C. More specifically, $R_T \geq CT/C = T$. Thus, $R_T/T \nrightarrow 0$ as $T \to \infty$.

We complete the proof.

D Proof of Theorem 2

Theorem 2 generalizes the optimization setting used in Theorem 1. In details, we construct the binary optimization problem with the bounded gradients and a more general case if adding a bias constant in the denominator of the update in Adam as follows:

$$\hat{w}_{t+1} = \tilde{w}_t - \eta_t m_t / \sqrt{V_t + \varepsilon \mathbb{I}}.$$

Here we provide the setting of the example for completeness.

Proof. Consider the setting where f_t are linear functions with the latent weights and $\mathcal{F} = [-1, 1]$. In details, we define the following function sequences:

$$\nabla f_t(w) = \begin{cases} -1 & \text{for } \tilde{w} = -1; \\ C\alpha_t & \text{for } -1 < \tilde{w} < 1 \text{ and } t \bmod C = 1; \\ -\alpha_t & \text{for } -1 < \tilde{w} < 1 \text{ and } t \bmod C \neq 1; \\ 0 & \text{otherwise,} \end{cases}$$

where$\gamma = \frac{\beta_1}{\sqrt{\beta_2}} < 1$. Hence, $C \in \mathbb{N}$ satisfies the large constant that depends on β_1, β_2, α and $\varepsilon \leq (1-\beta_2)(1-\beta_2^{(C-2)/2}C^2)$.

According to the proof given by Reddi et al. (Reddi, Kale, and Kumar (2018)) in their Appendix B, if $m_{kC} \leq 0;\ \forall k \in \mathbb{N} \cup \{0\}$, and for the general case the m_{kC+C} is observed as $m_{kC+C} = -(1-\beta_1^{C-1}) \sum_{t=1}^{C-1} \sum_{i=1}^{t} \beta_1^{t-i}\alpha_i + (1-\beta_1)\beta_1^{C-1} C\alpha_{C-1} + \beta_1^C m_{kC}$.

If $m_{kC} \leq 0$, it can be easily shown that the constraint of $m_{kC+C} \leq 0$ is still satisfied. Now we consider the case of $m_{kC} > 0$, assuming the $\beta_2^{(C-2)/2}C^2 \leq \alpha_{(C-2)/2} \leq 1$ and $(1-\beta_1)\beta_1^{C-1}C\alpha_{C-1} \leq -(1-\beta_1^{C-1}) \sum_{t=1}^{C-1} \sum_{i=1}^{t} \beta_1^{t-i}\alpha_i$ by using the principle of mathematical induction.

Consider an iterate at time ste t of the form kC after T', we further prove the following claim:

$$x_{t+C} \geq \min\{x_t + c_t, 1\} \tag{8}$$

for some $c_t > 0$, considering the updates of Adam for the particular sequence of functions with latent weights, we have

$$\delta_t = -\frac{\eta}{\sqrt{t}} \frac{(1-\beta_1)C\alpha_t + \beta_1 m_t}{\sqrt{(1-\beta_2)C^2\alpha_t^2 + \beta_2 v_t + \varepsilon}},$$

$$\delta_{t+i} = -\frac{\eta}{\sqrt{t+i}} \frac{(1-\beta_1)\alpha_t + \beta_1 m_{t+i-1}}{\sqrt{(1-\beta_2)\alpha_t^2 + \beta_2 v_{t+i} + \varepsilon}}.$$

where $i \in \{1, \ldots, C\}$. If $\delta_{t+j} \geq 0$ for some $j \in \{1, \ldots, C-1\}$ then $\delta_{t+s} \geq 0;\ \forall s \in \{j, \ldots, C-1\}$. Using Lemma 5, we have the following:

$$x_{t+C} \geq \min\{x_t + \sum_{i=t}^{t+C-1} \delta_i, 1\}$$

Let $i' = C/2$. In order to prove the claim in Eq. 8, we need to prove the following:

$$\delta = \sum_{i=t}^{t+C-1} \delta_i > 0$$

To this end, we have:

$$\begin{aligned} \sum_{i=t+1}^{t+C-1} \delta_i &= \sum_{i=1}^{C-1} -\frac{\eta}{\sqrt{t+i}} \frac{-(1-\beta_1)\alpha_t + \beta_1 m_{t+i}}{\sqrt{(1-\beta_2)\alpha_t^2 + \beta_2 v_{t+i}}} \\ &\geq \frac{\eta}{\rho\sqrt{t(1+\beta_2)}} \left(C - i' - \frac{\beta_1^{i'-1}}{1-\beta_1} \right) - \frac{\eta}{\sqrt{t}} \frac{\gamma(1-\beta_1)(1-\gamma^{C-1})}{(1-\gamma)\sqrt{(1-\beta_2)}} \geq 0. \end{aligned}$$

The above inequality is using the Lemma 2, the definition of $\rho^2 t \geq t + C;\ \forall t \geq T'$ and the constraints of C from Eq. 2. And it is further due to the following upper bound that applies for all $i' \leq i \leq C$:

$$\begin{aligned} v_{t+i-1} &= (1-\beta_2) \sum_{j=1}^{t+i-1} \beta_2^{t+i-1-j} g_j^2 \\ &\leq (1-\beta_2)\left[\frac{\beta_2^{i'-1} C^2}{1-\beta_2^C} + \frac{1}{1-\beta_2}\right] \leq 2 \end{aligned}$$

The above inequality follows from the online problem setting where gradient is $C\alpha_{i+C}$ once every C iterations and α_i for the rest, and the fact that $\beta_2^C \leq \beta_2$ Furthermore, from the above inequality and $(1 + \frac{\gamma(1-\gamma^{C-1})}{1-\gamma}) + \frac{\beta_1^{C/2-1}}{1-\beta_1} \leq \frac{C}{3}$, we have

$$\begin{aligned} \sum_{i=t}^{t+C-1} \delta_i &\geq \delta_t + \frac{\eta}{\rho\sqrt{t(1+\beta_2)}}\left(C - i' - \frac{\beta_1^{i'-1}}{1-\beta_1}\right) - \frac{\eta}{\sqrt{t}} \frac{\gamma(1-\beta_1)(1-\gamma^{C-1})}{(1-\gamma)\sqrt{(1-\beta_2)}} \\ &\geq \frac{\eta}{\rho\sqrt{t}}\left[\frac{C}{3} - \frac{\beta_1^{C/2-1}}{1-\beta_1} - \frac{3(1-\beta_1)}{2\sqrt{1-\beta_2}}\left(1 + \frac{\gamma(1-\gamma^{C-1})}{1-\gamma}\right)\right] = \frac{\eta}{\sqrt{t}}\gamma \end{aligned}$$

According to the conditions hold for Reddi et al. (Reddi, Kale, and Kumar (2018)), hence, when $t \geq T_1'$, for every C steps, Adam optimizing the deep binary model suffers a regret of at least 2. More specifically, $R_T \geq CT/C = T$. Thus, $R_T/T \nrightarrow 0$ as $T \to \infty$.

We complete the proof.

E Proof of Theorem 3

According to the examples proposed by Reddi et al. (Reddi, Kale, and Kumar (2018)) in their Appendix C, we extend it with the constraints to satisfy Theorem 3.

Proof. Let ξ be an arbitrary small positive constant. Considering the following one dimensional stochastic optimization setting over the domain $[-1, 1]$. At each time step t, the gradient of function $f_t(w)$ is chosen as following:

$$\nabla f_i(w) = \begin{cases} C\alpha_t\varpi_t & \text{for } -1 < \tilde{w} \leq 1 \text{ and with probability } p := \frac{1+\xi}{C+1}; \\ -\alpha_t\varpi_t & \text{for } -1 < \tilde{w} \leq 1 \text{ and with probability } 1-p; \end{cases}$$

The expected function is $F(w) = \xi w$. Thus the optimal point over $[-1, 1]$ is $w^* = -1$. At each time step t the gradient g_t equals $C\alpha_t\varpi_t$ with probability p and $-\alpha_t\varpi_t$ with probability $1-p$, where $\varpi_t = sign(\tilde{w_t}) = \{-1, +1\}$.

Then, the step taken by ADAM as

$$\triangle_t = \frac{-\eta_t\Big(\beta_1 m_{t-1} + (1-\beta_1)C\alpha_t\varpi_t\Big)}{\sqrt{\beta_2 v_{t-1} + (1-\beta_2)C^2\alpha_t^2}}$$

there exists a large enough C

$$-\frac{1+\xi}{C+1}\left(\frac{1}{\sqrt{1-\beta_2}} + \frac{-\beta_1(1-\beta_1)\lceil\frac{log(C\alpha^*+1)}{log(1/\beta_1)}\rceil}{\sqrt{(1-\beta_2)(\beta_2-\beta_1^2)}}\right) + \left(1 - \frac{1+\xi}{C+1}\right)\frac{1-\beta_1}{\sqrt{\beta_2(1+\xi)C\alpha^* + (1-\beta_2)}}.$$

where C as a function of ξ, β_1, β_2 and α^*, where α^* is an optimal solution for specific constraints of quantization, and the above expression can be made non-negative. According to the proof in Appendix C in Reddi et al. (Reddi, Kale, and Kumar (2018)), such that $\mathbb{E}[\triangle_t] \geq 0$, which then implies that the Adam's step keep drifting away from the optimal solution $w^* = -1$. It should be noticed that there is no limitation of the initial step size η currently, which completes the proof.

F Proof of Theorem 4

The proof of Theorem 4 presented below which provides a claim of convergence for BAMSProd. Since our examples show a non-convergence of Adam for optimizing the deep binary model, the main issues is that Υ_t causes a increasing step size even maintains the maximum of all v_t until the time step t. Hence, the following proofs fixes these issues and provide a proof convergence for BAMSProd.

Proof. Let $w^* = \arg\min_{\tilde{w}\in\mathcal{F}} \sum_{t=1}^{T} f_t(\tilde{w})$, which w^* exists as $\mathcal{F}$ is closed and convex. We begin with the following observation:

$$\tilde{w} = \prod\nolimits_{\mathcal{F},\sqrt{\hat{V}_t}}(\tilde{w}_t - \eta_t \hat{V}_t^{-1/2} m_t) = \min_{\tilde{w}\in\mathcal{F}} \|\hat{V}_t^{-1/2}(\tilde{w} - (\tilde{w}_t - \eta_t \hat{V}_t^{-1/2} m_t))\|$$

where $\forall w^* \in \mathcal{F};\ \prod_{\mathcal{F},\sqrt{\hat{V}_t}}(w^*) = w^*$. Using Lemma 4 with $u_1 = \tilde{w}_{t+1}$ and $u_2 = w^*$, we have the following:

$$\begin{aligned}
\|\hat{V}_t^{1/4}(\tilde{w}_{t+1} - w^*)\|^2 &\le \|\hat{V}_t^{1/4}(\tilde{w}_t - \eta_t \hat{V}_t^{-1/2} m_t - w^*)\|^2 \\
&= \|\hat{V}_t^{1/4}(\tilde{w}_t - w^*)\|^2 + \eta_t^2 \|\hat{V}_t^{-1/4} m_t\|^2 - 2\eta_t \langle m_t, \tilde{w}_t - w^*\rangle \\
&= \|\hat{V}_t^{1/4}(\tilde{w}_t - w^*)\|^2 + \eta_t^2 \|\hat{V}_t^{-1/4} m_t\|^2 - 2\eta_t \langle \beta_{1t} m_{t-1} + (1-\beta_{1t})\tilde{g}_t, \tilde{w}_t - w^*\rangle \\
&= \|\hat{V}_t^{1/4}(\tilde{w}_t - w^*)\|^2 + \eta_t^2 \|\hat{V}_t^{-1/4} m_t\|^2 - 2\eta_t \langle \beta_{1t} m_{t-1} + (1-\beta_{1t})\alpha_t \varpi_t, \tilde{w}_t - w^*\rangle
\end{aligned}$$

Rearranging the above inequality, we have

$$\begin{aligned}
\langle \alpha_t \varpi_t, \tilde{w}_t - w^*\rangle &\le \frac{1}{2\eta_t(1-\beta_{1t})}\Big[\|\hat{V}_t^{1/4}(\tilde{w}_t - w^*)\|^2 - \|\hat{V}_t^{1/4}(\tilde{w}_{t+1} - w^*)\|^2\Big] + \frac{\eta_t}{2(1-\beta_{1t})}\|\hat{V}_t^{-1/4} m_t\|^2 \\
&\quad + \frac{\beta_{1t}}{1-\beta 1t}\langle m_{t-1}, \tilde{w}_t - w^*\rangle \\
&\le \frac{1}{2\eta_t(1-\beta_{1t})}\Big[\|\hat{V}_t^{1/4}(\tilde{w}_t - w^*)\|^2 - \|\hat{V}_t^{1/4}(\tilde{w}_{t+1} - w^*)\|^2\Big] + \frac{\eta_t}{2(1-\beta_{1t})}\|\hat{V}_t^{-1/4} m_t\|^2 \\
&\quad + \frac{\eta_t \beta_{1t}}{2(1-\beta_{1t})}\|\hat{V}_t^{-1/4} m_{t-1}\|^2 + \frac{\beta_{1t}}{2\eta_t(1-\beta_{1t})}\|\hat{V}_t^{1/4}(\tilde{w}_t - w^*)\|^2
\end{aligned} \tag{9}$$

The second inequality follows from Cauchy-Schwarz and Young's inequality. We now use the standard approach of bounding the regret at each step using convexity of the function f_t in the following manner:

$$\begin{aligned}
&\sum_{t=1}^{T} f_t(\tilde{w}_t) - f_t(w^*) \le \sum_{t=1}^{T} \langle \alpha_t \varpi_t, \tilde{w}_t - w^*\rangle \\
&\le \sum_{t=1}^{T}\left[\frac{1}{2\eta_t(1-\beta_{1t})}\Big[\|\hat{V}_t^{1/4}(\tilde{w}_t - w^*)\|^2 - \|\hat{V}_t^{1/4}(\tilde{w}_{t+1} - w^*)\|^2\Big] + \frac{\eta_t}{2(1-\beta_{1t})}\|\hat{V}_t^{-1/4} m_t\|^2\right. \\
&\qquad \left. + \frac{\eta_t \beta_{1t}}{2(1-\beta_{1t})}\|\hat{V}_t^{-1/4} m_{t-1}\|^2 + \frac{\beta_{1t}}{2\eta_t(1-\beta_{1t})}\|\hat{V}_t^{1/4}(\tilde{w}_t - w^*)\|^2\right]
\end{aligned} \tag{10}$$

The first inequality is due to the convexity of function f_t. The second inequality follows from the bound in Eq. 9. For further bounding this inequality, using the Lemma 3, we then have

$$\begin{aligned}
\sum_{t=1}^{T} f_t(\tilde{w}_t) - f_t(w^*) &\le \sum_{t=1}^{T}\left[\frac{1}{2\eta_t(1-\beta_{1t})}\Big[\|\hat{V}_t^{1/4}(\tilde{w}_t - w^*)\|^2 - \|\hat{V}_t^{1/4}(\tilde{w}_{t+1} - w^*)\|^2\Big]\right. \\
&\quad \left. + \frac{\beta_{1t}}{2\eta_t(1-\beta_{1t})}\|\hat{V}_t^{1/4}(\tilde{w}_t - w^*)\|^2\right] + \frac{\eta\sqrt{1+\log T}}{(1-\beta_1)^2(1-\beta_1/\sqrt{\beta_2})\sqrt{(1-\beta_2)}}\sum_{i=1}^{d}\|\alpha_{1:T,i}\|_2 \\
&\le \frac{1}{2\eta_1(1-\beta_1)}\|\hat{V}_1^{1/4}(\tilde{w}_1 - w^*)\|^2 + \frac{1}{2(1-\beta_1)}\sum_{t=2}^{T}\Big[\frac{\|\hat{V}_t^{1/4}(\tilde{w}_t - w^*)\|^2}{\eta_t} - \frac{\|\hat{V}_{t-1}^{1/4}(\tilde{w}_t - w^*)\|^2}{\eta_{t-1}}\Big] \\
&\quad + \sum_{t=1}^{T}\Big[\frac{\beta_{1t}}{2\eta_t(1-\beta_{1t})}\|\hat{V}_t^{1/4}(\tilde{w}_t - w^*)\|^2\Big] + \frac{\eta\sqrt{1+\log T}}{(1-\beta_1)^2(1-\beta_1/\sqrt{\beta_2})\sqrt{(1-\beta_2)}}\sum_{i=1}^{d}\|\alpha_{1:T,i}\|_2
\end{aligned}$$

$$= \frac{1}{2\eta_1(1-\beta_1)}\sum_{i=1}^{d}\hat{v}_{1,i}^{1/2}(\tilde{w}_{1,i}-w_i^*)^2 + \frac{1}{2(1-\beta_1)}\sum_{t=2}^{T}\sum_{i=1}^{d}(\tilde{w}_{t,i}-w_i^*)^2\left[\frac{\hat{v}_{t,i}^{1/2}}{\eta_t}-\frac{\hat{v}_{t-1,i}^{1/2}}{\eta_{t-1}}\right] + \frac{1}{2(1-\beta_1)}\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\beta_{1t}(\tilde{w}_{t,i}-w_i^*)^2\hat{v}_{t,i}^{1/2}}{\eta_t} + \frac{\eta\sqrt{1+\log T}}{(1-\beta_1)^2(1-\beta_1/\sqrt{\beta_2})\sqrt{(1-\beta_2)}}\sum_{i=1}^{d}\|\alpha_{1:T,i}\|_2 \tag{11}$$

The first and second inequality use the fact that $\beta_{1t} \leq \beta_1 \leq 1$. Considering the definition in paper Section 4 with the constraint $\|\tilde{w}_x - \tilde{w}_y\|_\infty \leq D_\infty$ and the Lipschitz-continuous $\|f_t(\tilde{w}_x) - f_t(\tilde{w}_y)\| \leq C(\alpha)\|\tilde{w}_x - \tilde{w}_y\| \leq C(\alpha)D_\infty$, then we have

$$\sum_{t=1}^{T} f_t(\tilde{w}_t) - f_t(w^*) \leq \frac{1}{2\eta_1(1-\beta_1)}\sum_{i=1}^{d}\hat{v}_{1,i}^{1/2}(\tilde{w}_{1,i}-w_i^*)^2 + \frac{1}{2(1-\beta_1)}\sum_{t=2}^{T}\sum_{i=1}^{d}(\tilde{w}_{t,i}-w_i^*)^2\left[\frac{\hat{v}_{t,i}^{1/2}}{\eta_t}-\frac{\hat{v}_{t-1,i}^{1/2}}{\eta_{t-1}}\right] + \frac{\eta\sqrt{1+\log T}}{(1-\beta_1)^2(1-\beta_1/\sqrt{\beta_2})\sqrt{(1-\beta_2)}}\sum_{i=1}^{d}\|\alpha_{1:T,i}\|_2 + C(\alpha_t)\sum_{t=1}^{T}\sqrt{\|\tilde{w}_t - w^*\|_{\sqrt{\tilde{V}_{t-1}}}}$$

By the definition of $\hat{v}_{t,i}$ with $\frac{\hat{v}_{t,i}^{1/2}}{\eta_t} \geq \frac{\hat{v}_{t-1,i}^{1/2}}{\eta_{t-1}}$ and using the $\|C(\alpha_t)\| \leq L_\infty$ on the feasible region and making use of the above property in Eq. 11, we have

$$R_T = \sum_{t=1}^{T} f_t(\tilde{w}_t) - f_t(w^*) \leq \frac{D_\infty^2\sqrt{T}}{\eta(1-\beta_1)}\sum_{i=1}^{d}\hat{v}_{T,i}^{1/2} + \frac{D_\infty^2}{2(1-\beta_1)}\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\beta_{1t}\hat{v}_{T,i}^{1/2}}{\eta_t} + \frac{\eta\sqrt{1+\log T}}{(1-\beta_1)^2(1-\beta_1/\sqrt{\beta_2})\sqrt{(1-\beta_2)}}\sum_{i=1}^{d}\|\alpha_{1:T,i}\|_2 + L_\infty D_\infty\sum_{t=1}^{T}\sqrt{\|w_t - \alpha^*\|_{\sqrt{\tilde{V}_{t-1}}}}$$

we further consider the quantization property as

$$\tilde{w}_t = \alpha_t sign(w);\ s.t.\ sign(w) \in (\mathcal{S}^d)$$

Then we have

$$R_T = \sum_{t=1}^{T} f_t(\tilde{w}_t) - f_t(w^*) \leq \frac{D_\infty^2\sqrt{T}}{\eta(1-\beta_1)}\sum_{i=1}^{d}\hat{v}_{T,i}^{1/2} + \frac{D_\infty^2}{2(1-\beta_1)}\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\beta_{1t}\hat{v}_{T,i}^{1/2}}{\eta_t} + \frac{\eta\sqrt{1+\log T}}{(1-\beta_1)^2(1-\beta_1/\sqrt{\beta_2})\sqrt{(1-\beta_2)}}\sum_{i=1}^{d}\|\alpha_{1:T,i}\|_2 + L_\infty D_\infty\sum_{t=1}^{T}\sqrt{D_\infty + \alpha^2 d_{\tilde{w}}^2}$$

The equality follows from simple telescopic sum, which yields the regret of BAMSProd to be bounded by $O(G_\infty\sqrt{T})$. It is not hard to see that. Thus, the regret of BAMSProd is upper bounded by minimum of $O(G_\infty\sqrt{T})$ and bound in the Theorem 4 and therefore, the worst case dependence of regret on T in our case is $O(\sqrt{T})$.