Appendix for

Exploring the Inefficiency of Heavy Ball as Momentum Parameter Approaches 1

A Proofs

A.1 Proof of the Main Techniques in Section 3

Let

$$\mathbf{y}^k := \begin{bmatrix} \mathbf{w}^k - \mathbf{w}^* \\ \mathbf{w}^{k-1} - \mathbf{w}^* \end{bmatrix} \in \mathbb{R}^{2d}.$$

According to the fact $\nabla R_S(\mathbf{w}^*) = \mathbf{0}$ and the iterative format (1) of SHB, we have

$$\begin{aligned} \mathbf{w}^{k+1} - \mathbf{w}^* &= \mathbf{w}^k - \mathbf{w}^* - \gamma(\nabla R_S(\mathbf{w}^k) - \nabla R_S(\mathbf{w}^*)) + \beta(\mathbf{w}^k - \mathbf{w}^*) - \beta(\mathbf{w}^{k-1} - \mathbf{w}^*) - \gamma(\mathbf{g}^k - \nabla R_S(\mathbf{w}^k)) \\ &= \mathbf{w}^k - \mathbf{w}^* - \gamma\mathbf{A}(\mathbf{w}^k - \mathbf{w}^*) + \beta(\mathbf{w}^k - \mathbf{w}^*) - \beta(\mathbf{w}^{k-1} - \mathbf{w}^*) - \gamma(\mathbf{g}^k - \nabla R_S(\mathbf{w}^k)), \end{aligned} \tag{8}$$

where $\mathbf{A} := \nabla^2 R_S$. Then SHB can be reformulated as

$$\mathbf{y}^{k+1} = \mathcal{T}\mathbf{y}^k - \gamma\mathbf{e}^k,$$

where $\mathcal{T} := \begin{bmatrix} (1+\beta)\mathbf{I} - \gamma\mathbf{A} & -\beta\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix}$ and $\mathbf{e}^k := \begin{bmatrix} \mathbf{g}^k - \nabla R_S(\mathbf{w}^k) \\ \mathbf{0} \end{bmatrix}$. We then have

$$\mathbf{y}^{k+1} = \mathcal{T}^k\mathbf{y}^1 - \gamma\sum_{i=1}^{k}\mathcal{T}^{k-i}\mathbf{e}^i.$$

Using the fact that $\mathbb{E}\langle\mathbf{e}^i, \mathbf{e}^j\rangle = 0$ if $i \neq j$ and $\mathbb{E}[\mathbf{e}^i] = 0, \forall i$, we have

$$\mathbb{E}\|\mathbf{y}^{k+1}\|^2 = \mathbb{E}\|\mathcal{T}^k\mathbf{y}^1 - \gamma\sum_{i=1}^{k}\mathcal{T}^{k-i}\mathbf{e}^i\|^2 = \mathbb{E}\|\mathcal{T}^k\mathbf{y}^1\|^2 + \gamma^2\sum_{i=1}^{k}\|\mathcal{T}^{k-i}\mathbf{e}^i\|^2. \tag{9}$$

A.2 Proof of Lemma 1

We need to exploit the eigenvalues of the matrix $\mathcal{T} = \begin{bmatrix} (1+\beta)\mathbf{I} - \gamma\mathbf{A} & -\beta\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{2d\times 2d}$, i.e., the complex number λ satisfying

$$\det\begin{pmatrix} (\lambda - 1 - \beta)\mathbf{I} + \gamma\mathbf{A} & \beta\mathbf{I} \\ -\mathbf{I} & \lambda\mathbf{I} \end{pmatrix} = 0.$$

Then we have

$$\det\begin{pmatrix} (\lambda + \frac{\beta}{\lambda} - 1 - \beta)\mathbf{I} + \gamma\mathbf{A} & \mathbf{0} \\ -\mathbf{I} & \lambda\mathbf{I} \end{pmatrix} = 0 \quad \Longrightarrow \quad \det((\lambda + \frac{\beta}{\lambda})\mathbf{I} - [(1+\beta)\mathbf{I} - \gamma\mathbf{A}]) = 0.$$

If λ^* is an eigenvalue of $\mathbf{A}$, we just need to consider

$$\lambda + \frac{\beta}{\lambda} = (1+\beta) - \gamma\lambda^*. \tag{10}$$

Let $\mathbf{U} := [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_d]$ be the eigenvectors of $\mathbf{A}$, it then holds that

$$\langle\mathbf{u}_i, \mathbf{u}_j\rangle = 0 \quad \text{if } i \neq j, \tag{11}$$

since $\mathbf{A}$ is symmetric positive definite. It is easy to see that $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_d$ are the eigenvectors of $(1+\beta)\mathbf{I} - \gamma\mathbf{A}$. Let λ_i be the ith eigenvalue of $\mathbf{A}$. With $0 < \nu \leq \lambda_{\min}(\mathbf{A})$, $\beta = (1 - \sqrt{\gamma\nu})^2 + \varrho$ and $0 < \varrho \ll \epsilon$, we can derive

$$(1 + \beta - \gamma\lambda_i)^2 - 4\beta \leq (1 + \beta - \gamma\nu)^2 - 4\beta \leq 0.$$

Thus, we define $\overline{\lambda_i}$ and $\underline{\lambda_i}$ as follows

$$\begin{aligned} \overline{\lambda_i} &:= \frac{(1+\beta-\gamma\lambda_i) + \sqrt{4\beta - (1+\beta-\gamma\lambda_i)^2}\mathbf{i}}{2}, \\ \underline{\lambda_i} &:= \frac{(1+\beta-\gamma\lambda_i) - \sqrt{4\beta - (1+\beta-\gamma\lambda_i)^2}\mathbf{i}}{2}, \end{aligned}$$

where $\mathbf{i}^2 = -1$. Direct calculating gives us

$$\mathcal{T}\begin{pmatrix} \overline{\lambda_i}\mathbf{u}_i \\ \mathbf{u}_i \end{pmatrix} = \overline{\lambda_i}\begin{pmatrix} \overline{\lambda_i}\mathbf{u}_i \\ \mathbf{u}_i \end{pmatrix}, \mathcal{T}\begin{pmatrix} \underline{\lambda_i}\mathbf{u}_i \\ \mathbf{u}_i \end{pmatrix} = \underline{\lambda_i}\begin{pmatrix} \underline{\lambda_i}\mathbf{u}_i \\ \mathbf{u}_i \end{pmatrix}.$$

Therefore, all the eigenvectors of $\mathcal{T}$ can be written as

$$\left\{\begin{pmatrix} \overline{\lambda_i}\mathbf{u}_i \\ \mathbf{u}_i \end{pmatrix}, \begin{pmatrix} \underline{\lambda_i}\mathbf{u}_i \\ \mathbf{u}_i \end{pmatrix}\right\}_{1\leq i\leq d}.$$

From (11), if $i \neq j$, we have

$$\left\langle\begin{pmatrix} \overline{\lambda_i}\mathbf{u}_i \\ \mathbf{u}_i \end{pmatrix}, \begin{pmatrix} \underline{\lambda_i}\mathbf{u}_j \\ \mathbf{u}_j \end{pmatrix}\right\rangle = 0.$$

Since $\beta = (1-\sqrt{\gamma\nu})^2 + \varrho$, we know that $\overline{\lambda_i} \neq \underline{\lambda_i}$, which means the matrix $\mathcal{T}$ has $2d$ different eigenvalues. Denote that

$$\overline{\Lambda} := \mathrm{Diag}(\overline{\lambda_1}, \overline{\lambda_2}, \dots, \overline{\lambda_d}), \;\; \underline{\Lambda} := \mathrm{Diag}(\underline{\lambda_1}, \underline{\lambda_2}, \dots, \underline{\lambda_d}).$$

Let $\overline{\mathbf{u}}_i := \begin{pmatrix} \overline{\lambda_i}\mathbf{u}_i \\ \mathbf{u}_i \end{pmatrix}, \underline{\mathbf{u}}_i := \begin{pmatrix} \underline{\lambda_i}\mathbf{u}_i \\ \mathbf{u}_i \end{pmatrix}$, we have $\mathcal{T}\overline{\mathbf{u}}_i = \overline{\lambda_i}\overline{\mathbf{u}}_i, \mathcal{T}\underline{\mathbf{u}}_i = \underline{\mathbf{u}}_i$. Then we construct the following matrix

$$\mathcal{U} := [\overline{\mathbf{u}}_1, \overline{\mathbf{u}}_2, \dots, \overline{\mathbf{u}}_d, \underline{\mathbf{u}}_1, \underline{\mathbf{u}}_2, \dots, \underline{\mathbf{u}}_d] = \begin{pmatrix} \overline{\Lambda}\mathbf{U} & \underline{\Lambda}\mathbf{U} \\ \mathbf{U} & \mathbf{U} \end{pmatrix},$$

and the matrix $\mathcal{U}$ is invertible. Then the matrix $\mathcal{T}$ satisfies

$$\mathcal{T}\mathcal{U} = \mathcal{U}\begin{bmatrix} \overline{\Lambda} & \\ & \underline{\Lambda} \end{bmatrix} \quad \Longrightarrow \quad \mathcal{T} = \mathcal{U}\begin{bmatrix} \overline{\Lambda} & \\ & \underline{\Lambda} \end{bmatrix}\mathcal{U}^{-1}. \tag{12}$$

Further, we have

$$\mathcal{T}^k = \mathcal{U}\underbrace{\begin{bmatrix} \overline{\Lambda}^k & \\ & \underline{\Lambda}^k \end{bmatrix}}_{:=\Lambda^k}\mathcal{U}^{-1}.$$

We are then led to

$$\|\mathcal{T}^k\| = \|\mathcal{U}\Lambda^k\mathcal{U}^{-1}\| \leq \|\mathcal{U}\|_F\|\mathcal{U}^{-1}\|_F \cdot 2d|\lambda_{\max}|^k, \tag{13}$$

where $\lambda_{\max} = \max\{\overline{\lambda_i}, \underline{\lambda_i}\}_{1\leq i\leq d}$ and we use the fact that $\|\mathbf{M}\mathbf{N}\|_F \leq \max\{\|\mathbf{M}\|_F\|\mathbf{N}\|, \|\mathbf{M}\|\|\mathbf{N}\|_F\}$. When $\beta = (1-\sqrt{\gamma\nu})^2 + \varrho$ and $0 < \varrho \ll \epsilon$,

$$|\lambda_{\max}| \leq 1 - \sqrt{\gamma\nu} + \varrho.$$

Direct calculation yields

$$\mathcal{U}^{-1} := \begin{pmatrix} \mathbf{U}^\top(\overline{\Lambda}-\underline{\Lambda})^{-1} & -\mathbf{U}^\top(\overline{\Lambda}-\underline{\Lambda})^{-1}\underline{\Lambda} \\ -\mathbf{U}^\top(\overline{\Lambda}-\underline{\Lambda})^{-1} & \mathbf{U}^\top(\overline{\Lambda}-\underline{\Lambda})^{-1}\overline{\Lambda} \end{pmatrix}. \tag{14}$$

From the definition of $\overline{\Lambda}, \underline{\Lambda}$, we have

$$[(\overline{\Lambda}-\underline{\Lambda})^{-1}]_{i,i} = ([\overline{\Lambda}-\underline{\Lambda}]_{i,i})^{-1} = -\frac{1}{\sqrt{4\beta-(1+\beta-\gamma\lambda_i)^2}}\mathbf{i} \approx -\frac{1}{2\sqrt{\gamma\lambda_i}}\mathbf{i}.$$

The approximation here is due to that $\epsilon > 0$ is small enough and $\gamma = \Theta(\epsilon)$, and $0 \leq \beta = (1-\sqrt{\gamma\nu})^2 + \varrho < 1$ for $0 < \varrho \ll \epsilon$. That means

$$(\overline{\Lambda}-\underline{\Lambda})^{-1} = \Theta(\frac{-\mathbf{i}}{\sqrt{\gamma\nu}})\mathbf{I}.$$

Noticing that β is very close to 1 and ϵ is very small, we have $\overline{\lambda_i} \approx 1, \underline{\lambda_i} \approx 1$ and

$$\underline{\Lambda} \approx \mathbf{I}, \; \overline{\Lambda} \approx \mathbf{I}.$$

Turning back to $\mathcal{U}$ and $\mathcal{U}^{-1}$, we see that $\|\mathcal{U}\|_F = \mathcal{O}(1)$ and $\|\mathcal{U}^{-1}\|_F = \Theta(\frac{1}{\sqrt{\gamma\nu}})$. Substituting the above result into inequality (13), we have

$$\|\mathcal{T}^k\| \leq \frac{C_1}{\sqrt{\gamma\nu}} \cdot (1-\sqrt{\gamma\nu})^k,$$

where $C_1 > 0$ is a constant. The proof is completed.

A.3 Proof of Lemma 2

If $\beta = 1 - \Theta(\gamma^\tau)$ and $\tau \geq 1$, we have $\beta \geq (1-\sqrt{\gamma\nu})^2$ when γ is small. And $(1+\beta) - \gamma\nu \leq 2\sqrt{\beta}$ holds, then the equation (10) has complex roots whose norms are $\sqrt{\beta}$. That is, the eigenvalues $\{\overline{\lambda}_i, \underline{\lambda}_i\}_{1\leq i\leq d}$ of the matrix $\mathcal{T}$ satisfy

$$|\overline{\lambda}_i|, |\underline{\lambda}_i| = \sqrt{\beta} \geq 1 - \Theta(\gamma^\tau), \ \ 1 \leq i \leq d.$$

With such a choice, we still have $\overline{\Lambda} \approx \underline{\Lambda} \approx \mathbf{I}$. Let $\xi = (\xi_1, \mathbf{0})^\top \in \mathbb{R}^{2d}$ and $\xi_1 \in \mathbb{R}^d \sim \mathcal{E}$. Denote $\bar{\xi} := \mathcal{U}^{-1}\xi$, we then have

$$\mathbb{E}\|\mathcal{T}^k\xi\|^2 = \mathbb{E}\|\mathcal{U}\Lambda^k\bar{\xi}\|^2 \geq \mathbb{E}\|\Lambda^k\bar{\xi}\|^2/\|\mathcal{U}^{-1}\|_F^2 \geq [1-\Theta(\gamma^\tau)]^{2k}\mathbb{E}\|\bar{\xi}\|^2/\|\mathcal{U}^{-1}\|_F^2,$$

Here, the norm $\|\cdot\|_F$ and $\|\cdot\|$ are taken on the complex domain. Recall the definition of $\mathcal{U}^{-1}$ in (14) and Assumption 2, we know

$$\mathbb{E}\|\bar{\xi}\|^2 = \mathbb{E}\Big\|\begin{bmatrix} \mathbf{U}^\top(\overline{\Lambda}-\underline{\Lambda})^{-1}\xi_1 \\ -\mathbf{U}^\top(\overline{\Lambda}-\underline{\Lambda})^{-1}\xi_1 \end{bmatrix}\Big\|^2 \geq \mathbb{E}\|\mathbf{U}^\top(\overline{\Lambda}-\underline{\Lambda})^{-1}\xi_1\|^2 = \mathrm{Tr}(\mathbf{U}^\top(\overline{\Lambda}-\underline{\Lambda})^{-1}\xi_1\xi_1^\top(\overline{\Lambda}-\underline{\Lambda})^{-1}\mathbf{U}) = \mathrm{Tr}\big((\overline{\Lambda}-\underline{\Lambda})^{-2}\Sigma\big),$$

and $\|\mathcal{U}^{-1}\|_F^2 = \Theta(\|(\overline{\Lambda}-\underline{\Lambda})^{-1}\|^2)$. Then for some $C_2 \geq 0$, we are arrive at

$$\mathbb{E}\|\mathcal{T}^k\xi\|^2 \geq C_2(1-\Theta(\gamma^\tau))^{2k}.$$

A.4 Proof of Lemma 3

Let λ_i be the ith eigenvalue of $\mathbf{A}$ and $0 \leq \beta \leq \beta_0 < 1, 1-\beta_0 \gg \epsilon$, we can see that

$$(1+\beta-\gamma\lambda_i)^2 - 4\beta \geq (1+\beta-\gamma L)^2 - 4\beta \geq 0.$$

Thus, we define $\overline{\lambda_i}$ and $\underline{\lambda_i}$ as follows

$$\overline{\lambda_i} := \frac{(1+\beta-\gamma\lambda_i) + \sqrt{(1+\beta-\gamma\lambda_i)^2-4\beta}}{2},$$
$$\underline{\lambda_i} := \frac{(1+\beta-\gamma\lambda_i) - \sqrt{(1+\beta-\gamma\lambda_i)^2-4\beta}}{2}.$$

Noticed that $\gamma = \Theta(\epsilon)$, we have

$$[(\overline{\Lambda}-\underline{\Lambda})^{-1}]_{i,i} = ([\overline{\Lambda}-\underline{\Lambda}]_{i,i})^{-1} = \frac{1}{\sqrt{(1+\beta-\gamma\lambda_i)^2-4\beta}} \approx \frac{1}{1-\beta}.$$

The approximation here is due to that γ is small enough and β is not close to 1. That means

$$\|(\overline{\Lambda}-\underline{\Lambda})^{-1}\| = \Theta(\frac{1}{1-\beta_0}), \ \|\overline{\Lambda}\| = \mathcal{O}(1), \ \|\underline{\Lambda}\| = \mathcal{O}(1).$$

Then we have $\|\mathcal{U}\|_F = \mathcal{O}(1)$ and $\|\mathcal{U}^{-1}\|_F = \Theta(\frac{1}{1-\beta_0})$. On the other hand, the eigenvalues of the matrix $\mathcal{T}$ satisfy

$$\frac{(1+\beta-\gamma\lambda_i) + \sqrt{(1+\beta-\gamma\lambda_i)^2-4\beta}}{2} \leq 1 - \frac{\gamma\lambda_i}{1-\beta} + C_3\epsilon^2,$$

Here, we used $\gamma = \Theta(\epsilon)$ and the Taylor expansion for $\gamma\lambda_i$. Then we have

$$|\lambda_{\max}| \leq 1 - \frac{\gamma\nu}{1-\beta} + C_3\epsilon^2.$$

Now we can derive

$$\|\mathcal{T}^k\| \leq \|\mathcal{U}\|_F\|\mathcal{U}^{-1}\|_F \cdot 2d|\lambda_{\max}|^k \leq C_4(1 - \frac{\gamma\nu}{1-\beta} + C_3\epsilon^2)^k,$$

where constants $C_3, C_4 > 0$ are independent of k and γ.

A.5 Proof of Theorem 1

Noticed that $\mathbf{y}^k = \begin{bmatrix} \mathbf{w}^k - \mathbf{w}^* \\ \mathbf{w}^{k-1} - \mathbf{w}^* \end{bmatrix}$. Combing the equation (9) and Lemma 1, it follows

$$\begin{aligned}\mathbb{E}\|\mathbf{w}^{k+1}-\mathbf{w}^*\|^2 \leq \mathbb{E}\|\mathbf{y}^{k+1}\|^2 &= \mathbb{E}\|\mathcal{T}^k\mathbf{y}^1\|^2 + \gamma^2\sum_{i=1}^k\|\mathcal{T}^{k-i}\mathbf{e}^i\|^2 \\ &\leq \frac{C_1^2}{\gamma\nu}(1-\sqrt{\gamma\nu})^{2k}\|\mathbf{y}^1\|^2 + \frac{\gamma C_1^2\sigma^2}{\nu}\sum_{i=1}^k(1-\sqrt{\gamma\nu})^{2k-2i}. \end{aligned} \tag{15}$$

When γ is small, $\sum_{i=1}^k(1-\sqrt{\gamma\nu})^{2k-2i} \leq \frac{1}{\sqrt{\gamma\nu}}$, we then proved the result.

A.6 Proof of Theorem 2

Noticing that with Lemma 2, it holds $\mathbb{E}\|\mathcal{T}^{k-i}\mathbf{e}^i\|^2 \geq C_2(1-\Theta(\gamma^\tau))^{2k-2i}$. Stating from (9), we are then led to

$$\mathbb{E}\|\mathbf{y}^k\|^2 \geq \mathbb{E}\|\mathcal{T}^k\mathbf{y}^1\|^2 + C_2\gamma^2\sum_{i=1}^{k}[1-\Theta(\gamma^\tau)]^{2k-2i} = \Theta(\gamma^{2-\tau}).$$

The above equation indicates that

$$\mathbb{E}\|\mathbf{w}^k - \mathbf{w}^*\|^2 + \mathbb{E}\|\mathbf{w}^{k-1} - \mathbf{w}^*\|^2 \geq \Theta(\gamma^{2-\tau}). \tag{16}$$

According to (16), if we set $\gamma = \Theta(\epsilon)$, the lower bound is in the order of $\Theta(\epsilon^{2-\tau})$.

A.7 Proof of Theorem 3

Note that the equality (9) still holds. With Lemma 3, we have

$$\mathbb{E}\|\mathbf{w}^K - \mathbf{w}^*\|^2 \leq C_4^2\Big(1-\frac{\gamma\nu}{1-\beta}+C_3\epsilon^2\Big)^{2K}\|\mathbf{y}^1\|^2 + \gamma^2C_4^2\sigma^2\sum_{i=1}^{K}\Big(1-\frac{\gamma\nu}{1-\beta}+C_3\epsilon^2\Big)^{2K-2i}.$$

When $\Theta(\epsilon)$ and ϵ is small,

$$\gamma^2C_4^2\sigma^2\sum_{i=1}^{K}\Big(1-\frac{\gamma\nu}{1-\beta}+C_3\epsilon^2\Big)^{2K-2i} = C_4^2\sigma^2\frac{1-\beta}{\nu}\gamma + \mathcal{O}(\epsilon^2) = \mathcal{O}(\epsilon).$$

If $\mathbb{E}\|\mathbf{w}^K - \mathbf{w}^*\|^2 \leq \epsilon$, we then have

$$\Big(1-\frac{\gamma\nu}{1-\beta}+C_3\epsilon^2\Big)^{2K} = \mathcal{O}(\epsilon).$$

The worst case is then

$$\mathcal{O}(\frac{\ln\frac{1}{\epsilon}}{\frac{\gamma\nu}{1-\beta}-C_3\epsilon^2}) = \widetilde{\mathcal{O}}(\frac{1-\beta}{\epsilon\nu}).$$

B Experiments

Model \ Seed	Algorithm	1	19	31	42	80	Average	STDEV
ResNet18	SGD	92.23	92.33	92.82	92.67	93.26	92.66	0.41
	Adam	91.21	91.31	90.94	90.99	90.97	91.08	0.17
	SHB($\beta = 0.9$)	92.81	93.26	93.40	92.97	93.87	93.26	0.41
	SHB-DW	92.62	93.67	93.55	93.13	93.95	93.38	0.52
ResNet34	SGD	92.84	92.76	92.11	91.42	91.79	92.18	0.61
	Adam	91.65	91.44	91.60	91.23	91.53	91.49	0.17
	SHB($\beta = 0.9$)	93.58	93.44	92.51	92.52	92.70	92.95	0.52
	SHB-DW	93.43	93.72	93.36	92.80	92.78	93.22	0.41

Table 2: Supplement to Table 1. Test accuracy (%) of ResNet18 and ResNet34 for CIFAR10 classification, where the models are trained by SGD, Adam, SHB and SHB-DW algorithms. Each experiment was repeated five times for different seeds.

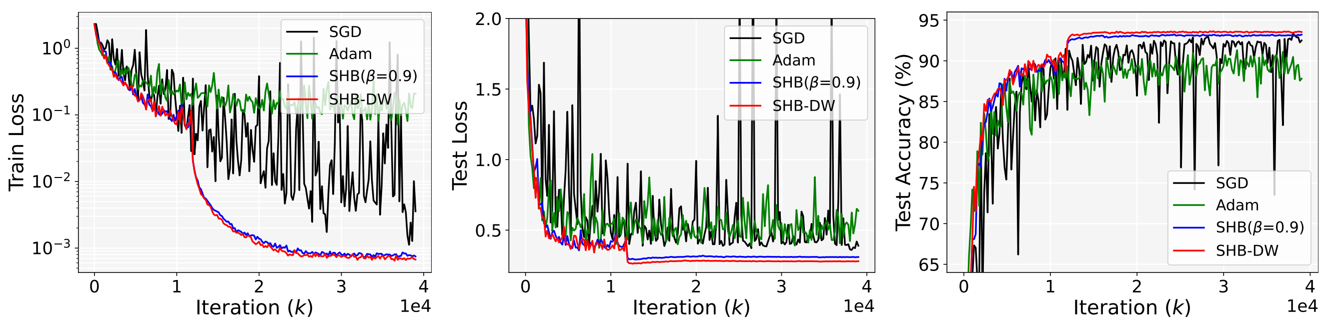

Figure 8: Supplement to Table 1. Training of ResNet18 for CIFAR10 classification using SGD, Adam, SHB ($\beta = 0.9$), and SHB-DW. All the algorithms are run for 200 epochs with a batch size of 256. The initial learning rate for Adam is set to 0.001, while the others are set to 0.1. All algorithms use a decreasing learning rate strategy, i.e., decreasing by a factor of 10 at the 60th, 120th and 180th epochs, respectively.

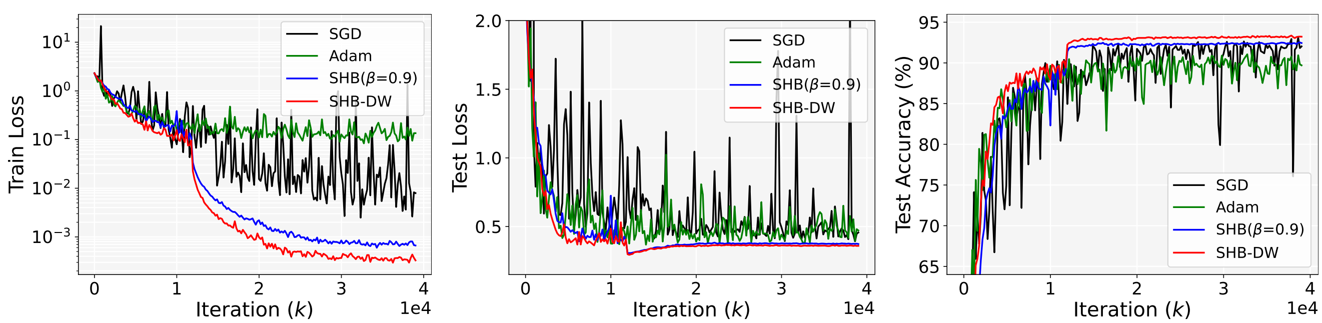

Figure 9: Supplement to table 1 and Figure 7. Training of ResNet34 for CIFAR10 classification using SGD, Adam, SHB ($\beta = 0.9$), and SHB-DW. All the algorithms are run for 200 epochs with a batch size of 256. The initial learning rate for Adam is set to 0.001, while the others are set to 0.1. All algorithms use a decreasing learning rate strategy, i.e., decreasing by a factor of 10 at the 60th, 120th and 180th epochs, respectively.