Analog In-memory Training on General Non-ideal Resistive Elements: The Impact of Response Functions

Zhaoxian Wu*, **Quan Xiao***, **Tayfun Gokmen**♯, **Omobayode Fagbohungbe**♯, **Tianyi Chen***
*Cornell University, New York, NY
♯IBM T. J. Watson Research Center, Yorktown Heights, NY
{zw868, qx232}@cornell.edu,
{tgokmen, Omobayode.Fagbohungbe}@us.ibm.com,
tianyi.chen@cornell.edu

Abstract

As the economic and environmental costs of training and deploying large vision or language models increase dramatically, analog in-memory computing (AIMC) emerges as a promising energy-efficient solution. However, the training perspective, especially its training dynamic, is underexplored. In AIMC hardware, the trainable weights are represented by the conductance of resistive elements and updated using consecutive electrical pulses. While the conductance changes by a constant in response to each pulse, in reality, the change is scaled by asymmetric and non-linear *response functions*, leading to a non-ideal training dynamic. This paper provides a theoretical foundation for gradient-based training on AIMC hardware with non-ideal response functions. We demonstrate that asymmetric response functions negatively impact `Analog SGD` by imposing an implicit penalty on the objective. To overcome the issue, we propose `residual learning` algorithm, which provably converges exactly to a critical point by solving a bilevel optimization problem. We demonstrate that the proposed method can be extended to address other hardware imperfections, such as limited response granularity. As we know, it is the first paper to investigate the impact of a class of generic non-ideal response functions. The conclusion is supported by simulations validating our theoretical insights.

1 Introduction

The remarkable success of large vision and language models is underpinned by advances in modern hardware accelerators, such as GPU, TPU [1], NPU [2], and NorthPole chip [3]. However, the computational demands of training these models are staggering. For instance, training LLaMA [4] cost \$2.4 million, while training GPT-3 [5] required \$4.6 million, highlighting the urgent need for more efficient computing hardware. Current mainstream hardware relies on the Von Neumann architecture, where the physical separation of memory and processing units creates a bottleneck due to frequent and costly data movement between them.

In this context, the industry has turned its attention to *analog in-memory computing (AIMC) accelerators* based on resistive crossbar arrays [6–10], which excel at accelerating the ubiquitous and computationally intensive matrix-vector multiplications (MVMs) operations. In AIMC hardware, the weights (matrices) are represented by the conductance states of the *resistive elements* in analog crossbar arrays [11, 12], while the input and output of MVM are analog signals like voltage and

*The work was done when the authors were at Rensselaer Polytechnic Institute, and was supported by IBM through the IBM-Rensselaer Future of Computing Research Collaboration.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

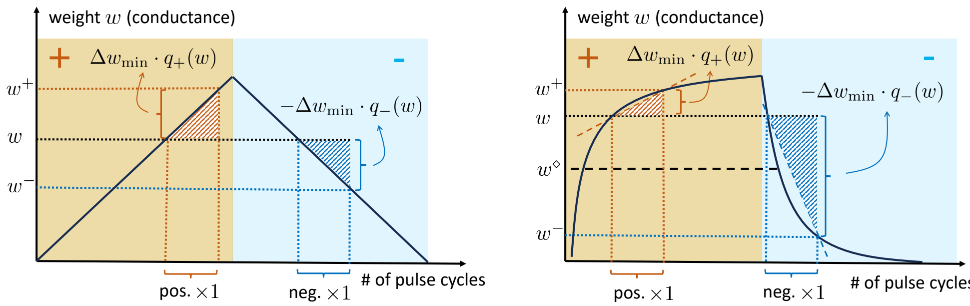

Figure 1: The weight's response curve. Positive and negative pulses are fired continuously on the left and right halves, respectively. One pulse is fired per cycle. Given w, the weight becomes w^+ or w^- after one positive and negative pulse, respectively. The response factors $q_+(w)$ and $q_-(w)$ are approximately the slope of the curve at w, and $\Delta w_{\min}$ is the response granularity. **(Left)** Ideal response functions $q_+(w) \equiv q_-(w)$. Every point is a symmetric point. **(Right)** Asymmetric response functions $q_+(w) \neq q_-(w)$ almost everywhere expect for the symmetric point $w^\diamond$.

current. Leveraging Kirchhoff's and Ohm's laws, AIMC hardware achieves 10×-10,000× energy efficiency than GPU [13–15] in model inference.

Despite its high efficiency, *analog training* is considerably more challenging than *inference* since it involves frequent weight updates. Unlike digital hardware, where the weight increment can be applied to the original weight in the memory cell, the weights in AIMC hardware are changed by the so-called *pulse update*. **Pulse update.** When receiving electrical pulses from its peripheral circuits, the resistive elements change their conductance in response to the pulse polarity [16]. Receiving a pulse at each pulse cycle, the conductance is updated by a small amount around $\Delta w_{\min}$, which is called *response granularity*. $\Delta w_{\min}$ is scaled by a *response functions* $q_+(w)$ or $q_-(w)$ depending on its polarity. Geometrically, $q_+(w)$ and $q_-(w)$ are the slopes of *response curves*; see Figure 1. All $\Delta w_{\min}$, $q_+(w)$, and $q_-(w)$ are element-specific parameters or functions that are set before training and hence remain fixed during training. Typically, $\Delta w_{\min}$ is known while $q_+(w)$ and $q_-(w)$ are not.

Gradient-based training implemented by analog update. Supported by pulse update, the gradient-based training algorithms are used to optimize the weights. Consider a standard training problem with objective $f(\cdot) : \mathbb{R}^D \to \mathbb{R}$ and a model parameterized by $W \in \mathbb{R}^D$

$$W^* := \underset{W \in \mathbb{R}^D}{\arg\min}\ f(W) := \mathbb{E}_\xi[f(W; \xi)] \tag{1}$$

where ξ is a random data sample. Similar to stochastic gradient descent (SGD) in digital training (`Digital SGD`), the gradient-based training algorithm on AIMC hardware, `Analog SGD`, updates the weights by stochastic gradients $\nabla f(W_k; \xi_k)$. `Digital SGD` updates the weight by $W_{k+1} = W_k - \alpha \nabla f(W_k; \xi_k)$ with learning rate α. Given a *desired update* $\Delta W = -\alpha \nabla f(W_k; \xi_k)$, AIMC hardware implements `Analog SGD` by sending about $|[\Delta W]_i| / \Delta w_{\min}$ pulses to the i-th element. With each pulse updating $[W_k]_i$ by $\Delta w_{\min}$, $[W_k]_i$ is ultimately updated by about $[\Delta W]_i$.

Challenges of analog training. Despite its ultra-efficiency, gradient-based training on AIMC hardware is challenging. First, the generic response functions are *asymmetric* (i.e. $q_+(w) \not\equiv q_-(w)$), and *non-linear* [17–19]. Due to the variation of response functions and conductance states, gradients are scaled by different magnitudes (i.e., $q_s([W_k]_i)$ for $s = \pm$) across different coordinates, leading to biased gradients. Furthermore, the response granularity $\Delta w_{\min}$ is a constant. When the gradients or the learning rate decay below $\Delta w_{\min}$, pulse update no longer provides sufficient precision to perform gradient descent [20]. Other imperfections include, but are not limited to, input/output (IO) of MVM operations and analog-digital conversion error [18]. This paper aims to investigate the impact of non-ideal response functions and develop a method to mitigate their negative effects. We also discuss extending the proposed method to deal with other hardware imperfections.

1.1 Main results

Complementing existing empirical studies in analog in-memory computing, this paper aims to build a rigorous theoretical foundation of analog training. By introducing bias to the gradient, the asymmetric response function plays a central role in differentiating digital and analog training. In contrast, the other non-idealities hinder the training process by causing precision-related issues. Therefore, we approach the problem progressively, beginning with a simplified case that involves only the asymmetric response functions, and extending the proposed methods to more general scenarios.

As a warm-up, building upon the pulse update mechanism, we propose the following discrete-time mathematical model to characterize the trajectory of `Analog SGD`

$$\texttt{Analog SGD} \qquad W_{k+1} = W_k - \alpha \nabla f(W_k; \xi_k) \odot F(W_k) - \alpha |\nabla f(W_k; \xi_k)| \odot G(W_k) \tag{2}$$

where $\alpha > 0$ is the learning rate and ξ_k is the data sample of iteration k; $|\cdot|$ and $\odot$ represent the element-wise absolute value and multiplication, respectively; and $F(\cdot)$ and $G(\cdot)$ are hardware-specific matrix which are defined by $q_+(\cdot)$ and $q_-(\cdot)$. In Section 2, we will explain the underlying rationale of (2). Compared with the standard `Digital SGD`, the gradients in (2) are scaled by $F(\cdot)$ and an extra bias term is introduced. Typically, hardware imperfections lead to non-ideal response functions, i.e., $F(\cdot) \not\equiv 1$ and $G(\cdot) \not\equiv 0$. Thus, we ask a natural question that

Q1) *What is the impact of non-ideal response functions and how to alleviate it?*

Recently, [21] partially answers the question by showing `Analog SGD` suffers from a convergence issue due to the asymmetric update, and a heuristic algorithm, `Tiki-Taka` [22–24], converges exactly by reducing the weight drift. However, their work is limited to a special case of *linear response*, which are in the form of $q_+(w) = 1 - w/\tau, q_-(w) = 1 + w/\tau$ with hardware-specific parameter $\tau > 0$. Given more general $q_+(w)$ and $q_-(w)$, the convergence of `Tiki-Taka` does not trivially holds, even though the response functions are still linear.

Gap between theory for special linear and generic response. Consider a more generic linear response setting $q_+(w) = (1 + c_{\texttt{Lin}})(1 - w/\tau), q_-(w) = (1 - c_{\texttt{Lin}})(1 + w/\tau)$ with a parameter $c_{\texttt{Lin}}$, which reduces to the setting in [21] when $c_{\texttt{Lin}} = 0$. Figure 2 shows the damage from a non-zero $c_{\texttt{Lin}}$ to `Tiki-Taka`. Consistent with the conclusion in [21], `Tiki-Taka` significantly outperforms `Analog SGD` when $c_{\texttt{Lin}} = 0$. However, when $c_{\texttt{Lin}}$ is perturbed from 0.1 to 0.3, `Tiki-Taka` degrades dramatically and even becomes worse than `Analog SGD` does. The modification is slight, but the convergence guarantee in [21] fails, and the convergence of `Tiki-Taka` is harmed significantly. This counter-example indicates a gap between the theory with special linear response and generic response, and necessitates the study of the analog training with generic response functions and the exploration of exact convergence conditions.

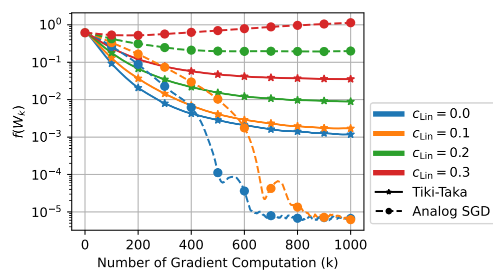

Figure 2: Comparison of `Analog SGD` and `Tiki-Taka` under different parameter $c_{\texttt{Lin}}$. The error plateau in the order 10^{-5} comes from the limited response granularity $\Delta w_{\min} = 10^{-4}$.

Ignoring other imperfections temporarily, we first analyze the impact of response functions. We show that `Analog SGD` suffers from asymptotic error due to the mismatch between the algorithmic *stationary point* and physical *symmetric point*. Inspired by that, we propose a novel algorithm framework that aligns two points, overcoming the asymmetric issues. Building on that, we endeavor to extend the proposed algorithm to more practical scenarios that involve other imperfections like limited granularity and noisy readings, prompting a second critical question:

Q2) *How to extend the framework to address the limited response granularity and noisy IO issues?*

To answer this question, we propose two mechanisms to further overcome these two issues.

Our contributions. This paper makes the following contributions:

- **C1)** Building upon the equation of pulse update, we propose an approximated discrete-time dynamics of analog update. Enabled by it, we study the impact of response functions directly without being limited to concrete element candidates.
- **C2)** Based on that, we show that instead of optimizing $f(\cdot)$, `Analog SGD` optimizes another penalized objective implicitly. An implicit penalty is introduced by the asymmetric response functions and attracts the weights towards symmetric points. Consequently, `Analog SGD` can only converge to the optimal point inexactly.
- **C3)** We propose a novel `Residual Learning` algorithm framework to alleviate the asymmetric update and implicit penalty issues. `Residual Learning` explicitly introduces another

residual array, which has a stationary point 0. By properly zero-shifting so that the stationary and symmetric points overlap, `Residual Learning` provably converges to a critical point.

C4) Building on C3), we propose a variant, `Residual Learning v2`, tailored for more practical training scenarios. We propose introducing a digital buffer to filter out the reading error due to the IO noise. Furthermore, we propose a threshold-based transfer rule to alleviate the instability due to the limited granularity.

1.2 Prior art

AIMC training. Analog training has shown promising early successes with tremendous energy advantage [25, 26]. Among them, on-chip training, which performs forward, backward, and update directly on analog chips [22–24, 27, 28] is considered to be the most efficient paradigm, but it is more sensitive to hardware imperfections. Sacrificing energy efficiency for robustness, hybrid digital-analog off-chip training is proposed [29–32], which offloads some computation burden to digital components. This paper focuses on the more challenging on-chip training setting.

Energy-based model and equilibrium propagation. AIMC training leverages back-propagation to compute the gradient signals. Recently, a class of energy-based models has been studied, which performs equilibrium propagation to compute gradient signals [33–37]. Focusing on the training dynamics instead of concrete gradient computing, our work is orthogonal to them and is expected to provide insight for algorithm designs of energy-based model training.

2 Analog Training with Generic Response Functions

This section examines the discrete-time dynamics of analog training and introduces the challenges posed by generic response functions. After that, we introduce a family of response functions that reflect crucial physical properties that interest us.

Compact formulations of analog update. We first investigate the dynamics of one element w in $W \in \mathbb{R}^D$. This paper adopts w to represent the element of the weight W_k without specifying its index. The notation makes the formulations more concise and indicates that all elements are updated in parallel. As we discuss in Section 1, the response granularity $\Delta w_{\min}$ is scaled by the response function $q_s(w)$. Naturally, given a desired update Δw, the increment is approximately scaled by $q_s(w)$ as well. Accordingly, we propose that an approximate dynamics of analog update is given by $w' \approx U_q(w, \Delta w)$, where $U_q(w, \Delta w)$ is defined by

$$U_q(w, \Delta w) := \begin{cases} w + \Delta w \cdot q_+(w), & \Delta w \geq 0, \\ w + \Delta w \cdot q_-(w), & \Delta w < 0. \end{cases} \tag{3}$$

The update (3) holds at each resistive element, while the model W contains D resistive elements with different response functions $q_+(\cdot)$ and $q_-(\cdot)$. We stack all the weights w_k and expected increment Δw_k together into vectors $W_k, \Delta W_k \in \mathbb{R}^D$, where k is the iteration index Similarly, the response functions $q_+(\cdot)$ and $q_-(\cdot)$ are stacked into $Q_+(\cdot)$ and $Q_-(\cdot)$, respectively. Let the notation $U_Q(W_k, \Delta W)$ on matrices W_k and ΔW denote the element-wise operation on W_k and ΔW, i.e. $[U_Q(W_k, \Delta W)]_i := U_{[Q]_i}([W_k]_i, [\Delta w]_i), \forall i \in [D]$ with $[D] := \{1, 2, \cdots, D\}$ denoting the index set. The element-wise update (3) can be expressed as $W_{k+1} = U_Q(W_k, \Delta W_k)$. Leveraging the symmetric decomposition [22, 21], we decompose $Q_-(W)$ and $Q_+(W)$ into symmetric component $F(\cdot)$ and asymmetric component $G(\cdot)$

$$F(W) := (Q_-(W) + Q_+(W))/2, \quad \text{and} \quad G(W) := (Q_-(W) - Q_+(W))/2, \tag{4}$$

which leads to a compact form of the `Analog Update`

$$\texttt{Analog Update} \qquad W_{k+1} = W_k + \Delta W_k \odot F(W_k) - |\Delta W_k| \odot G(W_k). \tag{5}$$

Gradient-based training algorithms on AIMC hardware. In (5), the desired update ΔW_k varies based on different algorithms. Replacing ΔW_k with the stochastic gradient $\nabla f(W_k; \xi_k)$, we obtain the dynamics of `Analog SGD` shown in (2). This update is reduced to the mathematical form for linear response functions in [21] as a special case; see Appendix B for details. **Response function class.** Before proceeding to the study of response functions, we first define the response function

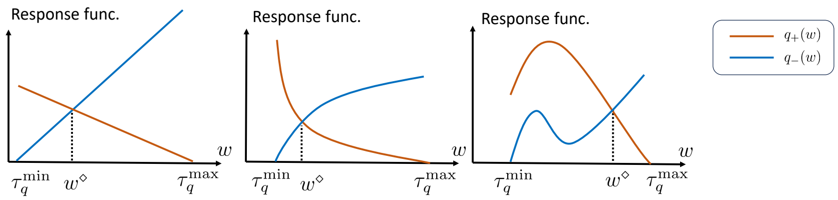

Figure 3: Examples of response functions from Definition 1; $w^\diamond$ is the symmetric point.

class that interests us. Since the behavior of resistive elements is always governed by physical laws, the function class should reflect certain crucial physical properties.

We first introduce a commonly observed *saturation* property across a wide range of resistive elements. Resistive elements get *saturated* when they keep receiving the same pulses to avoid reaching arbitrarily high or low conductance states. Constrained by that, the conductance of the resistive element with $q_+(\cdot)$ and $q_-(\cdot)$ is bounded inside a *dynamic range* $[\tau^{\min}, \tau^{\max}]$ where $\tau^{\min}$ and $\tau^{\max}$ are the saturation points with zero responses, i.e. $q_+(\tau^{\max}) = q_-(\tau^{\min}) = 0$. Near the saturation points, the asymmetric issue is significant, and thus, the update in one direction is suppressed, considerably impacting the convergence. On the contrary, if a point $w^\diamond$ satisfies $q_-(w^\diamond) = q_+(w^\diamond)$, the analog update behaves like a digital update. We refer to $w^\diamond$ as *symmetric point*. Symmetric points are typically located in the interior of the dynamic range and are far from saturation.

Stacking all $w^\diamond$ into a vector $W^\diamond \in \mathbb{R}^D$. Observe that the function $G(W)$ is large near the saturation points while almost zero around $W^\diamond$, implying it can reflect the saturation degree. At the same time, $F(W)$ is the average of the response functions in two directions. As we will see in Sections 3.2 and 4, the ratio $\frac{G(W)}{\sqrt{F(W)}}$ plays a critical role in the convergence behaviors.

Besides saturation, the function class should also enjoy other properties. First, the conductance ends up increasing when receiving a positive pulse and vice versa, leading to positive response functions. On top of that, we also assume the response functions are differentiable (and hence continuous) for mathematical tractability. Taking all into account, we define the following response function class.

Definition 1 (Response function class)**.** *$q_+(\cdot)$ and $q_-(\cdot)$ with dynamic range $[\tau^{\min}, \tau^{\max}]$ satisfy*

- ***(Positive-definiteness)*** *$q_+(w) > 0, \forall w < \tau^{\max}$ and $q_-(w) > 0, \forall w > \tau^{\min}$;*
- ***(Differentiable)*** *$q_+(\cdot)$ and $q_-(\cdot)$ are differentiable;*
- ***(Saturation)*** *$q_+(\tau^{\max}) = q_-(\tau^{\min}) = 0$.*

Definition 1 covers a wide range of response functions, including but not limited to PCM, ReRAM, ECRAM, and others mention in Section A. Figure 3 showcases three examples from the response functions class, including linear, non-linear but monotonic, and even non-monotonic functions.

3 Implicit Penalty and Inexact Convergence of `Analog SGD`

This section introduces a critical impact of the response functions, *implicit penalized objective*. Affected by it `Analog SGD` can only converge inexactly with a non-diminishing asymptotic error.

3.1 Implicit penalty

We first give an intuition through a situation where W_k is already a critical point, i.e., $\mathbb{E}_\xi[\nabla f(W_k; \xi)] = 0$. Recall that stochastic gradient descent on digital hardware (`Digital SGD`) is stable in expectation, i.e. $\mathbb{E}_{\xi_k}[W_{k+1}] = W_k - \mathbb{E}_{\xi_k}[\alpha \nabla f(W_k; \xi_k)] = W_k$. However, this does not work for `Analog SGD`

$$\begin{aligned}\mathbb{E}_{\xi_k}[W_{k+1}] &= W_k - \mathbb{E}_{\xi_k}[\alpha \nabla f(W_k; \xi_k) \odot F(W_k) - \alpha |\nabla f(W_k; \xi_k)| \odot G(W_k)] \\ &= W_k - \alpha \mathbb{E}_{\xi_k}[|\nabla f(W_k; \xi_k)| \odot G(W_k)] \neq W_k.\end{aligned} \tag{6}$$

Consider a simplified version that the weight is a scalar ($D = 1$) and the function $G(W)$ is strictly monotonically decreasing[2] to help us gain intuition on the impact of the drift in (6). Recall $G(W^\diamond) =$

[2]It happens when both $q_+(\cdot)$ and $q_-(\cdot)$ are strictly monotonic.

0 at the symmetric point $W^\diamond$. $G(W) > 0$ when $W > W^\diamond$ and $G(W) < 0$ otherwise. Consequently, (6) indicates that $\mathbb{E}_{\xi_k}[W_{k+1}] < W_k$ when $W_k > W^\diamond$ and $\mathbb{E}_{\xi_k}[W_{k+1}] > W_k$ otherwise. It implies that W_k suffers from a drift tendency towards $W^\diamond$. In addition, the penalty coefficient proportional to the noise level since the drift is proportional to $\mathbb{E}_{\xi_k}[|\nabla f(W_k;\xi_k)|]$, which is the first moment of noise $\mathbb{E}_{\xi_k}[|\nabla f(W_k;\xi_k) - \mathbb{E}_\xi[\nabla f(W_k;\xi)]|]$ in essence.

The following theorem formalizes the implicit penalty effect. Before that, we define an accumulated asymmetric function $R_c(\cdot) : \mathbb{R}^D \to \mathbb{R}^D$, whose derivative is $R(W) := \frac{G(W)}{F(W)}$, i.e. $\frac{\mathrm{d}[R_c(W)]_d}{\mathrm{d}[W]_d} = [R(W)]_d = \frac{[G(W)]_d}{[F(W)]_d}$. If $R(W)$ is strictly monotonic, $R_c(W)$ reaches its minimum at the symmetric point $W^\diamond$ where $R(W^\diamond) = 0$, so that it penalizes the weight away from the symmetric point.

Theorem 1 (Implicit penalty, short version)**.** *Suppose* $\mathbb{E}_{\xi_k}[|\nabla f(W_k;\xi_k) - \mathbb{E}_\xi[\nabla f(W_k;\xi)]|]$ *is a constant* $\Sigma \in \mathbb{R}^D$ *and let* $D = 1$. `Analog SGD` *implicitly optimizes the following penalized objective*

$$\min_W \; f_\Sigma(W) := f(W) + \langle \Sigma, R_c(W) \rangle \,. \tag{7}$$

The full version of Theorem 1 and its proof are deferred to Appendix G. In Theorem 1, $R_c(W)$ plays the role of a penalty to force the weight toward a symmetric point. As shown in Appendix G, $R_c(W)$ has a simple expression on linear response functions when $c_{\texttt{Lin}} = 0$, leading (7) to $\min_W \; f_\Sigma(W) := f(W) + \frac{\Sigma}{2\tau}\|W\|^2$ which is an ℓ_2 regularized objective. In addition, the implicit penalty has a coefficient proportional to the noise level Σ and inversely proportional to the dynamic range τ. It implies that the implicit penalty becomes active only when gradients are noisy, and the noise amplifies the effect.

> With noisy gradients, an **implicit penalty** attracts `Analog SGD` towards symmetric points.

3.2 Inexact Convergence of `Analog SGD` under generic devices

Due to the implicit penalty, `Analog SGD` only converges to a critical point inexactly. Before showing that, We introduce a series of assumptions on the objective, as well as noise.

Assumption 1 (Objective)**.** *The objective* $f(W)$ *is* L*-smooth and is lower bounded by* f^*.

Assumption 2 (Unbiasness and bounded variance)**.** *The stochastic gradient is unbiased and has bounded variance* σ^2.

Assumption 1–2 are standard in non-convex optimization [38, 21]. Additionally, similar to the setting in [21], we also assume that the saturation degree is bounded, i.e., all response functions are positive.

Assumption 3 (Bounded saturation)**.** *There exists a constant* $R_{\min} > 0$ *such that* $\min\{Q_+(W) \odot Q_-(W)\} > R_{\min}$.

Assumption 3 requires that W_k remains far from saturation points, which is a mild assumption in actual training. This paper considers the average gradient square norm as the convergence metric, given by $E_K := \frac{1}{K}\sum_{k=0}^{K-1}\|\nabla f(W_k)\|^2$. Now, we establish the convergence of `Analog SGD`.

Theorem 2 (Inexact convergence of `Analog SGD`)**.** *Under Assumption 1–3, if the learning rate is set as* $\alpha = O(1/\sqrt{K})$*, it holds that*

$$E_K \le O\left(\sqrt{\sigma^2/K} + \sigma^2 S_K^{\texttt{ASGD}}\right) \tag{8}$$

where $S_K^{\texttt{ASGD}}$ *denotes the amplification factor given by* $S_K^{\texttt{ASGD}} := \frac{1}{K}\sum_{k=0}^{K-1}\left\|\frac{G(W_k)}{\sqrt{F(W_k)}}\right\|_\infty^2$.

The proof of Theorem 2 is deferred to Appendix H. Theorem 2 suggests that the convergence metric E_K is upper bounded by two terms: the first term vanishes at a rate of $O(\sqrt{\sigma^2/K})$, which matches the `Digital SGD`'s convergence rate [38] up to a constant; the second term contributes to the *asymptotic error* of `Analog SGD`, which does not vanish with the number of iterations K.

Impact of saturation/asymmetric update. The exact expression of $S_K^{\texttt{ASGD}}$ depends on the specific noise distribution and thus is difficult to reach. However, $S_K^{\texttt{ASGD}}$ reflects the saturation degree near the critical point W^* when W_k converges to a neighborhood of W^*. If W^* is far from the symmetric point $W^\diamond$, $S_K^{\texttt{ASGD}}$ becomes large, leading to a large $E_K^{\texttt{ASGD}}$ and a large asymptotic error. In contrast, if W^* remains close to the symmetric point $W^\diamond$, the asymptotic error is small.

4 Mitigating Implicit Penalty by Residual Learning

The asymptotic error in `Analog SGD` is a fundamental issue that arises from the mismatch between the symmetric point and the critical point. An idealistic remedy for the inexact convergence is carefully shifting the weights to ensure the stationary point is close to a symmetric point. However, determining the appropriate adjustment is always challenging, as the critical point is difficult to pinpoint before the actual training. Therefore, an ideal solution to address this issue is to jointly construct a sequence with a predictable stationary point and a proper shift of the symmetric point.

Residual learning. Our solution overlaps the algorithmic stationary point and physical symmetric point on the special point 0. Besides the main analog array, W_k, we maintain another array, P_k, whose stationary point should be 0. A natural choice is the *residual* of the weight, $P^*(W)$, defined by the P that minimizes the objective $f(W+\gamma P)$ with a non-zero γ. Additionally, the goal of the main array is to minimize the residual so that the model W_k is approaching optimal. This process can be formulated as the following bilevel problem whose optimal points can be proved to be that of $f(W)$

$$\text{Residual Learning} \quad \underset{W\in\mathbb{R}^D}{\arg\min}\ \|P^*(W)\|^2, \quad \text{s.t. } P^*(W) \in \underset{P\in\mathbb{R}^D}{\arg\min}\ f(W+\gamma P). \tag{9}$$

Now we propose a gradient-based method to solve (9). The gradient of $f(W+\gamma P)$ with respect to P, given by $\nabla_P f(W+\gamma P) = \gamma\nabla f(W+\gamma P)$, is accessible with fair expense, enabling us to introduce a sequence P_k to track the residual of W_k by optimizing $f(W_k+\gamma P)$

$$P_{k+1} = P_k - \alpha\nabla f(\bar{W}_k;\xi_k)\odot F(P_k) - \alpha|\nabla f(\bar{W}_k;\xi_k)|\odot G(P_k). \tag{10}$$

where $\bar{W}_k := W_k + \gamma P_k$ is the mixed weight. We then derive the hyper-gradient of the upper-level objective. Notice $\nabla\|P^*(W)\|^2 = 2\nabla P^*(W)P^*(W)$. Assuming W^* is the unique minimum of $f(\cdot)$, we know $P^*(W)$ satisfies $P^*(W)+W = W^*$. Taking gradient with respective to W on both sides, we have $\nabla P^*(W) = -I$ and hence $\nabla\|P^*(W)\|^2 = -2P^*(W)$. Approximating $P^*(W)$ by P_k, we reach the update of the main array.

$$W_{k+1} = W_k + \beta P_{k+1}\odot F(W_k) - \beta|P_{k+1}|\odot G(W_k). \tag{11}$$

Featuring moving the residual P_k to W_k, (11) is referred to as *transfer* process. The updates (10) and (11) are performed alternatively until convergence. It is noticeable that `Residual Learning` has the same recursion format as `Tiki-Taka`. However, as we will discuss in Section 5, the proposed method can be naturally extended, and simulations in Section 6 show that the extension of `Residual Learning` outperforms the extension of `Tiki-Taka`.

On the response functions side, it is naturally required to let zero be a symmetric point, i.e., $G(0) = 0$, which can be implemented by the zero-shifting technique [39] by subtracting a reference array.

Convergence properties of `Residual Learning`. We begin by analyzing the convergence of `Residual Learning`without considering the zero-shift first, which enables us to understand how zero-shifted response functions affect convergence.

If the optimal point W^* exists and is unique, the solution of the lower-level objective has a closed form $P^*(W) := \frac{W^*-W}{\gamma}$. At that time, the upper-level objective equals $\|W^*-W\|^2$. However, the solutions of $f(\cdot)$ are non-unique in general, especially for non-convex objectives with multiple local minima. To ensure the existence and uniqueness of W^*, we assume the objective is strongly convex.

Assumption 4 (μ-strong convexity)**.** *The objective $f(W)$ is μ-strongly convex.*

Under the strongly convex assumption, the optimal point W^* is unique. Although the requirement of strong convexity is non-essential by the development of bilevel optimization [40–43], and the proof can be extended to more general cases, we leave it for future work.

Involving two sequences W_k and P_k, `Residual Learning` converges in different senses, including: (a) the residual array P_k converges to the optimal point $P^*(W_k)$; (b) W_k converges to the critical point of $f(\cdot)$ or the optimal point W^*; (c) the sum $\bar{W}_k = W_k + \gamma P_k$ converges to a critical point where $\nabla f(\bar{W}_k) = 0$. Taking all these into account, we define the convergence metric as

$$E_K^{\texttt{RL}} := \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\Big[\|\nabla f(\bar{W}_k)\|^2 + O(\|P_k - P^*(W_k)\|^2) + O(\|W_k - W^*\|^2)\Big]. \tag{12}$$

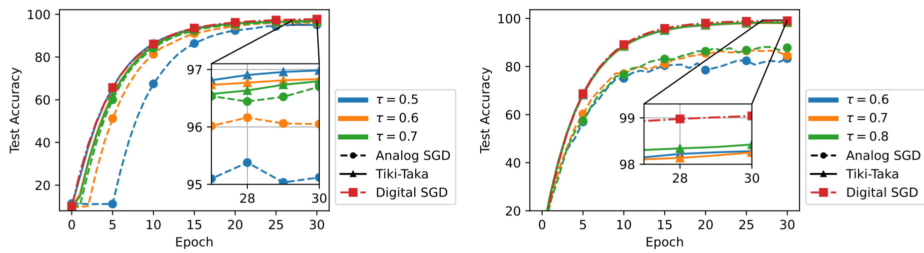

Figure 4: Test accuracy of training on MNIST dataset under different τ; **(Left)** FCN. **(Right)** CNN.

For simplicity, the constants in front of some terms in $E_K^{\texttt{RL}}$ are hidden. Now, we provide the convergence of `Residual Learning` with generic responses.

Theorem 3 (Convergence of `Residual Learning`)**.** *Under Assumptions 1–3 and 4, with the learning rate* $\alpha = O\left(\sqrt{1/\sigma^2 K}\right)$, $\beta = O(\alpha\gamma^{3/2})$, *it holds for* `Residual Learning` *that*

$$E_K^{\texttt{RL}} \leq O\left(\sqrt{\sigma^2/K} + \sigma^2 S_K^{\texttt{RL}}\right) \tag{13}$$

where $S_K^{\texttt{RL}}$ *denotes the amplification factor of* P_k *given by* $S_K^{\texttt{RL}} := \frac{1}{K}\sum_{k=0}^{K}\left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|_\infty^2$.

The proof of Theorem 3 is deferred to Appendix I. Theorem 3 claims that `Residual Learning` converges at the rate $O\left(\sqrt{\sigma^2/K}\right)$ to a neighbor of critical point with radius $O(\sigma^2 S_K^{\texttt{RL}})$, which share almost the same expression with the convergence of `Analog SGD`. The difference lies in the amplification factor $S_K^{\texttt{RL}}$ and $S_K^{\texttt{ASGD}}$, where the former depends on P_k while the latter depends on W_k.

Impact of response functions. Response function affects the `Analog SGD` and `Residual Learning` similarly. However, attributed to the residual array, constructing response functions to enable exact convergence of `Residual Learning` is viable.

As we have discussed, P_k tends to $P^*(W_k)$ which tends to 0 given W_k tends to W^*. Therefore, response functions with $G(P) = 0$ when $P = 0$ are required for the exact convergence.

Assumption 5. ***(Zero-shifted symmetric point)*** *$P = 0$ is a symmetric point, i.e.* $G(0) = 0$.

Under it and the Lipschitz continuity of the response functions, it holds directly that $\left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|_\infty \leq L_S\|P_k\|_\infty$ for a constant $L_S \geq 0$. Consequently, when $P_k \to P^*(W_k) \to 0$ as $W_k \to W^*$, the asymptotic error disappears. Formally, the following corollary holds true.

Corollary 1 (Exact convergence of `Residual Learning`)**.** *Under Assumption 5 and the conditions in Theorem 3, if* $\gamma \geq \Omega(R_{\min}^{-1/5})$, *it holds that* $E_K^{\texttt{RL}} \leq O\left(\sqrt{\sigma^2 L/K}\right)$.

The proof of Corollary 1 is deferred to Appendix I.5. Corollary 1 demonstrates the failure of `Tiki-Taka` in Figure 2. The symmetric point is $w^\diamond = c_{\texttt{Lin}}\tau$ in this example, which violates Assumption 5 when $c_{\texttt{Lin}} \neq 0$ and hence introduces asymptotic error into `Residual Learning`.

5 Extension of Residual Learning: limited granularity and noisy IO

This section extends `Residual Learning` to more practical scenarios involving more hardware imperfections. To be specific, we consider the *noisy IO* and *limited granularity* as examples. We highlight that we are not trying to diminish the importance of imperfection, but we focus on two of the primary ones that are known to be crucial.

Noisy IO introduces noise during the reading of P_{k+1} in the transfer process (11), given by $W_{k+1} = W_k + \beta(P_{k+1} + \varepsilon_k) \odot F(W_k) - \beta|P_{k+1} + \varepsilon_k| \odot G(W_k)$ with a noise ε_k. It incurs the implicit penalty issues again, leading to a penalized upper-level objective $\|P^*(W)\|^2 + \Sigma_\varepsilon\|W\|^2/2$, as claimed by Theorem 1, where Σ_ε is the standard variation of ε_k. To filter out the noise, we propose to use a digital buffer H_k to take a moving average of noisy P_{k+1} signals by

$$H_{k+1} = (1-\beta)H_k + \beta(P_{k+1} + \varepsilon_{k+1}) \tag{14}$$

Intuitively, with a fixed P_{k+1}, H_k will converge to a neighborhood of P_{k+1} with radius $O(\beta)$. Therefore, a sufficiently small β renders H_k a fair approximation of noiseless P_k, enabling optimizing the upper-level objective with clearer signals. After that, H_{k+1} is transferred to W_k as follows

$$W_{k+1} = W_k + \beta H_{k+1} \odot F(W_k) - \beta |H_{k+1}| \odot G(W_k). \tag{15}$$

Furthermore, the optimization process suffers from a constant error in the order of $O(\Delta w_{\min})$ as the pulses are fired discretely, with each changing the weight by $O(\Delta w_{\min})$. To overcome these issues, we propose introducing a threshold mechanism to train in limited precision. Instead of transferring H_{k+1} to W_k at each iteration, the transfer is triggered only if an element in H_{k+1} is greater than the response granularity $\Delta w_{\min}$. The proposed algorithms are referred to as `Residual Learning` v2.

	CIFAR10				
	`DSGD`	`ASGD`	`TT/RL`	`TTv2`	`RLv2`
ResNet18	95.43 $\pm$0.13	84.47 $\pm$3.40	94.81 $\pm$0.09	95.31 $\pm$0.05	95.12 $\pm$0.14
ResNet34	96.48 $\pm$0.02	95.43 $\pm$0.12	96.29 $\pm$0.12	96.60 $\pm$0.05	96.42 $\pm$0.13
ResNet50	96.57 $\pm$0.10	94.36 $\pm$1.16	96.34 $\pm$0.04	96.63 $\pm$0.09	96.56 $\pm$0.08
	CIFAR100				
	`DSGD`	`ASGD`	`TT/RL`	`TTv2`	`RLv2`
ResNet18	81.12 $\pm$0.25	68.98 $\pm$1.01	76.17 $\pm$0.23	78.56 $\pm$0.29	79.83 $\pm$0.13
ResNet34	83.86 $\pm$0.12	78.98 $\pm$0.55	80.58 $\pm$0.11	81.81 $\pm$0.15	82.85 $\pm$0.19
ResNet50	83.98 $\pm$0.11	79.88 $\pm$1.26	80.80 $\pm$0.22	82.82 $\pm$0.33	83.90 $\pm$0.20

Table 1: Fine-tuning ResNet models with the *power response* on CIFAR10/100. Test accuracy is reported. `DSGD`, `ASGD`, `TT/RL`, `TTv2`, and `RLv2` represent `Digital SGD`, `Analog SGD`, `Residual Learning/Tiki-Taka`, and `Residual Learning` v2, respectively.

6 Numerical Simulations

In this section, we verify the main theoretical results by simulations on both synthetic datasets and real datasets. We use the open source toolkit IBM Analog Hardware Acceleration Kit (AIHWKIT) [44] to simulate the behaviors of `Analog SGD`, `Residual Learning` (which reduces to `Tiki-Taka`). Each simulation is repeated three times, and the mean and standard deviation are reported. We consider two types of response functions in our simulations: power and exponential response functions with dynamic ranges $[-\tau, \tau]$, while the symmetric point is 0, as required by Corollary 1. More details, simulations, and ablation studies can be found in Appendix K.

MNIST FCN/CNN. We train a fully-connected network (FCN) and a convolutional neural network (CNN) on the MNIST dataset and compare the performance of `Analog SGD` and `Tiki-Taka` under various τ on power responses; see the results in Figure 4. By tracking residual, `Tiki-Taka` outperforms `Analog SGD` and reaches comparable accuracy with `Digital SGD`. For both architectures, the accuracy of `Tiki-Taka` drops by $< 1\%$. In contrast, `Analog SGD` takes a few epochs to achieve an observable accuracy increment in FCN training, rendering a slower convergence rate than `Tiki-Taka`. In CNN training, `Analog SGD`'s accuracy increases more slowly than `Tiki-Taka` does and finally gets stuck at about 80%. It is consistent with the theoretical claims.

CIFAR10/CIFAR100 ResNet. We fine-tune three ResNet models with different scales on CIFAR10/CIFAR100 datasets. The power response functions are used, whose results are shown in Table 1. The results show that the `Tiki-Taka` outperforms `Analog SGD` by about 1.0% in most of the cases in ResNet34/50, and the gap even reaches about 7.0% for ResNet18 training on the CIFAR100 dataset. On top of that, we also compare the proposed `Residual Learning v2` and `Tiki-Taka v2`. Both of them outperform `Residual Learning` since they introduce a digital buffer to filter out the reading noise. However, `Residual Learning v2` outperforms `Tiki-Taka v2` on the CIFAR100 dataset, demonstrating the benefit from the bilevel formulation.

7 Conclusions and Limitations

This paper studies the impact of a generic class of asymmetric and non-linear response functions on gradient-based training in analog in-memory computing hardware. We first formulate the dynamic of `Analog Update` based on the pulse update rule. Based on it, we show that `Analog SGD` implicitly optimizes a penalized objective and hence can only converge inexactly. To overcome this issue, we propose a `Residual Learning` framework which solves a bilevel optimization problem. Explicitly aligning the algorithmic stationary point and physical symmetric point, `Residual Learning` provably converges to the optimal point exactly. Furthermore, we demonstrate how to extend `Residual Learning` to overcome the noisy reading and limited update granularity issues. The efficiency of the proposed method is verified through simulations. One limitation of this work is that the current analysis considers only the three hardware imperfections. While they are known to be crucial for analog training, it is also important to extend our convergence analysis and methods to more practical scenarios involving more imperfections in future work.

References

[1] Norm Jouppi, George Kurian, Sheng Li, Peter Ma, Rahul Nagarajan, Lifeng Nai, Nishant Patil, Suvinay Subramanian, Andy Swing, Brian Towles, et al. TPU v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings. In *Annual International Symposium on Computer Architecture*, pages 1–14, 2023.

[2] Hadi Esmaeilzadeh, Adrian Sampson, Luis Ceze, and Doug Burger. Neural acceleration for general-purpose approximate programs. In *IEEE/ACM international symposium on microarchitecture*, pages 449–460. IEEE, 2012.

[3] Dharmendra S Modha, Filipp Akopyan, Alexander Andreopoulos, Rathinakumar Appuswamy, John V Arthur, Andrew S Cassidy, Pallab Datta, Michael V DeBole, Steven K Esser, Carlos Ortega Otero, et al. Neural inference at the frontier of energy, space, and time. *Science*, 382(6668):329–335, 2023.

[4] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] An Chen. A comprehensive crossbar array model with solutions for line resistance and nonlinear device characteristics. *IEEE Transactions on Electron Devices*, 60(4):1318–1326, 2013.

[7] Wilfried Haensch, Tayfun Gokmen, and Ruchir Puri. The next generation of deep learning hardware: Analog computing. *Proceedings of the IEEE*, 107(1):108–122, 2019.

[8] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S Emer. Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, 105(12):2295–2329, 2017.

[9] Abu Sebastian, Manuel Le Gallo, Riduan Khaddam-Aljameh, and Evangelos Eleftheriou. Memory devices and applications for in-memory computing. *Nature Nanotechnology*, 15:529–544, 2020.

[10] Manuel Le Gallo, Riduan Khaddam-Aljameh, Milos Stanisavljevic, Athanasios Vasilopoulos, Benedikt Kersting, Martino Dazzi, Geethan Karunaratne, Matthias Brändli, Abhairaj Singh, Silvia M Mueller, et al. A 64-core mixed-signal in-memory compute chip based on phase-change memory for deep neural network inference. *Nature Electronics*, 6(9):680–693, 2023.

[11] Geoffrey W Burr, Robert M Shelby, Abu Sebastian, Sangbum Kim, Seyoung Kim, Severin Sidler, Kumar Virwani, Masatoshi Ishii, Pritish Narayanan, Alessandro Fumarola, et al. Neuromorphic computing using non-volatile memory. *Advances in Physics: X*, 2(1):89–124, 2017.

[12] J Joshua Yang, Dmitri B Strukov, and Duncan R Stewart. Memristive devices for computing. *Nature nanotechnology*, 8(1):13, 2013.

[13] Shubham Jain et al. Neural network accelerator design with resistive crossbars: Opportunities and challenges. *IBM Journal of Research and Development*, 63(6):10–1, 2019.

[14] Stefan Cosemans, Bram-Ernst Verhoef, Jonas Doevenspeck, Ioannis A. Papistas, Francky Catthoor, Peter Debacker, Arindam Mallik, and Diederik Verkest. Towards 10000TOPS/W DNN inference with analog in-memory computing – a circuit blueprint, device options and requirements. In *IEEE International Electron Devices Meeting*, pages 22.2.1–22.2.4, 2019.

[15] Ioannis A Papistas, Stefan Cosemans, Bram Rooseleer, Jonas Doevenspeck, M-H Na, Arindam Mallik, Peter Debacker, and Diederik Verkest. A 22 nm, 1540 TOP/s/W, 12.1 TOP/s/mm 2 in-memory analog matrix-vector-multiplier for DNN acceleration. In *IEEE Custom Integrated Circuits Conference*, pages 1–2. IEEE, 2021.

[16] Tayfun Gokmen and Yurii Vlasov. Acceleration of deep neural network training with resistive cross-point devices: Design considerations. *Frontiers in neuroscience*, 10:333, 2016.

[17] Geoffrey W Burr, Robert M Shelby, Severin Sidler, Carmelo Di Nolfo, Junwoo Jang, Irem Boybat, Rohit S Shenoy, Pritish Narayanan, Kumar Virwani, Emanuele U Giacometti, et al. Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses) using phase-change memory as the synaptic weight element. *IEEE Transactions on Electron Devices*, 62(11):3498–3507, 2015.

[18] Sapan Agarwal, Steven J Plimpton, David R Hughart, Alexander H Hsia, Isaac Richter, Jonathan A Cox, Conrad D James, and Matthew J Marinella. Resistive memory device requirements for a neural algorithm accelerator. In *International Joint Conference on Neural Networks*, pages 929–938. IEEE, 2016.

[19] Paiyu Chen, Binbin Lin, I-Ting Wang, Tuohung Hou, Jieping Ye, Sarma Vrudhula, Jae-sun Seo, Yu Cao, and Shimeng Yu. Mitigating effects of non-ideal synaptic device characteristics for on-chip learning. In *IEEE/ACM International Conference on Computer-Aided Design*, pages 194–199. IEEE, 2015.

[20] Vinay Joshi, Manuel Le Gallo, Simon Haefeli, Irem Boybat, Sasidharan Rajalekshmi Nandakumar, Christophe Piveteau, Martino Dazzi, Bipin Rajendran, Abu Sebastian, and Evangelos Eleftheriou. Accurate deep neural network inference using computational phase-change memory. *Nature communications*, 11(1):2473, 2020.

[21] Zhaoxian Wu, Tayfun Gokmen, Malte J Rasch, and Tianyi Chen. Towards exact gradient-based training on analog in-memory computing. In *Advances in Neural Information Processing Systems*, 2024.

[22] Tayfun Gokmen and Wilfried Haensch. Algorithm for training neural networks on resistive device arrays. *Frontiers in Neuroscience*, 14, 2020.

[23] Tayfun Gokmen. Enabling training of neural networks on noisy hardware. *Frontiers in Artificial Intelligence*, 4:1–14, 2021.

[24] Malte J Rasch, Fabio Carta, Omobayode Fagbohungbe, and Tayfun Gokmen. Fast and robust analog in-memory deep neural network training. *Nature Communications*, 15(1):7133–7147, 2024.

[25] Peng Yao, Huaqiang Wu, Bin Gao, Sukru Burc Eryilmaz, Xueyao Huang, Wenqiang Zhang, Qingtian Zhang, Ning Deng, Luping Shi, H-S Philip Wong, et al. Face classification using electronic synapses. *Nature communications*, 8(1):15199, 2017.

[26] Zhongrui Wang, Can Li, Peng Lin, Mingyi Rao, Yongyang Nie, Wenhao Song, Qinru Qiu, Yunning Li, Peng Yan, John Paul Strachan, et al. In situ training of feed-forward and recurrent convolutional memristor networks. *Nature Machine Intelligence*, 1(9):434–442, 2019.

[27] Yaoyuan Wang, Shuang Wu, Lei Tian, and Luping Shi. SSM: a high-performance scheme for in situ training of imprecise memristor neural networks. *Neurocomputing*, 407:270–280, 2020.

[28] Shanshi Huang, Xiaoyu Sun, Xiaochen Peng, Hongwu Jiang, and Shimeng Yu. Overcoming challenges for achieving high in-situ training accuracy with emerging memories. In *Design, Automation & Test in Europe Conference & Exhibition*, pages 1025–1030. IEEE, 2020.

[29] Weier Wan, Rajkumar Kubendran, Clemens Schaefer, Sukru Burc Eryilmaz, Wenqiang Zhang, Dabin Wu, Stephen Deiss, Priyanka Raina, He Qian, Bin Gao, et al. A compute-in-memory chip based on resistive random-access memory. *Nature*, 608(7923):504–512, 2022.

[30] Peng Yao, Huaqiang Wu, Bin Gao, Jianshi Tang, Qingtian Zhang, Wenqiang Zhang, J Joshua Yang, and He Qian. Fully hardware-implemented memristor convolutional neural network. *Nature*, 577(7792):641–646, 2020.

[31] S. R. Nandakumar, Manuel Le Gallo, Irem Boybat, Bipin Rajendran, Abu Sebastian, and Evangelos Eleftheriou. Mixed-precision architecture based on computational memory for training deep neural networks. In *IEEE International Symposium on Circuits and Systems*, pages 1–5, 2018.

[32] S. R. Nandakumar, Manuel Le Gallo, Christophe Piveteau, Vinay Joshi, Giovanni Mariani, Irem Boybat, Geethan Karunaratne, Riduan Khaddam-Aljameh, Urs Egger, Anastasios Petropoulos, Theodore Antonakopoulos, Bipin Rajendran, Abu Sebastian, and Evangelos Eleftheriou. Mixed-precision deep learning based on computational memory. *Frontiers in Neuroscience*, 14, 2020.

[33] Benjamin Scellier and Yoshua Bengio. Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in computational neuroscience*, 11:24, 2017.

[34] Mohamed Watfa, Alberto Garcia-Ortiz, and Gilles Sassatelli. Energy-based analog neural network framework. *Frontiers in Computational Neuroscience*, 17:1114651, 2023.

[35] Benjamin Scellier, Maxence Ernoult, Jack Kendall, and Suhas Kumar. Energy-based learning algorithms for analog computing: a comparative study. *Advances in Neural Information Processing Systems*, 36, 2024.

[36] Jack Kendall, Ross Pantone, Kalpana Manickavasagam, Yoshua Bengio, and Benjamin Scellier. Training end-to-end analog neural networks with equilibrium propagation. *arXiv preprint arXiv:2006.01981*, 2020.

[37] Maxence Ernoult, Julie Grollier, Damien Querlioz, Yoshua Bengio, and Benjamin Scellier. Equilibrium propagation with continual weight updates. *arXiv preprint arXiv:2005.04168*, 2020.

[38] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

[39] Hyungjun Kim, Malte J Rasch, Tayfun Gokmen, Takashi Ando, Hiroyuki Miyazoe, Jae-Joon Kim, John Rozen, and Seyoung Kim. Zero-shifting technique for deep neural network training on resistive cross-point arrays. *arXiv preprint arXiv:1907.10228*, 2019.

[40] Quan Xiao, Songtao Lu, and Tianyi Chen. A generalized alternating method for bilevel learning under the polyak-łojasiewicz condition. In *Proc. Advances in Neural Info. Process. Syst.*, 2023.

[41] Michael Arbel and Julien Mairal. Non-convex bilevel games with critical point selection maps. In *Advances in Neural Information Processing Systems*, 2022.

[42] Han Shen, Quan Xiao, and Tianyi Chen. On penalty-based bilevel gradient descent method. In *Proc. of International Conference on Machine Learning*, 2023.

[43] Jeongyeol Kwon, Dohyun Kwon, Steve Wright, and Robert Nowak. On penalty methods for non-convex bilevel optimization and first-order stochastic approximation. In *Proc. of International Conference on Learning Representations*, 2024.

[44] Malte J Rasch, Diego Moreda, Tayfun Gokmen, Manuel Le Gallo, Fabio Carta, Cindy Goldberg, Kaoutar El Maghraoui, Abu Sebastian, and Vijay Narayanan. A flexible and fast PyTorch toolkit for simulating training and inference on analog crossbar arrays. *IEEE International Conference on Artificial Intelligence Circuits and Systems*, pages 1–4, 2021.

[45] Geoffrey W Burr, Matthew J BrightSky, Abu Sebastian, Huai-Yu Cheng, Jau-Yi Wu, Sangbum Kim, Norma E Sosa, Nikolaos Papandreou, Hsiang-Lan Lung, Haralampos Pozidis, Evangelos Eleftheriou, and Chung H Lam. Recent Progress in Phase-Change Memory Technology. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 6(2):146–162, 2016.

[46] Manuel Le Gallo and Abu Sebastian. An overview of phase-change memory device physics. *Journal of Physics D: Applied Physics*, 53(21):213002, 2020.

[47] Jun-Woo Jang, Sangsu Park, Yoon-Ha Jeong, and Hyunsang Hwang. ReRAM-based synaptic device for neuromorphic computing. In *IEEE International Symposium on Circuits and Systems*, pages 1054–1057, 2014.

[48] Jun-Woo Jang, Sangsu Park, Geoffrey W Burr, Hyunsang Hwang, and Yoon-Ha Jeong. Optimization of conductance change in $Pr_{1-x}Ca_xMnO_3$-based synaptic devices for neuromorphic systems. *IEEE Electron Device Letters*, 36(5):457–459, 2015.

[49] Tommaso Stecconi, Valeria Bragaglia, Malte J Rasch, Fabio Carta, Folkert Horst, Donato F Falcone, Sofieke C Ten Kate, Nanbo Gong, Takashi Ando, Antonis Olziersky, et al. Analog resistive switching devices for training deep neural networks with the novel Tiki-Taka algorithm. *Nano Letters*, 24(3):866–872, 2024.

[50] Seokjae Lim, Myounghoon Kwak, and Hyunsang Hwang. Improved synaptic behavior of CBRAM using internal voltage divider for neuromorphic systems. *IEEE Transactions on Electron Devices*, 65(9):3976–3981, 2018.

[51] Elliot J Fuller, Scott T Keene, Armantas Melianas, Zhongrui Wang, Sapan Agarwal, Yiyang Li, Yaakov Tuchman, Conrad D James, Matthew J Marinella, J Joshua Yang, Alberto Salleo, and A Alec Talin. Parallel programming of an ionic floating-gate memory array for scalable neuromorphic computing. *Science*, 364(6440):570–574, 2019.

[52] Jianshi Tang, Douglas Bishop, Seyoung Kim, Matt Copel, Tayfun Gokmen, Teodor Todorov, SangHoon Shin, Ko-Tao Lee, Paul Solomon, Kevin Chan, et al. ECRAM as scalable synaptic cell for high-speed, low-power neuromorphic computing. In *IEEE International Electron Devices Meeting*, pages 13–1. IEEE, 2018.

[53] Murat Onen, Nicolas Emond, Baoming Wang, Difei Zhang, Frances M Ross, Ju Li, Bilge Yildiz, and Jesús A Del Alamo. Nanosecond protonic programmable resistors for analog deep learning. *Science*, 377(6605):539–543, 2022.

[54] Seungchul Jung, Hyungwoo Lee, Sungmeen Myung, Hyunsoo Kim, Seung Keun Yoon, Soon-Wan Kwon, Yongmin Ju, Minje Kim, Wooseok Yi, Shinhee Han, et al. A crossbar array of magnetoresistive memory devices for in-memory computing. *Nature*, 601(7892):211–216, 2022.

[55] Zhihua Xiao, Vinayak Bharat Naik, Jia Hao Lim, Yaoru Hou, Zhongrui Wang, and Qiming Shao. Adapting magnetoresistive memory devices for accurate and on-chip-training-free in-memory computing. *Science Advances*, 10(38):eadp3710, 2024.

[56] Rui Guo, Weinan Lin, Xiaobing Yan, T Venkatesan, and Jingsheng Chen. Ferroic tunnel junctions and their application in neuromorphic networks. *Applied physics reviews*, 7(1), 2020.

[57] Panni Wang, Feng Xu, Bo Wang, Bin Gao, Huaqiang Wu, He Qian, and Shimeng Yu. Three-dimensional NAND flash for vector-matrix multiplication. *IEEE Transactions on Very Large Scale Integration Systems*, 27(4):988–991, 2018.

[58] Yachen Xiang, Peng Huang, Runze Han, Chu Li, Kunliang Wang, Xiaoyan Liu, and Jinfeng Kang. Efficient and robust spike-driven deep convolutional neural networks based on NOR flash computing array. *IEEE Transactions on Electron Devices*, 67(6):2329–2335, 2020.

[59] Farnood Merrikh-Bayat, Xinjie Guo, Michael Klachko, Mirko Prezioso, Konstantin K Likharev, and Dmitri B Strukov. High-performance mixed-signal neurocomputing with nanoscale floating-gate memory cell arrays. *IEEE Transactions on Nneural Networks and Learning Systems*, 29(10):4782–4790, 2017.

[60] Bonan Zhang, Peter Deaville, and Naveen Verma. Statistical computing framework and demonstration for in-memory computing systems. In *ACM/IEEE Design Automation Conference*, pages 979–984, 2022.

[61] Peter Deaville, Bonan Zhang, Lung-Yen Chen, and Naveen Verma. A maximally row-parallel MRAM in-memory-computing macro addressing readout circuit sensitivity and area. In *IEEE European Solid State Circuits Conference*, pages 75–78. IEEE, 2021.

[62] Jung-Hoon Lee, Dong-Hyeok Lim, Hongsik Jeong, Huimin Ma, and Luping Shi. Exploring cycle-to-cycle and device-to-device variation tolerance in mlc storage-based neural network training. *IEEE Transactions on Electron Devices*, 66(5):2172–2178, 2019.

[63] Jintao Zhang, Zhuo Wang, and Naveen Verma. In-memory computation of a machine-learning classifier in a standard 6t SRAM array. *IEEE Journal of Solid-State Circuits*, 52(4):915–924, 2017.

[64] Tayfun Gokmen, Malte J Rasch, and Wilfried Haensch. The marriage of training and inference for scaled deep learning analog hardware. In *IEEE International Electron Devices Meeting*, pages 22–3. IEEE, 2019.

[65] Corey Lammie, Athanasios Vasilopoulos, Julian Büchel, Giacomo Camposampiero, Manuel Le Gallo, Malte Rasch, and Abu Sebastian. Improving the accuracy of analog-based in-memory computing accelerators post-training. In *IEEE International Symposium on Circuits and Systems*, pages 1–5. IEEE, 2024.

[66] Qing Jin, Zhiyu Chen, Jian Ren, Yanyu Li, Yanzhi Wang, and Kaiyuan Yang. PIM-QAT: Neural network quantization for processing-in-memory (PIM) systems. *arXiv preprint arXiv:2209.08617*, 2022.

[67] Malte J Rasch, Charles Mackin, Manuel Le Gallo, An Chen, Andrea Fasoli, Frédéric Odermatt, Ning Li, S. R. Nandakumar, Pritish Narayanan, Hsinyu Tsai, et al. Hardware-aware training for large-scale and diverse deep learning inference workloads using in-memory computing-based accelerators. *Nature Communications*, 14(1):5282, 2023.

[68] Bonan Zhang, Chia-Yu Chen, and Naveen Verma. Reshape and adapt for output quantization (RAOQ): Quantization-aware training for in-memory computing systems. In *International Conference on Machine Learning*, 2024.

[69] Beiye Liu, Hai Li, Yiran Chen, Xin Li, Qing Wu, and Tingwen Huang. Vortex: Variation-aware training for memristor x-bar. In *Proceedings of the 52nd Annual Design Automation Conference*, pages 1–6, 2015.

[70] Abhiroop Bhattacharjee, Lakshya Bhatnagar, Youngeun Kim, and Priyadarshini Panda. NEAT: Non-linearity aware training for accurate and energy-efficient implementation of neural networks on 1t-1r memristive crossbars. *arXiv preprint arXiv:2012.00261*, 2020.

[71] Tayfun Gokmen, Murat Onen, and Wilfried Haensch. Training deep convolutional neural networks with resistive cross-point devices. *Frontiers in neuroscience*, 11:538, 2017.

[72] Zhongrui Wang, Saumil Joshi, Sergey Savel'Ev, Wenhao Song, Rivu Midya, Yunning Li, Mingyi Rao, Peng Yan, Shiva Asapu, Ye Zhuo, et al. Fully memristive neural networks for pattern classification with unsupervised learning. *Nature Electronics*, 1(2):137–145, 2018.

[73] Nanbo Gong, Malte Rasch, Soon-Cheon Seo, Arthur Gasasira, Paul Solomon, Valeria Bragaglia, Steven Consiglio, Hisashi Higuchi, Chanro Park, Kevin Brew, et al. Deep learning acceleration in 14nm CMOS compatible ReRAM array: device, material and algorithm co-optimization. In *IEEE International Electron Devices Meeting*, 2022.

[74] Zhaoxian Wu, Quan Xiao, Tayfun Gokmen, Hsinyu Tsai, Kaoutar El Maghraoui, and Tianyi Chen. Pipeline gradient-based model training on analog in-memory accelerators. *arXiv preprint arXiv:2410.15155*, 2024.

[75] Logan G Wright, Tatsuhiro Onodera, Martin M Stein, Tianyu Wang, Darren T Schachter, Zoey Hu, and Peter L McMahon. Deep physical neural networks trained with backpropagation. *Nature*, 601(7894):549–555, 2022.

[76] Ali Momeni, Babak Rahmani, Benjamin Scellier, Logan G Wright, Clara C McMahon, Peter L andWanjura, Yuhang Li, Anas Skalli, Natalia G. Berloff, Tatsuhiro Onodera, Ilker Oguz, Francesco Morichetti, Philipp del Hougne, Manuel Le Gallo, Abu Sebastian, Azalia Mirhoseini, Cheng Zhang, Danijela Marković, Daniel Brunner, Christophe Moser, Sylvain Gigan, Florian Marquardt, Aydogan Ozcan, Julie Grollier, Andrea J Liu, Demetri Psaltis, Andrea Alù, and Romain Fleury. Training of physical neural networks. *arXiv preprint arXiv:2406.03372*, 2024.

[77] Demetri Psaltis, David Brady, Xiang-Guang Gu, and Steven Lin. Holography in artificial neural networks. *Nature*, 343(6256):325–330, 1990.

[78] Tyler W Hughes, Ian AD Williamson, Momchil Minkov, and Shanhui Fan. Wave physics as an analog recurrent neural network. *Science advances*, 5(12):eaay6946, 2019.

[79] Alexander N Tait, Thomas Ferreira De Lima, Ellen Zhou, Allie X Wu, Mitchell A Nahmias, Bhavin J Shastri, and Paul R Prucnal. Neuromorphic photonic networks using silicon photonic weight banks. *Scientific reports*, 7(1):7430, 2017.

[80] Nanbo Gong, T Idé, S Kim, Irem Boybat, Abu Sebastian, V Narayanan, and Takashi Ando. Signal and noise extraction from analog memory elements for neuromorphic computing. *Nature communications*, 9(1):2102, 2018.

[81] Mingyi Rao, Hao Tang, Jiangbin Wu, Wenhao Song, Max Zhang, Wenbo Yin, Ye Zhuo, Fatemeh Kiani, Benjamin Chen, Xiangqi Jiang, et al. Thousands of conductance levels in memristors integrated on CMOS. *Nature*, 615(7954):823–829, 2023.

[82] Deepak Sharma, Santi Prasad Rath, Bidyabhusan Kundu, Anil Korkmaz, Damien Thompson, Navakanta Bhat, Sreebrata Goswami, R Stanley Williams, and Sreetosh Goswami. Linear symmetric self-selecting 14-bit kinetic molecular memristors. *Nature*, 633(8030):560–566, 2024.

[83] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2013.

Appendix for “Analog In-memory Training on Non-ideal Resistive Elements: Understanding the Impact of Response Functions”

Table of Contents

A Literature Review

This section briefly reviews literature that is related to this paper, as complementary to Section 1.

Training on AIMC hardware. Analog training has shown promising early successes in certain tasks like face classification [25] and digit classification [26], where $1,000\times$ lower energy consumption compared with digital implementation is presented. Researchers are also exploring approaches to mitigate the impact of hardware non-idealities. For example, [27, 28] proposes to leverage the momentum technique to stabilize the training by reducing the noise. In order to overcome other potential non-idealities, a hybrid training paradigm is also being explored. [29] leverages the chip-in-the-loop technique to train models layer-by-layer, while [30] proposes to train the backbone in the digital domain and train the last layer in the analog domain. In general, these works have provided valuable insights into analog training, shedding light on many critical technical challenges. However, their focus has largely been on experimental and simulation aspects, with limited systematic and theoretical analysis of how specific imperfections affect the training process. In our paper, we present an alternative viewpoint and novel tools to explore the effects of non-idealities.

Resistive element. A series of works seeks various resistive elements that have near-constant or at least symmetric responses. The leading candidates currently include PCM [45, 46], ReRAM [47–49], CBRAM [50, 51], ECRAM [52, 53], MRAM [54, 55], FTJ [56] or flash memory [57–59].

However, the resistive element possessing symmetric updates may not be the best option in terms of manufacturing. For example, although ECRAM provides almost symmetric updates, it is still less competitive than ReRAM as ReRAM has a faster response speed and lower pulse voltage [49]. The suitability of the resistive elements is evaluated based on metrics across multiple dimensions, such as the number of conductance states, retention, material endurance, switching energy, response speed, manufacturing cost, and cell size. Among them, this paper is only interested in the impact of response functions in the training.

Imperfection of AIMC hardware. Besides the response functions, analog training suffers from all kinds of hardware imperfection, especially when the task's scale increases, like asymmetric update [17, 19], reading/writing noise [18, 60, 61], device/cycle variations [62], non-linear current response due to IR drop [18, 6, 63]. This paper mainly focuses on asymmetric response functions. However, this paper is not trying to diminish the importance of other hardware imperfections but focuses on one of the primary ones that are known to be very important [19, 16].

Hardware-aware training. For inference on AIMC hardware purposes, models pretrained on digital hardware will be programmed on analog hardware. Due to the hardware imperfection, the pretrained models suffer from performance drops. Hardware-aware training (HWA) is a technique designed to bridge the gap between ideal pretrained models and non-ideal programmed models. In contrast to standard training methods, hardware-aware training explicitly incorporates device-specific imperfections, such as weight drift [20], device fail [64], bounded dynamic range [65], quantization error from ADC [66–68], device variation [69], and non-linear current output [70], into the training loop. By modeling these constraints during the training phase, the learned parameters become inherently more robust to real-world deployment conditions. It is worth highlighting that HWA is still performed on digital hardware, and the trained model will be programmed onto AIMC hardware. On the contrary, this paper considers a different and more challenging setting where the training is performed directly on analog hardware.

Gradient-based training on AIMC hardware. A series of works focuses on implementing back-propagation (BP) and gradient-based training on AIMC hardware. The seminal work [16, 71] leverages the rank-one structure of the gradient and implements `Analog SGD` by a stochastic pulse update scheme, *rank-update*. Rank-update significantly accelerates the gradient descent step by avoiding dealing the gradients with $O(N^2)$ elements directly but using two $O(N)$ vectors for update instead, where N is the numbers of matrix row and column. To alleviate the *asymmetric update issue*, researchers also design various of `Analog SGD` variants, `Tiki-Taka` algorithm family [22–24]. The key components of `Tiki-Taka` are introducing a *residual array* to stabilize the training. Apart from the rank-update, a hybrid scheme that performs forward and backward in the analog domain but computes the gradients in the digital domain has been proposed in [31, 32]. Their solution, referred to as *mixed-precision update*, provides a more accurate gradient signal but requires 5×-10× higher overhead compared to the rank-update scheme [24].

Attributed to these efforts, analog training has empirically shown great promise in achieving a similar level of accuracy as digital training on chip prototype, with reduced energy consumption and training time [72, 73]. Simultaneously, the parallel acceleration solution with AIMC hardware is also under exploration [74]. Despite its good performance, it is still mysterious about when and why they work.

Theoretical foundation of gradient-based training. The closely related result comes from the convergence study of `Tiki-Taka` [21]. Similar to our work, they attempt to model the dynamic and provide the convergence properties of `Analog SGD` and `Tiki-Taka`. However, their work is limited to a special linear response function. Furthermore, their paper considers a simplified version of `Tiki-Taka`, with a hyper-parameter $\gamma = 0$ (see Section 4). As we will show empirically and theoretically, `Tiki-Taka` benefits from a non-zero γ. Consequently, We compare the results briefly in Table 2 and comprehensively in Appendix B.

	γ	Generic response	Linear response
`Tiki-Taka` [21]	$= 0$	✗	$O\left(\sqrt{\frac{1}{K}\frac{1}{1-33P_{\max}^2/\tau^2}}\right)$
`Tiki-Taka` [Corollary 1]	$\neq 0$	$O\left(\sqrt{\frac{1}{K}\frac{1}{R_{\min}^{\texttt{RL}}}}\right)$	$O\left(\sqrt{\frac{1}{K}\frac{1}{1-P_{\max}^2/\tau^2}}\right)$

Table 2: Comparison between our paper and [21]. Mixing-coefficient γ is a hyper-parameter of `Tiki-Taka`. "Generic response" and "Linear response" columns are the convergence rates in the corresponding settings. K represents the number of iterations. $R_{\min}^{\texttt{RL}}$ and $P_{\max}^2/\tau^2 < 1$ measure the saturation while the former one reduces to the latter on linear response functions.

Energy-based models and equilibrium propagation. Apart from achieving explicit gradient signals by the BP, there are also attempts to train models based on *equilibrium propagation* (EP, [33]), which provides a biologically plausible alternative to traditional BP. EP is applicable to a series of energy-based models, where the forward pass is performed by minimizing an energy function [34, 35]. The update signal in EP is computed by measuring the output difference between a free phase and an active phase. EP eliminates the need for BP non-local weight transport mechanism, making it more compatible with neuromorphic and energy-efficient hardware [36, 37]. We highlight here that the approach to attain update signals (BP or EP) is orthogonal to the update mechanism (pulse update). Their difference lies in the objective $f(W_k)$, which is hidden in this paper. Therefore, building upon the pulse update, our work is applicable to both BP and EP.

Physical neural network. The model executing on AIMC hardware, which leverages resistive crossbar array to accelerate MVM operation, is a concrete implementation of physical neural networks (PNNs, [75, 76]). PNN is a generic concept of implementing neural networks via a physical system in which a set of tunable parameters, such as holographic grating [77], wave-based systems [78], and photonic networks [79]. Our work particularly focuses on training with AIMC hardware, but the methodology developed in this paper can be transferred to the study of other PNNs.

B Relation with the result in [21]

Similar to this paper, [21] also attempts to model the dynamic of analog training. They shows that `Analog SGD` converges to a critical point of problem (1) inexactly with an asymptotic error, and `Tiki-Taka` converges to a critical point exactly. In this section, we compare our results with our results and theirs.

As discussed in Section 1, [21] studies the analog training on special linear response functions

$$q_+(w) = 1 - \frac{w}{\tau}, \quad q_-(w) = 1 + \frac{w}{\tau}. \tag{16}$$

It can be checked that the symmetric point is 0 while the dynamic range of it is $[-\tau, \tau]$. The symmetric and asymmetric components is defined by $F(W) = 1$ and $G(W) = \frac{W}{\tau}$, respectively. It indicates $F_{\max} = 1$. Furthermore, they assume the bounded weight saturation by assuming bounded weights, i.e., $\|W_k\|_\infty \leq W_{\max}, \forall k \in [K]$ with a constant $W_{\max} < \tau$. Under this assumption, the lower bounds of response functions are given by

$$\min\{R(W_k)\} = \min\{Q_+(W_k) \odot Q_-(W_k)\} = 1 - \left(\frac{\|W_k\|_\infty}{\tau}\right)^2 \tag{17}$$

$$R_{\min}^{\texttt{ASGD}} = \min\{R(W_k)\} = 1 - \left(\frac{W_{\max}}{\tau}\right)^2. \tag{18}$$

Challenge of analyzing the convergence of `Tiki-Taka` with generic response functions. For linear response functions (16), the recursion of residual array P_k has a special structure, where the first and the biased term can be combined

$$\begin{aligned} P_{k+1} &= P_k - \alpha \nabla f(\bar{W}_k;\xi_k) - \frac{\alpha}{\tau} |\nabla f(\bar{W}_k;\xi_k)| \odot P_k \\ &= \Big(1 - \frac{\alpha}{\tau}|\nabla f(\bar{W}_k;\xi_k)|\Big) \odot P_k - \alpha \nabla f(\bar{W}_k;\xi_k) \end{aligned} \tag{19}$$

which is a weighted average of P_k and $\nabla f(\bar{W}_k;\xi_k)$. Consequently, P_k can be interpreted as an approximation of the average gradient. For this perspective, the transfer operation can be interpreted as biased gradient descent. However, given a generic $G(\cdot)$, the combination is no longer viable, bringing difficulties to the analysis.

Convergence of `Analog SGD`. As we will show in Remark 1 at the end of Appendix H, inequality (8) can be improved when the saturation never happens

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[\|\nabla f(W_k)\|^2] \tag{20}$$

$$\leq \frac{4F_{\max}^2}{R_{\min}^{\texttt{RL}}} \sqrt{\frac{(f(W_0)-f^*)\sigma^2 L}{K}} + 2F_{\max}\sigma^2 \times \frac{1}{K}\sum_{k=0}^{K-1} \left\| \frac{G(W_k)}{\sqrt{F(W_k)}} \right\|_\infty^2 \Bigg/ \min\{R(W_k)\} \tag{21}$$

$$\leq O\left(\sqrt{\frac{(f(W_0)-f^*)\sigma^2 L}{K}} \frac{1}{1 - W_{\max}^2/\tau^2} \right) + 2\sigma^2 \times \frac{1}{K}\sum_{k=0}^{K} \frac{\|W_k\|_\infty^2/\tau^2}{1 - \|W_k\|_\infty^2/\tau^2}$$

which is exactly the result in [21].

Convergence of `Tiki-Taka`. It is shown that a non-zero γ in (10) improves the training accuracy [22]. However, [21] considers a special case with $\gamma = 0$ while this paper considers a non-zero γ. As we will discuss latter in this section, different γ leads to different convergence behaviors of `Tiki-Taka`.

With the linear response, if we also assume the bounded saturation of P_k by letting $\|P_k\|_\infty \leq P_{\max}$, the minimal average response function is given by $R_{\min}^{\texttt{RL}} = 1 - \left(\frac{P_{\max}}{\tau}\right)^2$. The upper bound in Corollary 1 becomes

$$\frac{1}{K}\sum_{k=0}^{K-1} \|\nabla f(\bar{W}_k)\|^2 \leq O\left(\frac{1}{1 - P_{\max}^2/\tau^2} \sqrt{\frac{(f(W_0)-f^*)\sigma^2 L}{K}} \right). \tag{22}$$

As a comparison, without introducing a non-zero γ, [21] shows that convergence rate of `Tiki-Taka` is only

$$\frac{1}{K}\sum_{k=0}^{K-1} \|\nabla f(W_k)\|^2 \leq O\left(\frac{1}{1 - 33P_{\max}^2/\tau^2} \sqrt{\frac{(f(W_0)-f^*)\sigma^2 L}{K}} \right). \tag{23}$$

Even though it is not a completely fair comparison since two paper relies on different assumptions, it is still worthy to compare the analysis in two paper. [21] assumes the noise should be non-zero, i.e. $[\mathbb{E}_\xi[|\nabla f(W;\xi)|]]_i \geq c_{\texttt{noise}}\sigma, \forall i \in [D]$ holds for a non-zero constant $c_{\texttt{noise}}$. Instead, this paper does not make this assumption but assumes that the objective is strongly convex. As mentioned in Section 4, the strong convexity is introduced only to ensure the existence of $P^*(W_k)$. Therefore, we believe it can be relaxed, and the convergence rate can remain unchanged, which is left as a future work. Taking that into account, we believe the comparison can provide insight of how the non-zero γ improves the convergence rate of `Tiki-Taka`.

Why does non-zero γ improve the convergence rate of `Tiki-Taka`? As discussed in Section 4, P_k is interpreted as a residual array that optimizes $f(W_k + \gamma P)$. In the ideal setting that $F(W) = 1$ and $G(W) = 0$, it can be shown that P_k converges to $P^*(W_k)$ if W_k is fixed and P_k is kept updated, even though the $W_k \neq W^*$ (hence $\nabla f(W_k) \neq 0$).

Instead, without a non-zero γ, [21] points out that P_k is an approximation of clear gradient by showing

$$\mathbb{E}_{\xi_k}[\|P_{k+1} - C\nabla f(W_k)\|^2] \tag{24}$$

$$\leq \left(1-\frac{\beta}{C}\right)\|P_k - C\nabla f(W_k)\|^2 + O(\beta C')\|\nabla f(W_k)\|^2 + \text{remainder}$$

where C, C' are constants depending on the resistive element and model dimension, and the "remainder" is the non-essential terms. Consider the case that W_k is fixed and (10) is kept iterating, in which case the increment on P_k is constant since $\gamma = 0$. Telescoping (24) we find the upper bound above only guarantee that

$$\limsup_{k\to\infty} \mathbb{E}[\|P_{k+1} - C\nabla f(W_k)\|^2] \leq O(CC'\|\nabla f(W_k)\|^2) \tag{25}$$

which means that P_k tracks the gradient accurately only when $\nabla f(W_k)$ reaches zero asymptotically. Consequently, the less accurate approximation leads to a slower rate than this paper.

C Dynamic of Non-ideal Analog Update

This section presents details about how to get the dynamic of analog update (3) appearing in Section 2 and its error analysis. The primary distinction between digital and analog training is the weight update method. As discussed in Section 1, the weight update in AIMC hardware is implemented by `Analog Update`, which sends a series of pulses to the resistive elements.

Pulse update. Consider the response of one resistive element in one cycle, which involves only one pulse. Given the initial weight w, the updated weight increases or decreases by about $\Delta w_{\min}$ depending on the pulse polarity, where $\Delta w_{\min} > 0$ is the *response granularity* determined by elements. The granularity is further scaled by a factor, which varies by the update direction due to the *asymmetric update* property of resistive elements. The notations $q_+(\cdot)$ and $q_-(\cdot)$ are used to denote the *response functions* on positive or negative sides, respectively, to describe the dominating part of the factor. In practice, the analog noise also causes a deviation of the effective factor from the response functions, referred to as *cycle variation*. It is represented by the magnitude σ_c times a random variable ξ_c with expectation 0 and variance 1. Taking all of them into account, with $s \in \{+,-\}$ being the update direction, the updated weight after receiving one pulse is $\tilde{U}_q(w,s)$ where $\tilde{U}_q(\cdot,\cdot) : \mathbb{R}\times\{+,-\} \to \mathbb{R}$ is the element-dependent update that implements the resistive element, which can be expressed as

$$\begin{aligned}\tilde{U}_q(w,s) &:= w + \Delta w_{\min}\cdot(q_s(w) + \sigma_c\xi) \\ &= \begin{cases} w + \Delta w_{\min}\cdot(q_+(w)+\sigma_c\xi_c), & s=+, \\ w - \Delta w_{\min}\cdot(q_-(w)+\sigma_c\xi_c), & s=-. \end{cases}\end{aligned} \tag{26}$$

The typical signal and noise ratio $\sigma_c/q_s(w)$ is roughly 5%-100% [80, 49], varied by the type of resistive elements. Furthermore, the response functions also vary by elements due to the imperfection in fabrication, called *element variation* (also referred to as *device variation* in literature [16]).

Equation (26) is a resistive element level equation. Existing work exploring the candidates of resistive elements usually reports the response curves similar to Figure 1, [73, 52, 49]. Taking the difference between weights in two consecutive pulse cycles and adopting statistical approaches [80], all the element-dependent quantities, including $\Delta w_{\min}$, $q_+(\cdot)$, $q_-(\cdot)$ and σ_c, can be estimated from the response curves of the resistive elements.

Analog update implemented by pulse updates. Even though the update scheme has evolved over the years [16, 71], we discuss a simplified version, called `Analog Update`, to retain the essential properties. To update the weight w by Δw, a series of pulses are sent, whose *bit length (BL)* is computed by $\text{BL} := \left\lceil \frac{|\Delta w|}{\Delta w_{\min}} \right\rceil$. After received BL pulses, the updated weight w' can be expressed as the function composition of (26) by BL times

$$w' = \underbrace{\tilde{U}_q \circ \tilde{U}_q \circ \cdots \circ \tilde{U}_q}_{\times\,\text{BL}}(w,s) =: \tilde{U}_q^{\text{BL}}(w,s). \tag{27}$$

Roughly speaking, given an ideal response $q_+(w) = q_-(w) = 1$ and $\sigma_c = 0$, BL pulses, with $\Delta w_{\min}$ increment for each individual pulse, incur the weight update Δw. Since the response granularity $\Delta w_{\min}$ is scaled by the response function $q_s(w)$, the expected increment is approximately scaled by $q_s(w)$ as well. Accordingly, we propose an approximate dynamic of `Analog Update` is given by

$w' \approx U_q(w, \Delta w)$, where $U_q(w, \Delta w)$ is defined in (3). The following theorem provides an estimation of the approximation error. It has been shown empirically that the response granularity can be made sufficiently small for updating [81, 82], implying $\Delta w_{\min} \ll \Delta w$. Therefore, we establish the error estimation of the approximation based on a small response granularity condition.

Theorem 4 (Error from discrete pulse update)**.** *Suppose the response granularity is sufficiently small such that $\Delta w_{\min} \le o(\Delta w)$. With the update direction $s = \text{sign}(\Delta w)$, the error between the true update $\tilde{U}_q^{\text{BL}}(w, s)$ and the approximated $U_q(w, \Delta w)$ is bounded by*

$$\lim_{\Delta w \to 0} \frac{|\tilde{U}_q^{\text{BL}}(w, s) - U_q(w, \Delta w)|}{|\tilde{U}_q^{\text{BL}}(w, s) - w|} = 0. \tag{28}$$

The proof of Theorem 4 is deferred to Appendix F. In Theorem 4, $|\tilde{U}_q^{\text{BL}}(w, s) - U_q(w, \Delta w)|$ is the error between the true update and the proposed dynamic, while $|\tilde{U}_q^{\text{BL}}(w, s) - w|$ is the difference between original weight and the updated one. Theorem 4 shows that the proposed dynamic dominates the update, and the approximation error is negligible when Δw is small, which holds as Δw always includes a small learning rate in gradient-based training.

Takeaway. Theorem 4 enables us to discuss the impact of response functions directly without dealing with element-specific details like response granularity $\Delta w_{\min}$ and cycle variation σ_c. Response functions are the bridge between the resistive element level equation (pulse update (26)) and algorithm level equation (dynamic of `Analog Update` (3)).

D Comparison of `Residual Learning v2` and `Tiki-Taka v2`

Introducing a digital buffer, the proposed `Residual Learning v2` has a similar form of `Tiki-Taka v2` [23]. However, there are slight differences. `Tiki-Taka v2` updates the digital buffer by

$$H_{k+\frac{1}{2}} = H_k + \beta(P_{k+1} + \varepsilon_k) \tag{29}$$

which do not include a decay coefficient in front of H_k. Furthermore, `Tiki-Taka v2` uses the gradient $\nabla f(W_k; \xi_k)$ that are solely computed on the main array W_k. Instead, `Residual Learning v2` computes gradient on a mixed weight $\bar{W}_k = W_k + \gamma P_k$. As suggested by the ablation simulations in Section K.7, the training benefits from a non-zero γ.

E Useful Lemmas and Proofs

E.1 Lemma 1: Properties of weighted norm

Lemma 1. *$\|W\|_S$ has the following properties: (a) $\|W\|_S = \|W \odot \sqrt{S}\|$; (b) $\|W\|_S \le \|W\|\sqrt{\|S\|_\infty}$; (c) $\|W\|_S \ge \|W\|\sqrt{\min\{S\}}$.*

Proof of Lemma 1. The lemma can be proven easily by definition. □

E.2 Lemma 2: Properties of weighted norm

A direct property from Definition 1 is that all $q_+(\cdot)$, $q_-(\cdot)$, and $F(\cdot)$ are bounded, as guaranteed by the following lemma.

Lemma 2. *The following statements are valid for all $W \in \mathcal{R}$. (a) $F(\cdot)$ is element-wise upper bounded by a constant $F_{\max} > 0$, i.e., $\|F(W)\|_\infty \le F_{\max}$; (b) $Q_+(\cdot)$ and $\nabla Q_-(\cdot)$ are element-wise bounded by L_Q, i.e., $\|\nabla Q_+(W)\|_\infty \le L_Q$, $\|\nabla Q_-(W)\|_\infty \le L_Q$.*

E.3 Lemma 3: Lipschitz continuity of analog update

Lemma 3. *The increment defined in (5) is Lipschitz continuous with respect to ΔW under any weighted norm $\|\cdot\|_S$, i.e., for any $W, \Delta W, \Delta W' \in \mathbb{R}^D$ and $S \in \mathbb{R}_+^D$, it holds*

$$\|\Delta W \odot F(W) - |\Delta W| \odot G(W) - (\Delta W' \odot F(W) - |\Delta W'| \odot G(W))\|_S \tag{30}$$

$$\leq F_{\max}\|\Delta W - \Delta W'\|_S.$$

Proof of Lemma 3. We prove for the case where $D = 1$ and $S = 1$, and the general case can be proven similarly. Notice that the absolute value $|\cdot|$ and vector norm $\|\cdot\|$, scalar multiplication $\times$ and element-wise multiplication $\odot$, are equivalent at that situation. We adopt both notations just for readability.

$$\begin{aligned}&\|\Delta W \odot F(W) - |\Delta W| \odot G(W) - (\Delta W' \odot F(W) - |\Delta W'| \odot G(W))\| \\ = &\|(\Delta W - \Delta W') \odot F(W) - (|\Delta W| - |\Delta W'|) \odot G(W)\|.\end{aligned} \tag{31}$$

Since $\|\Delta W - \Delta W'\| \geq \||\Delta W| - |\Delta W'|\|$ and $|G(W)| \leq |F(W)|$, we have

$$\begin{aligned}&|(\Delta W - \Delta W') \odot F(W) - (|\Delta W| - |\Delta W'|) \odot G(W)| \\ \leq &|(\Delta W - \Delta W') \odot (F(W) - |G(W)|)| \\ \leq &|\Delta W - \Delta W'|\, |F(W) - |G(W)|| \\ \leq &F_{\max}|\Delta W - \Delta W'|\end{aligned} \tag{32}$$

which completes the proof. □

E.4 Lemma 4: Element-wise product error

Lemma 4. *Let $U, V, Q \in \mathbb{R}^D$ be vectors indexed by $[D]$. Then the following inequality holds*

$$\langle U, V \odot Q\rangle \geq C_+ \langle U, V\rangle - C_- \langle |U|, |V|\rangle \tag{33}$$

where the constant C_+ and C_- are defined by

$$C_+ := \frac{1}{2}\left(\max\{Q\} + \min\{Q\}\right), \tag{34}$$

$$C_- := \frac{1}{2}\left(\max\{Q\} - \min\{Q\}\right). \tag{35}$$

Proof of Lemma 4. For any vectors $U, V, Q \in \mathbb{R}^D$, it is always valid that

$$\begin{aligned}\langle U, V \odot Q\rangle &= \sum_{i\in[D]} [U]_i[V]_i[Q]_i \\ &= \sum_{i\in[D],[U]_i[V]_i\geq 0} [U]_i[V]_i[Q]_i + \sum_{i\in[D],[U]_i[V]_i< 0} [U]_i[V]_i[Q]_i \\ &\geq \min\{Q\} \times \left(\sum_{i\in[D],[U]_i[V]_i\geq 0} [U]_i[V]_i\right) + \max\{Q\} \times \left(\sum_{i\in[D],[U]_i[V]_i< 0} [U]_i[V]_i\right) \\ &\overset{(a)}{=} C_+\left(\sum_{i\in[D],[U]_i[V]_i\geq 0} [U]_i[V]_i\right) - C_-\left(\sum_{i\in[D],[U]_i[V]_i\geq 0} |[U]_i[V]_i|\right) \\ &\quad + C_+\left(\sum_{i\in[D],[U]_i[V]_i< 0} [U]_i[V]_i\right) - C_-\left(\sum_{i\in[D],[U]_i[V]_i< 0} |[U]_i[V]_i|\right) \\ &= C_+ \sum_{i\in[D]} [U]_i[V]_i - C_- \sum_{i\in[D]} |[U]_i[V]_i| \\ &= C_+ \langle U, V\rangle - C_- \langle |U|, |V|\rangle\end{aligned} \tag{36}$$

where (a) uses the following equality

$$\min\{Q\}[U]_i[V]_i = C_+[U]_i[V]_i - C_-|[U]_i[V]_i|, \quad \text{if } [U]_i[V]_i \geq 0, \tag{37}$$

$$\max\{Q\}[U]_i[V]_i = C_+[U]_i[V]_i - C_-|[U]_i[V]_i|, \quad \text{if } [U]_i[V]_i < 0. \tag{38}$$

This completes the proof. □

F Proof of Theorem 4: Error from Pulse Update

Theorem 4 (Error from discrete pulse update)**.** *Suppose the response granularity is sufficiently small such that $\Delta w_{\min} \leq o(\Delta w)$. With the update direction $s = \operatorname{sign}(\Delta w)$, the error between the true update $\tilde{U}_q^{\mathrm{BL}}(w, s)$ and the approximated $U_q(w, \Delta w)$ is bounded by*

$$\lim_{\Delta w \to 0} \frac{|\tilde{U}_q^{\mathrm{BL}}(w, s) - U_q(w, \Delta w)|}{|\tilde{U}_q^{\mathrm{BL}}(w, s) - w|} = 0. \tag{28}$$

Proof of Theorem 4. Recall the definition of the bit length is

$$\mathrm{BL} := \left\lceil \frac{|\Delta w|}{\Delta w_{\min}} \right\rceil = \Theta\left(\frac{|\Delta w|}{\Delta w_{\min}}\right) \tag{39}$$

leading to

$$|\,\mathrm{BL}\,\Delta w_{\min} - |\Delta w|| \leq \Delta w_{\min} \quad \text{or} \quad |s\,\mathrm{BL}\,\Delta w_{\min} - \Delta w| \leq \Delta w_{\min}. \tag{40}$$

Notice that the update responding to each pulse is a $\Theta(\Delta w_{\min})$ term. Directly manipulating $U_p^{\mathrm{BL}}(w, s)$ and expanding it in Taylor series to the first-order term yields

$$\begin{aligned}
U_p^{\mathrm{BL}}(w, s) &= w + s \cdot \Delta w_{\min} \sum_{t=0}^{\mathrm{BL}-1} q_s(w + \Theta(t\Delta w_{\min})) + \Delta w_{\min} \sum_{t=0}^{\mathrm{BL}-1} \sigma_c \xi_t \\
&= w + s \cdot \Delta w_{\min} \sum_{t=0}^{\mathrm{BL}-1} q_s(w) + \sum_{t=0}^{\mathrm{BL}-1} \Theta(t(\Delta w_{\min})^2) + \Delta w_{\min} \sum_{t=0}^{\mathrm{BL}-1} \sigma_c \xi_t \\
&= w + s \cdot \Delta w_{\min} \cdot \mathrm{BL} \cdot q_s(w) + \Theta(\mathrm{BL}^2 (\Delta w_{\min})^2) + \Delta w_{\min} \cdot \sqrt{\mathrm{BL}} \cdot \sigma_c \xi \\
&= w + \Delta w \cdot q_s(w) + (s\,\mathrm{BL}\,\Delta w_{\min} - \Delta w) + \Theta((\Delta w)^2) + \sqrt{\Delta w_{\min}} \cdot \sqrt{\Delta w} \cdot \sigma_c \xi \\
&= U_q(w, \Delta w) + \Theta(\Delta w_{\min}) + \Theta((\Delta w)^2) + \Theta(\sqrt{\Delta w_{\min}} \cdot \sqrt{\Delta w} \cdot \sigma_c)
\end{aligned} \tag{41}$$

where $\xi := \frac{1}{\sqrt{\mathrm{BL}}} \sum_{t=0}^{\mathrm{BL}-1} \xi_t$ is the accumulated noise with variance 1. The proof is completed. $\square$

G Proof of Theorem 1: Implicit Bias of Analog Training

In this section, we provide a full version of Theorem 1. Before that, we formally define the accumulated asymmetric function $R_c(W) : \mathbb{R}^D \to \mathbb{R}^D$ element-wise. Let $R(W) := \frac{G(W)}{F(W)}$ be the asymmetric ratio and $R_c(W)$ is defined by

$$[R_c(W)]_i := \int_{\tau_i^{\min}}^{[W]_i} [R(W)]_i \,\mathrm{d}[W]_i \tag{42}$$

which satisfies $\nabla R_c(W) = R(W)$ and $\nabla \langle \Sigma, R_c(W) \rangle = \Sigma \odot R(W)$. Since we do not further assume stronger properties for response functions, like monotonicity, it is hard to provide strong claims on the shape of $R(W)$ or $R_c(W)$. Here we provide the expression of $R_c(W)$ for the linear response functions $Q_+(W) = 1 - \frac{W}{\tau}$, $Q_-(W) = 1 + \frac{W}{\tau}$. In this case, $F(W) = 1$ and $G(W) = \frac{W}{\tau}$ based on definition (4); and hence $R(W) = \frac{G(W)}{F(W)} = \frac{W}{\tau}$. Accordingly, the accumulated asymmetric function is given by

$$\begin{aligned}
[R_c(W)]_i &= \int_{\tau_i^{\min}}^{[W]_i} [R(W)]_i \,\mathrm{d}[W]_i = \int_{\tau_i^{\min}}^{[W]_i} \frac{[W]_i}{\tau} \,\mathrm{d}[W]_i \\
&= \frac{1}{2\tau}([W]_i)^2 - \frac{1}{2\tau}(\tau_i^{\min})^2.
\end{aligned} \tag{43}$$

Therefore, the last term in the objective (7) becomes

$$\langle \Sigma, R_c(W) \rangle = \sum_{i=1}^{D} [\Sigma]_i [R_c(W)]_i = \sum_{i=1}^{D} [\Sigma]_i \left(\frac{1}{2\tau}([W]_i)^2 - \frac{1}{2\tau}(\tau_i^{\min})^2 \right) \tag{44}$$

$$= \frac{1}{2\tau}\|W\|_{\Sigma}^2 + \text{const.}$$

which is a weighted ℓ_2 norm regularization term. In the scalar case, it reduces to (45).

$$\min_{W} \ f_{\Sigma}(W) := f(W) + \frac{\Sigma}{2\tau}\|W\|^2 \tag{45}$$

It is noticeable that if the ratio $R(W)$ is monotonic at each coordinate, $R_c(W)$ reaches its minimum at $W^{\diamond}$. Therefore, $R_c(W)$ has no impact on the convergence only when the optimal point of $f(W)$ is $W^{\diamond}$.

Theorem 5 (Implicit Penalty, full version of Theorem 1)**.** *Let $T(w)$ denote the effective update of `Analog SGD`.*

$$T(W) := \left|\frac{\mathbb{E}_{\xi}\left[U_q(W, -\alpha f'(W;\xi))\right] - W}{\alpha}\right| = \left|\mathbb{E}_{\xi}[f'(w;\xi)] \odot F(W) - \mathbb{E}_{\xi}[|f'(W;\xi)|] \odot G(W)\right|. \tag{46}$$

`Analog SGD` implicitly optimizes the following penalized objective

$$\min_{W} \ f_{\Sigma}(W) := f(W) + \langle \Sigma, R_c(W)\rangle \tag{47}$$

in the sense that there exists a point W^S given by

$$W^S := (\nabla^2 f(W^*) - \nabla R(W^{\diamond})\Sigma)^{-1}(\nabla^2 f(W^*)\, W^* - \nabla R(W^{\diamond})\Sigma\, W^{\diamond}) \tag{48}$$

such that $\|\nabla f_{\Sigma}(W^S)\| \leq O((W^{\diamond} - W^*)^2)$ and $T(W^S) \leq O((W^{\diamond} - W^*)^2)$. Both $T(W^S)$ and $\|\nabla f_{\Sigma}(W^S)\|$ are significantly smaller than $T(W^{\diamond}) = O(|W^{\diamond} - W^*|)$ and $T(W^*) = O(|W^{\diamond} - W^*|)$ when $W^{\diamond}$ is close to W^*.

If W is a scalar, i.e. $D = 1$, (48) reduces to (45)

$$\min_{W} \ f_{\Sigma}(W) := f(W) + \frac{\Sigma}{2\tau}\|W\|^2 \tag{49}$$

with its solution

$$W^S := \frac{f''(W^*)\, W^* - R'(W^{\diamond})\Sigma\, W^{\diamond}}{f''(W^*) - R'(W^{\diamond})\Sigma}. \tag{50}$$

Proof of Theorem 1 and 5. We separately show that $\|\nabla f_{\Sigma}(W)\| \leq O((W^{\diamond} - W^*)^2)$ and $T(W^S) \leq O((W^{\diamond} - W^*)^2)$.

Proof of $\|\nabla f_{\Sigma}(W)\| \leq O((W^{\diamond} - W^*)^2)$. The gradient of the penalized objective $f_{\Sigma}(W)$ is given by

$$\nabla f_{\Sigma}(W) = \nabla f(W) + \Sigma \odot R(W). \tag{51}$$

Leveraging the fact that $\nabla f(W^*) = 0$, $\frac{G(W^{\diamond})}{F(W^{\diamond})} = 0$, as well as Taylor expansion given by

$$\nabla f(W^S) = \nabla f(W^*) + \nabla^2 f(W^*)(W^S - W^*) + O((W^S - W^*)^2), \tag{52}$$

$$\frac{G(W^S)}{F(W^S)} = \frac{G(W^{\diamond})}{F(W^{\diamond})} + \nabla R(W^{\diamond})(W^S - W^{\diamond}) + O((W^S - W^{\diamond})^2), \tag{53}$$

we bound the gradient of the penalized objective as follows

$$\|\nabla f_{\Sigma}(W)\| = \left\|\nabla f(W^S) - \Sigma \odot \frac{G(W^S)}{F(W^S)}\right\| \tag{54}$$

$$= \left\|\nabla^2 f(W^*)(W^S - W^*) + O((W^S - W^*)^2) - \Sigma \odot (\nabla R(W^{\diamond})(W^S - W^{\diamond})) + O((W^S - W^{\diamond})^2)\right\|$$

$$= O((W^S - W^*)^2) + O((W^S - W^{\diamond})^2)$$

$$= O((W^* - W^{\diamond})^2)$$

where the last inequality holds by the definition of W^S.

Proof of $T(w^S) \leq O((w^\diamond - w^*)^2)$**.** By the definition of effective update $T(W^S)$, we have

$$
\begin{aligned}
&\left\|\frac{\mathbb{E}_\xi\left[U_q(W^S, -\alpha\nabla f(W^S;\xi))\right] - W^S}{\alpha}\right\| \\
=& \left\|\mathbb{E}_\xi[\nabla f(W^S;\xi)] \odot F(W^S) - \mathbb{E}_\xi[|\nabla f(W^S;\xi)|] \odot G(W^S)\right\| \\
\leq& \left\|\left(\nabla f(W^S) - \mathbb{E}_\xi[|\nabla f(W^S;\xi)|] \odot \frac{G(W^S)}{F(W^S)}\right)\right\| F_{\max} \\
\leq& \left\|\nabla f(W^S) - \mathbb{E}_\xi[|\nabla f(W^S;\xi) - \nabla f(W^S)|] \odot \frac{G(W^S)}{F(W^S)}\right\| F_{\max} \\
&+ \left\|(\mathbb{E}_\xi[|\nabla f(W^S;\xi)|] - \mathbb{E}_\xi[|\nabla f(W^S;\xi) - \nabla f(W^S)|]) \odot \frac{G(W^S)}{F(W^S)}\right\| F_{\max} \\
\leq& \|\nabla f_\Sigma(W)\| F_{\max} + \left\|(\mathbb{E}_\xi[|\nabla f(W^S;\xi)|] - \mathbb{E}_\xi[|\nabla f(W^S;\xi) - \nabla f(W^S)|]) \odot \frac{G(W^S)}{F(W^S)}\right\| F_{\max}
\end{aligned}
\tag{55}
$$

The first term in the right-hand side (RHS) of (55) is already bounded by (54). By inequality $||x| - |y|| \leq |x - y|$ for any $x, y \in \mathbb{R}$, the second term in the RHS of (55) is bounded by

$$
\begin{aligned}
&\left\|(\mathbb{E}_\xi[|\nabla f(W^S;\xi)|] - \mathbb{E}_\xi[|\nabla f(W^S;\xi) - \nabla f(W^S)|]) \odot \frac{G(W^S)}{F(W^S)}\right\| \\
\leq& \left\||\nabla f(W^S)| \odot \frac{G(W^S)}{F(W^S)}\right\| \\
=& \left\||\nabla f(W^S) - \nabla f(W^*)| \odot \left(\frac{G(W^S)}{F(W^S)} - \frac{G(W^\diamond)}{F(W^\diamond)}\right)\right\| \\
\leq& O(|W^S - W^*|) \cdot O(|W^S - W^\diamond|) \\
=& O((W^* - W^\diamond)^2)
\end{aligned}
\tag{56}
$$

Plugging back (54) and (56) into (55) shows $T(w^S) \leq O((w^\diamond - w^*)^2)$. It is trivial to prove $T(W^\diamond) = O(|W^\diamond - W^*|)$ and $T(W^*) = O(|W^\diamond - W^*|)$ by the definition of W^S and (52). □

H Proof of Theorem 2: Convergence of `Analog SGD`

Theorem 2 (Inexact convergence of `Analog SGD`)**.** *Under Assumption 1–3, if the learning rate is set as* $\alpha = O(1/\sqrt{K})$*, it holds that*

$$
E_K \leq O\left(\sqrt{\sigma^2/K} + \sigma^2 S_K^{\texttt{ASGD}}\right)
\tag{8}
$$

where $S_K^{\texttt{ASGD}}$ *denotes the amplification factor given by* $S_K^{\texttt{ASGD}} := \frac{1}{K}\sum_{k=0}^{K-1}\left\|\frac{G(W_k)}{\sqrt{F(W_k)}}\right\|_\infty^2$.

Proof of Theorem 2. The L-smooth assumption (Assumption 1) implies that

$$
\mathbb{E}_{\xi_k}[f(W_{k+1})] \leq f(W_k) + \underbrace{\mathbb{E}_{\xi_k}[\langle\nabla f(W_k), W_{k+1} - W_k\rangle]}_{(a)} + \underbrace{\frac{L}{2}\mathbb{E}_{\xi_k}[\|W_{k+1} - W_k\|^2]}_{(b)}.
\tag{57}
$$

Next, we will handle the second and the third terms in the RHS of (57) separately.

Bound of the second term (a). To bound term (a) in the RHS of (57), we leverage the assumption that noise has expectation 0 (Assumption 2)

$$
\begin{aligned}
&\mathbb{E}_{\xi_k}[\langle\nabla f(W_k), W_{k+1} - W_k\rangle] \\
=& \alpha\mathbb{E}_{\xi_k}\left[\left\langle\nabla f(W_k) \odot \sqrt{F(W_k)}, \frac{W_{k+1} - W_k}{\alpha\sqrt{F(W_k)}} + (\nabla f(W_k;\xi_k) - \nabla f(W_k)) \odot \sqrt{F(W_k)}\right\rangle\right] \\
=& -\frac{\alpha}{2}\|\nabla f(W_k) \odot \sqrt{F(W_k)}\|^2
\end{aligned}
\tag{58}
$$

$$
-\frac{1}{2\alpha}\mathbb{E}_{\xi_k}\left[\left\|\frac{W_{k+1}-W_k}{\sqrt{F(W_k)}}+\alpha(\nabla f(W_k;\xi_k)-\nabla f(W_k))\odot\sqrt{F(W_k)}\right\|^2\right]
+\frac{1}{2\alpha}\mathbb{E}_{\xi_k}\left[\left\|\frac{W_{k+1}-W_k}{\sqrt{F(W_k)}}+\alpha\nabla f(W_k;\xi_k)\odot\sqrt{F(W_k)}\right\|^2\right].
$$

The second term of the RHS of (58) is bounded by

$$
\begin{aligned}
&\frac{1}{2\alpha}\mathbb{E}_{\xi_k}\left[\left\|\frac{W_{k+1}-W_k}{\sqrt{F(W_k)}}+\alpha(\nabla f(W_k;\xi_k)-\nabla f(W_k))\odot\sqrt{F(W_k)}\right\|^2\right] \qquad (59)\\
=&\frac{1}{2\alpha}\mathbb{E}_{\xi_k}\left[\left\|\frac{W_{k+1}-W_k+\alpha(\nabla f(W_k;\xi_k)-\nabla f(W_k))\odot F(W_k)}{\sqrt{F(W_k)}}\right\|^2\right]\\
\ge&\frac{1}{2\alpha F_{\max}}\mathbb{E}_{\xi_k}\left[\|W_{k+1}-W_k+\alpha(\nabla f(W_k;\xi_k)-\nabla f(W_k))\odot F(W_k)\|^2\right].
\end{aligned}
$$

The third term in the RHS of (58) can be bounded by variance decomposition and bounded variance assumption (Assumption 2)

$$
\begin{aligned}
&\frac{1}{2\alpha}\mathbb{E}_{\xi_k}\left[\left\|\frac{W_{k+1}-W_k}{\sqrt{F(W_k)}}+\alpha\nabla f(W_k;\xi_k)\odot\sqrt{F(W_k)}\right\|^2\right] \qquad (60)\\
=&\frac{\alpha}{2}\mathbb{E}_{\xi_k}\left[\left\||\nabla f(W_k;\xi_k)|\odot\frac{G(W_k)}{\sqrt{F(W_k)}}\right\|^2\right]\\
\le&\frac{\alpha}{2}\left\||\nabla f(W_k)|\odot\frac{G(W_k)}{\sqrt{F(W_k)}}\right\|^2+\frac{\alpha\sigma^2}{2}\left\|\frac{G(W_k)}{\sqrt{F(W_k)}}\right\|_\infty^2.
\end{aligned}
$$

Define the saturation vector $R(W_k)\in\mathbb{R}^D$ by

$$
\begin{aligned}
R(W_k):=&F(W_k)^{\odot 2}-G(W_k)^{\odot 2}=(F(W_k)+G(W_k))\odot(F(W_k)-G(W_k)) \qquad (61)\\
=&Q_+(W_k)\odot Q_-(W_k).
\end{aligned}
$$

Note that the first term in the RHS of (58) and the second term in the RHS of (60) can be bounded by

$$
\begin{aligned}
&-\frac{\alpha}{2}\|\nabla f(W_k)\odot\sqrt{F(W_k)}\|^2+\frac{\alpha}{2}\left\||\nabla f(W_k)|\odot\frac{G(W_k)}{\sqrt{F(W_k)}}\right\|^2 \qquad (62)\\
=&-\frac{\alpha}{2}\sum_{d\in[D]}\left([\nabla f(W_k)]_d^2\left([F(W_k)]_d-\frac{[G(W_k)]_d^2}{[F(W_k)]_d}\right)\right)\\
=&-\frac{\alpha}{2}\sum_{d\in[D]}\left([\nabla f(W_k)]_d^2\left(\frac{[F(W_k)]_d^2-[G(W_k)]_d^2}{[F(W_k)]_d}\right)\right)\\
\le&-\frac{\alpha}{2F_{\max}}\sum_{d\in[D]}\left([\nabla f(W_k)]_d^2\left([F(W_k)]_d^2-[G(W_k)]_d^2\right)\right)\\
=&-\frac{\alpha}{2F_{\max}}\|\nabla f(W_k)\|_{R(W_k)}^2\le 0.
\end{aligned}
$$

Plugging (59) to (62) into (58), we bound the term (a) by

$$
\begin{aligned}
&\mathbb{E}_{\xi_k}[\langle\nabla f(W_k),W_{k+1}-W_k\rangle] \qquad (63)\\
=&-\frac{\alpha}{2F_{\max}}\|\nabla f(W_k)\|_{R(W_k)}^2+\frac{\alpha\sigma^2}{2}\left\|\frac{G(W_k)}{\sqrt{F(W_k)}}\right\|_\infty^2
\end{aligned}
$$

$$
- \frac{1}{2\alpha F_{\max}} \mathbb{E}_{\xi_k}\left[\|W_{k+1} - W_k + \alpha(\nabla f(W_k;\xi_k) - \nabla f(W_k)) \odot F(W_k)\|^2\right].
$$

Bound of the third term (b). The third term (b) in the RHS of (57) is bounded by

$$
\begin{aligned}
&\frac{L}{2}\mathbb{E}_{\xi_k}[\|W_{k+1} - W_k\|^2] \qquad (64)\\
\le\ & L\mathbb{E}_{\xi_k}[\|W_{k+1} - W_k + \alpha(\nabla f(W_k;\xi_k) - \nabla f(W_k)) \odot F(W_k)\|^2]\\
&+ \alpha^2 L\mathbb{E}_{\xi_k}[\|(\nabla f(W_k;\xi_k) - \nabla f(W_k)) \odot F(W_k)\|^2]\\
\le\ & L\mathbb{E}_{\xi_k}[\|W_{k+1} - W_k + \alpha(\nabla f(W_k;\xi_k) - \nabla f(W_k)) \odot F(W_k)\|^2] + \alpha^2 L F_{\max}^2\sigma^2
\end{aligned}
$$

where the last inequality leverages the bounded variance of noise (Assumption 2) and the fact that $F(W_k)$ is bounded by $F_{\max}$ element-wise.

Substituting (63) and (64) back into (57), we have

$$
\begin{aligned}
\mathbb{E}_{\xi_k}[f(W_{k+1})] \le\ & f(W_k) - \frac{\alpha}{2F_{\max}}\|\nabla f(W_k)\|^2_{R(W_k)} + \alpha^2 L F_{\max}^2\sigma^2 + \frac{\alpha\sigma^2}{2}\left\|\frac{G(W_k)}{\sqrt{F(W_k)}}\right\|_\infty^2 \qquad (65)\\
&- \frac{1}{F_{\max}}\left(\frac{1}{2\alpha} - LF_{\max}\right)\mathbb{E}_{\xi_k}[\|W_{k+1} - W_k + \alpha(\nabla f(W_k;\xi_k) - \nabla f(W_k)) \odot F(W_k)\|^2].
\end{aligned}
$$

The third term in the RHS of (65) can be bounded by

$$
\begin{aligned}
&\mathbb{E}_{\xi_k}[\|W_{k+1} - W_k + \alpha(\nabla f(W_k;\xi_k) - \nabla f(W_k)) \odot F(W_k)\|^2] \qquad (66)\\
=\ & \alpha^2\mathbb{E}_{\xi_k}[\|\nabla f(W_k) \odot F(W_k) + |\nabla f(W_k;\xi_k)| \odot G(W_k)\|^2]\\
\ge\ & \frac{1}{2}\alpha^2\mathbb{E}_{\xi_k}[\|\nabla f(W_k) \odot F(W_k) + |\nabla f(W_k)| \odot G(W_k)\|^2]\\
&- \alpha^2\mathbb{E}_{\xi_k}[\|(|\nabla f(W_k)| - |\nabla f(W_k;\xi_k)|) \odot G(W_k)\|^2]\\
\ge\ & \frac{1}{2}\alpha^2\mathbb{E}_{\xi_k}[\|\nabla f(W_k) \odot F(W_k) + |\nabla f(W_k)| \odot G(W_k)\|^2]\\
&- \alpha^2\mathbb{E}_{\xi_k}[\|(\nabla f(W_k) - \nabla f(W_k;\xi_k)) \odot G(W_k)\|^2]\\
\ge\ & \frac{1}{2}\alpha^2\mathbb{E}_{\xi_k}[\|\nabla f(W_k) \odot F(W_k) + |\nabla f(W_k)| \odot G(W_k)\|^2] - \alpha^2 F_{\max}\sigma^2\left\|\frac{G(W_k)}{\sqrt{F(W_k)}}\right\|_\infty^2
\end{aligned}
$$

where the first inequality holds because $\|x\|^2 \ge \frac{1}{2}\|x-y\|^2 - \|y\|^2$ for any $x, y \in \mathbb{R}^D$, the second inequality comes from $||x| - |y|| \le |x - y|$ for any $x, y \in \mathbb{R}$, and the last inequality holds because

$$
\begin{aligned}
&\mathbb{E}_{\xi_k}[\|(\nabla f(W_k) - \nabla f(W_k;\xi_k)) \odot G(W_k)\|^2] \qquad (67)\\
=\ & \mathbb{E}_{\xi_k}\left[\left\|(\nabla f(W_k) - \nabla f(W_k;\xi_k)) \odot \frac{G(W_k)}{\sqrt{F(W_k)}} \odot \sqrt{F(W_k)}\right\|^2\right]\\
\le\ & F_{\max}\mathbb{E}_{\xi_k}\left[\left\|(\nabla f(W_k) - \nabla f(W_k;\xi_k)) \odot \frac{G(W_k)}{\sqrt{F(W_k)}}\right\|^2\right]\\
\le\ & F_{\max}\mathbb{E}_{\xi_k}\left[\|\nabla f(W_k) - \nabla f(W_k;\xi_k)\|^2\right]\left\|\frac{G(W_k)}{\sqrt{F(W_k)}}\right\|_\infty^2\\
\le\ & F_{\max}\sigma^2\left\|\frac{G(W_k)}{\sqrt{F(W_k)}}\right\|_\infty^2.
\end{aligned}
$$

The learning rate $\alpha \le \frac{1}{4LF_{\max}}$ implies that $\frac{1}{2\alpha} - LF_{\max} \le \frac{1}{4\alpha}$ in (65), which leads (57) to

$$
\mathbb{E}_{\xi_k}[f(W_{k+1})] \le f(W_k) - \frac{\alpha}{2F_{\max}}\|\nabla f(W_k)\|^2_{R(W_k)} + \alpha^2 L F_{\max}^2\sigma^2 + \alpha\sigma^2\left\|\frac{G(W_k)}{\sqrt{F(W_k)}}\right\|_\infty^2 \qquad (68)
$$

$$- \frac{\alpha}{8F_{\max}} \|\nabla f(W_k) \odot F(W_k) + |\nabla f(W_k)| \odot G(W_k)\|^2.$$

Reorganizing (68), taking expectation over all $\xi_K, \xi_{K-1}, \cdots, \xi_0$, and averaging them for k from 0 to $K-1$ deduce that

$$\begin{aligned}
E_K &= \frac{1}{K} \sum_{k=0}^{K} \mathbb{E}[\|\nabla f(W_k) \odot F(W_k) + |\nabla f(W_k)| \odot G(W_k)\|^2 + 4\|\nabla f(W_k)\|^2_{R(W_k)}] \\
&\le \frac{8F_{\max}(f(W_0) - \mathbb{E}[f(W_{k+1})])}{\alpha K} + 8\alpha L F_{\max}^3 \sigma^2 + 8F_{\max}\sigma^2 \times \frac{1}{K} \sum_{k=0}^{K-1} \left\| \frac{G(W_k)}{\sqrt{F(W_k)}} \right\|_\infty^2 \\
&\le \frac{8F_{\max}(f(W_0) - f^*)}{\alpha K} + 8\alpha L F_{\max}^3 \sigma^2 + 8F_{\max}\sigma^2 \times \frac{1}{K} \sum_{k=0}^{K-1} \left\| \frac{G(W_k)}{\sqrt{F(W_k)}} \right\|_\infty^2 \\
&= 16F_{\max}^2 \sqrt{\frac{(f(W_0) - f^*)\sigma^2 L}{K}} + 8F_{\max}\sigma^2 S_K^{\texttt{ASGD}}
\end{aligned} \tag{69}$$

where the last equality chooses the learning rate as $\alpha = \frac{1}{F_{\max}} \sqrt{\frac{f(W_0)-f^*}{\sigma^2 LK}}$. The proof is completed. $\square$

Remark 1 (Tighter bound without saturation)**.** *Assuming the saturation never happens during the training, i.e.* $R(W_k) \ge R_{\min}^{\texttt{RL}} > 0$ *for all* $k \in [K]$*, we get a tighter bound in* (68) *by leveraging* $\|\nabla f(W_k)\|^2_{R(W_k)} \ge \min\{R(W_k)\} \|\nabla f(W_k)\|^2 \ge R_{\min}^{\texttt{RL}} \|\nabla f(W_k)\|^2$

$$\mathbb{E}_{\xi_k}[f(W_{k+1})] \le f(W_k) - \frac{\alpha}{2F_{\max}} \|\nabla f(W_k)\|^2_{R(W_k)} + \alpha^2 L F_{\max}^2 \sigma^2 + \alpha\sigma^2 \left\| \frac{G(W_k)}{\sqrt{F(W_k)}} \right\|_\infty^2 \tag{70}$$

which leads to

$$\begin{aligned}
&\frac{1}{K} \sum_{k=0}^{K} [\|\nabla f(W_k)\|^2] \\
=& \frac{4F_{\max}^2}{R_{\min}^{\texttt{RL}}} \sqrt{\frac{(f(W_0) - f^*)\sigma^2 L}{K}} + 2F_{\max}\sigma^2 \times \frac{1}{K} \sum_{k=0}^{K-1} \left\| \frac{G(W_k)}{\sqrt{F(W_k)}} \right\|_\infty^2 \bigg/ \min\{R(W_k)\}.
\end{aligned} \tag{71}$$

It exactly reduces to the result for the convergence of `Analog SGD` *in [21] on special linear repsonse functions, as discussed in Appendix B.*

I Proof of Theorem 3: Convergence of `Residual Learning`

This section provides the convergence guarantee of the `Tiki-Taka` under the strongly convex assumption.

Theorem 3 (Convergence of `Residual Learning`)**.** *Under Assumptions 1–3 and 4, with the learning rate* $\alpha = O\left(\sqrt{1/\sigma^2 K}\right)$*,* $\beta = O(\alpha\gamma^{3/2})$*, it holds for* `Residual Learning` *that*

$$E_K^{\texttt{RL}} \le O\left(\sqrt{\sigma^2/K} + \sigma^2 S_K^{\texttt{RL}}\right) \tag{13}$$

where $S_K^{\texttt{RL}}$ *denotes the amplification factor of* P_k *given by* $S_K^{\texttt{RL}} := \frac{1}{K}\sum_{k=0}^{K} \left\| \frac{G(P_k)}{\sqrt{F(P_k)}} \right\|_\infty^2$.

I.1 Main proof

Proof of Theorem 3. The proof of the `Tiki-Taka` convergence relies on the following two lemmas, which provide the sufficient descent of W_k and $\bar{W}_k$, respectively.

Lemma 5 (Descent Lemma of $\bar{W}_k$)**.** *Suppose Assumptions 1–2 hold. It holds for* `Tiki-Taka` *that*

$$\mathbb{E}_{\xi_k}[f(\bar{W}_{k+1})] \leq f(\bar{W}_k) - \frac{\alpha}{4F_{\max}}\|\nabla f(\bar{W}_k)\|^2_{R(P_k)} + 2\alpha\sigma^2 \left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|^2_\infty + 2\alpha^2 L F_{\max}^2\sigma^2 \tag{72}$$
$$- \frac{\alpha\gamma}{8F_{\max}}\|\nabla f(\bar{W}_k) \odot F(P_k) + |\nabla f(\bar{W}_k)| \odot G(P_k)\|^2$$
$$+ \frac{F_{\max}}{\alpha}\mathbb{E}_{\xi_k}\left[\|W_{k+1} - W_k\|^2_{R(P_k)^\dagger}\right] + \mathbb{E}_{\xi_k}[\|W_{k+1} - W_k\|^2].$$

Lemma 6 (Descent Lemma of W_k)**.** *It holds for* `Tiki-Taka` *that*

$$\|W_{k+1} - W^*\|^2 \leq \|W_k - W^*\|^2 - \frac{\beta}{2\gamma F_{\max}}\|W_k - W^*\|^2_{R(W_k)} \tag{73}$$
$$- \frac{\beta\gamma}{2F_{\max}}\|P^*(W_k) \odot F(W_k) - |P^*(W_k)| \odot G(W_k)\|^2$$
$$+ \frac{2\beta F_{\max}^3}{\gamma}\|P_{k+1} - P^*(W_k)\|^2_{R(W_k)^\dagger} + 2\beta^2\|P_{k+1} - P^*(W_k)\|^2.$$

The proof of Lemma 5 and 6 are deferred to Section I.2 and I.3, respectively. For a sufficiently large γ, $P^*(W_k)$ is ensured to be located in the dynamic range of the analog array P_k. Therefore, we may assume both $q_+(P_k)$ and $q_-(P_k)$ are non-zero, equivalently, there exists a non-zero constant $R_{\min}^{\mathtt{RL}}$ such that $\min\{R(P_k)\} \geq R_{\min}^{\mathtt{RL}}$ for all k. Under this condition, we have the following inequalities

$$\frac{\alpha}{4F_{\max}}\|\nabla f(\bar{W}_k)\|^2_{R(P_k)} \geq \frac{\alpha R_{\min}^{\mathtt{RL}}}{4F_{\max}}\|\nabla f(\bar{W}_k)\|^2, \tag{74}$$

$$\frac{F_{\max}}{\alpha\gamma}\|W_{k+1} - W_k\|^2_{R(P_k)^\dagger} \leq \frac{F_{\max}}{\alpha\gamma R_{\min}^{\mathtt{RL}}}\|W_{k+1} - W_k\|^2. \tag{75}$$

Similarly, we bound the term $\|P_{k+1} - P^*(W_k)\|^2_{R(W_k)^\dagger}$ in (72) by

$$\frac{2\beta F_{\max}^3}{\gamma}\|P_{k+1} - P^*(W_k)\|^2_{R(W_k)^\dagger} \leq \frac{2\beta F_{\max}^3}{\gamma\min\{R(W_k)\}}\|P_{k+1} - P^*(W_k)\|^2. \tag{76}$$

Notice it is only required to have $\min\{R(W_k)\} > 0$ for the inequality to hold.

By inequality (75), the last two terms in the RHS of (72) is bounded by

$$\frac{F_{\max}}{\alpha}\mathbb{E}_{\xi_k}\left[\|W_{k+1} - W_k\|^2_{R(P_k)^\dagger}\right] + \mathbb{E}_{\xi_k}[\|W_{k+1} - W_k\|^2] \tag{77}$$
$$= \frac{F_{\max}}{\alpha R_{\min}^{\mathtt{RL}}}\mathbb{E}_{\xi_k}\left[\|W_{k+1} - W_k\|^2\right] + \mathbb{E}_{\xi_k}[\|W_{k+1} - W_k\|^2]$$
$$\overset{(a)}{\leq} \frac{2F_{\max}}{\alpha R_{\min}^{\mathtt{RL}}}\mathbb{E}_{\xi_k}\left[\|W_{k+1} - W_k\|^2\right] = \frac{2\beta^2 F_{\max}}{\alpha R_{\min}^{\mathtt{RL}}}\|P_{k+1} \odot F(W_k) - |P_{k+1}| \odot G(W_k)\|^2$$
$$\leq \frac{4\beta^2 F_{\max}}{\alpha R_{\min}^{\mathtt{RL}}}\|P^*(W_k) \odot F(W_k) - |P^*(W_k)| \odot G(W_k)\|^2$$
$$+ \frac{4\beta^2 F_{\max}}{\alpha R_{\min}^{\mathtt{RL}}}\|P_{k+1} \odot F(W_k) - |P_{k+1}| \odot G(W_k) - (P^*(W_k) \odot F(W_k) - |P^*(W_k)| \odot G(W_k))\|^2$$
$$\overset{(b)}{\leq} \frac{4\beta^2 F_{\max}}{\alpha R_{\min}^{\mathtt{RL}}}\|P^*(W_k) \odot F(W_k) - |P^*(W_k)| \odot G(W_k)\|^2 + \frac{4\beta^2 F_{\max}}{\alpha R_{\min}^{\mathtt{RL}}}\|P_{k+1} - P^*(W_k)\|^2.$$

where (a) holds if learning rate α is sufficiently small such that $\frac{F_{\max}}{\alpha\gamma R_{\min}^{\mathtt{RL}}} \geq 1$; (b) comes from the Lipschitz continuity of the analog update (c.f. Lemma 3).

With all the inequalities and lemmas above, we are ready to prove the main conclusion in Theorem 3 now. Define a Lyapunov function by

$$\mathbb{V}_k := f(\bar{W}_k) - f^* + C\|W_k - W^*\|^2. \tag{78}$$

By Lemmas 5 and 6, we show that $\mathbb{V}_k$ has sufficient descent in expectation

$$
\begin{aligned}
&\mathbb{E}_{\xi_k}[\mathbb{V}_{k+1}] \qquad (79)\\
=& \mathbb{E}_{\xi_k}\left[f(\bar{W}_{k+1}) - f^* + C\|W_{k+1} - W^*\|^2\right] \\
\le& f(\bar{W}_k) - f^* - \frac{\alpha}{4F_{\max}}\|\nabla f(\bar{W}_k)\|^2_{R(P_k)} + 2\alpha\sigma^2 \left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|_\infty^2 + 2\alpha^2 L F_{\max}^2\sigma^2 \\
&- \frac{\alpha}{8F_{\max}}\|\nabla f(\bar{W}_k)\odot F(P_k) + |\nabla f(\bar{W}_k)|\odot G(P_k)\|^2 \\
&+ \frac{4\beta^2 F_{\max}}{\alpha R_{\min}^{\mathsf{RL}}}\|P^*(W_k)\odot F(W_k) - |P^*(W_k)|\odot G(W_k)\|^2 + \frac{4\beta^2 F_{\max}}{\alpha R_{\min}^{\mathsf{RL}}}\mathbb{E}_{\xi_k}[\|P_{k+1} - P^*(W_k)\|^2] \\
&+ C\Bigg(\|W_k - W^*\|^2 - \frac{\beta}{2\gamma F_{\max}}\|W_k - W^*\|^2_{R(W_k)} + \frac{3\beta F_{\max}^3}{\gamma \min\{R(W_k)\}}\mathbb{E}_{\xi_k}[\|P_{k+1} - P^*(W_k)\|^2] \\
&\qquad - \frac{\beta\gamma}{2F_{\max}}\|P^*(W_k)\odot F(W_k) - |P^*(W_k)|\odot G(W_k)\|^2 \Bigg) \\
\le& \mathbb{V}_k - \frac{\alpha R_{\min}^{\mathsf{RL}}}{4F_{\max}}\|\nabla f(\bar{W}_k)\|^2 + 2\alpha\sigma^2 \left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|_\infty^2 + 2\alpha^2 L F_{\max}^2\sigma^2 \\
&- \frac{\alpha}{8F_{\max}}\|\nabla f(\bar{W}_k)\odot F(P_k) + |\nabla f(\bar{W}_k)|\odot G(P_k)\|^2 \\
&- \left(\frac{\beta\gamma}{2F_{\max}}C - \frac{4\beta^2 F_{\max}}{\alpha R_{\min}^{\mathsf{RL}}}\right)\|P^*(W_k)\odot F(W_k) - |P^*(W_k)|\odot G(W_k)\|^2 \\
&+ \left(\frac{3\beta F_{\max}^3}{\gamma \min\{R(W_k)\}}C + \frac{4\beta^2 F_{\max}}{\alpha R_{\min}^{\mathsf{RL}}}\right)\mathbb{E}_{\xi_k}[\|P_{k+1} - P^*(W_k)\|^2] - \frac{\beta}{2\gamma F_{\max}}C\|W_k - W^*\|^2_{R(W_k)}.
\end{aligned}
$$

Let $C = \frac{10\beta F_{\max}^2}{\alpha R_{\min}^{\mathsf{RL}}\gamma}$, which leads to the positive coefficient in front of $\|P_{k+1} - P^*(W_k)\|^2$, i.e.

$$
\begin{aligned}
&\mathbb{E}_{\xi_k}[\mathbb{V}_{k+1}] \qquad (80)\\
\le& \mathbb{V}_k - \frac{\alpha R_{\min}^{\mathsf{RL}}}{4F_{\max}}\|\nabla f(\bar{W}_k)\|^2 + 2\alpha\sigma^2 \left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|_\infty^2 + 2\alpha^2 L F_{\max}^2\sigma^2 \\
&- \frac{\alpha}{8F_{\max}}\|\nabla f(\bar{W}_k)\odot F(P_k) + |\nabla f(\bar{W}_k)|\odot G(P_k)\|^2 \\
&- \frac{\beta^2 F_{\max}}{\alpha R_{\min}^{\mathsf{RL}}}\|P^*(W_k)\odot F(W_k) - |P^*(W_k)|\odot G(W_k)\|^2 \\
&+ \left(\frac{30\beta^2 F_{\max}^5}{\alpha\gamma \min\{R(W_k)\} R_{\min}^{\mathsf{RL}}} + \frac{4\beta^2 F_{\max}}{\alpha R_{\min}^{\mathsf{RL}}}\right)\mathbb{E}_{\xi_k}[\|P_{k+1} - P^*(W_k)\|^2] - \frac{5\beta^2 F_{\max}}{\alpha R_{\min}^{\mathsf{RL}}}\|W_k - W^*\|^2_{R(W_k)}.
\end{aligned}
$$

Notice that the $\|P_{k+1} - P^*(W_k)\|^2$ appears in the RHS above, we also need the following lemma to bound it in terms of $\|P_k - P^*(W_k)\|^2$.

Lemma 7 (Descent Lemma of P_k)**.** *Suppose Assumptions 1-2 and 4 hold. It holds for* `Tiki-Taka` *that*

$$
\begin{aligned}
&\mathbb{E}_{\xi_k}[\|P_{k+1} - P^*(W_k)\|^2] \qquad (81)\\
\le& \left(1 - \frac{\alpha\gamma\mu L}{4(\mu+L)}\right)\|P_k - P^*(W_k)\|^2 + \frac{2\alpha(\mu+L)F_{\max}\sigma^2}{\gamma\mu L}\left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|_\infty^2 + \alpha^2 F_{\max}^2\sigma^2.
\end{aligned}
$$

The proof of Lemma 7 is deferred to Section I.4. By Lemma 7, we bound the $\|P_{k+1} - P^*(W_k)\|^2$ in terms of $\|P_k - P^*(W_k)\|^2$ as

$$
\left(\frac{30\beta^2 F_{\max}^5}{\alpha\gamma \min\{R(W_k)\} R_{\min}^{\mathsf{RL}}} + \frac{4\beta^2 F_{\max}}{\alpha R_{\min}^{\mathsf{RL}}}\right)\mathbb{E}_{\xi_k}[\|P_{k+1} - P^*(W_k)\|^2] \qquad (82)
$$

$$
\begin{aligned}
&\overset{(a)}{\leq} \frac{32\beta^2 F_{\max}^5}{\alpha\gamma \min\{R(W_k)\} R_{\min}^{\texttt{RL}}} \mathbb{E}_{\xi_k}[\|P_{k+1} - P^*(W_k)\|^2] \\
&\leq \frac{32\beta^2 F_{\max}^5}{\alpha\gamma \min\{R(W_k)\} R_{\min}^{\texttt{RL}}} \left(1 - \frac{\alpha}{4}\frac{\mu L}{\gamma(\mu+L)}\right) \|P_k - P^*(W_k)\|^2 \\
&\quad + \frac{32\beta^2 F_{\max}^5}{\alpha\gamma \min\{R(W_k)\} R_{\min}^{\texttt{RL}}} \left(\frac{2\alpha(\mu+L)F_{\max}\sigma^2}{\gamma\mu L} \left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|_\infty^2 + \alpha^2 F_{\max}^2\sigma^2\right) \\
&\leq \frac{32\beta^2 F_{\max}^5}{\alpha\gamma \min\{R(W_k)\} R_{\min}^{\texttt{RL}}} \|P_k - P^*(W_k)\|^2 + O\left(\beta^2\sigma^2 \left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|_\infty^2 + \alpha\beta^2 F_{\max}^2\sigma^2\right) \\
&\overset{(b)}{\leq} \frac{32\beta^2 F_{\max}^5}{\alpha\gamma \min\{R(W_k)\} R_{\min}^{\texttt{RL}}} \|P_k - P^*(W_k)\|^2 + \alpha\sigma^2 \left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|_\infty^2 + \alpha^2 L F_{\max}^2\sigma^2
\end{aligned}
$$

where (a) assumes $\frac{4\beta^2 F_{\max}}{\alpha R_{\min}^{\texttt{RL}}} \leq \frac{2\beta^2 F_{\max}^5}{\alpha\gamma \min\{R(W_k)\} R_{\min}^{\texttt{RL}}}$ with lost of generality to keep the formulations simple since $\gamma \min\{R(W_k)\}$ is typically small; (b) holds given α and β is sufficiently small. In addition, the strong convexity of the objective (c.f. Assumption 4) implies that

$$
\begin{aligned}
\frac{\alpha R_{\min}^{\texttt{RL}}}{8F_{\max}} \left\|\nabla f(\bar{W}_k)\right\|^2 &\geq \frac{\alpha\mu^2 R_{\min}^{\texttt{RL}}}{8F_{\max}} \left\|\bar{W}_k - W^*\right\|^2 = \frac{\alpha\mu^2 R_{\min}^{\texttt{RL}}}{8F_{\max}} \|W_k + \gamma P_k - W^*\|^2 \\
&= \frac{\alpha\mu^2\gamma^2 R_{\min}^{\texttt{RL}}}{8F_{\max}} \left\|P_k - \frac{W^* - W_k}{\gamma}\right\|^2 = \frac{\alpha\mu^2\gamma^2 R_{\min}^{\texttt{RL}}}{8F_{\max}} \|P_k - P^*(W_k)\|^2.
\end{aligned} \tag{83}
$$

Substituting (82) and (83) back into (80) yields

$$
\begin{aligned}
&\mathbb{E}_{\xi_k}[\mathbb{V}_{k+1}] \\
\leq\ &\mathbb{V}_k - \frac{\alpha R_{\min}^{\texttt{RL}}}{8F_{\max}} \|\nabla f(\bar{W}_k)\|^2 + 3\alpha\sigma^2 \left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|_\infty^2 + 3\alpha^2 L F_{\max}^2\sigma^2 \\
&- \frac{\alpha}{8F_{\max}} \|\nabla f(\bar{W}_k) \odot F(P_k) + |\nabla f(\bar{W}_k)| \odot G(P_k)\|^2 \\
&- \frac{\beta^2 F_{\max}}{\alpha R_{\min}^{\texttt{RL}}} \|P^*(W_k) \odot F(W_k) - |P^*(W_k)| \odot G(W_k)\|^2 \\
&- \left(\frac{\alpha\mu^2\gamma^2 R_{\min}^{\texttt{RL}}}{8F_{\max}} - \frac{32\beta^2 F_{\max}^5}{\alpha\gamma \min\{R(W_k)\} R_{\min}^{\texttt{RL}}}\right) \|P_k - P^*(W_k)\|^2 - \frac{5\beta^2 F_{\max}}{\alpha R_{\min}^{\texttt{RL}}} \|W_k - W^*\|_{R(W_k)}^2 \\
=\ &\mathbb{V}_k - \frac{\alpha R_{\min}^{\texttt{RL}}}{8F_{\max}} \|\nabla f(\bar{W}_k)\|^2 + 3\alpha\sigma^2 \left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|_\infty^2 + 3\alpha^2 L F_{\max}^2\sigma^2 \\
&- \frac{\alpha}{8F_{\max}} \|\nabla f(\bar{W}_k) \odot F(P_k) + |\nabla f(\bar{W}_k)| \odot G(P_k)\|^2 \\
&- \frac{\alpha\mu^2\gamma^3 \min\{R(W_k)\}}{512 F_{\max}^5 R_{\min}^{\texttt{RL}}} \|P^*(W_k) \odot F(W_k) - |P^*(W_k)| \odot G(W_k)\|^2 \\
&- \frac{\alpha\mu^2\gamma^2 R_{\min}^{\texttt{RL}}}{16F_{\max}} \|P_k - P^*(W_k)\|^2 - \frac{5\alpha\mu^2\gamma^3}{512 F_{\max}^5 R_{\min}^{\texttt{RL}}} \|W_k - W^*\|_{R(W_k)}^2
\end{aligned} \tag{84}
$$

where the last step chooses the transfer learning rate by

$$
\beta = \frac{\alpha\mu\gamma^{\frac{3}{2}} \sqrt{\min\{R(W_k)\}} R_{\min}^{\texttt{RL}}}{16\sqrt{2} F_{\max}^3}. \tag{85}
$$

Rearranging inequality (79) above, we have

$$
\begin{aligned}
&\frac{\alpha}{8F_{\max}} \|\nabla f(\bar{W}_k) \odot F(P_k) + |\nabla f(\bar{W}_k)| \odot G(P_k)\|^2 + \frac{\alpha}{8F_{\max} R_{\min}^{\texttt{RL}}} \|\nabla f(\bar{W}_k)\|^2 \\
&+ \frac{\alpha\mu^2\gamma^3 \min\{R(W_k)\}}{512 F_{\max}^5 R_{\min}^{\texttt{RL}}} \|P^*(W_k) \odot F(W_k) - |P^*(W_k)| \odot G(W_k)\|^2
\end{aligned} \tag{86}
$$

$$
+\frac{5\alpha\mu^2\gamma^3\min\{R(W_k)\}}{512F_{\max}^5R_{\min}^{\mathtt{RL}}}\|W_k-W^*\|^2_{R(W_k)}+\frac{\alpha\mu^2\gamma^2R_{\min}^{\mathtt{RL}}}{16F_{\max}}\|P_k-P^*(W_k)\|^2
$$

$$
\le \mathbb{V}_k-\mathbb{E}_{\xi_k}[\mathbb{V}_{k+1}]+3\alpha^2LF_{\max}^2\sigma^2+3\alpha\sigma^2\left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|_\infty^2.
$$

Define the convergence metric $E_K^{\mathtt{RL}}$ as

$$
\begin{aligned}
E_K^{\mathtt{RL}}:=\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\Bigg[&\|\nabla f(\bar{W}_k)\odot F(P_k)+|\nabla f(\bar{W}_k)|\odot G(P_k)\|^2+\frac{1}{R_{\min}^{\mathtt{RL}}}\|\nabla f(\bar{W}_k)\|^2\\
&+\frac{\mu^2\gamma^3\min\{R(W_k)\}}{64F_{\max}^4R_{\min}^{\mathtt{RL}}}\|P^*(W_k)\odot F(W_k)-|P^*(W_k)|\odot G(W_k)\|^2\\
&+\frac{5\mu^2\gamma^3}{64F_{\max}^4R_{\min}^{\mathtt{RL}}}\|W_k-W^*\|^2_{R(W_k)}+\frac{\mu^2\gamma^2R_{\min}^{\mathtt{RL}}}{2}\|P_k-P^*(W_k)\|^2\Bigg].
\end{aligned}
\tag{87}
$$

Taking expectation over all $\xi_K,\xi_{K-1},\cdots,\xi_0$, averaging (86) over k from 0 to $K-1$, and choosing the parameter α as $\alpha=O\left(\frac{1}{F_{\max}}\sqrt{\frac{\mathbb{V}_0}{\sigma^2LK}}\right)$ deduce that

$$
\begin{aligned}
E_K^{\mathtt{RL}}&\le 8F_{\max}\left(\frac{\mathbb{V}_0-\mathbb{E}[\mathbb{V}_{k+1}]}{\alpha K}+3\alpha LF_{\max}^2\sigma^2\right)+24F_{\max}\sigma^2\times\frac{1}{K}\sum_{k=0}^{K-1}\left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|_\infty^2\\
&\le 8F_{\max}\left(\frac{\mathbb{V}_0}{\alpha K}+3\alpha LF_{\max}^2\sigma^2\right)+24F_{\max}\sigma^2\times\frac{1}{K}\sum_{k=0}^{K-1}\left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|_\infty^2\\
&=O\left(F_{\max}^2\sqrt{\frac{\mathbb{V}_0\sigma^2L}{K}}\right)+24F_{\max}\sigma^2S_K^{\mathtt{RL}}.
\end{aligned}
\tag{88}
$$

The strong convexity of the objective (Assumption 4) implies that

$$
\mathbb{V}_0=f(\bar{W}_0)-f^*+C\|W_0-W^*\|^2\le\left(1+\frac{2C}{\mu}\right)(f(W_0)-f^*).
\tag{89}
$$

Plugging it back to the above inequality, we have

$$
E_K^{\mathtt{RL}}=O\left(F_{\max}^2\sqrt{\frac{(f(W_0)-f^*)\sigma^2L}{K}}\right)+24F_{\max}\sigma^2S_K^{\mathtt{RL}}.
\tag{90}
$$

The proof is completed. □

I.2 Proof of Lemma 5: Descent of sequence $\bar{W}_k$

Lemma 5 (Descent Lemma of $\bar{W}_k$)**.** *Suppose Assumptions 1–2 hold. It holds for* `Tiki-Taka` *that*

$$
\begin{aligned}
\mathbb{E}_{\xi_k}[f(\bar{W}_{k+1})]\le f(\bar{W}_k)&-\frac{\alpha}{4F_{\max}}\|\nabla f(\bar{W}_k)\|^2_{R(P_k)}+2\alpha\sigma^2\left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|_\infty^2+2\alpha^2LF_{\max}^2\sigma^2\\
&-\frac{\alpha\gamma}{8F_{\max}}\|\nabla f(\bar{W}_k)\odot F(P_k)+|\nabla f(\bar{W}_k)|\odot G(P_k)\|^2\\
&+\frac{F_{\max}}{\alpha}\mathbb{E}_{\xi_k}\left[\|W_{k+1}-W_k\|^2_{R(P_k)^\dagger}\right]+\mathbb{E}_{\xi_k}[\|W_{k+1}-W_k\|^2].
\end{aligned}
\tag{72}
$$

Proof of Lemma 5. The L-smooth assumption (Assumption 1) implies that

$$
\mathbb{E}_{\xi_k}[f(\bar{W}_{k+1})]\le f(\bar{W}_k)+\mathbb{E}_{\xi_k}[\langle\nabla f(\bar{W}_k),\bar{W}_{k+1}-\bar{W}_k\rangle]+\frac{L}{2}\mathbb{E}_{\xi_k}[\|\bar{W}_{k+1}-\bar{W}_k\|^2]
\tag{91}
$$

$$= f(\bar{W}_k) + \gamma \underbrace{\mathbb{E}_{\xi_k}[\langle \nabla f(\bar{W}_k), P_{k+1} - P_k \rangle]}_{(a)} + \underbrace{\mathbb{E}_{\xi_k}[\langle \nabla f(\bar{W}_k), W_{k+1} - W_k \rangle]}_{(b)} + \underbrace{\frac{L}{2}\mathbb{E}_{\xi_k}[\|\bar{W}_{k+1} - \bar{W}_k\|^2]}_{(c)}.$$

Next, we will handle the each term in the RHS of (91) separately.

Bound of the second term (a). To bound term (a) in the RHS of (91), we leverage the assumption that noise has expectation 0 (Assumption 2)

$$\begin{aligned}
&\mathbb{E}_{\xi_k}[\langle \nabla f(\bar{W}_k), P_{k+1} - P_k \rangle] \qquad (92)\\
=& \alpha \mathbb{E}_{\xi_k}\left[\left\langle \nabla f(\bar{W}_k) \odot \sqrt{F(P_k)}, \frac{P_{k+1} - P_k}{\alpha\sqrt{F(P_k)}} + (\nabla f(\bar{W}_k;\xi_k) - \nabla f(\bar{W}_k)) \odot \sqrt{F(P_k)} \right\rangle\right]\\
=& -\frac{\alpha}{2}\|\nabla f(\bar{W}_k) \odot \sqrt{F(P_k)}\|^2\\
&- \frac{1}{2\alpha}\mathbb{E}_{\xi_k}\left[\left\|\frac{P_{k+1} - P_k}{\sqrt{F(P_k)}} + \alpha(\nabla f(\bar{W}_k;\xi_k) - \nabla f(\bar{W}_k)) \odot \sqrt{F(P_k)}\right\|^2\right]\\
&+ \frac{1}{2\alpha}\mathbb{E}_{\xi_k}\left[\left\|\frac{P_{k+1} - P_k}{\sqrt{F(P_k)}} + \alpha\nabla f(\bar{W}_k;\xi_k) \odot \sqrt{F(P_k)}\right\|^2\right].
\end{aligned}$$

The second term in the RHS of (92) can be bounded by

$$\begin{aligned}
&\frac{1}{2\alpha}\mathbb{E}_{\xi_k}\left[\left\|\frac{P_{k+1} - P_k}{\sqrt{F(P_k)}} + \alpha(\nabla f(\bar{W}_k;\xi_k) - \nabla f(\bar{W}_k)) \odot \sqrt{F(P_k)}\right\|^2\right] \qquad (93)\\
=& \frac{1}{2\alpha}\mathbb{E}_{\xi_k}\left[\left\|\frac{P_{k+1} - P_k + \alpha(\nabla f(\bar{W}_k;\xi_k) - \nabla f(\bar{W}_k)) \odot F(P_k)}{\sqrt{F(P_k)}}\right\|^2\right]\\
\geq& \frac{1}{2\alpha F_{\max}}\mathbb{E}_{\xi_k}\left[\|P_{k+1} - P_k + \alpha(\nabla f(\bar{W}_k;\xi_k) - \nabla f(\bar{W}_k)) \odot F(P_k)\|^2\right].
\end{aligned}$$

The third term in the RHS of (92) can be bounded by variance decomposition and bounded variance assumption (Assumption 2)

$$\begin{aligned}
&\frac{1}{2\alpha}\mathbb{E}_{\xi_k}\left[\left\|\frac{P_{k+1} - P_k}{\sqrt{F(P_k)}} + \alpha\nabla f(\bar{W}_k;\xi_k) \odot \sqrt{F(P_k)}\right\|^2\right] \qquad (94)\\
\leq& \frac{\alpha}{2}\mathbb{E}_{\xi_k}\left[\left\||\nabla f(\bar{W}_k;\xi_k)| \odot \frac{G(P_k)}{\sqrt{F(P_k)}}\right\|^2\right]\\
\leq& \frac{\alpha}{2}\left\||\nabla f(\bar{W}_k)| \odot \frac{G(P_k)}{\sqrt{F(P_k)}}\right\|^2 + \frac{\alpha\sigma^2}{2}\left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|_\infty^2.
\end{aligned}$$

Notice that the first term in the RHS of (92) and the second term in the RHS of (94) can be bounded together

$$\begin{aligned}
&-\frac{\alpha}{2}\|\nabla f(\bar{W}_k) \odot \sqrt{F(P_k)}\|^2 + \frac{\alpha}{2}\left\||\nabla f(\bar{W}_k)| \odot \frac{G(P_k)}{\sqrt{F(P_k)}}\right\|^2 \qquad (95)\\
=& -\frac{\alpha}{2}\sum_{d\in[D]}\left([\nabla f(\bar{W}_k)]_d^2\left([F(P_k)]_d - \frac{[G(P_k)]_d^2}{[F(P_k)]_d}\right)\right)\\
=& -\frac{\alpha}{2}\sum_{d\in[D]}\left([\nabla f(\bar{W}_k)]_d^2\left(\frac{[F(P_k)]_d^2 - [G(P_k)]_d^2}{[F(P_k)]_d}\right)\right)
\end{aligned}$$

$$\leq -\frac{\alpha}{2F_{\max}}\sum_{d\in[D]}\left([\nabla f(\bar{W}_k)]_d^2\left([F(P_k)]_d^2-[G(P_k)]_d^2\right)\right)$$
$$= -\frac{\alpha}{2F_{\max}}\|\nabla f(\bar{W}_k)\|^2_{R(P_k)} \leq 0.$$

Plugging (93) to (95) into (92), we bound the term (a) by

$$\begin{aligned}
&\mathbb{E}_{\xi_k}[\langle \nabla f(\bar{W}_k), P_{k+1}-P_k\rangle] \\
\leq & -\frac{\alpha}{2F_{\max}}\|\nabla f(\bar{W}_k)\|^2_{R(P_k)} + \frac{\alpha\sigma^2}{2}\left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|^2_\infty \\
& -\frac{1}{2\alpha F_{\max}}\mathbb{E}_{\xi_k}\left[\left\|P_{k+1}-P_k+\alpha(\nabla f(\bar{W}_k;\xi_k)-\nabla f(\bar{W}_k))\odot F(P_k)\right\|^2\right].
\end{aligned} \tag{96}$$

Bound of the third term (b). By Young's inequality, we have

$$\mathbb{E}_{\xi_k}[\langle \nabla f(\bar{W}_k), W_{k+1}-W_k\rangle] \leq \frac{\alpha}{4F_{\max}}\|\nabla f(\bar{W}_k)\|^2_{R(P_k)} + \frac{F_{\max}}{\alpha}\mathbb{E}_{\xi_k}[\|W_{k+1}-W_k\|^2_{R(P_k)^\dagger}]. \tag{97}$$

Bound of the third term (c). Repeatedly applying inequality $\|U+V\|^2 \leq 2\|U\|^2+2\|V\|^2$ for any $U, V\in\mathbb{R}^D$, we have

$$\begin{aligned}
&\frac{L}{2}\mathbb{E}_{\xi_k}[\|\bar{W}_{k+1}-\bar{W}_k\|^2] \\
\leq & L\mathbb{E}_{\xi_k}[\|W_{k+1}-W_k\|^2] + L\mathbb{E}_{\xi_k}[\|P_{k+1}-P_k\|^2] \\
\leq & L\mathbb{E}_{\xi_k}[\|W_{k+1}-W_k\|^2] + 2L\mathbb{E}_{\xi_k}\left[\left\|P_{k+1}-P_k+\alpha(\nabla f(\bar{W}_k;\xi_k)-\nabla f(\bar{W}_k))\odot F(P_k)\right\|^2\right] \\
& + 2\alpha^2 L\mathbb{E}_{\xi_k}\left[\left\|(\nabla f(\bar{W}_k;\xi_k)-\nabla f(\bar{W}_k))\odot F(P_k)\right\|^2\right] \\
\leq & \mathbb{E}_{\xi_k}[\|W_{k+1}-W_k\|^2] + 2L\mathbb{E}_{\xi_k}\left[\left\|P_{k+1}-P_k+\alpha(\nabla f(\bar{W}_k;\xi_k)-\nabla f(\bar{W}_k))\odot F(P_k)\right\|^2\right] \\
& + 2\alpha^2 LF_{\max}^2\sigma^2
\end{aligned} \tag{98}$$

where the last inequality comes from the bounded variance assumption (Assumption 2)

$$\begin{aligned}
&2\alpha^2 L\mathbb{E}_{\xi_k}\left[\left\|(\nabla f(\bar{W}_k;\xi_k)-\nabla f(\bar{W}_k))\odot F(P_k)\right\|^2\right] \\
\leq & 2\alpha^2 LF_{\max}^2\mathbb{E}_{\xi_k}\left[\left\|\nabla f(\bar{W}_k;\xi_k)-\nabla f(\bar{W}_k)\right\|^2\right] \\
\leq & 2\alpha^2 LF_{\max}^2\sigma^2.
\end{aligned} \tag{99}$$

Combination of the upper bound (a), (b), and (c). Plugging (96), (97), (98) into (91), we derive

$$\begin{aligned}
\mathbb{E}_{\xi_k}[f(\bar{W}_{k+1})] \leq & f(\bar{W}_k) - \frac{\alpha}{4F_{\max}}\|\nabla f(\bar{W}_k)\|^2_{R(P_k)} + \frac{\alpha\sigma^2}{2}\left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|^2_\infty \\
& -\left(\frac{1}{2\alpha F_{\max}}-2L\right)\mathbb{E}_{\xi_k}\left[\left\|P_{k+1}-P_k+\alpha(\nabla f(\bar{W}_k;\xi_k)-\nabla f(\bar{W}_k))\odot F(P_k)\right\|^2\right] \\
& + \frac{F_{\max}}{\alpha}\mathbb{E}_{\xi_k}\left[\|W_{k+1}-W_k\|^2_{R(P_k)^\dagger}\right] + \mathbb{E}_{\xi_k}[\|W_{k+1}-W_k\|^2] + 2\alpha^2 LF_{\max}^2\sigma^2.
\end{aligned} \tag{100}$$

We bound the fourth term in the RHS of (100) using the similar technique as in (66)

$$\begin{aligned}
&\mathbb{E}_{\xi_k}\left[\left\|P_{k+1}-P_k+\alpha(\nabla f(\bar{W}_k;\xi_k)-\nabla f(\bar{W}_k))\odot F(P_k)\right\|^2\right] \\
\geq & \frac{\alpha^2}{2}\|\nabla f(\bar{W}_k)\odot F(P_k)+|\nabla f(\bar{W}_k)|\odot G(P_k)\|^2 - \alpha^2 F_{\max}\sigma^2\left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|^2_\infty.
\end{aligned} \tag{101}$$

Inequality (101) as well as the learning rate rule $\alpha \leq \frac{1}{4LF_{\max}}$ leads to the conclusion

$$\mathbb{E}_{\xi_k}[f(\bar{W}_{k+1})] \leq f(\bar{W}_k) - \frac{\alpha}{4F_{\max}}\|\nabla f(\bar{W}_k)\|^2_{R(P_k)} + 2\alpha\sigma^2 \left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|^2_\infty + 2\alpha^2 LF_{\max}^2\sigma^2 \tag{102}$$
$$- \frac{\alpha\gamma}{8F_{\max}}\|\nabla f(\bar{W}_k) \odot F(P_k) + |\nabla f(\bar{W}_k)| \odot G(P_k)\|^2$$
$$+ \frac{F_{\max}}{\alpha}\mathbb{E}_{\xi_k}\left[\|W_{k+1} - W_k\|^2_{R(P_k)^\dagger}\right] + \mathbb{E}_{\xi_k}[\|W_{k+1} - W_k\|^2].$$

The proof is completed. □

I.3 Proof of Lemma 6: Descent of sequence W_k

Lemma 6 (Descent Lemma of W_k)**.** *It holds for* `Tiki-Taka` *that*

$$\|W_{k+1} - W^*\|^2 \leq \|W_k - W^*\|^2 - \frac{\beta}{2\gamma F_{\max}}\|W_k - W^*\|^2_{R(W_k)} \tag{73}$$
$$- \frac{\beta\gamma}{2F_{\max}}\|P^*(W_k) \odot F(W_k) - |P^*(W_k)| \odot G(W_k)\|^2$$
$$+ \frac{2\beta F_{\max}^3}{\gamma}\|P_{k+1} - P^*(W_k)\|^2_{R(W_k)^\dagger} + 2\beta^2\|P_{k+1} - P^*(W_k)\|^2.$$

Proof of Lemma 6. The proof begins from manipulating the norm $\|W_{k+1} - W^*\|^2$

$$\|W_{k+1} - W^*\|^2 = \|W_k - W^*\|^2 + 2\langle W_k - W^*, W_{k+1} - W_k\rangle + \|W_{k+1} - W_k\|^2. \tag{103}$$

Revisit that we interpret P_k as the residual of W_k, namely $P^*(W) := \frac{W^* - W}{\gamma}$. Therefore, we bound the second term in the RHS of (103) by

$$2\langle W_k - W^*, W_{k+1} - W_k\rangle \tag{104}$$
$$= 2\langle W_k - W^*, \beta P_{k+1} \odot F(W_k) - \beta|P_{k+1}| \odot G(W_k)\rangle$$
$$= 2\beta\langle W_k - W^*, P^*(W_k) \odot F(W_k) - |P^*(W_k)| \odot G(W_k)\rangle$$
$$+ 2\beta\langle W_k - W^*, P_{k+1} \odot F(W_k) - |P_{k+1}| \odot G(W_k) - (P^*(W_k) \odot F(W_k) - |P^*(W_k)| \odot G(W_k))\rangle.$$

The first term in the RHS of (104) is bounded by

$$2\beta\langle W_k - W^*, P^*(W_k) \odot F(W_k) - |P^*(W_k)| \odot G(W_k)\rangle \tag{105}$$
$$= 2\beta\left\langle (W_k - W^*) \odot \sqrt{F(W_k)}, \frac{P^*(W_k) \odot F(W_k) - |P^*(W_k)| \odot G(W_k)}{\sqrt{F(W_k)}}\right\rangle$$
$$= -\frac{2\beta}{\gamma}\left\langle (W_k - W^*) \odot \sqrt{F(W_k)}, (W_k - W^*) \odot \sqrt{F(W_k)}\right\rangle$$
$$+ \frac{2\beta}{\gamma}\left\langle (W_k - W^*) \odot \sqrt{F(W_k)}, |W_k - W^*| \odot \frac{G(W_k)}{\sqrt{F(W_k)}}\right\rangle$$
$$\overset{(a)}{=} -\frac{\beta}{\gamma}\|(W_k - W^*) \odot \sqrt{F(W_k)}\|^2 + \frac{\beta}{\gamma}\left\||W_k - W^*| \odot \frac{G(W_k)}{\sqrt{F(W_k)}}\right\|^2$$
$$- \frac{\beta}{\gamma}\left\|(W_k - W^*) \odot \sqrt{F(W_k)} + |W_k - W^*| \odot \frac{G(W_k)}{\sqrt{F(W_k)}}\right\|^2$$
$$\overset{(b)}{\leq} -\frac{\beta}{\gamma F_{\max}}\|W_k - W^*\|^2_{R(W_k)} - \frac{\beta}{\gamma}\left\|(W_k - W^*) \odot \sqrt{F(W_k)} + |W_k - W^*| \odot \frac{G(W_k)}{\sqrt{F(W_k)}}\right\|^2$$
$$\overset{(c)}{\leq} -\frac{\beta}{\gamma F_{\max}}\|W_k - W^*\|^2_{R(W_k)} - \frac{\beta\gamma}{F_{\max}}\|P^*(W_k) \odot F(W_k) - |P^*(W_k)| \odot G(W_k)\|^2$$

where (a) leverages the equality $2\langle U, V\rangle = \|U\|^2 - \|V\|^2 - \|U-V\|^2$ for any $U, V \in \mathbb{R}^D$, (b) is achieved by similar technique (62), and (c) comes from

$$
\begin{aligned}
& -\frac{\beta}{\gamma}\left\|(W_k - W^*) \odot \sqrt{F(W_k)} + |W_k - W^*| \odot \frac{G(W_k)}{\sqrt{F(W_k)}}\right\|^2 \\
=& -\beta\gamma\left\|\frac{1}{\sqrt{F(W_k)}} \odot \left(\frac{W_k - W^*}{\gamma} \odot F(W_k) + \left|\frac{W_k - W^*}{\gamma}\right| \odot G(W_k)\right)\right\|^2 \\
\le& -\frac{\beta\gamma}{F_{\max}} \left\|P^*(W_k) \odot F(W_k) - |P^*(W_k)| \odot G(W_k)\right\|^2 .
\end{aligned} \tag{106}
$$

The second term in the RHS of (104) is bounded by the Lipschitz continuity of analog update (c.f. Lemma 3)

$$
\begin{aligned}
& \frac{2\beta}{\gamma}\left\langle W_k - W^*, P_{k+1} \odot F(W_k) - |P_{k+1}| \odot G(W_k) - (P^*(W_k) \odot F(W_k) - |P^*(W_k)| \odot G(W_k))\right\rangle \\
\le& \frac{\beta}{2\gamma F_{\max}} \|W_k - W^*\|^2_{R(W_k)} + \frac{2\beta F_{\max}}{\gamma} \\
& \times \|P_{k+1} \odot F(W_k) - |P_{k+1}| \odot G(W_k) - (P^*(W_k) \odot F(W_k) - |P^*(W_k)| \odot G(W_k))\|^2_{R(W_k)^\dagger} \\
\le& \frac{\beta}{2\gamma F_{\max}} \|W_k - W^*\|^2_{R(W_k)} + \frac{2\beta F_{\max}^3}{\gamma} \|P_{k+1} - P^*(W_k)\|^2_{R(W_k)^\dagger} .
\end{aligned} \tag{107}
$$

Substituting (105) and (107) into (104), we bound the second term in the RHS of (103) by

$$
\begin{aligned}
& 2\langle W_k - W^*, W_{k+1} - W_k\rangle \\
\le& -\frac{\beta}{\gamma F_{\max}} \|W_k - W^*\|^2_{R(W_k)} - \frac{\beta\gamma}{F_{\max}} \|P^*(W_k) \odot F(W_k) - |P^*(W_k)| \odot G(W_k)\|^2 \\
& + \frac{2\beta F_{\max}^3}{\gamma} \|P_{k+1} - P^*(W_k)\|^2_{R(W_k)^\dagger} .
\end{aligned} \tag{108}
$$

The third term in the RHS of (103) is bounded by the Lipschitz continuity of analog update (c.f. Lemma 3)

$$
\begin{aligned}
& \|W_{k+1} - W_k\|^2 = \beta^2 \|P_{k+1} \odot F(W_k) - |P_{k+1}| \odot G(W_k)\|^2 \\
\le& 2\beta^2 \|P^*(W_k) \odot F(W_k) - |P^*(W_k)| \odot G(W_k)\|^2 \\
& + 2\beta^2 \|P_{k+1} \odot F(W_k) - |P_{k+1}| \odot G(W_k) - (P^*(W_k) \odot F(W_k) - |P^*(W_k)| \odot G(W_k))\|^2 \\
\le& 2\beta^2 \|P^*(W_k) \odot F(W_k) - |P^*(W_k)| \odot G(W_k)\|^2 + 2\beta^2 \|P_{k+1} - P^*(W_k)\|^2 .
\end{aligned} \tag{109}
$$

Plugging (108) and (109) into (103) yields

$$
\begin{aligned}
\|W_{k+1} - W^*\|^2 \le& \|W_k - W^*\|^2 - \frac{\beta}{2\gamma F_{\max}} \|W_k - W^*\|^2_{R(W_k)} \\
& - \left(\frac{\beta\gamma}{F_{\max}} - 2\beta^2\right) \|P^*(W_k) \odot F(W_k) - |P^*(W_k)| \odot G(W_k)\|^2 \\
& + \frac{2\beta F_{\max}^3}{\gamma} \|P_{k+1} - P^*(W_k)\|^2_{R(W_k)^\dagger} + 2\beta^2 \|P_{k+1} - P^*(W_k)\|^2 .
\end{aligned} \tag{110}
$$

Notice the learning rate β is chosen as $\beta \le \frac{\gamma}{2F_{\max}}$, we have

$$
\begin{aligned}
\|W_{k+1} - W^*\|^2 \le& \|W_k - W^*\|^2 - \frac{\beta}{2\gamma F_{\max}} \|W_k - W^*\|^2_{R(W_k)} \\
& - \frac{\beta\gamma}{2F_{\max}} \|P^*(W_k) \odot F(W_k) - |P^*(W_k)| \odot G(W_k)\|^2 \\
& + \frac{2\beta F_{\max}^3}{\gamma} \|P_{k+1} - P^*(W_k)\|^2_{R(W_k)^\dagger} + 2\beta^2 \|P_{k+1} - P^*(W_k)\|^2
\end{aligned} \tag{111}
$$

which completes the proof. □

I.4 Proof of Lemma 7: Descent of sequence P_k

Lemma 7 (Descent Lemma of P_k)**.** *Suppose Assumptions 1-2 and 4 hold. It holds for* `Tiki-Taka` *that*

$$\begin{aligned}
&\mathbb{E}_{\xi_k}[\|P_{k+1} - P^*(W_k)\|^2] \qquad (81)\\
\le\ &\left(1 - \frac{\alpha\gamma\mu L}{4(\mu+L)}\right)\|P_k - P^*(W_k)\|^2 + \frac{2\alpha(\mu+L)F_{\max}\sigma^2}{\gamma\mu L}\left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|_\infty^2 + \alpha^2 F_{\max}^2\sigma^2.
\end{aligned}$$

Proof of Lemma 7. The proof begins from manipulating the norm $\|P_{k+1} - P^*(W_k)\|^2$

$$\|P_{k+1} - P^*(W_k)\|^2 = \|P_k - P^*(W_k)\|^2 + 2\langle P_k - P^*(W_k), P_{k+1} - P_k\rangle + \|P_{k+1} - P_k\|^2. \tag{112}$$

To bound the second term, we need the following equality.

$$\begin{aligned}
&2\mathbb{E}_{\xi_k}[\langle P_k - P^*(W_k), P_{k+1} - P_k\rangle] \qquad (113)\\
=\ & -2\alpha\mathbb{E}_{\xi_k}[\langle P_k - P^*(W_k), \nabla f(\bar{W}_k;\xi_k)\odot F(P_k) - |\nabla f(\bar{W}_k;\xi_k)|\odot G(P_k)\rangle]\\
=\ & -2\alpha\mathbb{E}_{\xi_k}[\langle P_k - P^*(W_k), \nabla f(\bar{W}_k;\xi_k)\odot F(P_k)\rangle]\\
&+2\alpha\mathbb{E}_{\xi_k}[\langle P_k - P^*(W_k), |\nabla f(\bar{W}_k;\xi_k)|\odot G(P_k)\rangle]\\
=\ & -2\alpha\langle P_k - P^*(W_k), \nabla f(\bar{W}_k)\odot F(P_k)\rangle + 2\alpha\langle P_k - P^*(W_k), |\nabla f(\bar{W}_k)|\odot G(P_k)\rangle\\
&+2\alpha\mathbb{E}_{\xi_k}[\langle P_k - P^*(W_k), (|\nabla f(\bar{W}_k)| - |\nabla f(\bar{W}_k;\xi_k)|)\odot G(P_k)\rangle]\\
=\ & -2\alpha\underbrace{\langle P_k - P^*(W_k), \nabla f(\bar{W}_k)\odot F(P_k) - |\nabla f(\bar{W}_k)|\odot G(P_k)\rangle}_{(T1)}\\
&+2\alpha\underbrace{\mathbb{E}_{\xi_k}[\langle P_k - P^*(W_k), (|\nabla f(\bar{W}_k)| - |\nabla f(\bar{W}_k;\xi_k)|)\odot G(P_k)\rangle]}_{(T2)}
\end{aligned}$$

Upper bound of the first term $(T1)$**.** With Lemma 4, the second term in the RHS of (112) can be bounded by

$$\begin{aligned}
&-2\alpha\langle P_k - P^*(W_k), \nabla f(\bar{W}_k)\odot F(P_k) - |\nabla f(\bar{W}_k)|\odot G(P_k)\rangle \qquad (114)\\
=\ & -2\alpha\langle P_k - P^*(W_k), \nabla f(\bar{W}_k)\odot q_s(P_k)\rangle\\
\le\ & -2\alpha C_{k,+}\langle P_k - P^*(W_k), \nabla f(\bar{W}_k)\rangle + 2\alpha C_{k,-}\langle |P_k - P^*(W_k)|, |\nabla f(\bar{W}_k)|\rangle
\end{aligned}$$

where $C_{k,+}$ and $C_{k,-}$ are defined by

$$C_{k,+} := \frac{1}{2}\left(\max_{i\in[D]}\{q_s([P_k]_i)\} + \min_{i\in[D]}\{q_s([P_k]_i)\}\right), \tag{115}$$

$$C_{k,-} := \frac{1}{2}\left(\max_{i\in[D]}\{q_s([P_k]_i)\} - \min_{i\in[D]}\{q_s([P_k]_i)\}\right). \tag{116}$$

In the inequality above, the first term can be bounded by the strong convexity of f. Let $\varphi(P) := f(W+\gamma P)$ which is $\gamma^2 L$-smooth and $\gamma^2\mu$-strongly convex. It can be verified that $\varphi(P)$ has gradient $\nabla\varphi(P_k) = \nabla_{P_k} f(W_k + \gamma P_k) = \gamma\nabla f(\bar{W}_k)$ and optimal point $P^*(W)$. Leveraging Theorem 2.1.9 in [83], we have

$$\begin{aligned}
&\langle\nabla f(\bar{W}_k), P_k - P^*(W_k)\rangle = \frac{1}{\gamma}\langle\nabla\varphi(P_k), P_k - P^*(W_k)\rangle \qquad (117)\\
\ge\ & \frac{1}{\gamma}\left(\frac{\gamma^2\mu\cdot\gamma^2 L}{\gamma^2\mu + \gamma^2 L}\|P_k - P^*(W_k)\|^2 + \frac{1}{\gamma^2\mu + \gamma^2 L}\|\nabla\varphi(P_k)\|^2\right)\\
=\ & \frac{\gamma\mu L}{\mu+L}\|P_k - P^*(W_k)\|^2 + \frac{1}{\gamma(\mu+L)}\|\nabla f(\bar{W}_k)\|^2.
\end{aligned}$$

The second term in the RHS of (114) can be bounded by Young's inequality $2\langle x, y\rangle \le u\|x\|^2 + \frac{1}{u}\|y\|^2$ with any $u > 0$ and $x, y \in \mathbb{R}^D$

$$2\alpha C_{k,-}\langle |P_k - P^*(W_k)|, |\nabla f(\bar{W}_k)|\rangle \tag{118}$$

$$\leq \frac{\alpha C_{k,-}^2 \gamma(\mu+L)}{C_{k,+}} \|P_k - P^*(W_k)\|^2 + \frac{\alpha C_{k,+}}{\gamma(\mu+L)} \|\nabla f(\bar{W}_k)\|^2$$

where u is chosen to align the coefficient in front of $\|\nabla f(\bar{W}_k)\|^2$. Therefore, $(T1)$ in (114) becomes

$$\begin{aligned}
& -2\alpha \left\langle P_k - P^*(W_k), \nabla f(\bar{W}_k) \odot F(P_k) - |\nabla f(\bar{W}_k)| \odot G(P_k) \right\rangle \\
\leq & -\left(\frac{2\alpha\gamma\mu L C_{k,+}}{\mu+L} - \frac{\alpha C_{k,-}^2 \gamma(\mu+L)}{C_{k,+}} \right) \|P_k - P^*(W_k)\|^2 - \frac{\alpha C_{k,+}}{\gamma(\mu+L)} \|\nabla f(\bar{W}_k)\|^2.
\end{aligned} \tag{119}$$

Upper bound of the second term $(T2)$**.** Leveraging the Young's inequality $2\langle x, y\rangle \leq u\|x\|^2 + \frac{1}{u}\|y\|^2$ with any $u > 0$ and $x, y \in \mathbb{R}^D$, we have

$$\begin{aligned}
& 2\alpha \mathbb{E}_{\xi_k}[\langle P_k - P^*(W_k), (|\nabla f(\bar{W}_k)| - |\nabla f(\bar{W}_k; \xi_k)|) \odot G(P_k)\rangle] \\
= & 2\alpha \mathbb{E}_{\xi_k}\left[\left\langle (P_k - P^*(W_k)) \odot \sqrt{F(P_k)}, (|\nabla f(\bar{W}_k)| - |\nabla f(\bar{W}_k; \xi_k)|) \odot \frac{G(P_k)}{\sqrt{F(P_k)}} \right\rangle\right] \\
\overset{(a)}{\leq} & \frac{\alpha\gamma\mu L C_{k,+}}{(\mu+L)F_{\max}} \|(P_k - P^*(W_k)) \odot \sqrt{F(P_k)}\|^2 \\
& + \frac{\alpha(\mu+L)F_{\max}}{\gamma\mu L C_{k,+}} \mathbb{E}_{\xi_k}\left[\left\| (|\nabla f(\bar{W}_k)| - |\nabla f(\bar{W}_k; \xi_k)|) \odot \frac{G(P_k)}{\sqrt{F(P_k)}} \right\|^2\right] \\
\overset{(b)}{\leq} & \frac{\alpha\gamma\mu L C_{k,+}}{(\mu+L)F_{\max}} \|(P_k - P^*(W_k)) \odot \sqrt{F(P_k)}\|^2 \\
& + \frac{\alpha(\mu+L)F_{\max}}{\gamma\mu L C_{k,+}} \mathbb{E}_{\xi_k}\left[\left\| (|\nabla f(\bar{W}_k) - \nabla f(\bar{W}_k; \xi_k)|) \odot \frac{G(P_k)}{\sqrt{F(P_k)}} \right\|^2\right] \\
\overset{(c)}{\leq} & \frac{\alpha\gamma\mu L C_{k,+}}{(\mu+L)F_{\max}} \|(P_k - P^*(W_k)) \odot \sqrt{F(P_k)}\|^2 + \frac{\alpha(\mu+L)F_{\max}\sigma^2}{\gamma\mu L C_{k,+}} \left\| \frac{G(P_k)}{\sqrt{F(P_k)}} \right\|_\infty^2 \\
\overset{(d)}{\leq} & \frac{\alpha\gamma\mu L C_{k,+}}{\mu+L} \|P_k - P^*(W_k)\|^2 + \frac{\alpha(\mu+L)F_{\max}\sigma^2}{\gamma\mu L C_{k,+}} \left\| \frac{G(P_k)}{\sqrt{F(P_k)}} \right\|_\infty^2
\end{aligned} \tag{120}$$

where (a) choose $u > 0$ to align the coefficient in front of $\|P_k - P^*(W_k)\|^2$ in the RHS of (119), (b) applies $||x| - |y|| \leq |x - y|$ for any $x, y \in \mathbb{R}$, (c) uses the bounded variance assumption (c.f. Assumption 2), and (d) leverages the fact that $F(P_k)$ is bounded by $F_{\max}$ element-wise.

Combining the upper bound of $(T1)$ and $(T2)$, we bound (113) by

$$\begin{aligned}
& 2\mathbb{E}_{\xi_k}[\langle P_k - P^*(W_k), P_{k+1} - P_k\rangle] \\
\leq & -\left(\frac{\alpha\gamma\mu L C_{k,+}}{\mu+L} - \frac{\alpha C_{k,-}^2 \gamma(\mu+L)}{C_{k,+}} \right) \|P_k - P^*(W_k)\|^2 \\
& - \frac{\alpha C_{k,+}}{\gamma(\mu+L)} \|\nabla f(\bar{W}_k)\|^2 + \frac{\alpha(\mu+L)F_{\max}\sigma^2}{\gamma\mu L C_{k,+}} \left\| \frac{G(P_k)}{\sqrt{F(P_k)}} \right\|_\infty^2 \\
\leq & -\frac{\alpha\gamma\mu L C_{k,+}}{2(\mu+L)} \|P_k - P^*(W_k)\|^2 - \frac{\alpha C_{k,+}}{\gamma(\mu+L)} \|\nabla f(\bar{W}_k)\|^2 + \frac{\alpha(\mu+L)F_{\max}\sigma^2}{\gamma\mu L C_{k,+}} \left\| \frac{G(P_k)}{\sqrt{F(P_k)}} \right\|_\infty^2
\end{aligned} \tag{121}$$

where the last inequality holds when γ is sufficiently large, P_k as well as $C_{k,-}$ are sufficiently closed to 0, and the following inequality holds

$$(\mu+L)\frac{C_{k,-}^2}{C_{k,+}^2} \leq \frac{\mu L}{2(\mu+L)}. \tag{122}$$

Furthermore, the last term in the RHS of (112) can be bounded by the Lipschitz continuity of analog update (c.f. Lemma 3) and the bounded variance assumption (c.f. Assumption 2)

$$\begin{aligned}
\mathbb{E}_{\xi_k}[\|P_{k+1}-P_k\|^2] &= \mathbb{E}_{\xi_k}[\|\alpha\nabla f(\bar{W}_k;\xi_k)\odot F(P_k)-\alpha|\nabla f(\bar{W}_k;\xi_k)|\odot G(P_k)\|^2] \\
&\le \alpha^2 F_{\max}^2\mathbb{E}_{\xi_k}[\|\nabla f(\bar{W}_k;\xi_k)\|^2] \\
&= \alpha^2 F_{\max}^2\|\nabla f(\bar{W}_k)\|^2+\alpha^2F_{\max}^2\sigma^2 \\
&\le \frac{\alpha C_{k,+}}{\gamma(\mu+L)}\|\nabla f(\bar{W}_k)\|^2+\alpha^2F_{\max}^2\sigma^2
\end{aligned} \tag{123}$$

where the last inequality holds if α is sufficiently small.

Plugging inequality (121) and (123) above into (112) yields

$$\begin{aligned}
&\mathbb{E}_{\xi_k}[\|P_{k+1}-P^*(W_k)\|^2] \\
\le& \left(1-\frac{\alpha\gamma\mu LC_{k,+}}{2(\mu+L)}\right)\|P_k-P^*(W_k)\|^2+\frac{\alpha(\mu+L)F_{\max}\sigma^2}{\gamma\mu LC_{k,+}}\left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|_\infty^2+\alpha^2F_{\max}^2\sigma^2.
\end{aligned} \tag{124}$$

By definition of $C_{k,+}$, when the saturation degree of P_k is properly limited, we have $C_{k,+}\ge\frac{1}{2}$. Therefore, we have

$$\begin{aligned}
&\mathbb{E}_{\xi_k}[\|P_{k+1}-P^*(W_k)\|^2] \\
\le& \left(1-\frac{\alpha\gamma\mu L}{4(\mu+L)}\right)\|P_k-P^*(W_k)\|^2+\frac{2\alpha(\mu+L)F_{\max}\sigma^2}{\gamma\mu L}\left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|_\infty^2+\alpha^2F_{\max}^2\sigma^2
\end{aligned} \tag{125}$$

which completes the proof. □

I.5 Proof of Corollary 1: Exact convergence of `Tiki-Taka`

Corollary 1 (Exact convergence of `Residual Learning`)**.** *Under Assumption 5 and the conditions in Theorem 3, if $\gamma\ge\Omega(R_{\min}^{-1/5})$, it holds that $E_K^{\mathtt{RL}}\le O\left(\sqrt{\sigma^2L/K}\right)$.*

Proof of Corollary 1. From Theorem 3, we have

$$\|\nabla f(\bar{W}_k)\|^2\le O(E_K^{\mathtt{RL}})\le O\left(F_{\max}^2\sqrt{\frac{(f(W_0)-f^*)\sigma^2L}{K}}\right)+24F_{\max}\sigma^2S_K^{\mathtt{RL}}. \tag{126}$$

Under the zero-shift assumption (Assumption 5) and the Lipschitz continuity of the response functions, it holds directly that

$$\left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|_\infty^2\le\left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|^2=\left\|\frac{G(P_k)}{\sqrt{F(P_k)}}-\frac{G(0)}{\sqrt{F(0)}}\right\|^2\le L_S^2\|P_k\|^2 \tag{127}$$

where $L_S\ge0$ is a constant. Using $\|U+V\|^2\le2\|U\|^2+2\|V\|^2$ for any $U,V\in\mathbb{R}^D$, we have

$$\|P_k\|^2\le2\|P_k-P^*(W_k)\|^2+2\|P^*(W_k)\|^2=2\|P_k-P^*(W_k)\|^2+\frac{2}{\gamma^2}\|W_k-W^*\|^2 \tag{128}$$

where the last inequality comes from the definition of $P^*(W_k)$, as well as the definition of $P^*(W)$. Recall that convergence metric $E_K^{\mathtt{RL}}$ defined in (87) is in the order of

$$\begin{aligned}
E_K^{\mathtt{RL}} &\ge\Omega\left(\gamma^3\|W_k-W^*\|^2_{R(W_k)}+\gamma^2\|P_k-P^*(W_k)\|^2\right) \\
&\ge\Omega\left(\min\{R(W_k)\}\gamma^3\|W_k-W^*\|^2+\gamma^2\|P_k-P^*(W_k)\|^2\right) \\
&\ge\Omega\left(\frac{1}{\gamma^2}\|W_k-W^*\|^2+\gamma^2\|P_k-P^*(W_k)\|^2\right).
\end{aligned} \tag{129}$$

Therefore, we have

$$S_K^{\mathtt{RL}} = \frac{1}{K}\sum_{k=0}^{K}\left\|\frac{G(P_k)}{\sqrt{F(P_k)}}\right\|_\infty^2 \leq \frac{1}{K}\sum_{k=0}^{K}\left(2\|P_k - P^*(W_k)\|^2 + \frac{2}{\gamma^2}\|W_k - W^*\|^2\right) \leq O(E_K^{\mathtt{RL}}) \tag{130}$$

where the last inequality holds if γ is sufficiently large. Considering that, $E_K^{\mathtt{RL}} - S_K^{\mathtt{RL}} \geq \Omega(E_K^{\mathtt{RL}}) \geq 0$ and the conclusion is reached directly from Theorem 3. □

J Proof of Theorem 6: Convergence of `Analog GD`

In Section 3.2, we showed that `Analog SGD` converges to a critical point inexactly with asymptotic error proportional to the noise variance σ^2. Intuitively, without the effect of noise, `Analog GD` converges to the critical point. Define the convergence metric by

$$E_K^{\mathtt{AGD}} := \frac{1}{K}\sum_{k=0}^{K-1}\left(\|\nabla f(W_k)\odot F(W_k) - |\nabla f(W_k)|\odot G(W_k)\|^2 + \|\nabla f(W_k)\|^2_{R(W_k)}\right). \tag{131}$$

The convergence is guaranteed by the following theorem.

Theorem 6 (Convergence of `Analog GD`)**.** *Under Assumption 1–2, it holds that*

$$E_K^{\mathtt{AGD}} \leq \frac{8L(f(W_0) - f^*)F_{\max}^2}{K}. \tag{132}$$

Further, if $R_{\min}^{\mathtt{ASGD}} := \min_{k\in[K]}\min\{Q_+(W_k)Q_-(W_k)\} > 0$*, it holds that*

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla f(W_k)\|^2 \leq \frac{2L(f(W_0) - f^*)F_{\max}^2}{KR_{\min}^{\mathtt{ASGD}}}. \tag{133}$$

Proof of Theorem 6. The L-smooth assumption (Assumption 1) implies that

$$\begin{aligned} f(W_{k+1}) &\leq f(W_k) + \langle\nabla f(W_k), W_{k+1} - W_k\rangle + \frac{L}{2}\|W_{k+1} - W_k\|^2 \qquad (134)\\ &= f(W_k) - \frac{\alpha}{2}\|\nabla f(W_k)\odot\sqrt{F(W_k)}\|^2 - \frac{1}{F_{\max}}\left(\frac{1}{2\alpha} - \frac{LF_{\max}}{2}\right)\|W_{k+1} - W_k\|^2\\ &\quad + \frac{1}{2\alpha}\left\|\frac{W_{k+1} - W_k}{\sqrt{F(W_k)}} + \alpha\nabla f(W_k)\odot\sqrt{F(W_k)}\right\|^2 \end{aligned}$$

where the second inequality comes from

$$\begin{aligned} \langle\nabla f(W_k), W_{k+1} - W_k\rangle &= \alpha\left\langle\nabla f(W_k)\odot\sqrt{F(W_k)}, \frac{W_{k+1} - W_k}{\alpha\sqrt{F(W_k)}}\right\rangle \qquad (135)\\ &= -\frac{\alpha}{2}\|\nabla f(W_k)\odot\sqrt{F(W_k)}\|^2 - \frac{1}{2\alpha}\left\|\frac{W_{k+1} - W_k}{\sqrt{F(W_k)}}\right\|^2\\ &\quad + \frac{1}{2\alpha}\left\|\frac{W_{k+1} - W_k}{\sqrt{F(W_k)}} + \alpha\nabla f(W_k)\odot\sqrt{F(W_k)}\right\|^2 \end{aligned}$$

as well as the inequality

$$\left\|\frac{W_{k+1} - W_k}{\sqrt{F(W_k)}}\right\|^2 \geq \frac{1}{F_{\max}}\|W_{k+1} - W_k\|^2. \tag{136}$$

The third term in the RHS of (134) can be bounded by

$$\frac{1}{2\alpha}\left\|\frac{W_{k+1} - W_k}{\sqrt{F(W_k)}} + \alpha\nabla f(W_k)\odot\sqrt{F(W_k)}\right\|^2 = \frac{\alpha}{2}\left\||\nabla f(W_k)|\odot\frac{G(W_k)}{\sqrt{F(W_k)}}\right\|^2. \tag{137}$$

Define the saturation vector $R(W_k) \in \mathbb{R}^D$ by

$$
\begin{aligned}
R(W_k) :=& F(W_k)^{\odot 2} - G(W_k)^{\odot 2} = (F(W_k) + G(W_k)) \odot (F(W_k) - G(W_k)) \\
=& q_+(W_k) \odot q_-(W_k).
\end{aligned} \tag{138}
$$

Notice the following inequality is valid

$$
\begin{aligned}
& -\frac{\alpha}{2}\|\nabla f(W_k) \odot \sqrt{F(W_k)}\|^2 + \frac{\alpha}{2}\left\| |\nabla f(W_k)| \odot \frac{G(W_k)}{\sqrt{F(W_k)}} \right\|^2 \\
=& -\frac{\alpha}{2} \sum_{d\in[D]} \left([\nabla f(W_k)]_d^2 \left([F(W_k)]_d - \frac{[G(W_k)]_d^2}{[F(W_k)]_d} \right) \right) \\
=& -\frac{\alpha}{2} \sum_{d\in[D]} \left([\nabla f(W_k)]_d^2 \left(\frac{[F(W_k)]_d^2 - [G(W_k)]_d^2}{[F(W_k)]_d} \right) \right) \\
\le& -\frac{\alpha}{2F_{\max}} \sum_{d\in[D]} \left([\nabla f(W_k)]_d^2 \left([F(W_k)]_d^2 - G(W_k)]_d^2 \right) \right) \\
=& -\frac{\alpha}{2F_{\max}} \|\nabla f(W_k)\|_{S_k}^2 \le 0.
\end{aligned} \tag{139}
$$

Substituting (137) and (139) back into (134) yields

$$
\frac{1}{F_{\max}} \left(\frac{1}{2\alpha} - \frac{LF_{\max}}{2} \right) \|W_{k+1} - W_k\|^2 \le f(W_k) - f(W_{k+1}). \tag{140}
$$

Noticing that $\|W_{k+1} - W_k\|^2 = \alpha^2 \|\nabla f(W_k) \odot F(W_k) - |\nabla f(W_k)| \odot G(W_k)\|^2$ and averaging for k from 0 to $K-1$, we have

$$
\begin{aligned}
E_K^{\texttt{AGD}} =& \frac{1}{K} \sum_{k=0}^{K-1} \left(\|\nabla f(W_k) \odot F(W_k) - |\nabla f(W_k)| \odot G(W_k)\|^2 + \|\nabla f(W_k)\|_{R(W_k)}^2 \right) \\
\le& \frac{2(f(W_0) - f(W_{K+1}))F_{\max}}{\alpha(1 - \alpha L F_{\max})K} \le \frac{8L(f(W_0) - f^*)F_{\max}^2}{K}
\end{aligned} \tag{141}
$$

where the last inequality choose $\alpha = \frac{1}{2LF_{\max}}$. Further, if the degree of saturation is bounded, (134)–(139) implies that

$$
\frac{\alpha R_{\min}^{\texttt{AGD}}}{2} \|\nabla f(W_k)\|^2 \le \frac{\alpha}{2} \|\nabla f(W_k)\|_{R(W_k)}^2 \le f(W_k) - f(W_{k+1}). \tag{142}
$$

Averaging (142) for k from 0 to K deduce that

$$
\frac{1}{K} \sum_{k=0}^{K-1} \|\nabla f(W_k)\|^2 \le \frac{2(f(W_0) - f(W_{K+1}))F_{\max}}{\alpha K R_{\min}^{\texttt{AGD}}} \le \frac{2L(f(W_0) - f^*)F_{\max}^2}{K R_{\min}^{\texttt{AGD}}} \tag{143}
$$

where the second inequality holds because the learning rate is selected as $\alpha = \frac{1}{LF_{\max}}$. $\square$

K Simulation Details and Additional Results

This section provides details about the experiments in Section 6. All simulation is performed under the PyTorch framework https://github.com/pytorch/pytorch. The analog training algorithms, including `Analog SGD` and `Tiki-Taka`, are provided by the open-source simulation toolkit AIHWKit [44], which has MIT license; see github.com/IBM/aihwkit.

Optimizer. The digital SGD optimizer is implemented by `FloatingPointRPUConfig` in AIHWKit, which is equivalent to the SGD implemented in PyTorch. The `Analog SGD` is implemented by selecting `SingleRPUConfig` as configuration and `Tiki-Taka` optimizers are implemented by `UnitCellRPUConfig` with `TransferCompound` devices in AIHWKit.

Hardware. We conduct our experiments on one NVIDIA RTX 3090 GPU, which has 24GB memory and a maximum power of 350W. The simulations take from 30 minutes to 5 hours, depending on model sizes and datasets.

Statistical Significance. The simulation data reported in all tables is repeated three times. The randomness originates from the data shuffling, random initialization, and random noise in the analog hardware. The mean and standard deviation are calculated using *statistics* library.

K.1 Power and Exponential Response Functions

The *power response* is a power function, given by

$$q_+(w) = \left(1 - \frac{w}{\tau}\right)^{\gamma_{\text{res}}}, \qquad q_-(w) = \left(1 + \frac{w}{\tau}\right)^{\gamma_{\text{res}}} \tag{144}$$

which can be changed by adjusting the dynamic radius τ and shape parameter $\gamma_{\texttt{res}}$. We also consider the *exponential response*, whose response is an exponential function, defined by

$$q_+(w) = \frac{\exp\left(\gamma_{\texttt{res}}(1 - w/\tau)\right) - 1}{\exp\left(\gamma_{\texttt{res}}\right) - 1}, \qquad q_-(w) = \frac{\exp\left(\gamma_{\texttt{res}}(1 + w/\tau)\right) - 1}{\exp\left(\gamma_{\texttt{res}}\right) - 1}. \tag{145}$$

It could be checked that the boundary of their dynamic ranges are $\tau^{\max} = \tau$ and $\tau^{\min} = -\tau$, while the symmetric point is 0, as required by Corollary 1. Figure 5 illustrates how the response functions change with different $\gamma_{\texttt{res}}$.

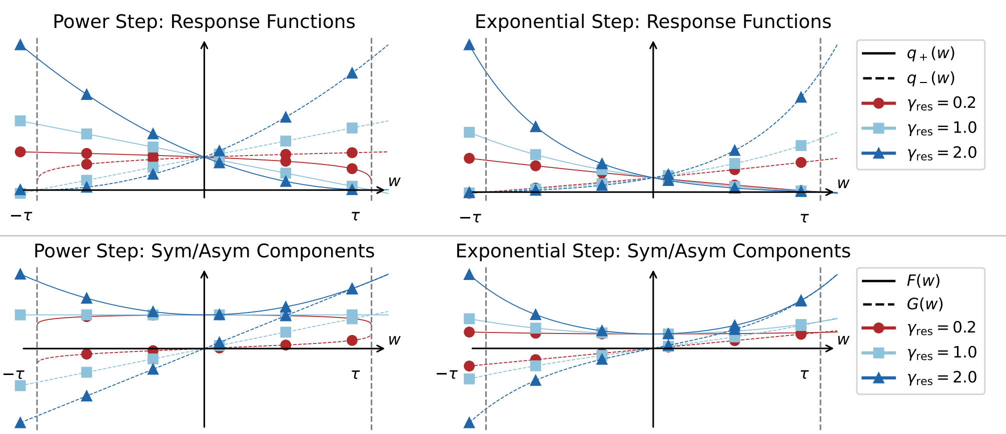

Figure 5: Examples of response functions. The dependence of the response function on the weight w can grow at various rates, including but not limited to power (Left) or exponential rate (Right). τ is the radius of the dynamic range, and $\gamma_{\texttt{res}}$ is a parameter that needs to be determined by physical measurements.

K.2 Least squares problem

In Figure 2 (see Section 1.1), we consider the least squares problem on a synthetic dataset and a ground truth $W^* \in \mathbb{R}^D$. The problem can be formulated by

$$\min_{W \in \mathbb{R}^D} f(W) := \frac{1}{2}\|AW - b\|^2 = \frac{1}{2}\|A(W - W^*)\|^2. \tag{146}$$

The elements of W^* are sampled from a Gaussian distribution with mean 0 and variance $\sigma^2_{W^*}$. Consider a matrix $A \in \mathbb{R}^{D_{\text{out}} \times D}$ of size $D = 50$ and $D_{\text{out}} = 100$ whose elements are sampled from a Gaussian distribution with variance σ^2_A. The label $b \in \mathbb{R}^{D_{\text{out}}}$ is generated by $b = AW^*$ where W^* are sampled from a standard Gaussian distribution with $\sigma^2_{W^*}$. The response granularity $\Delta w_{\min}$=1e-4 while $\tau = 3.5$. The maximum bit length is 8. The variance are set as $\sigma^2_A = 1.00^2$, $\sigma^2_{W^*} = 0.5^2$.

K.3 Classification problem

We conduct training simulations of image classification tasks on a series of real datasets. In the implementation of `Tiki-Taka`, only a few columns or rows of P_k are transferred per time to W_k in the recursion (11) to balance the communication and computation. In our simulations, we transfer 1 column every time. The response granularity is set as $\Delta w_{\min}$ =1e-3.

The other setting follows the settings of AIHWKIT, including output noise (0.5 % of the quantization bin width), quantization and clipping (output range set 20, output noise 0.1, and input and output quantization to 8 bits). Noise and bound management techniques are used in [71]. A learnable scaling factor is set after each analog layer, which is updated using SGD.

3-FC / MNIST. Following the setting in [16], we train a model with 3 fully connected layers. The hidden sizes are 256 and 128. The activation functions are Sigmoid. The learning rates are $\alpha = 0.1$ for `Digital SGD`, $\alpha = 0.05, \beta = 0.01$ for `Analog SGD` and `Tiki-Taka`. The batch size is 10 for all algorithms. In Figure 4, the power response functions with $\gamma_{\text{res}} = 0.5$ are used, and various τ are used as indicated in the legend.

CNN / MNIST. We train a convolution neural network, which contains 2-convolutional layers, 2-max-pooling layers, and 2-fully connected layers. The activation functions are Tanh. The first two convolutional layers use 5×5 kernels with 16 and 32 kernels, respectively. Each convolutional layer is followed by a subsampling layer implemented by the max pooling function over non-overlapping pooling windows of size 2×2. The output of the second pooling layer, consisting of 512 neuron activations, feeds into a fully connected layer consisting of 128 tanh neurons, which is then connected into a 10-way softmax output layer. The learning rates are set as $\alpha = 0.1$ for `Digital SGD`, $\alpha = 0.05, \beta = 0.01$ for `Analog SGD` are `Residual Learning/Tiki-Taka`. The batch size is 8 for all algorithms. In Figure 4, the power response functions with $\gamma_{\text{res}} = 0.5$ are used, and various τ are used as indicated in the legend.

ResNet & MobileNet / CIFAR10 & CIFAR100. We train different models from the ResNet family, including ResNet18, 34, and 50. The base model is pre-trained on ImageNet dataset. The last fully connected layer is replaced by an analog layer. The learning rates are set as $\alpha = 0.075$ for `Digital SGD`, $\alpha = 0.075, \beta = 0.01$ for `Analog SGD`, `Residual Learning/Tiki-Taka`, `Tiki-Taka v2`, and `Residual Learning v2`. `Tiki-Taka` adopts $\gamma = 0.4$ unless stated otherwise. The batch size is 128 for all algorithms.

K.4 Additional performance on real datasets

We train different models from the MobileNet family, including MobileNet2, MobileNetV3L, MobileNetV3S. The base model is pre-trained on ImageNet dataset. The last fully connected layer is replaced by an analog layer. The learning rates are set as $\alpha = 0.075$ for `Digital SGD`, $\alpha = 0.075, \beta = 0.01$ for `Analog SGD` or `Tiki-Taka`. `Tiki-Taka` adopts $\gamma = 0.4$ unless stated otherwise. The batch size is 128 for all algorithms. Power response function with $\gamma_{\texttt{res}} = 4.0$ and $\tau = 0.05$ is used in the simulations.

CIFAR10/CIFAR100 ResNet. We fine-tune three models from the ResNet family with different scales on CIFAR10/CIFAR100 datasets. The power response functions with $\gamma_{\text{res}} = 3.0$ and $\tau = 0.1$, and the exponential response functions with $\gamma_{\text{res}} = 4.0$ and $\tau = 0.1$ are used, whose results are shown in Table 1 and 3, respectively. The results show that the `Tiki-Taka` outperforms `Analog SGD` by about 1.0% in most of the cases in ResNet34/50, and the gap even reaches about 10.0% for ResNet18 training on the CIFAR100 dataset.

CIFAR10/CIFAR100 MobileNet. We fine-tune three MobileNet models with different scales on CIFAR10/CIFAR100 datasets. The response function is set as the power response with the parameter $\gamma_{\text{res}} = 4.0$ and $\tau = 0.05$, whose results are shown in Table 4. In the simulations, the accuracy of `Analog SGD` drops significantly by about 10% in most cases, while `Tiki-Taka` remains comparable to the `Digital SGD` with only a slight drop.

K.5 Ablation study on cycle variation

To verify the conclusion of Theorem 4 that the error introduced by cycle variation is a higher-order term, we conduct a numerical simulation training on an image classification task on the MNIST

	CIFAR10				
	DSGD	ASGD	TT/RL	TTv2	RLv2
ResNet18	95.43 $\pm$0.13	84.47 $\pm$3.40	94.81 $\pm$0.09	95.31 $\pm$0.05	95.12 $\pm$0.14
ResNet34	96.48 $\pm$0.02	95.43 $\pm$0.12	96.29 $\pm$0.12	96.60 $\pm$0.05	96.42 $\pm$0.13
ResNet50	96.57 $\pm$0.10	94.36 $\pm$1.16	96.34 $\pm$0.04	96.63 $\pm$0.09	96.56 $\pm$0.08
	CIFAR100				
	DSGD	ASGD	TT/RL	TTv2	RLv2
ResNet18	81.12 $\pm$0.25	68.98 $\pm$1.01	76.17 $\pm$0.23	78.56 $\pm$0.29	79.83 $\pm$0.13
ResNet34	83.86 $\pm$0.12	78.98 $\pm$0.55	80.58 $\pm$0.11	81.81 $\pm$0.15	82.85 $\pm$0.19
ResNet50	83.98 $\pm$0.11	79.88 $\pm$1.26	80.80 $\pm$0.22	82.82 $\pm$0.33	83.90 $\pm$0.20

Table 3: Fine-tuning ResNet models with the *exponential response* on CIFAR10/100 datasets. Test accuracy is reported. `DSGD`, `ASGD`, and `TT` represent `Digital SGD`, `Analog SGD`, `Tiki-Taka`, respectively.

dataset using Fully-connected network (FCN) or convolution neural network (CNN) network. In the pulse update (26), the parameter σ_c is varied from 10% to 120%, where the noise signal is already larger than the response function signal itself. The results are shown in Table 5. The results show that the test accuracy of both `Analog SGD` and `Tiki-Taka` is not significantly affected by the cycle variation, which complies with the theoretical analysis.

K.6 Ablation study on various response functions

We also train a FCN model on the MNIST dataset under various response functions. As shown in the figure, larger $\gamma_{\texttt{res}}$ leads to a steeper response function. The results are shown in Table 6. The accuracy < 15.00 in the table implies that `Analog SGD` fails completely at all trials, which is close to random guess. The results show that `Analog SGD` works well only when the asymmetric is mild, i.e. $\gamma_{\texttt{res}}$ is small and τ is large, while `Tiki-Taka` outperforms `Analog SGD` and achieves comparable accuracy with `Digital SGD`.

K.7 Ablation study on γ

We conduct a series of simulations to study the impact of mixing coefficient γ in (10) on CIFAR10 or CIFAR100 dataset in the ResNet training tasks. The results are presented in Figure 6, which shows that `Tiki-Taka` achieves a great accuracy gain from increasing γ from 0 to 0.1, while the gain saturates after that. Therefore, we conclude that `Tiki-Taka` benefits from a non-zero γ, and the performance is robust to the γ selection.

	CIFAR10				
	DSGD	ASGD	TT/RL	TTv2	RLv2
MobileNetV2	95.28 $\pm$0.20	94.34 $\pm$0.27	95.05 $\pm$0.11	95.20 $\pm$0.14	95.26 $\pm$0.03
MobileNetV3S	94.45 $\pm$0.10	80.66 $\pm$6.18	93.65 $\pm$0.24	93.54 $\pm$0.06	93.79 $\pm$0.00
MobileNetV3L	95.95 $\pm$0.08	80.79 $\pm$2.97	95.39 $\pm$0.27	95.27 $\pm$0.09	95.33 $\pm$0.08
	CIFAR100				
	DSGD	ASGD	TT/RL	TTv2	RLv2
MobileNetV2	80.60 $\pm$0.18	63.41 $\pm$1.20	73.33 $\pm$0.94	78.41 $\pm$0.15	79.60 $\pm$0.10
MobileNetV3S	78.94 $\pm$0.05	51.79 $\pm$1.05	71.14 $\pm$0.93	74.51 $\pm$0.37	75.39 $\pm$0.00
MobileNetV3L	82.16 $\pm$0.26	66.80 $\pm$1.40	78.81 $\pm$0.52	79.56 $\pm$0.10	80.18 $\pm$0.07

Table 4: Fine-tuning MobileNet models with *power response* on CIFAR10/100 datasets. Test accuracy is reported. `DSGD`, `ASGD`, and `TT` represent `Digital SGD`, `Analog SGD`, `Tiki-Taka`, respectively.

	FCN			CNN		
	DSGD	ASGD	TT	DSGD	ASGD	TT
$\sigma_c = 10\%$		97.22 $\pm$0.21	97.66 $\pm$0.04		92.68 $\pm$0.45	98.74 $\pm$0.07
$\sigma_c = 30\%$		96.97 $\pm$0.12	97.07 $\pm$0.12		93.36 $\pm$0.55	98.89 $\pm$0.05
$\sigma_c = 60\%$	98.17 $\pm$0.05	96.33 $\pm$0.21	97.70 $\pm$0.09	99.09 $\pm$0.04	93.07 $\pm$0.53	98.68 $\pm$0.09
$\sigma_c = 90\%$		95.99 $\pm$0.15	97.44 $\pm$0.15		91.87 $\pm$0.48	98.92 $\pm$0.02
$\sigma_c = 120\%$		96.19 $\pm$0.20	96.97 $\pm$0.20		91.57 $\pm$0.58	98.85 $\pm$0.04

Table 5: Test accuracy comparison under different cycle variation levels σ_c on MNIST dataset. `DSGD`, `ASGD`, and `TT` represent `Digital SGD`, `Analog SGD`, `Tiki-Taka`, respectively

		DSGD	Power response		Exponential response	
			ASGD	TT/RL	ASGD	TT/RL
$\gamma_{\texttt{res}} = 0.5$	$\tau = 0.6$		96.01 $\pm$0.26	96.92 $\pm$0.19	<15.00	97.27 $\pm$0.07
	$\tau = 0.7$		97.40 $\pm$0.15	97.05 $\pm$0.05	<15.00	97.39 $\pm$0.15
	$\tau = 0.8$		97.38 $\pm$0.10	96.82 $\pm$0.17	94.00 $\pm$0.63	97.16 $\pm$0.16
$\gamma_{\texttt{res}} = 1.0$	$\tau = 0.6$		<15.00	97.39 $\pm$0.05	<15.00	97.46 $\pm$0.08
	$\tau = 0.7$	98.17 $\pm$0.05	<15.00	97.33 $\pm$0.05	<15.00	97.49 $\pm$0.04
	$\tau = 0.8$		<15.00	97.34 $\pm$0.09	<15.00	97.25 $\pm$0.16
$\gamma_{\texttt{res}} = 2.0$	$\tau = 0.6$		<15.00	96.93 $\pm$0.15	<15.00	97.19 $\pm$0.16
	$\tau = 0.7$		<15.00	97.27 $\pm$0.02	<15.00	97.72 $\pm$0.07
	$\tau = 0.8$		<15.00	97.18 $\pm$0.04	<15.00	97.06 $\pm$0.10

Table 6: Test accuracy comparison under different response function parameters τ and $\gamma_{\texttt{res}}$ for FCN training on MNIST dataset with power or exponential response functions. `DSGD`, `ASGD`, and `TT` represent `Digital SGD`, `Analog SGD`, `Tiki-Taka`, respectively.

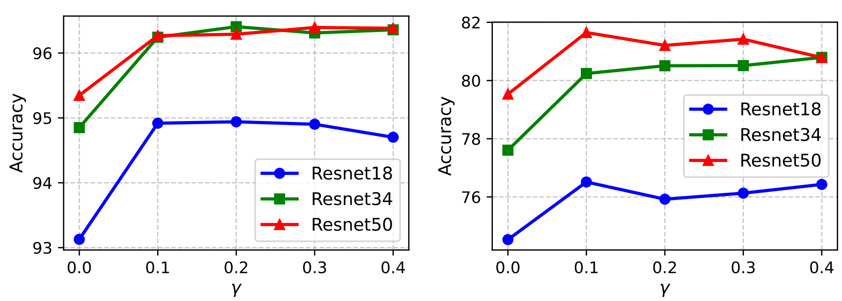

Figure 6: The test accuracy of ResNet family models after 100 epochs trained by `Tiki-Taka` under different γ in (10); **(Left)** CIFAR10. **(Right)** CIFAR100.

L Broader Impact

This paper focuses on developing a theoretical analysis for gradient-based training algorithms on a class of generic AIMC hardware, which can be leveraged to boost both energy and computational efficiency of training. While such efficiency gains could, in principle, enable broader and potentially unintended uses of machine learning models, we do not identify any specific societal risks that need to be highlighted in this context.