**Noname manuscript No.**
(will be inserted by the editor)

# Blended Coarse Gradient Descent for Full Quantization of Deep Neural Networks

**Penghang Yin · Shuai Zhang · Jiancheng Lyu · Stanley Osher · Yingyong Qi · Jack Xin***

Received: date / Accepted: date

**Abstract** Quantized deep neural networks (QDNNs) are attractive due to their much lower memory storage and faster inference speed than their regular full precision counterparts. To maintain the same performance level especially at low bit-widths, QDNNs must be retrained. Their training involves piecewise constant activation functions and discrete weights, hence mathematical challenges arise. We introduce the notion of coarse gradient and propose the blended coarse gradient descent (BCGD) algorithm, for training fully quantized neural networks. Coarse gradient is generally not a gradient of any function but an artificial ascent direction. The weight update of BCGD goes by coarse gradient correction of a weighted average of the full precision weights and their quantization (the so-called blending), which yields sufficient descent in the objective value and thus accelerates the training. Our experiments demonstrate that this simple blending technique is very effective for quantization at extremely low bit-width such as binarization. In full quantization of ResNet-18 for ImageNet classification task, BCGD gives 64.36% top-1 accuracy with binary weights across all layers and 4-bit adaptive activation. If the weights in the first and last layers are kept in full precision, this number increases to 65.46%. As theoretical justification, we show convergence analysis of coarse gradient descent for a two-linear-layer neural network model with Gaussian input data, and prove that the expected coarse gradient correlates positively with the underlying true gradient.

**Keywords** weight/activation quantization · blended coarse gradient descent · sufficient descent property · deep neural networks

**Mathematics Subject Classification (2010)** 90C35, 90C26, 90C52, 90C90.

P. Yin, S. Zhang and J. Lyu contributed equally.

Penghang Yin · Stanley Osher
Department of Mathematics, University of California at Los Angeles, Los Angeles, CA 90095
E-mail: yph@ucla.edu, sjo@math.ucla.edu

Shuai Zhang · Yingyong Qi
Qualcomm AI Research, San Diego, CA 92121
E-mail: shuazhan@qti.qualcomm.com, yingyong@qti.qualcomm.com

Jiancheng Lyu · Jack Xin
Department of Mathematics, University of California at Irvine, Irvine, CA 92697
E-mail: jianchel@uci.edu; jxin@math.uci.edu, *corresponding author, (949)-331-6314.

## 1 Introduction

Deep neural networks (DNNs) have seen enormous success in image and speech classification, natural language processing, health sciences among other big data driven applications in recent years. However, DNNs typically require hundreds of megabytes of memory storage for the trainable full-precision floating-point parameters, and billions of FLOPs (floating point operations per second) to make a single inference. This makes the deployment of DNNs on mobile devices a challenge. Some considerable recent efforts have been devoted to the training of low precision (quantized) models for substantial memory savings and computation/power efficiency, while nearly maintaining the performance of full-precision networks. Most works to date are concerned with weight quantization (WQ) [8,22,28,36,5,35]. In [13], He et al. theoretically justified for the applicability of WQ models by investigating their expressive power. Some also studied activation function quantization (AQ) [17,28,18,4,26,37], which utilize an external process outside of the network training. This is different from WQ at 4 bit or under, which must be achieved through network training. Learning activation function $\sigma$ as a parametrized family ($\sigma = \sigma(x, \alpha)$) and part of network training has been studied in [15] for parametric rectified linear unit, and was recently extended to uniform AQ in [6]. In uniform AQ, $\sigma(x, \alpha)$ is a step (or piecewise constant) function in $x$, and the parameter $\alpha$ determines the height and length of the steps. In terms of the partial derivative of $\sigma(x, \alpha)$ in $\alpha$, a two-valued proxy derivative of the parametric activation function (PACT) was proposed [6], although we will present an almost everywhere (a.e.) exact one in this paper.

The mathematical difficulty in training activation quantized networks is that the loss function becomes a piecewise constant function with sampled stochastic gradient a.e. zero, which is undesirable for back-propagation. A simple and effective way around this problem is to use a (generalized) straight-through (ST) estimator or derivative of a related (sub)differentiable function [16,1,17,18] such as clipped rectified linear unit (clipped ReLU) [4]. The idea of ST estimator dates back to the perceptron algorithm [29,30] proposed in 1950s for learning single-layer perceptrons with binary output. For multi-layer networks with hard threshold activation (a.k.a. binary neuron), Hinton [16] proposed to use the derivative of identity function as a proxy in back-propagation or chain rule, similar to the perceptron algorithm. The proxy derivative used in backward pass only was referred as straight-through estimator in [1], and several variants of ST estimator [17,18,4] have been proposed for handling quantized activation functions since then. A similar situation, where the derivative of certain layer composited in the loss function is unavailable for back-propagation, has also been brought up by [33] recently while improving accuracies of DNNs by replacing the softmax classifier layer with an implicit weighted nonlocal Laplacian layer. For the training of the latter, the derivative of a pre-trained fully-connected layer was used as a surrogate [33].

On the theoretical side, while the convergence of the single-layer perception algorithm has been extensively studied [34,11], there is almost no theoretical understanding of the unusual 'gradient' output from the modified chain rule based on ST estimator. Since this unusual 'gradient' is certainly not the gradient of the objective function, then a question naturally arises: *how does it correlate to the objective function?* One of the contributions in this paper is to answer this question. Our main contributions are threefold:

1. Firstly, we introduce the notion of coarse derivative and cast the early ST estimators or proxy partial derivatives of $\sigma(x, \alpha)$ in $\alpha$ including the two-valued PACT of [6] as examples. The coarse derivative is non-unique. We propose a three-valued coarse partial derivative of the quantized activation function $\sigma(x, \alpha)$ in $\alpha$ that can outperform the two-valued one [6] in network training. We find that unlike the partial derivative $\frac{\partial \sigma}{\partial x}(x, \alpha)$ which vanishes, the a.e. partial derivative of $\sigma(x, \alpha)$ in $\alpha$ is actually multi-valued (piecewise constant). Surprisingly, this a.e. accurate derivative is empirically less useful than the coarse ones in fully quantized network training.

2. Secondly, we propose a novel accelerated training algorithm for fully quantized networks, termed blended coarse gradient descent method (BCGD). Instead of correcting the current full precision weights with coarse gradient at their quantized values like in the popular BinaryConnect scheme [8,17,28,22,37,4,23,35], the BCGD weight update goes by coarse gradient correction of a suitable average of the full precision weights and their quantization. We shall show that BCGD satisfies the sufficient descent property for objectives with Lipschitz gradients, while BinaryConnect does not unless an approximate orthogonality condition holds for the iterates [35].

3. Our third contribution is the mathematical analysis of coarse gradient descent for a two-layer network with binarized ReLU activation function and i.i.d. unit Gaussian data. We provide an explicit form of coarse gradient based on proxy derivative of regular ReLU, and show that when there are infinite training data, the negative expected coarse gradient gives a descent direction for minimizing the expected training loss. Moreover, we prove that a normalized coarse gradient descent algorithm only converges to either a global minimum or a potential spurious local minimum. This answers the question.

The rest of the paper is organized as follows. In section 2, we discuss the concept of coarse derivative and give examples for quantized activation functions. In section 3, we present the joint weight and activation quantization problem, and BCGD algorithm satisfying the sufficient descent property. For readers' convenience, we also review formulas on 1-bit, 2-bit and 4-bit weight quantization used later in our numerical experiments. In section 4, we give details of fully quantized network training, including the disparate learning rates on weight and $\alpha$. We illustrate the enhanced validation accuracies of BCGD over BinaryConnect, and 3-valued coarse $\alpha$ partial derivative of $\sigma$ over 2-valued and a.e. $\alpha$ partial derivative in case of 4-bit activation, and (1,2,4)-bit weights on CIFAR-10 image datasets. We show top-1 and top-5 validation accuracies of ResNet-18 with all convolutional layers quantized at 1-bit weight/4-bit activation (1W4A), 4-bit weight/4-bit activation (4W4A), and 4-bit weight/8-bit activation (4W8A), using 3-valued and 2-valued $\alpha$ partial derivatives. The 3-valued $\alpha$ partial derivative out-performs the two-valued with larger margin in the low bit regime. The accuracies degrade gracefully from 4W8A to 1W4A while all the convolutional layers are quantized. The 4W8A accuracies with either the 3-valued or the 2-valued $\alpha$ partial derivatives are within 1% of those of the full precision network. If the first and last convolutional layers are in full precision, our top-1 (top-5) accuracy of ResNet-18 at 1W4A with 3-valued coarse $\alpha$-derivative is 4.7 % (3%) higher than that of HWGQ [4] on ImageNet dataset. This is in part due to the value of parameter $\alpha$ being learned without any statistical assumption.

**Notations.** $\|\cdot\|$ denotes the Euclidean norm of a vector or the spectral norm of a matrix; $\|\cdot\|_\infty$ denotes the $\ell_\infty$-norm. $\mathbf{0} \in \mathbb{R}^n$ represents the vector of zeros, whereas $\mathbf{1} \in \mathbb{R}^n$ the vector of all ones. We denote vectors by bold small letters and matrices by bold capital ones. For any $\mathbf{w}, \mathbf{z} \in \mathbb{R}^n$, $\mathbf{w}^\top \mathbf{z} = \langle \mathbf{w}, \mathbf{z} \rangle = \sum_i w_i z_i$ is their inner product. $\mathbf{w} \odot \mathbf{z}$ denotes the Hadamard product whose $i$-th entry is given by $(\mathbf{w} \odot \mathbf{z})_i = w_i z_i$.

## 2 Activation Quantization

In a network with quantized activation, given a training sample of input $\mathbf{Z}$ and label $u$, the associated sample loss is a composite function of the form:

$$\ell(\mathbf{w}, \boldsymbol{\alpha}; \{\mathbf{Z}, u\}) := \ell(\mathbf{w}_l * \sigma(\mathbf{w}_{l-1} * \cdots \mathbf{w}_2 * \sigma(\mathbf{w}_1 * \mathbf{Z}, \alpha_1) \cdots, \alpha_{l-1}); u), \tag{1}$$

where $\mathbf{w}_j$ contains the weights in the $j$-th linear (fully-connected or convolutional) layer, '$*$' denotes either matrix-vector product or convolution operation; reshaping is necessary to avoid mismatch in dimensions. The $j$-th quantized ReLU $\sigma(\mathbf{x}_j, \alpha_j)$ acts element-wise on the vector/tensor $\mathbf{x}_j$ output from the previous linear layer, which is parameterized by a trainable scalar $\alpha_j > 0$ known as the resolution. For practical hardware-level implementation, we are most interested in uniform quantization:

$$\sigma(x, \alpha) = \begin{cases} 0, & \text{if} \quad x \le 0, \\ k\alpha, & \text{if} \quad (k-1)\,\alpha < x \le k\alpha,\ k = 1, 2, \ldots, 2^{b_a} - 1, \\ \left(2^{b_a} - 1\right)\alpha, & \text{if} \quad x > \left(2^{b_a} - 1\right)\alpha, \end{cases} \tag{2}$$

where $x$ is the scalar input, $\alpha > 0$ the resolution, $b_a \in \mathbb{Z}_+$ the bit-width of activation and $k$ the quantization level. For example, in 4-bit activation quantization (4A), we have $b_a = 4$ and $2^{b_a} = 16$ quantization levels including the zero.

Given $N$ training samples, we train the network with quantized ReLU by solving the following empirical risk minimization

$$\min_{\mathbf{w}, \boldsymbol{\alpha}} f(\mathbf{w}, \boldsymbol{\alpha}) := \frac{1}{N} \sum_{i=1}^{N} \ell(\mathbf{w}, \boldsymbol{\alpha}; \{\mathbf{Z}^{(i)}, u^{(i)}\}) \tag{3}$$

In gradient-based training framework, one needs to evaluate the gradient of the sample loss (1) using the so-called back-propagation (a.k.a. chain rule), which involves the computation of partial derivatives $\frac{\partial \sigma}{\partial x}$ and $\frac{\partial \sigma}{\partial \alpha}$. Apparently, the partial derivative of $\sigma(x, \alpha)$ in $x$ is almost everywhere (a.e.) zero. After composition, this results in a.e. zero gradient of $\ell$ with respect to (w.r.t.) $\{\mathbf{w}_j\}_{j=1}^{l-1}$ and $\{\alpha_j\}_{j=1}^{l-2}$ in (1), causing their updates to become stagnant. To see this, we abstract the partial gradients $\frac{\partial \ell}{\partial \mathbf{w}_{l-1}}$ and $\frac{\partial \ell}{\partial \alpha_{l-2}}$, for instances, through the chain rule as follows:

$$\frac{\partial \ell}{\partial \mathbf{w}_{l-1}}(\mathbf{w}, \boldsymbol{\alpha}; \{\mathbf{Z}, u\}) = \sigma(\mathbf{x}_{l-2}, \alpha_{l-2}) \circ \frac{\partial \sigma}{\partial x}(\mathbf{x}_{l-1}, \alpha_{l-1}) \circ \mathbf{w}_l^\top \circ \nabla \ell(\mathbf{x}_l; u)$$

and

$$\frac{\partial \ell}{\partial \alpha_{l-2}}(\mathbf{w}, \boldsymbol{\alpha}; \{\mathbf{Z}, u\}) = \frac{\partial \sigma}{\partial \alpha}(\mathbf{x}_{l-2}, \alpha_{l-2}) \circ \mathbf{w}_{l-1}^\top \circ \frac{\partial \sigma}{\partial x}(\mathbf{x}_{l-1}, \alpha_{l-1}) \circ \mathbf{w}_l^\top \circ \nabla \ell(\mathbf{x}_l; u),$$

where we recursively define $\mathbf{x}_1 = \mathbf{w}_1 * \mathbf{Z}$, and $\mathbf{x}_j = \mathbf{w}_j * \sigma(\mathbf{x}_{j-1}, \alpha_{j-1})$ for $j \geq 2$ as the output from the $j$-th linear layer, and '$\circ$' denotes some sort of proper composition in the chain rule. It is clear that the two partial gradients are zeros a.e. because of the term $\frac{\partial \sigma}{\partial x}(\mathbf{x}_{l-1}, \alpha_{l-1})$. In fact, the automatic differentiation embedded in deep learning platforms such as PyTorch [27] would produce precisely zero gradients.

To get around this, we use a proxy derivative or so-called ST estimator for back-propagation. By overloading the notation '$\approx$', we denote the proxy derivative by

$$\frac{\partial \sigma}{\partial x}(x, \alpha) \approx \begin{cases} 0, & \text{if} \quad x \leq 0, \\ 1, & \text{if} \quad 0 < x \leq \left(2^{b_a} - 1\right) \alpha, \\ 0, & \text{if} \quad x > \left(2^{b_a} - 1\right) \alpha. \end{cases}$$

The proxy partial derivative has a non-zero value in the middle to reflect the overall variation of $\sigma$, which can be viewed as the derivative of the large scale (step-back) view of $\sigma$ in $x$, or the derivative of the clipped ReLU [4]:

$$\tilde{\sigma}(x, \alpha) = \begin{cases} 0, & \text{if} \quad x \leq 0, \\ x, & \text{if} \quad 0 < x \leq (2^{b_a} - 1)\alpha, \\ \left(2^{b_a} - 1\right) \alpha, & \text{if} \quad x > \left(2^{b_a} - 1\right) \alpha. \end{cases} \tag{4}$$

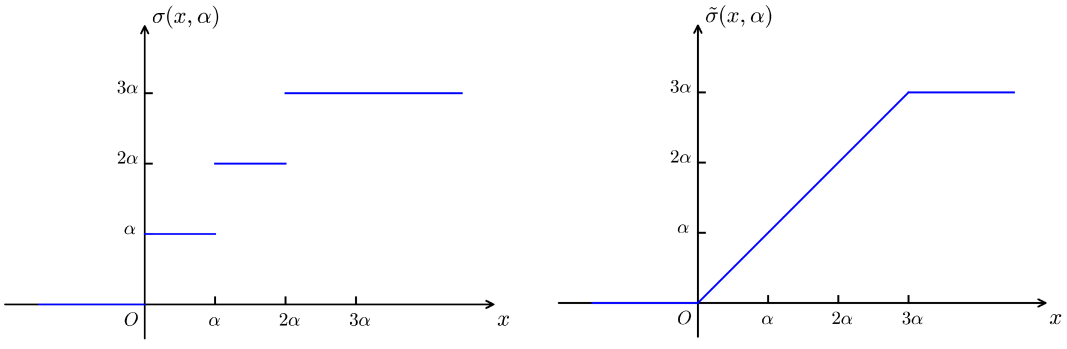


**Fig. 1** Left: plot of 2-bit quantized ReLU $\sigma(x, \alpha)$ in $x$. Right: plot of the associated clipped ReLU $\tilde{\sigma}(x, \alpha)$ in $x$.

On the other hand, we find the a.e. partial derivative of $\sigma(x, \alpha)$ w.r.t. $\alpha$ to be

$$\frac{\partial \sigma}{\partial \alpha}(x, \alpha) = \begin{cases} 0, & \text{if} \quad x \leq 0, \\ k, & \text{if} \quad (k-1)\,\alpha < x \leq k\alpha, \quad k = 1, 2, \cdots, 2^{b_a} - 1; \\ 2^{b_a} - 1, & \text{if} \quad x > \left(2^{b_a} - 1\right) \alpha. \end{cases}$$

Surprisingly, this a.e. derivative is not the best in terms of accuracy or computational cost in training, as will be reported in section 4. We propose an empirical three-valued proxy partial derivative in $\alpha$ as follows

$$\frac{\partial \sigma}{\partial \alpha}(x,\alpha) \approx \begin{cases} 0, & \text{if} \quad x \le 0, \\ 2^{(b_a-1)}, & \text{if} \quad 0 < x \le \left(2^{b_a}-1\right)\alpha, \\ 2^{b_a}-1, & \text{if} \quad x > \left(2^{b_a}-1\right)\alpha. \end{cases}$$

The middle value $2^{b_a-1}$ is the arithmetic mean of the intermediate $k$ values of the a.e. partial derivative above. Similarly, a more coarse two-valued proxy, same as PACT [6] which was derived differently, follows by zeroing out all the nonzero values except their maximum:

$$\frac{\partial \sigma}{\partial \alpha}(x,\alpha) \approx \begin{cases} 0, & \text{if} \quad x \le \left(2^{b_a}-1\right)\alpha, \\ 2^{b_a}-1, & \text{if} \quad x > \left(2^{b_a}-1\right)\alpha. \end{cases}$$

This turns out to be exactly the partial derivative $\dfrac{\partial \tilde{\sigma}}{\partial \alpha}(x,\alpha)$ of the clipped ReLU defined in (4).

We shall refer to the resultant composite 'gradient' of $f$ through the modified chain rule and averaging as coarse gradient. While given the name 'gradient', we believe it is generally not the gradient of any smooth function. It, nevertheless, somehow exploits the essential information of the piecewise constant function $f$, and its negation provides a descent direction for the minimization. In section 5, we will validate this claim by examining a two-layer network with i.i.d. Gaussian data. We find that when there are infinite number of training samples, the overall training loss $f$ (i.e., population loss) becomes pleasantly differentiable whose gradient is non-trivial and processes certain Lipschitz continuity. More importantly, we shall show an example of expected coarse gradient that provably forms an acute angle with the underlying true gradient of $f$ and only vanishes at the possible local minimizers of the original problem.

During the training process, the vector $\boldsymbol{\alpha}$ (one component per activation layer) should be prevented from being either too small or too large. Due to the sensitivity of $\boldsymbol{\alpha}$, we propose a two-scale training and set the learning rate of $\boldsymbol{\alpha}$ to be the learning rate of weight $\mathbf{w}$ multiplied by a rate factor far less than 1, which may be varied depending on network architectures. That rate factor effectively helps quantized network converge steadily and prevents $\boldsymbol{\alpha}$ from vanishing.

## 3 Full Quantization

Imposing the quantized weights amounts to adding a discrete set-constraint $\mathbf{w} \in \mathcal{Q}$ to the optimization problem (3). Suppose $M$ is the total number of weights in the network. For commonly used $b_w$-bit layer-wise quantization, $\mathcal{Q} \subset \mathbb{R}^M$ takes the form of $\mathcal{Q}_1 \times \mathcal{Q}_2 \cdots \times \mathcal{Q}_l$, meaning that the weight tensor in the $j$-th linear layer is constrained in the form $\mathbf{w}_j = \delta_j \mathbf{q}_j \in \mathcal{Q}_j$ for some adjustable scaling factor $\delta_j > 0$ shared by weights in the same layer. Each component of $\mathbf{q}_j$ is drawn from the quantization set given by $\{\pm 1\}$ for $b_w = 1$ (binarization) and $\{0, \pm 1, \cdots, \pm(2^{b_w-1}-1)\}$ for $b_w \ge 2$. This assumption on $\mathcal{Q}$ generalizes those of the 1-bit BWN [28] and the 2-bit TWN [22]. As such, the layer-wise weight and activation quantization problem here can be stated abstractly as follows

$$\min_{\mathbf{w},\boldsymbol{\alpha}} f(\mathbf{w},\boldsymbol{\alpha}) \ \text{ subject to } \ \mathbf{w} \in \mathcal{Q} = \mathcal{Q}_1 \times \mathcal{Q}_2 \cdots \times \mathcal{Q}_l, \tag{5}$$

where the training loss $f(\mathbf{w},\boldsymbol{\alpha})$ is defined in (3). Different from activation quantization, one bit is taken to represent the signs. For ease of presentation, we only consider the network-wise weight quantization throughout this section, i.e., weights across all the layers share the same (trainable) floating scaling factor $\delta > 0$, or simply, $\mathcal{Q} = \mathbb{R}_+ \times \{\pm 1\}^M$ for $b_w = 1$ and $\mathcal{Q} = \mathbb{R}_+ \times \left\{0, \pm 1, \ldots, \pm(2^{b_w-1}-1)\right\}^M$ for $b_w \ge 2$.

### 3.1 Weight Quantization

Given a float weight vector $\mathbf{w}_f$, the quantization of $\mathbf{w}_f$ is basically the following optimization problem for computing the projection of $\mathbf{w}_f$ onto set $\mathcal{Q}$

$$\text{proj}_{\mathcal{Q}}(\mathbf{w}_f) := \arg\min_{\mathbf{w} \in \mathcal{Q}} \ \|\mathbf{w} - \mathbf{w}_f\|^2. \tag{6}$$

Note that $\mathcal{Q}$ is a non-convex set, so the solution of (6) may not be unique. When $b_w = 1$, we have the binarization problem

$$\min_{\delta,\mathbf{q}} \|\delta\,\mathbf{q} - \mathbf{w}_f\|^2 \quad \text{subject to} \quad \delta > 0,\ \mathbf{q} \in \{\pm 1\}^M. \tag{7}$$

For $b_w \geq 2$, the projection/quantization problem (6) can be reformulated as

$$\min_{\delta,\mathbf{q}} \|\delta\,\mathbf{q} - \mathbf{w}_f\|^2 \quad \text{subject to} \quad \delta > 0,\ \mathbf{q} \in \left\{0, \pm 1, \cdots, \pm(2^{b_w-1}-1)\right\}^M. \tag{8}$$

It has been shown that the closed form (exact) solution of (7) can be computed at $O(M)$ complexity for (1-bit) binarization [28] and at $O(M\log(M))$ complexity for (2-bit) ternarization [36]. An empirical ternarizer of $O(M)$ complexity has also been proposed [22]. At wider bit-width $b_w \geq 3$, accurately solving (8) becomes computationally intractable due to the combinatorial nature of the problem [36].

The problem (8) is basically a constrained $K$-means clustering problem of 1-D points [35] with the centroids being $\delta$-spaced. It in principle can be solved by a variant of the classical Lloyd's algorithm [25] via an alternating minimization procedure. It iterates between the assignment step ($\mathbf{q}$-update) and centroid step ($\delta$-update). In the $i$-th iteration, fixing the scaling factor $\delta^{i-1}$, each entry of $\mathbf{q}^i$ is chosen from the quantization set, so that $\delta^{i-1}\mathbf{q}^i$ is as close as possible to $\mathbf{w}_f$. In the $\delta$-update, the following quadratic problem

$$\min_{\delta\in\mathbb{R}} \|\,\delta\,\mathbf{q}^i - \mathbf{w}_f\|^2$$

is solved by $\delta^i = \frac{(\mathbf{q}^i)^\top \mathbf{w}_f}{\|\mathbf{q}^i\|^2}$. Since quantization (6) is required in every iteration, to make this procedure practical, we just perform a single iteration of Lloyd's algorithm by empirically initializing $\delta$ to be $\frac{2}{2^{b_w}-1}\|\mathbf{w}_f\|_\infty$, which is derived by setting

$$\frac{\delta}{2}\left((2^{b_w-1}-1) + 2^{b_w-1}\right) = \|\mathbf{w}_f\|_\infty.$$

This makes the large components in $\mathbf{w}_f$ well clustered.

First introduced in [8] by Courbariaux et al., the BinaryConnect (BC) scheme has drawn much attention in training DNNs with quantized weight and regular ReLU. It can be summarized as

$$\mathbf{w}_f^{t+1} = \mathbf{w}_f^t - \eta\nabla f(\mathbf{w}^t),\ \mathbf{w}^{t+1} = \mathrm{proj}_{\mathcal{Q}}(\mathbf{w}_f^{t+1}),$$

where $\{\mathbf{w}^t\}$ denotes the sequence of the desired quantized weights, and $\{\mathbf{w}_f^t\}$ is an auxiliary sequence of floating weights. BC can be readily extended to full quantization regime by including the update of $\boldsymbol{\alpha}^t$ and replacing the true gradient $\nabla f(\mathbf{w}^t)$ with the coarse gradients from section 2. With a subtle change to the standard projected gradient descent algorithm (PGD) [7], namely

$$\mathbf{w}_f^{t+1} = \mathbf{w}^t - \eta\nabla f(\mathbf{w}^t),\ \mathbf{w}^{t+1} = \mathrm{proj}_{\mathcal{Q}}(\mathbf{w}_f^{t+1}),$$

BC significantly outperforms PGD and effectively bypasses spurious the local minima in $\mathcal{Q}$ [23]. An intuitive explanation is that the constraint set $\mathcal{Q}$ is basically a finite union of isolated one-dimensional subspaces (i.e., lines that pass through the origin) [35]. Since $\mathbf{w}_f^t$ is obtained near the projected point $\mathbf{w}^t$, the sequence $\{\mathbf{w}_f^t\}$ generated by PGD can get stuck in some line subspace easily when updated with a small learning rate $\eta$; see Figure 2 for graphical illustrations.

### 3.2 Blended Gradient Descent and Sufficient Descent Property

Despite the superiority of BC over PGD, we point out a drawback in regard to its convergence. While Yin et al. provided the convergence proof of BC scheme in the recent papers [35], their analysis hinges on an approximate orthogonality condition which may not hold in practice; see Lemma 4.4 and Theorem 4.10 of [35]. Suppose $f$ has $L$-Lipschitz gradient[1]. In light of the convergence proof in Theorem 4.10 of [35], we have

$$f(\mathbf{w}^{t+1}) - f(\mathbf{w}^t) \leq -\frac{1}{2}\left(\frac{1}{\eta}(\|\mathbf{w}^{t+1} - \mathbf{w}_f^t\|^2 - \|\mathbf{w}^t - \mathbf{w}_f^t\|^2) - L\|\mathbf{w}^{t+1} - \mathbf{w}^t\|^2\right). \tag{9}$$

[1] This assumption is valid for the population loss function; we refer readers to Lemma 2 in section 5.

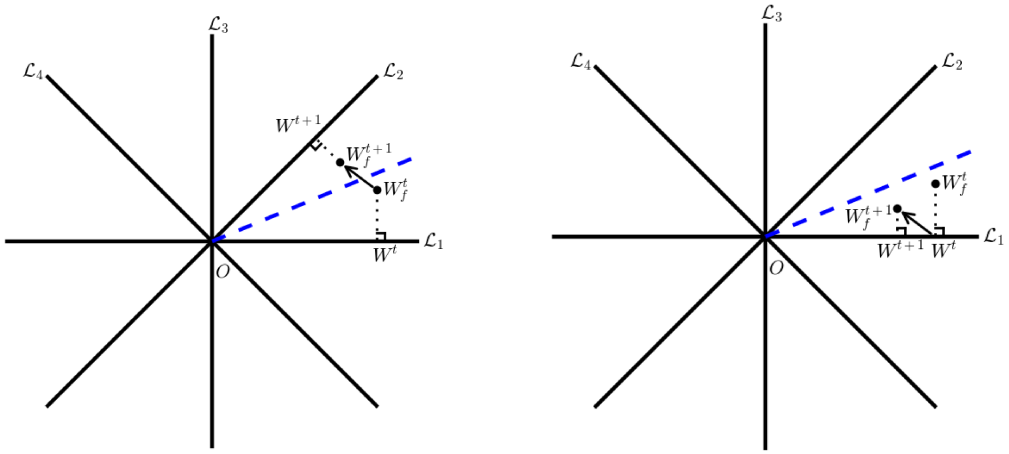


**Fig. 2** The ternarization of two weights. The one-dimensional subspaces $\mathcal{L}_i$'s constitute the constraint set $\mathcal{Q} = \mathbb{R}_+ \times \{0, \pm 1\}^2$. When updated with small learning rate, BC keeps searching among the subspaces (left), whereas PGD can get stagnated in $\mathcal{L}_1$ (right).

.

For the objective sequence $\{f(\mathbf{w}^t)\}$ to be monotonically decreasing and $\{\mathbf{w}^k\}$ converging to a critical point, it is crucial to have the sufficient descent property [12] hold for sufficiently small learning rate $\eta > 0$:

$$f(\mathbf{w}^{t+1}) - f(\mathbf{w}^t) \leq -c\,\|\mathbf{w}^{t+1} - \mathbf{w}^t\|^2, \tag{10}$$

with some positive constant $c > 0$.

Since $\mathbf{w}^t = \arg\min_{\mathbf{w}\in\mathcal{Q}} \|\mathbf{w} - \mathbf{w}_f^t\|^2$ and $\mathbf{w}^{t+1} \in \mathcal{Q}$, it holds in (9) that

$$\frac{1}{\eta}(\|\mathbf{w}^{t+1} - \mathbf{w}_f^t\|^2 - \|\mathbf{w}^t - \mathbf{w}_f^t\|^2) \geq 0.$$

Due to non-convexity of the set $\mathcal{Q}$, the above term can be as small as zero even when $\mathbf{w}^t$ and $\mathbf{w}^{t+1}$ are distinct. So it is not guaranteed to dominate the right hand side of (9). Consequently given (9), the inequality (10) does not necessarily hold. Without sufficient descent, even if $\{f(\mathbf{w}^t)\}$ converges, the iterates $\{\mathbf{w}^t\}$ may not converge well to a critical point. To fix this issue, we blend the ideas of PGD and BC, and propose the following blended gradient descent (BGD)

$$\mathbf{w}_f^{t+1} = (1-\rho)\mathbf{w}_f^t + \rho\mathbf{w}^t - \eta\nabla f(\mathbf{w}^t),\ \mathbf{w}^{t+1} = \mathrm{proj}_{\mathcal{Q}}(\mathbf{w}_f^{t+1}) \tag{11}$$

for some blending parameter $\rho \ll 1$. In contrast, the blended gradient descent satisfies (10) for small enough $\eta$.

**Proposition 1.** *For $\rho \in (0,1)$, the BGD (11) satisfies*

$$f(\mathbf{w}^{t+1}) - f(\mathbf{w}^t) \leq -\frac{1}{2}\left(\frac{1-\rho}{\eta}(\|\mathbf{w}^{t+1} - \mathbf{w}_f^t\|^2 - \|\mathbf{w}^t - \mathbf{w}_f^t\|^2) + \left(\frac{\rho}{\eta} - L\right)\|\mathbf{w}^{t+1} - \mathbf{w}^t\|^2\right).$$

Choosing the learning rate $\eta$ small enough so that $\rho/\eta \geq L + c$. Then inequality (10) follows from the above proposition, which will guarantee the convergence of (11) to a critical point by using similar arguments as in the proofs from [35].

**Corollary 1.** *The blended gradient descent iteration (11) satisfies the sufficient descent property (10).*

## 4 Experiments

We tested BCGD, as summarized in Algorithm 1, on the CIFAR-10 [20] and ImageNet [9,21] color image datasets. We coded up the BCGD in PyTorch platform [27]. In all experiments, we fix the blending factor in (11) to be $\rho = 10^{-5}$. All runs with quantization are warm started with a float pre-trained model, and the resolutions $\boldsymbol{\alpha}$ are initialized by $\frac{1}{2^{b_a}-1}$ of the maximal values in the

corresponding feature maps generated by a random mini-batch. The learning rate for weight $\mathbf{w}$ starts from 0.01. Rate factor for the learning rate of $\boldsymbol{\alpha}$ is 0.01, i.e., the learning rate for $\boldsymbol{\alpha}$ starts from $10^{-4}$. The decay factor for the learning rates is 0.1. The weights $\mathbf{w}$ and resolutions $\boldsymbol{\alpha}$ are updated jointly. In addition, we used momentum and batch normalization [19] to promote training efficiency. We mainly compare the performances of the proposed BCGD and the state-of-the-art BC (adapted for full quantization) on *layer-wise* quantization. The experiments were carried out on machines with 4 Nvidia GeForce GTX 1080 Ti GPUs.

**Algorithm 1** One iteration of BCGD for full quantization

**Input**: mini-batch loss function $f_t(\mathbf{w}, \boldsymbol{\alpha})$, blending parameter $\rho = 10^{-5}$, learning rate $\eta_{\mathbf{w}}^t$ for the weights $\mathbf{w}$, learning rate $\eta_{\boldsymbol{\alpha}}^t$ for the resolutions $\boldsymbol{\alpha}$ of AQ (one component per activation layer).
**Do**:
Evaluate the mini-batch coarse gradient $(\tilde{\nabla}_{\mathbf{w}} f_t, \tilde{\nabla}_{\boldsymbol{\alpha}} f_t)$ at $(\mathbf{w}^t, \boldsymbol{\alpha}^t)$ according to section 2.
$\mathbf{w}_f^{t+1} = (1-\rho)\mathbf{w}_f^t + \rho \mathbf{w}^t - \eta_{\mathbf{w}}^t \tilde{\nabla}_{\mathbf{w}} f_t(\mathbf{w}^t, \boldsymbol{\alpha}^t)$ // blended gradient update for weights
$\boldsymbol{\alpha}^{t+1} = \boldsymbol{\alpha}^t - \eta_{\boldsymbol{\alpha}}^t \tilde{\nabla}_{\boldsymbol{\alpha}} f_t(\mathbf{w}^t, \boldsymbol{\alpha}^t)$ // $\eta_{\boldsymbol{\alpha}}^t = 0.01 \cdot \eta_{\mathbf{w}}^t$
$\mathbf{w}^{t+1} = \text{proj}_{\mathcal{Q}}(\mathbf{w}_f^{t+1})$ // quantize the weights as per section 3.1

The CIFAR-10 dataset consists of 60,000 32×32 color images of 10 classes, with 6,000 images per class. There dataset is split into 50,000 training images and 10,000 test images. In the experiments, we used the testing images for validation. The mini-batch size was set to be 128 and the models were trained for 200 epochs with learning rate decaying at epoch 80 and 140. In addition, we used weight decay of $10^{-4}$ and momentum of 0.95. The a.e derivative, 3-valued and 2-valued coarse derivatives of $\boldsymbol{\alpha}$ are compared on the VGG-11 [31] and ResNet-20 [14] architectures, and the results are listed in Tables 1, 2 and 3, respectively. It can be seen that the 3-valued coarse $\boldsymbol{\alpha}$ derivative gives the best overall performance in terms of accuracy. Figure 3 shows that in weight binarization, BCGD converges faster and better than BC.

| Network | Float | 32W4A | 1W4A | 2W4A | 4W4A |
|---|---|---|---|---|---|
| VGG-11 + BC | 92.13 | 91.74 | 88.12 | 89.78 | **91.51** |
| VGG-11+BCGD | | | **88.74** | **90.08** | 91.38 |
| ResNet-20 + BC | 92.41 | 91.90 | 89.23 | 90.89 | 91.53 |
| ResNet-20+BCGD | | | **90.10** | **91.15** | 91.56 |

**Table 1** CIFAR-10 validation accuracies in % with the a.e. $\boldsymbol{\alpha}$ derivative.

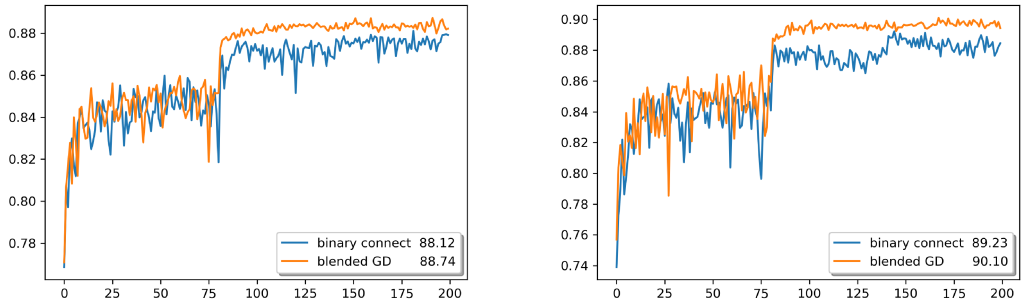


**Fig. 3** CIFAR-10 validation accuracies vs. epoch numbers with a.e. $\boldsymbol{\alpha}$ derivative and 1W4A quantization on VGG-11 (left) and ResNet-20 (right), with (orange) and without (blue) blending which speeds up training towards higher accuracies.

ImageNet (ILSVRC12) dataset [9] is a benchmark for large-scale image classification task, which has 1.2 million images for training and 50,000 for validation of 1,000 categories. We set mini-batch size to 256 and trained the models for 80 epochs with learning rate decaying at epoch 50 and 70. The weight decay of $10^{-5}$ and momentum of 0.9 were used. The ResNet-18 accuracies 65.46%/86.36% at 1W4A in Table 4 outperformed HWGQ [4] where top-1/top-5 accuracies are

| Network | Float | 32W4A | 1W4A | 2W4A | 4W4A |
|---|---|---|---|---|---|
| VGG-11 + BC | 92.13 | 92.08 | 89.12 | 90.52 | **91.89** |
| VGG-11+BCGD | | | **89.59** | **90.71** | 91.70 |
| ResNet-20 + BC | 92.41 | 92.14 | 89.37 | 91.02 | 91.71 |
| ResNet-20+BCGD | | | **90.05** | 91.03 | **91.97** |

**Table 2** CIFAR-10 validation accuracies with the 3-valued $\boldsymbol{\alpha}$ derivative.

| Network | Float | 32W4A | 1W4A | 2W4A | 4W4A |
|---|---|---|---|---|---|
| VGG-11 + BC | 92.13 | 91.66 | 88.50 | 89.99 | 91.31 |
| VGG-11+BCGD | | | **89.12** | 90.00 | 91.31 |
| ResNet-20 + BC | 92.41 | 91.73 | 89.22 | 90.64 | 91.37 |
| ResNet-20+BCGD | | | **89.98** | **90.75** | **91.65** |

**Table 3** CIFAR-10 validation accuracies with the 2-valued $\boldsymbol{\alpha}$ derivative (PACT [6]).

60.8%/83.4% with non-quantized first/last convolutional layers. The results in the Table 4 and Table 5 show that using the 3-valued coarse $\boldsymbol{\alpha}$ partial derivative appears more effective than the 2-valued as quantization bit precision is lowered. We also observe that the accuracies degrade gracefully from 4W8A to 1W4A for ResNet-18 while quantizing all convolutional layers. Again, BCGD converges much faster than BC towards higher accuracy as illustrated by Figure 4.

| | Float | 1W4A | | 4W4A | | 4W8A | |
|---|---|---|---|---|---|---|---|
| | | 3 valued | 2 valued | 3 valued | 2 valued | 3 valued | 2 valued |
| top-1 | 69.64 | 64.36/65.46* | 63.37/64.57* | 67.36 | 66.97 | 68.85 | 68.83 |
| top-5 | 88.98 | 85.65/86.36* | 84.93/85.75* | 87.76 | 87.41 | 88.71 | 88.84 |

**Table 4** ImageNet validation accuracies with BCGD on ResNet-18. Starred accuracies are with first and last convolutional layers in float precision as in [4]. The accuracies are for quantized weights across all layers otherwise.

| | Float | 1W4A | | 4W4A | | 4W8A | |
|---|---|---|---|---|---|---|---|
| | | 3 valued | 2 valued | 3 valued | 2 valued | 3 valued | 2 valued |
| top-1 | 73.27 | 68.43 | 67.51 | 70.81 | 70.01 | 72.07 | 72.18 |
| top-5 | 91.43 | 88.29 | 87.72 | 90.00 | 89.49 | 90.71 | 90.73 |

**Table 5** ImageNet validation accuracies with BCGD on ResNet-34. The accuracies are for quantized weights across all layers.

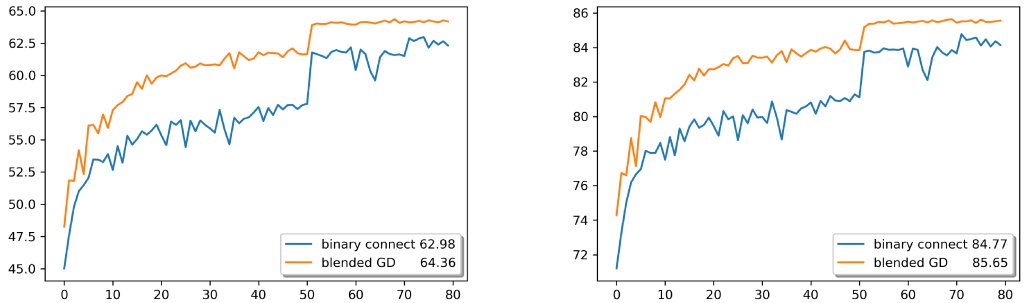


**Fig. 4** ImageNet validation accuracies (left: top-1, right: top-5) vs. number of epochs with 3-valued derivative on 1W4A quantization on ResNet-18 with (orange) and without (blue) blending which substantially speeds up training towards higher accuracies.

## 5 Analysis of Coarse Gradient Descent for Activation Quantization

As a proof of concept, we analyze a simple two-layer network with binarized ReLU activation. Let $\sigma$ be the binarized ReLU function, same as hard threshold activation [16], with the bit-width $b_a = 1$ and the resolution $\alpha \equiv 1$ in (2):

$$\sigma(x) = \begin{cases} 0 & \text{if } x \le 0, \\ 1 & \text{if } x > 0. \end{cases}$$

We define the training sample loss by

$$\ell(\mathbf{v}, \mathbf{w}; \mathbf{Z}) := \frac{1}{2}\Big(\mathbf{v}^\top \sigma(\mathbf{Z}\mathbf{w}) - (\mathbf{v}^*)^\top \sigma(\mathbf{Z}\mathbf{w}^*)\Big)^2,$$

where $\mathbf{v}^* \in \mathbb{R}^m$ and $\mathbf{w}^* \in \mathbb{R}^n$ are the underlying (nonzero) teacher parameters in the second and first layers, respectively. Same as in the literature that analyze the conventional ReLU nets [10,24, 32,3], we assume the entries of $\mathbf{Z} \in \mathbb{R}^{m\times n}$ are i.i.d. sampled from the standard normal distribution $\mathcal{N}(0,1)$. Note that $\ell(\mathbf{v}, \mathbf{w}; \mathbf{Z}) = \ell(\mathbf{v}, \mathbf{w}/c; \mathbf{Z})$ for any scalar $c > 0$. Without loss of generality, we fix $\|\mathbf{w}^*\| = 1$.

### 5.1 Population Loss Minimization

Suppose we have $N$ independent training samples $\{\mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(N)}\}$, then the associated empirical risk minimization reads

$$\min_{\mathbf{v}\in\mathbb{R}^m, \mathbf{w}\in\mathbb{R}^n} \ \frac{1}{N}\sum_{i=1}^{N} \ell(\mathbf{v}, \mathbf{w}; \mathbf{Z}^{(i)}). \tag{12}$$

The major difficulty of analysis here is that the empirical risk function in (12) is still piecewise constant and has a.e. zero partial $\mathbf{w}$ gradient. This issue can be resolved by instead considering the following population loss minimization [24,3,10,32]:

$$\min_{\mathbf{v}\in\mathbb{R}^m, \mathbf{w}\in\mathbb{R}^n} \ f(\mathbf{v}, \mathbf{w}) := \mathbb{E}_{\mathbf{Z}}\left[\ell(\mathbf{v}, \mathbf{w}; \mathbf{Z})\right]. \tag{13}$$

Specifically, in the limit $N \to \infty$, the objective function $f$ becomes favorably smooth with non-trivial gradient. For nonzero vector $\mathbf{w}$, let us define the angle between $\mathbf{w}$ and $\mathbf{w}^*$ by

$$\theta(\mathbf{w}, \mathbf{w}^*) := \arccos\left(\frac{\mathbf{w}^\top \mathbf{w}^*}{\|\mathbf{w}\|\|\mathbf{w}^*\|}\right) = \arccos\left(\frac{\mathbf{w}^\top \mathbf{w}^*}{\|\mathbf{w}\|}\right),$$

then we have

**Lemma 1.** *If every entry of $\mathbf{Z}$ is i.i.d. sampled from $\mathcal{N}(0,1)$, $\|\mathbf{w}^*\| = 1$, and $\|\mathbf{w}\| \neq 0$, then the population loss is*

$$f(\mathbf{v}, \mathbf{w}) = \frac{1}{8}\left[\mathbf{v}^\top(\mathbf{I} + \mathbf{1}\mathbf{1}^\top)\mathbf{v} - 2\mathbf{v}^\top\left(\left(1 - \frac{2}{\pi}\theta(\mathbf{w}, \mathbf{w}^*)\right)\mathbf{I} + \mathbf{1}\mathbf{1}^\top\right)\mathbf{v}^* + (\mathbf{v}^*)^\top(\mathbf{I} + \mathbf{1}\mathbf{1}^\top)\mathbf{v}^*\right]. \tag{14}$$

*Moreover, the gradients of $f(\mathbf{v}, \mathbf{w})$ w.r.t. $\mathbf{v}$ and $\mathbf{w}$ are*

$$\frac{\partial f}{\partial \mathbf{v}}(\mathbf{v}, \mathbf{w}) = \frac{1}{4}(\mathbf{I} + \mathbf{1}\mathbf{1}^\top)\mathbf{v} - \frac{1}{4}\left(\left(1 - \frac{2}{\pi}\theta(\mathbf{w}, \mathbf{w}^*)\right)\mathbf{I} + \mathbf{1}\mathbf{1}^\top\right)\mathbf{v}^* \tag{15}$$

*and*

$$\frac{\partial f}{\partial \mathbf{w}}(\mathbf{v}, \mathbf{w}) = -\frac{\mathbf{v}^\top \mathbf{v}^*}{2\pi\|\mathbf{w}\|} \frac{\left(\mathbf{I} - \frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|^2}\right)\mathbf{w}^*}{\left\|\left(\mathbf{I} - \frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|^2}\right)\mathbf{w}^*\right\|}, \quad \text{for } \theta(\mathbf{w}, \mathbf{w}^*) \in (0, \pi), \tag{16}$$

*respectively.*

When $\mathbf{w} \neq \mathbf{0}$, the possible (local) minimizers of problem (13) are located at

1. Stationary points where the gradients defined in (15) and (16) vanish simultaneously (which may not be possible), i.e.,

$$\mathbf{v}^\top \mathbf{v}^* = 0 \text{ and } \mathbf{v} = \left(\mathbf{I} + \mathbf{1}\mathbf{1}^\top\right)^{-1} \left(\left(1 - \frac{2}{\pi}\theta(\mathbf{w}, \mathbf{w}^*)\right)\mathbf{I} + \mathbf{1}\mathbf{1}^\top\right)\mathbf{v}^*. \tag{17}$$

2. Non-differentiable points where $\theta(\mathbf{w}, \mathbf{w}^*) = 0$ and $\mathbf{v} = \mathbf{v}^*$, or $\theta(\mathbf{w}, \mathbf{w}^*) = \pi$ and $\mathbf{v} = \left(\mathbf{I} + \mathbf{1}\mathbf{1}^\top\right)^{-1}(\mathbf{1}\mathbf{1}^\top - \mathbf{I})\mathbf{v}^*$.

Among them, $\{(\mathbf{v}, \mathbf{w}) : \mathbf{v} = \mathbf{v}^*, \theta(\mathbf{w}, \mathbf{w}^*) = 0\}$ are the global minimizers with $f(\mathbf{v}, \mathbf{w}) = 0$.

**Proposition 2.** *If* $(\mathbf{1}^\top \mathbf{v}^*)^2 < \frac{m+1}{2}\|\mathbf{v}^*\|^2$, *then*

$$\left\{(\mathbf{v}, \mathbf{w}) \in \mathbb{R}^{m+n} : \mathbf{v} = (\mathbf{I} + \mathbf{1}\mathbf{1}^\top)^{-1}\left(\frac{-(\mathbf{1}^\top \mathbf{v}^*)^2}{(m+1)\|\mathbf{v}^*\|^2 - (\mathbf{1}^\top \mathbf{v}^*)^2}\mathbf{I} + \mathbf{1}\mathbf{1}^\top\right)\mathbf{v}^*, \right.$$
$$\left. \theta(\mathbf{w}, \mathbf{w}^*) = \frac{\pi}{2}\frac{(m+1)\|\mathbf{v}^*\|^2}{(m+1)\|\mathbf{v}^*\|^2 - (\mathbf{1}^\top \mathbf{v}^*)^2}\right\}$$

*gives the stationary points obeying (17). Otherwise, problem (13) has no stationary points.*

The gradient of the population loss, $\left(\frac{\partial f}{\partial \mathbf{v}}, \frac{\partial f}{\partial \mathbf{w}}\right)(\mathbf{v}, \mathbf{w})$, holds Lipschitz continuity under a boundedness condition.

**Lemma 2.** *For any* $(\mathbf{v}, \mathbf{w})$ *and* $(\tilde{\mathbf{v}}, \tilde{\mathbf{w}})$ *with* $\min\{\|\mathbf{w}\|, \|\tilde{\mathbf{w}}\|\} = c > 0$ *and* $\max\{\|\mathbf{v}\|, \|\tilde{\mathbf{v}}\|\} = C$, *there exists a constant* $L > 0$ *depending on* $c$ *and* $C$, *such that*

$$\left\|\left(\frac{\partial f}{\partial \mathbf{v}}, \frac{\partial f}{\partial \mathbf{w}}\right)(\mathbf{v}, \mathbf{w}) - \left(\frac{\partial f}{\partial \mathbf{v}}, \frac{\partial f}{\partial \mathbf{w}}\right)(\tilde{\mathbf{v}}, \tilde{\mathbf{w}})\right\| \le L\|(\mathbf{v}, \mathbf{w}) - (\tilde{\mathbf{v}}, \tilde{\mathbf{w}})\|.$$

### 5.2 Convergence Analysis of Normalized Coarse Gradient Descent

The partial gradients $\frac{\partial f}{\partial \mathbf{v}}$ and $\frac{\partial f}{\partial \mathbf{w}}$, however, are not available in the training. What we really have access to are the expectations of the sample gradients, namely,

$$\mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell}{\partial \mathbf{v}}(\mathbf{v}, \mathbf{w}; \mathbf{Z})\right] \text{ and } \mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell}{\partial \mathbf{w}}(\mathbf{v}, \mathbf{w}; \mathbf{Z})\right].$$

If $\sigma$ was differentiable, then the back-propagation reads

$$\frac{\partial \ell}{\partial \mathbf{v}}(\mathbf{v}, \mathbf{w}; \mathbf{Z}) = \sigma(\mathbf{Z}\mathbf{w})\Big(\mathbf{v}^\top \sigma(\mathbf{Z}\mathbf{w}) - (\mathbf{v}^*)^\top \sigma(\mathbf{Z}\mathbf{w}^*)\Big). \tag{18}$$

and

$$\frac{\partial \ell}{\partial \mathbf{w}}(\mathbf{v}, \mathbf{w}; \mathbf{Z}) = \mathbf{Z}^\top\big(\sigma'(\mathbf{Z}\mathbf{w}) \odot \mathbf{v}\big)\Big(\mathbf{v}^\top \sigma(\mathbf{Z}\mathbf{w}) - (\mathbf{v}^*)^\top \sigma(\mathbf{Z}\mathbf{w}^*)\Big). \tag{19}$$

Now that $\sigma$ has zero derivative a.e., which makes (19) inapplicable. We study the coarse gradient descent with $\sigma'$ in (19) being replaced by the (sub)derivative $\mu'$ of regular ReLU $\mu(x) := \max(x, 0)$. More precisely, we use the following surrogate of $\frac{\partial \ell}{\partial \mathbf{w}}(\mathbf{v}, \mathbf{w}; \mathbf{Z})$:

$$\mathbf{g}(\mathbf{v}, \mathbf{w}; \mathbf{Z}) = \mathbf{Z}^\top\big(\mu'(\mathbf{Z}\mathbf{w}) \odot \mathbf{v}\big)\Big(\mathbf{v}^\top \sigma(\mathbf{Z}\mathbf{w}) - (\mathbf{v}^*)^\top \sigma(\mathbf{Z}\mathbf{w}^*)\Big) \tag{20}$$

with $\mu'(x) = \sigma(x)$, and consider the following coarse gradient descent with weight normalization:

$$\begin{cases} \mathbf{v}^{t+1} = \mathbf{v}^t - \eta \mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell}{\partial \mathbf{v}}(\mathbf{v}^t, \mathbf{w}^t; \mathbf{Z})\right] \\ \mathbf{w}^{t+\frac{1}{2}} = \mathbf{w}^t - \eta \mathbb{E}_{\mathbf{Z}}\left[\mathbf{g}(\mathbf{v}^t, \mathbf{w}^t; \mathbf{Z})\right] \\ \mathbf{w}^{t+1} = \frac{\mathbf{w}^{t+1/2}}{\|\mathbf{w}^{t+1/2}\|} \end{cases} \tag{21}$$

**Lemma 3.** *The expected gradient of* $\ell(\mathbf{v},\mathbf{w};\mathbf{Z})$ *w.r.t.* $\mathbf{v}$ *is*

$$\mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell}{\partial \mathbf{v}}(\mathbf{v},\mathbf{w};\mathbf{Z})\right]=\frac{\partial f}{\partial \mathbf{v}}(\mathbf{v},\mathbf{w})=\frac{1}{4}(\mathbf{I}+\mathbf{1}\mathbf{1}^\top)\mathbf{v}-\frac{1}{4}\left(\left(1-\frac{2}{\pi}\theta(\mathbf{w},\mathbf{w}^*)\right)\mathbf{I}+\mathbf{1}\mathbf{1}^\top\right)\mathbf{v}^*. \tag{22}$$

*The expected coarse gradient w.r.t.* $\mathbf{w}$ *is*

$$\mathbb{E}_{\mathbf{Z}}\Big[\mathbf{g}(\mathbf{v},\mathbf{w};\mathbf{Z})\Big]=\frac{h(\mathbf{v},\mathbf{v}^*)}{2\sqrt{2\pi}}\frac{\mathbf{w}}{\|\mathbf{w}\|}-\cos\left(\frac{\theta(\mathbf{w},\mathbf{w}^*)}{2}\right)\frac{\mathbf{v}^\top\mathbf{v}^*}{\sqrt{2\pi}}\frac{\frac{\mathbf{w}}{\|\mathbf{w}\|}+\mathbf{w}^*}{\left\|\frac{\mathbf{w}}{\|\mathbf{w}\|}+\mathbf{w}^*\right\|},^{2} \tag{23}$$

*where* $h(\mathbf{v},\mathbf{v}^*)=\|\mathbf{v}\|^2+(\mathbf{1}^\top\mathbf{v})^2-(\mathbf{1}^\top\mathbf{v})(\mathbf{1}^\top\mathbf{v}^*)+\mathbf{v}^\top\mathbf{v}^*$. *In particular,* $\mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell}{\partial \mathbf{v}}(\mathbf{v},\mathbf{w};\mathbf{Z})\right]$ *and* $\mathbb{E}_{\mathbf{Z}}\Big[\mathbf{g}(\mathbf{v},\mathbf{w};\mathbf{Z})\Big]$ *vanish simultaneously only in one of the following cases*

1. *(17) is satisfied according to Proposition 2.*
2. $\mathbf{v}=\mathbf{v}^*$, $\theta(\mathbf{w},\mathbf{w}^*)=0$, *or* $\mathbf{v}=(\mathbf{I}+\mathbf{1}\mathbf{1}^\top)^{-1}(\mathbf{1}\mathbf{1}^\top-\mathbf{I})\mathbf{v}^*$, $\theta(\mathbf{w},\mathbf{w}^*)=\pi$.

What is interesting is that the coarse partial gradient $\mathbb{E}_{\mathbf{Z}}\Big[\mathbf{g}(\mathbf{v},\mathbf{w};\mathbf{Z})\Big]=\mathbf{0}$ is properly defined at global minimizers of the population loss minimization problem (13) with $\mathbf{v}=\mathbf{v}^*$, $\theta(\mathbf{w},\mathbf{w}^*)=0$, whereas the true gradient $\frac{\partial f}{\partial \mathbf{w}}(\mathbf{v},\mathbf{w})$ does not exist there. Our key finding is that the coarse gradient $\mathbb{E}_{\mathbf{Z}}\Big[\mathbf{g}(\mathbf{v},\mathbf{w};\mathbf{Z})\Big]$ *has positive correlation with the true gradient* $\frac{\partial f}{\partial \mathbf{w}}(\mathbf{v},\mathbf{w})$, *and consequently,* $-\mathbb{E}_{\mathbf{Z}}\Big[\mathbf{g}(\mathbf{v},\mathbf{w};\mathbf{Z})\Big]$ *together with* $-\mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell}{\partial \mathbf{v}}(\mathbf{v},\mathbf{w};\mathbf{Z})\right]$ *give a descent direction in algorithm (21).*

**Lemma 4.** *If* $\theta(\mathbf{w},\mathbf{w}^*)\in(0,\pi)$ *, and* $\|\mathbf{w}\|\neq 0$, *then the inner product between the expected coarse and true gradients w.r.t.* $\mathbf{w}$ *is*

$$\left\langle \mathbb{E}_{\mathbf{Z}}\Big[\mathbf{g}(\mathbf{v},\mathbf{w};\mathbf{Z})\Big],\frac{\partial f}{\partial \mathbf{w}}(\mathbf{v},\mathbf{w})\right\rangle=\frac{\sin\left(\theta(\mathbf{w},\mathbf{w}^*)\right)}{2(\sqrt{2\pi})^3\|\mathbf{w}\|}(\mathbf{v}^\top\mathbf{v}^*)^2\geq 0.$$

Moreover, the following lemma asserts that $\mathbb{E}_{\mathbf{Z}}\Big[\mathbf{g}(\mathbf{v},\mathbf{w};\mathbf{Z})\Big]$ is sufficiently correlated with $\frac{\partial f}{\partial \mathbf{w}}(\mathbf{v},\mathbf{w})$, which will secure sufficient descent in objective values $\{f(\mathbf{v}^t,\mathbf{w}^t)\}$ and thus the convergence of $\{(\mathbf{v}^t,\mathbf{w}^t)\}$.

**Lemma 5.** *Suppose* $\|\mathbf{w}\|=1$ *and* $\|\mathbf{v}\|\leq C$. *There exists a constant* $A>0$ *depending on* $C$, *such that*

$$\left\|\mathbb{E}_{\mathbf{Z}}\Big[\mathbf{g}(\mathbf{v},\mathbf{w};\mathbf{Z})\Big]\right\|^2\leq A\left(\left\|\frac{\partial f}{\partial \mathbf{v}}(\mathbf{v},\mathbf{w})\right\|^2+\left\langle \mathbb{E}_{\mathbf{Z}}\Big[\mathbf{g}(\mathbf{v},\mathbf{w};\mathbf{Z})\Big],\frac{\partial f}{\partial \mathbf{w}}(\mathbf{v},\mathbf{w})\right\rangle\right).$$

Equipped with Lemma 2 and Lemma 5, we are able to show the convergence result of iteration (21).

**Theorem 1.** *Given the initialization* $(\mathbf{v}^0,\mathbf{w}^0)$ *with* $\|\mathbf{w}^0\|=1$, *and let* $\{(\mathbf{v}^t,\mathbf{w}^t)\}$ *be the sequence generated by iteration (21). There exists* $\eta_0>0$, *such that for any step size* $\eta<\eta_0$, $\{f(\mathbf{v}^t,\mathbf{w}^t)\}$ *is monotonically decreasing, and both* $\left\|\mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell}{\partial \mathbf{v}}(\mathbf{v}^t,\mathbf{w}^t;\mathbf{Z})\right]\right\|$ *and* $\left\|\mathbb{E}_{\mathbf{Z}}\Big[\mathbf{g}(\mathbf{v}^t,\mathbf{w}^t;\mathbf{Z})\Big]\right\|$ *converge to 0, as* $t\to\infty$.

**Remark 1.** *Combining the treatment of [10] for analyzing two-layer networks with regular ReLU and the positive correlation between* $\mathbb{E}_{\mathbf{Z}}\left[\mathbf{g}(\mathbf{w},\mathbf{v};\mathbf{Z})\right]$ *and* $\frac{\partial f}{\partial \mathbf{w}}(\mathbf{v},\mathbf{w})$, *one can further show that if the initialization* $(\mathbf{v}^0,\mathbf{w}^0)$ *satisfies* $(\mathbf{v}^0)^\top\mathbf{v}^*>0$, $\theta(\mathbf{w}^0,\mathbf{w}^*)<\frac{\pi}{2}$ *and* $(\mathbf{1}^\top\mathbf{v}^*)(\mathbf{1}^\top\mathbf{v}^0)\leq(\mathbf{1}^\top\mathbf{v}^*)^2$, *then* $\{(\mathbf{v}^t,\mathbf{w}^t)\}$ *converges to the global minimizer* $(\mathbf{v}^*,\mathbf{w}^*)$.

[2] We redefine the second term as $\mathbf{0}$ in the case $\theta(\mathbf{w},\mathbf{w}^*)=\pi$.

## 6 Concluding Remarks

We introduced the concept of coarse gradient for activation quantization problem of DNNs, for which the a.e. gradient is inapplicable. Coarse gradient is generally not a gradient but an artificial ascent direction. We further proposed BCGD algorithm, for training fully quantized neural networks. The weight update of BCGD goes by coarse gradient correction of a weighted average of the float weights and their quantization, which yields sufficient descent in objective and thus acceleration. Our experiments demonstrated that BCGD is very effective for quantization at extremely low bit-width such as binarization. Finally, we analyzed the coarse gradient descent for a two-layer neural network model with Gaussian input data, and proved that the expected coarse gradient essentially correlates positively with the underlying true gradient.

**Acknowledgements.** This work was partially supported by NSF grants DMS-1522383, IIS-1632935; ONR grant N00014-18-1-2527, AFOSR grant FA9550-18-0167, DOE grant DE-SC0013839 and STROBE STC NSF grant DMR-1548924.

## Conflict of Interest Statement

On behalf of all authors, the corresponding author states that there is no conflict of interest.

## References

1. Bengio, Y., Léonard, N., Courville, A.: Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432 (2013)
2. Bertsekas, D.P.: Nonlinear programming. Athena scientific Belmont (1999)
3. Brutzkus, A., Globerson, A.: Globally optimal gradient descent for a convnet with gaussian inputs. arXiv preprint arXiv:1702.07966 (2017)
4. Cai, Z., He, X., Sun, J., Vasconcelos, N.: Deep learning with low precision by half-wave gaussian quantization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
5. Carreira-Perpinán, M.: Model compression as constrained optimization, with application to neural nets. part i: General framework. arXiv preprint arXiv:1707.01209 (2017)
6. Choi, J., Wang, Z., Venkataramani, S., Chuang, P.I.J., Srinivasan, V., Gopalakrishnan, K.: Pact: Parameterized clipping activation for quantized neural networks. arXiv preprint arXiv:1805.06085 (2018)
7. Combettes, P.L., Pesquet, J.C.: Stochastic approximations and perturbations in forward-backward splitting for monotone operators. Pure and Applied Functional Analysis **1**, 13–37 (2016)
8. Courbariaux, M., Bengio, Y., David, J.: Binaryconnect: Training deep neural networks with binary weights during propagations. In: Advances in Neural Information Processing Systems (NIPS), p. 3123–3131 (2015)
9. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255 (2009)
10. Du, S.S., Lee, J.D., Tian, Y., Poczos, B., Singh, A.: Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minimum. arXiv preprint arXiv:1712.00779 (2018)
11. Freund, Y., Schapire, R.E.: Large margin classification using the perceptron algorithm. Machine learning **37**(3), 277–296 (1999)
12. Gilbert, J.C., Nocedal, J.: Global convergence properties of conjugate gradient methods for optimization. SIAM Journal on Optimization **2**(1), 21–42 (1992)
13. He, J., Li, L., Xu, J., Zheng, C.: Relu deep neural networks and linear finite elements. arXiv preprint arXiv:1807.03973 (2018)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)
15. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: IEEE International Conference on Computer Vision (ICCV) (2015)
16. Hinton, G.: Neural networks for machine learning, coursera. Coursera, video lectures (2012)
17. Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., Bengio, Y.: Binarized neural networks: Training neural networks with weights and activations constrained to +1 or -1. arXiv preprint arXiv:1602.02830 (2016)
18. Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., Bengio, Y.: Quantized neural networks: Training neural networks with low precision weights and activations. Journal of Machine Learning Research **18**, 1–30 (2018)
19. Ioffe, S., Szegedy, C.: Normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
20. Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech Report (2009)
21. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS), pp. 1097–1105 (2012)
22. Li, F., Zhang, B., Liu, B.: Ternary weight networks. arXiv preprint arXiv:1605.04711 (2016)
23. Li, H., De, S., Xu, Z., Studer, C., Samet, H., Goldstein, T.: Training quantized nets: A deeper understanding. In: NIPS, pp. 5813–5823 (2017)
24. Li, Y., Yuan, Y.: Convergence analysis of two-layer neural networks with relu activation. In: Advances in Neural Information Processing Systems, pp. 597–607 (2017)
25. Lloyd, S.: Least squares quantization in pcm. IEEE Trans. Info. Theory **28**, 129–137 (1982)

26. Park, E., Ahn, J., Yoo, S.: Weighted-entropy-based quantization for deep neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5456–5464 (2017)
27. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. Tech Report (2017)
28. Rastegari, M., Ordonez, V., Redmon, J., Farhadi, A.: Xnor-net: Imagenet classification using binary convolutional neural networks. In: European Conference on Computer Vision (ECCV) (2016)
29. Rosenblatt, F.: The perceptron, a perceiving and recognizing automaton Project Para. Cornell Aeronautical Laboratory (1957)
30. Rosenblatt, F.: Principles of neurodynamics. Spartan Book (1962)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2015)
32. Tian, Y.: An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. arXiv preprint arXiv:1703.00560 (2017)
33. Wang, B., Luo, X., Li, Z., Zhu, W., Shi, Z., Osher, S.J.: Deep neural nets with interpolating function as output activation. arXiv preprint arXiv:1802.00168 (2018)
34. Widrow, B., Lehr, M.A.: 30 years of adaptive neural networks: perceptron, madaline, and backpropagation. Proceedings of the IEEE **78**(9), 1415–1442 (1990)
35. Yin, P., Zhang, S., Lyu, J., Osher, S., Qi, Y., Xin, J.: Binaryrelax: A relaxation approach for training deep neural networks with quantized weights. arXiv preprint arXiv:1801.06313; SIAM Journal on Imaging Sciences, to appear (2018)
36. Yin, P., Zhang, S., Qi, Y., Xin, J.: Quantization and training of low bit-width convolutional neural networks for object detection. arXiv preprint arXiv:1612.06052; J. Comput. Math., to appear (2018)
37. Zhou, S., Wu, Y., Ni, Z., Zhou, X., Wen, H., Zou, Y.: Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. arXiv preprint arXiv: 1606.06160 (2016)

# Appendix

## A. Additional Preliminaries

**Lemma 6.** *Let* $\mathbf{z}$ *be a Gaussian random vector with entries i.i.d. sampled from* $\mathcal{N}(0,1)$*. Given nonzero vectors* $\mathbf{w}$ *and* $\tilde{\mathbf{w}}$ *with angle* $\theta$*, we have*

$$\mathbb{E}\left[1_{\{\mathbf{z}^\top\mathbf{w}>0\}}\right]=\frac{1}{2},\ \mathbb{E}\left[1_{\{\mathbf{z}^\top\mathbf{w}>0,\,\mathbf{z}^\top\tilde{\mathbf{w}}>0\}}\right]=\frac{\pi-\theta}{2\pi},$$

*and*

$$\mathbb{E}\left[\mathbf{z}1_{\{\mathbf{z}^\top\mathbf{w}>0\}}\right]=\frac{1}{\sqrt{2\pi}}\frac{\mathbf{w}}{\|\mathbf{w}\|},\ \mathbb{E}\left[\mathbf{z}1_{\{\mathbf{z}^\top\mathbf{w}>0,\,\mathbf{z}^\top\mathbf{w}^*>0\}}\right]=\frac{\cos(\theta/2)}{\sqrt{2\pi}}\frac{\frac{\mathbf{w}}{\|\mathbf{w}\|}+\frac{\tilde{\mathbf{w}}}{\|\tilde{\mathbf{w}}\|}}{\left\|\frac{\mathbf{w}}{\|\mathbf{w}\|}+\frac{\tilde{\mathbf{w}}}{\|\tilde{\mathbf{w}}\|}\right\|}.^{3}$$

*Proof.* The third identity was proved in Lemma A.1 of [10]. To show the first one, since Gaussian distribution is rotation-invariant, without loss of generality we assume $\mathbf{w}=[w_1,0,\mathbf{0}^\top]^\top$ with $w_1>0$, then $\mathbb{E}\left[1_{\{\mathbf{z}^\top\mathbf{w}>0\}}\right]=\mathbb{P}(z_1>0)=\frac{1}{2}$.

We further assume $\tilde{\mathbf{w}}=[\tilde{w}_1,\tilde{w}_2,\mathbf{0}^\top]^\top$. It is easy to see

$$\mathbb{E}\left[1_{\{\mathbf{z}^\top\mathbf{w}>0,\,\mathbf{z}^\top\tilde{\mathbf{w}}>0\}}\right]=\mathbb{P}(\mathbf{z}^\top\mathbf{w}>0,\,\mathbf{z}^\top\tilde{\mathbf{w}}>0)=\frac{\pi-\theta}{2\pi},$$

which is the probability that $\mathbf{z}$ forms an acute angle with both $\mathbf{w}$ and $\mathbf{w}^*$.

To prove the last identity, we use polar representation of 2-D Gaussian random variables, where $r$ is the radius and $\phi$ is the angle with $\mathrm{d}\mathbb{P}_r=r\exp(-r^2/2)\mathrm{d}r$ and $\mathrm{d}\mathbb{P}_\phi=\frac{1}{2\pi}\mathrm{d}\phi$. Then $\mathbb{E}\left[z_i1_{\{\mathbf{z}^\top\mathbf{w}>0,\,\mathbf{z}^\top\mathbf{w}^*>0\}}\right]=0$ for $i\geq 3$. Moreover,

$$\mathbb{E}\left[z_11_{\{\mathbf{z}^\top\mathbf{w}>0,\,\mathbf{z}^\top\mathbf{w}^*>0\}}\right]=\frac{1}{2\pi}\int_0^\infty r^2\exp\left(-\frac{r^2}{2}\right)\mathrm{d}r\int_{-\frac{\pi}{2}+\theta}^{\frac{\pi}{2}}\cos(\phi)\mathrm{d}\phi=\frac{1+\cos(\theta)}{2\sqrt{2\pi}}$$

and

$$\mathbb{E}\left[z_21_{\{\mathbf{z}^\top\mathbf{w}>0,\,\mathbf{z}^\top\mathbf{w}^*>0\}}\right]=\frac{1}{2\pi}\int_0^\infty r^2\exp\left(-\frac{r^2}{2}\right)\mathrm{d}r\int_{-\frac{\pi}{2}+\theta}^{\frac{\pi}{2}}\sin(\phi)\mathrm{d}\phi=\frac{\sin(\theta)}{2\sqrt{2\pi}}.$$

Therefore,

$$\mathbb{E}\left[\mathbf{z}1_{\{\mathbf{z}^\top\mathbf{w}>0,\,\mathbf{z}^\top\mathbf{w}^*>0\}}\right]=\frac{\cos(\theta/2)}{\sqrt{2\pi}}[\cos(\theta/2),\sin(\theta/2),\mathbf{0}^\top]^\top=\frac{\cos(\theta/2)}{\sqrt{2\pi}}\frac{\frac{\mathbf{w}}{\|\mathbf{w}\|}+\frac{\tilde{\mathbf{w}}}{\|\tilde{\mathbf{w}}\|}}{\left\|\frac{\mathbf{w}}{\|\mathbf{w}\|}+\frac{\tilde{\mathbf{w}}}{\|\tilde{\mathbf{w}}\|}\right\|},$$

where the last equality holds because $\frac{\mathbf{w}}{\|\mathbf{w}\|}$ and $\frac{\tilde{\mathbf{w}}}{\|\tilde{\mathbf{w}}\|}$ are two unit-normed vectors with angle $\theta$. □

**Lemma 7.** *For any nonzero vectors* $\mathbf{w}$ *and* $\tilde{\mathbf{w}}$ *with* $\|\tilde{\mathbf{w}}\|\geq\|\mathbf{w}\|=c>0$*, we have*

1. $|\theta(\mathbf{w},\mathbf{w}^*)-\theta(\tilde{\mathbf{w}},\mathbf{w}^*)|\leq\frac{\pi}{2c}\|\mathbf{w}-\tilde{\mathbf{w}}\|$.
2. $\left\|\frac{1}{\|\mathbf{w}\|}\frac{\left(\mathbf{I}-\frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|^2}\right)\mathbf{w}^*}{\left\|\left(\mathbf{I}-\frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|^2}\right)\mathbf{w}^*\right\|}-\frac{1}{\|\tilde{\mathbf{w}}\|}\frac{\left(\mathbf{I}-\frac{\tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top}{\|\tilde{\mathbf{w}}\|^2}\right)\mathbf{w}^*}{\left\|\left(\mathbf{I}-\frac{\tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top}{\|\tilde{\mathbf{w}}\|^2}\right)\mathbf{w}^*\right\|}\right\|\leq\frac{1}{c^2}\|\mathbf{w}-\tilde{\mathbf{w}}\|$.

*Proof.* 1. Since by Cauchy-Schwarz inequality,

$$\left\langle\tilde{\mathbf{w}},\mathbf{w}-\frac{c\tilde{\mathbf{w}}}{\|\tilde{\mathbf{w}}\|}\right\rangle=\tilde{\mathbf{w}}^\top\mathbf{w}-c\|\tilde{\mathbf{w}}\|\leq 0,$$

we have

$$\begin{aligned}\|\tilde{\mathbf{w}}-\mathbf{w}\|^2&=\left\|\left(1-\frac{c}{\|\tilde{\mathbf{w}}\|}\right)\tilde{\mathbf{w}}-\left(\mathbf{w}-\frac{c\tilde{\mathbf{w}}}{\|\tilde{\mathbf{w}}\|}\right)\right\|^2\geq\left\|\left(1-\frac{c}{\|\tilde{\mathbf{w}}\|}\right)\tilde{\mathbf{w}}\right\|^2+\left\|\mathbf{w}-\frac{c\tilde{\mathbf{w}}}{\|\tilde{\mathbf{w}}\|}\right\|^2\\&\geq\left\|\mathbf{w}-\frac{c\tilde{\mathbf{w}}}{\|\tilde{\mathbf{w}}\|}\right\|^2=c^2\left\|\frac{\mathbf{w}}{\|\mathbf{w}\|}-\frac{\tilde{\mathbf{w}}}{\|\tilde{\mathbf{w}}\|}\right\|^2.\end{aligned}\tag{24}$$

[3] Same as in Lemma 3, we redefine $\mathbb{E}\left[\mathbf{z}1_{\{\mathbf{z}^\top\mathbf{w}>0,\,\mathbf{z}^\top\mathbf{w}^*>0\}}\right]=\mathbf{0}$ in the case $\theta(\mathbf{w},\mathbf{w}^*)=\pi$.

Therefore,

$$|\theta(\mathbf{w},\mathbf{w}^*)-\theta(\tilde{\mathbf{w}},\mathbf{w}^*)| \le \theta(\mathbf{w},\tilde{\mathbf{w}}) = \theta\left(\frac{\mathbf{w}}{\|\mathbf{w}\|},\frac{\tilde{\mathbf{w}}}{\|\tilde{\mathbf{w}}\|}\right)$$
$$\le \pi \sin\left(\frac{\theta\left(\frac{\mathbf{w}}{\|\mathbf{w}\|},\frac{\tilde{\mathbf{w}}}{\|\tilde{\mathbf{w}}\|}\right)}{2}\right) = \frac{\pi}{2}\left\|\frac{\mathbf{w}}{\|\mathbf{w}\|}-\frac{\tilde{\mathbf{w}}}{\|\tilde{\mathbf{w}}\|}\right\| \le \frac{\pi}{2c}\|\mathbf{w}-\tilde{\mathbf{w}}\|,$$

where we used the fact $\sin(x) \ge \frac{2x}{\pi}$ for $x \in [0, \frac{\pi}{2}]$ and the estimate in (24).

2. Since $\left(\mathbf{I}-\frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|^2}\right)\mathbf{w}^*$ is the projection of $\mathbf{w}^*$ onto the complement space of $\mathbf{w}$, and likewise for $\left(\mathbf{I}-\frac{\tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top}{\|\tilde{\mathbf{w}}\|^2}\right)\mathbf{w}^*$, the angle between $\left(\mathbf{I}-\frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|^2}\right)\mathbf{w}^*$ and $\left(\mathbf{I}-\frac{\tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top}{\|\tilde{\mathbf{w}}\|^2}\right)\mathbf{w}^*$ is equal to the angle between $\mathbf{w}$ and $\tilde{\mathbf{w}}$. Therefore,

$$\left\langle \frac{\left(\mathbf{I}-\frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|^2}\right)\mathbf{w}^*}{\left\|\left(\mathbf{I}-\frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|^2}\right)\mathbf{w}^*\right\|}, \frac{\left(\mathbf{I}-\frac{\tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top}{\|\tilde{\mathbf{w}}\|^2}\right)\mathbf{w}^*}{\left\|\left(\mathbf{I}-\frac{\tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top}{\|\tilde{\mathbf{w}}\|^2}\right)\mathbf{w}^*\right\|} \right\rangle = \left\langle \frac{\mathbf{w}}{\|\mathbf{w}\|}, \frac{\tilde{\mathbf{w}}}{\|\tilde{\mathbf{w}}\|} \right\rangle,$$

and thus

$$\left\| \frac{1}{\|\mathbf{w}\|}\frac{\left(\mathbf{I}-\frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|^2}\right)\mathbf{w}^*}{\left\|\left(\mathbf{I}-\frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|^2}\right)\mathbf{w}^*\right\|} - \frac{1}{\|\tilde{\mathbf{w}}\|}\frac{\left(\mathbf{I}-\frac{\tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top}{\|\tilde{\mathbf{w}}\|^2}\right)\mathbf{w}^*}{\left\|\left(\mathbf{I}-\frac{\tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top}{\|\tilde{\mathbf{w}}\|^2}\right)\mathbf{w}^*\right\|} \right\| = \left\| \frac{\mathbf{w}}{\|\mathbf{w}\|^2} - \frac{\tilde{\mathbf{w}}}{\|\tilde{\mathbf{w}}\|^2} \right\| = \frac{\|\mathbf{w}-\tilde{\mathbf{w}}\|}{\|\mathbf{w}\|\|\tilde{\mathbf{w}}\|} \le \frac{1}{c^2}\|\mathbf{w}-\tilde{\mathbf{w}}\|.$$

The second equality above holds because

$$\left\| \frac{\mathbf{w}}{\|\mathbf{w}\|^2} - \frac{\tilde{\mathbf{w}}}{\|\tilde{\mathbf{w}}\|^2} \right\|^2 = \frac{1}{\|\mathbf{w}\|^2} + \frac{1}{\|\tilde{\mathbf{w}}\|^2} - \frac{2\langle \mathbf{w}, \tilde{\mathbf{w}}\rangle}{\|\mathbf{w}\|^2\|\tilde{\mathbf{w}}\|^2} = \frac{\|\mathbf{w}-\tilde{\mathbf{w}}\|^2}{\|\mathbf{w}\|^2\|\tilde{\mathbf{w}}\|^2}.$$

□

## B. Proofs

**Proof of Proposition 1.** We rewrite the update (11) as

$$\mathbf{w}^{t+1} = \arg\min_{\mathbf{w}\in\mathcal{Q}} \langle \mathbf{w}, \nabla f(\mathbf{w}^t)\rangle + \frac{1-\rho}{2\eta}\|\mathbf{w}-\mathbf{w}_f^t\|^2 + \frac{\rho}{2\eta}\|\mathbf{w}-\mathbf{w}^t\|^2.$$

Then since $\mathbf{w}^t$, $\mathbf{w}^{t+1} \in \mathcal{Q}$, we have

$$\langle \mathbf{w}^{t+1}, \nabla f(\mathbf{w}^t)\rangle + \frac{1-\rho}{2\eta}\|\mathbf{w}^{t+1}-\mathbf{w}_f^t\|^2 + \frac{\rho}{2\eta}\|\mathbf{w}^{t+1}-\mathbf{w}^t\|^2 \le \langle \mathbf{w}^t, \nabla f(\mathbf{w}^t)\rangle + \frac{1-\rho}{2\eta}\|\mathbf{w}^t-\mathbf{w}_f^t\|^2,$$

or equivalently,

$$\langle \mathbf{w}^{t+1}-\mathbf{w}^t, \nabla f(\mathbf{w}^t)\rangle + \frac{1-\rho}{2\eta}\left(\|\mathbf{w}^{t+1}-\mathbf{w}_f^t\|^2 - \|\mathbf{w}^t-\mathbf{w}_f^t\|^2\right) + \frac{\rho}{2\eta}\|\mathbf{w}^{t+1}-\mathbf{w}^t\|^2 \le 0. \tag{25}$$

On the other hand, since $f$ has $L$-Lipschitz gradient, the descent lemma [2] gives

$$f(\mathbf{w}^{t+1}) \le f(\mathbf{w}^t) + \langle \nabla f(\mathbf{w}^t), \mathbf{w}^{t+1}-\mathbf{w}^t\rangle + \frac{L}{2}\|\mathbf{w}^{t+1}-\mathbf{w}^t\|^2. \tag{26}$$

Combining (25) and (26) completes the proof. □

**Proof of Lemma 1.** We first evaluate $\mathbb{E}_{\mathbf{Z}}\left[\sigma(\mathbf{Z}\mathbf{w})\sigma(\mathbf{Z}\mathbf{w})^\top\right]$, $\mathbb{E}_{\mathbf{Z}}\left[\sigma(\mathbf{Z}\mathbf{w})\sigma(\mathbf{Z}\mathbf{w}^*)^\top\right]$, and $\mathbb{E}_{\mathbf{Z}}\left[\sigma(\mathbf{Z}\mathbf{w}^*)\sigma(\mathbf{Z}\mathbf{w}^*)^\top\right]$. Let $\mathbf{Z}_i^\top$ be the $i$-th row vector of $\mathbf{Z}$. Since $\mathbf{w}\neq\mathbf{0}$, using Lemma 6, we have

$$\mathbb{E}_{\mathbf{Z}}\left[\sigma(\mathbf{Z}\mathbf{w})\sigma(\mathbf{Z}\mathbf{w})^\top\right]_{ii}=\mathbb{E}\left[\sigma(\mathbf{Z}_i^\top\mathbf{w})\sigma(\mathbf{Z}_i^\top\mathbf{w})\right]=\mathbb{E}\left[1_{\{\mathbf{Z}_i^\top\mathbf{w}>0\}}\right]=\frac{1}{2},$$

and for $i\neq j$,

$$\mathbb{E}_{\mathbf{Z}}\left[\sigma(\mathbf{Z}\mathbf{w})\sigma(\mathbf{Z}\mathbf{w})^\top\right]_{ij}=\mathbb{E}\left[\sigma(\mathbf{Z}_i^\top\mathbf{w})\sigma(\mathbf{Z}_j^\top\mathbf{w})\right]=\mathbb{E}\left[1_{\{\mathbf{Z}_i^\top\mathbf{w}>0\}}\right]\mathbb{E}\left[1_{\{\mathbf{Z}_j^\top\mathbf{w}>0\}}\right]=\frac{1}{4}.$$

Therefore, $\mathbb{E}_{\mathbf{Z}}\left[\sigma(\mathbf{Z}\mathbf{w})\sigma(\mathbf{Z}\mathbf{w})^\top\right]=\mathbb{E}_{\mathbf{Z}}\left[\sigma(\mathbf{Z}\mathbf{w}^*)\sigma(\mathbf{Z}\mathbf{w}^*)^\top\right]=\frac{1}{4}\left(\mathbf{I}+\mathbf{1}\mathbf{1}^\top\right)$. Furthermore,

$$\mathbb{E}_{\mathbf{Z}}\left[\sigma(\mathbf{Z}\mathbf{w})\sigma(\mathbf{Z}\mathbf{w}^*)^\top\right]_{ii}=\mathbb{E}\left[1_{\{\mathbf{Z}_i^\top\mathbf{w}>0,\mathbf{Z}_i^\top\mathbf{w}^*>0\}}\right]=\frac{\pi-\theta(\mathbf{w},\mathbf{w}^*)}{2\pi},$$

and $\mathbb{E}_{\mathbf{Z}}\left[\sigma(\mathbf{Z}\mathbf{w})\sigma(\mathbf{Z}\mathbf{w}^*)^\top\right]_{ij}=\frac{1}{4}$. So,

$$\mathbb{E}_{\mathbf{Z}}\left[\sigma(\mathbf{Z}\mathbf{w})\sigma(\mathbf{Z}\mathbf{w}^*)^\top\right]=\frac{1}{4}\left(\left(1-\frac{2\theta(\mathbf{w},\mathbf{w}^*)}{\pi}\right)\mathbf{I}+\mathbf{1}\mathbf{1}^\top\right).$$

We thus have proved (14) by noticing that

$$\begin{aligned}f(\mathbf{v},\mathbf{w})=\frac{1}{2}\big(&\mathbf{v}^\top\mathbb{E}_{\mathbf{Z}}[\sigma(\mathbf{Z}\mathbf{w})^\top\sigma(\mathbf{Z}\mathbf{w})]\mathbf{v}-2\mathbf{v}^\top\mathbb{E}_{\mathbf{Z}}[\sigma(\mathbf{Z}\mathbf{w})^\top\sigma(\mathbf{Z}\mathbf{w}^*)]\mathbf{v}^*\\&+(\mathbf{v}^*)^\top\mathbb{E}_{\mathbf{Z}}[\sigma(\mathbf{Z}\mathbf{w}^*)^\top\sigma(\mathbf{Z}\mathbf{w}^*)]\mathbf{v}^*\big).\end{aligned}$$

Next, since (15) is trivial, we only show (16). Since $\theta(\mathbf{w},\mathbf{w}^*)=\arccos\left(\frac{\mathbf{w}^\top\mathbf{w}^*}{\|\mathbf{w}\|}\right)$ is differentiable w.r.t. $\mathbf{w}$ at $\theta(\mathbf{w},\mathbf{w}^*)\in(0,\pi)$, we have

$$\frac{\partial f}{\partial\mathbf{w}}(\mathbf{v},\mathbf{w})=\frac{\mathbf{v}^\top\mathbf{v}^*}{2\pi}\frac{\partial\theta}{\partial\mathbf{w}}(\mathbf{w},\mathbf{w}^*)=-\frac{\mathbf{v}^\top\mathbf{v}^*}{2\pi}\frac{\|\mathbf{w}\|^2\mathbf{w}^*-(\mathbf{w}^\top\mathbf{w}^*)\mathbf{w}}{\|\mathbf{w}\|^3\sqrt{1-\frac{(\mathbf{w}^\top\mathbf{w}^*)^2}{\|\mathbf{w}\|^2}}}=-\frac{\mathbf{v}^\top\mathbf{v}^*}{2\pi\|\mathbf{w}\|}\frac{\left(\mathbf{I}-\frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|^2}\right)\mathbf{w}^*}{\left\|\left(\mathbf{I}-\frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|^2}\right)\mathbf{w}^*\right\|}.$$

□

**Proof of Proposition 2.** Suppose $\mathbf{v}^\top\mathbf{v}^*=0$ and $\frac{\partial f}{\partial\mathbf{v}}(\mathbf{v},\mathbf{w})=\mathbf{0}$, then by Lemma 1,

$$0=\mathbf{v}^\top\mathbf{v}^*=(\mathbf{v}^*)^\top(\mathbf{I}+\mathbf{1}\mathbf{1}^\top)^{-1}\left(\left(1-\frac{2}{\pi}\theta(\mathbf{w},\mathbf{w}^*)\right)\mathbf{I}+\mathbf{1}\mathbf{1}^\top\right)\mathbf{v}^*. \tag{27}$$

From (27) it follows that

$$\frac{2}{\pi}\theta(\mathbf{w},\mathbf{w}^*)(\mathbf{v}^*)^\top(\mathbf{I}+\mathbf{1}\mathbf{1}^\top)^{-1}\mathbf{v}^*=(\mathbf{v}^*)^\top(\mathbf{I}+\mathbf{1}\mathbf{1}^\top)^{-1}\left(\mathbf{I}+\mathbf{1}\mathbf{1}^\top\right)\mathbf{v}^*=\|\mathbf{v}^*\|^2. \tag{28}$$

On the other hand, from (27) it also follows that

$$\left(\frac{2}{\pi}\theta(\mathbf{w},\mathbf{w}^*)-1\right)(\mathbf{v}^*)^\top(\mathbf{I}+\mathbf{1}\mathbf{1}^\top)^{-1}\mathbf{v}^*=(\mathbf{v}^*)^\top(\mathbf{I}+\mathbf{1}\mathbf{1}^\top)^{-1}\mathbf{1}(\mathbf{1}^\top\mathbf{v}^*)=\frac{(\mathbf{1}^\top\mathbf{v}^*)^2}{m+1},$$

where $\mathbf{I}$ is an $m$-by-$m$ identity matrix, and we used $(\mathbf{I}+\mathbf{1}\mathbf{1}^\top)\mathbf{1}=(m+1)\mathbf{1}$. Taking the difference of the two equalities above gives

$$(\mathbf{v}^*)^\top(\mathbf{I}+\mathbf{1}\mathbf{1}^\top)^{-1}\mathbf{v}^*=\|\mathbf{v}^*\|^2-\frac{(\mathbf{1}^\top\mathbf{v}^*)^2}{m+1}.$$

By (28), we have $\theta(\mathbf{w},\mathbf{w}^*)=\frac{\pi}{2}\frac{(m+1)\|\mathbf{v}^*\|^2}{(m+1)\|\mathbf{v}^*\|^2-(\mathbf{1}^\top\mathbf{v}^*)^2}$, which requires

$$\frac{\pi}{2}\frac{(m+1)\|\mathbf{v}^*\|^2}{(m+1)\|\mathbf{v}^*\|^2-(\mathbf{1}^\top\mathbf{v}^*)^2}<\pi,\ \text{or equivalently,}\ (\mathbf{1}^\top\mathbf{v}^*)^2<\frac{m+1}{2}\|\mathbf{v}^*\|^2.$$

Otherwise, $\frac{\partial f}{\partial\mathbf{v}}(\mathbf{v},\mathbf{w})$ and $\frac{\partial f}{\partial\mathbf{w}}(\mathbf{v},\mathbf{w})$ do not vanish simultaneously, and there is no critical point.

□

**Proof of Lemma 2.** It is easy to check that $\|\mathbf{I}+\mathbf{1}\mathbf{1}^\top\| = m+1$. Invoking Lemma 7.1 gives

$$\begin{aligned}\left\|\frac{\partial f}{\partial \mathbf{v}}(\mathbf{v},\mathbf{w}) - \frac{\partial f}{\partial \mathbf{v}}(\tilde{\mathbf{v}},\tilde{\mathbf{w}})\right\| &= \frac{1}{4}\left\|\left(\mathbf{I}+\mathbf{1}\mathbf{1}^\top\right)(\mathbf{v}-\tilde{\mathbf{v}}) + \frac{2}{\pi}(\theta(\mathbf{w},\mathbf{w}^*)-\theta(\tilde{\mathbf{w}},\mathbf{w}^*))\mathbf{v}^*\right\| \\ &\le \frac{1}{4}\left((m+1)\|\mathbf{v}-\tilde{\mathbf{v}}\| + \frac{2\|\mathbf{v}^*\|}{\pi}|\theta(\mathbf{w},\mathbf{w}^*)-\theta(\tilde{\mathbf{w}},\mathbf{w}^*)|\right) \\ &\le \frac{1}{4}\left((m+1)\|\mathbf{v}-\tilde{\mathbf{v}}\| + \frac{\|\mathbf{v}^*\|}{c}\|\mathbf{w}-\tilde{\mathbf{w}}\|\right) \\ &\le \frac{1}{4}\left(m+1+\frac{\|\mathbf{v}^*\|}{c}\right)\|(\mathbf{v},\mathbf{w})-(\tilde{\mathbf{v}},\tilde{\mathbf{w}})\|.\end{aligned}$$

Using Lemma 7.2, we further have

$$\begin{aligned}\left\|\frac{\partial f}{\partial \mathbf{w}}(\mathbf{v},\mathbf{w}) - \frac{\partial f}{\partial \mathbf{w}}(\tilde{\mathbf{v}},\tilde{\mathbf{w}})\right\| &= \left\|\frac{\mathbf{v}^\top\mathbf{v}^*}{2\pi\|\mathbf{w}\|}\frac{\left(\mathbf{I}-\frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|^2}\right)\mathbf{w}^*}{\left\|\left(\mathbf{I}-\frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|^2}\right)\mathbf{w}^*\right\|} - \frac{\tilde{\mathbf{v}}^\top\mathbf{v}^*}{2\pi\|\tilde{\mathbf{w}}\|}\frac{\left(\mathbf{I}-\frac{\tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top}{\|\tilde{\mathbf{w}}\|^2}\right)\mathbf{w}^*}{\left\|\left(\mathbf{I}-\frac{\tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top}{\|\tilde{\mathbf{w}}\|^2}\right)\mathbf{w}^*\right\|}\right\| \\ &\le \left\|\frac{\mathbf{v}^\top\mathbf{v}^*}{2\pi\|\mathbf{w}\|}\frac{\left(\mathbf{I}-\frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|^2}\right)\mathbf{w}^*}{\left\|\left(\mathbf{I}-\frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|^2}\right)\mathbf{w}^*\right\|} - \frac{\mathbf{v}^\top\mathbf{v}^*}{2\pi\|\tilde{\mathbf{w}}\|}\frac{\left(\mathbf{I}-\frac{\tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top}{\|\tilde{\mathbf{w}}\|^2}\right)\mathbf{w}^*}{\left\|\left(\mathbf{I}-\frac{\tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top}{\|\tilde{\mathbf{w}}\|^2}\right)\mathbf{w}^*\right\|}\right\| \\ &\quad + \left\|\frac{\mathbf{v}^\top\mathbf{v}^*}{2\pi\|\tilde{\mathbf{w}}\|}\frac{\left(\mathbf{I}-\frac{\tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top}{\|\tilde{\mathbf{w}}\|^2}\right)\mathbf{w}^*}{\left\|\left(\mathbf{I}-\frac{\tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top}{\|\tilde{\mathbf{w}}\|^2}\right)\mathbf{w}^*\right\|} - \frac{\tilde{\mathbf{v}}^\top\mathbf{v}^*}{2\pi\|\tilde{\mathbf{w}}\|}\frac{\left(\mathbf{I}-\frac{\tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top}{\|\tilde{\mathbf{w}}\|^2}\right)\mathbf{w}^*}{\left\|\left(\mathbf{I}-\frac{\tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top}{\|\tilde{\mathbf{w}}\|^2}\right)\mathbf{w}^*\right\|}\right\| \\ &\le \frac{|\mathbf{v}^\top\mathbf{v}^*|}{2\pi c^2}\|\mathbf{w}-\tilde{\mathbf{w}}\| + \frac{\|\mathbf{v}^*\|}{2\pi c}\|\mathbf{v}-\tilde{\mathbf{v}}\| \\ &\le \frac{(C+c)\|\mathbf{v}^*\|}{2\pi c^2}\|(\mathbf{v},\mathbf{w})-(\tilde{\mathbf{v}},\tilde{\mathbf{w}})\|.\end{aligned}$$

Combining the two inequalities above validates the claim. □

**Proof of Lemma 3.** (22) is true because $\frac{\partial \ell}{\partial \mathbf{v}}(\mathbf{v},\mathbf{w};\mathbf{Z})$ is linear in $\mathbf{v}$. To show (23), by (20) and the fact that $\mu' = \sigma$, we have

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}}\left[\mathbf{g}(\mathbf{v},\mathbf{w};\mathbf{Z})\right] &= \mathbb{E}_{\mathbf{Z}}\left[\left(\sum_{i=1}^m v_i\sigma(\mathbf{Z}_i^\top\mathbf{w}) - \sum_{i=1}^m v_i^*\sigma(\mathbf{Z}_i^\top\mathbf{w}^*)\right)\left(\sum_{i=1}^m \mathbf{Z}_i v_i\sigma(\mathbf{Z}_i^\top\mathbf{w})\right)\right] \\ &= \mathbb{E}_{\mathbf{Z}}\left[\left(\sum_{i=1}^m v_i 1_{\{\mathbf{Z}_i^\top\mathbf{w}>0\}} - \sum_{i=1}^m v_i^* 1_{\{\mathbf{Z}_i^\top\mathbf{w}^*>0\}}\right)\left(\sum_{i=1}^m 1_{\{\mathbf{Z}_i^\top\mathbf{w}>0\}} v_i\mathbf{Z}_i\right)\right].\end{aligned}$$

Invoking Lemma 6, we have

$$\mathbb{E}\left[\mathbf{Z}_i 1_{\{\mathbf{Z}_i^\top\mathbf{w}>0,\mathbf{Z}_j^\top\mathbf{w}>0\}}\right] = \begin{cases}\frac{1}{\sqrt{2\pi}}\frac{\mathbf{w}}{\|\mathbf{w}\|} & \text{if } i=j, \\ \frac{1}{2\sqrt{2\pi}}\frac{\mathbf{w}}{\|\mathbf{w}\|} & \text{if } i\neq j,\end{cases} \tag{29}$$

and

$$\mathbb{E}\left[\mathbf{Z}_i 1_{\{\mathbf{Z}_i^\top\mathbf{w}>0,\mathbf{Z}_j^\top\mathbf{w}^*>0\}}\right] = \begin{cases}\frac{\cos(\theta(\mathbf{w},\mathbf{w}^*)/2)}{\sqrt{2\pi}}\frac{\frac{\mathbf{w}}{\|\mathbf{w}\|}+\mathbf{w}^*}{\left\|\frac{\mathbf{w}}{\|\mathbf{w}\|}+\mathbf{w}^*\right\|} & \text{if } i=j, \\ \frac{1}{2\sqrt{2\pi}}\frac{\mathbf{w}}{\|\mathbf{w}\|} & \text{if } i\neq j.\end{cases} \tag{30}$$

Therefore,

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}}\left[\mathbf{g}(\mathbf{v},\mathbf{w};\mathbf{Z})\right] &= \sum_{i=1}^m v_i^2\mathbb{E}\left[\mathbf{Z}_i 1_{\{\mathbf{Z}_i^\top\mathbf{w}>0\}}\right] + \sum_{i=1}^m\sum_{\substack{j=1\\j\neq i}}^m v_iv_j\mathbb{E}\left[\mathbf{Z}_i 1_{\{\mathbf{Z}_i^\top\mathbf{w}>0,\mathbf{Z}_j^\top\mathbf{w}>0\}}\right]\\ &\quad - \sum_{i=1}^m v_iv_i^*\mathbb{E}\left[\mathbf{Z}_i 1_{\{\mathbf{Z}_i^\top\mathbf{w}>0,\mathbf{Z}_i^\top\mathbf{w}^*>0\}}\right] - \sum_{i=1}^m\sum_{\substack{j=1\\j\neq i}}^m v_iv_j^*\mathbb{E}\left[\mathbf{Z}_i 1_{\{\mathbf{Z}_i^\top\mathbf{w}>0,\mathbf{Z}_j^\top\mathbf{w}^*>0\}}\right]\\ &= \frac{1}{2\sqrt{2\pi}}\left(\|\mathbf{v}\|^2+(\mathbf{1}^\top\mathbf{v})^2\right)\frac{\mathbf{w}}{\|\mathbf{w}\|} - \cos\left(\frac{\theta(\mathbf{w},\mathbf{w}^*)}{2}\right)\frac{\mathbf{v}^\top\mathbf{v}^*}{\sqrt{2\pi}}\frac{\frac{\mathbf{w}}{\|\mathbf{w}\|}+\mathbf{w}^*}{\left\|\frac{\mathbf{w}}{\|\mathbf{w}\|}+\mathbf{w}^*\right\|}\\ &\quad - \frac{1}{2\sqrt{2\pi}}\left((\mathbf{1}^\top\mathbf{v})(\mathbf{1}^\top\mathbf{v}^*)-\mathbf{v}^\top\mathbf{v}^*\right)\frac{\mathbf{w}}{\|\mathbf{w}\|},\end{aligned}$$

which is exactly (23). □

**Proof of Lemma 4.** Notice that $\left(\mathbf{I}-\frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|^2}\right)\mathbf{w}=\mathbf{0}$ and $\|\mathbf{w}^*\|=1$, if $\theta(\mathbf{w},\mathbf{w}_*)\neq 0,\pi$, then we have

$$\begin{aligned}&\left\langle \mathbb{E}_{\mathbf{Z}}\Big[\mathbf{g}(\mathbf{v},\mathbf{w};\mathbf{Z})\Big], \frac{\partial f}{\partial \mathbf{w}}(\mathbf{v},\mathbf{w})\right\rangle\\ =\ &\cos\left(\frac{\theta(\mathbf{w},\mathbf{w}^*)}{2}\right)\frac{(\mathbf{v}^\top\mathbf{v}^*)^2}{(\sqrt{2\pi})^3}\left\langle\frac{1}{\|\mathbf{w}\|}\frac{\left(\mathbf{I}-\frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|^2}\right)\mathbf{w}^*}{\left\|\left(\mathbf{I}-\frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|^2}\right)\mathbf{w}^*\right\|}, \frac{\mathbf{w}^*}{\left\|\frac{\mathbf{w}}{\|\mathbf{w}\|}+\mathbf{w}^*\right\|}\right\rangle\\ =\ &\cos\left(\frac{\theta(\mathbf{w},\mathbf{w}^*)}{2}\right)\frac{(\mathbf{v}^\top\mathbf{v}^*)^2}{(\sqrt{2\pi})^3}\frac{\|\mathbf{w}\|^2-(\mathbf{w}^\top\mathbf{w}^*)^2}{\|\|\mathbf{w}\|^2\mathbf{w}^*-\mathbf{w}(\mathbf{w}^\top\mathbf{w}^*)\|\,\|\mathbf{w}+\|\mathbf{w}\|\mathbf{w}^*\|}\\ =\ &\cos\left(\frac{\theta(\mathbf{w},\mathbf{w}^*)}{2}\right)\frac{(\mathbf{v}^\top\mathbf{v}^*)^2}{(\sqrt{2\pi})^3}\frac{\|\mathbf{w}\|^2-(\mathbf{w}^\top\mathbf{w}^*)^2}{\sqrt{\|\mathbf{w}\|^4-\|\mathbf{w}\|^2(\mathbf{w}^\top\mathbf{w}^*)^2}\sqrt{2(\|\mathbf{w}\|^2+\|\mathbf{w}\|(\mathbf{w}^\top\mathbf{w}^*))}}\\ =\ &\cos\left(\frac{\theta(\mathbf{w},\mathbf{w}^*)}{2}\right)\frac{(\mathbf{v}^\top\mathbf{v}^*)^2}{4(\sqrt{\pi\|\mathbf{w}\|})^3}\frac{\|\mathbf{w}\|^2-(\mathbf{w}^\top\mathbf{w}^*)^2}{\sqrt{\|\mathbf{w}\|^2-(\mathbf{w}^\top\mathbf{w}^*)^2}\sqrt{\|\mathbf{w}\|+(\mathbf{w}^\top\mathbf{w}^*)}}\\ =\ &\cos\left(\frac{\theta(\mathbf{w},\mathbf{w}^*)}{2}\right)\frac{(\mathbf{v}^\top\mathbf{v}^*)^2\sqrt{1-\frac{\mathbf{w}^\top\mathbf{w}^*}{\|\mathbf{w}\|}}}{4(\sqrt{\pi})^3\|\mathbf{w}\|}\\ =\ &\cos\left(\frac{\theta(\mathbf{w},\mathbf{w}^*)}{2}\right)\frac{(\mathbf{v}^\top\mathbf{v}^*)^2\sqrt{1-\cos(\theta(\mathbf{w},\mathbf{w}^*))}}{4(\sqrt{\pi})^3\|\mathbf{w}\|}\\ =\ &\frac{\sin\left(\theta(\mathbf{w},\mathbf{w}^*)\right)}{2(\sqrt{2\pi})^3\|\mathbf{w}\|}(\mathbf{v}^\top\mathbf{v}^*)^2.\end{aligned}$$

□

**Proof of Lemma 5.** Denote $\theta:=\theta(\mathbf{w},\mathbf{w}^*)$. By Lemma 1, we have

$$\frac{\partial f}{\partial \mathbf{v}}(\mathbf{v},\mathbf{w}) = \frac{1}{4}(\mathbf{I}+\mathbf{1}\mathbf{1}^\top)\mathbf{v} - \frac{1}{4}\left(\left(1-\frac{2\theta}{\pi}\right)\mathbf{I}+\mathbf{1}\mathbf{1}^\top\right)\mathbf{v}^*.$$

Since $\|\mathbf{w}\|=1$, Lemma 3 gives

$$\mathbb{E}_{\mathbf{Z}}\Big[\mathbf{g}(\mathbf{v},\mathbf{w};\mathbf{Z})\Big] = \frac{h(\mathbf{v},\mathbf{v}^*)}{2\sqrt{2\pi}}\mathbf{w} - \cos\left(\frac{\theta}{2}\right)\frac{\mathbf{v}^\top\mathbf{v}^*}{\sqrt{2\pi}}\frac{\mathbf{w}+\mathbf{w}^*}{\|\mathbf{w}+\mathbf{w}^*\|}, \tag{31}$$

where

$$\begin{aligned} h(\mathbf{v},\mathbf{v}^*) &= \|\mathbf{v}\|^2 + (\mathbf{1}^\top\mathbf{v})^2 - (\mathbf{1}^\top\mathbf{v})(\mathbf{1}^\top\mathbf{v}^*) + \mathbf{v}^\top\mathbf{v}^* \\ &= \mathbf{v}^\top\left(\mathbf{I}+\mathbf{1}\mathbf{1}^\top\right)\mathbf{v} - \mathbf{v}^\top(\mathbf{1}\mathbf{1}^\top - \mathbf{I})\mathbf{v}^* \\ &= \mathbf{v}^\top\left(\mathbf{I}+\mathbf{1}\mathbf{1}^\top\right)\mathbf{v} - \mathbf{v}^\top\left(\mathbf{1}\mathbf{1}^\top + \left(1-\frac{2\theta}{\pi}\right)\mathbf{I}\right)\mathbf{v}^* + 2\left(1-\frac{\theta}{\pi}\right)\mathbf{v}^\top\mathbf{v}^* \\ &= 4\mathbf{v}^\top\frac{\partial f}{\partial \mathbf{v}}(\mathbf{v},\mathbf{w}) + 2\left(1-\frac{\theta}{\pi}\right)\mathbf{v}^\top\mathbf{v}^*, \end{aligned} \tag{32}$$

and by Lemma 4,

$$\left\langle \mathbb{E}_{\mathbf{Z}}\left[\mathbf{g}(\mathbf{v},\mathbf{w};\mathbf{Z})\right], \frac{\partial f}{\partial \mathbf{w}}(\mathbf{v},\mathbf{w})\right\rangle = \frac{\sin(\theta)}{2(\sqrt{2\pi})^3}(\mathbf{v}^\top\mathbf{v}^*)^2.$$

Hence, for some $A$ depending only on $C$, we have

$$\begin{aligned} &\left\|\mathbb{E}_{\mathbf{Z}}\left[\mathbf{g}(\mathbf{v},\mathbf{w};\mathbf{Z})\right]\right\|^2 \\ =& \left\| \frac{2\mathbf{v}^\top\frac{\partial f}{\partial \mathbf{v}}(\mathbf{v},\mathbf{w})}{\sqrt{2\pi}}\mathbf{w} + \cos\left(\frac{\theta}{2}\right)\frac{\mathbf{v}^\top\mathbf{v}^*}{\sqrt{2\pi}}\left(\mathbf{w} - \frac{\mathbf{w}+\mathbf{w}^*}{\|\mathbf{w}+\mathbf{w}^*\|}\right) + \left(1-\frac{\theta}{\pi}-\cos\left(\frac{\theta}{2}\right)\right)\frac{\mathbf{v}^\top\mathbf{v}^*}{\sqrt{2\pi}}\mathbf{w}\right\|^2 \\ \le& \frac{6C^2}{\pi}\left\|\frac{\partial f}{\partial \mathbf{v}}(\mathbf{v},\mathbf{w})\right\|^2 + \cos^2\left(\frac{\theta}{2}\right)\frac{3(\mathbf{v}^\top\mathbf{v}^*)^2}{2\pi}\left\|\mathbf{w} - \frac{\mathbf{w}+\mathbf{w}^*}{\|\mathbf{w}+\mathbf{w}^*\|}\right\|^2 \\ &+ \left(1-\frac{\theta}{\pi}-\cos\left(\frac{\theta}{2}\right)\right)^2\frac{3(\mathbf{v}^\top\mathbf{v}^*)^2}{2\pi} \\ \le& \frac{6C^2}{\pi}\left\|\frac{\partial f}{\partial \mathbf{v}}(\mathbf{v},\mathbf{w})\right\|^2 + \cos^2\left(\frac{\theta}{2}\right)\frac{3\theta^2}{8\pi}(\mathbf{v}^\top\mathbf{v}^*)^2 + \left(1-\frac{\theta}{\pi}-\cos\left(\frac{\theta}{2}\right)\right)^2\frac{3(\mathbf{v}^\top\mathbf{v}^*)^2}{2\pi} \\ \le& \frac{6C^2}{\pi}\left\|\frac{\partial f}{\partial \mathbf{v}}(\mathbf{v},\mathbf{w})\right\|^2 + \frac{3\pi}{8}\cos^2\left(\frac{\theta}{2}\right)\sin^2\left(\frac{\theta}{2}\right)(\mathbf{v}^\top\mathbf{v}^*)^2 + \frac{3\sin(\theta)}{2\pi}(\mathbf{v}^\top\mathbf{v}^*)^2 \\ \le& A\left(\left\|\frac{\partial f}{\partial \mathbf{v}}(\mathbf{v},\mathbf{w})\right\|^2 + \left\langle \mathbb{E}_{\mathbf{Z}}\left[\mathbf{g}(\mathbf{v},\mathbf{w};\mathbf{Z})\right], \frac{\partial f}{\partial \mathbf{w}}(\mathbf{v},\mathbf{w})\right\rangle\right), \end{aligned}$$

where the equality is due to (31) and (32), the first inequality is due to Cauchy-Schwarz inequality, the second inequality holds because the angle between $\mathbf{w}$ and $\frac{\mathbf{w}+\mathbf{w}^*}{\|\mathbf{w}+\mathbf{w}^*\|}$ is $\frac{\theta}{2}$ and $\left\|\mathbf{w} - \frac{\mathbf{w}+\mathbf{w}^*}{\|\mathbf{w}+\mathbf{w}^*\|}\right\| \le \frac{\theta}{2}$, whereas the third inequality is due to $\sin(x) \ge \frac{2x}{\pi}$, $\cos(x) \ge 1-\frac{2x}{\pi}$, and

$$\left(1-\frac{2x}{\pi}-\cos(x)\right)^2 \le \left(\cos(x)-1+\frac{2x}{\pi}\right)\left(\cos(x)+1-\frac{2x}{\pi}\right) \le \sin(x)(2\cos(x)) = \sin(2x),$$

for all $x \in [0, \frac{\pi}{2}]$. □

**Proof of Theorem 1** . To leverage Lemma 2 and Lemma 5, we would need the boundedness of $\{\mathbf{v}^t\}$. Due to the coerciveness of $f$ w.r.t $\mathbf{v}$, there exists $C_0 > 0$, such that $\|\mathbf{v}\| \le C_0$ for any $\mathbf{v} \in \{\mathbf{v} \in \mathbb{R}^m : f(\mathbf{v},\mathbf{w}) \le f(\mathbf{v}^0,\mathbf{w}^0) \text{ for some } \mathbf{w}\}$. In particular, $\|\mathbf{v}^0\| \le C_0$. Using induction, suppose we already have $f(\mathbf{v}^t,\mathbf{w}^t) \le f(\mathbf{v}^0,\mathbf{w}^0)$ and $\|\mathbf{v}^t\| \le C_0$. If $\mathbf{w}^t = \pm\mathbf{w}^*$, then $\mathbf{w}^{t+1} = \mathbf{w}^{t+2} = \cdots = \pm\mathbf{w}^*$, and the original problem reduces to a quadratic program in terms of $\mathbf{v}$. So $\{\mathbf{v}^t\}$ will converge to $\mathbf{v}^*$ or $(\mathbf{I}+\mathbf{1}\mathbf{1}^\top)^{-1}(\mathbf{1}\mathbf{1}^\top - \mathbf{I})\mathbf{v}^*$ by choosing a suitable step size $\eta$. In either case, we have $\left\|\mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell}{\partial \mathbf{v}}(\mathbf{v}^t,\mathbf{w}^t;\mathbf{Z})\right]\right\|$ and $\left\|\mathbb{E}_{\mathbf{Z}}\left[\mathbf{g}(\mathbf{v}^t,\mathbf{w}^t;\mathbf{Z})\right]\right\|$ both converge to 0. Else if $\mathbf{w}^t \ne \pm\mathbf{w}^*$, we define for $a \in [0,1]$ that

$$\mathbf{v}^t(a) := \mathbf{v}^t - a(\mathbf{v}^{t+1} - \mathbf{v}^t) = \mathbf{v}^t - a\eta\mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell}{\partial \mathbf{v}}(\mathbf{v}^t,\mathbf{w}^t;\mathbf{Z})\right]$$

and

$$\mathbf{w}^t(a) := \mathbf{w}^t - a(\mathbf{w}^{t+1/2} - \mathbf{w}^t) = \mathbf{w}^t - a\eta\mathbb{E}_{\mathbf{Z}}\left[\mathbf{g}(\mathbf{v}^t,\mathbf{w}^t;\mathbf{Z})\right],$$

which satisfy

$$\mathbf{v}^t(0)=\mathbf{v}^t,\ \mathbf{v}^t(1)=\mathbf{v}^{t+1},\ \mathbf{w}^t(0)=\mathbf{w}^t,\ \mathbf{w}^t(1)=\mathbf{w}^{t+1/2}.$$

Let us fix $0<c<1$ and $C\geq C_0$. By the expressions of $\mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell}{\partial \mathbf{v}}(\mathbf{v}^t,\mathbf{w}^t;\mathbf{Z})\right]$ and $\mathbb{E}_{\mathbf{Z}}\left[\mathbf{g}(\mathbf{v}^t,\mathbf{w}^t;\mathbf{Z})\right]$ given in Lemma 3, and since $\|\mathbf{w}^t\|=1$, for sufficiently small $\tilde{\eta}$ depending on $C_0$, with $\eta\leq\tilde{\eta}$, it holds that $\|\mathbf{v}^t(a)\|\leq C$ and $\|\mathbf{w}^t(a)\|\geq c$ for all $a\in[0,1]$. Possibly at some point $a_0$ where $\theta(\mathbf{w}^t(a_0),\mathbf{w}^*)=0$ or $\pi$, such that $\frac{\partial f}{\partial \mathbf{w}}(\mathbf{v}^t(a_0),\mathbf{w}^t(a_0))$ does not exist. Otherwise, $\left\|\frac{\partial f}{\partial \mathbf{w}}(\mathbf{v}^t(a),\mathbf{w}^t(a))\right\|$ is uniformly bounded for all $a\in[0,1]/\{a_0\}$, which makes it integrable over the interval $[0,1]$. Then we have

$$\begin{aligned}
f(\mathbf{v}^{t+1},\mathbf{w}^{t+1}) &= f(\mathbf{v}^{t+1},\mathbf{w}^{t+1/2}) = f(\mathbf{v}^t+(\mathbf{v}^{t+1}-\mathbf{v}^t),\mathbf{w}^t+(\mathbf{w}^{t+1/2}-\mathbf{w}^t))\\
&= f(\mathbf{v}^t,\mathbf{w}^t)+\int_0^1\left\langle\frac{\partial f}{\partial \mathbf{v}}(\mathbf{v}^t(a),\mathbf{w}^t(a)),\mathbf{v}^{t+1}-\mathbf{v}^t\right\rangle \mathrm{d}a\\
&\quad+\int_0^1\left\langle\frac{\partial f}{\partial \mathbf{w}}(\mathbf{v}^t(a),\mathbf{w}^t(a)),\mathbf{w}^{t+1/2}-\mathbf{w}^t\right\rangle \mathrm{d}a\\
&= f(\mathbf{v}^t,\mathbf{w}^t)+\left\langle\frac{\partial f}{\partial \mathbf{v}}(\mathbf{v}^t,\mathbf{w}^t),\mathbf{v}^{t+1}-\mathbf{v}^t\right\rangle+\left\langle\frac{\partial f}{\partial \mathbf{w}}(\mathbf{v}^t,\mathbf{w}^t),\mathbf{w}^{t+1/2}-\mathbf{w}^t\right\rangle\\
&\quad+\int_0^1\left\langle\frac{\partial f}{\partial \mathbf{v}}(\mathbf{v}^t(a),\mathbf{w}^t(a))-\frac{\partial f}{\partial \mathbf{v}}(\mathbf{v}^t,\mathbf{w}^t),\mathbf{v}^{t+1}-\mathbf{v}^t\right\rangle \mathrm{d}a\\
&\qquad+\int_0^1\left\langle\frac{\partial f}{\partial \mathbf{w}}(\mathbf{v}^t(a),\mathbf{w}^t(a))-\frac{\partial f}{\partial \mathbf{w}}(\mathbf{v}^t,\mathbf{w}^t),\mathbf{w}^{t+1/2}-\mathbf{w}^t\right\rangle \mathrm{d}a\\
&\leq f(\mathbf{v}^t,\mathbf{w}^t)-\left(\eta-\frac{L\eta^2}{2}\right)\left\|\frac{\partial f}{\partial \mathbf{v}}(\mathbf{v}^t,\mathbf{w}^t)\right\|^2-\eta\left\langle\frac{\partial f}{\partial \mathbf{w}}(\mathbf{v}^t,\mathbf{w}^t),\mathbb{E}_{\mathbf{Z}}\left[\mathbf{g}(\mathbf{v}^t,\mathbf{w}^t;\mathbf{Z})\right]\right\rangle\\
&\quad+\frac{L\eta^2}{2}\left\|\mathbb{E}_{\mathbf{Z}}\left[\mathbf{g}(\mathbf{v}^t,\mathbf{w}^t;\mathbf{Z})\right]\right\|^2\\
&\leq f(\mathbf{v}^t,\mathbf{w}^t)-\left(\eta-(1+A)\frac{L\eta^2}{2}\right)\left\|\frac{\partial f}{\partial \mathbf{v}}(\mathbf{v}^t,\mathbf{w}^t)\right\|^2\\
&\quad-\left(\eta-\frac{AL\eta^2}{2}\right)\left\langle\frac{\partial f}{\partial \mathbf{w}}(\mathbf{v}^t,\mathbf{w}^t),\mathbb{E}_{\mathbf{Z}}\left[\mathbf{g}(\mathbf{v}^t,\mathbf{w}^t;\mathbf{Z})\right]\right\rangle. \qquad (33)
\end{aligned}$$

The third equality is due to the fundamental theorem of calculus. In the first inequality, we called Lemma 2 for $(\mathbf{v}^t,\mathbf{w}^t)$ and $(\mathbf{v}^t(a),\mathbf{w}^t(a))$ with $a\in[0,1]/\{a_0\}$. In the last inequality, we used Lemma 5. So when $\eta<\eta_0:=\min\left\{\frac{2}{(1+A)L},\tilde{\eta}\right\}$, we have $f(\mathbf{v}^{t+1},\mathbf{w}^{t+1})\leq f(\mathbf{v}^0,\mathbf{w}^0)$ and thus $\|\mathbf{v}^{t+1}\|\leq C_0$.

Summing up the inequality (33) over $t$ from 0 to $\infty$ and using $f\geq 0$, we have

$$\begin{aligned}
&\eta\sum_{t=0}^{\infty}\left(1-(1+A)\frac{L\eta}{2}\right)\left\|\frac{\partial f}{\partial \mathbf{v}}(\mathbf{v}^t,\mathbf{w}^t)\right\|^2+\left(1-\frac{AL\eta}{2}\right)\left\langle\frac{\partial f}{\partial \mathbf{w}}(\mathbf{v}^t,\mathbf{w}^t),\mathbb{E}_{\mathbf{Z}}\left[\mathbf{g}(\mathbf{v}^t,\mathbf{w}^t;\mathbf{Z})\right]\right\rangle\\
&\leq f(\mathbf{v}^0,\mathbf{w}^0)<\infty.
\end{aligned}$$

Hence,

$$\lim_{t\to\infty}\left\|\frac{\partial f}{\partial \mathbf{v}}(\mathbf{v}^t,\mathbf{w}^t)\right\|=0$$

and

$$\lim_{t\to\infty}\left\langle\frac{\partial f}{\partial \mathbf{w}}(\mathbf{v}^t,\mathbf{w}^t),\mathbb{E}_{\mathbf{Z}}\left[\mathbf{g}(\mathbf{v}^t,\mathbf{w}^t;\mathbf{Z})\right]\right\rangle=0.$$

Invoking Lemma 5 again, we further have

$$\lim_{t\to\infty}\left\|\mathbb{E}_{\mathbf{Z}}\left[\mathbf{g}(\mathbf{v}^t,\mathbf{w}^t;\mathbf{Z})\right]\right\|=0,$$

which completes the proof. □