

## A EXPERIMENTAL DETAILS

We summarize additional details for the experiments in our paper here regarding preprocessing, augmentation, and hyperparameters. All of the adaptive methods tested in our work have a learning rate  $\alpha$  and a denominator constant  $\epsilon$  associated with them. Unless otherwise stated, these were the only hyperparameters we changed; any other hyperparameters were implicitly left as the default values recommended by the methods’ original papers. This also includes Expectigrad’s momentum constant, which we generally left fixed at  $\beta = 0.9$  for the experiments—although an obvious exception to this was the MNIST momentum experiment in Figure 2.

**Reddi Problem** (Section 4.1, Figure 1.) We replicated both the “online” (below, left-hand side) and “stochastic” (below, right-hand side) variants of the synthetic convex optimization problem from Reddi et al. (2019).

$$h_t(x) = \begin{cases} 1010x & \text{if } t \equiv 0 \pmod{101} \\ -10x & \text{otherwise} \end{cases} \quad \text{or} \quad h_t(x) = \begin{cases} 1010x & \text{with probability 0.01} \\ -10x & \text{otherwise} \end{cases}$$

The online variant corresponds to the scaled function in (4), and therefore Theorem 1 guarantees that Expectigrad does not diverge in this case. We ran each adaptive method for 100M timesteps with  $\alpha = 3 \times 10^{-4}$  and  $\epsilon = 10^{-3}$ .

**MNIST** (Section 4.2, Figure 2.) We preprocessed the data by subtracting the mean image over the training data. Random horizontal flipping was applied during training for augmentation purposes. For the momentum experiments (Figure 2, left), we used a batch size of 1,024 images such that the single-trial epoch consisted of just 58 steps. The equations for each momentum variant are given in Table 2. For the 150-epoch comparison (Figure 2, right), minibatches consisted of 128 images to match the MNIST experiment in Kingma & Ba (2014). The default values of the learning rate  $\alpha$  and denominator constant  $\epsilon$  are borrowed from TensorFlow 1.x:  $\alpha = 10^{-3}$  and  $\epsilon = 10^{-8}$ .

Method	Update
Inner	$\mathbf{m}_t \leftarrow \beta \mathbf{m}_{t-1} + (1 - \beta) \mathbf{g}_t$ $\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \alpha \cdot \frac{\mathbf{m}_t}{\epsilon + \sqrt{\frac{s_t}{n_t}}}$
Bias-corrected Inner	$\mathbf{m}_t \leftarrow \beta \mathbf{m}_{t-1} + (1 - \beta) \mathbf{g}_t$ $\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \frac{\alpha}{1 - \beta^t} \cdot \frac{\mathbf{m}_t}{\epsilon + \sqrt{\frac{s_t}{n_t}}}$
Bias-corrected Outer	$\mathbf{m}_t \leftarrow \beta \mathbf{m}_{t-1} + (1 - \beta) \frac{\mathbf{g}_t}{\epsilon + \sqrt{\frac{s_t}{n_t}}}$ $\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \frac{\alpha}{1 - \beta^t} \cdot \mathbf{m}_t$

Table 2: Update equations for the alternative forms of Expectigrad with momentum that we tested in Section 4.2. Bias-corrected outer momentum is our contribution and represents the standard form of Expectigrad (Algorithm 1).

**CIFAR-10** We subtracted the mean RGB channel of the *ImageNet* training data for preprocessing (default procedure in TensorFlow). The data were augmented by applying random horizontal flips, rotations (max  $\pm 15^\circ$ ), and horizontal/vertical shifting (max  $\pm 25\%$  per dimension) during training. The minibatch size was 256. We searched for hyperparameter values via a small grid search with  $\alpha \in \{1 \times 10^{-8}, 3 \times 10^{-8}, 1 \times 10^{-7}, 3 \times 10^{-7}, \dots, 1 \times 10^0\}$  and  $\epsilon \in \{10^{-8}, 10^{-3}\}$ , where the “best” values were determined by the lowest final training loss (see Table 3).

**IWSLT’15 English-Vietnamese Translation** (Section 6, Table 1.) We trained the Transformer for 50,000 minibatches and the LSTM for 100,000 minibatches, where a single minibatch consisted of 4,096 tokens. The sentences were prepared using byte-pair encoding. After training, we decoded

Method	VGGNet-16		ResNet-50	
	$\alpha$	$\epsilon$	$\alpha$	$\epsilon$
AdaGrad	$3 \times 10^{-3}$	$10^{-3}$	$10^{-2}$	$10^{-3}$
Adam	$3 \times 10^{-4}$	$10^{-3}$	$10^{-2}$	$10^{-8}$
AMSGrad	$10^{-3}$	$10^{-3}$	$10^{-3}$	$10^{-8}$
Yogi	$10^{-3}$	$10^{-3}$	$10^{-3}$	$10^{-8}$
Expectigrad	$3 \times 10^{-4}$	$10^{-3}$	$10^{-3}$	$10^{-8}$

Table 3: Hyperparameters for the CIFAR-10 experiments.

the model outputs using a beam search of size 4 and a length penalty of 0.6 following Vaswani et al. (2017). We did not use checkpoint averaging, nor did we search over hyperparameter values; we chose  $\alpha = \epsilon = 10^{-3}$  for all methods based on the strongest results for the similar experiment conducted in Zaheer et al. (2018).

**ImageNet** (Section 6, Figure 4.) To preprocess the data, we first centrally cropped each image along its longer side to make it square and then resized it to  $96 \times 96$ . We then subtracted the mean RGB channel over the training set. We applied the same augmentation techniques that were used for the CIFAR-10 experiments, with additional random cropping of size  $84 \times 84$ . Note that the choice of cropping  $84 \times 84$  from  $96 \times 96$  is proportional to the commonly used dimensions  $224 \times 224$  from  $256 \times 256$ . The minibatch size was 512. We did not conduct a formal hyperparameter search due to the large-scale nature of this task. Instead, we chose a learning rate of  $\alpha = 10^{-2}$  and  $\epsilon = 10^{-3}$  for Yogi based on the recommendation in Zaheer et al. (2018). We then divided both of these values by 10 for Expectigrad.

## B LEMMAS FROM ASSUMPTIONS 1-3

In this section, we prove three important consequences of our assumptions in Section 2 that are necessary for establishing the convergence guarantees in Theorem 2. While the resulting properties that we establish for  $l$  and  $f$  below are often simply covered by additional assumptions in optimization research, we have chosen to derive them from our own assumptions for the sake of generality. In particular, we avoid defining arbitrary bound constants in Lemmas 2 and 3, allowing us to operate exclusively in terms of the Lipschitz constant  $L$  throughout our work.

**Lemma 1.**  $f$  is Lipschitz smooth:  $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$ ,  $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ .

**Lemma 2.**  $\nabla f$  has bounded norm:  $\|\nabla f(\mathbf{x})\| \leq L$ ,  $\forall \mathbf{x} \in \mathbb{R}^d$ .

**Lemma 3.**  $\nabla l$  has finite variance:  $\sigma^2 = \mathbb{E}[\|\nabla l(\mathbf{x}, \xi) - \nabla f(\mathbf{x})\|^2] < \infty$ ,  $\forall \mathbf{x} \in \mathbb{R}^d$ ,  $\forall \xi \in \Xi$ .

### B.1 PROOF OF LEMMA 1

*Proof.* From Assumption 3,  $\nabla l$  is Lipschitz continuous. Therefore,  $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$  and  $\forall \xi \in \Xi$ ,

$$L\|\mathbf{x} - \mathbf{y}\| \geq \|\nabla l(\mathbf{x}, \xi) - \nabla l(\mathbf{y}, \xi)\|$$

Taking the expectation of both sides and then invoking Jensen’s inequality,

$$\begin{aligned} L\|\mathbf{x} - \mathbf{y}\| &\geq \mathbb{E}[\|\nabla l(\mathbf{x}, \xi) - \nabla l(\mathbf{y}, \xi)\|] \\ &\geq \|\mathbb{E}[\nabla l(\mathbf{x}, \xi) - \nabla l(\mathbf{y}, \xi)]\| \\ &= \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \end{aligned}$$

It immediately follows that  $\nabla f$  is Lipschitz continuous with constant  $L$ .  $\square$

### B.2 PROOF OF LEMMA 2

*Proof.* From Assumption 2,  $l$  is Lipschitz continuous. Therefore,  $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$  and  $\forall \xi \in \Xi$ ,

$$L\|\mathbf{x} - \mathbf{y}\| \geq \|l(\mathbf{x}, \xi) - l(\mathbf{y}, \xi)\|$$

Taking the expectation of both sides and invoking Jensen’s inequality,

$$\begin{aligned} L\|\mathbf{x} - \mathbf{y}\| &\geq \mathbb{E}[\|l(\mathbf{x}, \boldsymbol{\xi}) - l(\mathbf{y}, \boldsymbol{\xi})\|] \\ &\geq \|\mathbb{E}[l(\mathbf{x}, \boldsymbol{\xi}) - l(\mathbf{y}, \boldsymbol{\xi})]\| \\ &= \|f(\mathbf{x}) - f(\mathbf{y})\| \end{aligned}$$

It immediately follows that  $f$  is Lipschitz continuous with constant  $L$ . This is sufficient to show that  $\|\nabla f(\mathbf{x})\| \leq L$  for all  $\mathbf{x} \in \mathbb{R}^d$  by the mean value theorem (e.g. Sohrab, 2003, Corollary 6.4.5).  $\square$

### B.3 PROOF OF LEMMA 3

*Proof.* We begin by deriving an expression for the second moment  $\mathbb{E}[\|\nabla l(\mathbf{x}, \boldsymbol{\xi})\|^2]$ , which will be used for our proof of Theorem 2. We will then prove that the variance—and, therefore, the second moment—is finite for all  $\mathbf{x} \in \mathbb{R}^d$  and  $\boldsymbol{\xi} \in \Xi$ . Assuming that each component of  $\nabla l(\mathbf{x}, \boldsymbol{\xi})$  is uncorrelated, and invoking the linearity of the expectation,

$$\begin{aligned} \mathbb{E}[\|\nabla l(\mathbf{x}, \boldsymbol{\xi})\|^2] &= \sum_{i=1}^d \mathbb{E}[\nabla l(\mathbf{x}, \boldsymbol{\xi})_i^2] \\ &= \sum_{i=1}^d \mathbb{E}[\nabla l(\mathbf{x}, \boldsymbol{\xi})_i^2] + \mathbb{E}[(\nabla l(\mathbf{x}, \boldsymbol{\xi})_i - \mathbb{E}[\nabla l(\mathbf{x}, \boldsymbol{\xi})_i])^2] \\ &= \sum_{i=1}^d \nabla f(\mathbf{x})_i^2 + \mathbb{E}[(\nabla l(\mathbf{x}, \boldsymbol{\xi})_i - \nabla f(\mathbf{x})_i)^2] \\ &= \|\nabla f(\mathbf{x})\|^2 + \mathbb{E}\left[\sum_{i=1}^d (\nabla l(\mathbf{x}, \boldsymbol{\xi})_i - \nabla f(\mathbf{x})_i)^2\right] \\ &= \|\nabla f(\mathbf{x})\|^2 + \underbrace{\mathbb{E}[\|\nabla l(\mathbf{x}, \boldsymbol{\xi}) - \nabla f(\mathbf{x})\|^2]}_{\sigma^2} \end{aligned} \tag{5}$$

Recall that  $\|\nabla f(\mathbf{x})\| \leq L$  by Lemma 2. Similarly,  $l$  is Lipschitz continuous by Assumption 2, implying that  $\|\nabla l(\mathbf{x}, \boldsymbol{\xi})\| \leq L$ . We can therefore use the triangle inequality to obtain a worst-case upper bound for the variance:

$$\begin{aligned} \sigma^2 &= \mathbb{E}[\|\nabla l(\mathbf{x}, \boldsymbol{\xi}) - \nabla f(\mathbf{x})\|^2] \\ &\leq \mathbb{E}[(\|\nabla l(\mathbf{x}, \boldsymbol{\xi})\| + \|\nabla f(\mathbf{x})\|)^2] \\ &\leq \mathbb{E}[(L + L)^2] \\ &= 4L^2 \end{aligned}$$

Since  $4L^2 < \infty$ , the variance  $\sigma^2$  must also be finite.  $\square$

## C PROOF OF THEOREM 1

*Proof.* We begin by computing the derivative with respect to  $x$  of the one-dimensional online optimization problem in (4):

$$g_t = \begin{cases} C & \text{if } t \equiv 1 \pmod{N} \\ -1 & \text{otherwise} \end{cases}$$

Recall that  $C > N - 1$  makes the correct solution  $-\infty$  (since the function should be minimized). Additionally, we must have  $N \geq 2$  to ensure the periodic function is nontrivial. In order to establish that Expectigrad does not diverge, we will show that the signed displacement of Expectigrad over a full period of (4) is always negative when  $C > N - 1$ . Note that the derivative is always nonzero, which means we can ignore sparsity in our analysis and reduce Expectigrad’s denominator average to  $\frac{s_t}{t}$ . We also will assume  $\beta = 0$  to simplify the analysis further. Suppose that  $t$  timesteps have already

elapsed, where  $t = nN$  for some positive integer  $n$ . We can directly calculate the total displacement of Expectigrad over the next period:

$$\begin{aligned}
\sum_{k=1}^N \left( -\alpha \frac{g_{t+k}}{\epsilon + \sqrt{\frac{s_{t+k}}{t+k}}} \right) &= -\alpha \left( \sum_{k=1}^N \frac{g_{t+k}}{\epsilon + \sqrt{\frac{s_{t+k}}{t+k}}} \right) \\
&= -\alpha \left( \frac{g_{t+1}}{\epsilon + \sqrt{\frac{s_{t+1}}{t+1}}} + \sum_{k=2}^N \frac{g_{t+k}}{\epsilon + \sqrt{\frac{s_{t+k}}{t+k}}} \right) \\
&= -\alpha \left( \frac{C}{\epsilon + \sqrt{\frac{s_{t+1}}{t+1}}} - \sum_{k=2}^N \frac{1}{\epsilon + \sqrt{\frac{s_{t+k}}{t+k}}} \right) \\
&\leq -\alpha \left( \frac{C}{\epsilon + \sqrt{\frac{s_{t+1}}{t+1}}} - \sum_{k=2}^N \frac{1}{\epsilon + \sqrt{\frac{s_{t+N}}{t+N}}} \right) \\
&= -\alpha \left( \frac{C}{\epsilon + \sqrt{\frac{s_{t+1}}{t+1}}} - \frac{N-1}{\epsilon + \sqrt{\frac{s_{t+N}}{t+N}}} \right) \\
&= -\alpha \cdot \frac{C \left( \epsilon + \sqrt{\frac{s_{t+N}}{t+N}} \right) - (N-1) \left( \epsilon + \sqrt{\frac{s_{t+1}}{t+1}} \right)}{\left( \epsilon + \sqrt{\frac{s_{t+1}}{t+1}} \right) \left( \epsilon + \sqrt{\frac{s_{t+N}}{t+N}} \right)} \quad (6)
\end{aligned}$$

The inequality is due to  $\frac{s_{t+N}}{t+N} = \frac{s_{t+k} + (N-k)}{t+k + (N-k)} > \frac{s_{t+k}}{t+k}$  as long as  $k < N$ , since  $\frac{a}{b} > 1 \implies \frac{a}{b} > \frac{a+c}{b+c}$  for any positive numbers  $a, b, c$ . It is easy to verify that  $\frac{s_{t+k}}{t+k}$  is greater than 1:

$$\frac{s_{t+k}}{t+k} = \frac{\sum_{i=1}^{t+k} g_i^2}{t+k} = \frac{\sum_{i=1}^t g_i^2 + C^2 + (k-1)}{t+k} > \frac{t + C^2 + (k-1)}{t+k} > \frac{t+k}{t+k} = 1$$

The inequalities result from the deduction that  $C^2 > 1$  (because  $C > N-1$  and  $N \geq 2$ ) and therefore  $g_i^2 \geq 1$ . To complete the proof, we must now show that the quantity in (6) is strictly negative. Equivalently, we can show that the numerator of the fraction is strictly positive:

$$\begin{aligned}
C \left( \epsilon + \sqrt{\frac{s_{t+N}}{t+N}} \right) - (N-1) \left( \epsilon + \sqrt{\frac{s_{t+1}}{t+1}} \right) &> 0 \\
C \left( \epsilon + \sqrt{\frac{s_{t+N}}{t+N}} \right) &> (N-1) \left( \epsilon + \sqrt{\frac{s_{t+1}}{t+1}} \right) \\
C &> (N-1) \cdot \frac{\epsilon + \sqrt{\frac{s_{t+1}}{t+1}}}{\epsilon + \sqrt{\frac{s_{t+N}}{t+N}}} \quad (7)
\end{aligned}$$

Because we already showed that  $\frac{s_{t+1}}{t+1} < \frac{s_{t+N}}{t+N}$ , it immediately follows that

$$\frac{\epsilon + \sqrt{\frac{s_{t+1}}{t+1}}}{\epsilon + \sqrt{\frac{s_{t+N}}{t+N}}} < 1$$

Combining this with (7), we arrive at the conclusion that Expectigrad's displacement over any period of (4) is negative whenever  $C > N-1$ . Hence, Expectigrad approaches  $-\infty$  after an arbitrarily large number of periods, which is the correct solution for this problem.  $\square$

## D PROOF OF THEOREM 2

*Proof.* From Lemma 1,  $f$  is Lipschitz smooth with constant  $L$ . Therefore, the following inequality holds for any two iterates  $\mathbf{x}_{t-1}, \mathbf{x}_t \in \mathbb{R}^d$  generated by Expectigrad:

$$f(\mathbf{x}_t) \leq f(\mathbf{x}_{t-1}) + \langle \nabla f(\mathbf{x}_{t-1}), \mathbf{x}_t - \mathbf{x}_{t-1} \rangle + \frac{L}{2} \|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 \quad (8)$$

Just as we did in our proof of Theorem 2, we will ignore sparsity in the gradient and also assume  $\beta = 0$ . Observe that Expectigrad’s denominator can be rewritten to resemble a time-variant EMA-based method:

$$\mathbf{v}_t = \beta_t \mathbf{v}_{t-1} + (1 - \beta_t) \mathbf{g}_t^2 \quad (9)$$

where  $\mathbf{v}_t = \frac{1}{t} \mathbf{s}_t$  and  $\beta_t = 1 - \frac{1}{t}$ . Using these identities to rewrite (1) provides a relationship between successive iterates:  $\mathbf{x}_t - \mathbf{x}_{t-1} = \frac{-\alpha \mathbf{g}_t}{\epsilon + \sqrt{\mathbf{v}_t}}$ . Substituting this into (8) gives

$$\begin{aligned} f(\mathbf{x}_t) &\leq f(\mathbf{x}_{t-1}) + \langle \nabla f(\mathbf{x}_{t-1}), \frac{-\alpha \mathbf{g}_t}{\epsilon + \sqrt{\mathbf{v}_t}} \rangle + \frac{L}{2} \left\| \frac{-\alpha \mathbf{g}_t}{\epsilon + \sqrt{\mathbf{v}_t}} \right\|^2 \\ &= f(\mathbf{x}_{t-1}) - \underbrace{\alpha \langle \nabla f(\mathbf{x}_{t-1}), \frac{\mathbf{g}_t}{\epsilon + \sqrt{\mathbf{v}_t}} \rangle}_A + \underbrace{\frac{L\alpha^2}{2} \left\| \frac{\mathbf{g}_t}{\epsilon + \sqrt{\mathbf{v}_t}} \right\|^2}_B \end{aligned} \quad (10)$$

Following Zaheer et al. (2018), we will expand the inner product by adding and subtracting a term.

$$A = -\alpha \underbrace{\langle \nabla f(\mathbf{x}_{t-1}), \frac{\mathbf{g}_t}{\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}}} \rangle}_{A_1} - \alpha \underbrace{\langle \nabla f(\mathbf{x}_{t-1}), \frac{\mathbf{g}_t}{\epsilon + \sqrt{\mathbf{v}_t}} - \frac{\mathbf{g}_t}{\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}}} \rangle}_{A_2}$$

To proceed with bounding the expression in (10), we will bound  $A_2$  and  $B$ . Because  $A_2$  is nonpositive, we can introduce an upper bound by taking the absolute value:

$$\begin{aligned} A_2 &= -\alpha \left\langle \nabla f(\mathbf{x}_{t-1}), \frac{\mathbf{g}_t}{\epsilon + \sqrt{\mathbf{v}_t}} - \frac{\mathbf{g}_t}{\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}}} \right\rangle \\ &\leq \left| -\alpha \left\langle \nabla f(\mathbf{x}_{t-1}), \frac{\mathbf{g}_t}{\epsilon + \sqrt{\mathbf{v}_t}} - \frac{\mathbf{g}_t}{\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}}} \right\rangle \right| \\ &\leq \alpha \left\langle |\nabla f(\mathbf{x}_{t-1})|, \left| \frac{\mathbf{g}_t}{\epsilon + \sqrt{\mathbf{v}_t}} - \frac{\mathbf{g}_t}{\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}}} \right| \right\rangle \end{aligned} \quad (11)$$

$$\leq L\alpha \left\langle \mathbf{1}, \left| \frac{\mathbf{g}_t}{\epsilon + \sqrt{\mathbf{v}_t}} - \frac{\mathbf{g}_t}{\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}}} \right| \right\rangle \quad (12)$$

$$\begin{aligned} &= L\alpha \left\langle \mathbf{1}, |\mathbf{g}_t| \cdot \left| \frac{1}{\epsilon + \sqrt{\mathbf{v}_t}} - \frac{1}{\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}}} \right| \right\rangle \\ &= L\alpha \left\langle \mathbf{1}, \frac{|\mathbf{g}_t|}{(\epsilon + \sqrt{\mathbf{v}_t})(\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}})} \cdot \frac{|\mathbf{v}_t - \beta_t \mathbf{v}_{t-1}|}{\sqrt{\mathbf{v}_t} + \sqrt{\beta_t \mathbf{v}_{t-1}}} \right\rangle \\ &= L\alpha \left\langle \mathbf{1}, \frac{|\mathbf{g}_t|}{(\epsilon + \sqrt{\mathbf{v}_t})(\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}})} \cdot \frac{(1 - \beta_t)g_t^2}{\sqrt{\beta_t \mathbf{v}_{t-1}} + (1 - \beta_t)g_t^2 + \sqrt{\beta_t \mathbf{v}_{t-1}}} \right\rangle \quad (13) \\ &\leq L\alpha \left\langle \mathbf{1}, \frac{|\mathbf{g}_t|}{(\epsilon + \sqrt{\mathbf{v}_t})(\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}})} \cdot |\mathbf{g}_t| \sqrt{1 - \beta_t} \right\rangle \\ &\leq L\alpha \left\langle \mathbf{1}, \frac{|\mathbf{g}_t|}{\epsilon(\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}})} \cdot |\mathbf{g}_t| \sqrt{1 - \beta_t} \right\rangle \\ &= \frac{L\alpha \sqrt{1 - \beta_t}}{\epsilon(\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}})} \|\mathbf{g}_t\|^2 \end{aligned}$$

Recall that the absolute value of a vector represents the element-wise absolute value in our work, which justifies the step in (11). Lemma 2 states that  $\|\nabla f(\mathbf{x}_{t-1})\| \leq L$ , allowing us to deduce that  $|\nabla f(\mathbf{x}_{t-1})_i| \leq L$  in (12). The equality in (13) follows from substituting the identity in (9) twice. The subsequent inequalities are obtained by removing positive terms from the denominators to yield upper bounds and then simplifying.

Next, because (9) implies that  $\beta_t \mathbf{v}_{t-1} \leq \mathbf{v}_t$ , we have the following bound on  $B$ :

$$B = \frac{L\alpha^2}{2} \left\| \frac{\mathbf{g}_t}{\epsilon + \sqrt{\mathbf{v}_t}} \right\|^2 = \frac{L\alpha^2}{2(\epsilon + \sqrt{\mathbf{v}_t})^2} \|\mathbf{g}_t\|^2 \leq \frac{L\alpha^2}{2\epsilon(\epsilon + \sqrt{\mathbf{v}_t})} \|\mathbf{g}_t\|^2 \leq \frac{L\alpha^2}{2\epsilon(\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}})} \|\mathbf{g}_t\|^2$$

Substituting these bounds for  $A_2$  and  $B$  into (10),

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) - \frac{\alpha \langle \nabla f(\mathbf{x}_t), \mathbf{g}_t \rangle}{\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}}} + \frac{L\alpha\sqrt{1-\beta_t}}{\epsilon(\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}})} \|\mathbf{g}_t\|^2 + \frac{L\alpha^2}{2\epsilon(\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}})} \|\mathbf{g}_t\|^2 \\ &= f(\mathbf{x}_t) - \frac{\alpha \langle \nabla f(\mathbf{x}_t), \mathbf{g}_t \rangle}{\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}}} + \frac{L\alpha(\frac{1}{2}\alpha + \sqrt{1-\beta_t})}{\epsilon(\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}})} \|\mathbf{g}_t\|^2 \end{aligned}$$

We are now ready to begin expressing this bound in terms of deterministic quantities. Recall that  $\mathbb{E}[\mathbf{g}_t] = \nabla f(\mathbf{x}_{t-1})$  by definition. Additionally, we showed earlier in (5) that  $\mathbb{E}[\|\mathbf{g}_t\|^2] = \|\nabla f(\mathbf{x}_{t-1})\|^2 + \sigma^2$  for some finite variance  $\sigma^2$ . Let us generalize this result by supposing that training is conducted using minibatches of size  $b$ . Because training samples are *i.i.d.*, this simply amounts to dividing the variance by the minibatch size. We can use these identities to take the conditional expectation of both sides of our inequality.

$$\begin{aligned} \mathbb{E}[f(\mathbf{x}_t) \mid \mathbf{x}_{t-1}] &\leq f(\mathbf{x}_{t-1}) - \frac{\alpha \langle \nabla f(\mathbf{x}_{t-1}), \mathbb{E}[\mathbf{g}_t \mid \mathbf{x}_{t-1}] \rangle}{\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}}} + \frac{L\alpha(\frac{1}{2}\alpha + \sqrt{1-\beta_t})}{\epsilon(\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}})} \mathbb{E}[\|\mathbf{g}_t\|^2 \mid \mathbf{x}_{t-1}] \\ &= f(\mathbf{x}_{t-1}) - \frac{\alpha}{\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}}} \|\nabla f(\mathbf{x}_{t-1})\|^2 + \frac{L\alpha(\frac{1}{2}\alpha + \sqrt{1-\beta_t})}{\epsilon(\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}})} \left( \|\nabla f(\mathbf{x}_{t-1})\|^2 + \frac{\sigma^2}{b} \right) \\ &= f(\mathbf{x}_{t-1}) - \frac{\alpha}{\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}}} \|\nabla f(\mathbf{x}_{t-1})\|^2 + \frac{L\alpha(\frac{1}{2}\alpha + \frac{1}{\sqrt{t}})}{\epsilon(\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}})} \left( \|\nabla f(\mathbf{x}_{t-1})\|^2 + \frac{\sigma^2}{b} \right) \\ &= f(\mathbf{x}_{t-1}) - \frac{\alpha[1 - \frac{L}{\epsilon}(\frac{1}{2}\alpha + \frac{1}{\sqrt{t}})]}{\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}}} \|\nabla f(\mathbf{x}_{t-1})\|^2 + \frac{L\alpha(\frac{1}{2}\alpha + \frac{1}{\sqrt{t}})\sigma^2}{\epsilon(\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}})b} \\ &\leq f(\mathbf{x}_{t-1}) - \frac{\alpha[1 - \frac{L}{\epsilon}(\frac{1}{2}\alpha + \frac{1}{\sqrt{t}})]}{\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}}} \|\nabla f(\mathbf{x}_{t-1})\|^2 + \frac{L\alpha(\frac{1}{2}\alpha + \frac{1}{\sqrt{t}})\sigma^2}{\epsilon^2 b} \end{aligned}$$

We made the substitution  $\sqrt{1-\beta_t} = \frac{1}{\sqrt{t}}$ . We will select  $\alpha \leq \frac{\epsilon}{L}$  to guarantee that  $\frac{L\alpha}{2\epsilon} \geq \frac{1}{2}$ . We impose no restriction on  $\epsilon$  except the implicit assumption that  $\epsilon > 0$ . Therefore,

$$\begin{aligned} \mathbb{E}[f(\mathbf{x}_t) \mid \mathbf{x}_{t-1}] &\leq f(\mathbf{x}_{t-1}) - \frac{\alpha[\frac{1}{2} - \frac{L}{\epsilon\sqrt{t}}]}{\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}}} \|\nabla f(\mathbf{x}_{t-1})\|^2 + \frac{L\alpha(\frac{1}{2}\alpha + \frac{1}{\sqrt{t}})\sigma^2}{\epsilon^2 b} \\ &= f(\mathbf{x}_{t-1}) - \frac{\alpha}{2(\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}})} \|\nabla f(\mathbf{x}_{t-1})\|^2 + \frac{L\alpha}{\epsilon\sqrt{t}} \left( \frac{\|\nabla f(\mathbf{x}_{t-1})\|^2}{\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}}} + \frac{\sigma^2}{\epsilon b} \right) + \frac{L\alpha^2\sigma^2}{2\epsilon^2 b} \\ &\leq f(\mathbf{x}_{t-1}) - \frac{\alpha}{2(\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}})} \|\nabla f(\mathbf{x}_{t-1})\|^2 + \frac{L\alpha}{\epsilon\sqrt{t}} \left( \frac{L^2}{\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}}} + \frac{\sigma^2}{\epsilon b} \right) + \frac{L\alpha^2\sigma^2}{2\epsilon^2 b} \end{aligned} \tag{14}$$

$$\leq f(\mathbf{x}_{t-1}) - \frac{\alpha}{2(\epsilon + L)} \|\nabla f(\mathbf{x}_{t-1})\|^2 + \frac{L\alpha}{\epsilon^2\sqrt{t}} \left( L^2 + \frac{\sigma^2}{b} \right) + \frac{L\alpha^2\sigma^2}{2\epsilon^2 b} \tag{15}$$

$$\leq f(\mathbf{x}_{t-1}) - \frac{\alpha}{2(\epsilon + L)} \|\nabla f(\mathbf{x}_{t-1})\|^2 + \frac{L\alpha}{\epsilon^2\sqrt{t}} \left( L^2 + \frac{\sigma^2}{b} \right) + \frac{\alpha\sigma^2}{2\epsilon b} \tag{16}$$

We applied Lemma 2 to (14).  $\sqrt{\beta_t \mathbf{v}_{t-1}} \leq L$  and  $\epsilon^2 \leq \epsilon(\epsilon + \sqrt{\beta_t \mathbf{v}_{t-1}})$ . In (16), we substituted  $\alpha \leq \frac{\epsilon}{L}$  into the rightmost term.

We can now rearrange the inequality, average over  $T$  timesteps, and take the total expectation of both sides.

$$\begin{aligned}
\frac{\alpha}{2(\epsilon + L)} \|\nabla f(\mathbf{x}_{t-1})\|^2 &\leq f(\mathbf{x}_{t-1}) - \mathbb{E}[f(\mathbf{x}_t) | \mathbf{x}_{t-1}] + \frac{L\alpha}{\epsilon^2\sqrt{t}} \left( L^2 + \frac{\sigma^2}{b} \right) + \frac{\alpha\sigma^2}{2\epsilon b} \\
\|\nabla f(\mathbf{x}_{t-1})\|^2 &\leq 2(\epsilon + L) \left[ \frac{f(\mathbf{x}_{t-1}) - \mathbb{E}[f(\mathbf{x}_t) | \mathbf{x}_{t-1}]}{\alpha} + \frac{L}{\epsilon^2\sqrt{t}} \left( L^2 + \frac{\sigma^2}{b} \right) + \frac{\sigma^2}{2\epsilon b} \right] \\
\mathbb{E}[\|\nabla f(\mathbf{x}_{t-1})\|^2] &\leq 2(\epsilon + L) \left[ \frac{\mathbb{E}[f(\mathbf{x}_{t-1})] - \mathbb{E}[f(\mathbf{x}_t)]}{\alpha} + \frac{L}{\epsilon^2\sqrt{t}} \left( L^2 + \frac{\sigma^2}{b} \right) + \frac{\sigma^2}{2\epsilon b} \right] \\
\frac{1}{T} \sum_{t=1}^T \mathbb{E}[\|\nabla f(\mathbf{x}_{t-1})\|^2] &\leq 2(\epsilon + L) \left[ \frac{f(\mathbf{x}_0) - \mathbb{E}[f(\mathbf{x}_T)]}{\alpha T} + \frac{L}{\epsilon^2} \left( L^2 + \frac{\sigma^2}{b} \right) \left( \frac{1}{T} \sum_{t=1}^T \frac{1}{\sqrt{t}} \right) + \frac{\sigma^2}{2\epsilon b} \right] \tag{17}
\end{aligned}$$

We can use a telescoping sum to obtain the final bound on this expression. Note the following inequality:

$$\frac{1}{\sqrt{t}} = \frac{2}{\sqrt{t} + \sqrt{t}} < \frac{2}{\sqrt{t} + \sqrt{t-1}} = 2(\sqrt{t} - \sqrt{t-1})$$

Therefore,

$$\frac{1}{T} \sum_{t=1}^T \frac{1}{\sqrt{t}} < \frac{1}{T} \sum_{t=1}^T 2(\sqrt{t} - \sqrt{t-1}) = \frac{1}{T} \cdot 2\sqrt{T} = \frac{2}{\sqrt{T}}$$

Substituting this bound into (17), we arrive at our final regret bound:

$$\begin{aligned}
\frac{1}{T} \sum_{t=1}^T \mathbb{E}[\|\nabla f(\mathbf{x}_{t-1})\|^2] &< 2(\epsilon + L) \left[ \frac{f(\mathbf{x}_0) - \mathbb{E}[f(\mathbf{x}_T)]}{\alpha T} + \frac{2L}{\epsilon^2\sqrt{T}} \left( L^2 + \frac{\sigma^2}{b} \right) + \frac{\sigma^2}{2\epsilon b} \right] \\
&\leq 2(\epsilon + L) \left[ \frac{f(\mathbf{x}_0) - f(\mathbf{x}^*)}{\alpha T} + \frac{2L}{\epsilon^2\sqrt{T}} \left( L^2 + \frac{\sigma^2}{b} \right) + \frac{\sigma^2}{2\epsilon b} \right]
\end{aligned}$$

where  $\mathbf{x}^*$  is local minimizer of  $f$ . Hence, Expectigrad achieves a bounded average regret on the order of  $O(\frac{1}{T} + \frac{1}{\sqrt{T}} + \frac{1}{b})$ .  $\square$