Optimization for Machine Learning HW 3

SOLUTIONS

All parts of each question are equally weighted. When solving one question/part, you may assume the results of all previous questions/parts.

1. This question foreshadows the idea of "adaptive learning rates" that we will discuss in more detail later. Suppose $\mathcal{L}(\mathbf{w}) = \mathbb{E}_z[\ell(\mathbf{w}, z)]$ is a convex function, and suppose $D \geq \|\mathbf{w}_1 - \mathbf{w}_\star\|$ for some $\mathbf{w}_1$ and $\mathbf{w}_\star = \operatorname{argmin} \mathcal{L}(\mathbf{w})$. In class, we showed that if $\|\nabla \ell(\mathbf{w}, z)\| \leq G$ for all z and $\mathbf{w}$, then stochastic gradient descent with learning rate $\eta = \frac{D}{G\sqrt{T}}$ satisfies

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)\right] \leq \frac{DG}{\sqrt{T}}$$

However, in order to set this learning rate, we needed to use knowledge of D, G and T. This question helps show a way to avoid needing to know T, although we still need to know G and D.

(a) First, we'll deal with unknown T. To do this, we will consider *projected* stochastic gradient descent with *varying learning rate*. Suppose we start at $\mathbf{w}_1 = 0$. Then the update is:

$$\mathbf{w}_{t+1} = \Pi_{\|\mathbf{w}\| \leq D}\left[\mathbf{w}_t - \eta_t \nabla \ell(\mathbf{w}_t, z_t)\right]$$

where $\Pi_{\|\mathbf{w}\| \leq D}[x] = \operatorname{argmin}_{\|\mathbf{w}\| \leq D} \|x - \mathbf{w}\|$. Notice that $\Pi_{\|\mathbf{w}\| \leq D}[\mathbf{w}_\star] = \mathbf{w}_\star$ by definition of D. Show that

$$\langle \nabla \ell(\mathbf{w}_t, z_t), \mathbf{w}_t - \mathbf{w}_\star \rangle \leq \frac{\|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2\eta_t} + \frac{\eta_t \|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}$$

And conclude:

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} \frac{\|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2\eta_t} + \frac{\eta_t \|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}\right]$$

(You may use without proof the identity $\|\Pi_{\|\mathbf{w}\| \leq D}[x] - \mathbf{w}_t\|^2 \leq \|x - \mathbf{w}_t\|^2$ for all t and all vectors x. This follows because $\|\mathbf{w}_t\| \leq D$.)

Solution:

You did not need to prove the identity provided in the hint, but if you are curious, here is a complete proof of the fact that $\|\Pi_{\|w\| \leq D}[x] - y\| \leq \|x - y\|$ for all y with $\|y\| \leq D$. First, observe that if $\|x\| \leq D$, then $\Pi_{\|w\| \leq D}[x] = x$, so the statement is immediate. Next, consider $\|x\| > D$. We can write $\Pi_{\|w\| \leq D}[x] = D\frac{x}{\|x\|}$. Let us define this quantity as $\overline{x}$. Then we have $x = (1 + r)\overline{x}$ for some positive scalar $r = \frac{\|x\| - D}{D}$. Then:

$$\begin{aligned}
\|x - y\|^2 &= \|x\|^2 - 2\langle x, y \rangle + \|y\|^2 \\
&= (1 + r)^2 \|\overline{x}\|^2 - 2\langle x, y \rangle + \|y\|^2 \\
&= (1 + 2r + r^2)\|\overline{x}\|^2 - 2(1 + r)\langle \overline{x}, y \rangle + \|y\|^2
\end{aligned}$$

From Cauchy-Schwarz we have $-\langle \overline{x}, y\rangle \geq -\|\overline{x}\|\|y\| \geq -D^2$, so:

$$\begin{aligned}
&\geq \|\overline{x}\|^2 + (2r + r^2)D^2 - 2\langle \overline{x}, y\rangle - 2rD^2 + \|y\|^2 \\
&\geq \|\overline{x}\|^2 - 2\langle \overline{x}, y\rangle + \|y\|^2 \\
&= \|\overline{x} - y\|^2
\end{aligned}$$

Now, armed with this identity we proceed:

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 &= \left\|\Pi_{\|\mathbf{w}\|\leq D}\left[\mathbf{w}_t - \eta_t \nabla \ell(\mathbf{w}_t, z_t)\right] - \mathbf{w}_\star\right\|^2 \\
&\leq \|\mathbf{w}_t - \eta_t \nabla \ell(\mathbf{w}_t, z_t) - \mathbf{w}_\star \\
&= \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - 2\eta_t\langle \nabla \ell(\mathbf{w}_t, z_t), \mathbf{w}_t - \mathbf{w}_\star\rangle + \eta_t^2\|\nabla \ell(\mathbf{w}_t.z_t)\|^2
\end{aligned}$$

rearranging:

$$\langle \nabla \ell(\mathbf{w}_t, z_t), \mathbf{w}_t - \mathbf{w}_\star\rangle \leq \frac{\|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2\eta_t} + \frac{\eta_t\|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}$$

This shows the first part of the question. Now, we notice that since $\mathbb{E}[\nabla \ell(\mathbf{w}_t, z_t)|\mathbf{w}_t] = \nabla \mathcal{L}(\mathbf{w}_t)$, we have by convexity:

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] &\leq \mathbb{E}[\langle \nabla \mathcal{L}(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_\star\rangle] \\
&= \mathbb{E}[\langle \nabla \ell(\mathbf{w}_t, z_t), \mathbf{w}_t - \mathbf{w}_\star\rangle] \\
&\leq \mathbb{E}\left[\frac{\|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2\eta_t} + \frac{\eta_t\|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}\right]
\end{aligned}$$

So, now summing over t yields:

$$\mathbb{E}\left[\sum_{t=1}^T \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)\right] \leq \mathbb{E}\left[\sum_{t=1}^T \frac{\|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2\eta_t} + \frac{\eta_t\|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}\right]$$

(b) Next, show that so long as η_t satisfies $\eta_t \leq \eta_{t-1}$ for all t, we have:

$$\mathbb{E}\left[\sum_{t=1}^T \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)\right] \leq \mathbb{E}\left[\frac{2D^2}{\eta_T} + \frac{\sum_{t=1}^T \eta_t\|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}\right]$$

(hint: at some point you will probably need to show $\|\mathbf{w}_t - \mathbf{w}_\star\|^2(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}}) \leq 2D^2(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}})$).

Solution:
Let's start by showing the hint. Notice that since $\eta_t \leq \eta_{t-1}$, we have $\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \geq 0$. Further, $\|\mathbf{w}_t\| \leq D$ since $\mathbf{w}_t$ is obtained by projecting to the ball of radius D, and $\|\mathbf{w}_\star\| \leq D$ by assumption, so that $\|\mathbf{w}_t - \mathbf{w}_\star\|^2 \leq (\|\mathbf{w}_t\| + \|\mathbf{w}_\star\|)^2 \leq 4D^2$. Therefore

$$\|\mathbf{w}_t - \mathbf{w}_\star\|^2(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}}) \leq 2D^2(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}})$$

as desired.
Now, from the previous part we have:

$$\mathbb{E}\left[\sum_{t=1}^T \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)\right] \leq \mathbb{E}\left[\sum_{t=1}^T \frac{\|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2\eta_t} + \frac{\eta_t\|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}\right]$$

reordering the sum:

$$= \mathbb{E}\left[\frac{\|\mathbf{w}_1 - \mathbf{w}_\star\|^2}{2\eta_1} - \frac{\|\mathbf{w}_{T+1} - \mathbf{w}_\star\|^2}{\eta_T} + \sum_{t=2}^{T} \|\mathbf{w}_t - \mathbf{w}_\star\|^2 \left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}}\right) + \sum_{t=1}^{T} \frac{\eta_t \|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}\right]$$

dropping the negative term and using the proved hint identity:

$$\leq \mathbb{E}\left[\frac{\|\mathbf{w}_1 - \mathbf{w}_\star\|^2}{2\eta_1} + 2D^2 \sum_{t=2}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \sum_{t=1}^{T} \frac{\eta_t \|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}\right]$$

telescoping, and dropping another negative term:

$$\leq \mathbb{E}\left[\frac{2D^2}{\eta_T} + \sum_{t=1}^{T} \frac{\eta_t \|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}\right]$$

(c) Next, consider the update

$$\mathbf{w}_{t+1} = \Pi_{\|\mathbf{w}\| \leq D}\left[\mathbf{w}_t - \eta_t \nabla \ell(\mathbf{w}_t, z_t)\right]$$

where we set $\eta_t = \frac{D}{G\sqrt{t}}$. Recalling our assumption that $\|\nabla \ell(\mathbf{w}_t, z_t)\| \leq G$ with probability 1, Show that

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)\right] \leq O(DG\sqrt{T})$$

This allows you to handle any T value without having the algorithm know T ahead of time. (Hint: you may want to show that $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 1 + \int_1^T \frac{dx}{\sqrt{x}}$).

Solution:
First, let's show the hint. Since $\frac{1}{\sqrt{x}}$ is decreasing as a function of x, we have

$$\begin{aligned}
\frac{1}{\sqrt{t}} &\leq \int_{t-1}^{t} \frac{dx}{\sqrt{x}} \\
\sum_{t=2}^{T} \frac{1}{\sqrt{t}} &\leq \int_{1}^{T} \frac{dx}{\sqrt{x}} \\
\sum_{t=1}^{T} \frac{1}{\sqrt{t}} &\leq 1 + \int_{1}^{T} \frac{dx}{\sqrt{x}} \\
&= 2\sqrt{T} - 1
\end{aligned}$$

Now, from part (b) (and noticing that this schedule for learning rates is always decreasing), we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)\right] &\leq \mathbb{E}\left[\frac{2D^2}{\eta_T} + \frac{\sum_{t=1}^{T} \eta_t \|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}\right] \leq \mathbb{E}\left[2DG\sqrt{T} + \frac{D}{2G}\sum_{t=1}^{T} \frac{\|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{\sqrt{t}}\right] \\
&\leq \mathbb{E}\left[2DG\sqrt{T} + \frac{D}{2G}\sum_{t=1}^{T} \frac{G^2}{\sqrt{t}}\right] \leq \mathbb{E}\left[2DG\sqrt{T} + \frac{DG}{2}\sum_{t=1}^{T} \frac{1}{\sqrt{t}}\right] \\
&\leq 3DG\sqrt{T} - \frac{DG}{2} \leq O(DG\sqrt{T})
\end{aligned}$$

(d) Finally, let's provide a learning rate schedule η_t such that η_t can be set *without prior knowledge of* G. Set $G_t = \max_{i\le t} \|\nabla \ell(\mathbf{w}_i, z_i)\|$ and set $\eta_t = \frac{D}{G_t\sqrt{t}}$. Show that:

$$\sum_{t=1}^{T} \frac{\|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{G_t\sqrt{t}} \le G\sum_{t=1}^{T}\frac{1}{\sqrt{t}}$$

Then show that this setting of η_t guarantees:

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)\right] \le O(DG\sqrt{T})$$

(You may use the hint of the previous part as given, even if you did not show it).

Solution:
First, notice that by definition of G_t, $\|\nabla \ell(\mathbf{w}_t, z_t)\|^2 \le G_t^2$. Therefore:

$$\begin{aligned}\sum_{t=1}^{T} \frac{\|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{G_t\sqrt{t}} &\le \sum_{t=1}^{T}\frac{G_t^2}{G_t\sqrt{t}} \\ &= \sum_{t=1}^{T}\frac{G_t}{\sqrt{t}}\end{aligned}$$

Using $G_t \le G$:

$$\le \sum_{t=1}^{T}\frac{G}{\sqrt{t}}$$

Now, also notice that since G_t is monotonically increasing, η_t is decreasing. Therefore we can apply the result of part (b) to obtain:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)\right] &\le \mathbb{E}\left[\frac{2D^2}{\eta_T} + \frac{\sum_{t=1}^{T}\eta_t\|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}\right] \\ &\le \mathbb{E}\left[2DG_T\sqrt{T} + \frac{D}{2}\sum_{t=1}^{T}\frac{\|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{G_t\sqrt{t}}\right] \\ &\le \mathbb{E}\left[2DG\sqrt{T} + \frac{GD}{2}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\right]\end{aligned}$$

Use the bound on the sum of $1/\sqrt{t}$ from part (c):

$$\begin{aligned}&\le \mathbb{E}\left[2DG\sqrt{T} + \frac{GD}{2}(2\sqrt{T}-1)\right] \\ &= O(DG\sqrt{T})\end{aligned}$$

2. This question is an exercise in understanding the non-convex SGD analysis. In class, we discussed setting a varying learning rate η_t proportional to $\frac{1}{\sqrt{t}}$ to obtain a non-convex convergence rate of:

$$\mathbb{E}[\|\nabla\mathcal{L}(\hat{\mathbf{w}})\|] \le O\left(\frac{\log(T)}{\sqrt{T}}\right)$$

In this question, we will remove the logarithmic factor by adding an extra assumption.

(a) Suppose that $\mathcal{L}$ is H-smooth, $\|\nabla \ell(\mathbf{w}, z)\| \le G$ for all $\mathbf{w}$ and z, and further that $\mathcal{L}(\mathbf{w}) \in [0, M]$ for all $\mathbf{w}$ (this last assumption is slightly stronger than we have assumed in class). Consider the SGD update:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla \ell(\mathbf{w}_t, z_t)$$

Suppose η_t is an arbitrary deterministic learning rate schedule satisfying $\eta_{t+1} \le \eta_t$ for all t (i.e. the learning rate never increases). Show that for all $\tau \le T$:

$$\frac{1}{T-\tau} \mathbb{E}\left[\sum_{t=\tau+1}^{T} \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2\right] \le \frac{1}{\eta_T (T-\tau)} \left(M + \frac{HG^2}{2} \sum_{t=\tau+1}^{T} \eta_t^2\right)$$

Solution:
By smoothness, we have:

$$\mathcal{L}(\mathbf{w}_{t+1}) \le \mathcal{L}(\mathbf{w}_t) + \langle \nabla \mathcal{L}(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t \rangle + \frac{H}{2} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2$$

taking expectations:

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_{t+1})] \le \mathbb{E}[\mathcal{L}(\mathbf{w}_t)] - \eta_t \, \mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] + \frac{H}{2} \eta_t^2 G^2$$
$$\eta_t \, \mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] \le \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_{t+1})] + \frac{H}{2} \eta_t^2 G^2$$

Now, sum from $t = \tau + 1$ to t and telescope:

$$\sum_{t=\tau+1}^{T} \eta_t \, \mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] \le \mathbb{E}[\mathcal{L}(\mathbf{w}_{\tau+1}) - \mathcal{L}(\mathbf{w}_{T+1})] + \frac{HG^2}{2} \sum_{t=\tau+1}^{T} \eta_t^2$$

Use $\mathcal{L}(\mathbf{w}) \le [0, M]$ to conclude $\mathcal{L}(\mathbf{w}_{\tau+1}) - \mathcal{L}(\mathbf{w}_{T+1}) \le M$:

$$\le M + \frac{HG^2}{2} \sum_{t=\tau}^{T} \eta_t^2$$

Next, since η_t is decreasing, $\eta_T \le \eta_t$ for all $t \le T$. Thus:

$$\begin{aligned} \eta_T \sum_{t=\tau}^{T} \mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] &\le \sum_{t=\tau}^{T} \eta_t \, \mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] \\ &\le M + \frac{HG^2}{2} \sum_{t=\tau}^{T} \eta_t^2 \end{aligned}$$

Divide both sides by $\eta_T (T - \tau)$ to conclude the desired result.

(b) Next, consider $\eta_t = \frac{1}{\sqrt{t}}$. In class, we considered choosing $\hat{\mathbf{w}}$ *uniformly* at random from $\mathbf{w}_1, \dots, \mathbf{w}_T$. Instead, produce a *non-uniform* distribution over $\mathbf{w}_1, \dots, \mathbf{w}_T$ such that choosing $\mathbf{w}_T$ from this distribution satisfies:

$$\mathbb{E}[\|\nabla \mathcal{L}(\hat{\mathbf{w}})\|^2] \le O\left(\frac{1}{\sqrt{T}}\right)$$

Consider the distribution that is uniform over the last $T/2$ iterates. That is, the probability that $\hat{b}w = \mathbf{w}_t$ is 0 if $t \le T/2$ and $2/T$ otherwise. Then we have:

$$\mathbb{E}[\|\nabla\mathcal{L}(\hat{\mathbf{w}})\|^2] = \frac{2}{T}\sum_{t=T/2+1}^{T}$$

Now, by the previous problem, with $\tau = T/2$, we have:

$$\mathbb{E}[\|\nabla\mathcal{L}(\hat{\mathbf{w}})\|^2] \le \frac{2}{T\eta_{T/2+1}}\left(M + \frac{HG^2}{2}\sum_{t=T/2+1}^{T}\eta_t^2\right) \tag{1}$$

To finish, we consider the sum $\sum_{t=T/2+1}^{T}\eta_t^2$. Notice that for $t \ge T/2$, $\eta_t \le \frac{\sqrt{2}}{\sqrt{T}}$. Thus,

$$\sum_{t=T/2+1}^{T}\eta_t^2 \le \sum_{t=T/2+1}^{T}\frac{2c^2}{T} = c^2$$

Putting this into (1), we have:

$$\begin{aligned}\mathbb{E}[\|\nabla\mathcal{L}(\hat{\mathbf{w}})\|^2] &\le \frac{2}{T\eta_T}\left(M + \frac{HG^2}{2}\right)\\ &\le \sqrt{2}\sqrt{T}\left(M + \frac{HG^2c^2}{2}\right)\\ &= O(1/\sqrt{T})\end{aligned}$$

BONUS (c) Assume that $\mathcal{L}$ is H-smooth, $\|\nabla\ell(\mathbf{w}, z)\| \le G$ for all $\mathbf{w}$ and z, and $\mathbf{w}_1$ is such that $\mathcal{L}(\mathbf{w}_1) - \inf_{\mathbf{w}}\mathcal{L} \le \Delta$ (note that this is *the same* as our normal assumptions in class). Devise sequence of learning rates such that:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2\right] \le O\left(\frac{(HG^2\log\log(T) + \Delta)\sqrt{\log(T)}}{\sqrt{T}}\right)$$

where the $O(\cdot)$ notation hides constants that may depend on G, Δ and H but *not* T.

Solution:
First, we establish a bound on the sum $\sum_{t=1}^{T}\frac{1}{(t+1)\log(t+1)}$. Observe that $\frac{1}{(x+1)\log(x+1)}$ is decreasing, so

$$\begin{aligned}\frac{1}{(t+1)\log(t+1)} &\le \int_{t-1}^{t}\frac{dx}{(x+1)\log(x+1)}\\ \sum_{t=2}^{T}\frac{1}{(t+1)\log(t+1)} &\le \int_{t=1}^{T}\frac{dx}{(x+1)\log(x+1)}\\ &= \log\log(T+1) - \log\log(2)\\ \sum_{t=1}^{T}\frac{1}{(t+1)\log(t+1)} &\le \frac{1}{2\log(2)} + \log\log(T+1) - \log\log(2)\end{aligned}$$

Now, from the lecture notes (Theorem 5.2), we have that for any sequence of learning rates:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2\right] \le \frac{\Delta}{T\eta_T} + \frac{HG^2}{2T\eta_T}\sum_{t=1}^{T}\eta_t^2$$

Let us set $\eta_t = \frac{1}{\sqrt{(t+1)\log(t+1)}}$. Then this result implies:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2\right] \leq \frac{\Delta\sqrt{\log(T+1)}}{\sqrt{T}} + \frac{HG^2\sqrt{\log(T+1)}}{2\sqrt{T}}\sum_{t=1}^{T}\frac{1}{(t+1)\log(t+1)}$$

using the result of part (a):

$$\leq \frac{\Delta\sqrt{\log(T+1)}}{\sqrt{T}} + \frac{HG^2\sqrt{\log(T+1)}}{2\sqrt{T}}\left(\frac{1}{2\log(2)} + \log\log(T+1) - \log\log(2)\right)$$

dropping constants:

$$= O\left(\frac{(HG^2\log\log(T) + \Delta)\sqrt{\log(T)}}{\sqrt{T}}\right)$$