Introduction to Computational Learning Theory **Spring 2021**

Lecture 10: Optimization algorithms

Lecturer: Roi Livni *Scribe:*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

We are still discussing convex learning problems of the form

$$\begin{aligned} \text{minimize} \quad & \underbrace{\mathbf{E}_{z\sim D}\left[\ell(w,z)\right]}_{\text{true loss}}. \\ \text{s.t} \quad & w \in K \end{aligned}$$

This also led us to consider the empirical version (with or without regularization). So we also consider empirical problems of the form:

$$\begin{aligned} \text{minimize} \quad & \underbrace{\lambda \cdot R(w)}_{\text{regularization}} + \underbrace{\frac{1}{m}\sum_{i=1}^{m} \ell(w,z_i)}_{\text{empirical loss}}. \\ \text{s.t} \quad & w \in K \end{aligned}$$

R here is a function that controls the *complexity* of the parameter w, and the RHS is the empirical loss. For example, we showed that by regularizing the ℓ_2 norm (i.e. $R(w) = \|w\|^2$) we can control the gap between the true loss and empirical loss if ℓ has the form

$$\ell(w, \underbrace{(x,y)}_{=z}) = \ell(w \cdot x, y).$$

Both the empirical objective, as well as the true loss are convex problems (if R, and ℓ are convex and K is a convex set). Therefore, we now move on to the algorithmic question: How to optimize a convex objective?

10.1 The general setup for convex optimization

In general, convex optimization are considered "easy" in the sense that we can efficiently optimize them. We first want to emphasize that this is not necessarily true. In fact, any computational problem can be formalized as a convex problem (in fact as a linear program). Thus, if we could solve any convex problem we could in fact solve NP-hard problems. Nevertheless, under certain mild assumptions the above is indeed true.

10.1.1 Assumptions

First-order oracle - The first underlying assumption we will make is that we can compute gradients. More accurately given a convex function F and a parameter $\mathbf{w}$, we assume that we have access to $\nabla F(\mathbf{w})$- the gradient of F at point $\mathbf{w}$. In other words, we assume that, to optimize F, we have acess to a first order oracle - a routine which, given input $\mathbf{w}$ returns on output a value $F(\mathbf{w})$ and a gradient $\nabla F(\mathbf{w})$ of F at $\mathbf{w}$.

Exercise 10.1. *Suppose*

$$F(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m} \ell(\mathbf{w}\cdot\mathbf{x}_i; y_i) \tag{10.1}$$

where $\ell(a; y)$ is some function. Calculate $\nabla F(\mathbf{w})$ in all of the above cases

- *(Ridge Regression) $\ell(a; y) = (a - y)^2$.*
- *(SVM) $\ell(a; y) = \max(0, 1 - a \cdot y)$.*
- *(Logistic Loss) $\ell(a; y) = \log(1 + \exp(-a \cdot y))$.*

Linear projections The second assumption we will make concerns the case we wish to optimize over a constrained set K. Namely, consider a general optimization problem of the form:

$$\begin{aligned} \text{minimize} \quad & F(\mathbf{w}) \\ \text{s.t} \quad & \mathbf{w} \in K \end{aligned}$$

The linear projection of an element $\mathbf{w}'$ on the set K is defined as

$$\Pi_K(\mathbf{w}') = \arg\min_{\mathbf{w}\in K} \|\mathbf{w} - \mathbf{w}'\| \tag{10.2}$$

Exercise 10.2. *Show how to compute* $\Pi_K(\mathbf{w})$ *for the case* $K = \{\mathbf{w} : \|\mathbf{w}\| \leq 1\}$.

An important property of the linear projection is that it reduces the distance between *all* elements in K, namely, for any $\mathbf{w}^* \in K$:

$$\|\Pi_K(\mathbf{w}) - \mathbf{w}^*\| \leq \|\mathbf{w} - \mathbf{w}^*\|. \tag{10.3}$$

Proof. : One can show that for any $\mathbf{w}^* \in K$: we have that

$$(\mathbf{w}^* - \Pi_K(\mathbf{w})) \cdot (\Pi_K(\mathbf{w}) - \mathbf{w}) \geq 0.$$

Assume otherwise, then we can prove that $h(\lambda) := \frac{1}{2}\|(1-\lambda)\Pi_K(\mathbf{w}) + \lambda\mathbf{w}^* - \mathbf{w}\|^2$, is a monotonically decreasing function of λ around $\lambda = 0$:

$$h'(0) = (\Pi_K(\mathbf{w}) - \mathbf{w}) \cdot (\mathbf{w}^\star - \Pi_K(\mathbf{w})) \leq 0.$$

By convexity this contradicts the optimality of $\Pi_K(\mathbf{w})$ w.r.t to eq. (10.2). Next, we have that

$$\begin{aligned}\|\mathbf{w} - \mathbf{w}^*\|^2 &= \|\mathbf{w}^* - \Pi_K(\mathbf{w})\|^2 + \|\Pi_K(\mathbf{w}) - \mathbf{w}\|^2 + 2(\mathbf{w}^* - \Pi_K(\mathbf{w})) \cdot (\Pi_K(\mathbf{w}) - \mathbf{w}) \\ &\geq \|\mathbf{w}^* - \Pi_K(\mathbf{w})\|^2\end{aligned}$$

□

Further structual assumptions We will further make the following two assumptions on the optimization problem.

1. First, we assume that F is G-Lipschitz. Formally, this means that

$$|F(\mathbf{w}_1) - F(\mathbf{w}_2)| \leq G\|\mathbf{w}_1 - \mathbf{w}_2\|.$$

Using the mean-value theorem, an equivalent formulation of this property is a bound on the norm of the gradient. In other words, we leave it as an exercise to prove that a convex function F is G-Lipschitz if and only if:

$$\|\nabla F(\mathbf{w})\| \leq G$$

for every $\mathbf{w}$.

2. Second we will assume a bound on the diameter of K. In particular we assume that for some fixed B:

$$\sup_{\mathbf{w}_1, \mathbf{w}_2 \in K} \|\mathbf{w}_1 - \mathbf{w}_2\| < B.$$

10.2 Gradient-Descent

We will analyze the following algorithm:

Algorithm 1 Gradient Descent

Input: A convex function F, a convex set K

Initialize $\mathbf{w}_1 = 0$

for $t = 1, 2 \ldots T$ **do**

Set $\mathbf{g}_t = \nabla F(\mathbf{w}_t)$.

Update:

$$\mathbf{w}_{t+\frac{1}{2}} = \mathbf{w}_t - \eta_t \mathbf{g}_t$$

$$\mathbf{w}_{t+1} = \Pi_K(\mathbf{w}_{t+\frac{1}{2}}).$$

end for

return $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t$.

Theorem 10.1. *Run Algorithm 1 with step sizes* $\eta_t = \frac{B}{G\sqrt{t}}$ *and assume that* $T \geq \frac{2G^2B^2}{3\epsilon^2}$ *then*

$$F(\bar{\mathbf{w}}_T) \leq \frac{1}{T}\sum_{t=1}^{T} F(\mathbf{w}_t) \leq \min_{\mathbf{w}^* \in \mathcal{K}} F(\mathbf{w}^*) + \epsilon$$

Proof. First, note that the first inequality simply follows from convexity of F, therefore we only need to prove the second inequality: namely,

$$\frac{1}{T}\sum_{t=1}^{T} F(\mathbf{w}_t) \leq \min_{\mathbf{w}^* \in \mathcal{K}} F(\mathbf{w}^*) + \epsilon$$

Next, we apply convexity again to obtain for any $\mathbf{w}^*$ (and in particular the minimizer above):

$$\frac{1}{T}\sum_{t=1}^{T} F(\mathbf{w}_t) - F(\mathbf{w}^*) \leq \frac{1}{T}\sum_{t=1}^{T} \nabla F(\mathbf{w}_t) \cdot (\mathbf{w}_t - \mathbf{w}^*) = \frac{1}{T}\sum_{t=1}^{T} \quad [\mathbf{g}_t \cdot (\mathbf{w}_t - \mathbf{w}^*)]. \tag{10.4}$$

We now upper bound $\mathbf{g}_t \cdot (\mathbf{w}_t - \mathbf{w}^*)$ using the update rule for $\mathbf{w}_{t+1}$

$$\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 = \|\Pi_K(\mathbf{w}_{t+1/2}) - \mathbf{w}^*\|^2 \tag{10.5}$$

$$= \|\Pi_{\mathcal{K}}(\mathbf{w}_t - \eta_t \mathbf{g}_t) - \mathbf{w}^*\|^2 \tag{10.6}$$

$$\leq \|\mathbf{w}_t - \eta_t \mathbf{g}_t - \mathbf{w}^*\|^2 \qquad \textit{eq. } (10.3) \tag{10.7}$$

Hence,

$$[\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2] \leq \quad [\|\mathbf{w}_t - \mathbf{w}^*\|^2] + \eta_t^2 \quad [\|\mathbf{g}_t\|^2] - 2 \quad [\eta_t \mathbf{g}_t \cdot (\mathbf{w}_t - \mathbf{w}^*)]$$

Rearranging terms we get:

$$2 \quad [\mathbf{g}_t \cdot (\mathbf{w}_t - \mathbf{w}^*)] \leq \quad \left[\frac{\|\mathbf{w}_t - \mathbf{w}^*\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2}{\eta_t}\right] + \eta_t G^2 \tag{10.8}$$

Summing Eq. 10.12 and Eq. 10.15 from $t = 1$ to T, and setting $\eta_t = \frac{B}{G\sqrt{t}}$, (and defining $\frac{1}{\eta_0} \equiv 0$):

$$\begin{aligned}
\left[\sum \mathbf{g}_t \cdot (\mathbf{w}_t - \mathbf{w}^*)\right] &\leq \sum_{t=1}^{T} \left[\frac{\|\mathbf{w}_t - \mathbf{w}^*\| - \|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2}{2\eta_t}\right] + \frac{G^2}{2}\sum_{t=1}^{T}\eta_t \\
&\leq \sum_{t=1}^{T} \left[\frac{\|\mathbf{w}_t - \mathbf{w}^*\|^2}{2}\right]\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \frac{G^2}{2}\sum_{t=1}^{T}\eta_t \\
&\leq \frac{B^2}{2}\sum_{t=1}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \frac{G^2}{2}\sum_{t=1}^{T}\eta_t && \text{Telescoping Series} \\
&\leq \frac{B^2}{2}\frac{1}{\eta_T} + \frac{G^2}{2}\sum_{t=1}^{T}\eta_t \\
&\leq \frac{3}{2}BG\sqrt{T}
\end{aligned}$$

The last inequality follows from our choice of η_t and $\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq 2\sqrt{T}$. Finally, we obtain the result by plugging in our choice of T, and dividing both sides by T. □

10.2.1 Learning by optimizing the empirical risk

Now, let's utilize the optimization guarantees from this lecture as well as the generalization guarantees from previous lecture to conclude what learning guarantees we can accomplish:

10.2.1.1 General Convex Problems

Suppose we have an unknown distribution $z \sim D$, and we want to minimize

$$\begin{aligned}
&\text{minimize}\mathcal{L}_D = \mathop{\mathbb{E}}_{z\sim D}[\ell(\mathbf{w}, z)] \\
&\text{s.t.}\mathbf{w} \in \mathcal{W}
\end{aligned}$$

We assume here that K is a set of diameter B (and let's assume the functions f are 1-Lipschitz, for concretness). We also assume that we have a first order oracle for the function $\ell(\mathbf{w}, z)$ for every z.

Our strategy will be similar to the PAC-learning setting. Namely: Minimize the empirical Risk. i.e. sample

$S =< z_1, \ldots, z_m >$ from the distribution D and consider the objective function:

$$\text{minimize}\mathcal{L}_S = \frac{1}{m}\sum_{i=1}^{m}[\ell(\mathbf{w}, z_i)]$$
$$\text{s.t.}\mathbf{w} \in \mathcal{W}$$

Minimizing the above, using say GD with $T = O(\frac{B^2}{\epsilon^2})$ iterations, we can find a solution $\mathbf{w}_S$, such that

$$F_S(w_s) - \min_{w\in\mathcal{W}} F_S(w) \leq \epsilon.$$

What is the performance of w_s over the true risk? Recall theorem 9.2, and let

$$w^\star = \arg\min_{w\in\mathcal{W}} \mathcal{L}_D(w).$$

if $m = \tilde{O}\left(\frac{d}{\epsilon^2}\log B/\delta\right)$ then for any $w \in \mathcal{W}$ (in particular, for w_S and $w^\star$,

$$|\mathcal{L}_D(w) - \mathcal{L}_S(w)| \leq \epsilon.$$

To conclude: Run GD for $\tilde{O}(\frac{B^2}{\epsilon^2})$ iterations, over an empirical objective of size $m = O(\frac{d}{\epsilon^2}\log\frac{B}{\delta}$, then w.p. $1-\delta$:

$$\begin{aligned}\mathcal{L}_D(w_S) - \mathcal{L}_D(w^\star) &\leq \underbrace{\mathcal{L}_D(w_S) - \mathcal{L}_S(w_S)}_{\leq\epsilon} + \underbrace{\mathcal{L}_S(w_S) - \mathcal{L}_S(w^*)}_{\leq\epsilon} + \underbrace{\mathcal{L}_S(w^*) - \mathcal{L}_D(w^*)}_{\leq\epsilon} \\ &\leq 3\cdot\epsilon\end{aligned}$$

Question 10.1. *The number of iterations of GD does not depend on the dimension, note that only the sample complexity depends on the dimension. Does this mean the algorithm's complexity is dimension independent?*

Notice, that every gradient step requires us to calculate all the m iterations. In other words, calculate the gradient of the empirical loss takes $O(m)$ time, meaning if m is dimension dependent then so is the algorithm's overall complexity..

10.2.1.2 Improved rates for Linear Losses

Next let us assume that

$$\ell(w, z) = \ell(w \cdot x, y).$$

We've shown that, under this structual assumption, we can improve the sample complexity rate. Recall theorem 9.13. For simplicity let us now assume that $\mathcal{W} = \{w : \|w\| \leq B\}$ then one can show that if $m = O(\frac{B \log 1/\delta}{\epsilon^2})$ then w.p. $1 - \delta$

$$\sup_{w \in \mathcal{W}} \mathcal{L}_D(w) - \mathcal{L}_S(w) \leq \epsilon.$$

Hence, again run the same number of iterations as before but with $m = O(\frac{B \log 1/\delta}{\epsilon^2})$ then w.p. $1 - \delta$, using the same argument as before

$$\mathcal{L}_S(w_s) - \mathcal{L}(w^\star) \leq \epsilon.$$

Question 10.2. *Note that the dimension does not appear in the sample complexity, nor in the number of iterations. Is this algorithm truly dimension independent?*

Note that to calculate the gradient, we need to calculate $w \cdot x$, which generally depends on the dimension. Though as we will discuss, one can sometimes consider spaces where the scalar product computation is also dimension free

10.3 Stochastic Gradient Descent Algorithm

We next analyze a slightly "lighter" approach for optimization, which would turn out to be particularly well-equipped for the setting of learning.

When performing GD, we at each iteration compute the full gradient ∇F. If F is an empirical risk minimization problem (with perhaps regularization term) that enforces us to compute at each iteration the gradient of a function of the form given in eq. (10.1). More generally, we will compute the gradient of a function of the form

$$F(\mathbf{w}) = \lambda R(\mathbf{w}) + \frac{1}{m} \sum_{i=1}^{m} \ell(\mathbf{w}; z_i).$$

Roughly, we would need to compute at each iteration m-gradients of the form $\nabla\ell(\mathbf{w}; z_i)$, overall applying theorem 10.1 we would consider $O(\frac{m}{\epsilon^2})$ computations of the gradient to achieve error ϵ. This may be very expensive in real-life applications.

Stochastic first order oracle: The idea behind Stochastic Gradient Descent is to avoid the heavy computation of the gradient at each iteration and instead consider the computation of a *noisy* version of the gradient.

In more detail, let us relax the first-order oracle and we will assume we have acess to a *noisy* first order oracle - a routine which, given input $\mathbf{w}$ returns on output a value $F(\mathbf{w})$ and a value $\mathbf{g}$ such that $\mathbb{E}(\mathbf{g}) = \nabla F(\mathbf{w})$.

Algorithm 2 Stochastic Gradient Descent

0: **Input:** A convex function F, a convex domain $\mathcal{K}$ and a sequence of learning rates $\{\eta_t\}$.

for $t = 1, 2 \dots T$ **do**

Compute $\mathbf{g}_t$ such that $\mathbb{E}[\mathbf{g}_t] = \nabla F(\mathbf{w}_t)$

Update and project:

$$\mathbf{w}_{t+\frac{1}{2}} = \mathbf{w}_t - \eta_t \mathbf{g}_t$$

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{K}}(\mathbf{w}_{t+\frac{1}{2}})$$

end for

return $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t$.

Theorem 10.2. *Run Algorithm 2 with step sizes* $\eta_t = \frac{B}{G\sqrt{t}}$. *Assume that* $\mathbb{E}[\|\mathbf{g}_t\|^2] \leq G^2$ *then for* $T \geq \frac{2G^2B^2}{3\epsilon^2}$*:*

$$\mathbb{E}\left[F(\bar{\mathbf{w}}_T)\right] \leq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[F(\mathbf{w}_t)\right] \leq \min_{\mathbf{x}^* \in \mathcal{K}} F(\mathbf{w}^*) + \epsilon$$

The proof is an almost exact repetition of the analysis of algorithm 1, with the difference that we add expectations in the relevant places (see the end of these lecture notes for completeness).

The implication to learning though is a little bit more interesting:

Application to learning First, let us return to the general form of an (regularized) empirical risk minimization problem of the form

$$\begin{aligned} \text{minimize} \qquad & F(\mathbf{w}) = \lambda R(\mathbf{w}) + \frac{1}{m}\sum_{i=1}^{m} \ell(\mathbf{w}; z_i) \\ \text{s.t.} \qquad & \mathbf{w} \in K \end{aligned}$$

We will now exploit the fact that the objective function, F, to be minimized is described as the sum of functions and readily apply Stochastic Gradient Descent, which is potentially much cheaper then Gradient Descent.

For this, instead of computing the full gradient of F at each iteration, we pick z_i from the sample (uniformly, with replacement) and consider the gradient step at direction

$$\mathbf{g}_t = \lambda \nabla R(\mathbf{w}) + \nabla \ell(\mathbf{w}; z_i).$$

In expectation over the choice of z_i we have

$$\mathop{\mathbb{E}}_{z_i}[\mathbf{g}_t] = \frac{1}{m}\sum_{i=1}^{m} \lambda \nabla R(\mathbf{w}) + \nabla \ell(\mathbf{w}; z_i) = \lambda \nabla R(\mathbf{w}) + \frac{1}{m}\sum_{i=1}^{m} \ell(\mathbf{w}; z_i) = \nabla F(\mathbf{w}).$$

Thus, if we can compute $\nabla \ell(\mathbf{w}; z)$ for any arbitrary z then we can compute a noisy gradient of $\nabla F(\mathbf{w})$, and minimize the empirical risk (with or without regularization).

Optimizing the true risk instead of empirical risk There is however a different view point that we can take on SGD. Let us now return to the true risk objective, and consider an optimization problem of the form:

$$\text{minimize} \qquad F(\mathbf{w}) = \mathop{\mathbb{E}}_{z \sim D}[\ell(\mathbf{w}; z)] \tag{10.9}$$

$$\text{s.t.} \qquad \mathbf{w} \in K \tag{10.10}$$

Note that by linearity of expectation (and the derivative): $\nabla F(\mathbf{w}) = \mathbb{E}_{z \sim D}[\nabla \ell(\mathbf{w}; z)]$.

Thus, if we pick randomly $z \sim D$, we obtain a noisy gradient as before. The main difference from the previous approach is that now we do not draw z with replacement from a sample set S. Instead, at each iteration we draw a fresh example from S without replacement.

Applying our result to the optimization procedure we now obtain a guarantee over the *true* risk and not the empirical risk. In other words, a direct application of theorem 10.2, with $T = O(1/\epsilon^2)$ and $\eta_t = O(1/\sqrt{t})$:

$$\mathbf{E}[\ell(\bar{\mathbf{w}}_T; z)] \leq \min_{\mathbf{w} \in K} \mathbf{E}[\ell(\mathbf{w}; z)] + \epsilon.$$

Here the expectation is over both the output of the algorithm and a fresh new example $z \sim D$.

Thus, unlike theorem 10.1, here obtain a guarantee on the *true risk* and not a guarantee on the empirical risk. In contrast, we cannot apply Gradient descent directly to eq. (10.9) as we don't know how to compute the gradient of the true risk.

Because we directly minimize the true risk, here the use of SGD leads to generalization bounds *independent of the Rademacher complexity or any other structural properties of the set K* or the loss function. In fact, this difference is inherent as was shown in [2] there are convex problems that don't exhibit any type of uniform convergence, but SGD provably can learn them. More recently, [1] showed that there are in fact convex problems where the generalization bound of gradient descent is inherently *dimension dependent*, while the generalization bound of sgd is not (as we've seen here).

10.3.1 Proof of theorem 10.2

Proof. First, note that the first inequality simply follows from convexity of F, therefore we only need to prove the second inequality: namely,

$$\frac{1}{T}\sum_{t=1}^{T} F(\mathbf{w}_t) \leq \min_{\mathbf{w}^* \in \mathcal{K}} F(\mathbf{w}^*) + \epsilon$$

Next, we apply convexity again to obtain for any $\mathbf{w}^*$ (and in particular the minimizer above):

$$\frac{1}{T}\sum_{t=1}^{T} F(\mathbf{w}_t) - F(\mathbf{w}^*) \leq \frac{1}{T}\sum_{t=1}^{T} \nabla F(\mathbf{w}_t)\cdot(\mathbf{w}_t - \mathbf{w}^*) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\mathbf{g}_t \cdot (\mathbf{w}_t - \mathbf{w}^*)]. \tag{10.11}$$

The last equality is true since g_t is independent of w_t and $w^\star$.

We now upper bound $\mathbf{g}_t \cdot (\mathbf{w}_t - \mathbf{w}^*)$ using the update rule for $\mathbf{w}_{t+1}$

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 &= \|\Pi_K(\mathbf{w}_{t+1/2}) - \mathbf{w}^*\|^2 && (10.12)\\
&= \|\Pi_{\mathcal{K}}(\mathbf{w}_t - \eta_t \mathbf{g}_t) - \mathbf{w}^*\|^2 && (10.13)\\
&\leq \|\mathbf{w}_t - \eta_t \mathbf{g}_t - \mathbf{w}^*\|^2 && \textit{eq. } (10.3) \quad (10.14)
\end{aligned}$$

Hence,

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2] \leq \mathbb{E}[\|\mathbf{w}_t - \mathbf{w}^*\|^2] + \eta_t^2\,\mathbb{E}[\|\mathbf{g}_t\|^2] - 2\,\mathbb{E}[\eta_t \mathbf{g}_t \cdot (\mathbf{w}_t - \mathbf{w}^*)]$$

Rearranging terms we get:

$$2\,\mathbb{E}\left[\mathbf{g}_t \cdot (\mathbf{w}_t - \mathbf{w}^*)\right] \leq \mathbb{E}\left[\frac{\|\mathbf{w}_t - \mathbf{w}^*\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2}{\eta_t}\right] + \eta_t G^2 \tag{10.15}$$

Summing Eq. 10.12 and Eq. 10.15 from $t = 1$ to T, and setting $\eta_t = \frac{B}{G\sqrt{t}}$, (and defining $\frac{1}{\eta_0} \equiv 0$):

$$\begin{aligned}
\mathbb{E}\left[\sum \mathbf{g}_t \cdot (\mathbf{w}_t - \mathbf{w}^*)\right] &\leq \sum_{t=1}^{T} \mathbb{E}\left[\frac{\|\mathbf{w}_t - \mathbf{w}^*\| - \|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2}{2\eta_t}\right] + \frac{G^2}{2}\sum_{t=1}^{T}\eta_t\\
&\leq \sum_{t=1}^{T} \mathbb{E}\left[\frac{\|\mathbf{w}_t - \mathbf{w}^*\|^2}{2}\right]\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \frac{G^2}{2}\sum_{t=1}^{T}\eta_t\\
&\leq \frac{B^2}{2}\sum_{t=1}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \frac{G^2}{2}\sum_{t=1}^{T}\eta_t && \text{Telescoping Series}\\
&\leq \frac{B^2}{2}\frac{1}{\eta_T} + \frac{G^2}{2}\sum_{t=1}^{T}\eta_t\\
&\leq \frac{3}{2}BG\sqrt{T}
\end{aligned}$$

The last inequality follows from our choice of η_t and $\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq 2\sqrt{T}$. Finally, we obtain the result by plugging in our choice of T, and dividing both sides by T. □

References

[1] Idan Amir, Tomer Koren, and Roi Livni. Sgd generalizes better than gd (and regularization doesn't help). *arXiv preprint arXiv:2102.01117*, 2021.

[2] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT*, 2009.