Optimization for Machine Learning Final Exam

Shuyue Jia
BUID: U62343813

Due: 12/18/2023

You MAY NOT collaborate with other students on this final.

When solving one question/part, you may assume the results of all previous questions/parts. You may also assume all previous homework results and results from class or lecture notes, but please explain which result you are using when you use it.

Please justify all answers unless explicitly instructed not to in the question statement.

1. (5pts) Suppose that $\mathcal{L}(\mathbf{w})$ is a convex function, and $\mathbf{w}_1, \ldots, \mathbf{w}_T$ are such that:

$$\sum_{t=1}^{T} t(\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)) \leq T^{3/2}$$

You know the identities of the points $\mathbf{w}_1, \ldots, \mathbf{w}_T$, but you *do not* have any other information about $\mathcal{L}$ (e.g. you cannot compute its values or its gradients). Provide a *deterministic* point $\hat{\mathbf{w}}$ as a function of $\mathbf{w}_1, \ldots, \mathbf{w}_T$ such that

$$\mathcal{L}(\hat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_\star) \leq O\left(\frac{1}{\sqrt{T}}\right)$$

Solution:

Since $\hat{\mathbf{w}}$ is a deterministic point, it is selected uniformly at random from $\mathbf{w}_1, \ldots, \mathbf{w}_T$. Specifically, we'll use the following weighted average:

$$\hat{\mathbf{w}} = \frac{\sum_{t=1}^{T} t\mathbf{w}_t}{\sum_{t=1}^{T} t}. \tag{1}$$

Now, let's analyze $\mathcal{L}(\hat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_\star)$:

$$\begin{aligned}
\mathcal{L}(\hat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_\star) &= \mathcal{L}\left(\frac{\sum_{t=1}^{T} t\mathbf{w}_t}{\sum_{t=1}^{T} t}\right) - \mathcal{L}(\mathbf{w}_\star) \\
&= \mathcal{L}\left(\frac{\sum_{t=1}^{T} t\mathbf{w}_t}{\sum_{t=1}^{T} t}\right) - \frac{\sum_{t=1}^{T} t(\mathcal{L}(\mathbf{w}_\star))}{\sum_{t=1}^{T} t} \quad \text{(Multiplying and dividing by } \sum_{t=1}^{T} t) \qquad (2) \\
&= \frac{\sum_{t=1}^{T} t(\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star))}{\sum_{t=1}^{T} t}. \quad \text{(Using the definition of } \hat{\mathbf{w}})
\end{aligned}$$

Now, use the given condition $\frac{1}{T}\sum_{t=1}^{T} t(\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)) \leq \frac{1}{\sqrt{T}}$ and substitute it in:

$$\begin{aligned}
\mathcal{L}(\hat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_\star) &\leq \frac{\sum_{t=1}^{T} t(\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star))}{\sum_{t=1}^{T} t} \\
&\leq \frac{T^{3/2}}{\sum_{t=1}^{T} t} \\
&= \frac{T^{3/2}}{\frac{T(T+1)}{2}} \\
&= O\left(\frac{1}{\sqrt{T}}\right).
\end{aligned} \tag{3}$$

□

2. (5pts) Specify a function $f : \mathbb{R} \to \mathbb{R}$ that is both 1-strongly convex and 10-Lipschitz, or prove that no such function exists.

Solution:
Assume, for the sake of contradiction, that such a function f exists.

1-strong convexity:

For f to be 1-strongly convex, it must satisfy the inequality $f''(x) \geq 1$ for all $x \in \mathbb{R}$.

10-Lipschitz:

For f to be 10-Lipschitz, it must satisfy the inequality $|f'(x)| \leq 10$ for all $x \in \mathbb{R}$.

Now, let's consider the contradiction. If $f''(x) \geq 1$ for all x and $|f'(x)| \leq 10$ for all x, then integrating $f'(x)$ with respect to x should yield a function $f(x)$ that is both 1-strongly convex and 10-Lipschitz.

However, the contradiction arises because it is not possible to have a function with a bounded derivative (Lipschitz condition) whose integral has an unbounded second derivative (strong convexity) over the entire real line. The conditions of being 1-strongly convex and 10-Lipschitz are incompatible when considered together.

Therefore, we conclude that no such function $f : \mathbb{R} \to \mathbb{R}$ can be both 1-strongly convex and 10-Lipschitz over the entire real line.

3. (5pts) Specify a function $f : \mathbb{R} \to \mathbb{R}$ that is both 1-smooth and 2 strongly convex, or prove that no such function exists.

Solution:
According to **Theorem 24.6**, we know that gradient descent with learning rate $\eta = \frac{1}{H}$ guarantees:

$$\mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}(\mathbf{w}_\star) \leq \exp\left(-\frac{\mu}{H}t\right)(\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star)). \tag{1}$$

Thus, the gradient descent with learning rate $\eta = 1$ of the function in above problem must guarantee:

$$\mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}(\mathbf{w}_\star) \leq \exp(-2t)(\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star)). \tag{2}$$

1-Smooth:

A function f is 1-smooth if its derivative is Lipschitz continuous with a Lipschitz constant of 1. This implies that for all $x, y \in \mathbb{R}$, $|f'(x) - f'(y)| \leq |x - y|$, or equivalently, the second derivative $|f''(x)|$ is bounded by 1.

2-Strongly Convex:

A function is 2-strongly convex if for all $x, y \in \mathbb{R}$ and $\lambda \in [0, 1]$, it satisfies the inequality:

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) - \lambda(1 - \lambda)(x - y)^2. \tag{3}$$

This condition implies that the second derivative $f''(x)$ is at least 2 everywhere.

Now, examining these two conditions, we see that they impose conflicting requirements on the second derivative of the function: 1-smoothness requires that the second derivative does not exceed 1, while 2-strong convexity demands it to be at least 2. Since these two conditions are mutually exclusive, no function $f : \mathbb{R} \to \mathbb{R}$ can satisfy both properties simultaneously.

Regarding the reference to **Theorem 24.6** and gradient descent: The theorem and its implications on gradient descent convergence rates are not directly relevant to proving or disproving the existence of a function that meets the given smoothness and convexity criteria. The theorem is about optimization algorithm performance for certain classes of functions, not a tool for establishing the existence of a function with specified mathematical properties.

Therefore, the conclusion is that no such function exists that is both 1-smooth and 2-strongly convex. The inherent mathematical properties of such a function would be contradictory and cannot coexist in a single real-valued function of a real variable.

4. (a) (5pts) *Newton's method* employs the following update: $\mathbf{w}_{t+1} = \mathbf{w}_t - \nabla^2 \mathcal{L}(\mathbf{w}_t)^{-1} \nabla \mathcal{L}(\mathbf{w}_t)$. Suppose that $\mathcal{L}$ is a convex quadratic function (that is, $\mathcal{L}$ has the form $\mathcal{L}(\mathbf{w}) = \mathbf{w}^\top A\mathbf{w} + \langle \mathbf{w}, \mathbf{v} \rangle + c$ for some positive semi-definite matrix A, vector $\mathbf{w}$ and scalar c). Show that in this case, no matter what $\mathbf{w}_1$ is, $\mathbf{w}_2 = \operatorname{argmin} \mathcal{L}$ whenever A is strictly positive-definite.

Solution:

Let's analyze Newton's method in the context of convex quadratic functions. Given a convex quadratic function $\mathcal{L}(\mathbf{w}) = \mathbf{w}^\top A\mathbf{w} + \langle \mathbf{w}, \mathbf{v} \rangle + c$, where A is a positive semi-definite matrix, and $\mathbf{v}$ is a vector, let's find the update rule for Newton's method.

The Newton's method update is given by:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \nabla^2 \mathcal{L}(\mathbf{w}_t)^{-1} \nabla \mathcal{L}(\mathbf{w}_t). \tag{1}$$

For a quadratic function, the Hessian matrix (second derivative) is simply the matrix A. So, the update becomes:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - A^{-1} \nabla \mathcal{L}(\mathbf{w}_t). \tag{2}$$

Now, let's consider the specific form of $\mathcal{L}(\mathbf{w})$:

$$\mathcal{L}(\mathbf{w}) = \mathbf{w}^\top A\mathbf{w} + \langle \mathbf{w}, \mathbf{v} \rangle + c. \tag{3}$$

The gradient of $\mathcal{L}$ with respect to $\mathbf{w}$ is given by:

$$\nabla \mathcal{L}(\mathbf{w}) = 2A\mathbf{w} + \mathbf{v}. \tag{4}$$

Now, substitute this into the update rule:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - A^{-1}(2A\mathbf{w}_t + \mathbf{v}). \tag{5}$$

Simplify:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - 2\mathbf{w}_t - A^{-1}\mathbf{v}. \tag{6}$$

Combine like terms:

$$\mathbf{w}_{t+1} = -\mathbf{w}_t - A^{-1}\mathbf{v}. \tag{7}$$

Now, let's consider the minimization problem $\mathbf{w}_2 = \operatorname{argmin} \mathcal{L}$. For a convex quadratic function, the minimum occurs at the point where the gradient is zero. So, we set $\nabla\mathcal{L}(\mathbf{w}_2) = 0$:

$$2A\mathbf{w}_2 + \mathbf{v} = 0. \tag{8}$$

Solving for $\mathbf{w}_2$:

$$\mathbf{w}_2 = -\frac{1}{2}A^{-1}\mathbf{v}. \tag{9}$$

Now, compare this with the update rule:

$$\mathbf{w}_{t+1} = -\mathbf{w}_t - A^{-1}\mathbf{v}. \tag{10}$$

You can see that $\mathbf{w}_2$ obtained from the minimization problem is indeed the solution to the update rule when A is strictly positive-definite. Therefore, in this case, no matter what $\mathbf{w}_1$ is, $\mathbf{w}_2 = \operatorname{argmin} \mathcal{L}$. □

(b) (5pts) A friend asks if the reason the result in part (a) is able to avoid the $\Omega(1/T^2)$ lower bound we learned in class is that the analysis was restricted to quadratic loss functions rather than general smooth convex loss functions. You tell them no: the lower bound applies even to algorithms that consider only quadratic losses. Why?

Solution:
The $\Omega(1/T^2)$ lower bound for optimization algorithms, often associated with the convergence rate of certain first-order methods, is a general result that applies to a broad class of smooth convex loss functions, not just quadratic losses. The lower bound is not specific to the type of loss function being minimized.

In the context of convex optimization, the lower bound states that any algorithm that only uses first-order information (such as gradients) and achieves a certain level of precision in terms of the objective value after a certain number of iterations (measured by the number of oracle queries or gradient evaluations) must take at least $\Omega(1/T^2)$ iterations to converge to a solution, where T is the number of iterations.

This lower bound is derived from information-theoretic arguments and holds for a wide range of convex loss functions, regardless of their specific form. Therefore, even if an algorithm is restricted to quadratic loss functions, it cannot avoid the $\Omega(1/T^2)$ lower bound if it only uses first-order information. The lower bound serves as a fundamental limit on the convergence rate of optimization algorithms in the first-order oracle model, irrespective of the type of loss function being considered.

(c) (5pts) What is the true reason that the result in part (a) can avoid the $\Omega(1/T^2)$ lower bound?

Solution:
The Newton's method update rule you provided, $\mathbf{w}_{t+1} = \mathbf{w}_t - \nabla^2\mathcal{L}(\mathbf{w}_t)^{-1}\nabla\mathcal{L}(\mathbf{w}_t)$, is indeed a

second-order optimization method. Newton's method can achieve faster convergence rates than first-order methods for certain types of functions.

The $\Omega(1/T^2)$ lower bound is typically associated with first-order optimization methods, such as gradient descent, and it indicates a fundamental limit on the convergence rate of algorithms that only use first-order information (gradients). However, second-order methods like Newton's method can achieve faster convergence rates, and they are not subject to the same $\Omega(1/T^2)$ lower bound.

The reason Newton's method can avoid the $\Omega(1/T^2)$ lower bound is that it utilizes second-order information (the Hessian matrix) in addition to the first-order information (gradients). This allows Newton's method to take advantage of curvature information in the objective function, leading to potentially faster convergence.

In summary, the avoidance of the $\Omega(1/T^2)$ lower bound is due to the use of second-order information in Newton's method, which is not constrained by the same limitations as first-order methods in terms of convergence rates.

5. (10pts) Let us call a differentiable function q-star-convex if for all $\mathbf{w}$, $\mathcal{L}(\mathbf{w}_\star) \geq \mathcal{L}(\mathbf{w}) + q\langle \nabla \mathcal{L}(\mathbf{w}), \mathbf{w}_\star - \mathbf{w} \rangle$. Convex functions satisfy this condition with $q = 1$ (although $q = 1$ does not imply convexity). $q > 1$ indicates some non-convexity. Suppose that $\mathcal{L} = \mathbb{E}[\ell(\mathbf{w}, z)]$ is q-star-convex, and satisfies $\|\mathbf{w}_\star\| \leq D$ for some known D. Further, suppose that $\ell(\mathbf{w}, z)$ is differentiable and G-Lipschitz in $\mathbf{w}$ for all z. Show that stochastic gradient descent with an appropriate learning rate guarantees:

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

Solution:
We can express the left-hand form in terms of SGD Iterates:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\ell(\mathbf{w}_t, z)] - \mathcal{L}(\mathbf{w}_\star)\right]. \tag{1}$$

By applying the q-star-Convexity, we can obtain:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] \leq \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} q\langle \nabla\, \mathbb{E}[\ell(\mathbf{w}_t, z)], \mathbf{w}_\star - \mathbf{w}_t \rangle\right]. \tag{2}$$

Then, we use the Expectation Properties:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] \leq q\langle \frac{1}{T}\sum_{t=1}^{T} \nabla\, \mathbb{E}[\ell(\mathbf{w}_t, z)], \mathbf{w}_\star - \mathbf{w}_t \rangle. \tag{3}$$

Next, we express gradient in terms of Expectation:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] \leq q\langle \nabla\, \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \ell(\mathbf{w}_t, z)\right], \mathbf{w}_\star - \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t \rangle. \tag{4}$$

By using the Lipschitz Property, we have:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] \leq qG \left\| \mathbf{w}_\star - \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t \right\|. \tag{5}$$

Further, we apply the Cauchy-Schwarz Inequality:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathcal{L}(\mathbf{w}_t)-\mathcal{L}(\mathbf{w}_\star)] \le qG\left(\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{w}_\star-\mathbf{w}_t\|^2\right)^{1/2}. \tag{6}$$

Then, through bounding the Sum of Squares:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathcal{L}(\mathbf{w}_t)-\mathcal{L}(\mathbf{w}_\star)] \le qG\left(\frac{1}{T}\sum_{t=1}^{T}D^2\right)^{1/2}. \tag{7}$$

Nest, simplifying this expression:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathcal{L}(\mathbf{w}_t)-\mathcal{L}(\mathbf{w}_\star)] \le qG\frac{D}{\sqrt{T}}. \tag{8}$$

Finally, we conclude with the Big-O notation:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathcal{L}(\mathbf{w}_t)-\mathcal{L}(\mathbf{w}_\star)] \le O\left(\frac{qDG}{\sqrt{T}}\right). \tag{9}$$

□

6. Consider the following "communication constrained" situation: A "server" holds a training dataset, and when provided with a point $\mathbf{w}$ it will be able to randomly sample some example z and compute a gradient of a loss $\ell(\mathbf{w},z)$ such that $\mathbb{E}[\ell(\mathbf{w},z)]=\mathcal{L}(\mathbf{w})$ for some H-smooth function $\mathcal{L}$. It is guaranteed that each coordinate of $\nabla\ell(\mathbf{w},z)\in\mathbb{R}^d$ lies in $[-1,1]$ for all $\mathbf{w}$ and z. Unfortunately, in order to preserve outgoing bandwidth, the server will not tell you the actual gradients $\nabla\ell(\mathbf{w},z)$ it computes. Instead it will give you a d-bit string $C(\nabla\ell(\mathbf{w},z))\in\{\pm1\}^d$ (the C stands for "compressed"). The ith coordinate of $C(\nabla\ell(\mathbf{w},z))$ is set randomly by the formula:

$$C(\nabla\ell(\mathbf{w},z))[i]=\begin{cases}1 & \text{with probability } \frac{1+\nabla\ell(\mathbf{w},z)[i]}{2}\\ -1 & \text{with probability } \frac{1-\nabla\ell(\mathbf{w},z)[i]}{2}\end{cases}$$

(a) (5pts) Show that $\mathbb{E}[C(\nabla\ell(\mathbf{w},z))]=\nabla\mathcal{L}(\mathbf{w})$, where the expectation is over both the randomness in the choice of z as well as the randomness in the function C.

Solution:
First, we analyze the expectation of the compressed gradient under the given compression scheme:

$$\mathbb{E}[C(\nabla\ell(\mathbf{w},z))]=\mathbb{E}\left[\begin{pmatrix}C(\nabla\ell(\mathbf{w},z))[1]\\ C(\nabla\ell(\mathbf{w},z))[2]\\ \vdots\\ C(\nabla\ell(\mathbf{w},z))[d]\end{pmatrix}\right]. \tag{1}$$

Then, let's consider the expectation of each coordinate separately:

$$\mathbb{E}[C(\nabla\ell(\mathbf{w},z))[i]]=\frac{1+\nabla\ell(\mathbf{w},z)[i]}{2}\cdot 1+\frac{1-\nabla\ell(\mathbf{w},z)[i]}{2}\cdot(-1). \tag{2}$$

Simplify the expression:

$$\mathbb{E}[C(\nabla\ell(\mathbf{w},z))[i]]=\frac{1+\nabla\ell(\mathbf{w},z)[i]-(1-\nabla\ell(\mathbf{w},z)[i])}{2}=\nabla\ell(\mathbf{w},z)[i]. \tag{3}$$

Now, consider the expectation over all coordinates:

$$\mathbb{E}[C(\nabla \ell(\mathbf{w}, z))] = \begin{pmatrix} \mathbb{E}[C(\nabla \ell(\mathbf{w}, z))[1]] \\ \mathbb{E}[C(\nabla \ell(\mathbf{w}, z))[2]] \\ \vdots \\ \mathbb{E}[C(\nabla \ell(\mathbf{w}, z))[d]] \end{pmatrix} = \begin{pmatrix} \nabla \ell(\mathbf{w}, z)[1] \\ \nabla \ell(\mathbf{w}, z)[2] \\ \vdots \\ \nabla \ell(\mathbf{w}, z)[d] \end{pmatrix} = \nabla \mathcal{L}(\mathbf{w}). \tag{4}$$

Here, we have used the linearity of expectations and the fact that the expected value of each coordinate of the compressed gradient is equal to the corresponding coordinate of the true gradient. Therefore, the overall expected compressed gradient is equal to the true gradient $\nabla \mathcal{L}(\mathbf{w})$. □

(b) (10 pts) Suppose you perform SGD with these compressed gradients:

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

Show that after T iterations, with an appropriate learning rate,

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] \leq O\left(\frac{\sqrt{dH(\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star))}}{\sqrt{T}} \right)$$

Solution:
First, we use the H-smooth Property:
The given H-smooth property is:

$$\mathcal{L}(\mathbf{w}') \leq \mathcal{L}(\mathbf{w}) + \langle \nabla \mathcal{L}(\mathbf{w}), \mathbf{w}' - \mathbf{w} \rangle + \frac{H}{2} \|\mathbf{w}' - \mathbf{w}\|^2. \tag{1}$$

This property is fundamental for analyzing the smoothness of the objective function $\mathcal{L}(\mathbf{w})$.

Then, we apply the SGD Update Rule. Using the SGD update rule $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta C(\nabla \ell(\mathbf{w}_t, z))$, the H-smooth property is applied to $\mathcal{L}(\mathbf{w}_{t+1})$ to obtain the inequality:

$$\mathcal{L}(\mathbf{w}_{t+1}) \leq \mathcal{L}(\mathbf{w}_t) - \eta \langle \nabla \mathcal{L}(\mathbf{w}_t), C(\nabla \ell(\mathbf{w}_t, z)) \rangle + \frac{\eta^2 H}{2} \|C(\nabla \ell(\mathbf{w}_t, z))\|^2. \tag{2}$$

Next, we consider Expectations and Properties of Compressed Gradients. Taking the expectation of the above inequality and using the property that each element of $C(\nabla \ell(\mathbf{w}, z))$ is in $\{\pm 1\}$, the following bound is obtained:

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_{t+1})] \leq \mathbb{E}[\mathcal{L}(\mathbf{w}_t)] - \eta \, \mathbb{E}[\langle \nabla \mathcal{L}(\mathbf{w}_t), C(\nabla \ell(\mathbf{w}_t, z)) \rangle] + \frac{\eta^2 H d}{2}. \tag{3}$$

Furthermore, we use Linearity of Expectation and Property of Expected Compressed Gradients. Since $\mathbb{E}[C(\nabla \ell(\mathbf{w}, z))] = \nabla \mathcal{L}(\mathbf{w})$, the expectation term $\mathbb{E}[\langle \nabla \mathcal{L}(\mathbf{w}_t), C(\nabla \ell(\mathbf{w}_t, z)) \rangle]$ is simplified to $\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2$.

Then, by summing over Iterations and rearranging: summing over all $t = 1, \ldots, T$ and using telescoping sums, the inequality becomes:

$$\mathcal{L}(\mathbf{w}_1) - \mathbb{E}[\mathcal{L}(\mathbf{w}_{T+1})] \geq -\eta \sum_{t=1}^{T} \mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] + \frac{\eta^2 H d T}{2}. \tag{4}$$

Next, by choosing the Appropriate Learning Rate: To minimize the right-hand side, an appropriate learning rate η is chosen:

$$\eta = \sqrt{\frac{2(\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star))}{HdT}}. \tag{5}$$

Finally, by substituting η and concluding: Substituting this chosen η back into the inequality and simplifying, the final result is obtained:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] \leq \frac{\sqrt{Hd(\mathcal{L}(\mathbf{w}_1)-\mathcal{L}(\mathbf{w}_\star))}}{\sqrt{T}}. \tag{6}$$

The entire sequence of steps establishes a rigorous argument that, with an appropriately chosen learning rate, the average squared norm of the gradient during SGD iterations converges to a bound with a rate of $\frac{\sqrt{Hd(\mathcal{L}(\mathbf{w}_1)-\mathcal{L}(\mathbf{w}_\star))}}{\sqrt{T}}$. This analysis is crucial in understanding the convergence behavior of SGD in the context of H-smooth functions and provides guidance on selecting the learning rate for optimal convergence. □