ELEG/CISC 867: Advanced Machine Learning **Spring 2019**

Lecture 10: Stochastic Gradient Descent

Lecturer: Xiugang Wu *04/23/2019, 04/25/2019*

Recall that our ultimate goal in an inference problem is to apply the Bayes predictor

$$\operatorname*{argmin}_h L(h, P)$$

to an instance X; however, since the underlying distribution P is unknown we cannot directly minimize the true risk $L(h, P)$ to find the Bayes predictor. Hence, the problem of learning arises where we want to find an approximation to the Bayes predictor based on an i.i.d. sequence of n training examples $Z^n = \{(X_i, Y_i)\}_{i=1}^n$. To achieve this, we have discussed the Empirical Risk Minimization (ERM) rule where one picks hypothesis h that minimizes the empirical risk $L(h, Z^n)$ over the hypothesis class $\mathcal{H}$; we will also see the Regularized Risk Minimization (RLM) rule where one picks hypothesis h that jointly minimizes $L(h, Z^n)$ and a regularization function. In this lecture, we introduce a different approach, called Stochastic Gradient Descent (SGD), where we try to minimize the true risk using a gradient descent procedure.

1 Gradient Descent

We first describe the standard gradient descent (GD) approach for minimizing a differentiable convex function $f(\mathbf{w})$. Gradient descent is an iterative algorithm. We start with an initial value of $\mathbf{w}$, say, $\mathbf{w}^{(1)} = \mathbf{0}$. Then at each iteration, we take a step in the direction of the negative of the gradient at the current point, i.e.,

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \nabla f(\mathbf{w}^{(t)}) \tag{1}$$

where $\eta > 0$ is called the step size and will be discussed shortly; intuitively, since the gradient points in the direction of the greatest rate of increase of f around $\mathbf{w}^{(t)}$, the algorithm makes a small step in the opposite direction, thus decreasing the value of the function. After T iterations, the algorithm outputs a vector that is close to the minimizer of the function f; this output vector could be the average vector

$$\bar{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{w}^{(t)},$$

or the last vector

$$\mathbf{w}^{(T)},$$

or the best performing vector

$$\operatorname*{argmin}_{t \in [1:T]} \mathbf{w}^{(t)}.$$

Here we adopt the average vector, which turns out to be quite useful especially when we generalize gradient descent to non-differentiable functions and to the stochastic case.

One can interpret gradient descent by looking at the Taylor approximation of f. At iteration t, the first order Taylor approximation of f around $\mathbf{w}^{(t)}$ is given by:

$$f(\mathbf{w}) \approx f(\mathbf{w}^{(t)}) + \nabla f(\mathbf{w}^{(t)}) \cdot (\mathbf{w} - \mathbf{w}^{(t)});$$

and how tight this approximation is depends on how close $\mathbf{w}$ is to $\mathbf{w}^{(t)}$. Hence, to minimize $f(\mathbf{w})$, we can minimize jointly the distance between $\mathbf{w}$ and $\mathbf{w}^{(t)}$, and the approximation of f around $\mathbf{w}^{(t)}$:

$$\mathbf{w}^{(t+1)} = \operatorname*{argmin}_{\mathbf{w}} \frac{1}{2}\|\mathbf{w} - \mathbf{w}^{(t)}\|^2 + \eta \left[f(\mathbf{w}^{(t)}) + \nabla f(\mathbf{w}^{(t)}) \cdot (\mathbf{w} - \mathbf{w}^{(t)}) \right].$$

Taking the derivative of the above and letting it be zero, we obtain the update rule (1).

1.1 Analysis of GD for Convex-Lipschitz Functions

Suppose f is a convex-Lipschitz function. Fix an arbitrarily $\mathbf{w}^*$ with $\|\mathbf{w}^*\| \leq B$; in particular, $\mathbf{w}^*$ here could be the minimizer of f. We now derive an upper bound on the suboptimality of the GD solution with respect to $\mathbf{w}^*$, i.e. $f(\bar{\mathbf{w}}) - f(\mathbf{w}^*)$. From the definition of $\bar{\mathbf{w}}$ and using Jensen's inequality, we have

$$\begin{aligned}
f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) &= f\left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{w}^{(t)}\right) - f(\mathbf{w}^*) \\
&\leq \frac{1}{T}\sum_{t=1}^{T} f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*) \\
&= \frac{1}{T}\sum_{t=1}^{T} \left[f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*) \right] \\
&\leq \frac{1}{T}\sum_{t=1}^{T} \nabla f(\mathbf{w}^{(t)}) \cdot (\mathbf{w}^{(t)} - \mathbf{w}^*)
\end{aligned}$$

where the last inequality follows because f is convex and hence

$$f(\mathbf{w}^*) \geq f(\mathbf{w}^{(t)}) + \nabla f(\mathbf{w}^{(t)}) \cdot (\mathbf{w}^* - \mathbf{w}^{(t)}).$$

To proceed we need the following lemma, whose proof can be found in the textbook; see Lemma 14.1 on Page 187.

Lemma 1.1 *Let $\mathbf{v}_1, \ldots, \mathbf{v}_T$ be an arbitrary sequence of vectors. Any algorithm with an initialization $\mathbf{w}^{(1)} = \mathbf{0}$ and an update rule*

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \mathbf{v}_t$$

satisfies

$$\sum_{t=1}^{T} \mathbf{v}_t \cdot (\mathbf{w}^{(t)} - \mathbf{w}^*) \leq \frac{\|\mathbf{w}^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \|\mathbf{v}_t\|^2.$$

In particular, for every $B, \rho > 0$, if for all t we have $\|\mathbf{v}_t\| \leq \rho$ and if we set

$$\eta = \frac{B}{\rho\sqrt{T}},$$

then for every $\mathbf{w}^*$ with $\|\mathbf{w}^*\| \leq B$ we have

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{v}_t \cdot (\mathbf{w}^{(t)} - \mathbf{w}^*) \leq \frac{B\rho}{\sqrt{T}}.$$

We now apply the above lemma to the GD algorithm. In particular, replace the sequence of vectors $\mathbf{v}_t$ in the lemma with the gradients $\nabla f(\mathbf{w}^{(t)})$ in our GD algorithm, and note that $\|\nabla f(\mathbf{w}^{(t)})\| \leq \rho$ if f is ρ-Lipschitz (see Prop. 2.2). This allows us to reach the following result:

Theorem 1.1 *Let f be a convex, ρ-Lipschitz function and let $\mathbf{w}^* \in \operatorname{argmin}_{\mathbf{w}:\|\mathbf{w}\|\leq B} f(\mathbf{w})$. If we run the GD algorithm on f for T steps with step size $\eta = \frac{B}{\rho\sqrt{T}}$, then the output vector $\bar{\mathbf{w}}$ satisfies that*

$$f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) \leq \frac{B\rho}{\sqrt{T}}.$$

Therefore, to achieve $f(\bar{\mathbf{w}}) - f(\mathbf{w}^*) \leq \epsilon$ for any $\epsilon > 0$, it suffices to run GD with step size $\eta = \frac{B}{\rho\sqrt{T}}$ for $T \geq \frac{B^2\rho^2}{\epsilon^2}$ iterations.

2 Subgradients

GD can be generalized to nondifferentiable functions by using the so-called subgradient of f instead of the gradient. To motivate the definition of subgradients, recall from the last lecture that the existence of a tangent that lies below f is an important property of convex functions, i.e. if f is convex and differentiable, then

$$f(\mathbf{u}) \geq f(\mathbf{w}) + \nabla f(\mathbf{w}) \cdot (\mathbf{u} - \mathbf{w}), \forall \mathbf{u}.$$

The following proposition says that this property is in fact an alternative characterization of convexity for general functions that may not be differentiable.

Proposition 2.1 *Let S be an open convex set. A function $f : S \to \mathbb{R}$ is convex if and only if for any $\mathbf{w} \in S$ there exists some $\mathbf{v}$ such that*

$$f(\mathbf{u}) \geq f(\mathbf{w}) + \mathbf{v} \cdot (\mathbf{u} - \mathbf{w}), \forall \mathbf{u} \in S. \tag{2}$$

This leads us to the following definition of subgradients:

Definition 2.1 *A vector that satisfies (2) is called a subgradient of f at $\mathbf{w}$. The set of subgradients of f at $\mathbf{w}$ is called the differential set at $\mathbf{w}$ and denoted $\partial f(\mathbf{w})$.*

2.1 Calculating Subgradients

If f is differentiable at $\mathbf{w}$ then $\partial f(\mathbf{w})$ contains a single element, i.e. $\nabla f(\mathbf{w})$. For example, the absolute value function $f(x) = |x|$ has the differential set $\partial f(x)$ equal to $\{1\}$ for $x > 0$ and $\{-1\}$ for $x < 0$; however, for $x = 0$, its subdifferential can be any number between -1 and 1 and therefore $\partial f(x) = [-1, 1]$.

For many practical uses, it suffices to find just one subgradient at a given point and there is no need to determine the whole differential set. For pointwise maximum of convex functions, this task can be easily achieved. In particular, let $g(\mathbf{w}) = \max_{i\in[1:r]} g_i(\mathbf{w})$ where g_i is convex differentiable for any $i \in [1 : r]$. Then $\nabla g_j(\mathbf{w}) \in \partial g(\mathbf{w})$ for any $j \in \operatorname{argmax}_{i\in[1:r]} g_i(\mathbf{w})$. See Page 189 in the textbook for a simple proof of this result.

2.2 Subgradients of Lipschitz Functions

Recall that a function $f : A \to \mathbb{R}$ is ρ-Lipschitz if for all $\mathbf{u}, \mathbf{v} \in A$,

$$|f(\mathbf{u}) - f(\mathbf{v})| \leq \rho \|\mathbf{u} - \mathbf{v}\|.$$

The following proposition give an equivalent definition of Lipschitzness for convex functions using norms of subgradients. Proof of this proposition can be found on Page 190 of the textbook.

Proposition 2.2 *Let A be an open convex set. A convex function $f : A \to \mathbb{R}$ is ρ-Lipschitz if and only if for any $\mathbf{w} \in A$ and $\mathbf{v} \in \partial f(\mathbf{w})$ it holds that $\|\mathbf{v}\| \leq \rho$.*

2.3 Subgradient Descent

The gradient descent algorithm can be generalized to nondifferentiable functions by using a subgradient of $f(\mathbf{w})$ at $\mathbf{w}^{(t)}$, i.e. $\mathbf{v} \in \partial f(\mathbf{w}^{(t)})$, instead of the gradient. The analysis of the convergence rate remains unchanged—this is because the crucial step in the analysis is to use the existence of a tangent that lies below f, which can also be yielded by using subgradients for nondifferentiable functions.

3 Stochastic Gradient Descent

In stochastic gradient descent we don't require the update direction $\mathbf{v}_t$ to be exactly based on the gradient. Instead, we allow the direction $\mathbf{V}_t$ to be a random vector and only require that its expected value at each iteration equals the gradient (or more generally, a subgradient) at vector $\mathbf{W}^{(t)}$. As we will see shortly, in the context of learning problems, it is easy to find a random vector whose expectation is a subgradient of the risk function.

Algorithm 1 SGD for minimizing $f(\mathbf{w})$

input: step size $\eta > 0$, number of iterations T
initialize: $\mathbf{W}^{(1)} \leftarrow \mathbf{0}$
for $t = 1, \ldots, T$ **do**
    choose $\mathbf{V}_t$ at random from a distribution such that $\mathbb{E}[\mathbf{V}_t | \mathbf{W}^{(t)}] \in \partial f(\mathbf{W}^{(t)})$
    update $\mathbf{W}^{(t+1)} \leftarrow \mathbf{W}^{(t)} - \eta \mathbf{V}_t$
end for
output $\bar{\mathbf{W}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{W}^{(t)}$

3.1 Analysis of SGD for Convex-Lipschitz-Bounded Functions

Similar to the bound on the performance of the GD output, we can prove a bound on the expected performance of the SGD output.

Theorem 3.1 *Let $B, \rho > 0$. Let f be a convex function and let $\mathbf{w}^* \in \operatorname{argmin}_{\mathbf{w}:\|\mathbf{w}\| \leq B} f(\mathbf{w})$. Assume that SGD is run on f for T steps with step size $\eta = \frac{B}{\rho\sqrt{T}}$, and assume that $\|\mathbf{V}_t\| \leq \rho$ with probability 1 for all t. Then the output vector $\bar{\mathbf{W}}$ satisfies that*

$$\mathbb{E}[f(\bar{\mathbf{W}}) - f(\mathbf{w}^*)] \leq \frac{B\rho}{\sqrt{T}}.$$

Therefore, to achieve $\mathbb{E}[f(\bar{\mathbf{W}}) - f(\mathbf{w}^*)] \leq \epsilon$ *for any* $\epsilon > 0$*, it suffices to run SGD with step size* $\eta = \frac{B}{\rho\sqrt{T}}$ *for* $T \geq \frac{B^2\rho^2}{\epsilon^2}$ *iterations.*

Proof: Recall from the analysis of GD that

$$f(\bar{\mathbf{W}}) - f(\mathbf{w}^*) \leq \frac{1}{T}\sum_{t=1}^{T}\left[f(\mathbf{W}^{(t)}) - f(\mathbf{w}^*)\right].$$

Taking expectation on both sides of the above inequality yields that

$$\mathbb{E}\left[f(\bar{\mathbf{W}}) - f(\mathbf{w}^*)\right] \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[f(\mathbf{W}^{(t)}) - f(\mathbf{w}^*)\right].$$

On the other hand, applying Lemma 1.1 to SGD, we obtain

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathbf{V}_t \cdot (\mathbf{W}^{(t)} - \mathbf{w}^*)\right] \leq \frac{B\rho}{\sqrt{T}}.$$

Therefore, to prove the theorem it suffices to show that

$$\mathbb{E}\left[f(\mathbf{W}^{(t)}) - f(\mathbf{w}^*)\right] \leq \mathbb{E}\left[\mathbf{V}_t \cdot (\mathbf{W}^{(t)} - \mathbf{w}^*)\right].$$

This is indeed true because

$$\begin{aligned}
\mathbb{E}\left[\mathbf{V}_t \cdot (\mathbf{W}^{(t)} - \mathbf{w}^*)\right] &= \mathbb{E}\left[\mathbb{E}[\mathbf{V}_t \cdot (\mathbf{W}^{(t)} - \mathbf{w}^*)|\mathbf{W}^{(t)}]\right] \\
&= \mathbb{E}\left[\mathbb{E}[\mathbf{V}_t|\mathbf{W}^{(t)}] \cdot (\mathbf{W}^{(t)} - \mathbf{w}^*)\right] \\
&\geq \mathbb{E}\left[f(\mathbf{W}^{(t)}) - f(\mathbf{w}^*)\right],
\end{aligned}$$

where the last inequality follows because by assumption $\mathbb{E}[\mathbf{V}_t|\mathbf{W}^{(t)}] \in \partial f(\mathbf{W}^{(t)})$. □

4 Learning with SGD

Having analyzed SGD for general convex functions, we now consider applying it to learning tasks. Recall that in learning we face the problem of minimizing the risk $L(\mathbf{w}, P) = \mathbb{E}_{Z\sim P}[\ell(\mathbf{w}, Z)]$. We have seen the ERM learning rule where one minimizes the empirical risk $L(\mathbf{w}, Z^n) = \frac{1}{n}\ell(\mathbf{w}, Z_i)$ as an estimate to minimizing the true risk $L(\mathbf{w}, P)$. We now take a different approach, that is, we will minimize $L(\mathbf{w}, P)$ directly using SGD without explicitly calculating $L(\mathbf{w}, P)$. Indeed, since we don't know P, we don't have the analytical form of $L(\mathbf{w}, P)$ and GD is not applicable here for minimizing $L(\mathbf{w}, P)$; however, with SGD, all we need is to find an unbiased estimate of the gradient $L(\mathbf{w}, P)$, i.e. a random vector $\mathbf{V}_t$ whose conditional expectation given $\mathbf{W}^{(t)}$ is the gradient of $L(\mathbf{w}, P)$ at $\mathbf{W}^{(t)}$. We now describe how to construct such an estimate.

First consider the case when the loss function $\ell(\mathbf{w}, z)$ is differentiable for any $z \in \mathcal{Z}$; in this case, $L(\mathbf{w}, P)$ is obviously also differentiable. Suppose an i.i.d. sequence of training examples Z^n is sampled from distribution P. At each iteration t of SGD where $t \leq n$, choose $\mathbf{V}_t = \nabla\ell(\mathbf{W}^{(t)}, Z_t)$ and update $\mathbf{W}^{(t)}$ to $\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \eta\mathbf{V}_t$. Note that here $\mathbf{W}^{(t)}$ is determined by $\mathbf{V}_{[1:t-1]}$ (which is further determined by $Z_{[1:t-1]}$), and is independent of Z_t; hence we have

$$\mathbb{E}[\mathbf{V}_t|\mathbf{W}^{(t)}] = \mathbb{E}[\nabla\ell(\mathbf{W}^{(t)}, Z_t)|\mathbf{W}^{(t)}] = \nabla\mathbb{E}[\ell(\mathbf{W}^{(t)}, Z_t)|\mathbf{W}^{(t)}] = \nabla L(\mathbf{W}^{(t)}, P),$$

i.e. the gradient of the loss function $\nabla \ell(\mathbf{w}, Z_t)$ at $\mathbf{W}^{(t)}$ is an unbiased estimate of the gradient of the risk function $\nabla L(\mathbf{W}^{(t)}, P)$.

The above argument can be extended to nondifferentiable convex loss functions by taking $\mathbf{V}_t \in \partial \ell(\mathbf{W}^{(t)}, Z_t)$, i.e. choosing $\mathbf{V}_t$ to be a subgradient of $\ell(\mathbf{w}, Z_t)$ at $\mathbf{W}^{(t)}$. As such, we have

$$\ell(\mathbf{u}, Z_t) - \ell(\mathbf{W}^{(t)}, Z_t) \geq \mathbf{V}_t \cdot (\mathbf{u} - \mathbf{W}^{(t)}), \ \forall \mathbf{u},$$

and hence, by taking conditional expectation given $\mathbf{W}^{(t)}$ at both sides of the above inequality,

$$L(\mathbf{u}, P) - L(\mathbf{W}^{(t)}, P) \geq \mathbb{E}[\mathbf{V}_t | \mathbf{W}^{(t)}] \cdot (\mathbf{u} - \mathbf{W}^{(t)}), \ \forall \mathbf{u}.$$

Therefore, $\mathbb{E}[\mathbf{V}_t | \mathbf{W}^{(t)}]$ is indeed a subgradient of $L(\mathbf{w}, P)$ at $\mathbf{W}^{(t)}$, i.e. $\mathbb{E}[\mathbf{V}_t | \mathbf{W}^{(t)}] \in \partial \ell(\mathbf{W}^{(t)}, P)$. This leads us to the following SGD framework for directly minimizing the risk function $L(\mathbf{w}, P)$ without explicitly knowing the analytical expression of $L(\mathbf{w}, P)$.

Algorithm 2 SGD for minimizing $L(\mathbf{w}, P)$

1: **input:** training data Z^n, step size $\eta > 0$, number of iterations $T \in [1 : n]$
2: **initialize:** $\mathbf{W}^{(1)} \leftarrow \mathbf{0}$
3: **for** $t = 1, \ldots, T$ **do**
4: pick $\mathbf{V}_t \in \partial \ell(\mathbf{W}^{(t)}, Z_t)$
5: update $\mathbf{W}^{(t+1)} \leftarrow \mathbf{W}^{(t)} - \eta \mathbf{V}_t$
6: **end for**
7: **output** $\bar{\mathbf{W}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{W}^{(t)}$

4.1 SGD for Convex-Lipschitz-Bounded Learning Problems

The following result is immediate given the previous analysis of SGD for convex-Lipschitz-bounded functions.

Theorem 4.1 *Consider a convex-Lipschitz-bounded learning problem with parameter ρ, B. If we run SGD to minimize $L(\mathbf{w}, P)$ with step size $\eta = \frac{B}{\rho\sqrt{T}}$ for T iterations, then the output vector $\bar{\mathbf{W}}$ satisfies that*

$$\mathbb{E}[L(\bar{\mathbf{W}}, P)] \leq L(\mathbf{w}^*, P) + \frac{B\rho}{\sqrt{T}}.$$

Therefore, to achieve

$$\mathbb{E}[L(\bar{\mathbf{W}}, P)] \leq L(\mathbf{w}^*, P) + \epsilon$$

for any $\epsilon > 0$, it suffices to run SGD with step size $\eta = \frac{B}{\rho\sqrt{T}}$ for $T \geq \frac{B^2\rho^2}{\epsilon^2}$ iterations.

Note that the above theorem not only tells us convex-Lipschitz-bounded learning problems are indeed learnable, but also provides an upper bound on the sample complexity; in particular, we have shown that with $n \geq \frac{B^2\rho^2}{\epsilon^2}$ training examples one can guarantee the expected risk of the learned predictor to be bounded by the approximation error of the class $\min_{\mathbf{w} \in \mathcal{H}} L(\mathbf{w}, P)$ plus an arbitrarily small ϵ irrespective of the underlying distribution P.

4.2 SGD for Convex-Smooth-Bounded Learning Problems

Since the loss function is differentiable in convex-smooth-bounded problems, we simply take $\mathbf{V}_t = \nabla \ell(\mathbf{W}^{(t)}, Z_t)$. The following theorem shows that convex-smooth-bounded problems are also learnable with SGD.

Theorem 4.2 *Consider a convex-smooth-bounded learning problem with parameter β, B. If we run SGD to minimize $L(\mathbf{w}, P)$ with step size η for T iterations, then the output vector $\bar{\mathbf{W}}$ satisfies that*

$$\mathbb{E}[L(\bar{\mathbf{W}}, P)] \leq \frac{1}{1-\eta\beta}\left(L(\mathbf{w}^*, P) + \frac{B^2}{2\eta T}\right).$$

Therefore, if $L(\mathbf{w}^*, P) \leq 1$ then running SGD with $\eta = \frac{1}{\beta(1+3/\epsilon)}$ and $T \geq \frac{12B^2\beta}{\epsilon^2}$ yields

$$\mathbb{E}[L(\bar{\mathbf{W}}, P)] \leq L(\mathbf{w}^*, P) + \epsilon.$$

Proof: Along the similar lines as before, we have

$$\begin{aligned}\sum_{t=1}^{T}[\ell(\mathbf{W}^{(t)}, Z_t) - \ell(\mathbf{w}^*, Z_t)] &\leq \sum_{t=1}^{T}\nabla\ell(\mathbf{W}^{(t)}, Z_t)\cdot(\mathbf{W}^{(t)} - \mathbf{w}^*) \\ &\leq \frac{B^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|\nabla\ell(\mathbf{W}^{(t)}, Z_t)\|^2.\end{aligned}$$

Since $\ell(\mathbf{W}^{(t)}, Z_t)$ is β-smooth and nonnegative, it is self-bounded:

$$\|\nabla\ell(\mathbf{W}^{(t)}, Z_t)\|^2 \leq 2\beta\ell(\mathbf{W}^{(t)}, Z_t).$$

Therefore, we have

$$\sum_{t=1}^{T}[\ell(\mathbf{W}^{(t)}, Z_t) - \ell(\mathbf{w}^*, Z_t)] \leq \frac{B^2}{2\eta} + \eta\beta\sum_{t=1}^{T}\ell(\mathbf{W}^{(t)}, Z_t),$$

i.e.,

$$\frac{1}{T}\sum_{t=1}^{T}\ell(\mathbf{W}^{(t)}, Z_t) \leq \frac{1}{1-\eta\beta}\left(\frac{1}{T}\sum_{t=1}^{T}\ell(\mathbf{w}^*, Z_t) + \frac{B^2}{2\eta T}\right).$$

Taking expectation at both sides of the above inequality proves the theorem. □