

Online Learning

The Perceptron algorithm accesses training data in a sequential fashion, processing each training example in time $\Theta(d)$ where d is the number of features. This fact, and the observation that a linear model \mathbf{w} can be stored in space $\Theta(d)$, make the Perceptron very competitive on large training sets, when we cannot afford a training time growing faster than linear in the number of data points.

Algorithms that learn sequentially, like the Perceptron, are also very good at dealing with scenarios in which new training data are generated at all times. For example: sensor data, financial data, user interaction data, and so on. In these cases, the traditional learning protocol, where predictors are generated by feeding a fixed-size training set to a learning algorithm, becomes inefficient. This happens because everytime new train data are available we would have to run again the algorithm from scratch.

This sequential learning protocol, which we call **online learning**, can be summarized as follows.

Parameters: Class \mathcal{H} of predictors, loss function ℓ .

The algorithm outputs a default initial predictor $h_1 \in \mathcal{H}$

For $t = 1, 2, \dots$

1. The next example (\mathbf{x}_t, y_t) is observed
2. The loss $\ell(h_t(\mathbf{x}_t), y_t)$ of the current predictor h_t is computed
3. The online learner updates h_t generating a new predictor $h_{t+1} \in \mathcal{H}$

A characterizing feature of online learning is that the model update $h_t \rightarrow h_{t+1}$ is typically *local*. That is, it only involves the current predictor h_t and the current example (\mathbf{x}_t, y_t) .

Note that an online learner A generates a sequence $h_1, h_2, \dots \in \mathcal{H}$ of predictors. We evaluate the performance of A through the notion of **sequential risk**,

$$\frac{1}{T} \sum_{t=1}^T \ell(h_t(\mathbf{x}_t), y_t)$$

measuring, as a function of T , the average loss of the predictor sequence over the first T examples. The sequential risk is the online learning counterpart of the notion of statistical risk in statistical learning.

In what follows, we use the notation $\ell_t(h) = \ell(h(\mathbf{x}_t), y_t)$ when the sequence $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots$ is understood from the context. This defines a sequence ℓ_1, ℓ_2, \dots of loss functions.

In keeping with the analogy between online and statistical learning, we also define the *regret*

$$\frac{1}{T} \sum_{t=1}^T \ell_t(h_t) - \min_{h \in \mathcal{H}} \frac{1}{T} \sum_{t=1}^T \ell_t(h)$$

which measures the difference between the sequential risk of h_1, \dots, h_T —generated by some online algorithm A — and the sequential risk of the best predictor in the class \mathcal{H} for the loss functions ℓ_1, \dots, ℓ_T . Regret can be viewed as the sequential counterpart of variance error in statistical learning.

We now introduce the online version of gradient descent, *online gradient descent* (OGD), which can be applied to any convex and differentiable loss function. Gradient descent is the workhorse of convex optimization. Given a convex and differentiable function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ which we want to minimize, gradient descent works by iterating $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t)$ starting from some initial point \mathbf{w}_1 in the domain of f , where $\eta > 0$ is a parameter. If the current point \mathbf{w}_t is not a minimum of f , then $\nabla f(\mathbf{w}_t) \neq \mathbf{0}$ and $\mathbf{w}_{t+1} - \mathbf{w}_t$ points in the direction opposite to $\nabla f(\mathbf{w}_t)$, which is—by definition of gradient—the direction where f decreases the most when moving away from \mathbf{w}_t . In order to analyze OGD, we must understand the behavior of gradient descent when the function to minimize changes at every step.

We focus on OGD applied to linear predictors $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, for $\mathbf{w} \in \mathbb{R}^d$. We use the notation $\ell_t(\mathbf{w}) = \ell(\mathbf{w}^\top \mathbf{x}_t, y_t)$ and assume that losses ℓ_1, ℓ_2, \dots are all convex and differentiable. For example, $\ell_t(\mathbf{w}) = (\mathbf{w}^\top \mathbf{x}_t - y_t)^2$.

Projected OGD

Parameters: $\eta > 0, U > 0$

Initialization: $\mathbf{w}_1 = \mathbf{0}$

For $t = 1, 2, \dots$

1. $\mathbf{w}'_{t+1} = \mathbf{w}_t - \frac{\eta}{\sqrt{t}} \nabla \ell_t(\mathbf{w}_t)$
2. $\mathbf{w}_{t+1} = \operatorname{argmin}_{\mathbf{w} : \|\mathbf{w}\| \leq U} \|\mathbf{w} - \mathbf{w}'_{t+1}\|$

In step 2, we project \mathbf{w}'_{t+1} in an Euclidean sphere of radius U . If $\|\mathbf{w}'_{t+1}\| \leq U$, then $\mathbf{w}_{t+1} = \mathbf{w}'_{t+1}$. Let $\eta_t = \eta/\sqrt{t}$, where $\eta > 0$ will be determined by the analysis.

Our goal is to control the regret

$$\frac{1}{T} \sum_{t=1}^T \ell_t(\mathbf{w}_t) - \frac{1}{T} \sum_{t=1}^T \ell_t(\mathbf{u}_T^*) \quad \text{where} \quad \mathbf{u}_T^* = \operatorname{argmin}_{\mathbf{u} : \|\mathbf{u}\| \leq U} \frac{1}{T} \sum_{t=1}^T \ell_t(\mathbf{u})$$

Note that \mathbf{u}_T^* is the predictor in the ball of radius U with smallest average loss over the first T steps. In what follows, we use the notation $R_T(\mathbf{u}) = \sum_{t=1}^T (\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{u}))$. The analysis of OGD is based on the following well-known result.

Lemma 1 (Taylor's formula for multivariate functions). *Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a twice differentiable function. Then, for all $\mathbf{w}, \mathbf{u} \in \mathbb{R}^d$,*

$$f(\mathbf{u}) = f(\mathbf{w}) + \nabla f(\mathbf{w})^\top (\mathbf{u} - \mathbf{w}) + \frac{1}{2} (\mathbf{u} - \mathbf{w})^\top \nabla^2 f(\boldsymbol{\xi}) (\mathbf{u} - \mathbf{w})$$

where $\nabla^2 f(\boldsymbol{\xi})$ is the Hessian matrix of f evaluated at a point $\boldsymbol{\xi}$ on the segment joining \mathbf{u} and \mathbf{w} .

If f is convex, then $\nabla^2 f$ is positive semidefinite, and so $\mathbf{z}^\top \nabla^2 f(\boldsymbol{\xi}) \mathbf{z} \geq 0$ for all $\mathbf{z}, \boldsymbol{\xi} \in \mathbb{R}^d$. This in turn implies

$$f(\mathbf{w}) - f(\mathbf{u}) \leq \nabla f(\mathbf{w})^\top (\mathbf{w} - \mathbf{u}). \quad (1)$$

This actually holds for any convex and differentiable f (i.e., f need not be twice differentiable). Now fix T , let $\mathbf{u} = \mathbf{u}_T^*$, and note that, for each $t = 1, 2, \dots$,

$$\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{u}) \leq \nabla \ell_t(\mathbf{w}_t)^\top (\mathbf{w}_t - \mathbf{u}) \quad (2)$$

$$= -\frac{1}{\eta_t} (\mathbf{w}'_{t+1} - \mathbf{w}_t)^\top (\mathbf{w}_t - \mathbf{u}) \quad (3)$$

$$= \frac{1}{\eta_t} \left(\frac{1}{2} \|\mathbf{w}_t - \mathbf{u}\|^2 - \frac{1}{2} \|\mathbf{w}'_{t+1} - \mathbf{u}\|^2 + \frac{1}{2} \|\mathbf{w}'_{t+1} - \mathbf{w}_t\|^2 \right) \quad (4)$$

$$\leq \frac{1}{\eta_t} \left(\frac{1}{2} \|\mathbf{w}_t - \mathbf{u}\|^2 - \frac{1}{2} \|\mathbf{w}_{t+1} - \mathbf{u}\|^2 + \frac{1}{2} \|\mathbf{w}'_{t+1} - \mathbf{w}_t\|^2 \right). \quad (5)$$

Inequality (2) is due to (1). Equality (3) uses $\mathbf{w}'_{t+1} - \mathbf{w}_t = -\eta_t \nabla \ell_t(\mathbf{w}_t)$. Equality (4) is an easily verified algebraic identity. Finally, inequality (5) holds because \mathbf{u} belong to the sphere of radius U centered at the origin. Hence, by projecting \mathbf{w}'_{t+1} onto this sphere, the distance to \mathbf{u} can not increase.

We now add and subtract the same term $\frac{1}{2\eta_{t+1}} \|\mathbf{w}_{t+1} - \mathbf{u}\|^2$ to the last member of the above chain of inequalities. Then, we regroup terms as indicated below here

$$\underbrace{\frac{1}{2\eta_t} \|\mathbf{w}_t - \mathbf{u}\|^2 - \frac{1}{2\eta_{t+1}} \|\mathbf{w}_{t+1} - \mathbf{u}\|^2}_{\text{telescoping}} - \underbrace{\frac{1}{2\eta_t} \|\mathbf{w}_{t+1} - \mathbf{u}\|^2 + \frac{1}{2\eta_{t+1}} \|\mathbf{w}_{t+1} - \mathbf{u}\|^2}_{\text{common factor}} + \frac{1}{2\eta_t} \|\mathbf{w}'_{t+1} - \mathbf{w}_t\|^2.$$

Summing over $t = 1, \dots, T$ we observe that the first pair of terms forms a telescopic sum, while the terms in the second pair have a common factor,

$$\begin{aligned} R_T(\mathbf{u}) &\leq \frac{1}{2\eta_1} \|\mathbf{w}_1 - \mathbf{u}\|^2 - \frac{1}{2\eta_{T+1}} \|\mathbf{w}_{T+1} - \mathbf{u}\|^2 \\ &\quad + \frac{1}{2} \sum_{t=1}^T \|\mathbf{w}_{t+1} - \mathbf{u}\|^2 \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) + \frac{1}{2} \sum_{t=1}^T \frac{1}{\eta_t} \|\mathbf{w}'_{t+1} - \mathbf{w}_t\|^2. \end{aligned} \quad (6)$$

Next, we make use of the following facts:

$$\begin{aligned} \mathbf{w}_1 &= \mathbf{0} && \text{by construction} \\ \|\mathbf{w}_{t+1} - \mathbf{u}\|^2 &\leq 4U^2 && \text{since both } \mathbf{w}_{t+1} \text{ and } \mathbf{u} \text{ belong to a sphere of radius } U \\ \|\mathbf{w}'_{t+1} - \mathbf{w}_t\|^2 &= \eta_t^2 \|\nabla \ell_t(\mathbf{w}_t)\|^2 && \text{by construction.} \end{aligned}$$

Substituting these relations in (6), and choosing G so that $\|\nabla\ell_t(\mathbf{w}_t)\| \leq G$ for all $t \leq T$, we obtain

$$\begin{aligned} R_T(\mathbf{u}) &\leq \frac{U^2}{2\eta_1} - \frac{1}{2\eta_{T+1}} \|\mathbf{w}_{T+1} - \mathbf{u}\|^2 \\ &\quad + 2U^2 \sum_{t=1}^{T-1} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) + \frac{1}{2\eta_{T+1}} \|\mathbf{w}_{T+1} - \mathbf{u}\|^2 - \frac{1}{2\eta_T} \|\mathbf{w}_{T+1} - \mathbf{u}\|^2 + \frac{G^2}{2} \sum_{t=1}^T \eta_t . \end{aligned}$$

We proceed by simplifying the telescopic sum, deleting terms with opposite signs, and dropping the term $-\frac{1}{2\eta_T} \|\mathbf{w}_{T+1} - \mathbf{u}\|^2$,

$$R_T(\mathbf{u}) \leq \frac{U^2}{2\eta_1} + \frac{2U^2}{\eta_T} - \frac{2U^2}{\eta_1} + \frac{G^2}{2} \sum_{t=1}^T \eta_t \leq \frac{2U^2\sqrt{T}}{\eta} + \frac{G^2\eta}{2} \sum_{t=1}^T \frac{1}{\sqrt{t}} \leq \frac{2U^2\sqrt{T}}{\eta} + G^2\eta\sqrt{T}$$

where we used the upper bound

$$\sum_{t=1}^T \frac{1}{\sqrt{t}} \leq 2\sqrt{T} .$$

Choosing $\eta = (U/G)\sqrt{2}$ and dividing everything by T we obtain the final regret bound

$$\frac{1}{T} \sum_{t=1}^T \ell_t(\mathbf{w}_t) \leq \min_{\mathbf{u}: \|\mathbf{u}\| \leq U} \frac{1}{T} \sum_{t=1}^T \ell_t(\mathbf{u}) + UG\sqrt{\frac{8}{T}} . \quad (7)$$

Explicit values for G may be obtained under specific assumptions. For example, in case of regression with square loss $\ell_t(\mathbf{w}) = (\mathbf{w}^\top \mathbf{x}_t - y_t)^2$, assuming $\|\mathbf{x}_t\| \leq X$ and $|y_t| \leq UX$ for all t we can compute

$$\|\nabla\ell_t(\mathbf{w}_t)\| \leq 2|\mathbf{w}^\top \mathbf{x}_t - y_t| \|\mathbf{x}_t\| \leq 2(\|\mathbf{w}_t\| \|\mathbf{x}_t\| + |y_t|) \|\mathbf{x}_t\| \leq 4UX^2 .$$

Substituting this value for G in the previous upper bound we get

$$\frac{1}{T} \sum_{t=1}^T \ell_t(\mathbf{w}_t) \leq \min_{\mathbf{u}: \|\mathbf{u}\| \leq U} \frac{1}{T} \sum_{t=1}^T \ell_t(\mathbf{u}) + 8(UX)^2 \sqrt{\frac{2}{T}} .$$

OGD and the Perceptron algorithm. Next, we show that Perceptron can be viewed as a variant of OGD without projection step. Observe that the Perceptron update rule can be written as

$$\mathbf{w}_{t+1} = \mathbf{w}_t + y_t \mathbf{x}_t \mathbb{I}\{y_t \mathbf{w}_t^\top \mathbf{x}_t \leq 0\} \quad (8)$$

Introduce the convex loss function $\ell_t(\mathbf{w}) = [-y_t \mathbf{w}^\top \mathbf{x}_t]_+$, where $[z]_+ = \max\{0, z\}$. Then

$$\nabla\ell_t(\mathbf{w}) = -y_t \mathbf{x}_t \mathbb{I}\{y_t \mathbf{w}^\top \mathbf{x}_t \leq 0\}$$

and thus we can write (8) as $\mathbf{w}_{t+1} = \mathbf{w}_t - \nabla\ell_t(\mathbf{w}_t)$. This solution has a problem though: we want to upper bound the Perceptron mistake count

$$\sum_t \mathbb{I}\{y_t \mathbf{w}_t^\top \mathbf{x}_t < 0\}$$

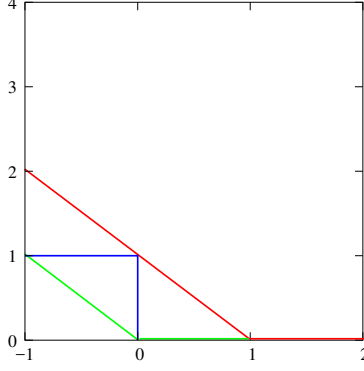


Figure 1: The derivative of $\ell(z) = [-z]_+$ (green line) corresponds to the Perceptron update rule. However, $\ell(z)$ does not upper bound the zero-one loss $\mathbb{I}\{z \leq 0\}$ (blue line). On the other hand, the hinge loss $h(z) = [1 - z]_+$ (red line) is a convex upper bound of the zero-one loss.

However, $\ell_t(\mathbf{w}) = [-y_t \mathbf{w}^\top \mathbf{x}_t]_+$ is not an upper bound of $\mathbb{I}\{y_t \mathbf{w}_t^\top \mathbf{x}_t < 0\}$. In order to fix this problem, we replace ℓ_t with the *hinge loss* function $h_t(\mathbf{w}) = [1 - y_t \mathbf{w}^\top \mathbf{x}_t]_+$ —see Figure 1. The hinge loss satisfies $\mathbb{I}\{y_t \mathbf{w}_t^\top \mathbf{x}_t \leq 0\} \leq h_t(\mathbf{w}_t)$. On the other hand, its gradient is $-y_t \mathbf{x}_t \mathbb{I}\{y_t \mathbf{w}_t^\top \mathbf{x}_t \leq 1\}$ which is different from $-y_t \mathbf{x}_t \mathbb{I}\{y_t \mathbf{w}_t^\top \mathbf{x}_t \leq 0\}$.

In order to express Perceptron as a variant of OGD, we then add the condition $\mathbb{I}\{y_t \mathbf{w}_t^\top \mathbf{x}_t \leq 0\}$. Noting that this condition implies $\mathbb{I}\{y_t \mathbf{w}_t^\top \mathbf{x}_t \leq 1\}$, we can finally write

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla h_t(\mathbf{w}_t) \mathbb{I}\{y_t \mathbf{w}_t^\top \mathbf{x}_t \leq 0\} = \mathbf{w}_t + \eta y_t \mathbf{x}_t \mathbb{I}\{y_t \mathbf{w}_t^\top \mathbf{x}_t \leq 0\} \quad (9)$$

where we also added the parameter $\eta > 0$ which will be used in the analysis. Because \mathbf{w}_t changes only when $y_t \mathbf{w}_t^\top \mathbf{x}_t \leq 0$, we can apply the OGD analysis only to the steps t where \mathbf{w}_t errs. That is, to the steps t when $y_t \mathbf{w}_t^\top \mathbf{x}_t \leq 0$. Inequality (6), with the negative term $-\frac{1}{2\eta} \|\mathbf{w}_{T+1} - \mathbf{u}\|^2$ dropped, gives

$$\begin{aligned} \sum_{t=1}^T \left(h_t(\mathbf{w}_t) - h_t(\mathbf{u}) \right) \mathbb{I}\{y_t \mathbf{w}_t^\top \mathbf{x}_t \leq 0\} &\leq \frac{1}{2\eta} \|\mathbf{u}\|^2 \\ &+ \frac{1}{2} \sum_{t=1}^T \|\mathbf{w}_{t+1} - \mathbf{u}\|^2 \left(\frac{1}{\eta} - \frac{1}{\eta} \right) \mathbb{I}\{y_t \mathbf{w}_t^\top \mathbf{x}_t \leq 0\} + \frac{\eta G^2}{2} \sum_{t=1}^T \mathbb{I}\{y_t \mathbf{w}_t^\top \mathbf{x}_t \leq 0\} \end{aligned}$$

for all $\mathbf{u} \in \mathbb{R}^d$. Note that the terms containing the factor $\frac{1}{\eta} - \frac{1}{\eta}$ are all zero (and this is the reason why we do not need projections). Since $y_t \mathbf{w}_t^\top \mathbf{x}_t \leq 0$ implies $h_t(\mathbf{w}_t) \geq 1$, we can write $\mathbb{I}\{y_t \mathbf{w}_t^\top \mathbf{x}_t \leq 0\} \leq h_t(\mathbf{w}_t) \mathbb{I}\{y_t \mathbf{w}_t^\top \mathbf{x}_t \leq 0\}$. So, setting $X = \max_t \|\mathbf{x}_t\| \geq \max_t \|\nabla h_t(\mathbf{w}_t)\|$ so that we have $G = X$, we obtain

$$\sum_{t=1}^T \mathbb{I}\{y_t \mathbf{w}_t^\top \mathbf{x}_t \leq 0\} \leq \sum_{t=1}^T h_t(\mathbf{u}) + \frac{1}{2\eta} \|\mathbf{u}\|^2 + \frac{\eta X^2}{2} \sum_{t=1}^T \mathbb{I}\{y_t \mathbf{w}_t^\top \mathbf{x}_t \leq 0\}$$

where we also used $h_t(\mathbf{u}) \mathbb{I}\{y_t \mathbf{w}_t^\top \mathbf{x}_t \leq 0\} \leq h_t(\mathbf{u})$.

Let $M_T = \sum_{t=1}^T \mathbb{I}\{y_t \mathbf{w}_t^\top \mathbf{x}_t \leq 0\}$ the number of mistakes made by Perceptron in the first T steps. Choosing $\eta = \|\mathbf{u}\| / (X\sqrt{M_T})$, solving for M_T , and upper bounding we get

$$M_T \leq \sum_{t=1}^T h_t(\mathbf{u}) + (\|\mathbf{u}\| X)^2 + \|\mathbf{u}\| X \sqrt{\sum_{t=1}^T h_t(\mathbf{u})} .$$

This shows a bound on the number of mistakes made by the Perceptron on any data sequence, including those sequences that are not linearly separable. When the sequence is linearly separable, then there exists $\mathbf{u} \in \mathbb{R}^d$ such that $y_t \mathbf{u}^\top \mathbf{x}_t \geq 1$ for all t , which in turn implies $h_t(\mathbf{u}) = 0$ for all t . Hence, the bound reduces to the one already proved in the Perceptron convergence theorem,

$$M_T \leq (\|\mathbf{u}\| X)^2$$

The last remaining issue is the choice η . The Perceptron runs with $\eta = 1$ while here η is chosen based on $\|\mathbf{u}\|$, X and M_T , which may not be all known. However, this discrepancy is an artifact of the proof. Indeed, the update rule (9) shows that the weight \mathbf{w}_t has the form

$$\mathbf{w}_t = \eta \sum_{s=1}^{t-1} y_s \mathbf{x}_s \mathbb{I}\{y_s \mathbf{w}_s^\top \mathbf{x}_s \leq 0\} .$$

Because the Perceptron predicts with $\text{sgn}(\mathbf{w}_t^\top \mathbf{x}_t)$, the actual choice of $\eta > 0$ is fully immaterial. In other words, running the algorithm with $\eta = 1$ as opposed to running it with $\eta = \|\mathbf{u}\| / (X\sqrt{M_T})$ does not change the predictions. Therefore, we may use in our analysis the value of $\eta > 0$ best suited to our purposes.

OGD with strongly convex losses. The upper bound (7) holds for any sequence ℓ_1, ℓ_2, \dots of convex and differentiable loss functions, including linear functions such as $\ell_t(\mathbf{w}) = |y_t - \mathbf{w}^\top \mathbf{x}_t|$ for $\mathbf{x}_t \in \mathbb{R}^d$ and $y_t \in \mathbb{R}$. It can be shown that (7) can not be significantly improved if the loss functions are all linear. But what if all loss functions are convex and never flat? To formalize this scenario, we use the notion of *strong convexity*. A differentiable function ℓ is σ -strongly convex, for some $\sigma > 0$, if

$$\ell(\mathbf{w}) - \ell(\mathbf{u}) \leq \nabla \ell(\mathbf{w})^\top (\mathbf{w} - \mathbf{u}) - \frac{\sigma}{2} \|\mathbf{u} - \mathbf{w}\|^2 . \quad (10)$$

If ℓ is also twice-differentiable, then (10) is equivalent to saying that the Hessian matrix of ℓ has full rank, that is, all of its eigenvalues are positive. A simple example of strongly convex function is $\ell(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|^2$. Indeed,

$$\frac{1}{2} \|\mathbf{w}\|^2 - \frac{1}{2} \|\mathbf{u}\|^2 = \mathbf{w}^\top (\mathbf{w} - \mathbf{u}) - \frac{1}{2} \|\mathbf{w} - \mathbf{u}\|^2$$

Hence, this function is strongly convex for $\sigma = 1$.

As we see later, OGD with strongly convex functions can be applied to a vast and important class of learning algorithms, including Support Vector Machines, corresponding to regularized forms of ERM.

When run on a sequence of strongly convex function, OGD does not need the projection step.

The OGD algorithm for σ -strongly convex functions

Initialization: $\mathbf{w}_1 = \mathbf{0}$

For $t = 1, 2, \dots$

1. $\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{1}{\sigma t} \nabla \ell_t(\mathbf{w}_t)$

In order to prove a regret bound, we apply (4) to the analysis of OGD under the assumption that ℓ_1, ℓ_2, \dots are all σ -strongly convex functions. Setting $\eta_t = \frac{1}{\sigma t}$ we get

$$\begin{aligned} \ell_t(\mathbf{w}_t) - \ell_t(\mathbf{u}) &\leq \nabla \ell_t(\mathbf{w}_t)^\top (\mathbf{w}_t - \mathbf{u}) - \frac{\sigma}{2} \|\mathbf{u} - \mathbf{w}_t\|^2 \\ &= -\frac{1}{\eta_t} (\mathbf{w}_{t+1} - \mathbf{w}_t)^\top (\mathbf{w}_t - \mathbf{u}) - \frac{\sigma}{2} \|\mathbf{u} - \mathbf{w}_t\|^2 \\ &= \frac{1}{\eta_t} \left(\frac{1}{2} \|\mathbf{w}_t - \mathbf{u}\|^2 - \frac{1}{2} \|\mathbf{w}_{t+1} - \mathbf{u}\|^2 + \frac{1}{2} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \right) - \frac{\sigma}{2} \|\mathbf{u} - \mathbf{w}_t\|^2. \end{aligned}$$

Proceeding just like we did in the proof of OGD with projection, while exploiting the additional terms $-\frac{\sigma}{2} \|\mathbf{u} - \mathbf{w}_t\|^2$, we obtain

$$\begin{aligned} R_T(\mathbf{u}) &\leq \left(\frac{1}{\eta_1} - \sigma \right) \frac{1}{2} \|\mathbf{w}_1 - \mathbf{u}\|^2 - \frac{1}{2\eta_{T+1}} \|\mathbf{w}_{T+1} - \mathbf{u}\|^2 \\ &\quad + \frac{1}{2} \sum_{t=1}^{T-1} \|\mathbf{w}_{t+1} - \mathbf{u}\|^2 \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} - \sigma \right) + \|\mathbf{w}_{T+1} - \mathbf{u}\|^2 \frac{1}{2} \left(\frac{1}{\eta_{T+1}} - \frac{1}{\eta_T} \right) + \frac{G^2}{2} \sum_{t=1}^T \eta_t \end{aligned}$$

where, similarly to before, $G \geq \max_t \|\nabla \ell_t(\mathbf{w}_t)\|$.

Dropping the negative term $-\frac{1}{2\eta_T} \|\mathbf{w}_{T+1} - \mathbf{u}\|^2$, simplifying the term $\frac{1}{2\eta_{T+1}} \|\mathbf{w}_{T+1} - \mathbf{u}\|^2$ which occurs with opposite signs, using the choice $\eta_t = \frac{1}{\sigma t}$, and making some further cancellations leads us to

$$R_T(\mathbf{u}) \leq \frac{G^2}{2\sigma} \sum_{t=1}^T \frac{1}{t} \leq \frac{G^2}{2\sigma} \ln(T+1)$$

where we used a simple logarithmic upper bound to the harmonic sum $1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{T}$.

This gives the final result

$$\frac{1}{T} \sum_{t=1}^T \ell_t(\mathbf{w}_t) \leq \min_{\mathbf{u} \in \mathbb{R}^d} \frac{1}{T} \sum_{t=1}^T \ell_t(\mathbf{u}) + \frac{G^2 \ln(T+1)}{2\sigma T}$$

Note the improved dependence $\frac{\ln T}{T}$ compared to $\frac{1}{\sqrt{T}}$ obtained in (7) for convex (as opposed to strongly convex) loss functions.