**Machine Learning — Statistical Methods for Machine Learning**

# Online Gradient Descent

Instructor: *Nicolò Cesa-Bianchi* version of May 9, 2023

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 $\boldsymbol{w}$ 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 $\ell$.

The algorithm outputs a default initial predictor $h_1 \in \mathcal{H}$
For $t = 1, 2, \ldots$

1. The next example $(\boldsymbol{x}_t, y_t)$ is observed
2. The loss $\ell\big(h_t(\boldsymbol{x}_t), y_t\big)$ 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 $h_t \to h_{t+1}$ is typically *local*. That is, it only involves the current predictor $h_t$ and the current example $(\boldsymbol{x}_t, y_t)$.

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

$$\frac{1}{T}\sum_{t=1}^{T} \ell\big(h_t(\boldsymbol{x}_t), y_t\big) \approx e_D(h) = \mathbb{E}\big[\ell(h(X), Y) \,\big|\, X\big]$$

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

In what follows, we use the notation $\ell_t(h) = \ell\big(h(\boldsymbol{x}_t), y_t\big)$ when the sequence $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots$ is understood from the context. This defines a sequence $\ell_1, \ell_2, \ldots$ 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 $\ell_1,\dots,\ell_T$. 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 $f:\mathbb{R}^d\to\mathbb{R}$ which we want to minimize, gradient descent works by iterating $\boldsymbol{w}_{t+1}=\boldsymbol{w}_t-\eta_t\nabla f(\boldsymbol{w}_t)$ starting from some initial point $\boldsymbol{w}_1$ in the domain of $f$, where $\eta>0$ is a parameter. If the current point $\boldsymbol{w}_t$ is not a minimum of $f$, then $\nabla f(\boldsymbol{w}_t)\neq\boldsymbol{0}$ and $\boldsymbol{w}_{t+1}-\boldsymbol{w}_t$ points in the direction opposite to $\nabla f(\boldsymbol{w}_t)$, which is —by definition of gradient— the direction where $f$ decreases the most when moving away from $\boldsymbol{w}_t$. 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 $h(\boldsymbol{x})=\boldsymbol{w}^\top\boldsymbol{x}$, for $\boldsymbol{w}\in\mathbb{R}^d$. We use the notation $\ell_t(\boldsymbol{w})=\ell\big(\boldsymbol{w}^\top\boldsymbol{x}_t,y_t\big)$ and assume that losses $\ell_1,\ell_2,\dots$ are all convex and differentiable. For example, $\ell_t(\boldsymbol{w})=\big(\boldsymbol{w}^\top\boldsymbol{x}_t-y_t\big)^2$.

**Projected OGD**
Parameters: $\eta>0$, $U>0$
Initialization: $\boldsymbol{w}_1=\boldsymbol{0}$
For $t=1,2,\dots$

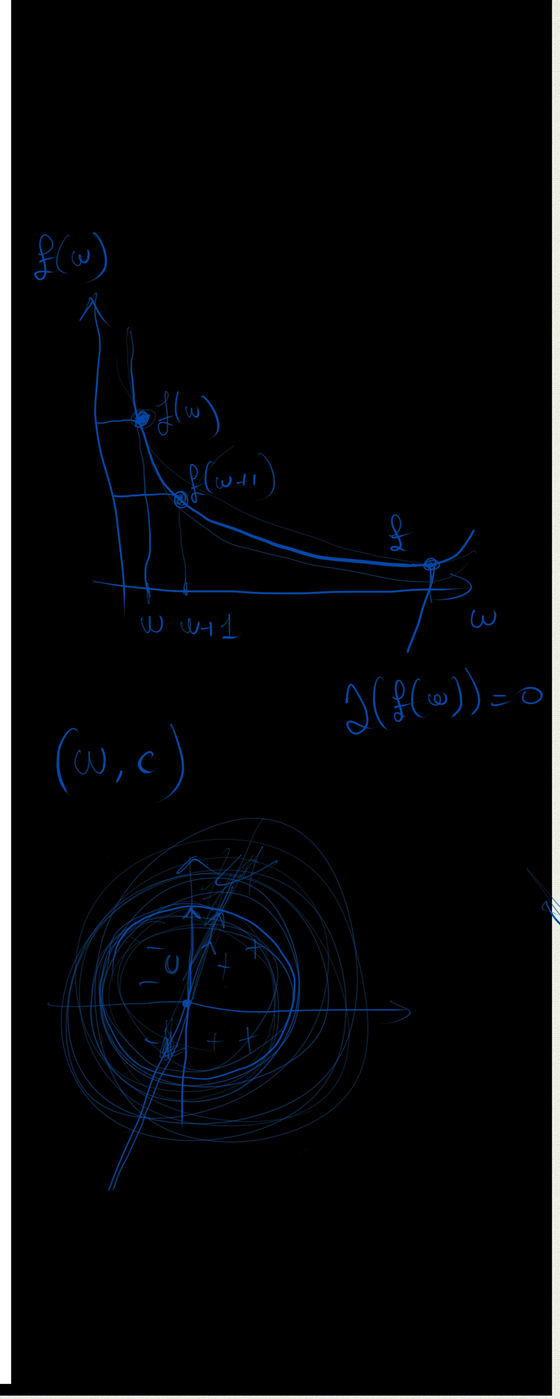
1. $\boldsymbol{w}'_{t+1}=\boldsymbol{w}_t-\dfrac{\eta}{\sqrt{t}}\nabla\ell_t(\boldsymbol{w}_t)$
2. $\boldsymbol{w}_{t+1}=\underset{\boldsymbol{w}\,:\,\|\boldsymbol{w}\|\le U}{\operatorname{argmin}}\ \|\boldsymbol{w}-\boldsymbol{w}'_{t+1}\|$

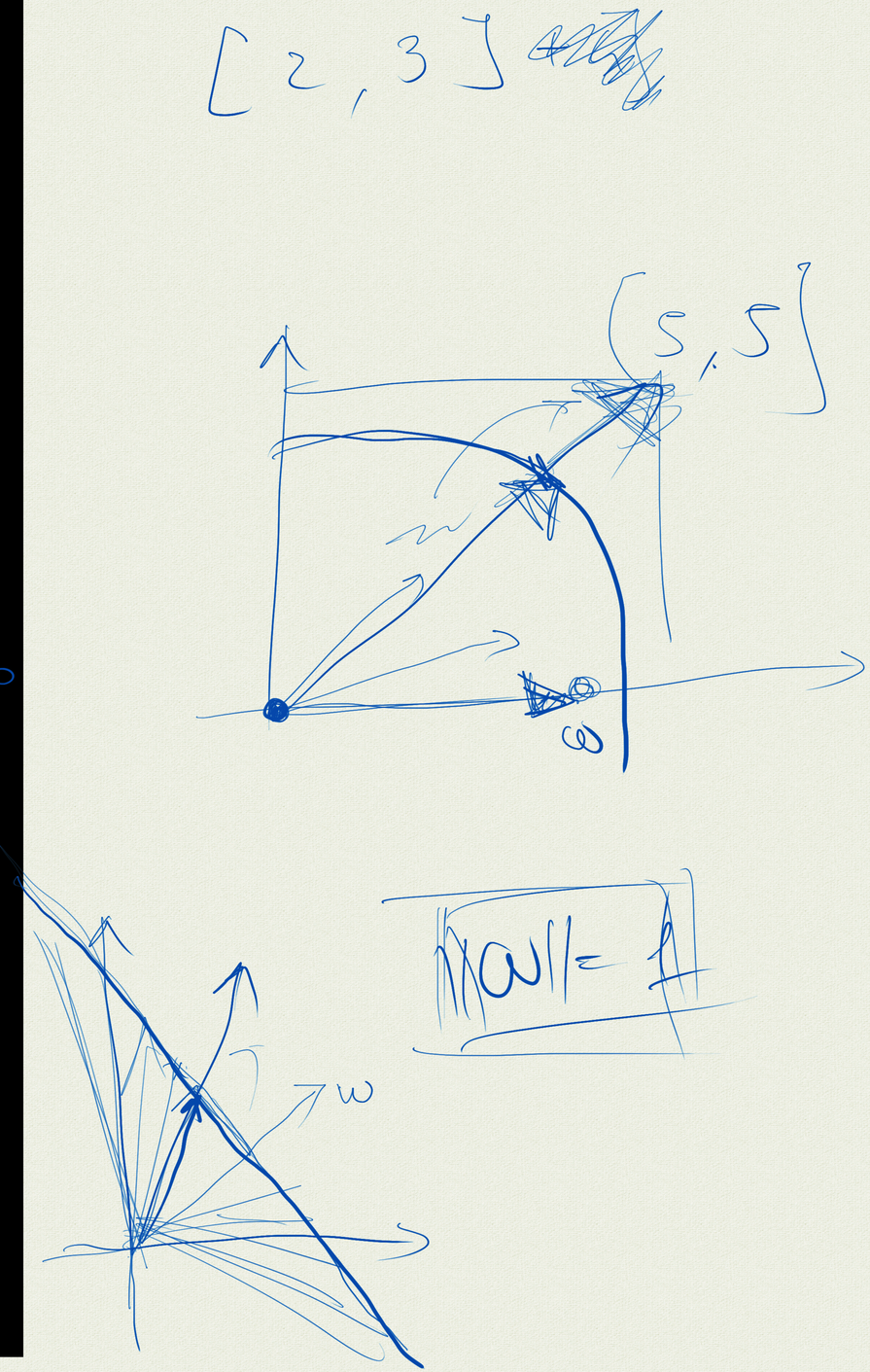
In step 2, we project $\boldsymbol{w}'_{t+1}$ in an Euclidean sphere of radius $U$. If $\|\boldsymbol{w}'_{t+1}\|\le U$, then $\boldsymbol{w}_{t+1}=\boldsymbol{w}'_{t+1}$. Let $\eta_t=\eta/\sqrt{t}$, where $\eta>0$ will be determined by the analysis.

Our goal is to control the regret

$$\frac{1}{T}\sum_{t=1}^{T}\ell_t(\boldsymbol{w}_t)-\frac{1}{T}\sum_{t=1}^{T}\ell_t(\boldsymbol{u}_T^*)\qquad\text{where}\qquad\boldsymbol{u}_T^*=\underset{\boldsymbol{u}\,:\,\|\boldsymbol{u}\|\le U}{\operatorname{argmin}}\frac{1}{T}\sum_{t=1}^{T}\ell_t(\boldsymbol{u})$$

Note that $\boldsymbol{u}_T^*$ 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 $R_T(\boldsymbol{u})=\sum_{t=1}^{T}\big(\ell_t(\boldsymbol{w}_t)-\ell_t(\boldsymbol{u})\big)$. The analysis of OGD is based on the following well-known result.



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

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

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

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

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

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

$$\begin{aligned}
\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u}) &\leq \nabla \ell_t(\boldsymbol{w}_t)^\top (\boldsymbol{w}_t - \boldsymbol{u}) && (2)\\
&= -\frac{1}{\eta_t}(\boldsymbol{w}'_{t+1} - \boldsymbol{w}_t)^\top (\boldsymbol{w}_t - \boldsymbol{u}) && (3)\\
&= \frac{1}{\eta_t}\left(\frac{1}{2}\|\boldsymbol{w}_t - \boldsymbol{u}\|^2 - \frac{1}{2}\|\boldsymbol{w}'_{t+1} - \boldsymbol{u}\|^2 + \frac{1}{2}\|\boldsymbol{w}'_{t+1} - \boldsymbol{w}_t\|^2\right) && (4)\\
&\leq \frac{1}{\eta_t}\left(\frac{1}{2}\|\boldsymbol{w}_t - \boldsymbol{u}\|^2 - \frac{1}{2}\|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2 + \frac{1}{2}\|\boldsymbol{w}'_{t+1} - \boldsymbol{w}_t\|^2\right) . && (5)
\end{aligned}$$

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

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

$$\underbrace{\frac{1}{2\eta_t}\|\boldsymbol{w}_t - \boldsymbol{u}\|^2 - \frac{1}{2\eta_{t+1}}\|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2}\ \underbrace{-\frac{1}{2\eta_t}\|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2 + \frac{1}{2\eta_{t+1}}\|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2}\ + \frac{1}{2\eta_t}\|\boldsymbol{w}'_{t+1} - \boldsymbol{w}_t\|^2 .$$

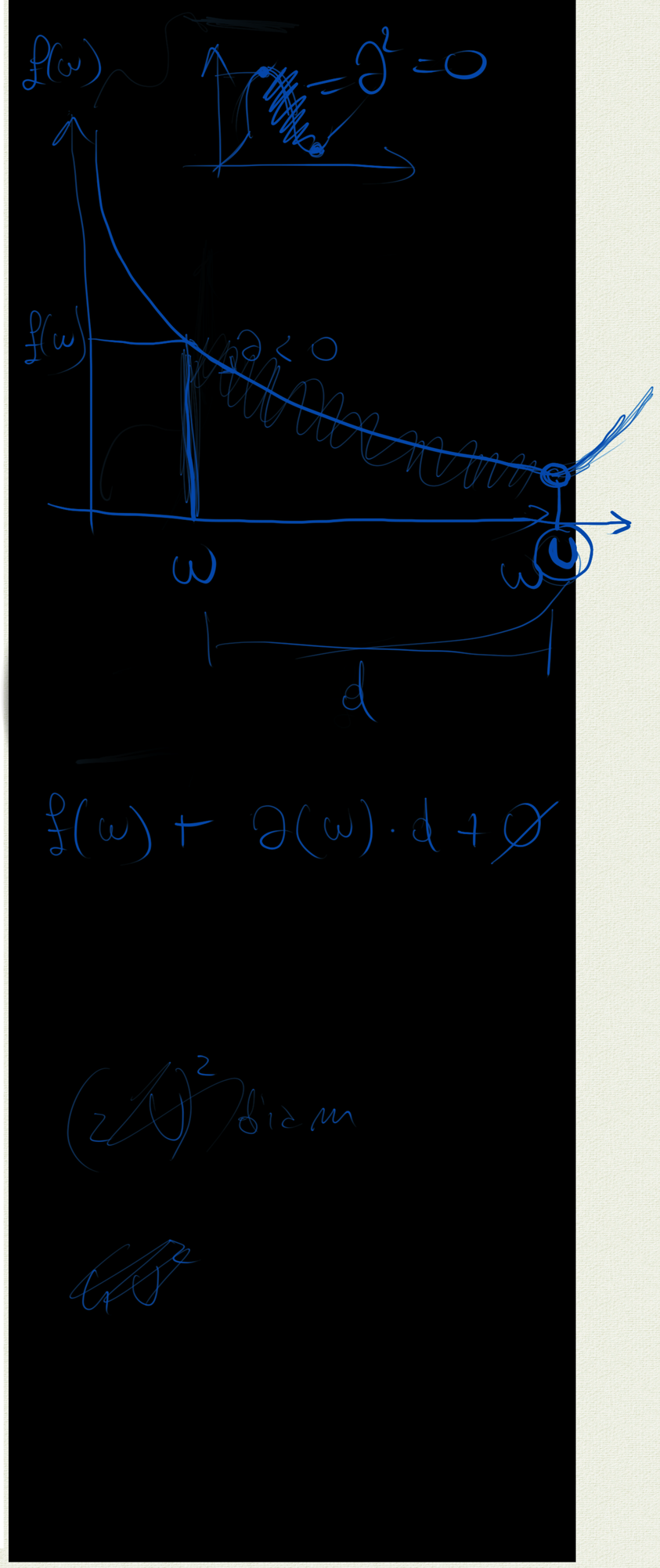
Summing over $t = 1, \ldots, T$ we observe that the first pair of terms forms a telescopic sum, while the terms in the second pair have a common factor,

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

Next, we make use of the following facts:

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





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

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

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

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

where we used the upper bound

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T} \ . \quad \text{Harmonic function}$$

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

$R_T(u) = \sum_{t=1}^{T} \ell_t(w_t) - \ell_T(u)$ — $\ell_T(w)$ — $\ell_T(u^*)$ — Regret

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

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

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

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

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

$U$ = maximum radius of euclidean sphere
$X$ = bound on the $\|x_i\|$
$u$ = best predictor

**OGD with strongly convex losses.** The upper bound (7) holds for any sequence $\ell_1, \ell_2, \ldots$ of convex and differentiable loss functions, including linear functions such as $\ell_t(\boldsymbol{w}) = \left|y_t - \boldsymbol{w}^\top \boldsymbol{x}_t\right|$ for $\boldsymbol{x}_t \in \mathbb{R}^d$ 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 $\ell$ is $\sigma$-strongly convex, for some $\sigma > 0$, if

$$\ell(\boldsymbol{w}) - \ell(\boldsymbol{u}) \leq \nabla \ell(\boldsymbol{w})^\top (\boldsymbol{w} - \boldsymbol{u}) - \frac{\sigma}{2} \|\boldsymbol{u} - \boldsymbol{w}\|^2 \ . \tag{8}$$

If $\ell$ is also twice-differentiable, then (8) is equivalent to saying that the Hessian matrix of $\ell$ has full rank, that is, all of its eigenvalues are positive. A simple example of strongly convex function is $\ell(\boldsymbol{w}) = \frac{1}{2} \|\boldsymbol{w}\|^2$. Indeed,

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

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 $\sigma$-strongly convex functions**
Initialization: $\boldsymbol{w}_1 = \mathbf{0}$
For $t = 1, 2, \ldots$

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

In order to prove a regret bound, we apply (4) to the analysis of OGD under the assumption that $\ell_1, \ell_2, \ldots$ are all $\sigma$-strongly convex functions. Setting $\eta_t = \frac{1}{\sigma t}$ we get

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

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

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

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

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

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

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

This gives the final result

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

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.

**A mistake bound for the Perceptron algorithm.** We now prove an upper bound on the number of prediction mistakes made by the Perceptron on an arbitrary stream. Because the zero-one loss is not convex, we cannot directly apply the machinery developed for OGD. Instead, we adapt the proof of the Perceptron convergence theorem and use a convex upper bound on the zero-one loss to compensate for the lack of convexity.

Let $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots \in \mathbb{R}^d \times \{-1, 1\}$ be a stream of data points with binary labels and let $M$ be the number of prediction mistakes made by the Perceptron in the first $T$ examples of the stream. Let $\boldsymbol{w}_M$ be the Perceptron hyperplane after these $M$ prediction mistakes and let $t_M \in \{1, \ldots, T\}$ be the index of the example $(\boldsymbol{x}_{t_M}, y_{t_M})$ in the stream that caused the $M$-th mistake $\boldsymbol{w}_M = \boldsymbol{w}_{M-1} + y_{t_M} \boldsymbol{x}_{t_M}$. Now fix any $\boldsymbol{u} \in \mathbb{R}^d$. This $\boldsymbol{u}$ is not necessarily a separator, because we are not making any assumption on the stream. The first part of the proof of the Perceptron convergence theorem does not use any special property of $\boldsymbol{u}$. Therefore, proceeding in exactly the same way, we have that

$$\|\boldsymbol{w}_M\| \, \|\boldsymbol{u}\| \leq \|\boldsymbol{u}\| \left( \max_{t=1,\ldots,m} \|\boldsymbol{x}_t\| \right) \sqrt{M} \; .$$

In order to prove a lower bound on $\|\boldsymbol{w}_M\| \, \|\boldsymbol{u}\|$ and finish the proof, we proceed as follows

$$\begin{aligned}
\|\boldsymbol{w}_M\| \, \|\boldsymbol{u}\| &\geq \boldsymbol{w}_M^\top \boldsymbol{u} \\
&= \left( \boldsymbol{w}_{M-1} + y_{t_M} \boldsymbol{x}_{t_M} \right)^\top \boldsymbol{u} \\
&= \boldsymbol{w}_{M-1}^\top \boldsymbol{u} + y_{t_M} \boldsymbol{u}^\top \boldsymbol{x}_{t_M} \\
&= \boldsymbol{w}_{M-1}^\top \boldsymbol{u} + 1 - 1 + y_{t_M} \boldsymbol{u}^\top \boldsymbol{x}_{t_M} \\
&\geq \boldsymbol{w}_{M-1}^\top \boldsymbol{u} + 1 - \left[ 1 + y_{t_M} \boldsymbol{u}^\top \boldsymbol{x}_{t_M} \right]_+
\end{aligned}$$

where $[z]_+ = \max\{0, z\}$. Iterating $M$ times we get

$$\|\boldsymbol{w}_M\| \, \|\boldsymbol{u}\| \geq M + \sum_{i=1}^{M} \left[ 1 + y_{t_i} \boldsymbol{u}^\top \boldsymbol{x}_{t_i} \right]_+$$

Where we used $\boldsymbol{w}_0^\top \boldsymbol{u} = 0$ since $\boldsymbol{w}_0 = (0, \ldots, 0)$. Let $X = \max_t \|\boldsymbol{x}_t\|$. Combining upper and lower bound we obtain

$$M \leq \sum_{i=1}^{M} \left[ 1 + y_{t_i} \boldsymbol{u}^\top \boldsymbol{x}_{t_i} \right]_+ + \|\boldsymbol{u}\| \, X \sqrt{M} \tag{9}$$

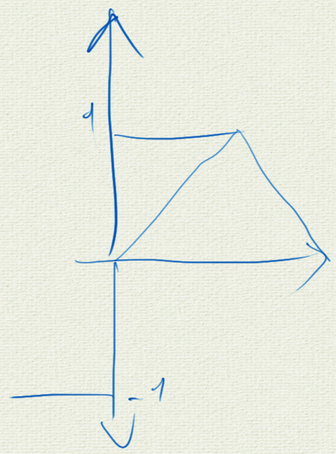
The function $h_t(\boldsymbol{u}) = \left[ 1 + y_t \boldsymbol{u}^\top \boldsymbol{x}_t \right]_+$ is a loss function called **hinge loss**. Since $\mathbb{I}\{\operatorname{sgn}(z) \neq y\} \leq [1 - zy]_+$ for all $z \in \mathbb{R}$ and $y \in \{-1, 1\}$, the hinge loss is a convex upper bound on the zero-one loss. Because $\{t_1, \ldots, t_M\} \subseteq \{1, \ldots, T\}$,

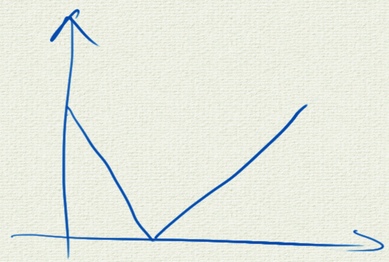
$$\sum_{i=1}^{M} h_{t_i(\boldsymbol{u})} \leq \sum_{t=1}^{T} h_t(\boldsymbol{u})$$

we can rewrite (9) as

$$M \leq \sum_{t=1}^{T} h_t(\boldsymbol{u}) + \|\boldsymbol{u}\| \, X \sqrt{M}$$







Solving with respect to $M$ and overapproximating, we get

$$M \leq \sum_{t=1}^{T} h_t(\boldsymbol{u}) + \left(\|\boldsymbol{u}\| X\right)^2 + \|\boldsymbol{u}\| X \sqrt{\sum_{t=1}^{T} h_t(\boldsymbol{u})} \qquad \text{for all } \boldsymbol{u} \in \mathbb{R}^d$$

This shows a bound on the number of mistakes made by the Perceptron on any data sequence of arbitrary length $T$, including those sequences that are not linearly separable. When the sequence is linearly separable, then there exists $\boldsymbol{u} \in \mathbb{R}^d$ such that $y_t \, \boldsymbol{u}^\top \boldsymbol{x}_t \geq 1$ for all $t$, which in turn implies $h_t(\boldsymbol{u}) = 0$ for all $t$. Hence, the bound reduces to the one already proved in the Perceptron convergence theorem, $M_T \leq \left(\|\boldsymbol{u}\| X\right)^2$.