Machine Learning Fundamentals Homework 3

Predrag Pilipovic (predrag.pilipovic@grenoble-inp.org)

November 17, 2019

An analysis of the PEGASOS Algorithm

Different learning algorithms for binary classification have been proposed for the minimization of the following learning objective over the class of linear functions, $\mathcal{H} = \{h : \mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle\}$:

$$\hat{\mathcal{L}}_m(S, \mathbf{w}) = \frac{1}{m} \sum_{(\mathbf{x},y)\in S} l(h(\mathbf{x}), y) + \frac{\lambda}{2}||\mathbf{w}||^2, \tag{1}$$

where $S = (\mathbf{x}_i, y_i)_{1\leqslant i \leqslant m}$ is a training set of size m, $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ a vector representation of an observation, $y \in \{1, +1\}$ its associated class label, and l an instantanious loss (called the hinge loss) defined as:

$$l(h(\mathbf{x}), y) = \max(0, 1 - yh(\mathbf{x})). \tag{2}$$

In the following we will analysis the algorithm called PEGASOS (*Primal Estimated sub-Gradient SOlver for SVM*)[1] which procedure is summarized below.

Algorithm 1 Pegasos

```
Input: Training set $S = (\mathbf{x}_i, y_i)_{1\leqslant i \leqslant m}$, constant $\lambda > 0$ and maximum number of iterations $T$
Initialize: Set $\mathbf{w}^{(1)} \leftarrow \mathbf{0}$
for $t = 1, 2, ..., T$ do
    Set $S_t^+ = \left\{(\mathbf{x}, y) \in S \mid y \left\langle \mathbf{w}^{(t)}, \mathbf{x} \right\rangle < 1\right\}$
    Set $\eta_t = \frac{1}{\lambda \cdot t}$
    Update $\mathbf{w}^{(t+1)} \leftarrow (1 - \lambda \cdot \eta_t) \cdot \mathbf{w}^{(t)} + \frac{\eta_t}{m} \sum_{(\mathbf{x},y)\in S_t^+} y \cdot \mathbf{x}$
end for
Output: $\mathbf{w}^{(T+1)}$
```

Begining from a null weight vector, the algorithm iteratively updates the weights over the subset of misclassified training examples S_t^+ by applying the following rule

$$(\forall t \in \{1, 2, ..., T\})\ \mathbf{w}^{(t+1)} \leftarrow (1 - \lambda \cdot \eta_t) \cdot \mathbf{w}^{(t)} + \frac{\eta_t}{m} \sum_{(\mathbf{x},y)\in S_t^+} y \cdot \mathbf{x}, \tag{3}$$

where, $\eta_t = \frac{1}{\lambda \cdot t}$ is the learning rate. In the following we will analysis the convergence property of the algorithm.

Question 1.

For an observation $(\mathbf{x}, y)$ and a prediction function $h \in \mathcal{H}$, why the sign of the product $yh(\mathbf{x}) = y \cdot \langle \mathbf{w}, \mathbf{x} \rangle$ is an indicator of good/bad classification?

[1] S. Shalev-Shwartz, Y. Singer, N. Srebro and A. Cotter. Primal Estimated sub-Gradient SOlver for SVM (Pegasos) *Mathematical Programming* March 2011, Volume 127, Issue 1, pp 330

The sign of $yh(\mathbf{x})$ will determine if $\mathbf{x}$ is well classified. If $\mathbf{x}$ is well classified then the sign of y and $\langle \mathbf{w}, \mathbf{x} \rangle$ should be the same. For example, if $y = -1$, then $\langle \mathbf{w}, \mathbf{x} \rangle < 0$, because h is a prediction function. On the other hand, if $\mathbf{x}$ is missclassified, then the sign of $yh(\mathbf{x})$ should be negative. For example, if $y = -1$, then $\langle \mathbf{w}, \mathbf{x} \rangle > 0$, so the sign will be negative. To conclude, for good classification we will have positive sign of $yh(\mathbf{x})$, and negative for bad classification.

Question 2.

Which other learning algorithm updates the learning weights over misclassified training examples? In the case where $S_t^+ = (\mathbf{x}_t, y_t)$ is a singleton what is the update rule of this other learning algorithm and what is the difference with the one proposed in PEGASOS (3)?

The other learning algorithm that updates the learning weights over misclassified training examples is Perceptron. Let $S_t^+ = \{(\mathbf{x}, y)\}$, so now we have a Pegasos update rule

$$\mathbf{w}^{(t+1)} \leftarrow (1 - \lambda \cdot \eta_t) \cdot \mathbf{w}^{(t)} + \frac{\eta_t}{m} \cdot y \cdot \mathbf{x}.$$

On the other hand, in the Perceptron algorithm we have an update rule

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} + \eta_t \cdot y \cdot \mathbf{x}.$$

As we already know for Perceptron algorithm, the previous rule is Gradient descent, but we will see in Question 7 the Pegasos update rule is also Gradient descent, so both rules can be written as

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta_t \nabla \hat{\mathcal{L}}.$$

The only difference between those two algorithms (when $S_t^+ = (\mathbf{x}_t, y_t)$) is in the learning objective function $\hat{\mathcal{L}}$.

Question 3.

Draw the binary classification loss $l_b : (h(\mathbf{x}), y) \mapsto \mathbb{1}_{yh(x)<0}$, and the hinge loss (Eq. 2) with respect to the product $yh(\mathbf{x})$, i.e. the loss on the y-axis and $yh(\mathbf{x})$ on the x-axis

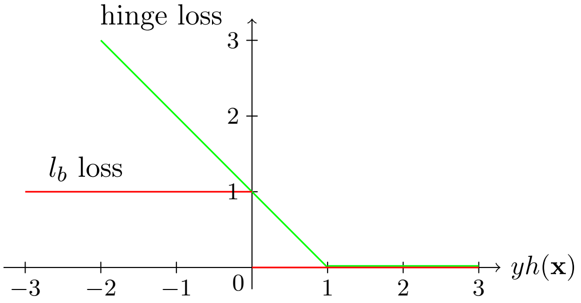

Figure 1: Loss functions with respect to $yh(\mathbf{x})$

Question 4.

For a given example $(\mathbf{x}, y)$, what does $\frac{|h(\mathbf{x})|}{||\mathbf{w}||}$ represent?

The prediction function is defined as a hyperplane, and $\mathbf{w}$ is normal to this hyperplane. Any wector $\mathbf{x}$ in the vector space has a unique decomposition

$$\mathbf{x} = \mathbf{x}_P + \mathbf{x}_H,$$

where $\mathbf{x}_P$ is projection of $\mathbf{x}$ on the hyperplane and $\mathbf{x}_H$ is orthogonal to the hyperplane. So, we have

$$\langle \mathbf{w}, \mathbf{x} \rangle = \langle \mathbf{w}, \mathbf{x}_P + \mathbf{x}_H \rangle = \underbrace{\langle \mathbf{w}, \mathbf{x}_P \rangle}_{0} + \langle \mathbf{w}, \mathbf{x}_H \rangle = ||\mathbf{w}|| \cdot ||\mathbf{x}_H|| \cdot \underbrace{\cos\angle(\mathbf{w}, \mathbf{x}_H)}_{1}.$$

Note here that $\langle \mathbf{w}, \mathbf{x}_P \rangle = 0$ because $\mathbf{x}_P$ is orthogonal to $\mathbf{w}$, and $\cos\angle(\mathbf{w}, \mathbf{x}_H)$, because both $\mathbf{w}$ and $\mathbf{x}_H$ are orthogonal to $\mathbf{x}_P$, so the angle is then 0. Now, we have

$$||\mathbf{x}_H|| = \frac{\langle \mathbf{w}, \mathbf{x} \rangle}{||\mathbf{w}||} = \frac{h(\mathbf{x})}{||\mathbf{w}||},$$

which means that $\frac{h(\mathbf{x})}{||\mathbf{w}||}$ represents the distance from $\mathbf{x}$ to the hyperplane.

Question 5.

Why the learning objective (1) admits a single minimizer $\mathbf{w}^* \in \mathbb{R}^d$?

Minimizing the learning objective function $\hat{\mathcal{L}}$ is quadratic optimization problem subject to linear constraints and there is a unique minimum

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \hat{\mathcal{L}}(S, \mathbf{w}).$$

Question 6.

Explain why at the first iteration, S_1 is the whole training set; $S_1 = S$?

In the first iteration we have $\mathbf{w}^{(1)} = 0$, so $y \cdot \langle \mathbf{w}^{(1)}, \mathbf{x} \rangle = 0$, for any $(\mathbf{x}, y) \in S$, which means that $S_1 = S$.

Question 7.

Show that the update (3) follows the gradient descente rule:

$$(\forall t \in \{1, 2, ..., T\}) \ \mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta_t \nabla_t$$

where $\nabla_t = \nabla_t \hat{\mathcal{L}}_m\left(S, \mathbf{w}^{(t)}\right)$ denotes the gradient of the learning objective (Eq. 1) at $\mathbf{w}^{(t)}$.

If we work with data in S_t^+, i.e. data for which we know

$$\left((\mathbf{x}, y) \in S_t^+\right) \ y \cdot \left\langle \mathbf{w}^{(t)}, \mathbf{x} \right\rangle < 1,$$

then for $(\mathbf{x}, y) \in S_t^+$ must be

$$\max(0, 1 - yh(\mathbf{x})) = 1 - yh(\mathbf{x}).$$

So, we have

$$\begin{aligned}\hat{\mathcal{L}}_m\left(S_t^+, \mathbf{w}^{(t)}\right) &= \frac{1}{m} \sum_{(\mathbf{x},y)\in S_t^+} \left(1 - y \cdot \left\langle \mathbf{w}^{(t)}, \mathbf{x}\right\rangle\right) + \frac{\lambda}{2} \left|\left|\mathbf{w}^{(t)}\right|\right|^2 \\ &= \frac{1}{m} \sum_{(\mathbf{x},y)\in S_t^+} \left(1 - y \sum_{i=1}^{d} w_i^{(t)} x_i\right) + \frac{\lambda}{2} \sum_{i=1}^{d} \left(w_i^{(t)}\right)^2.\end{aligned}$$

After this, we know how to calculate gradient, i.e.

$$\nabla_t = \nabla \hat{\mathcal{L}}_m\left(S_t^+, \mathbf{w}^{(t)}\right) = \begin{bmatrix} \dfrac{\partial \hat{\mathcal{L}}}{\partial w_1^{(t)}} \\ \dfrac{\partial \hat{\mathcal{L}}}{\partial w_2^{(t)}} \\ \vdots \\ \dfrac{\partial \hat{\mathcal{L}}}{\partial w_d^{(t)}} \end{bmatrix} = \begin{bmatrix} \frac{1}{m} \sum_{(\mathbf{x},y)\in S_t^+} (-yx_1) + \lambda w_1^{(t)} \\ \frac{1}{m} \sum_{(\mathbf{x},y)\in S_t^+} (-yx_2) + \lambda w_2^{(t)} \\ \vdots \\ \frac{1}{m} \sum_{(\mathbf{x},y)\in S_t^+} (-yx_d) + \lambda w_d^{(t)} \end{bmatrix} = -\frac{1}{m} \sum_{(\mathbf{x},y)\in S_t^+} y\mathbf{x} + \lambda \mathbf{w}^{(t)}.$$

Our update rule (3) now can be transformed as

$$\begin{aligned}\mathbf{w}^{(t+1)} &= (1 - \lambda \cdot \eta_t) \cdot \mathbf{w}^{(t)} + \frac{\eta_t}{m} \sum_{(\mathbf{x},y)\in S_t^+} y \cdot \mathbf{x} \\ &= \mathbf{w}^{(t)} - \eta_t \left(-\frac{1}{m} \sum_{(\mathbf{x},y)\in S_t^+} y\mathbf{x} + \lambda \mathbf{w}^{(t)}\right) \\ &= \mathbf{w}^{(t)} - \eta_t \nabla_t.\end{aligned}$$

Question 8.

For two consecutive weights $\mathbf{w}^{(t)}$ and $\mathbf{w}^{(t+1)}$, show that

$$\left|\left|\mathbf{w}^{(t)} - \mathbf{w}^*\right|\right|^2 - \left|\left|\mathbf{w}^{(t+1)} - \mathbf{w}^*\right|\right|^2 = 2\eta_t \cdot \left\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla_t \right\rangle - \eta_t^2 \cdot \left|\left|\nabla_t\right|\right|^2.$$

We are using knowledge from the previous question to perform calculations. We have

$$\begin{aligned}\left|\left|\mathbf{w}^{(t)} - \mathbf{w}^*\right|\right|^2 - \left|\left|\mathbf{w}^{(t+1)} - \mathbf{w}^*\right|\right|^2 &= \left|\left|\mathbf{w}^{(t)}\right|\right|^2 - 2\left\langle \mathbf{w}^{(t)}, \mathbf{w}^*\right\rangle + \left|\left|\mathbf{w}^*\right|\right|^2 \\ &- \left|\left|\mathbf{w}^{(t+1)}\right|\right|^2 + 2\left\langle \mathbf{w}^{(t+1)}, \mathbf{w}^*\right\rangle - \left|\left|\mathbf{w}^*\right|\right|^2 \\ &= \left|\left|\mathbf{w}^{(t)}\right|\right|^2 - 2\left\langle \mathbf{w}^{(t)}, \mathbf{w}^*\right\rangle - \left|\left|\mathbf{w}^{(t)} - \eta_t\nabla_t\right|\right|^2 + 2\left\langle \mathbf{w}^{(t)} - \eta_t\nabla_t, \mathbf{w}^*\right\rangle \\ &= \left|\left|\mathbf{w}^{(t)}\right|\right|^2 - 2\left\langle \mathbf{w}^{(t)}, \mathbf{w}^*\right\rangle - \left|\left|\mathbf{w}^{(t)}\right|\right|^2 + 2\left\langle \mathbf{w}^{(t)}, \eta_t\nabla_t\right\rangle - \eta_t^2 \left|\left|\nabla_t\right|\right|^2 \\ &- 2\left\langle \mathbf{w}^{(t)}, \mathbf{w}^*\right\rangle + 2\left\langle \eta_t\nabla_t, \mathbf{w}^*\right\rangle \\ &2\eta_t \cdot \left\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla_t\right\rangle - \eta_t^2 \cdot \left|\left|\nabla_t\right|\right|^2.\end{aligned}$$

Question 9.

The objective learning function is λ-strongly convex (admitted), that is

$$(\forall u \in \mathbb{R}^d) \ \left\langle \mathbf{w}^{(t)} - u, \nabla_t \right\rangle \geqslant \hat{\mathcal{L}}\left(\mathbf{w}^{(t)}\right) - \hat{\mathcal{L}}(u) + \frac{\lambda}{2} \cdot \left|\left|\mathbf{w}^{(t)} - u\right|\right|^2.$$

From this property and the previous question, deduce then

$$\sum_{t=1}^{T}\left(\hat{\mathcal{L}}\left(\mathbf{w}^{(t)}\right) - \hat{\mathcal{L}}(\mathbf{w}^*)\right) \leqslant \sum_{t=1}^{T}\left(\frac{\left|\left|\mathbf{w}^{(t)} - \mathbf{w}^*\right|\right|^2 - \left|\left|\mathbf{w}^{(t+1)} - \mathbf{w}^*\right|\right|^2}{2\eta_t} - \frac{\lambda}{2} \cdot \left|\left|\mathbf{w}^{(t)} - \mathbf{w}^*\right|\right|^2\right) + \frac{1}{2}\sum_{t=1}^{T}\eta_t \left|\left|\nabla_t\right|\right|^2$$

Letting u be $\mathbf{w}^*$ in λ-strongly convex definition and transforming it a little bit, we have

$$\hat{\mathcal{L}}\left(\mathbf{w}^{(t)}\right) - \hat{\mathcal{L}}(\mathbf{w}^*) \leqslant \left\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \nabla_t \right\rangle - \frac{\lambda}{2} \cdot \left|\left|\mathbf{w}^{(t)} - \mathbf{w}^*\right|\right|^2.$$

Now using the equation from the previous question we have

$$\hat{\mathcal{L}}\left(\mathbf{w}^{(t)}\right) - \hat{\mathcal{L}}(\mathbf{w}^*) \leqslant \frac{\left|\left|\mathbf{w}^{(t)} - \mathbf{w}^*\right|\right|^2 - \left|\left|\mathbf{w}^{(t+1)} - \mathbf{w}^*\right|\right|^2}{2\eta_t} + \frac{\eta_t}{2}||\nabla_t||^2 - \frac{\lambda}{2} \cdot \left|\left|\mathbf{w}^{(t)} - \mathbf{w}^*\right|\right|^2.$$

After summing the previous equation over all $t = 1, 2, ..., T$ we will get the desired inequality.

Question 10.

Show that for two consecutive iterations t and $t+1$, we have

$$\sum_{j=t}^{t+1}\left(\frac{\left|\left|\mathbf{w}^{(j)} - \mathbf{w}^*\right|\right|^2 - \left|\left|\mathbf{w}^{(j+1)} - \mathbf{w}^*\right|\right|^2}{2\eta_j} - \frac{\lambda}{2} \cdot \left|\left|\mathbf{w}^{(j)} - \mathbf{w}^*\right|\right|^2\right) = \frac{\lambda(t-1)}{2} \cdot \left|\left|\mathbf{w}^{(t)} - \mathbf{w}^*\right|\right|^2 - \frac{\lambda(t+1)}{2} \cdot \left|\left|\mathbf{w}^{(t+2)} - \mathbf{w}^*\right|\right|^2$$

Note here, that the sum in the formula is just addition of two elements. So, we can rewrite it and some elements will cancel themselves.

$$\sum_{j=t}^{t+1}\left(\frac{\left|\left|\mathbf{w}^{(j)} - \mathbf{w}^*\right|\right|^2 - \left|\left|\mathbf{w}^{(j+1)} - \mathbf{w}^*\right|\right|^2}{2\eta_j} - \frac{\lambda}{2} \cdot \left|\left|\mathbf{w}^{(j)} - \mathbf{w}^*\right|\right|^2\right)$$
$$= \frac{\left|\left|\mathbf{w}^{(t)} - \mathbf{w}^*\right|\right|^2 - \left|\left|\mathbf{w}^{(t+1)} - \mathbf{w}^*\right|\right|^2}{2\eta_t} - \frac{\lambda}{2} \cdot \left|\left|\mathbf{w}^{(t)} - \mathbf{w}^*\right|\right|^2$$
$$+ \frac{\left|\left|\mathbf{w}^{(t+1)} - \mathbf{w}^*\right|\right|^2 - \left|\left|\mathbf{w}^{(t+2)} - \mathbf{w}^*\right|\right|^2}{2\eta_{t+1}} - \frac{\lambda}{2} \cdot \left|\left|\mathbf{w}^{(t+1)} - \mathbf{w}^*\right|\right|^2.$$

After using the definition that $\eta_t = \frac{1}{\lambda\cdot t}$ and rearranging the previous two lines we have

$$\begin{aligned}
&\sum_{j=t}^{t+1}\left(\frac{\left|\left|\mathbf{w}^{(j)}-\mathbf{w}^*\right|\right|^2-\left|\left|\mathbf{w}^{(j+1)}-\mathbf{w}^*\right|\right|^2}{2\eta_j}-\frac{\lambda}{2}\cdot\left|\left|\mathbf{w}^{(j)}-\mathbf{w}^*\right|\right|^2\right)\\
&=\frac{\lambda\cdot t}{2}\cdot\left|\left|\mathbf{w}^{(t)}-\mathbf{w}^*\right|\right|^2-\cancel{\frac{\lambda\cdot t}{2}\cdot\left|\left|\mathbf{w}^{(t+1)}-\mathbf{w}^*\right|\right|^2}-\frac{\lambda}{2}\cdot\left|\left|\mathbf{w}^{(t)}-\mathbf{w}^*\right|\right|^2\\
&+\cancel{\frac{\lambda\cdot(t+1)}{2}\cdot\left|\left|\mathbf{w}^{(t+1)}-\mathbf{w}^*\right|\right|^2}-\frac{\lambda\cdot(t+1)}{2}\cdot\left|\left|\mathbf{w}^{(t+2)}-\mathbf{w}^*\right|\right|^2-\cancel{\frac{\lambda}{2}\cdot\left|\left|\mathbf{w}^{(t+1)}-\mathbf{w}^*\right|\right|^2}\\
&=\frac{\lambda(t-1)}{2}\cdot\left|\left|\mathbf{w}^{(t)}-\mathbf{w}^*\right|\right|^2-\frac{\lambda(t+1)}{2}\cdot\left|\left|\mathbf{w}^{(t+2)}-\mathbf{w}^*\right|\right|^2.
\end{aligned}$$

Question 11.

From the two previous questions deduce then

$$\sum_{t=1}^{T}\left(\hat{\mathcal{L}}\left(\mathbf{w}^{(t)}\right)-\hat{\mathcal{L}}\left(\mathbf{w}^*\right)\right)\leqslant\frac{-\lambda T}{2}\cdot\left|\left|\mathbf{w}^{(T+1)}-\mathbf{w}^*\right|\right|^2+\frac{1}{2}\sum_{t=1}^{T}\eta_t\left|\left|\nabla_t\right|\right|^2\leqslant\frac{1}{2}\sum_{t=1}^{T}\eta_t\left|\left|\nabla_t\right|\right|^2.$$

From the previous question we can find general equation for the sum by applying inductive step. For this part we need to presume that T is even in order to group elements form sum in groups of two. This is not really important because we choose T, so we can always chose even T. So, we have

$$\begin{aligned}
&\sum_{t=1}^{T}\left(\frac{\left|\left|\mathbf{w}^{(t)}-\mathbf{w}^*\right|\right|^2-\left|\left|\mathbf{w}^{(t+1)}-\mathbf{w}^*\right|\right|^2}{2\eta_t}-\frac{\lambda}{2}\cdot\left|\left|\mathbf{w}^{(t)}-\mathbf{w}^*\right|\right|^2\right)\\
&=\sum_{t=1}^{\frac{T}{2}}\sum_{j=2t-1}^{2t}\left(\frac{\left|\left|\mathbf{w}^{(j)}-\mathbf{w}^*\right|\right|^2-\left|\left|\mathbf{w}^{(j+1)}-\mathbf{w}^*\right|\right|^2}{2\eta_j}-\frac{\lambda}{2}\cdot\left|\left|\mathbf{w}^{(j)}-\mathbf{w}^*\right|\right|^2\right)\\
&=\sum_{t=1}^{\frac{T}{2}}\left((\lambda(t-1))\cdot\left|\left|\mathbf{w}^{(2t-1)}-\mathbf{w}^*\right|\right|^2-\lambda t\cdot\left|\left|\mathbf{w}^{(2t+1)}-\mathbf{w}^*\right|\right|^2\right)\\
&=\lambda\left(\sum_{t=1}^{\frac{T}{2}}(t-1)\cdot\left|\left|\mathbf{w}^{(2t-1)}-\mathbf{w}^*\right|\right|^2-\sum_{t=1}^{\frac{T}{2}}t\cdot\left|\left|\mathbf{w}^{(2t+1)}-\mathbf{w}^*\right|\right|^2\right)\\
&=\lambda\left(\sum_{t=1}^{\frac{T}{2}}(t-1)\cdot\left|\left|\mathbf{w}^{(2t-1)}-\mathbf{w}^*\right|\right|^2-\sum_{t=1}^{\frac{T}{2}}t\cdot\left|\left|\mathbf{w}^{(2t+1)}-\mathbf{w}^*\right|\right|^2\right)\\
&=\lambda\left(\sum_{t=1}^{\frac{T}{2}-1}t\cdot\left|\left|\mathbf{w}^{(2t+1)}-\mathbf{w}^*\right|\right|^2-\sum_{t=1}^{\frac{T}{2}-1}t\cdot\left|\left|\mathbf{w}^{(2t+1)}-\mathbf{w}^*\right|\right|^2-\frac{T}{2}\cdot\left|\left|\mathbf{w}^{(T+1)}-\mathbf{w}^*\right|\right|^2\right)\\
&=-\lambda\frac{T}{2}\left|\left|\mathbf{w}^{(T+1)}-\mathbf{w}^*\right|\right|^2<0.
\end{aligned}$$

To calculate the last line in previous calculation we just need to rewrite the second sum as sum plus the last element. Also, we can see that the first sum starts with 0, because $t-1=0$ for $t=1$, so we can translate the sum to start from 1 and to end with $\frac{T}{2} - 1$. Then, two sums will cancel themselves, and we would only have the last part. Also, the last line is obviously lower than zero, because everything is positive and we have a minus. Finally, using the Question 9 we have

$$\sum_{t=1}^{T} \left(\hat{\mathcal{L}} \left(\mathbf{w}^{(t)} \right) - \hat{\mathcal{L}} \left(\mathbf{w}^{*} \right) \right) \leqslant \frac{-\lambda T}{2} \cdot \left| \left| \mathbf{w}^{(T+1)} - \mathbf{w}^{*} \right| \right|^{2} + \frac{1}{2} \sum_{t=1}^{T} \eta_t \left| \left| \nabla_t \right| \right|^2 \leqslant \frac{1}{2} \sum_{t=1}^{T} \eta_t \left| \left| \nabla_t \right| \right|^2 .$$

Question 12.

Suppose that the learning rate $\eta_t = \frac{1}{\lambda \cdot t}$, for all t and that the training data are contained in a ball of radius R; if at each iteration, we normalize the weights $\mathbf{w}^{(t)}$ such that $\left| \left| \mathbf{w}^{(t)} \right| \right| \leqslant \frac{1}{\sqrt{\lambda}}$, show that

$$\left| \left| \nabla_t \right| \right| \leqslant \sqrt{\lambda} + R,$$

and deduce that for $T \geqslant 3$

$$\frac{1}{T} \sum_{t=1}^{T} \hat{\mathcal{L}} \left(\mathbf{w}^{(t)} \right) \leqslant \hat{\mathcal{L}} \left(\mathbf{w}^{*} \right) + \frac{c \cdot (1 + \log T)}{2\lambda \cdot T},$$

where $c = (\sqrt{\lambda} + R)^2$.

For the first part of the question we have

$$\begin{aligned} \left| \left| \nabla_t \right| \right| &= \left| \left| -\frac{1}{m} \sum_{(\mathbf{x},y) \in S_t^{+}} y\mathbf{x} + \lambda \mathbf{w}^{(t)} \right| \right| \leqslant \frac{1}{m} \sum_{(\mathbf{x},y) \in S_t^{+}} y \left| \left| \mathbf{x} \right| \right| + \lambda \cdot \left| \left| \mathbf{w}^{(t)} \right| \right| \\ &\leqslant \frac{1}{m} \sum_{(\mathbf{x},y) \in S_t^{+}} yR + \frac{\lambda}{\sqrt{\lambda}} = R + \sqrt{\lambda}. \end{aligned}$$

Now, using the transformed inequality from the Question 11 and the previous, we have

$$\frac{1}{T} \sum_{t=1}^{T} \hat{\mathcal{L}} \left(\mathbf{w}^{(t)} \right) \leqslant \hat{\mathcal{L}} \left(\mathbf{w}^{*} \right) + \frac{1}{2T} \sum_{t=1}^{T} \eta_t \left| \left| \nabla_t \right| \right|^2 .$$

So, we need to estimate $\frac{1}{2T} \sum\limits_{t=1}^{T} \eta_t \left| \left| \nabla_t \right| \right|^2$. We are using the inequality from this question and definition of η_t, so we have

$$\frac{1}{2T} \sum_{t=1}^{T} \eta_t \left| \left| \nabla_t \right| \right|^2 \leqslant \frac{1}{2T} \sum_{t=1}^{T} \frac{c}{\lambda \cdot t} = \frac{c}{2\lambda \cdot T} \sum_{t=1}^{T} \frac{1}{t} = \frac{c}{2\lambda \cdot T} \cdot H_T,$$

where H_T is T^{th} harmonic number. We will use well known inequality for harmonic numbers $H_T \leqslant \log T + 1$, so we finally have

$$\frac{1}{T} \sum_{t=1}^{T} \hat{\mathcal{L}} \left(\mathbf{w}^{(t)} \right) \leqslant \hat{\mathcal{L}} \left(\mathbf{w}^{*} \right) + \frac{c \cdot (\log T + 1)}{2\lambda \cdot T}.$$

Question 13.

As the learning objective is convex we have from the Jensen inequality that

$$\hat{\mathcal{L}}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)}\right) \leqslant \frac{1}{T}\sum_{t=1}^{T}\hat{\mathcal{L}}\left(\mathbf{w}^{(t)}\right).$$

Using the above inequality and question 12, prove that

$$\hat{\mathcal{L}}(\mathbf{w}^*) \leqslant \hat{\mathcal{L}}(\overline{\mathbf{w}}) \leqslant \hat{\mathcal{L}}(\mathbf{w}^*) + \frac{c\cdot(1+\log T)}{2\lambda\cdot T}$$

where $\overline{\mathbf{w}} = \frac{1}{T}\sum\limits_{t=1}^{T}\mathbf{w}^{(t)}$, and finally

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)} = \mathbf{w}^*.$$

Firstly, remember that

$$\mathbf{w}^* = \arg\min_{\mathbf{w}}\hat{\mathcal{L}}(S,\mathbf{w}),$$

which means that $\hat{\mathcal{L}}(\mathbf{w}^*) \leqslant \hat{\mathcal{L}}(\overline{\mathbf{w}})$. As far as the second inequality consider we are using Jensen inequality and question 12 and have

$$\hat{\mathcal{L}}(\overline{\mathbf{w}}) = \hat{\mathcal{L}}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)}\right) \leqslant \frac{1}{T}\sum_{t=1}^{T}\hat{\mathcal{L}}\left(\mathbf{w}^{(t)}\right) \leqslant \hat{\mathcal{L}}(\mathbf{w}^*) + \frac{c\cdot(1+\log T)}{2\lambda\cdot T}.$$

Now if we apply limit when T tends to infinity on the

$$\hat{\mathcal{L}}(\mathbf{w}^*) \leqslant \hat{\mathcal{L}}(\overline{\mathbf{w}}) \leqslant \hat{\mathcal{L}}(\mathbf{w}^*) + \frac{c\cdot(1+\log T)}{2\lambda\cdot T}$$

we will have

$$\lim_{T\to\infty}\hat{\mathcal{L}}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)}\right) = \hat{\mathcal{L}}\left(\mathbf{w}^*\right).$$

Reason for that is because linear function faster converge to infinity than logarithmic one, so the second element on the right will go to the zero as $T \to \infty$. Finally, having in mind that $\hat{\mathcal{L}}$ is continuous function and that $\mathbf{w}^*$ unique we have

$$\hat{\mathcal{L}}\left(\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)}\right) = \hat{\mathcal{L}}\left(\mathbf{w}^*\right),$$

i.e.

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)} = \mathbf{w}^*.$$