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

# Support Vector Machines

Instructor: *Nicolò Cesa-Bianchi* version of May 23, 2023

The Support Vector Machine (SVM) is an algorithm for learning linear classifiers. Given a linearly separable training set $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_m, y_m) \in \mathbb{R}^d \times \{-1, 1\}$, SVM outputs the linear classifier corresponding to the unique solution $\boldsymbol{w}^* \in \mathbb{R}^d$ of the following convex optimization problem with linear constraints

$$\begin{aligned} \min_{\boldsymbol{w} \in \mathbb{R}^d} \quad & \frac{1}{2} \|\boldsymbol{w}\|^2 \\ \text{s.t.} \quad & y_t \, \boldsymbol{w}^\top \boldsymbol{x}_t \geq 1 \quad t = 1, \dots, m. \end{aligned} \tag{1}$$

Geometrically, $\boldsymbol{w}^*$ corresponds to the **maximum margin separating hyperplane**. For every linearly separable set $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_m, y_m) \in \mathbb{R}^d \times \{-1, 1\}$, the maximum margin is defined by

$$\gamma^* = \max_{\boldsymbol{u} \,:\, \|\boldsymbol{u}\|=1} \; \min_{t=1,\dots,m} y_t \boldsymbol{u}^\top \boldsymbol{x}_t$$

and the vector $\boldsymbol{u}^*$ achieving the maximum margin is the maximum margin separator.

**Theorem 1.** *For every linearly separable set $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_m, y_m) \in \mathbb{R}^d \times \{-1, 1\}$, the maximum margin separator $\boldsymbol{u}^*$ satisfies $\boldsymbol{u}^* = \gamma^* \boldsymbol{w}^*$, where $\boldsymbol{w}^*$ is the unique solution of (1).*

Proof. Note that $\boldsymbol{u}^*$ is the solution of the following optimization problem

$$\begin{aligned} \max_{\boldsymbol{u} \in \mathbb{R}^d, \gamma > 0} \quad & \gamma^2 \\ \text{s.t.} \quad & \|\boldsymbol{u}\|^2 = 1 \\ & y_t \, \boldsymbol{u}^\top \boldsymbol{x}_t \geq \gamma \quad t = 1, \dots, m. \end{aligned}$$

Indeed, $\boldsymbol{u}$ maximizing the margin $\gamma$ is the same $\boldsymbol{u}$ maximizing $\gamma^2$ because the function $f(\gamma) = \gamma^2$, is monotone for $\gamma > 0$. Dividing by $\gamma > 0$ both sides of each constraint $y_t \, \boldsymbol{u}^\top \boldsymbol{x}_t \geq \gamma$, we obtain the equivalent constraint $y_t\big(\boldsymbol{u}^\top \boldsymbol{x}_t\big)/\gamma \geq 1$. Introducing $\boldsymbol{w} = \boldsymbol{u}/\gamma$, and noting that $\|\boldsymbol{w}\|^2 = 1/\gamma^2$ because of the constraint $\|\boldsymbol{u}\|^2 = 1$, we obtain the equivalent problem

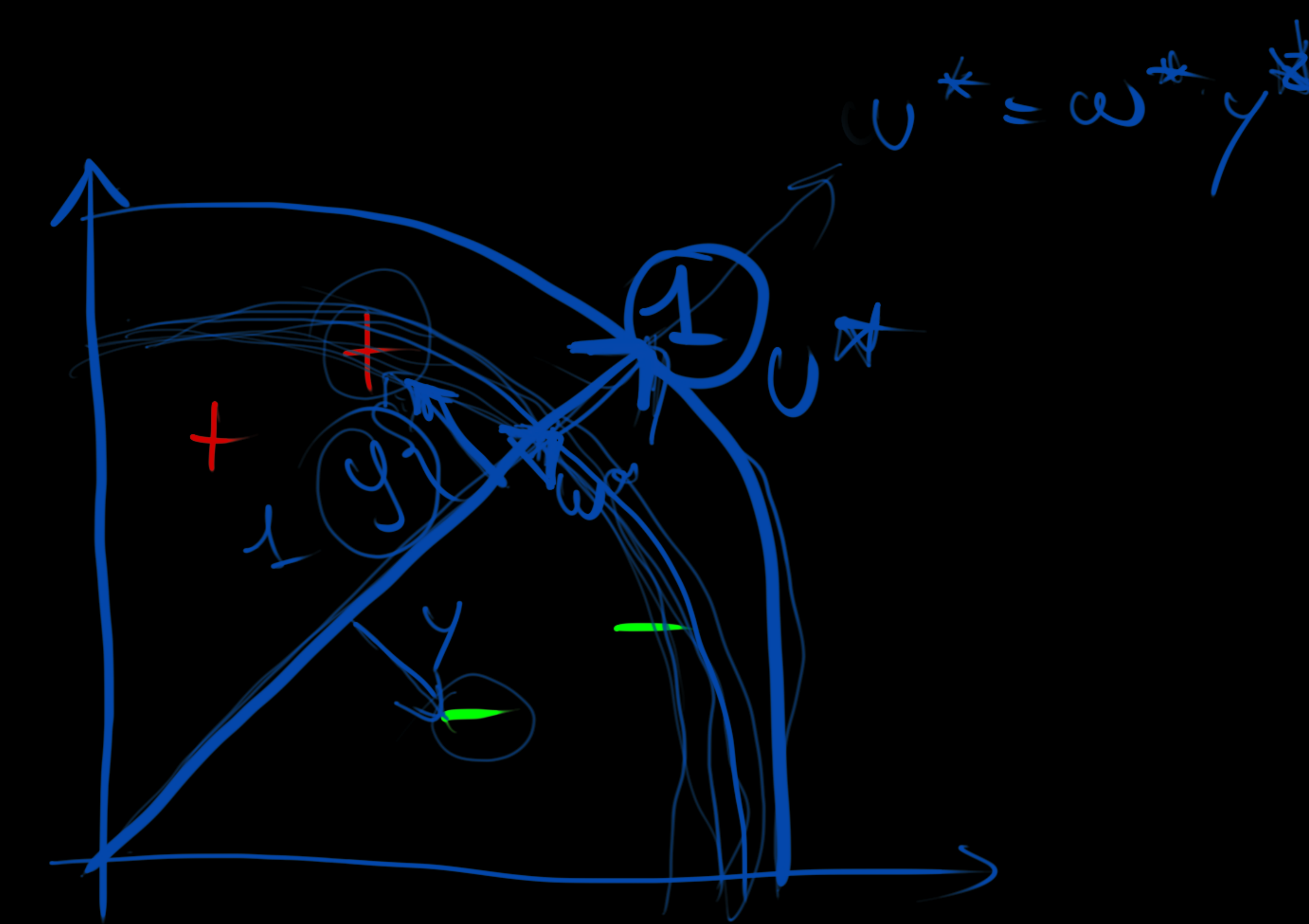
$$\begin{aligned} \min_{\boldsymbol{w} \in \mathbb{R}^d, \gamma > 0} \quad & \|\boldsymbol{w}\|^2 \\ \text{s.t.} \quad & \cancel{\gamma^2 \|\boldsymbol{w}\|^2 = 1} \\ & y_t \, \boldsymbol{w}^\top \boldsymbol{x}_t \geq 1 \quad t = 1, \dots, m. \end{aligned}$$

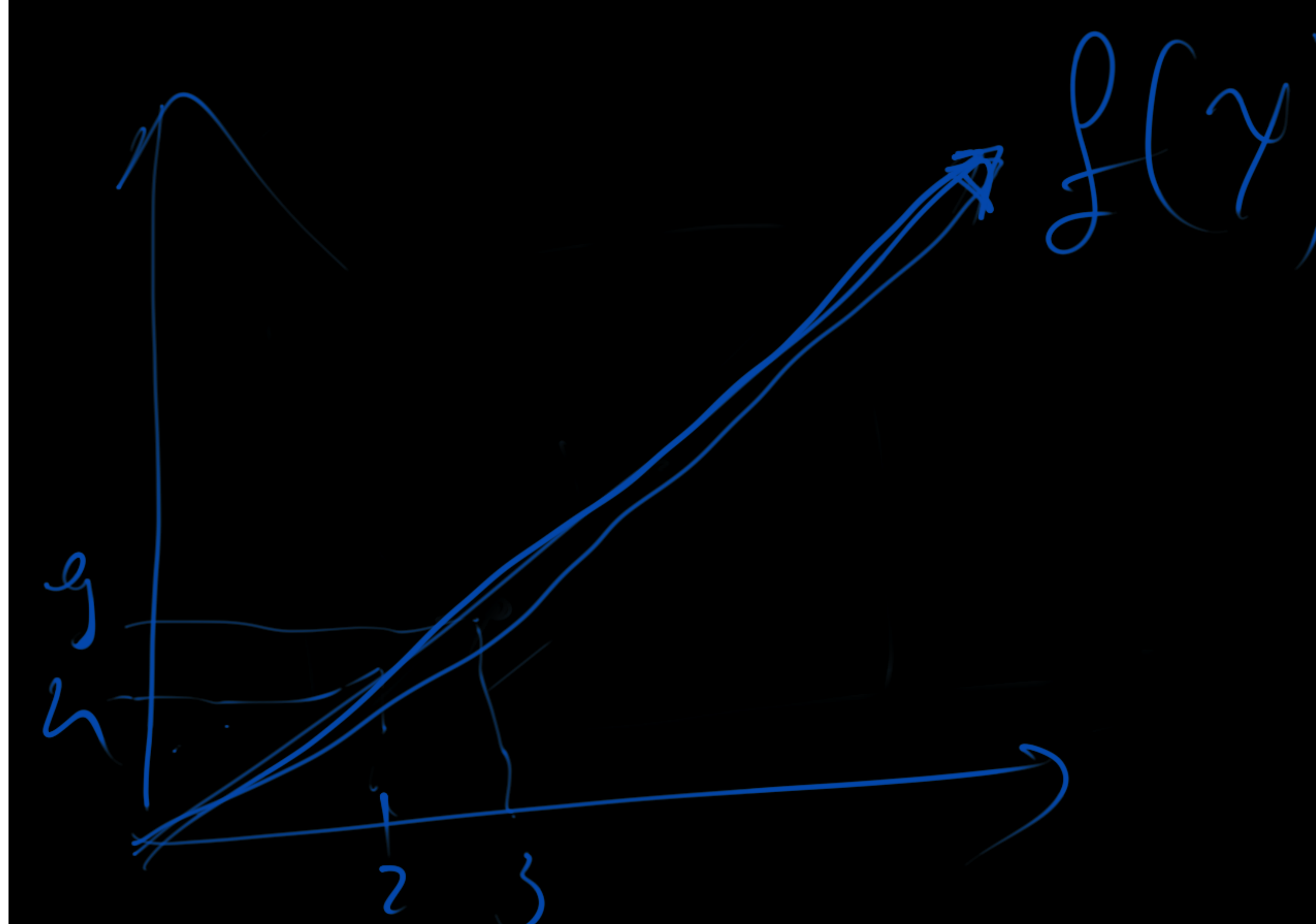
Now observe that the constraint $\gamma^2 \|\boldsymbol{w}\|^2 = 1$ is redundant and can be eliminated. Indeed, for all $\boldsymbol{w} \in \mathbb{R}^d$ we can find $\gamma > 0$ such that the constraint is satisfied. Multiplying the objective function by $\frac{1}{2}$, we obtain

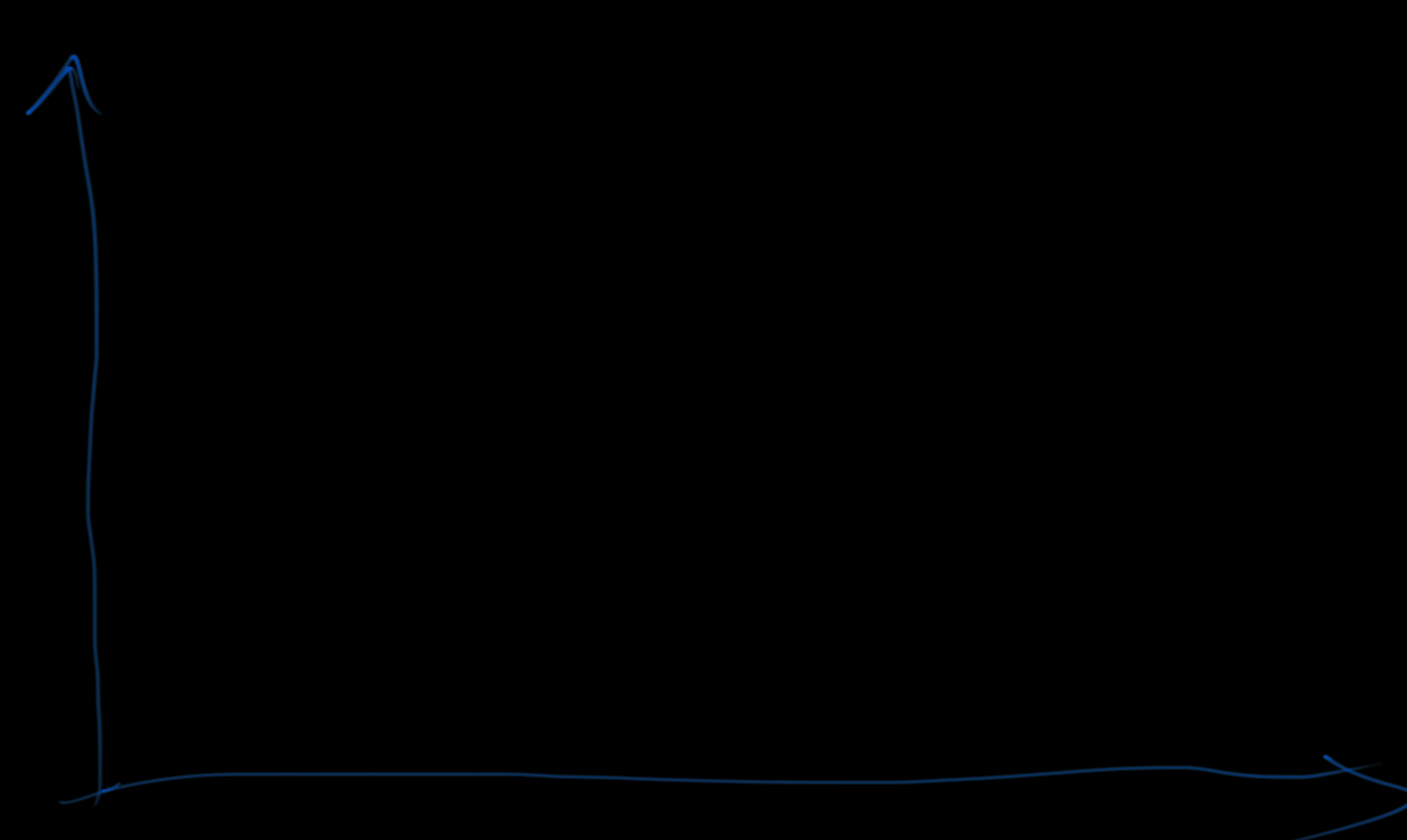
$$\begin{aligned} \min_{\boldsymbol{w} \in \mathbb{R}^d} \quad & \frac{1}{2} \|\boldsymbol{w}\|^2 \\ \text{s.t.} \quad & y_t \, \boldsymbol{w}^\top \boldsymbol{x}_t \geq 1 \quad t = 1, \dots, m \end{aligned}$$

concluding the proof. $\square$





We have thus shown the equivalence between the problem of maximizing the margin of $\boldsymbol{u}$ while keeping the norm $\|\boldsymbol{u}\|$ constant, and the problem of minimizing the norm $\|\boldsymbol{w}\|$ while keeping the margin of $\boldsymbol{w}$ constant.

The following result helps us compute the form of the optimal solution $\boldsymbol{w}^*$.

**Lemma 2** (Fritz John optimality condition). *Consider the problem*

$$\begin{array}{ll} \min\limits_{\boldsymbol{w}\in\mathbb{R}^d} & f(\boldsymbol{w}) \\ s.t. & g_t(\boldsymbol{w}) \le 0 \quad t = 1,\dots,m \end{array}$$

*where the functions $f, g_1, \dots, g_m$ are all differentiable. If $\boldsymbol{w}_0$ is an optimal solution, then there exists a nonnegative vector $\boldsymbol{\alpha} \in \mathbb{R}^m$ such that*

$$\boxed{\nabla f(\boldsymbol{w}_0)} + \sum_{t\in I} \alpha_t \nabla g_t(\boldsymbol{w}_0) = \boldsymbol{0}$$

*where $I = \{1 \le t \le m \,:\, g_t(\boldsymbol{w}_0) = 0\}$.*

By applying the Fritz John optimality condition to the SVM objective with $f(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2$ and $g_t(\boldsymbol{w}) = 1 - y_t\,\boldsymbol{w}^\top \boldsymbol{x}_t$ we obtain

$\boldsymbol{w}^*$

$g_t(w) = 0$ if our pred. is ok

$$\boldsymbol{w}^* - \sum_{t\in I} \alpha_t y_t\, \boldsymbol{x}_t = \boldsymbol{0}\ .$$

Hence, the optimal solution has form

$$\boldsymbol{w}^* = \sum_{t\in I} \alpha_t y_t\, \boldsymbol{x}_t$$

where $I$ denotes the set of training examples $(\boldsymbol{x}_t, y_t)$ such that $\boxed{y_t(\boldsymbol{w}^*)^\top \boldsymbol{x}_t = 1.}$ These $\boldsymbol{x}_t$ are called **support vectors**, and are all those training points for which the margin of $\boldsymbol{w}^*$ is exactly 1. If we removed all training examples except for the support vectors, the SVM solution would not change.

We now move on to consider the case of a training set that is not linearly separable. How should we change the SVM objective? Conside the following formulation

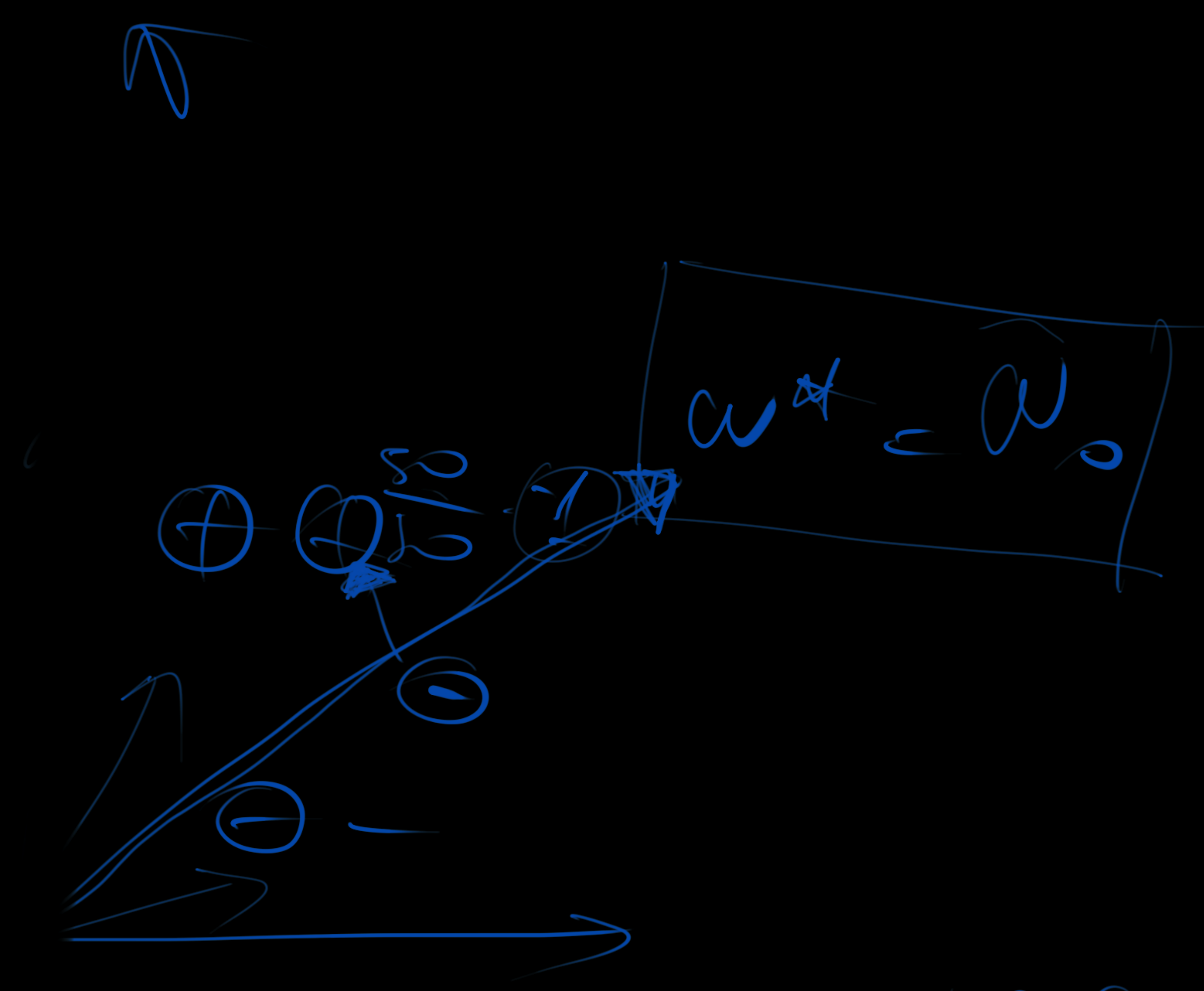
$$\begin{array}{lll} \min\limits_{(\boldsymbol{w},\boldsymbol{\xi})\in\mathbb{R}^{d+m}} & \dfrac{\lambda}{2}\|\boldsymbol{w}\|^2 + \dfrac{1}{m}\displaystyle\sum_{t=1}^{m} \xi_t & \\ \text{s.t.} & y_t\,\boldsymbol{w}^\top \boldsymbol{x}_t \ge 1 - \xi_t & t = 1,\dots,m \\ & \xi_t \ge 0 & t = 1,\dots,m. \end{array}$$

The components of $\boldsymbol{\xi} = (\xi_1, \dots, \xi_m)$ are called **slack variables** and measure how much each margin constraint is violated by a potential solution $\boldsymbol{w}$. The average of these violations is then added to the objective function. Finally, a regularization parameter $\lambda > 0$ is introduced to balance the two terms.

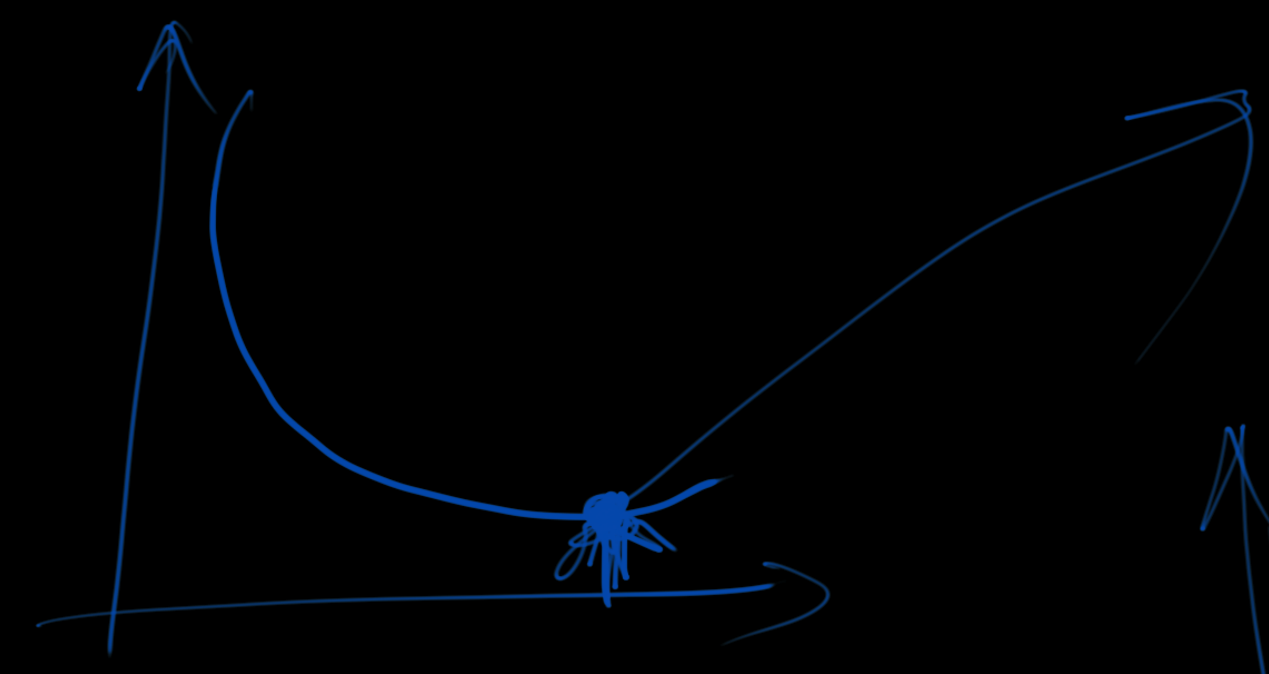
We now consider the constraints involving the slack variables $\xi_t$. That is, $\xi_t \ge 1 - y_t\,\boldsymbol{w}^\top \boldsymbol{x}_t$ and $\xi_t \ge 0$. In order to minimize each $\xi_t$, we can set

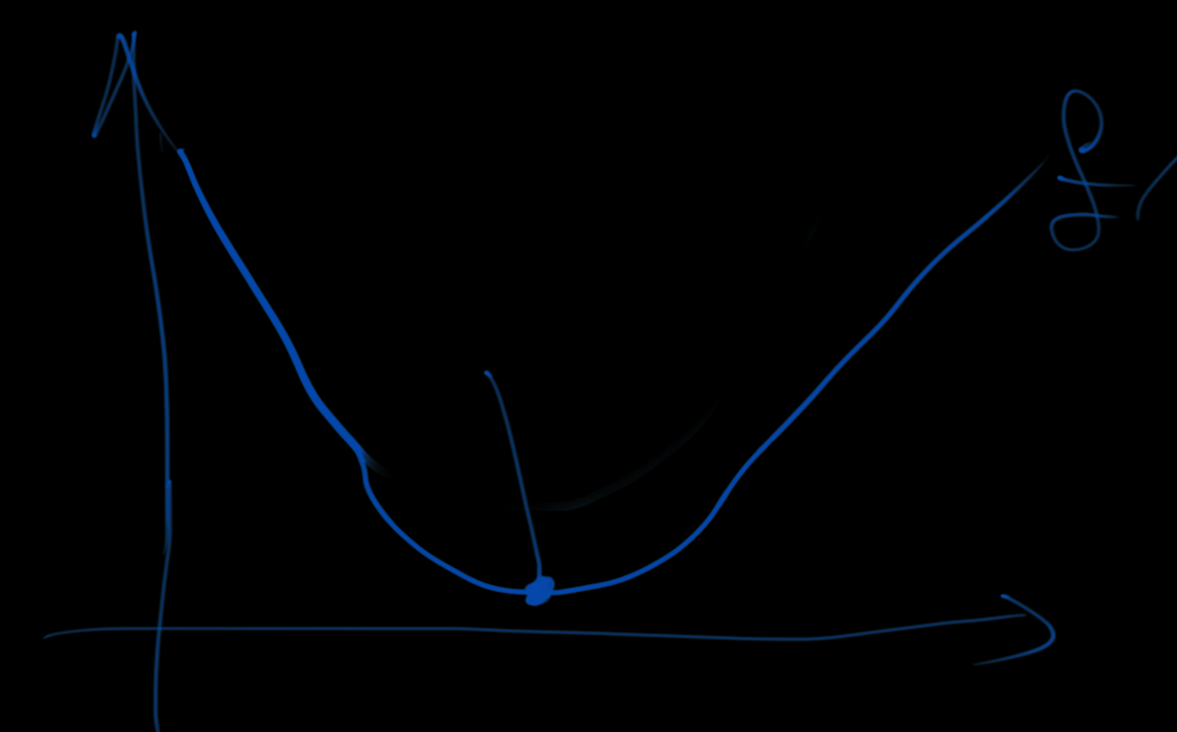
$$\xi_t = \begin{cases} 1 - y_t\,\boldsymbol{w}^\top \boldsymbol{x}_t & \text{if } y_t\,\boldsymbol{w}^\top \boldsymbol{x}_t < 1 \\ 0 & \text{otherwise.} \end{cases}$$






$\nabla f(w^*) = 0$






[1,2,3] [1]

[1,2,3] · α

To see this, fix $\boldsymbol{w} \in \mathbb{R}^d$. If the constraint $y_t \, \boldsymbol{w}^\top \boldsymbol{x}_t \geq 1$ is satisfied by $\boldsymbol{w}$, then $\xi_t$ can be set to zero. Otherwise, if the constraint is not satisfied by $\boldsymbol{w}$, then we set $\xi_t$ to the smallest value such that the constraint becomes satisfied, namely $1 - y_t \, \boldsymbol{w}^\top \boldsymbol{x}_t$. Summarizing, $\xi_t = \left[1 - y_t \, \boldsymbol{w}^\top \boldsymbol{x}_t\right]_+$, which is exactly the hinge loss $h_t(\boldsymbol{w})$ of $\boldsymbol{w}$.

The SVM problem can then be re-formulated as $\min\limits_{\boldsymbol{w} \in \mathbb{R}^d} F(\boldsymbol{w})$, where

$$F(\boldsymbol{w}) = \overset{\ell_S(w)}{\boxed{\frac{1}{m} \sum_{t=1}^{m} h_t(\boldsymbol{w})}} + \frac{\lambda}{2} \|\boldsymbol{w}\|^2 \ .$$

We now show that, even when the training set is not linearly separable, the solution $\boldsymbol{w}^*$ belongs to the subspace defined by linear combinations of training points multiplied by their labels.

**Theorem 3.** *The minimizer $\boldsymbol{w}^*$ of $F$ can be written as a linear combination of $y_1\boldsymbol{x}_1, \dots, y_m\boldsymbol{x}_m$.*

$\alpha_1$ $\alpha_m$

PROOF. By contradiction, assume

$$\boldsymbol{w}^* = \overbrace{\sum_{t=1}^{m} \alpha_t \, y_t \, \boldsymbol{x}_t}^{w^*} + \boldsymbol{u} \tag{2}$$

where $\boldsymbol{u} \in \mathbb{R}^d$ is the component of $\boldsymbol{w}^*$ orthogonal to the subspace spanned by $\boldsymbol{x}_1, \dots, \boldsymbol{x}_m$. Therefore,

$$y_t \boldsymbol{u}^\top \boldsymbol{x}_t = 0 \qquad t = 1, \dots, m. \tag{3}$$

Now, let $\boldsymbol{v} = \boldsymbol{w}^* - \boldsymbol{u}$. First, $\|\boldsymbol{v}\|^2 \leq \|\boldsymbol{w}^*\|^2$ because in (2) we wrote $\boldsymbol{w}^*$ as a sum of two orthogonal components and we removed one of them, and so its length decreased. Second,

$$h_t(\boldsymbol{v}) = \left[1 - y_t \boldsymbol{v}^\top \boldsymbol{x}_t\right]_+ = \left[1 - y_t \left(\boldsymbol{w}^* - \boldsymbol{u}\right)^\top \boldsymbol{x}_t\right]_+ = \left[1 - y_t (\boldsymbol{w}^*)^\top \boldsymbol{x}_t + y_t \boldsymbol{u}^\top \boldsymbol{x}_t\right]_+ = h_t(\boldsymbol{w}^*)$$

using (3). Therefore $F(\boldsymbol{v}) \leq F(\boldsymbol{w}^*)$, contradicting the optimality of $\boldsymbol{w}^*$. Hence $\boldsymbol{u} = \boldsymbol{0}$ and the proof is concluded. □

Note that, as in the linearly separable case, $\boldsymbol{w}^*$ generally depends on a subset of support vectors. However, unlike the linearly separable case, these support vectors also include the training points associated with positive slack variables.

We proceed by showing how $F$ can be minimized using Online Gradient Descent (OGD). First, observe that

$$F(\boldsymbol{w}) = \frac{1}{m} \sum_{t=1}^{m} \ell_t(\boldsymbol{w})$$

where $\ell_t(\boldsymbol{w}) = h_t(\boldsymbol{w}) + \frac{\lambda}{2} \|\boldsymbol{w}\|^2$ is a strongly convex function. Indeed, $\frac{\lambda}{2} \|\boldsymbol{w}\|^2$ is $\lambda$-strongly convex, and $h_t$ is convex (and also piecewise linear). This implies that their sum is $\lambda$-strongly convex. We can then apply the OGD algorithm for strongly convex functions to the set of losses $\ell_1, \dots, \ell_m$. This instance of OGD, which is known as **Pegasos**, can be described as follows.

$\min \frac{\lambda}{2} \|w\|^2 + \frac{1}{m}\sum_{i}^{m} \xi_i$

> **Parameters:** number $T$ of rounds, regularization coefficient $\lambda > 0$
> **Input:** Training set $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_m, y_m) \in \mathbb{R}^d \times \{-1, 1\}$
> Set $\boldsymbol{w}_1 = \boldsymbol{0}$
> For $t = 1, \dots, T$
>
> 1. Draw uniformly at random an element $(\boldsymbol{x}_{Z_t}, y_{Z_t})$ from the training set
> 2. Set $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \nabla \ell_{Z_t}(\boldsymbol{w}_t)$
>
> Output: $\overline{\boldsymbol{w}} = \frac{1}{T}(\boldsymbol{w}_1 + \cdots + \boldsymbol{w}_T)$.

Pegasos is an example of a class of algorithms known as **stochastic gradient descent**. These are OGD-like algorithms that are run over a sequence of examples randomly drawn from the training set.

We now move on to analyze Pegasos. Let $(\boldsymbol{x}_{Z_1}, y_{Z_1}), \dots, (\boldsymbol{x}_{Z_T}, y_{Z_T})$ the sequence of training set examples that were drawn at random in step 1 of the algorithm, and let $\ell_{Z_1}, \dots, \ell_{Z_T}$ the corresponding sequence of loss functions. Namely, $\ell_{Z_t}(\boldsymbol{w}) = h_{Z_t}(\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2$ where $h_{Z_t}(\boldsymbol{w}) = \left[1 - y_{Z_t} \boldsymbol{w}^\top \boldsymbol{x}_{Z_t}\right]_+$.

Let $\boldsymbol{w}^*$ be the optimal SVM solution,

ERM
$$\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^d} \left( \frac{1}{m} \sum_{t=1}^{m} h_t(\boldsymbol{w}) + \frac{\lambda}{2} \|\boldsymbol{w}\|^2 \right) . \tag{4}$$

For every realization $s_1, \dots, s_T$ of the random variables $Z_1, \dots, Z_T$, OGD analysis for strongly convex losses immediately gives

$$\frac{1}{T} \sum_{t=1}^{T} \ell_{s_t}(\boldsymbol{w}_t) \le \frac{1}{T} \sum_{t=1}^{T} \ell_{s_t}(\boldsymbol{w}^*) + \frac{G^2}{2\lambda T} \ln(T+1) \tag{5}$$

where $G = \max_{t=1,\dots,T} \|\nabla \ell_{s_t}(\boldsymbol{w}_t)\|$ is also a random variable.

In order to show how this result can be used to bound $F(\overline{\boldsymbol{w}})$, we use the following fact

$$\mathbb{E}\left[\ell_{Z_t}(\boldsymbol{w}_t) \mid Z_1, \dots, Z_{t-1}\right] = \frac{1}{m} \sum_{s=1}^{m} \ell_s(\boldsymbol{w}_t) = F(\boldsymbol{w}_t) . \tag{6}$$

In other words, conditioned on the first $t-1$ random draws (which determine $\boldsymbol{w}_t$), the expected value of $\ell_{Z_t}(\boldsymbol{w}_t)$ is equal to $F(\boldsymbol{w}_t)$. We also use the fact that for every pair of random variables

$X, Y$ the following holds $\mathbb{E}[X] = \mathbb{E}\big[\mathbb{E}[X \mid Y]\big]$. Hence, we can write

$$
\begin{aligned}
\mathbb{E}\big[F(\overline{\boldsymbol{w}})\big] &= \mathbb{E}\left[F\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{w}_t\right)\right] \\
&\le \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}F(\boldsymbol{w}_t)\right] \quad \text{using Jensen inequality, since } F \text{ is convex} \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\ell_{Z_t}(\boldsymbol{w}_t) \mid Z_1,\dots,Z_{t-1}\big]\right] \quad \text{using (6)} \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\ell_{Z_t}(\boldsymbol{w}_t)\right] \quad \text{using } \mathbb{E}[X] = \mathbb{E}\big[\mathbb{E}[X \mid Y]\big] \\
&\le \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\ell_{Z_t}(\boldsymbol{w}^*)\right] + \frac{\mathbb{E}\big[G^2\big]}{2\lambda T}\big(\ln T + 1\big) \quad \text{using (5)} \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\ell_{Z_t}(\boldsymbol{w}^*) \mid Z_1,\dots,Z_{t-1}\big]\right] + \frac{\mathbb{E}\big[G^2\big]}{2\lambda T}\big(\ln T + 1\big) \quad \text{using } \mathbb{E}[X] = \mathbb{E}\big[\mathbb{E}[X \mid Y]\big] \\
&= F(\boldsymbol{w}^*) + \frac{\mathbb{E}\big[G^2\big]}{2\lambda T}\ln(T+1) \quad \text{using (6).}
\end{aligned}
$$

We thus obtained

$$
\mathbb{E}\big[F(\overline{\boldsymbol{w}})\big] \le F(\boldsymbol{w}^*) + \frac{\mathbb{E}\big[G^2\big]}{2\lambda T}\big(\ln T + 1\big) \ . \tag{7}
$$

Therefore, if $\mathbb{E}\big[G^2\big]$ can be upper bounded by a constant, the average $\overline{\boldsymbol{w}}$ of the vectors generated by OGD converges (in expectation with respect to the random draw of the elements from the training set) to $\boldsymbol{w}^*$ with rate $\frac{\ln T}{T}$. With a bit more work, one can show that $\overline{\boldsymbol{w}}$ converges to $\boldsymbol{w}^*$ not only in expectation but also in probability.

We now bound $G$ for every realization $s_1,\dots,s_T$ of the random variables $Z_1,\dots,Z_T$. We have $\nabla\ell_{s_t}(\boldsymbol{w}_t) = -y_{s_t}\,\boldsymbol{x}_{s_t}\,\mathbb{I}\{h_{s_t}(\boldsymbol{w}_t) > 0\} + \lambda\,\boldsymbol{w}_t$. Let $\boldsymbol{v}_t = y_{s_t}\,\boldsymbol{x}_{s_t}\,\mathbb{I}\{h_{s_t}(\boldsymbol{w}_t) > 0\}$. Because $\eta_t = 1/(\lambda t)$, the update rule for $\boldsymbol{w}_t$ takes the following simple form,

$$
\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t\nabla\ell_t(\boldsymbol{w}_t) = \boldsymbol{w}_t + \eta_t\boldsymbol{v}_t - \eta_t\lambda\boldsymbol{w}_t = \left(1 - \frac{1}{t}\right)\boldsymbol{w}_t + \frac{1}{\lambda t}\boldsymbol{v}_t \ .
$$

Let $X = \max_{s=1,\dots,m}\|\boldsymbol{x}_s\|$. Since $\|\nabla\ell_{s_t}(\boldsymbol{w}_t)\| \le \|\boldsymbol{v}_t\| + \lambda\,\|\boldsymbol{w}_t\| \le X + \lambda\,\|\boldsymbol{w}_t\|$, we are left with the task of computing an upper bound for $\|\boldsymbol{w}_t\|$. In order to do so, we look at the recurrence

$$
\boldsymbol{w}_{t+1} = \left(1 - \frac{1}{t}\right)\boldsymbol{w}_t + \frac{1}{\lambda t}\boldsymbol{v}_t \ .
$$

As one can easily show by induction, $\boldsymbol{w}_{t+1}$ can be written as a linear combination of $\boldsymbol{v}_1,\dots,\boldsymbol{v}_t$. In order to determine the coefficients of this linear combination, we fix $s \le t$ and observe that $\boldsymbol{v}_s$ is added to the sum with coefficient $1/(\lambda s)$. When $\boldsymbol{w}_{t+1}$, is computed, the coefficient of $\boldsymbol{v}_s$ has become

$$
\frac{1}{\lambda s}\prod_{r=s+1}^{t}\left(1 - \frac{1}{r}\right) = \frac{1}{\lambda s}\prod_{r=s+1}^{t}\frac{r-1}{r} = \frac{1}{\lambda t} \ .
$$

We thus obtain a simple expression for $\boldsymbol{w}_{t+1}$,

$$\boldsymbol{w}_{t+1} = \frac{1}{\lambda t} \sum_{s=1}^{t} \boldsymbol{v}_s \ . \tag{8}$$

Because $\boldsymbol{w}_{t+1}$ is an average of $\boldsymbol{v}_s$ divided by $\lambda$, we finally have $\|\boldsymbol{w}_{t+1}\| \le \frac{1}{\lambda} \max_s \|\boldsymbol{v}_s\| \le \frac{1}{\lambda} X$. This allows us to conclude that $\|\nabla \ell_t(\boldsymbol{w}_t)\| \le X + \lambda \|\boldsymbol{w}_t\| \le 2X$. Substituting this bound for $G$ in (7) we get

$$\mathbb{E}\big[F(\overline{\boldsymbol{w}})\big] \le F(\boldsymbol{w}^*) + \frac{2X^2}{\lambda T} \ln(T+1) \ .$$

Theorem 3 states that the solution $\boldsymbol{w}^*$ to the SVM problem can be written as

$$\boldsymbol{w}^* = \sum_{s \in S} y_s \alpha_s \boldsymbol{x}_s$$

where $\alpha_s > 0$ and $S \equiv \{t = 1, \dots, m : h_t(\boldsymbol{w}^*) > 0\}$. An important consequence of this result is that we can solve the problem (4) in a RKHS $\mathcal{H}_K$, where the objective function $F$ becomes

$$F_K(g) = \frac{1}{m} \sum_{t=1}^{m} h_t(g) + \frac{\lambda}{2} \|g\|_K^2 \qquad g \in \mathcal{H}_K$$

with $h_t(g) = \big[1 - y_t g(\boldsymbol{x}_t)\big]_+$. In $\mathcal{H}_K$, the SVM solution can therefore be written as

$$\sum_{s \in S} y_s \alpha_s K(\boldsymbol{x}_s, \cdot)$$

which is clearly an element of the RKHS

$$\mathcal{H}_K \equiv \left\{ \sum_{i=1}^{N} \alpha_i K(\boldsymbol{x}_i, \cdot) \ : \ \boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \mathbb{R}^d, \, \alpha_1, \dots, \alpha_N \in \mathbb{R}, \, N \in \mathbb{N} \right\}$$

As we did for the Perceptron, we can run Pegasos in the RKHS $\mathcal{H}_K$. The gradient update in kernel Pegasos on some training example $(\boldsymbol{x}_{s_t}, y_{s_t})$ can be written as

$$g_{t+1} = \left(1 - \frac{1}{t}\right) g_t + \frac{y_{s_t}}{\lambda t} \mathbb{I}\{h_{s_t}(g_t) > 0\} K(\boldsymbol{x}_{s_t}, \cdot)$$

where $h_{s_t}(g_t) = \big[1 - y_{s_t} g_t(\boldsymbol{x}_{s_t})\big]_+$.