

## 1. Regularized polynomial regression

We derived in class the solution for a zero-degree polynomial regression. Consider the problem of regularized polynomial regression.

$$Err(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} (h_{\mathbf{w}}(\mathbf{x}_i) - y_i)^2 + \lambda ||\mathbf{w}||^2 .$$

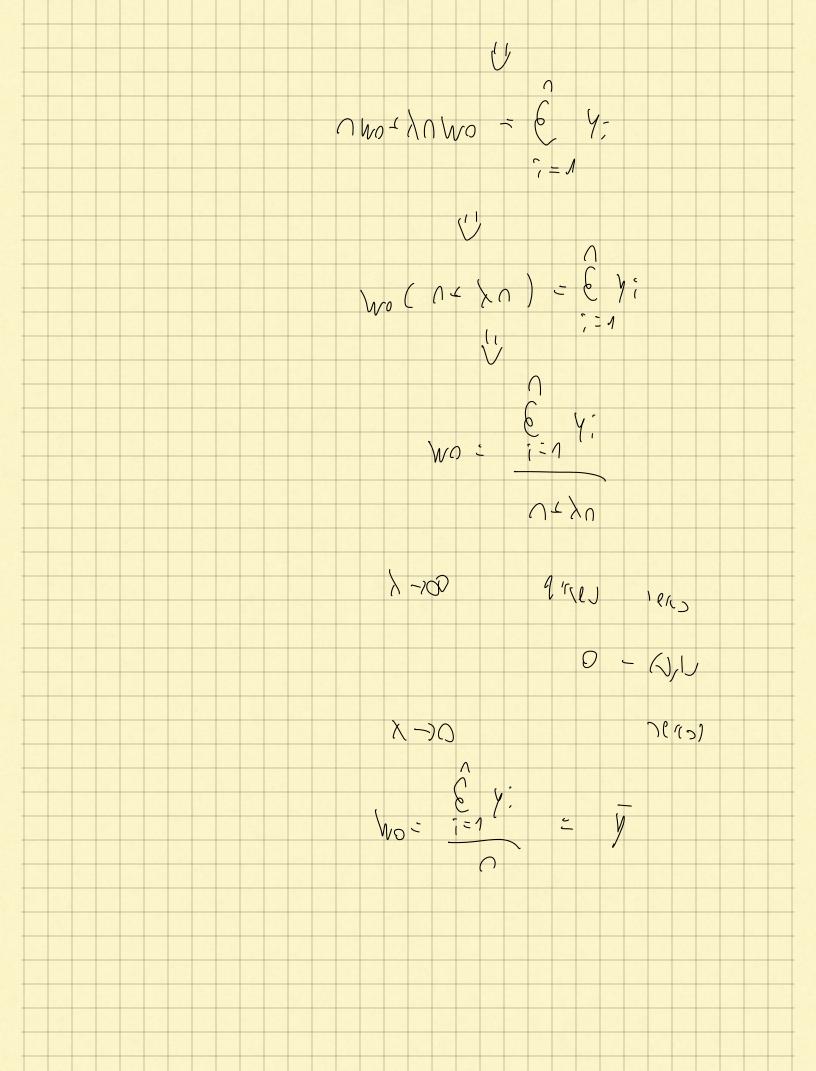
- 1. Derive the solution for a polynomial of degree 0:  $h_{\mathbf{w}}(\mathbf{x}) = w_0$ . Analyze the solution in the limit of  $\lambda \to \infty$  and  $\lambda \to 0$ .
- 2. Derive the solution for a polynomial of degree 1:  $h_{\mathbf{w}}(\mathbf{x}) = w_0 + w_1 x$ , by computing the derivatives w.r.t.  $w_0$  and  $w_1$  and writing a system of two linear equations in  $w_0$  and  $w_1$ . No need to solve the system. Analyze the solution in the limit of  $\lambda \to \infty$  and  $\lambda \to 0$ .

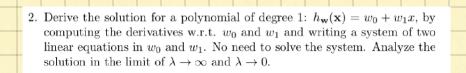
$$E_{V_{1}}(w) = \frac{1}{16} \frac{E}{E} |w(x_{1}) - y_{1}|^{2} + \lambda ||w|||^{2}$$

$$\frac{1}{16} \frac{E}{E} |w(x_{1}) - y_{1}|^{2} + \lambda ||w||^{2}$$

$$\frac{1}{16} \frac{E}{E} |w(x_{1}) - y_{1}|^{2} + \lambda ||w||^{2} + \lambda ||w||^{2}$$

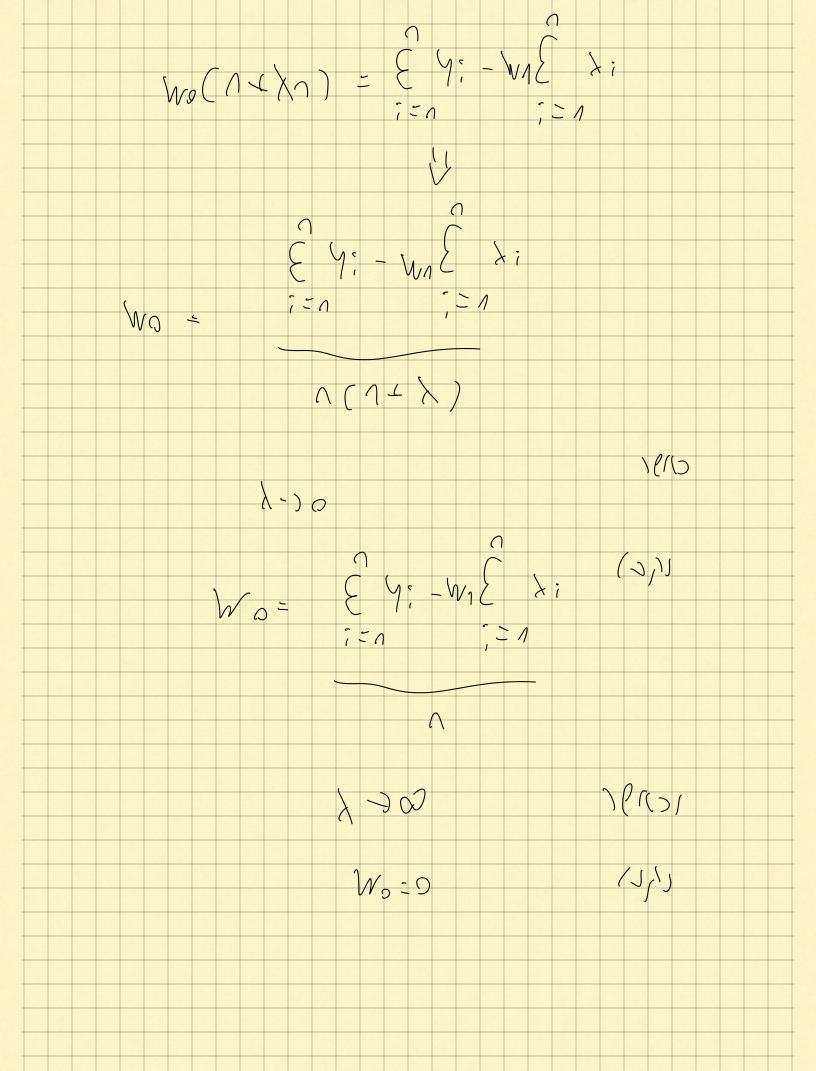
$$\frac{1}{16} \frac{E}{E} |w(x_{1}) - y_{1}|^{2} + \lambda ||w||^{2} + \lambda ||w||^{2}$$





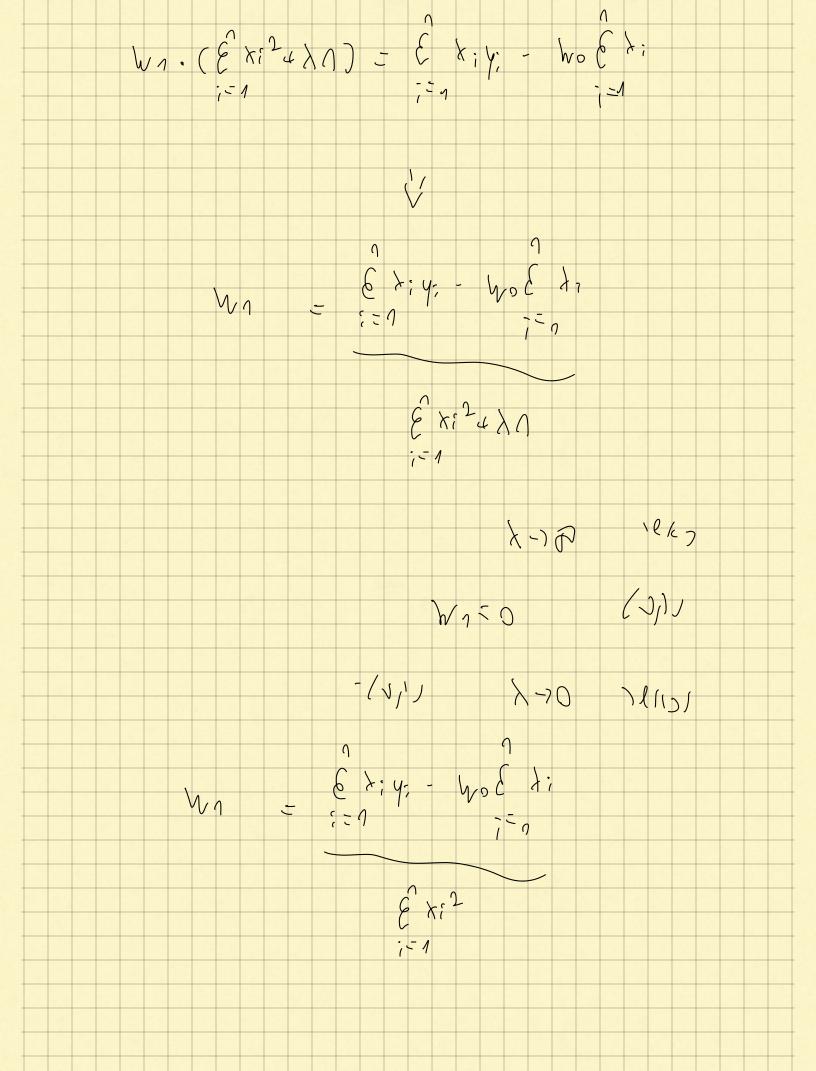






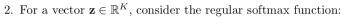
$$\begin{array}{c} W1 & (1/257) \\ \times V & \times V \end{array}$$

$$\begin{array}{c} EV(W) = \frac{1}{7} \stackrel{?}{E} & (W_0 \cdot W_1 X_1 - Y_1) & (W_0^2 \cdot W_1^2) \\ \times V & (W_0^2 \cdot W_1^2 X_1 - Y_1) &$$



## 2. Logistic regression

1. Prove that the logistic regression classifier is equivalent to a softmax over a linear multiclass classifier for two classes y = "a", y = "b", when their separating hyperplanes obey  $\mathbf{w}_a = -\mathbf{w}_b$ .



$$\operatorname{softmax}_{i}(\mathbf{z}) = \frac{\exp(z_{i})}{\sum_{k=1}^{K} \exp(z_{k})}$$
 (1)

For any vector  $\mathbf{b} = (b, \dots, b) \in \mathbb{R}^K$  for some  $b \in \mathbb{R}$ , prove that softmax<sub>i</sub>( $\mathbf{z}$ ) = softmax<sub>i</sub>( $\mathbf{z} - \mathbf{b}$ ) for any  $1 \le i \le K$ .

3. For a vector  $\mathbf{z} \in \mathbb{R}^K,$  consider the softmax function that is scaled by a constant  $T \in \mathbb{R}:$ 

$$f_i(\mathbf{z}) = \frac{\exp(Tz_i)}{\sum_{k=1}^K \exp(Tz_k)}$$
 (2)

Further, for a vector  $\mathbf{z} = (z_1, \dots, z_K) \in \mathbb{R}^K$ , instead of considering the  $\arg \max(z_1, \dots, z_K)$  function as a function with categorical output  $1, \dots, K$  (corresponding to the index of a vector's largest element), consider the  $\arg \max$  function with **one-hot** representation of the output (assuming there is a unique maximum element):

$$\arg\max(z_1,\dots,z_K) = (y_1,\dots,y_K) = (0,\dots,0,1,0,\dots,0)$$
(3)

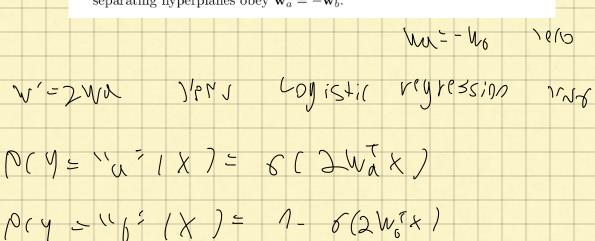
where  $y_i = 1$  if and only if  $i = \arg\max(z_1, \dots, z_K)$ , meaning that  $z_i$  is the unique maximum value of  $\mathbf{z} = (z_1, \dots, z_K)$ .

(a) For any vector  $\mathbf{z} \in \mathbb{R}^K$  whose maximum element is unique, show that the softmax converges to the arg max function as  $T \to \infty$ , i.e., prove that:

$$\lim_{T \to \infty} (f_1(\mathbf{z}), \dots, f_K(\mathbf{z})) = \arg \max(z_1, \dots, z_K)$$
 (4)

when arg max is in one-hot encoding.

- (b) For any vector  $\mathbf{z} \in \mathbb{R}^K$  whose maximum element is **not necessarily** unique, compute  $\lim_{T\to\infty}(f_1(\mathbf{z}),\ldots,f_K(\mathbf{z}))$  and provide a literal interpretation for your result.
- (c) For any vector  $\mathbf{z} \in \mathbb{R}^K$ , what happens when  $T \to 0$ ? Namely, compute the limit  $\lim_{T \to 0} (f_1(\mathbf{z}), \dots, f_K(\mathbf{z}))$ .
- 4. Write the gradient update rule for a logisitic regression model, when the usual loss of the negative log likelihood is now regularized with the square of the  $L_2$  norm over the weight vector  $\frac{1}{2}||\mathbf{w}||^2$ .
- 1. Prove that the logistic regression classifier is equivalent to a softmax over a linear multiclass classifier for two classes y = "a", y = "b", when their separating hyperplanes obey  $\mathbf{w}_a = -\mathbf{w}_b$ .



- SOFT MINX

