Loss function

$$f_{\mathbf{W}_1\mathbf{W}_2}(\mathbf{x}) = \mathbf{W}_2\,\sigma(\mathbf{W}_1\mathbf{x})$$

$$\mathscr{L}(\mathbf{x}, \mathbf{y}, \mathbf{W}_1, \mathbf{W}_2) = \left\| f_{\mathbf{W}_1\mathbf{W}_2}(\mathbf{x}) - \mathbf{y} \right\|^2$$

Stochastic gradient descent update

$$\mathbf{W}_1 \leftarrow \mathbf{W}_1 - \alpha \nabla_{\mathbf{W}_1} \mathscr{L}(\mathbf{x}, \mathbf{y}, \mathbf{W}_1, \mathbf{W}_2)$$

$$\mathbf{W}_2 \leftarrow \mathbf{W}_2 - \alpha \nabla_{\mathbf{W}_2} \mathscr{L}(\mathbf{x}, \mathbf{y}, \mathbf{W}_1, \mathbf{W}_2)$$

How do we calculate the gradients?

Approach

Training loss

$$\mathscr{L}(\mathbf{W}_1, \mathbf{W}_2) = \frac{1}{n} \sum_{i=1}^{n} \mathscr{L}(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}, \mathbf{W}_1, \mathbf{W}_2)$$

Objective

$$\hat{\mathbf{W}}_1, \hat{\mathbf{W}}_2 = \arg \min_{\mathbf{W}_1, \mathbf{W}_2} \mathscr{L}(\mathbf{W}_1, \mathbf{W}_2)$$

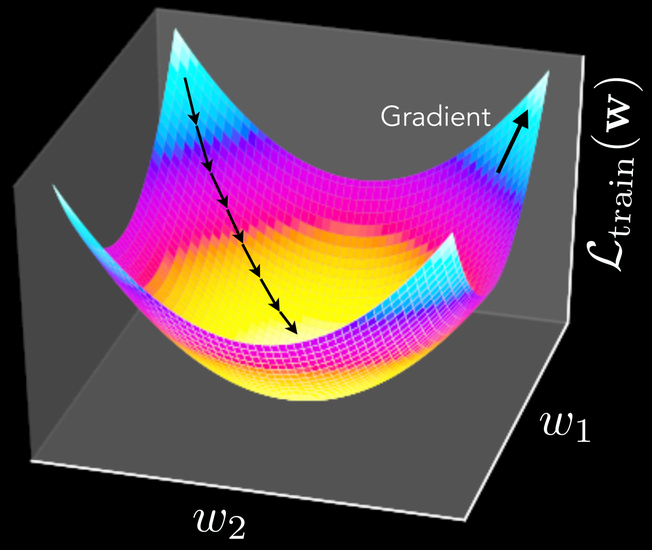
Optimal predictor

$$f_{\hat{\mathbf{W}}_1 \hat{\mathbf{W}}_2}(\mathbf{x}) = \hat{\mathbf{W}}_2 \sigma(\hat{\mathbf{W}}_1 \mathbf{x})$$

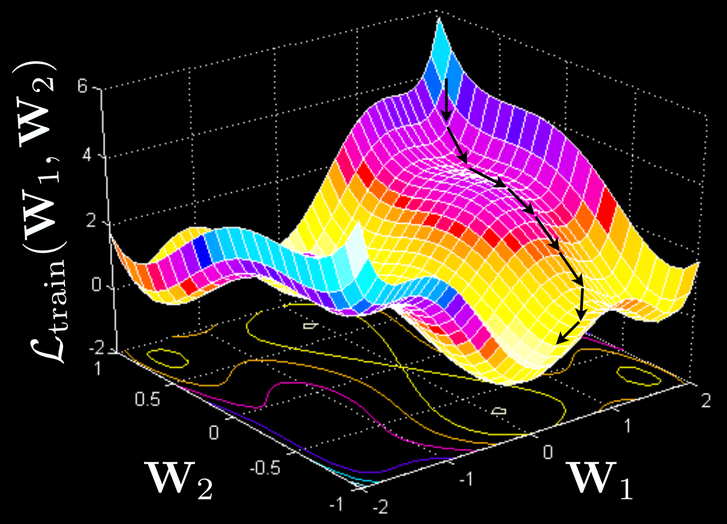
Non-convexity

$$\hat{\mathbf{W}}_1, \hat{\mathbf{W}}_2 = \arg\min_{\mathbf{W}_1, \mathbf{W}_2} \mathscr{L}(\mathbf{W}_1, \mathbf{W}_2)$$

Linear predictor loss



Neural network loss



Flatten

$\mathbf{x}$

$$\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{x} + \mathbf{b}^{[1]}$$
$$\mathbf{v}^{[1]} = \sigma(\mathbf{z}^{[1]})$$

$$\mathbf{z}^{[2]} = \mathbf{W}^{[2]}\mathbf{v}^{[1]} + \mathbf{b}^{[2]}$$
$$\mathbf{v}^{[2]} = \sigma(\mathbf{z}^{[2]})$$

$$\mathbf{z}^{[3]} = \mathbf{W}^{[3]}\mathbf{v}^{[2]} + \mathbf{b}^{[3]}$$
$$\mathbf{v}^{[3]} = \sigma(\mathbf{z}^{[3]})$$

$z_1^{[1]}$ $v_1^{[1]}$

$z_2^{[1]}$ $v_2^{[1]}$

$z_3^{[1]}$ $v_3^{[1]}$

$\mathbf{b}^{[1]}$

$z_1^{[2]}$ $v_1^{[2]}$

$z_2^{[2]}$ $v_1^{[2]}$

$\mathbf{b}^{[2]}$

$z_1^{[3]}$ $v_1^{[3]}$

$\mathbf{b}^{[3]}$

CAT

Hypothesis

$$\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{x} + \mathbf{b}^{[1]}$$

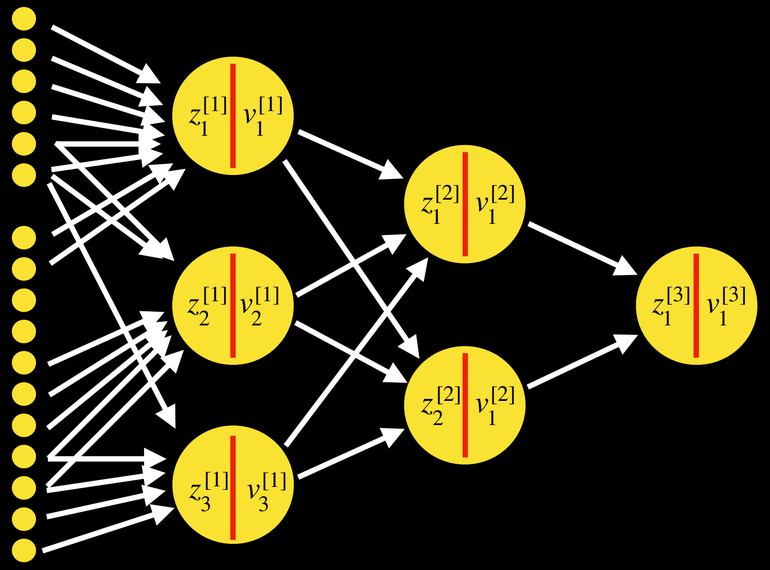
$$\mathbf{v}^{[1]} = \sigma(\mathbf{z}^{[1]})$$

$$\mathbf{z}^{[2]} = \mathbf{W}^{[2]}\mathbf{v}^{[1]} + \mathbf{b}^{[2]}$$

$$\mathbf{v}^{[2]} = \sigma(\mathbf{z}^{[2]})$$

$$\mathbf{z}^{[3]} = \mathbf{W}^{[3]}\mathbf{v}^{[2]} + \mathbf{b}^{[3]}$$

$$\mathbf{v}^{[3]} = \sigma(\mathbf{z}^{[3]})$$



$$\hat{y} \equiv \mathbf{v}^{[3]}$$

$$\hat{y} = f_{\mathbf{W}^{[1]}\mathbf{W}^{[2]}\mathbf{W}^{[3]}}(\mathbf{x})$$

Loss (Binary output): $$\mathscr{L}(\hat{y}, y) = -\sum_{i=1}^{n} y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})$$

Gradient Descent

Update the i^{th} layer:

$$\mathbf{W}^{[i]} = \mathbf{W}^{[i]} - \alpha \frac{\partial \mathscr{L}}{\partial \mathbf{W}^{[i]}}$$

$$\mathbf{b}^{[i]} = \mathbf{b}^{[i]} - \alpha \frac{\partial \mathscr{L}}{\partial \mathbf{b}^{[i]}}$$

What are the gradients?

Hypothesis

$$\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{x} + \mathbf{b}^{[1]}$$

$$\mathbf{v}^{[1]} = \sigma(\mathbf{z}^{[1]})$$

$$\mathbf{z}^{[2]} = \mathbf{W}^{[2]}\mathbf{v}^{[1]} + \mathbf{b}^{[2]}$$

$$\mathbf{v}^{[2]} = \sigma(\mathbf{z}^{[2]})$$

$$\mathbf{z}^{[3]} = \mathbf{W}^{[3]}\mathbf{v}^{[2]} + \mathbf{b}^{[3]}$$

$$\hat{y} = \sigma(\mathbf{z}^{[3]})$$

What are the gradients?

$$\mathscr{L}(\hat{y}, y) = -\sum_{i=1}^{n} y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})$$

$$\frac{\partial \mathscr{L}}{\partial \mathbf{W}^{[3]}} = \frac{\partial \mathscr{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{z}^{[3]}} \frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{W}^{[3]}}$$

$$\frac{\partial \mathscr{L}}{\partial \mathbf{W}^{[2]}} = \frac{\partial \mathscr{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{z}^{[3]}} \frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{v}^{[2]}} \frac{\partial \mathbf{v}^{[2]}}{\partial \mathbf{z}^{[2]}} \frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{W}^{[2]}}$$

$$\frac{\partial \mathscr{L}}{\partial \mathbf{b}^{[2]}} = \frac{\partial \mathscr{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{z}^{[3]}} \frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{v}^{[2]}} \frac{\partial \mathbf{v}^{[2]}}{\partial \mathbf{z}^{[2]}} \frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{b}^{[2]}}$$

Hypothesis

$$\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{x} + \mathbf{b}^{[1]}$$

$$\mathbf{v}^{[1]} = \sigma(\mathbf{z}^{[1]})$$

$$\mathbf{z}^{[2]} = \mathbf{W}^{[2]}\mathbf{v}^{[1]} + \mathbf{b}^{[2]}$$

$$\mathbf{v}^{[2]} = \sigma(\mathbf{z}^{[2]})$$

$$\mathbf{z}^{[3]} = \mathbf{W}^{[3]}\mathbf{v}^{[2]} + \mathbf{b}^{[3]}$$

$$\hat{y} = \sigma(\mathbf{z}^{[3]})$$

What are the gradients?

$$\mathscr{L}(\hat{y}, y) = -\sum_{i=1}^{n} y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})$$

$$\frac{\partial \mathscr{L}}{\partial \mathbf{W}^{[3]}} = \frac{\partial \mathscr{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{z}^{[3]}} \frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{W}^{[3]}}$$

$$\frac{\partial \mathscr{L}}{\partial \hat{y}} = -y^{(i)} \frac{1}{\hat{y}^{(i)}} + (1 - y^{(i)}) \frac{1}{1 - \hat{y}^{(i)}}$$

$$\frac{\partial \hat{y}}{\partial \mathbf{z}^{[3]}} = \sigma'(\mathbf{z}^{[3]}) = \sigma(\mathbf{z}^{[3]})(1 - \sigma(\mathbf{z}^{[3]}))$$

$$\frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{W}^{[3]}} = \mathbf{v}^{[2]^\top}$$

Hypothesis

$$\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{x} + \mathbf{b}^{[1]}$$

$$\mathbf{v}^{[1]} = \sigma(\mathbf{z}^{[1]})$$

$$\mathbf{z}^{[2]} = \mathbf{W}^{[2]}\mathbf{v}^{[1]} + \mathbf{b}^{[2]}$$

$$\mathbf{v}^{[2]} = \sigma(\mathbf{z}^{[2]})$$

$$\mathbf{z}^{[3]} = \mathbf{W}^{[3]}\mathbf{v}^{[2]} + \mathbf{b}^{[3]}$$

$$\hat{y} = \sigma(\mathbf{z}^{[3]})$$

$$\frac{\partial \mathscr{L}}{\partial \mathbf{W}^{[3]}} = \frac{\partial \mathscr{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{z}^{[3]}} \frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{W}^{[3]}}$$

$$\frac{\partial \mathscr{L}}{\partial \hat{y}} = -y^{(i)} \frac{1}{\hat{y}^{(i)}} + (1 - y^{(i)}) \frac{1}{1 - \hat{y}^{(i)}}$$

$$\frac{\partial \hat{y}}{\partial \mathbf{z}^{[3]}} = \sigma'(\mathbf{z}^{[3]}) = \sigma(\mathbf{z}^{[3]})(1 - \sigma(\mathbf{z}^{[3]}))$$

$$\frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{W}^{[3]}} = \mathbf{v}^{[2]\top}$$

$$\vdots$$

$$\frac{\partial \mathscr{L}}{\partial \mathbf{W}^{[3]}} = (y^{(i)} - \hat{y}^{(i)})\mathbf{v}^{[2]\top}$$

Hypothesis

$$\mathbf{z}^{[1]} = \mathbf{W}^{[1]}\mathbf{x} + \mathbf{b}^{[1]}$$

$$\mathbf{v}^{[1]} = \sigma(\mathbf{z}^{[1]})$$

$$\mathbf{z}^{[2]} = \mathbf{W}^{[2]}\mathbf{v}^{[1]} + \mathbf{b}^{[2]}$$

$$\mathbf{v}^{[2]} = \sigma(\mathbf{z}^{[2]})$$

$$\mathbf{z}^{[3]} = \mathbf{W}^{[3]}\mathbf{v}^{[2]} + \mathbf{b}^{[3]}$$

$$\hat{y} = \sigma(\mathbf{z}^{[3]})$$

SGD Update →

$$\mathbf{W}^{[3]} = \mathbf{W}^{[3]} - \alpha \frac{\partial \mathscr{L}}{\partial \mathbf{W}^{[3]}}$$

$$\frac{\partial \mathscr{L}}{\partial \mathbf{W}^{[3]}} = \frac{\partial \mathscr{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{z}^{[3]}} \frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{W}^{[3]}} = (y^{(i)} - \hat{y}^{(i)}) \mathbf{v}^{[2]\top}$$

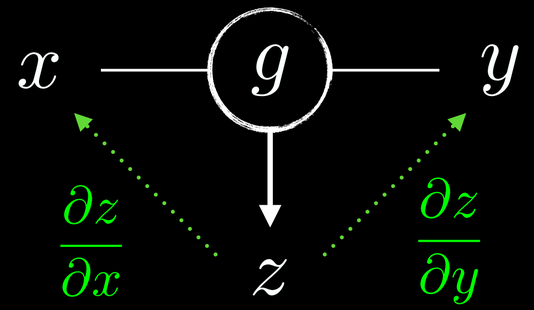
$$\frac{\partial \mathscr{L}}{\partial \mathbf{W}^{[2]}} = \frac{\partial \mathscr{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \mathbf{z}^{[3]}} \frac{\partial \mathbf{z}^{[3]}}{\partial \mathbf{v}^{[2]}} \frac{\partial \mathbf{v}^{[2]}}{\partial \mathbf{z}^{[2]}} \frac{\partial \mathbf{z}^{[2]}}{\partial \mathbf{W}^{[2]}}$$

common terms across gradients

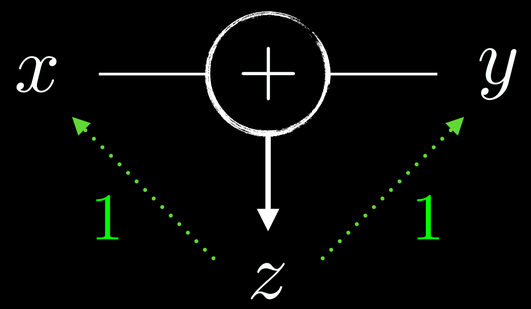
Can we save the gradients to be reused for computing other gradients?

Computation graphs

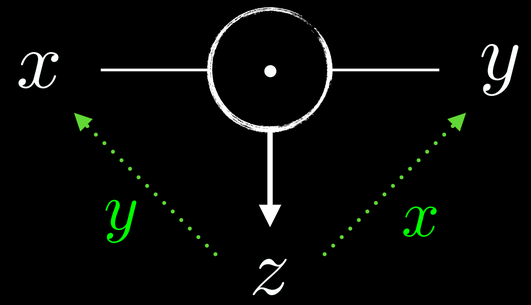
$z = g(x, y)$



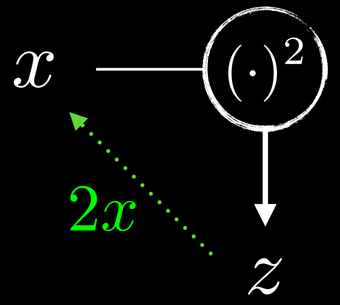
$z = x + y$



$z = xy$

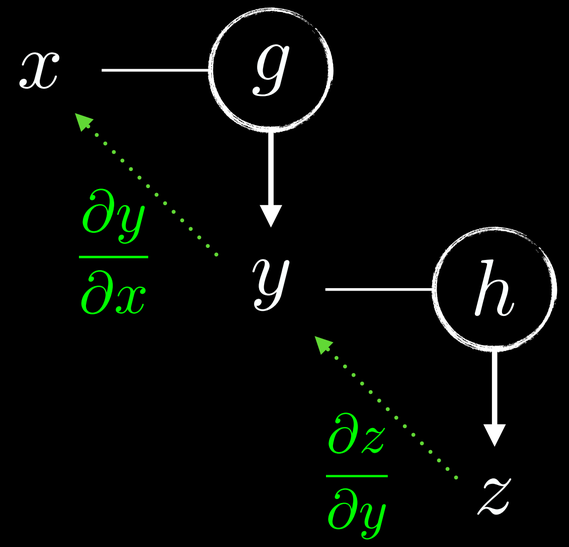


$z = x^2$



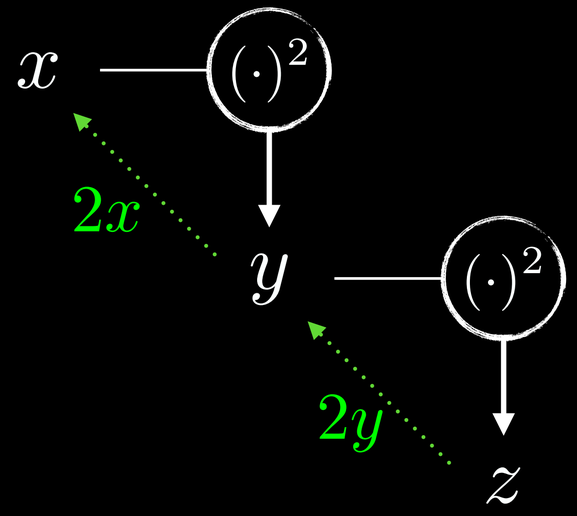
Chain rule

$$y = g(x)$$
$$z = h(y)$$



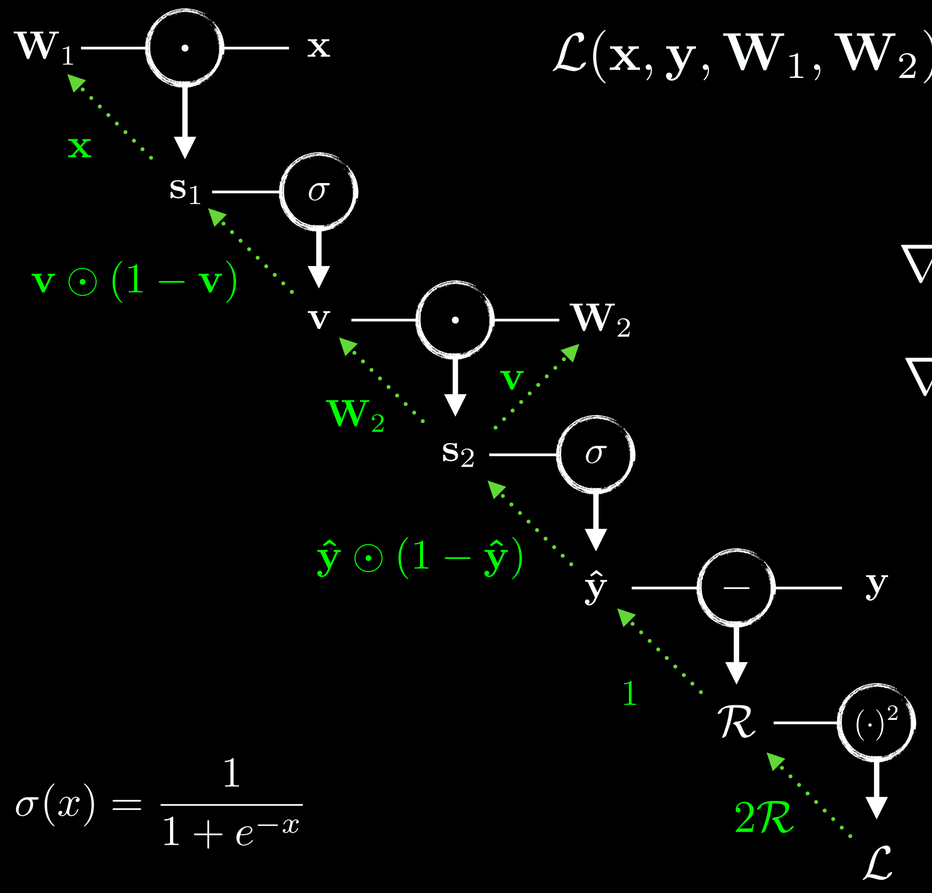
$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y}\frac{\partial y}{\partial x}$$

$$y = x^2$$
$$z = y^2$$



$$\frac{\partial z}{\partial x} = 4xy = 4x^3$$

Backpropagation



$$\mathcal{L}(\mathbf{x}, \mathbf{y}, \mathbf{W}_1, \mathbf{W}_2) = \|\sigma(\mathbf{W}_2\,\sigma(\mathbf{W}_1\mathbf{x})) - \mathbf{y}\|^2$$

$$\nabla_{\mathbf{W}_1}\mathcal{L} = 2\mathbf{W}_2^\top \mathcal{R} \odot \hat{\mathbf{y}} \odot (1 - \hat{\mathbf{y}}) \odot \mathbf{v} \odot (1 - \mathbf{v})\mathbf{x}^\top$$

$$\nabla_{\mathbf{W}_2}\mathcal{L} = 2\mathcal{R} \odot \hat{\mathbf{y}} \odot (1 - \hat{\mathbf{y}})\mathbf{v}^\top$$

assuming $\sigma(x) = \dfrac{1}{1 + e^{-x}}$