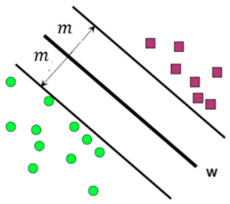
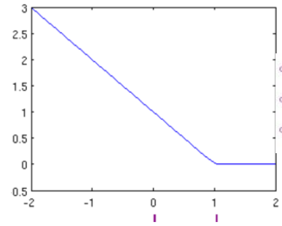
$$\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i$$

❑Define hinge loss or soft margin:
- $L_i(\boldsymbol{w}, b) = \max(0, 1 - y_i z_i)$

$\mathrm{m} = \frac{\gamma}{\|\boldsymbol{w}\|}$





$y_i z_i$

- $y_i z_i \geq 1 \Rightarrow$ Sample meets margin target,
- $y_i z_i \in [0,1) \Rightarrow$ Sample margin too small
- $y_i z_i \leq 0 \Rightarrow$ Sample misclassified

❑Define Lagrangian: $L(\boldsymbol{w}, \lambda) = f(\boldsymbol{w}) + \lambda g(\boldsymbol{w})$
- $\boldsymbol{w}$ is called the primal variable
- λ is called the dual variable

❑KKT Conditions: $\widehat{\boldsymbol{w}}, \hat{\lambda}$ satisfy:
- $\widehat{\boldsymbol{w}}$ minimizes the Lagrangian: $\widehat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} L(\boldsymbol{w}, \hat{\lambda})$
- Either
  - $g(\widehat{\boldsymbol{w}}) = 0$ and $\hat{\lambda} \geq 0$ [active constraint]
  - $g(\widehat{\boldsymbol{w}}) < 0$ and $\hat{\lambda} = 0$ [inactive constraint]

$$J(\boldsymbol{w}, b) = C \sum_{i=1}^{N} \max(0, 1 - y_i(\boldsymbol{w}^T \phi(\boldsymbol{x}_i) + b)) + \frac{1}{2}\|\boldsymbol{w}\|^2$$

$$\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \phi(\boldsymbol{x}_i)$$

❑Perfectly linearly separable if there exists a $\boldsymbol{\theta} = (b, w_1, \dots, w_d)$
- $b + w_1 x_{i1} + \cdots w_d x_{id} > \gamma$ when $y_i = 1$
- $b + w_1 x_{i1} + \cdots w_d x_{id} < -\gamma$ when $y_i = -1$

$y_i(b + w_1 x_{i1} + \cdots w_d x_{id}) > \gamma$

❑Support vectors: Samples that either:
- Are exactly on margin: $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) = 1$
- Or, on wrong side of margin: $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \leq 1$

$$z = b + \boldsymbol{w}^T \boldsymbol{x} = b + \sum_{i=1}^{N} \alpha_i y_i \underbrace{\phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x})}_{K(\boldsymbol{x}_i, \boldsymbol{x}) = \text{“kernel”}}$$

$$J(\boldsymbol{w}, b) = C \sum_{i=1}^{N} \max(0, 1 - y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b)) + \frac{1}{2}\|\boldsymbol{w}\|^2$$

❑Parameter ϵ_i called the slack variable
- $\epsilon_i = 0 \Rightarrow$ Sample on correct side of margin
- $\epsilon_i \geq 0 \Rightarrow$ Sample violates the margin
- $\epsilon_i \geq 1 \Rightarrow$ Sample misclassified (wrong side of hyperplane)

❑Parameter C:
- C large: Forces minimum number of violations. Highly fit to data. Low bias, higher variance
- C small: Enables more samples violations. Higher bias, lower variance

Hidden layer: $z_j^{\mathrm{H}} = \sum_{k=1}^{N_i} W_{jk}^{\mathrm{H}} x_k + b_j^{\mathrm{H}}, \quad u_j^{\mathrm{H}} = g_{\mathrm{act}}(z_j^{\mathrm{H}}), \quad j = 1, \dots, N_h$

Output layer: $z_j^{\mathrm{O}} = \sum_{k=1}^{N_h} W_{jk}^{\mathrm{O}} u_k^{\mathrm{H}} + b_j^{\mathrm{O}}, \quad u^{\mathrm{O}} = g_{\mathrm{out}}(\mathbf{z}^{\mathrm{O}}). \quad j = 1, \dots, N_o.$

In the hidden layer, the function $g_{\mathrm{act}}(z)$ is called the *activation function*.

Hard threshold:

$$g_{\mathrm{act}}(z) = \begin{cases} 1, & \text{if } z \geq 0 \\ 0, & \text{if } z < 0. \end{cases}$$

Sigmoid: $g_{\mathrm{act}}(z) = 1/(1 + e^{-z})$.

Rectified linear unit (ReLU): $g(z) = \max\{0, z\}$.

output map $g_{\mathrm{out}}(z)$

$$P(y = 1|\mathbf{x}) = u^{\mathrm{O}} = g_{\mathrm{out}}(z^{\mathrm{O}}) = \frac{1}{1 + e^{-z^{\mathrm{O}}}}.$$

$$P(y = k|\mathbf{x}) = u_k^{\mathrm{O}} = g_{\mathrm{out,k}}(z^{\mathrm{O}}) = \frac{e^{z_k^{\mathrm{O}}}}{\sum_{\ell=1}^{K} e^{z_\ell^{\mathrm{O}}}}.$$

$$\hat{\mathbf{y}} = \mathbf{u}^{\mathrm{O}} = g_{\mathrm{out}}(\mathbf{z}^{\mathrm{O}}) = \mathbf{z}^{\mathrm{O}}.$$

Gradient tensors: Suppose that $\mathbf{Y} = f(\mathbf{X})$ where:

– The input $\mathbf{X}$ is a tensor of size $(N_1, \dots, N_r)$,

The output $\mathbf{Y}$ is a tensor of size $(M_1, \dots, M_s)$.

$$\left[\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}\right]_{i_1,\dots,i_s,j_1,\dots,j_r} = \frac{\partial Y_{i_1,\dots,i_r}}{\partial X_{j_1,\dots,j_s}}.$$

The input $\mathbf{X}$ is a tensor of size $(K_1, \dots, K_t)$;

The intermediate variable $\mathbf{Y} = g(\mathbf{X})$ is a tensor of size $(N_1, \dots, N_r)$;

The output $\mathbf{Z} = h(\mathbf{Y})$ is a tensor of size $(M_1, \dots, M_s)$.

$$\frac{\partial Z_{i_1,\dots,i_r}}{\partial X_{k_1,\dots,k_t}} = \sum_{j_1=0}^{N_1-1} \cdots \sum_{j_r=0}^{N_r-1} \frac{\partial Z_{i_1,\dots,i_r}}{\partial Y_{j_1,\dots,j_s}} \frac{\partial Y_{j_1,\dots,j_s}}{\partial X_{k_1,\dots,k_t}}.$$

$$\mathbf{u} = f(\mathbf{z}) = (f(z_1), \dots, f(z_N)), \quad f(z_i) = \frac{1}{1 + e^{-z_i}},$$

$$\frac{\partial \mathbf{u}}{\partial \mathbf{z}} = \begin{bmatrix} f'(z_1) & 0 & \cdots & 0 \\ 0 & f'(z_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f'(z_N) \end{bmatrix}$$

$$\frac{\partial u_i}{\partial W_{ij}} = \frac{\partial u_i}{\partial z_i} \frac{\partial z_i}{\partial W_{ij}} = f'(z_i) x_j,$$

$$\frac{\partial u_i}{\partial b_i} = \frac{\partial u_i}{\partial z_i} \frac{\partial z_i}{\partial b_i} = f'(z_i).$$

$$\frac{\partial u_i}{\partial x_j} = \frac{\partial u_i}{\partial z_i} \frac{\partial z_i}{\partial x_j} = f'(z_i) W_{ij}.$$

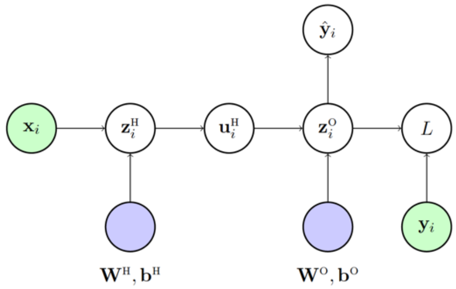
The loss function factor $g_{\mathrm{loss}}(z_i^{\mathrm{O}}, y_i)$

$$g_{\mathrm{loss}}(z_i^{\mathrm{O}}, y_i) = \ln\left[1 + e^{z_i^{\mathrm{O}}}\right] - y_i z_i.$$

$$g_{\mathrm{loss}}(\mathbf{z}_i^{\mathrm{O}}, y_i) := \ln\left[\sum_{\ell=1}^{K} e^{z_{i\ell}^{\mathrm{O}}}\right] - \sum_{k=1}^{K} r_{ik} z_{ik}^{\mathrm{O}},$$

$$r_{ik} = \begin{cases} 1 & \text{if } y_i = k, \\ 0 & \text{if } y_i \neq k. \end{cases}$$

$$g_{\mathrm{loss}}(\mathbf{z}_i^{\mathrm{O}}, \mathbf{y}_i) = \|\mathbf{z}_i^{\mathrm{O}} - \mathbf{y}_i\|^2 = \sum_{k=1}^{d} (z_{ik}^{\mathrm{O}} - y_{ik})^2.$$



$$\mathbf{z}_i^{\mathrm{H}} = \mathbf{W}^{\mathrm{H}}\mathbf{x}_i + \mathbf{b}^{\mathrm{H}}, \quad \mathbf{u}_i^{\mathrm{H}} = g_{\mathrm{act}}(\mathbf{z}_i^{\mathrm{H}}),$$
$$\mathbf{z}_i^{\mathrm{O}} = \mathbf{W}^{\mathrm{O}}\mathbf{u}_i^{\mathrm{H}} + \mathbf{b}^{\mathrm{O}}, \quad \hat{y}_i = g_{\mathrm{out}}(\mathbf{z}_i^{\mathrm{O}}).$$
$$L = \sum_{i=1}^{N} g_{\mathrm{loss}}(\mathbf{z}_i^{\mathrm{O}}, y_i).$$

$$u_{ij}^{\mathrm{H}} = g_{\mathrm{act}}(z_{ij}^{\mathrm{H}}) = \max(0, z_{ij}^{\mathrm{H}}).$$

$$\frac{\partial u_{ij}^{\mathrm{H}}}{\partial z_{ij}^{\mathrm{H}}} = \begin{cases} 1, & \text{if } z_{ij}^{\mathrm{H}} > 0 \\ 0, & \text{if } z_{ij}^{\mathrm{H}} < 0. \end{cases}$$

$$\frac{\partial L}{\partial z_{ij}^{\mathrm{H}}} = \frac{\partial L}{\partial u_{ij}^{\mathrm{H}}} \frac{\partial u_{ij}^{\mathrm{H}}}{\partial z_{ij}^{\mathrm{H}}}.$$

$$z_{ij}^{\mathrm{H}} = \sum_k W_{jk}^{\mathrm{H}} x_{ik} + b_j^{\mathrm{H}}.$$

$$\frac{\partial L}{\partial z_{ij}^{\mathrm{O}}} = \frac{\partial g_{\mathrm{loss}}(\mathbf{z}_i^{\mathrm{O}}, y_i)}{\partial z_{ij}^{\mathrm{O}}} = \frac{e^{z_{ij}^{\mathrm{O}}}}{\sum_{\ell=1}^{K} e^{z_{i\ell}^{\mathrm{O}}}} - r_{ij}.$$

$$z_{ij}^{\mathrm{O}} = \sum_k W_{jk}^{\mathrm{O}} u_{ik}^{\mathrm{H}} + b_j^{\mathrm{O}}.$$

$$\frac{\partial z_{ij}^{\mathrm{O}}}{\partial W_{jk}^{\mathrm{O}}} = u_{ik}^{\mathrm{H}}, \quad \frac{\partial z_{ij}^{\mathrm{O}}}{\partial b_j^{\mathrm{O}}} = 1, \quad \frac{\partial z_{ij}^{\mathrm{O}}}{\partial u_{ik}^{\mathrm{H}}} = W_{jk}.$$

$$\frac{\partial L}{\partial W_{jk}^{\mathrm{O}}} = \sum_{i=1}^{N} \frac{\partial L}{\partial z_{ik}^{\mathrm{O}}} \frac{\partial z_{ij}^{\mathrm{O}}}{\partial W_{jk}^{\mathrm{O}}} = \sum_{i=1}^{N} \frac{\partial L}{\partial z_{ik}^{\mathrm{O}}} u_{ik}^{\mathrm{H}},$$
$$\frac{\partial L}{\partial b_j^{\mathrm{O}}} = \sum_{i=1}^{N} \frac{\partial L}{\partial z_{ik}^{\mathrm{O}}} \frac{\partial z_{ij}^{\mathrm{O}}}{\partial b_j^{\mathrm{O}}} = \sum_{i=1}^{N} \frac{\partial L}{\partial z_{ik}^{\mathrm{O}}},$$
$$\frac{\partial L}{\partial u_{ik}^{\mathrm{H}}} = \sum_{j=1}^{N_o} \frac{\partial L}{\partial z_{ij}^{\mathrm{O}}} \frac{\partial z_{ij}^{\mathrm{O}}}{\partial u_{ik}^{\mathrm{H}}} \sum_{j=1}^{N_o} \frac{\partial L}{\partial z_{ij}^{\mathrm{O}}} W_{jk}.$$

$$\frac{\partial z_{ij}^{\mathrm{H}}}{\partial W_{jk}^{\mathrm{H}}} = x_{ik}, \quad \frac{\partial z_{ij}^{\mathrm{H}}}{\partial b_j^{\mathrm{H}}} = 1.$$

$$\frac{\partial L}{\partial W_{jk}^{\mathrm{H}}} = \sum_{i=1}^{N} \frac{\partial L}{\partial z_{ik}^{\mathrm{H}}} \frac{\partial z_{ij}^{\mathrm{H}}}{\partial W_{jk}^{\mathrm{H}}} = \sum_{i=1}^{N} \frac{\partial L}{\partial z_{ik}^{\mathrm{H}}} x_{ik},$$
$$\frac{\partial L}{\partial b_{jk}^{\mathrm{H}}} = \sum_{i=1}^{N} \frac{\partial L}{\partial z_{ik}^{\mathrm{H}}} \frac{\partial z_{ij}^{\mathrm{H}}}{\partial b_j^{\mathrm{H}}} = \sum_{i=1}^{N} \frac{\partial L}{\partial z_{ik}^{\mathrm{H}}}.$$

- Suppose inputs are
  - x, size $N_1 \times N_2$, w: size $K_1 \times K_2$, $K_1 \le N_1$, $K_2 \le N_2$
  - $z = x * w$ (without reversal)

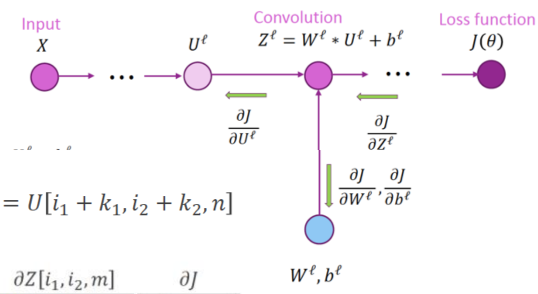
$$z[n_1, n_2] = \sum_{k_2=0}^{K_2-1} \sum_{k_1=0}^{K_1-1} w[k_1, k_2] x[n_1 + k_1, n_2 + k_2]$$

- Valid mode: $0 \le n_1 < N_1 - K_1 + 1$, $0 \le n_2 < N_2 - K_2 + 1$
  - Requires no zero padding
- Same mode: Output size $N_1 \times N_2$
  - Usually use zero padding for neural networks

- Weight and bias:
  - W: Weight tensor, size $(K_1, K_2, N_{in}, N_{out})$
  - b: Bias vector, size N_{out}

$$Z[i_1, i_2, m] = \sum_{k_1=0}^{K_1-1} \sum_{k_2=0}^{K_2-1} \sum_{n=0}^{N_{in}-1} W[k_1, k_2, n, m] X[i_1 + k_1, i_2 + k_2, n] + b[m]$$

$$Z^\ell = W^\ell * U^\ell + b^\ell$$



$$\frac{\partial Z[i_1, i_2, m]}{\partial W[k_1, k_2, n, m]} = U[i_1 + k_1, i_2 + k_2, n]$$

$$\frac{\partial J}{\partial W[k_1, k_2, n, m]} = \sum_{i_1=1}^{N_1} \sum_{i_2=1}^{N_2} \frac{\partial Z[i_1, i_2, m]}{\partial W[k_1, k_2, n, m]} \frac{\partial J}{\partial Z[i_1, i_2, m]}$$
$$= \sum_{i_1=1}^{N_1} \sum_{i_2=1}^{N_2} U[i_1 + k_1, i_2 + k_2, n] \frac{\partial J}{\partial Z[i_1, i_2, m]}$$