SVMs for Structured Output

Andrea Vedaldi

SVM*struct*

Tsochantaridis Hofmann Joachims Altun 04

Extending SVMs

Extending SVMs

- SVM = parametric function

$$f(\mathbf{x}|w) = \text{sign}\left(\langle w, \Psi(\mathbf{x})\rangle + b\right)$$

  - ***arbitrary input***
  - ***binary output***

Extending SVMs

- SVM = parametric function

$$f(\mathbf{x}|w) = \text{sign}\left(\langle w, \Psi(\mathbf{x})\rangle + b\right)$$

  - ***arbitrary input***
  - ***binary output***
- Regression by convex optimization

Extending SVMs

- SVM = parametric function

$$f(\mathbf{x}|w) = \text{sign}\left(\langle w, \Psi(\mathbf{x})\rangle + b\right)$$

  - ***arbitrary input***
  - ***binary output***
- Regression by convex optimization

Today:

➡ how to handle ***arbitrary output***

➡ (while keeping the computation efficient)

Arbitrary input *and* output

Arbitrary input *and* output

- SVM = sign of the projection $f(\mathbf{x}|w) = \text{sign}\langle w, \Psi(\mathbf{x})\rangle$

Arbitrary input *and* output

- SVM = sign of the projection $f(\mathbf{x}|w) = \text{sign}\langle w, \Psi(\mathbf{x})\rangle$
- Equivalently: *"output that matches best"*

$$F(\mathbf{x}, y|w) = y\langle w, \Psi(\mathbf{x})\rangle = \langle w, \Psi(\mathbf{x})y\rangle$$

$$f(\mathbf{x}|w) = \underset{y\in\{-1,+1\}}{\text{argmax}} \; F(\mathbf{x}, y|w)$$

Arbitrary input *and* output

- SVM = sign of the projection $f(\mathbf{x}|w) = \text{sign}\langle w, \Psi(\mathbf{x})\rangle$
- Equivalently: *"output that matches best"*

$$F(\mathbf{x}, y|w) = y\langle w, \Psi(\mathbf{x})\rangle = \langle w, \Psi(\mathbf{x})y\rangle$$

$$f(\mathbf{x}|w) = \underset{y\in\{-1,+1\}}{\text{argmax}}\ F(\mathbf{x}, y|w)$$

- Extend to ***arbitrary input*** and ***output***

$$F(\mathbf{x}, \mathbf{y}|w) = \langle w, \Psi(\mathbf{x}, \mathbf{y})\rangle$$

$$f(\mathbf{x}; w) = \underset{\mathbf{y}\in\mathcal{Y}}{\text{argmax}}\, F(\mathbf{x}, \mathbf{y}|w)$$

Arbitrary input *and* output

- SVM = sign of the projection $f(\mathbf{x}|w) = \text{sign}\langle w, \Psi(\mathbf{x})\rangle$
- Equivalently: *"output that matches best"*

$$F(\mathbf{x}, y|w) = y\langle w, \Psi(\mathbf{x})\rangle = \langle w, \Psi(\mathbf{x})y\rangle$$

$$f(\mathbf{x}|w) = \underset{y\in\{-1,+1\}}{\text{argmax}}\ F(\mathbf{x}, y|w)$$

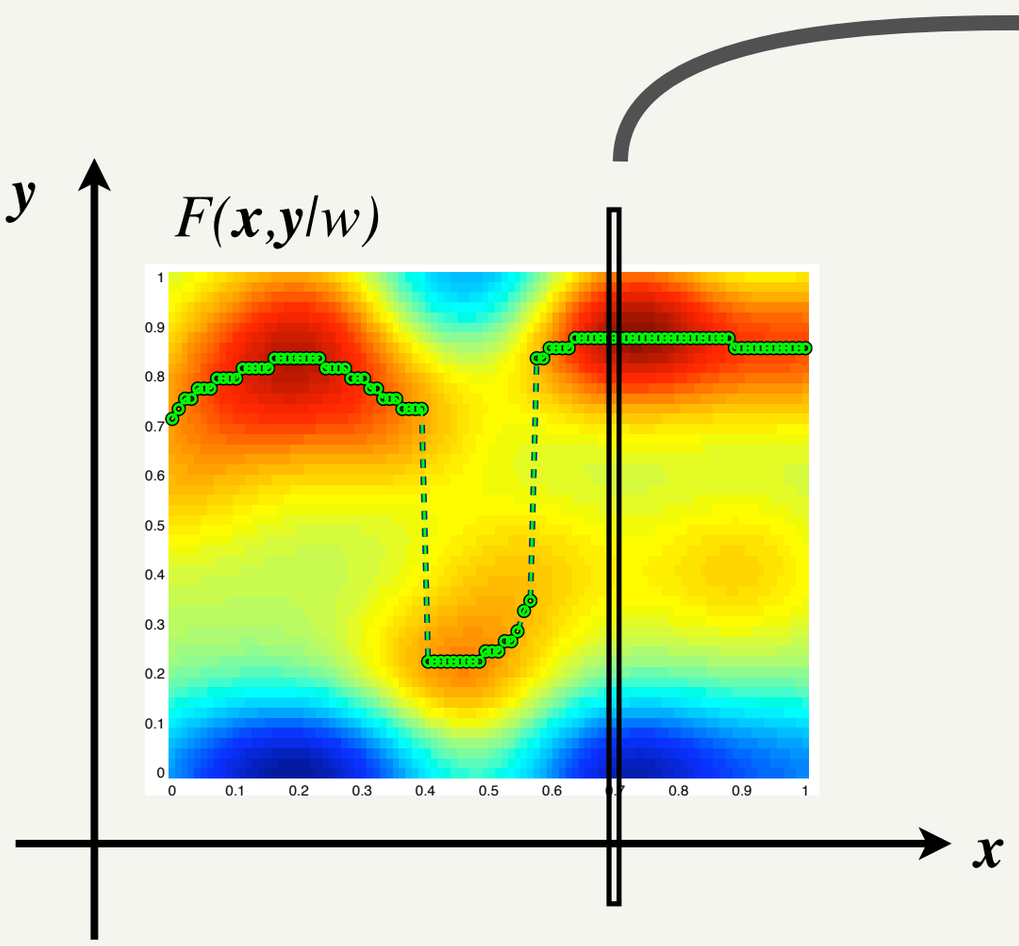
- Extend to ***arbitrary input*** and ***output***

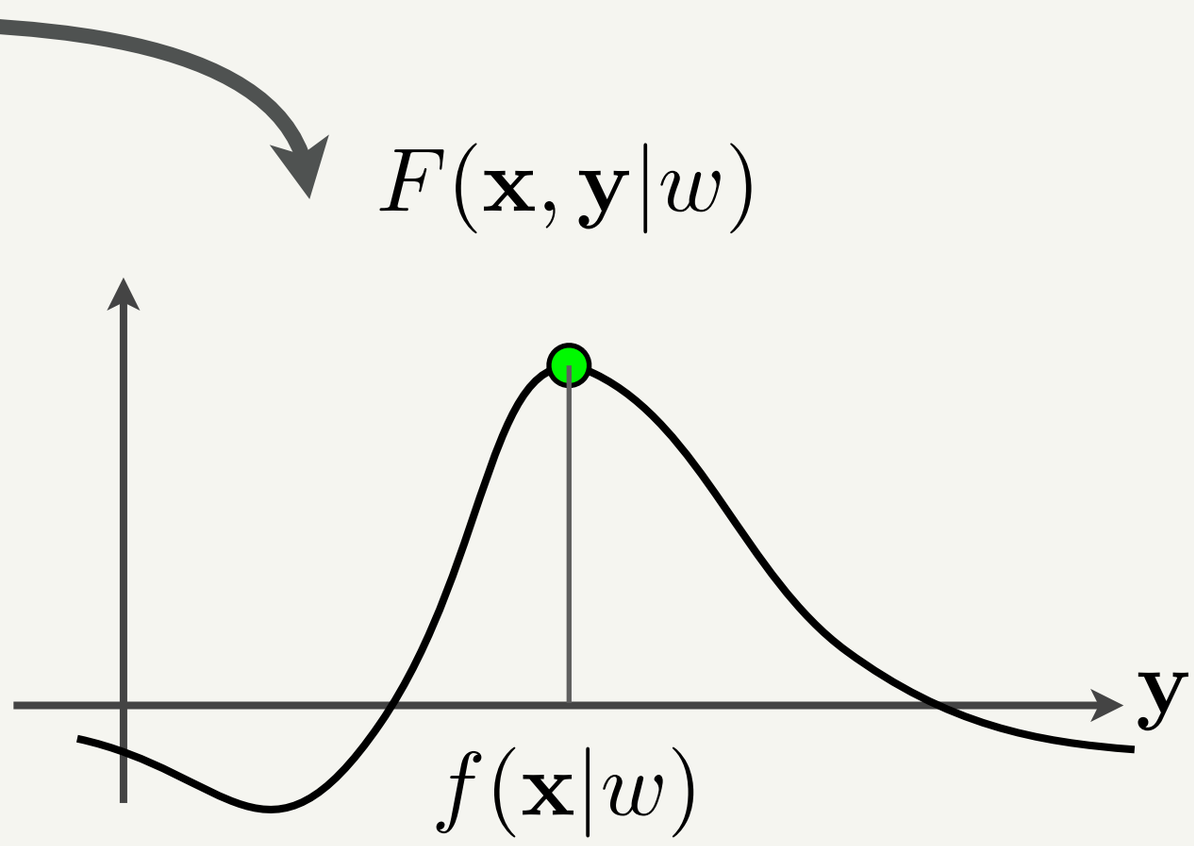
$$F(\mathbf{x}, \mathbf{y}|w) = \langle w, \Psi(\mathbf{x}, \mathbf{y})\rangle$$

$$f(\mathbf{x}; w) = \underset{\mathbf{y}\in\mathcal{Y}}{\text{argmax}}\ F(\mathbf{x}, \mathbf{y}|w)$$

Example: real function

- Encode $f(\mathbf{x}|w) : \mathbb{R} \rightarrow \mathbb{R}$
- $F(\mathbf{x}, \mathbf{y}|w) : \mathbb{R}^2 \rightarrow \mathbb{R}$





Regression

Regression problem

Regression problem

- SVM = function $y = f(x/w)$ parametrized in *w.*

Regression problem

- SVM = function $y = f(x/w)$ parametrized in w.
- **Goal.** Fit the function to data (x_1, y_1), ..., (x_N, y_N).

Regression problem

- SVM = function $y = f(x/w)$ parametrized in *w.*
- **Goal.** Fit the function to data (x_1, y_1), ..., (x_N, y_N).
- Formulated as *optimization problem:*
  - ***Loss***

$$\Delta(\mathbf{y}_i, \widehat{\mathbf{y}}_i) \geq 0, \quad \Delta(\mathbf{y}_i, \widehat{\mathbf{y}}_i) = 0 \Leftrightarrow \mathbf{y}_i = \widehat{\mathbf{y}}_i$$

Regression problem

- SVM = function $y = f(x|w)$ parametrized in w.
- **Goal.** Fit the function to data (x_1, y_1), ..., (x_N, y_N).
- Formulated as *optimization problem:*
  - ***Loss***

$$\Delta(\mathbf{y}_i, \widehat{\mathbf{y}}_i) \geq 0, \quad \Delta(\mathbf{y}_i, \widehat{\mathbf{y}}_i) = 0 \Leftrightarrow \mathbf{y}_i = \widehat{\mathbf{y}}_i$$

  - ***Risk*** (empirical)

$$R(w) = \frac{1}{N} \sum_{i=1}^{N} \Delta(\mathbf{y}_i, f(\mathbf{x}_i|w))$$

Regression problem

- SVM = function $y = f(x|w)$ parametrized in w.
- **Goal.** Fit the function to data (x_1, y_1), ..., (x_N, y_N).
- Formulated as *optimization problem:*
  - ***Loss***

$$\Delta(\mathbf{y}_i, \widehat{\mathbf{y}}_i) \geq 0, \quad \Delta(\mathbf{y}_i, \widehat{\mathbf{y}}_i) = 0 \Leftrightarrow \mathbf{y}_i = \widehat{\mathbf{y}}_i$$

  - ***Risk*** (empirical)

$$R(w) = \frac{1}{N} \sum_{i=1}^{N} \Delta(\mathbf{y}_i, f(\mathbf{x}_i|w))$$

- **Problem:** Find w that minimizes $R(w)$.

Separable case

Separable case

- Separable case ≡ exact fit ≡ *R(w) = 0*

$$R(w) = 0 \quad \Leftrightarrow \quad \forall i : \mathbf{y}_i = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} F(\mathbf{x}_i, \mathbf{y} | w)$$

Separable case

- Separable case ≡ exact fit ≡ $R(w) = 0$

$$R(w) = 0 \quad \Leftrightarrow \quad \forall i : \mathbf{y}_i = \operatorname*{argmax}_{y \in \mathcal{Y}} F(\mathbf{x}_i, \mathbf{y}|w)$$

- Necessary and sufficient condition:
  For each point $\mathbf{x}_i$ the maximum is reached at $\mathbf{y}_i$

$$\langle w, \delta\Psi_i(\mathbf{y}) \rangle = F(\mathbf{x}_i, \mathbf{y}_i|w) - F(\mathbf{x}_i, \mathbf{y}|w)$$

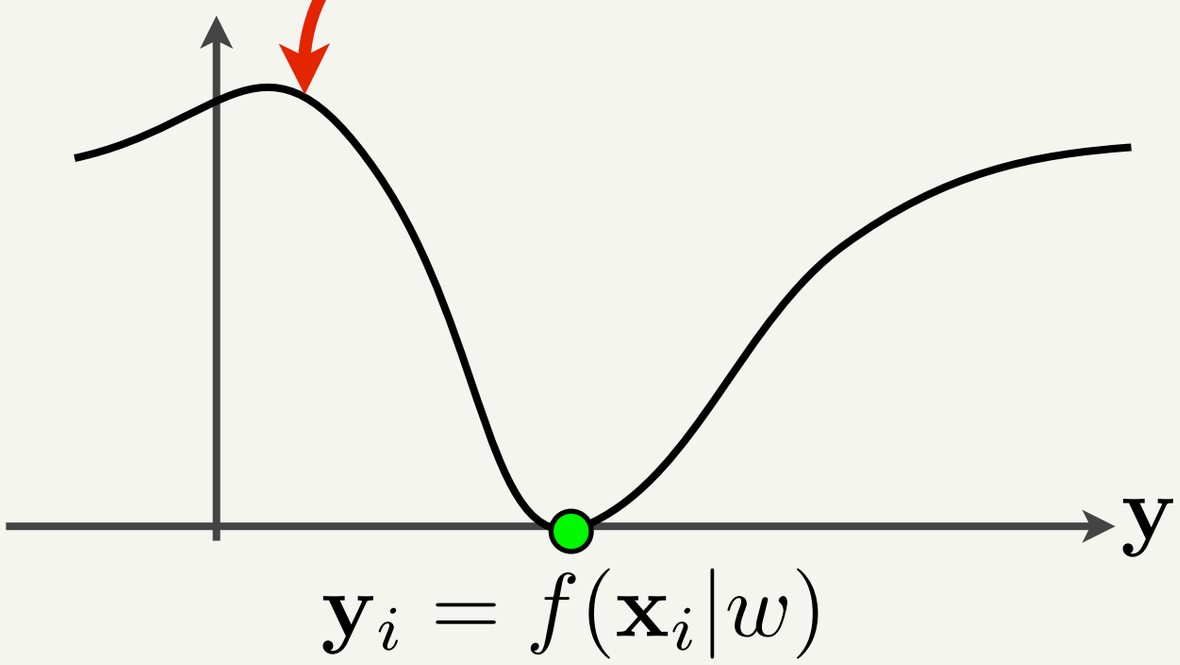
Separable case

- Separable case ≡ exact fit ≡ $R(w) = 0$

$$R(w) = 0 \quad \Leftrightarrow \quad \forall i : \mathbf{y}_i = \operatorname*{argmax}_{y \in \mathcal{Y}} F(\mathbf{x}_i, \mathbf{y}|w)$$

- Necessary and sufficient condition:
  For each point $\mathbf{x}_i$ the maximum is reached at $\mathbf{y}_i$

$$\langle w, \delta\Psi_i(\mathbf{y}) \rangle = F(\mathbf{x}_i, \mathbf{y}_i|w) - F(\mathbf{x}_i, \mathbf{y}|w)$$



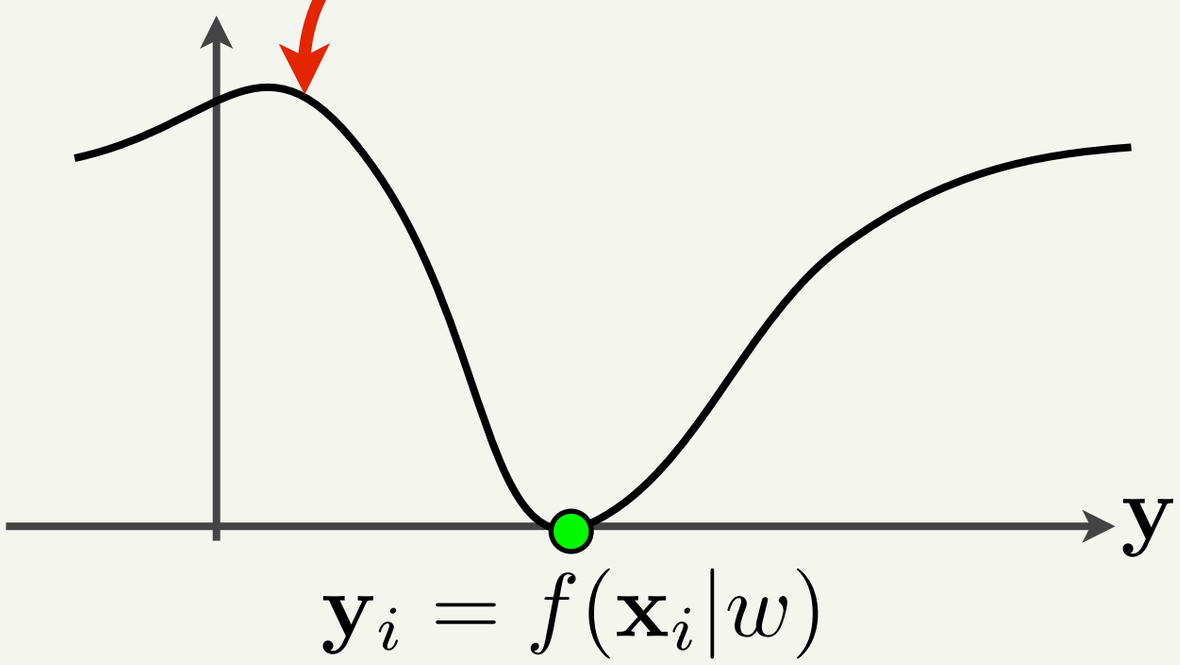
Separable case

- Separable case ≡ exact fit ≡ $R(w) = 0$

$$R(w) = 0 \quad \Leftrightarrow \quad \forall i : \mathbf{y}_i = \operatorname*{argmax}_{y \in \mathcal{Y}} F(\mathbf{x}_i, \mathbf{y}|w)$$

- Necessary and sufficient condition:
  For each point $\mathbf{x}_i$ the maximum is reached at $\mathbf{y}_i$

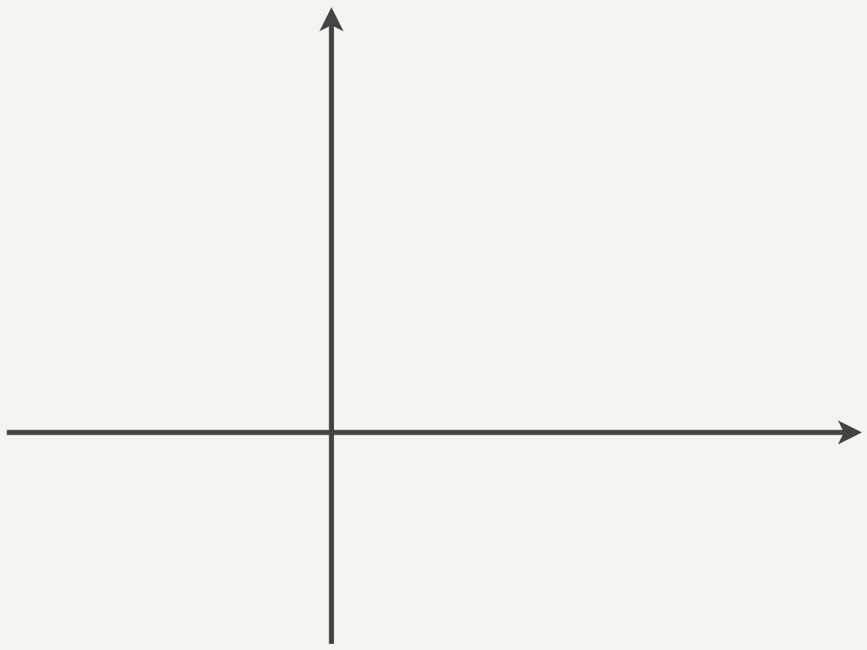
$$\langle w, \delta\Psi_i(\mathbf{y}) \rangle = F(\mathbf{x}_i, \mathbf{y}_i|w) - F(\mathbf{x}_i, \mathbf{y}|w)$$



$$\forall i, \mathbf{y} \neq \mathbf{y}_i : \ \langle w, \delta\Psi_i(\mathbf{y}) \rangle > 0$$

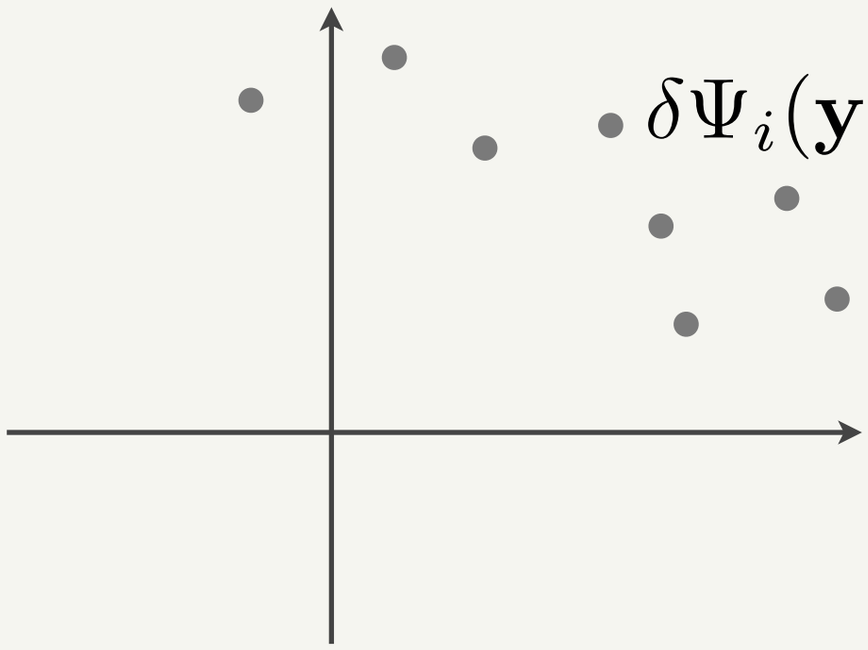
Margin → max-margin

$$\langle w, \delta\Psi_i(\mathbf{y}) \rangle > 0$$



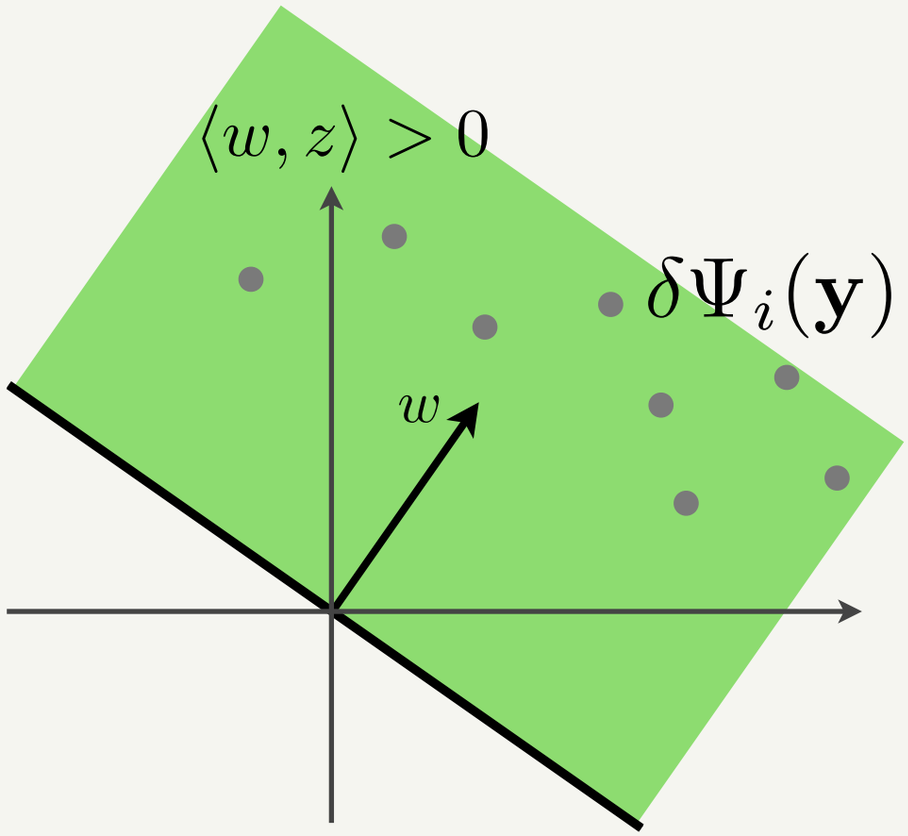
Margin → max-margin

$$\langle w, \delta\Psi_i(\mathbf{y}) \rangle > 0$$



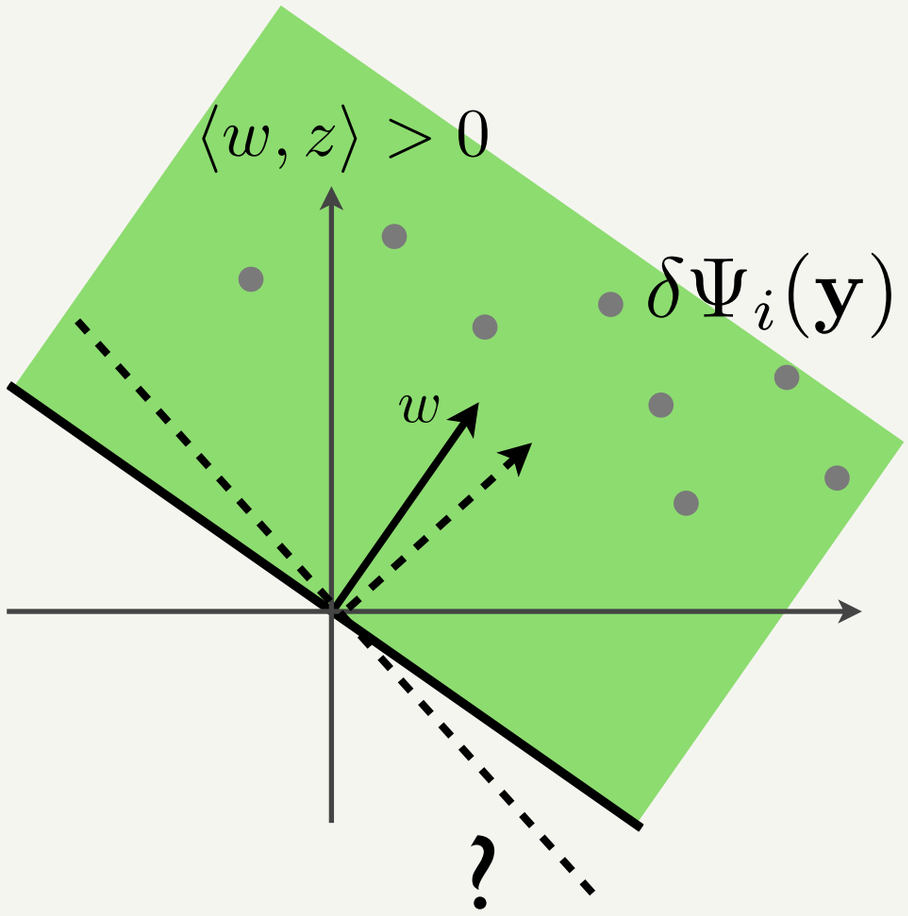
Margin → max-margin

$$\langle w, \delta\Psi_i(\mathbf{y}) \rangle > 0$$



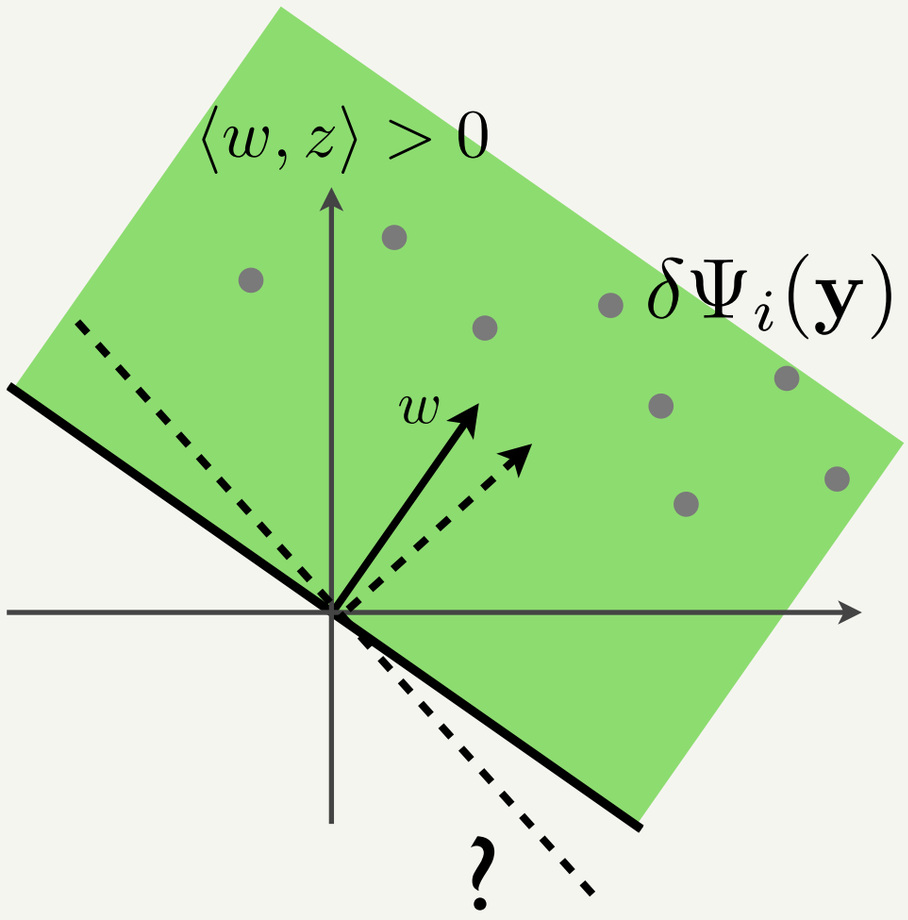
Margin → max-margin

$$\langle w, \delta\Psi_i(\mathbf{y}) \rangle > 0$$

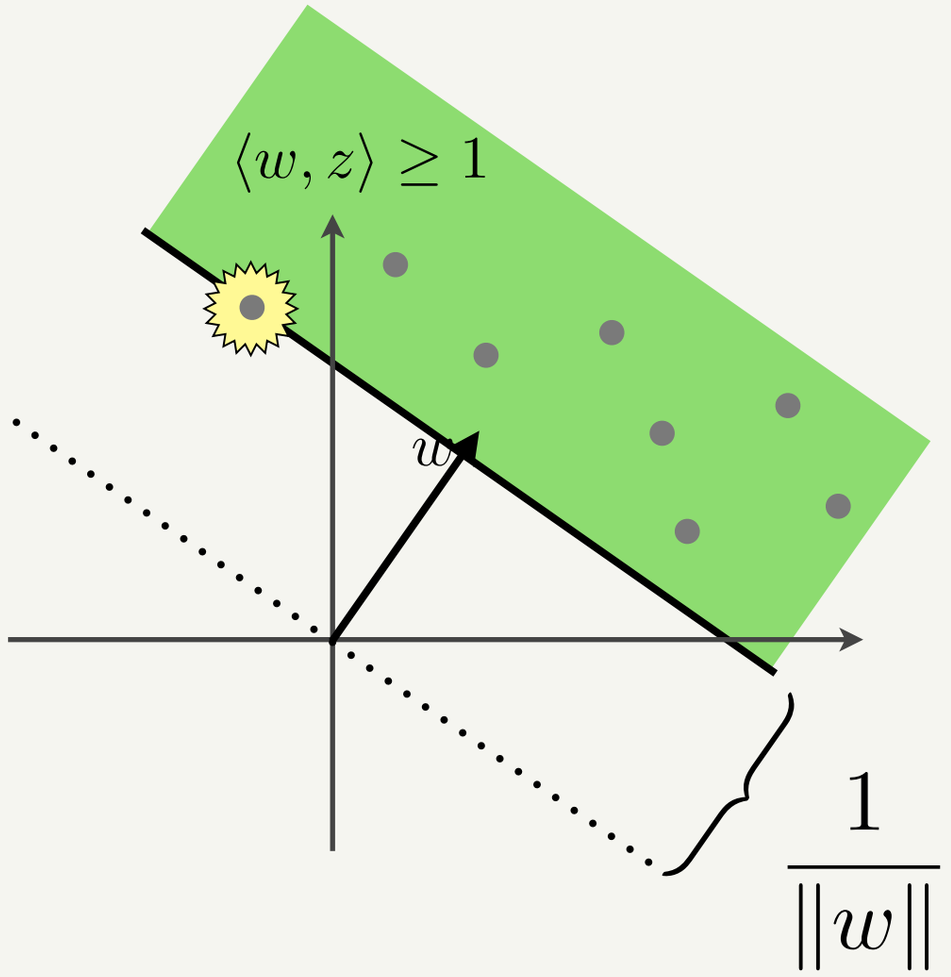


Margin → max-margin

$$\langle w, \delta\Psi_i(\mathbf{y}) \rangle > 0$$

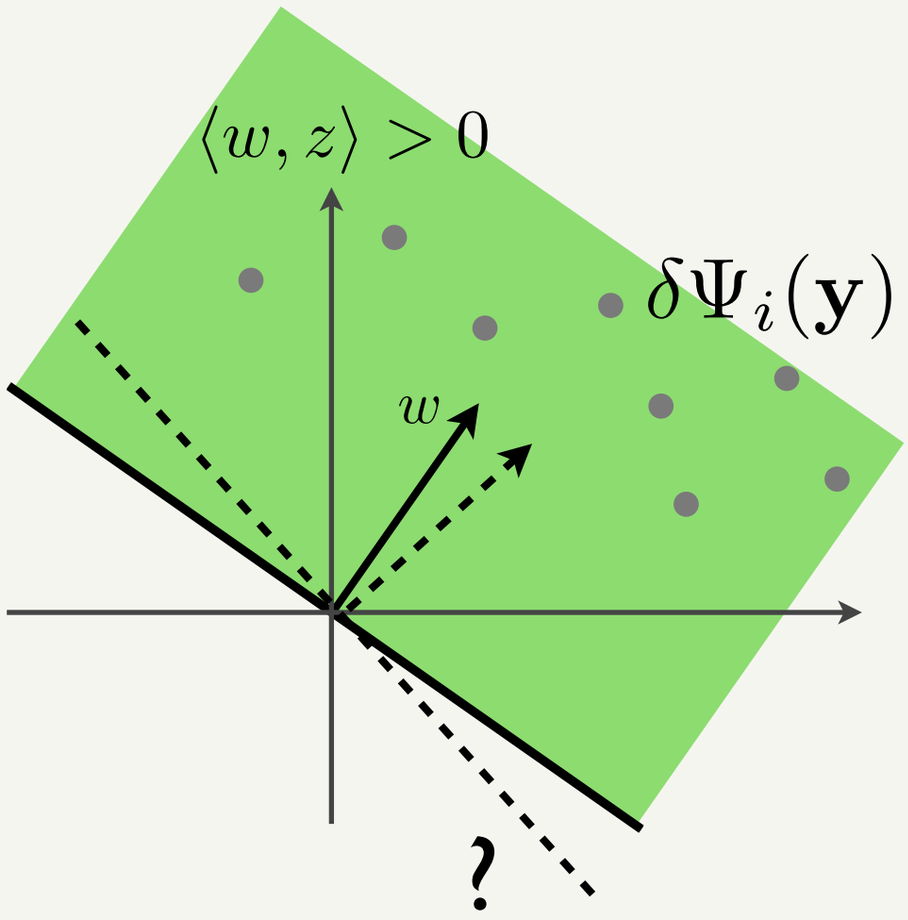


$$\langle w, \delta\Psi_i(\mathbf{y}) \rangle \geq 1$$

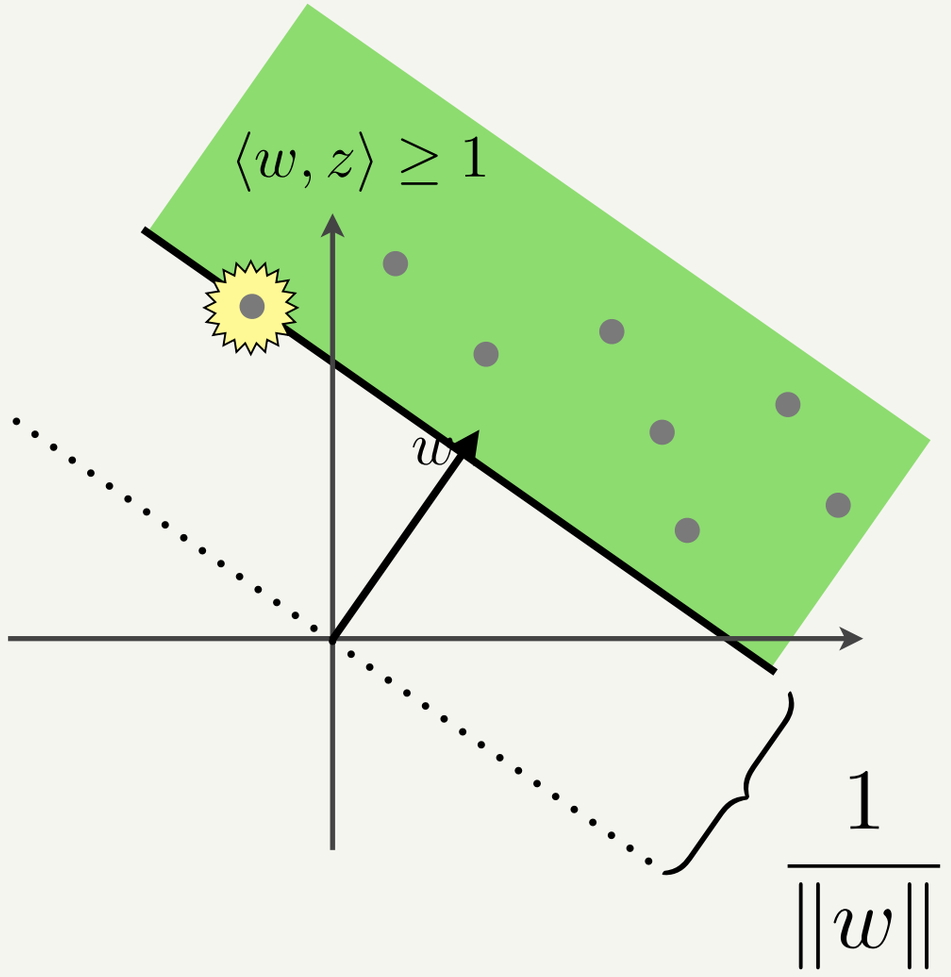


Margin → max-margin

$$\langle w, \delta\Psi_i(\mathbf{y}) \rangle > 0$$

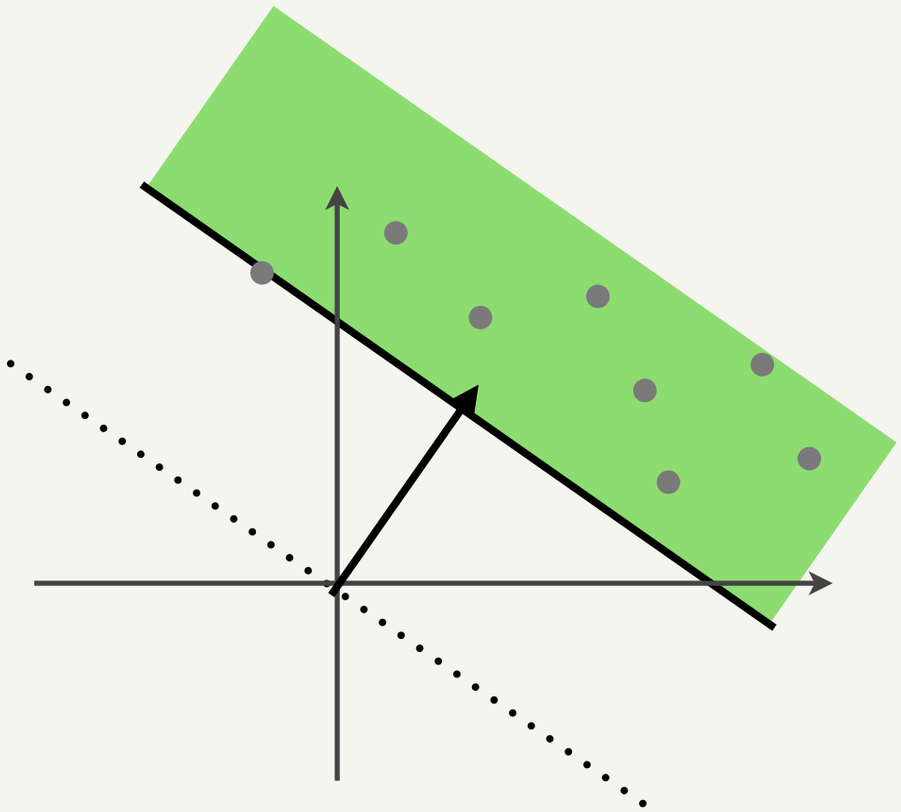


$$\langle w, \delta\Psi_i(\mathbf{y}) \rangle \geq 1$$

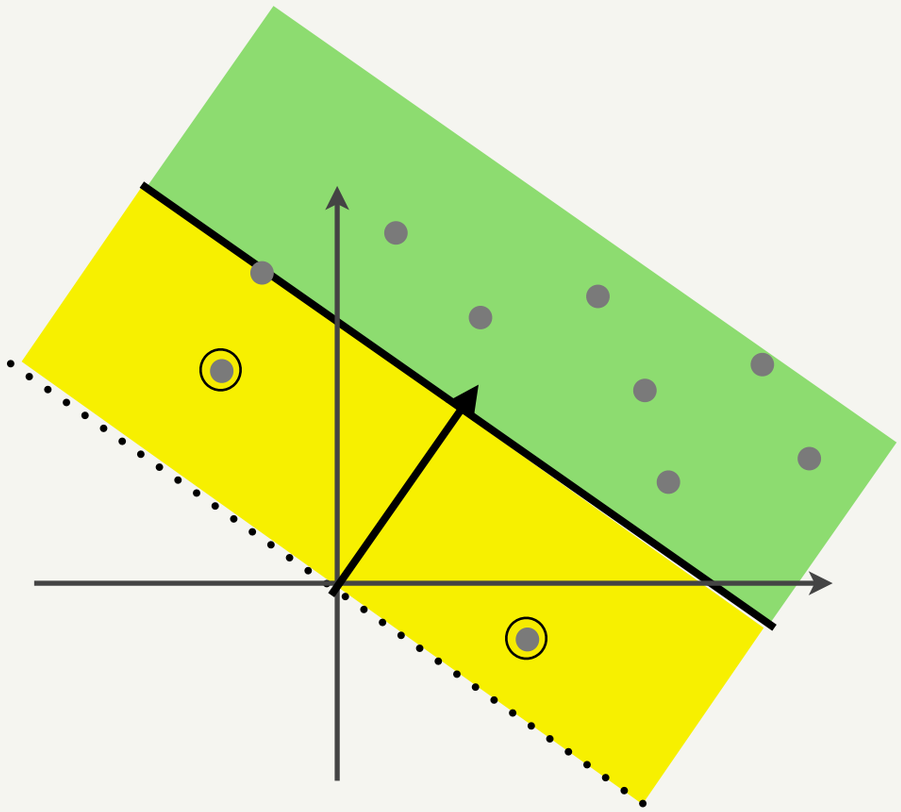


$$\min \tfrac{1}{2}\|w\|^2,$$
$$\langle w, \delta\Psi_i(\mathbf{y}) \rangle \geq 1 \quad \forall i, \mathbf{y} \neq \mathbf{y}_i$$

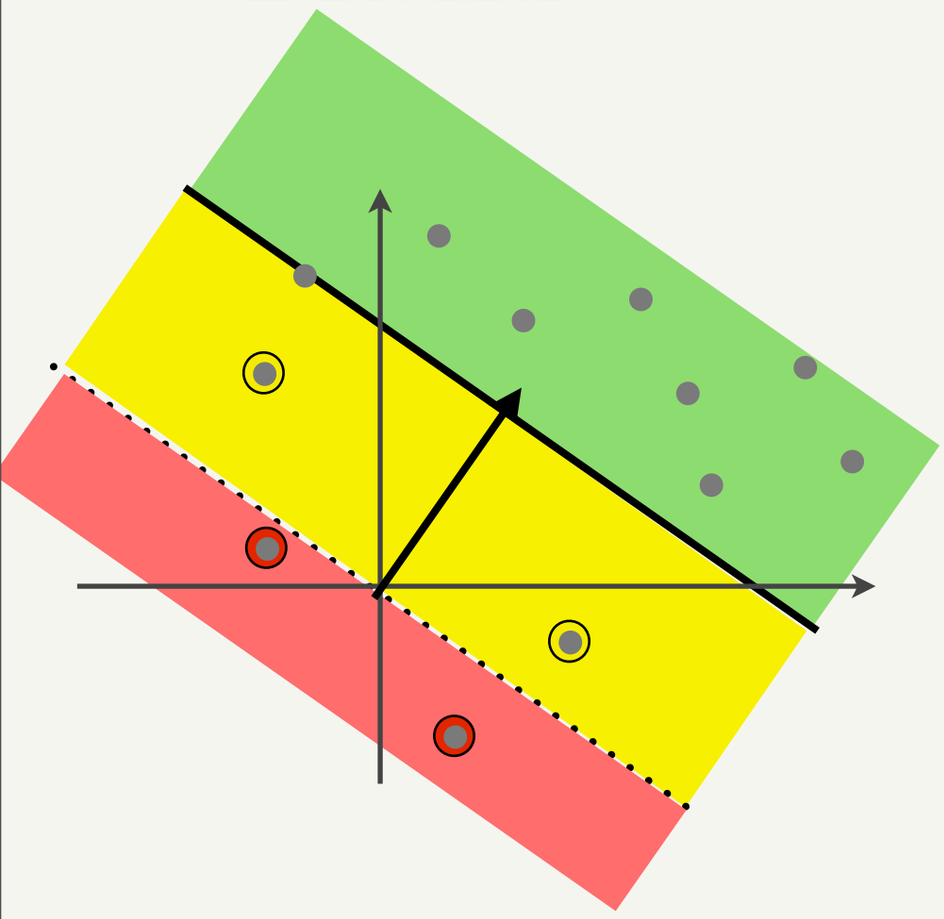
Non-separable case



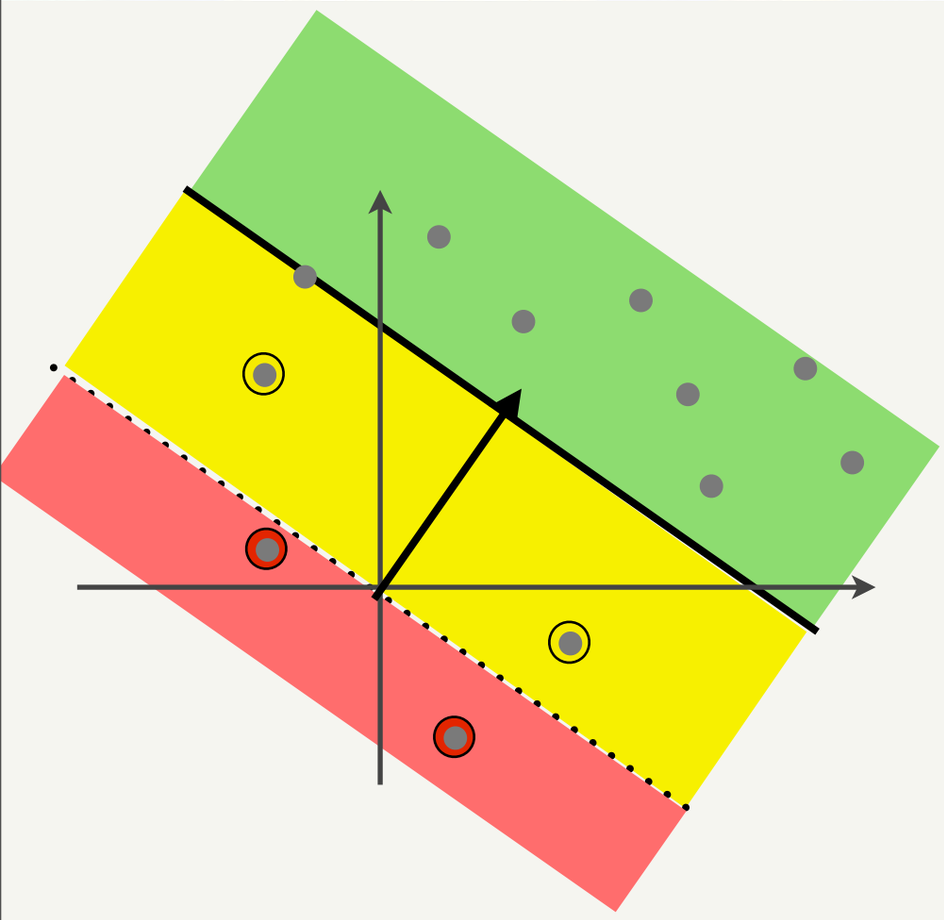
Non-separable case

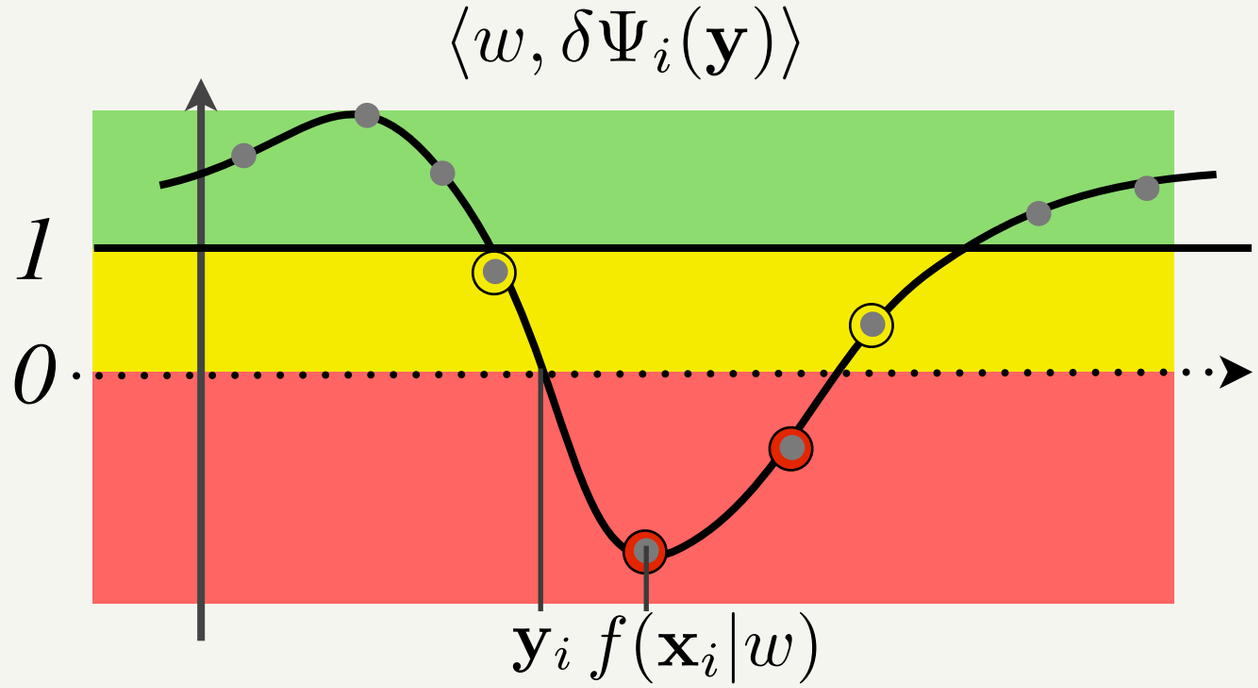


Non-separable case

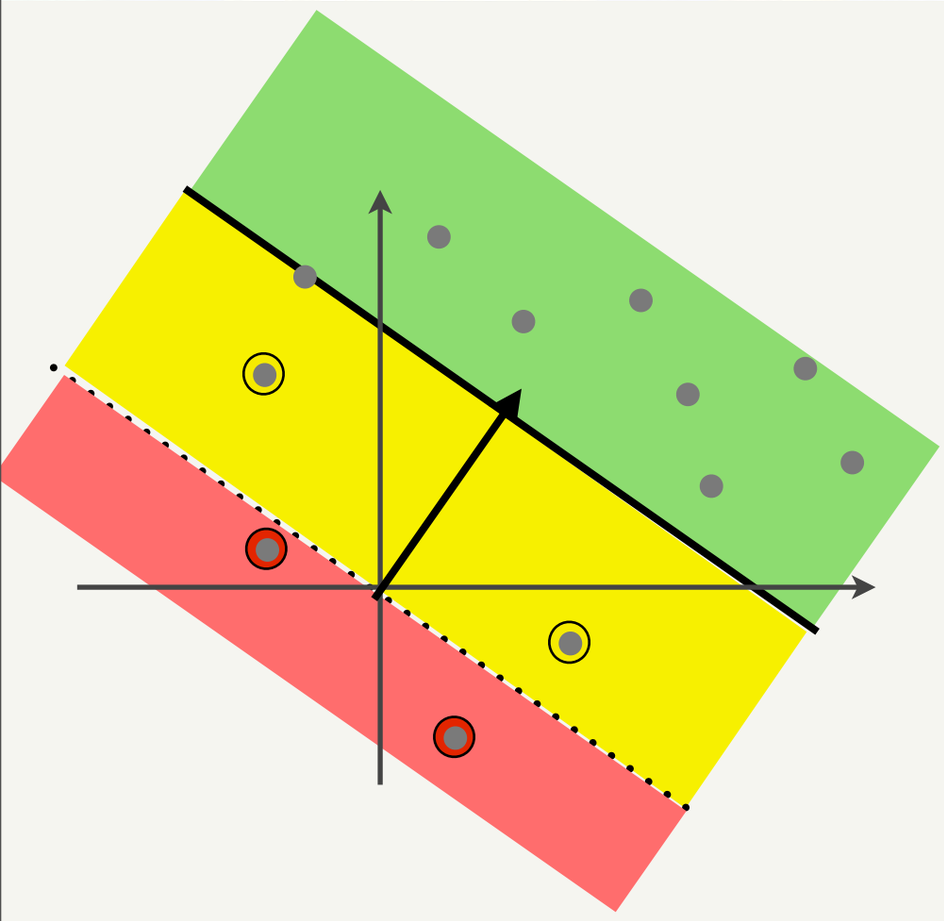


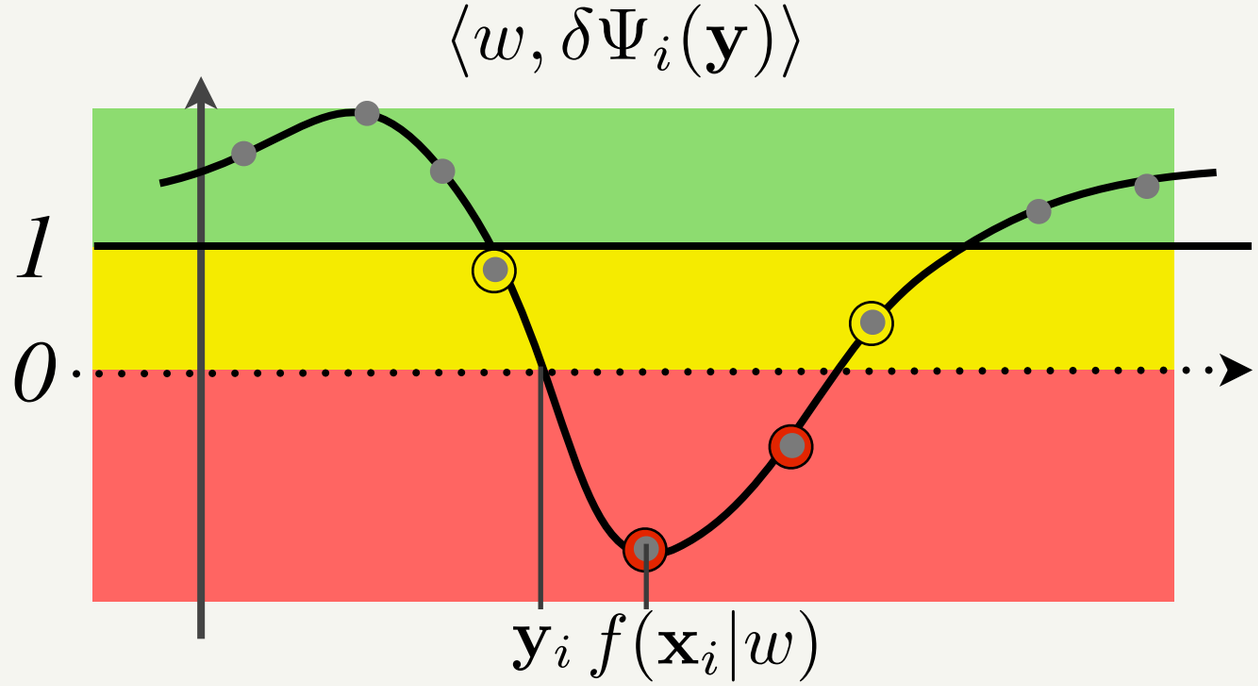
Non-separable case





Non-separable case

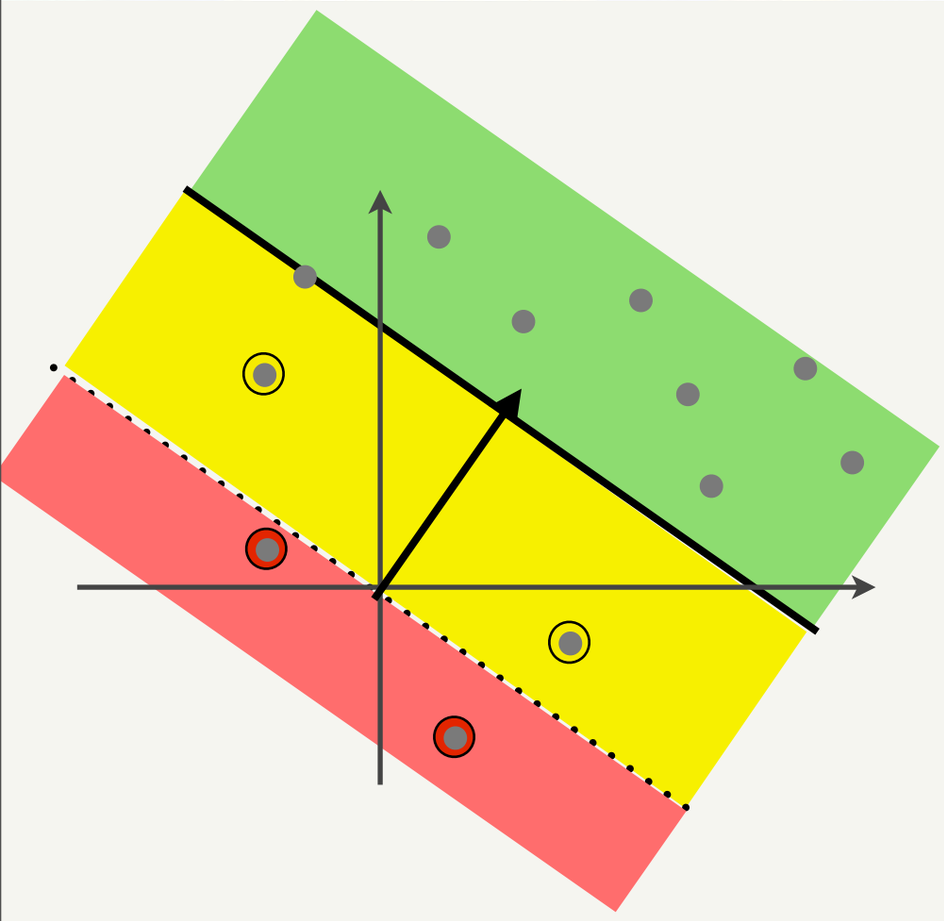


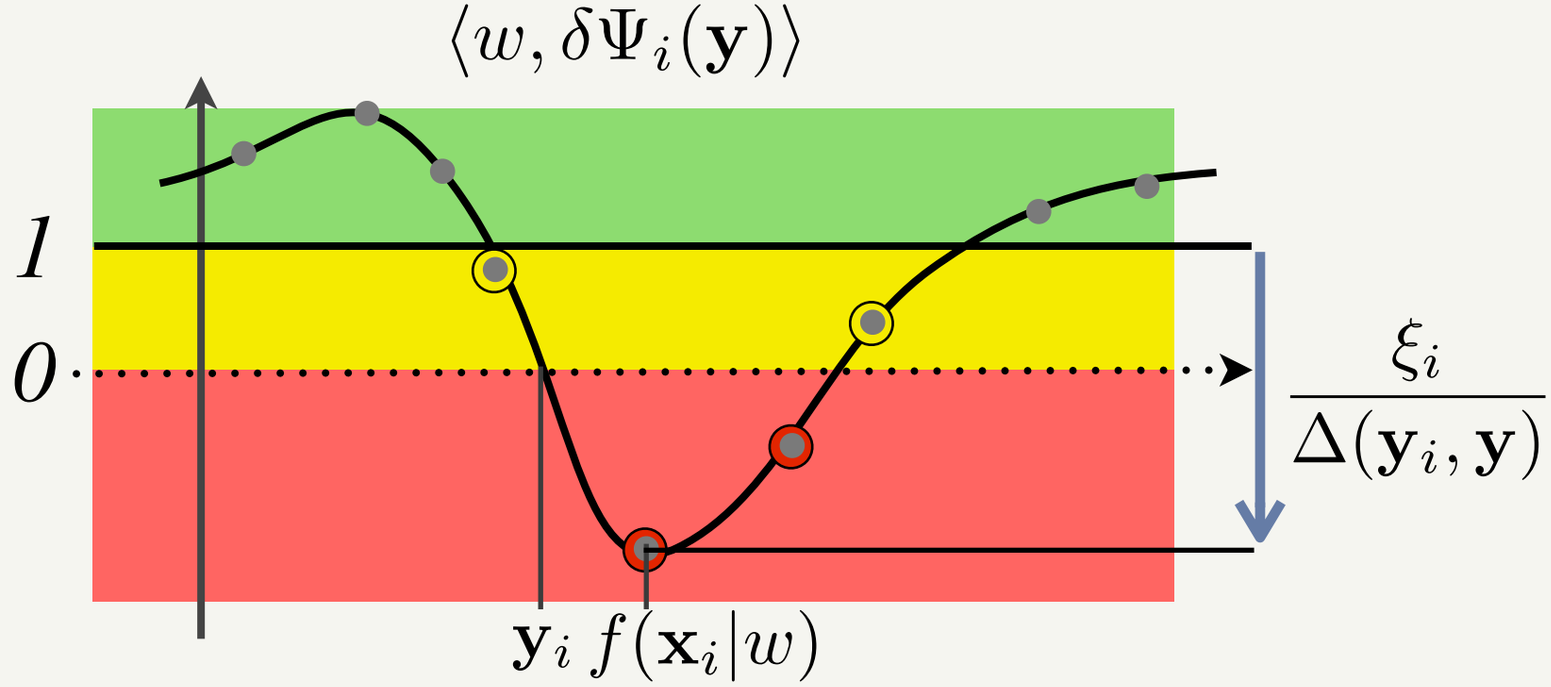


$$\langle w, \Psi_i(\mathbf{y}) \rangle \geq 1 - \xi_i / \Delta(\mathbf{y}_i, \mathbf{y})$$

$$\xi_i \geq 0$$

Non-separable case

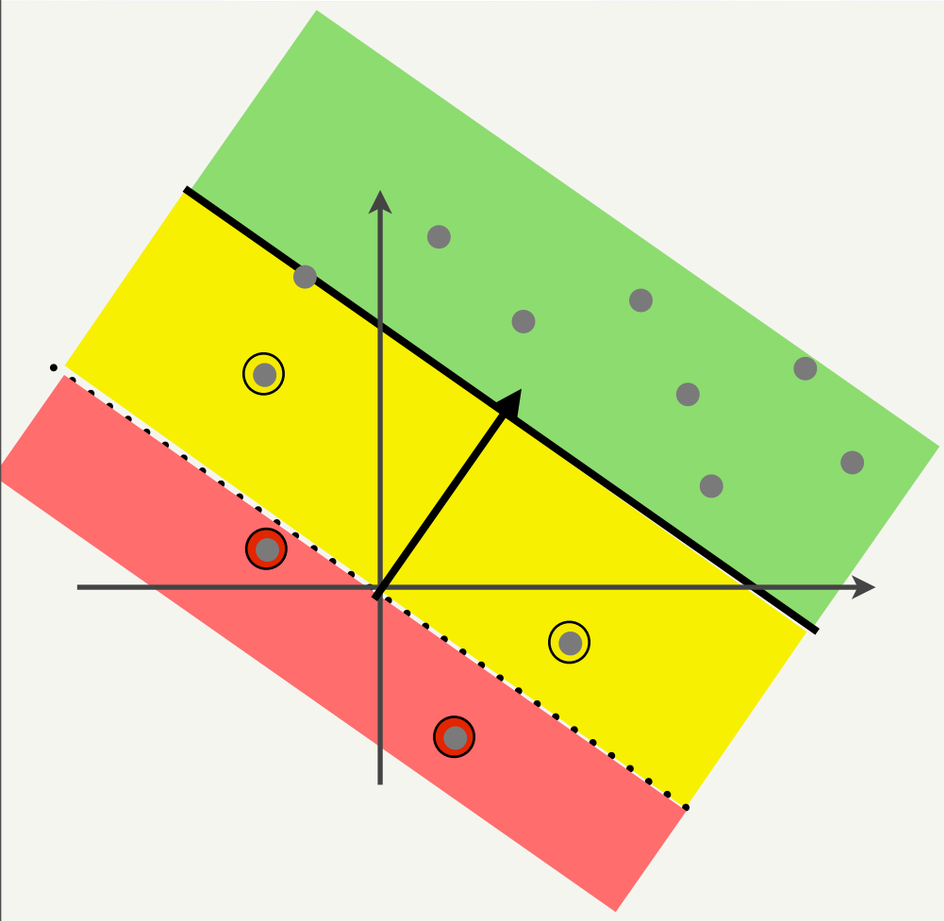


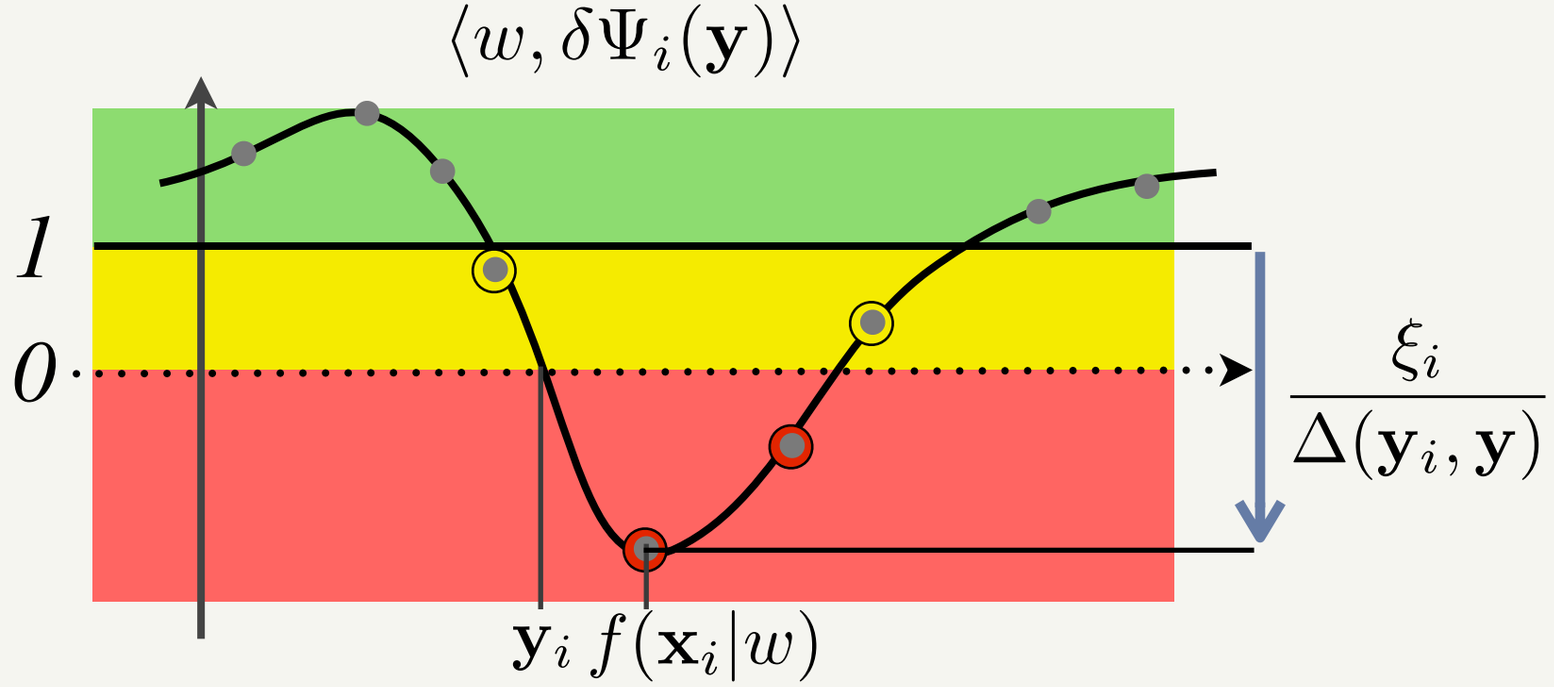


$$\langle w, \Psi_i(\mathbf{y})\rangle \geq 1 - \xi_i/\Delta(\mathbf{y}_i, \mathbf{y})$$

$$\xi_i \geq 0$$

Non-separable case



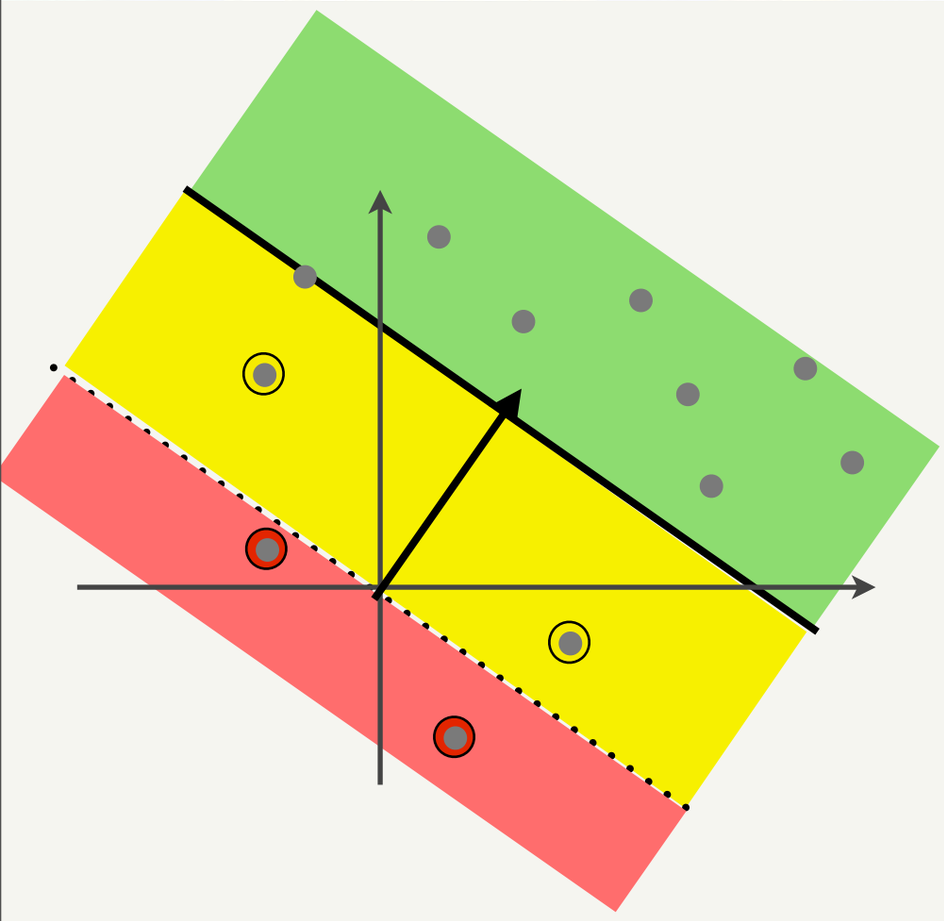


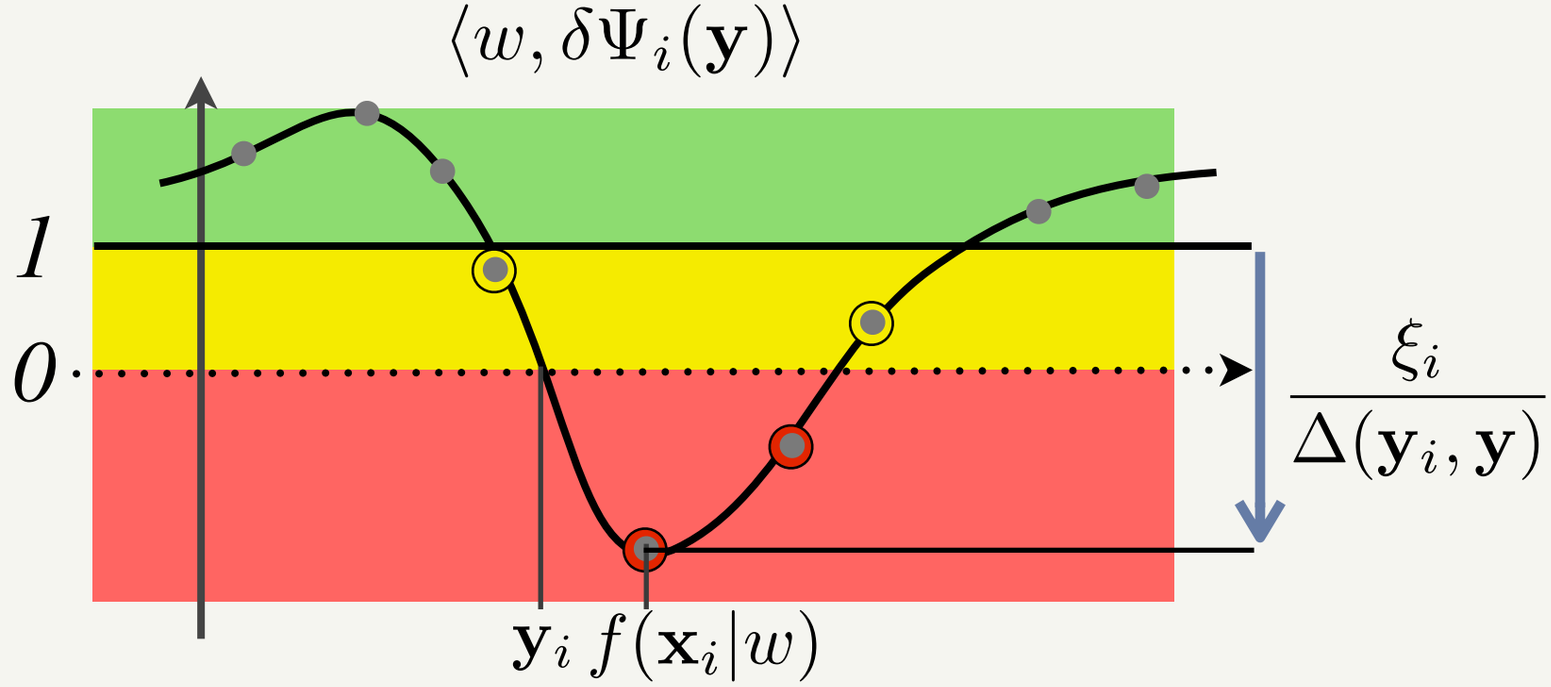
$$\langle w, \Psi_i(\mathbf{y}) \rangle \geq 1 - \xi_i / \Delta(\mathbf{y}_i, \mathbf{y})$$

$$\xi_i \geq 0$$

ξ_i UB on Loss: $\Delta(\mathbf{y}_i, f(\mathbf{x}_i|w)) \leq \xi_i$

Non-separable case





$$\langle w, \Psi_i(\mathbf{y})\rangle \geq 1 - \xi_i/\Delta(\mathbf{y}_i, \mathbf{y})$$
$$\xi_i \geq 0$$

ξ_i UB on Loss: $\Delta(\mathbf{y}_i, f(\mathbf{x}_i|w)) \leq \xi_i$

$$\min \tfrac{1}{2}\|w\|^2 + \tfrac{C}{N}\sum_{i=1}^{N}\xi_i,$$
$$\xi_i \geq 0 \quad \forall i,$$
$$\langle w, \delta\Psi_i(\mathbf{y})\rangle \geq 1 - \xi_i/\Delta(\mathbf{y}_i, \mathbf{y}) \quad \forall i, \mathbf{y} \neq \mathbf{y}_i$$

Recap

SVM*struct* regression

$$\min \tfrac{1}{2}\|w\|^2 + \tfrac{C}{N}\sum_{i=1}^{N}\xi_i,$$
$$\xi_i \geq 0 \quad \forall i,$$
$$\langle w, \delta\Psi_i(\mathbf{y})\rangle \geq 1 - \xi_i/\Delta(\mathbf{y}_i, \mathbf{y}) \quad \forall i, \mathbf{y} \neq \mathbf{y}_i$$

- Max-margin: regularized (and sparse) solution
- Soft-margin: non-separable case
- One constraint for each data point $\boldsymbol{x}_i$ and output value $\mathbf{y}$
- One slack variable for each data point $\boldsymbol{x}_i$
- Minimize upper bound on empirical error
- (Generalization results)

Kernelization

- Compute a kernel $k((\mathbf{x},\mathbf{y}),(\mathbf{x}',\mathbf{y}'))$ instead of $\psi(\mathbf{x},\mathbf{y})$

$$\begin{aligned}
&\min \tfrac{1}{2}\|w\|^2 + \tfrac{C}{N}\textstyle\sum_{i=1}^{N}\xi_i, \\
&\xi_i \geq 0 \quad \forall i, \\
&\langle w, \delta\Psi_i(\mathbf{y})\rangle \geq 1 - \xi_i/\Delta(\mathbf{y}_i,\mathbf{y}) \quad \forall i, \mathbf{y} \neq \mathbf{y}_i
\end{aligned}$$

Kernelization

- Compute a kernel $k((\boldsymbol{x},\boldsymbol{y}),(\boldsymbol{x}',\boldsymbol{y}'))$ instead of $\psi(\boldsymbol{x},\boldsymbol{y})$

$$\begin{aligned}
&\min \tfrac{1}{2}\|w\|^2 + \tfrac{C}{N}\textstyle\sum_{i=1}^{N} \xi_i, \\
&\xi_i \geq 0 \quad \forall i, \\
&\langle w, \delta\Psi_i(\mathbf{y})\rangle \geq 1 - \xi_i/\Delta(\mathbf{y}_i, \mathbf{y}) \quad \forall i, \mathbf{y} \neq \mathbf{y}_i
\end{aligned}$$

$$w = \sum_{i\mathbf{y}} c_{i\mathbf{y}} \Psi(\mathbf{x}_i, \mathbf{y})$$

$$k(j\mathbf{y}', i\mathbf{y}) = \langle \Psi(\mathbf{x}_j, \mathbf{y}'), \Psi(\mathbf{x}_i, \mathbf{y})\rangle$$

Kernelization

- Compute a kernel $k((\mathbf{x},\mathbf{y}),(\mathbf{x}',\mathbf{y}'))$ instead of $\psi(\mathbf{x},\mathbf{y})$

$$\begin{aligned}
&\min \tfrac{1}{2}\|w\|^2 + \tfrac{C}{N}\textstyle\sum_{i=1}^{N} \xi_i, \\
&\xi_i \geq 0 \quad \forall i, \\
&\langle w, \delta\Psi_i(\mathbf{y})\rangle \geq 1 - \xi_i/\Delta(\mathbf{y}_i,\mathbf{y}) \quad \forall i, \mathbf{y} \neq \mathbf{y}_i
\end{aligned}$$

$$w = \sum_{i\mathbf{y}} c_{i\mathbf{y}} \Psi(\mathbf{x}_i, \mathbf{y})$$

$$k(j\mathbf{y}', i\mathbf{y}) = \langle \Psi(\mathbf{x}_j, \mathbf{y}'), \Psi(\mathbf{x}_i, \mathbf{y})\rangle$$

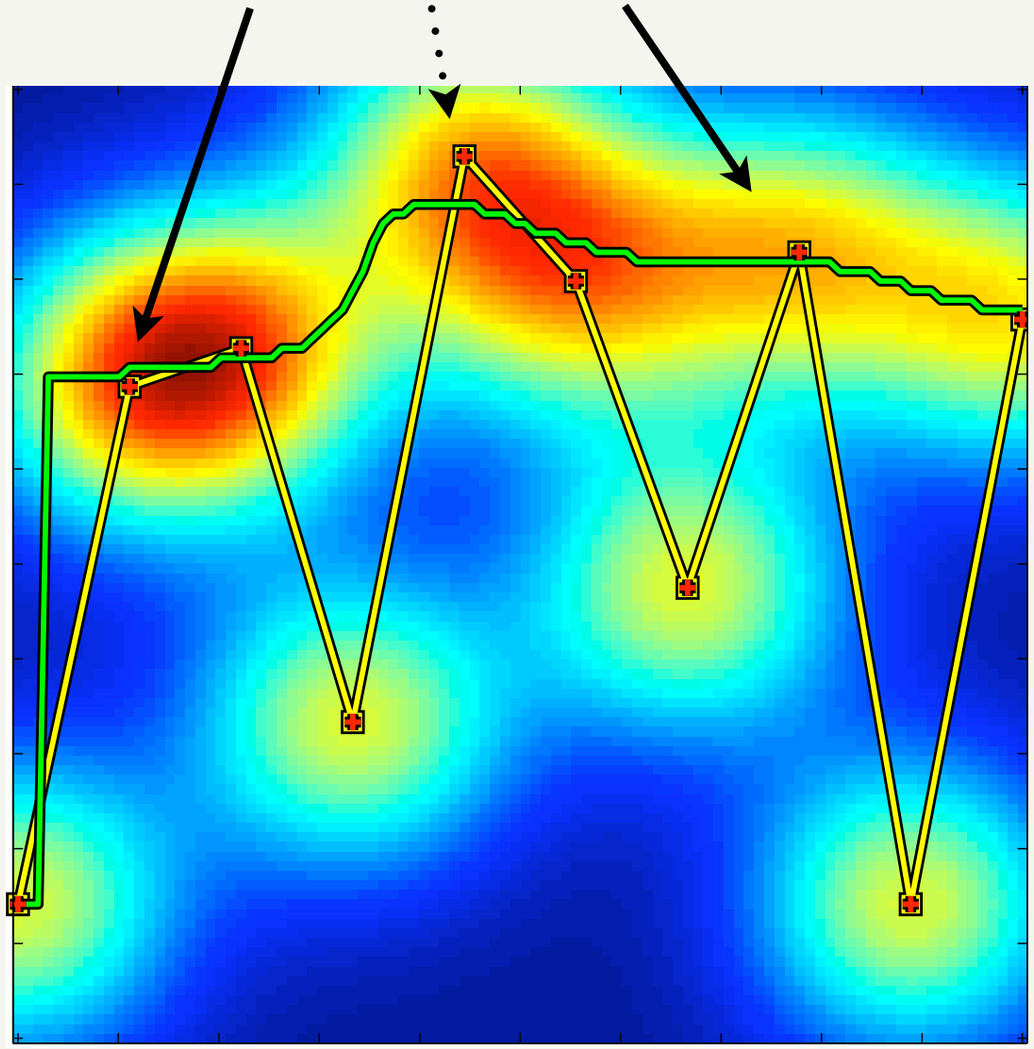
$$\sum_{j\mathbf{y}'i\mathbf{y}} c_{j\mathbf{y}'} c_{i\mathbf{y}} k(j\mathbf{y}', i\mathbf{y})$$

$$\sum_{j\mathbf{y}'} c_{j\mathbf{y}'} (k(j\mathbf{y}', i\mathbf{y}_i) - k(j\mathbf{y}', i\mathbf{y}))$$

Example: Real function

- Fit the real function to data $f(x|w) : \mathbb{R} \to \mathbb{R}$

data $(x_1, y_1), \ldots, (x_N, y_N)$

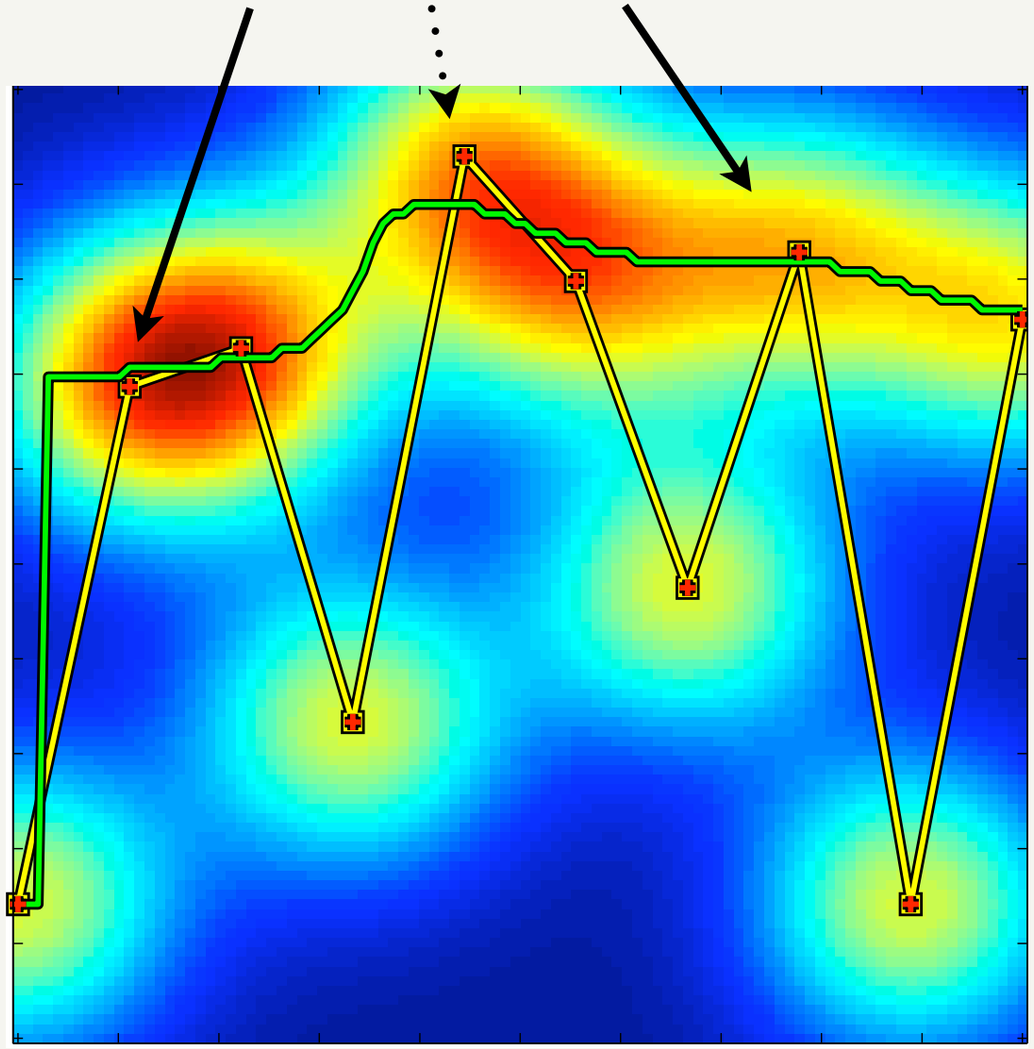


$$F(x, y|c) = \sum_{jy'} c_{jy'} k((x_j, y'), (x, y))$$

Example: Real function

- Fit the real function to data $f(x|w) : \mathbb{R} \to \mathbb{R}$

data $(x_1, y_1), \ldots, (x_N, y_N)$



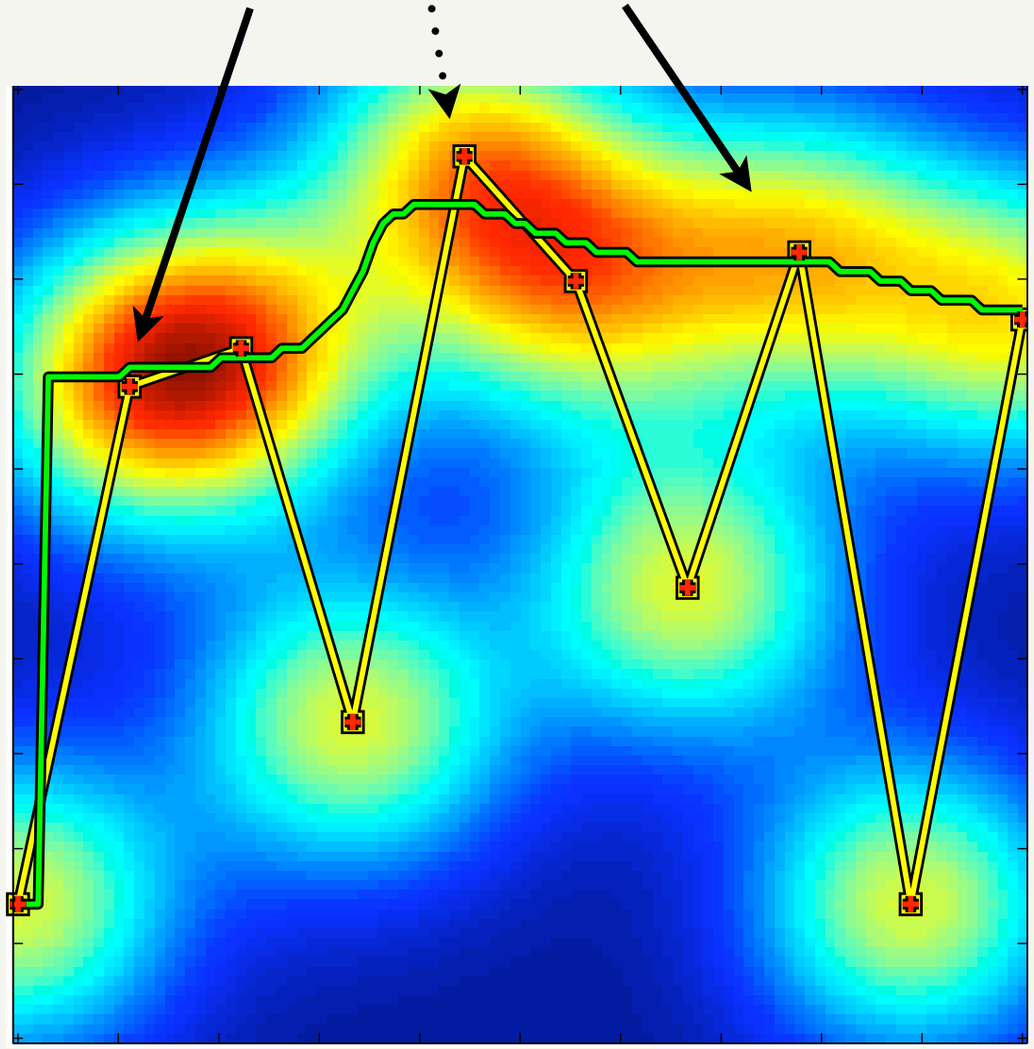
loss $\Delta(y, y_i) = (y - y_i)^2$

$$F(x, y|c) = \sum_{jy'} c_{jy'} k((x_j, y'), (x, y))$$

Example: Real function

- Fit the real function to data $f(x|w) : \mathbb{R} \to \mathbb{R}$

data $(x_1, y_1), \ldots, (x_N, y_N)$



loss $\Delta(y, y_i) = (y - y_i)^2$

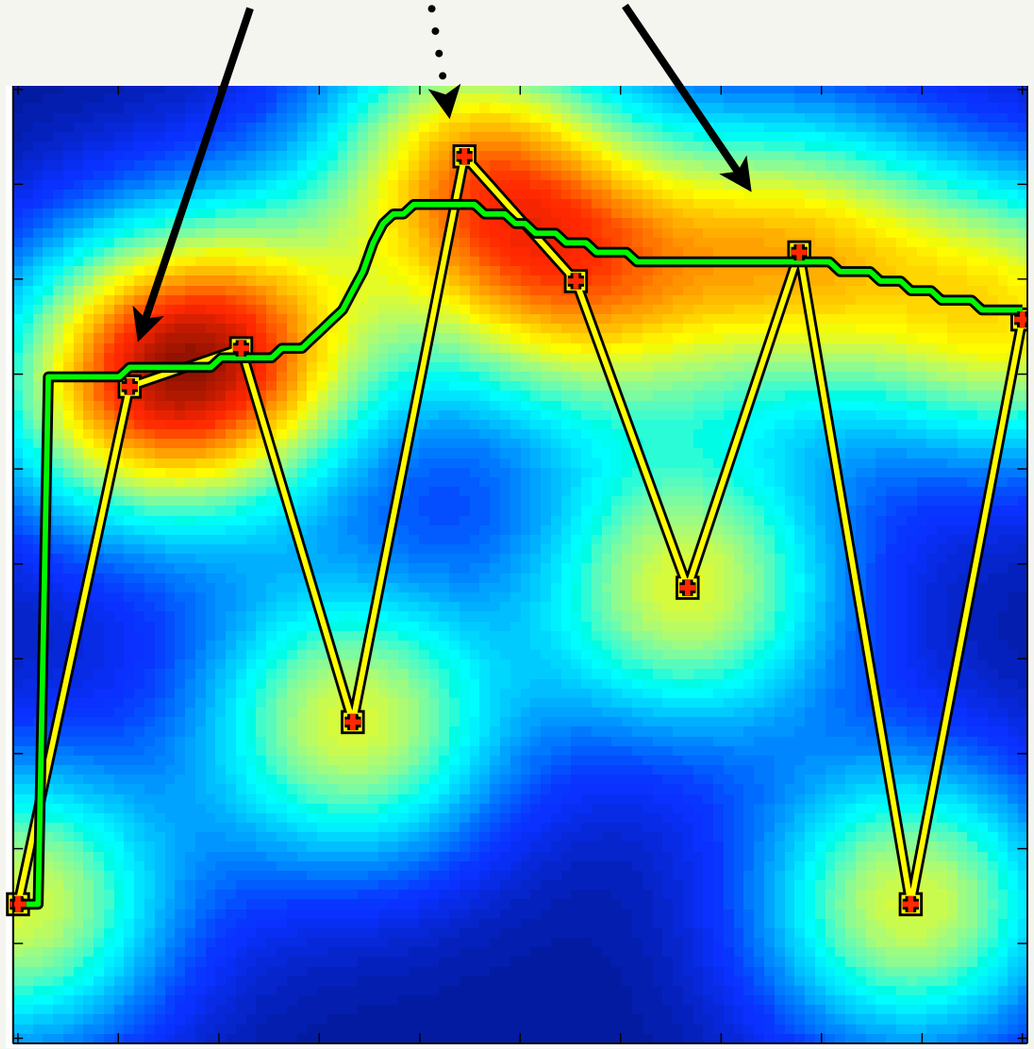
risk $R(w) = \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i|w))^2$

$$F(x, y|c) = \sum_{jy'} c_{jy'} k((x_j, y'), (x, y))$$

Example: Real function

- Fit the real function to data $f(x|w) : \mathbb{R} \to \mathbb{R}$

data $(x_1, y_1), \ldots, (x_N, y_N)$



loss $\Delta(y, y_i) = (y - y_i)^2$

risk $R(w) = \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i|w))^2$

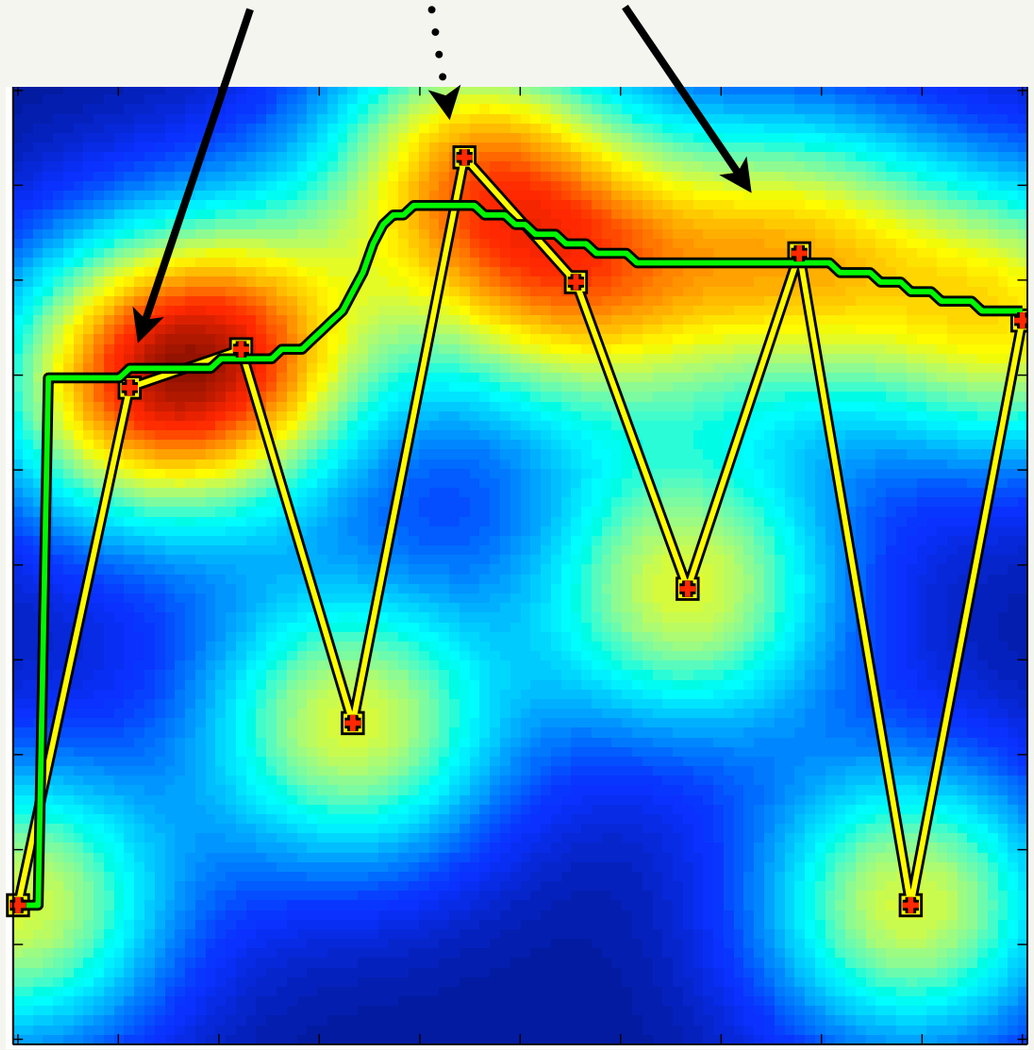
kernel $k(jy', iy) = e^{-\gamma_1 (x_j - x_i)^2 - \gamma_2 (y' - y)^2}$

$$F(x, y|c) = \sum_{jy'} c_{jy'} k((x_j, y'), (x, y))$$

Example: Real function

- Fit the real function to data $f(x|w) : \mathbb{R} \to \mathbb{R}$

data $(x_1, y_1), \ldots, (x_N, y_N)$



loss $\Delta(y, y_i) = (y - y_i)^2$

risk $R(w) = \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i|w))^2$

kernel $k(jy', iy) = e^{-\gamma_1 (x_j - x_i)^2 - \gamma_2 (y' - y)^2}$

coeff. $c_{iy} = \delta_{y=y_i}$

$$F(x, y|c) = \sum_{jy'} c_{jy'} k((x_j, y'), (x, y))$$

Dual

Primal in matrix notation

$$\min \tfrac{1}{2}\|w\|^2 + \tfrac{C}{N}\sum_{i=1}^{N}\xi_i,$$
$$\xi_i \geq 0 \quad \forall i,$$
$$\langle w, \delta\Psi_i(\mathbf{y})\rangle \geq 1 - \xi_i/\Delta(\mathbf{y}_i, \mathbf{y}) \quad \forall i, \mathbf{y} \neq \mathbf{y}_i$$

$$w = \sum_{i\mathbf{y}} c_{i\mathbf{y}} \delta\Psi(\mathbf{x}_i, \mathbf{y})$$

$$k(j\mathbf{y}', i\mathbf{y}) = \langle \Psi(\mathbf{x}_j, \mathbf{y}'), \Psi(\mathbf{x}_i, \mathbf{y})\rangle$$

Primal in matrix notation

- Assume finite range $Y = \{1, \ldots, M\}$

$$\min \tfrac{1}{2}\|w\|^2 + \tfrac{C}{N}\sum_{i=1}^{N}\xi_i,$$
$$\xi_i \geq 0 \quad \forall i,$$
$$\langle w, \delta\Psi_i(\mathbf{y})\rangle \geq 1 - \xi_i/\Delta(\mathbf{y}_i, \mathbf{y}) \quad \forall i, \mathbf{y} \neq \mathbf{y}_i$$

$$w = \sum_{i\mathbf{y}} c_{i\mathbf{y}}\delta\Psi(\mathbf{x}_i, \mathbf{y})$$

$$k(j\mathbf{y}', i\mathbf{y}) = \langle \Psi(\mathbf{x}_j, \mathbf{y}'), \Psi(\mathbf{x}_i, \mathbf{y})\rangle$$

Primal in matrix notation

- Assume finite range $Y = \{1, \ldots, M\}$
- Lexicographic ordering of pairs $(i, \mathbf{y})$

$$\min \tfrac{1}{2}\|w\|^2 + \tfrac{C}{N}\sum_{i=1}^{N}\xi_i,$$
$$\xi_i \geq 0 \quad \forall i,$$
$$\langle w, \delta\Psi_i(\mathbf{y})\rangle \geq 1 - \xi_i/\Delta(\mathbf{y}_i, \mathbf{y}) \quad \forall i, \mathbf{y} \neq \mathbf{y}_i$$

$$w = \sum_{i\mathbf{y}} c_{i\mathbf{y}}\delta\Psi(\mathbf{x}_i, \mathbf{y})$$

$$k(j\mathbf{y}', i\mathbf{y}) = \langle \Psi(\mathbf{x}_j, \mathbf{y}'), \Psi(\mathbf{x}_i, \mathbf{y})\rangle$$

Primal in matrix notation

- Assume finite range $Y = \{1, \ldots, M\}$
- Lexicographic ordering of pairs $(i, \mathbf{y})$

$$\begin{aligned}
&\min \tfrac{1}{2}\|w\|^2 + \tfrac{C}{N}\textstyle\sum_{i=1}^{N} \xi_i, \\
&\xi_i \geq 0 \quad \forall i, \\
&\langle w, \delta\Psi_i(\mathbf{y})\rangle \geq 1 - \xi_i/\Delta(\mathbf{y}_i, \mathbf{y}) \quad \forall i, \mathbf{y} \neq \mathbf{y}_i
\end{aligned}$$

$$w = \sum_{i\mathbf{y}} c_{i\mathbf{y}} \delta\Psi(\mathbf{x}_i, \mathbf{y})$$

$$k(j\mathbf{y}', i\mathbf{y}) = \langle \Psi(\mathbf{x}_j, \mathbf{y}'), \Psi(\mathbf{x}_i, \mathbf{y})\rangle$$

$$\mathbf{c} = \begin{bmatrix} \vdots \\ c_{i\mathbf{y}} \\ \vdots \end{bmatrix}$$

Primal in matrix notation

- Assume finite range $Y = \{1, \ldots, M\}$
- Lexicographic ordering of pairs $(i, \mathbf{y})$

$$\begin{aligned}
&\min \tfrac{1}{2}\|w\|^2 + \tfrac{C}{N}\textstyle\sum_{i=1}^{N} \xi_i, \\
&\xi_i \geq 0 \quad \forall i, \\
&\langle w, \delta\Psi_i(\mathbf{y})\rangle \geq 1 - \xi_i/\Delta(\mathbf{y}_i, \mathbf{y}) \quad \forall i, \mathbf{y} \neq \mathbf{y}_i
\end{aligned}$$

$$w = \sum_{i\mathbf{y}} c_{i\mathbf{y}} \delta\Psi(\mathbf{x}_i, \mathbf{y})$$

$$k(j\mathbf{y}', i\mathbf{y}) = \langle \Psi(\mathbf{x}_j, \mathbf{y}'), \Psi(\mathbf{x}_i, \mathbf{y})\rangle$$

$$\mathbf{c} = \begin{bmatrix} \vdots \\ c_{i\mathbf{y}} \\ \vdots \end{bmatrix} \qquad K = \begin{bmatrix} \ddots & & \\ & k(i\mathbf{y}, j\mathbf{y}') & \\ & & \ddots \end{bmatrix}$$

Primal in matrix notation

- Assume finite range $Y = \{1, \ldots, M\}$
- Lexicographic ordering of pairs $(i, \mathbf{y})$

$$\begin{aligned}
&\min \tfrac{1}{2}\|w\|^2 + \tfrac{C}{N}\textstyle\sum_{i=1}^{N} \xi_i, \\
&\xi_i \geq 0 \quad \forall i, \\
&\langle w, \delta\Psi_i(\mathbf{y})\rangle \geq 1 - \xi_i/\Delta(\mathbf{y}_i, \mathbf{y}) \quad \forall i, \mathbf{y} \neq \mathbf{y}_i
\end{aligned}$$

$$w = \sum_{i\mathbf{y}} c_{i\mathbf{y}} \delta\Psi(\mathbf{x}_i, \mathbf{y})$$

$$k(j\mathbf{y}', i\mathbf{y}) = \langle \Psi(\mathbf{x}_j, \mathbf{y}'), \Psi(\mathbf{x}_i, \mathbf{y})\rangle$$

$$\mathbf{c} = \begin{bmatrix} \vdots \\ c_{i\mathbf{y}} \\ \vdots \end{bmatrix} \qquad K = \begin{bmatrix} \ddots & & \\ & k(i\mathbf{y}, j\mathbf{y}') & \\ & & \ddots \end{bmatrix} \qquad E = \begin{bmatrix} \ldots & e_i \otimes (e_{\mathbf{y}_i} - e_{\mathbf{y}}) & \ldots \end{bmatrix}$$

$$A = \begin{bmatrix} \ldots\ \Delta(\mathbf{y}_1\mathbf{y})^{-1}\ \ldots & & \ldots 0 \ldots \\ & \ddots & \\ \ldots 0 \ldots & & \ldots\ \Delta(\mathbf{y}_N\mathbf{y})^{-1}\ \ldots \end{bmatrix} \qquad \mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

Primal in matrix notation

- Assume finite range $Y = \{1, \ldots, M\}$
- Lexicographic ordering of pairs $(i, \mathbf{y})$

$$\begin{aligned}
&\min \tfrac{1}{2}\|w\|^2 + \tfrac{C}{N}\textstyle\sum_{i=1}^{N} \xi_i, \\
&\xi_i \geq 0 \quad \forall i, \\
&\langle w, \delta\Psi_i(\mathbf{y})\rangle \geq 1 - \xi_i/\Delta(\mathbf{y}_i, \mathbf{y}) \quad \forall i, \mathbf{y} \neq \mathbf{y}_i
\end{aligned}$$

$$w = \sum_{i\mathbf{y}} c_{i\mathbf{y}} \delta\Psi(\mathbf{x}_i, \mathbf{y})$$

$$k(j\mathbf{y}', i\mathbf{y}) = \langle \Psi(\mathbf{x}_j, \mathbf{y}'), \Psi(\mathbf{x}_i, \mathbf{y})\rangle$$

$$\mathbf{c} = \begin{bmatrix} \vdots \\ c_{i\mathbf{y}} \\ \vdots \end{bmatrix} \qquad K = \begin{bmatrix} \ddots & & \\ & k(i\mathbf{y}, j\mathbf{y}') & \\ & & \ddots \end{bmatrix} \qquad E = \begin{bmatrix} \ldots & e_i \otimes (e_{\mathbf{y}_i} - e_{\mathbf{y}}) & \ldots \end{bmatrix}$$

$$A = \begin{bmatrix} \ldots\ \Delta(\mathbf{y}_1\mathbf{y})^{-1}\ \ldots & & \ldots 0 \ldots \\ & \ddots & \\ \ldots 0 \ldots & & \ldots\ \Delta(\mathbf{y}_N\mathbf{y})^{-1}\ \ldots \end{bmatrix} \qquad \mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$\begin{aligned}
&\min \tfrac{1}{2}\mathbf{c}^\top K \mathbf{c} + \tfrac{C}{N}\mathbf{1}^\top \xi \\
&\xi \geq 0, \\
&E^\top K \mathbf{c} \geq \mathbf{1} - A^\top \xi
\end{aligned}$$

Primal → dual

$$\begin{aligned}
&\min \tfrac{1}{2}\mathbf{c}^\top K\mathbf{c} + \tfrac{C}{N}\mathbf{1}^\top\xi \\
&\xi \geq 0, \\
&E^\top K\mathbf{c} \geq \mathbf{1} - A^\top\xi
\end{aligned}$$

Primal → dual

$$\begin{aligned}&\min \tfrac{1}{2}\mathbf{c}^\top K\mathbf{c} + \tfrac{C}{N}\mathbf{1}^\top\xi\\&\xi \geq 0,\\&E^\top K\mathbf{c} \geq \mathbf{1} - A^\top\xi\end{aligned}$$

- Introduce dual variables $\boldsymbol{\alpha}$

$$\max_{\alpha}\ \min_{\mathbf{c},\xi}\ \frac{1}{2}\mathbf{c}^\top K\mathbf{c} + \frac{C}{N}\mathbf{1}^\top\xi + \alpha^\top(\mathbf{1} - A^\top\xi - E^\top K\mathbf{c})$$

Primal → dual

$$\begin{aligned}&\min \tfrac{1}{2}\mathbf{c}^\top K\mathbf{c} + \tfrac{C}{N}\mathbf{1}^\top\xi\\&\xi \geq 0,\\&E^\top K\mathbf{c} \geq \mathbf{1} - A^\top\xi\end{aligned}$$

- Introduce dual variables α

$$\max_{\alpha}\ \min_{\mathbf{c},\xi} \frac{1}{2}\mathbf{c}^\top K\mathbf{c} + \frac{C}{N}\mathbf{1}^\top\xi + \alpha^\top(\mathbf{1} - A^\top\xi - E^\top K\mathbf{c})$$

- Optimize primal variables w, ξ
  - partial = 0 yields $\mathbf{c} = E\alpha$
  - bounded minimum yields $A\alpha \leq \frac{C}{N}\mathbf{1}$

$$\begin{aligned}&\max \mathbf{1}^\top\alpha - \tfrac{1}{2}\alpha^\top E^\top K E\alpha\\&\alpha \geq 0,\\&A\alpha \leq \tfrac{C}{N}\mathbf{1}\end{aligned}$$

Solution and Support Vectors

Solution and Support Vectors

- Support Vector (SV): pairs (x_i, y) with coefficient $c_{iy} > 0$

Solution and Support Vectors

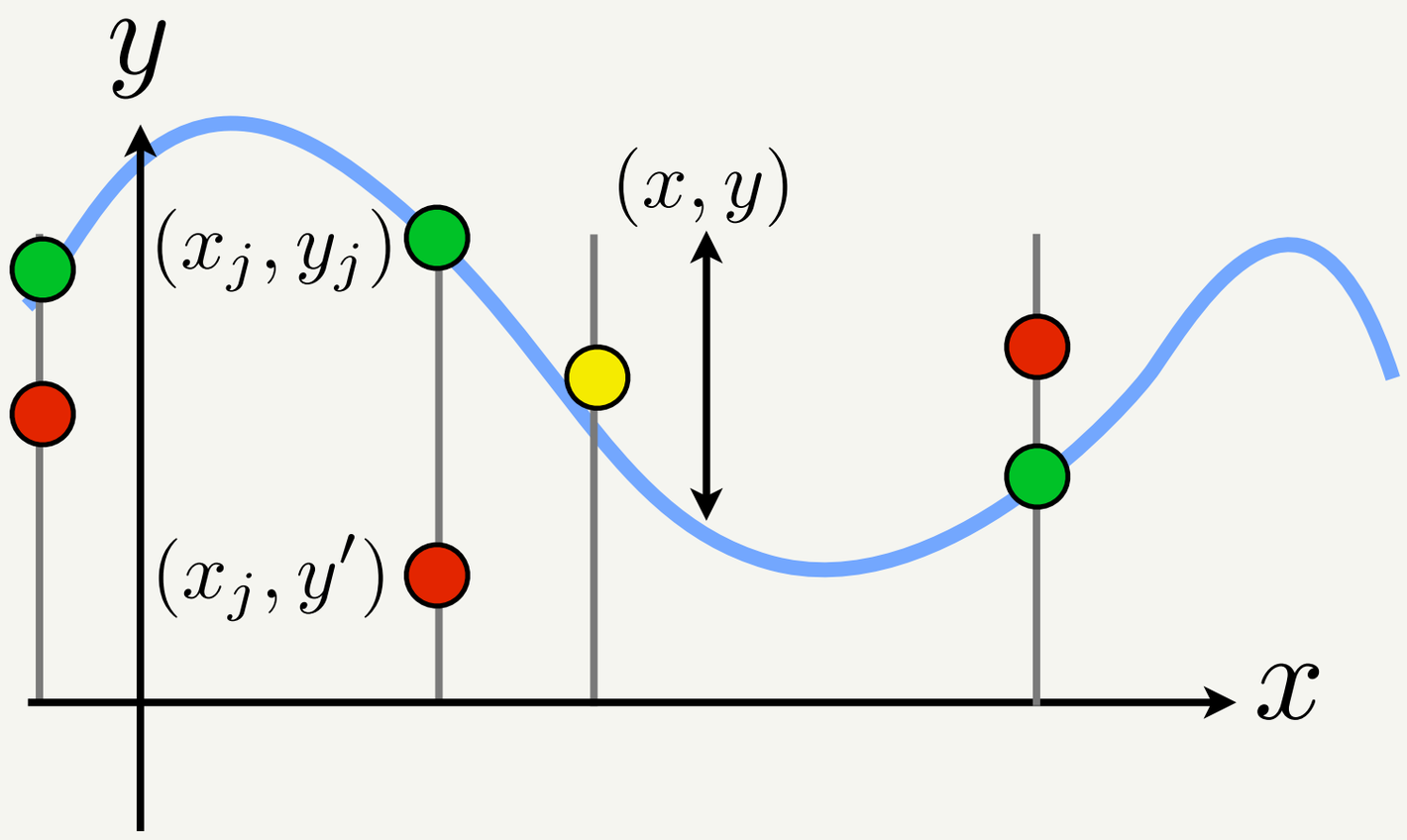
- Support Vector (SV): pairs (x_i, y) with coefficient $c_{iy} > 0$
- Dual variable $\alpha_{iy} > 0$ results in difference of 2 SVs c_{iy} and c_{iyi}

$$
\begin{aligned}
f(\mathbf{x}|w) &= \operatorname*{argmax}_{\mathbf{y}\in\mathcal{Y}} \sum_{j\mathbf{y}'} c_{j\mathbf{y}'} \langle \Psi(\mathbf{x}_j, \mathbf{y}'), \Psi(\mathbf{x}, \mathbf{y}) \rangle \\
&= \operatorname*{argmax}_{\mathbf{y}\in\mathcal{Y}} \sum_{j\mathbf{y}'} \alpha_{j\mathbf{y}'} \langle \Psi(\mathbf{x}_j, \mathbf{y}_j) - \Psi(\mathbf{x}_j, \mathbf{y}'), \Psi(\mathbf{x}, \mathbf{y}) \rangle
\end{aligned}
$$

Solution and Support Vectors

- Support Vector (SV): pairs (x_i, y) with coefficient $c_{iy} > 0$
- Dual variable $\alpha_{iy} > 0$ results in difference of 2 SVs c_{iy} and c_{iyi}

$$
\begin{aligned}
f(\mathbf{x}|w) &= \operatorname*{argmax}_{\mathbf{y}\in\mathcal{Y}} \sum_{j\mathbf{y}'} c_{j\mathbf{y}'} \langle \Psi(\mathbf{x}_j, \mathbf{y}'), \Psi(\mathbf{x}, \mathbf{y}) \rangle \\
&= \operatorname*{argmax}_{\mathbf{y}\in\mathcal{Y}} \sum_{j\mathbf{y}'} \alpha_{j\mathbf{y}'} \langle \Psi(\mathbf{x}_j, \mathbf{y}_j) - \Psi(\mathbf{x}_j, \mathbf{y}'), \Psi(\mathbf{x}, \mathbf{y}) \rangle
\end{aligned}
$$



Optimization

Dealing with many constraints

Dealing with many constraints

- Number of constraints $\propto |Y|$
  - $|Y|$ can be very large or ∞
  - infinite number of dual variables?

Dealing with many constraints

- Number of constraints $\propto |Y|$
  - $|Y|$ can be very large or ∞
  - infinite number of dual variables?
- Exploit sparsity of solutions

Dealing with many constraints

- Number of constraints $\propto |Y|$
  - $|Y|$ can be very large or ∞
  - infinite number of dual variables?
- Exploit sparsity of solutions
- Add one constraint/dual variable per time
  - *"the most violated constraint"*
  - until violation increases less than ε

Dealing with many constraints

- Number of constraints $\propto |Y|$
  - $|Y|$ can be very large or ∞
  - infinite number of dual variables?
- Exploit sparsity of solutions
- Add one constraint/dual variable per time
  - *"the most violated constraint"*
  - until violation increases less than ε
- Nice facts:
  - stops with an ε - optimal solution ...
  - ... in polynomial time

ε-satisfaction ⇒ ε-optimality

ε-satisfaction ⇒ ε-optimality

full primal

$$\min \tfrac{1}{2}\mathbf{c}^\top K\mathbf{c} + \tfrac{C}{N}\mathbf{1}^\top\xi$$
$$\xi \geq 0,$$
$$E^\top K\mathbf{c} \geq \mathbf{1} - A^\top\xi$$

ε-satisfaction ⇒ ε-optimality

full primal

$$\min \tfrac{1}{2}\mathbf{c}^\top K\mathbf{c} + \tfrac{C}{N}\mathbf{1}^\top\xi$$
$$\xi \geq 0,$$
$$E^\top K\mathbf{c} \geq \mathbf{1} - A^\top\xi$$

partially constrained primal

$$\min \tfrac{1}{2}\mathbf{c}^\top K\mathbf{c} + \tfrac{C}{N}\mathbf{1}^\top\xi$$
$$\xi \geq 0,$$
$$[E^\top K\mathbf{c}]_S \geq [\mathbf{1} - A^\top\xi]_S$$

ε-satisfaction ⇒ ε-optimality

full primal

$$\min \tfrac{1}{2}\mathbf{c}^\top K\mathbf{c} + \tfrac{C}{N}\mathbf{1}^\top\xi$$
$$\xi \geq 0,$$
$$E^\top K\mathbf{c} \geq \mathbf{1} - A^\top\xi$$

partially constrained primal

$$\min \tfrac{1}{2}\mathbf{c}^\top K\mathbf{c} + \tfrac{C}{N}\mathbf{1}^\top\xi$$
$$\xi \geq 0,$$
$$[E^\top K\mathbf{c}]_S \geq [\mathbf{1} - A^\top\xi]_S$$

active set

ε-satisfaction ⇒ ε-optimality

full primal

$$\min \tfrac{1}{2}\mathbf{c}^\top K\mathbf{c} + \tfrac{C}{N}\mathbf{1}^\top\xi$$
$$\xi \geq 0,$$
$$E^\top K\mathbf{c} \geq \mathbf{1} - A^\top\xi$$

optim: $E(\mathbf{c}^*, \xi^*)$

partially constrained primal

$$\min \tfrac{1}{2}\mathbf{c}^\top K\mathbf{c} + \tfrac{C}{N}\mathbf{1}^\top\xi$$
$$\xi \geq 0,$$
$$[E^\top K\mathbf{c}]_S \geq [\mathbf{1} - A^\top\xi]_S$$

active set

optim: $E(\mathbf{c}^*_S, \xi^*_S)$

ε-satisfaction ⇒ ε-optimality

full primal

$$\min \tfrac{1}{2}\mathbf{c}^\top K\mathbf{c} + \tfrac{C}{N}\mathbf{1}^\top\xi$$
$$\xi \geq 0,$$
$$E^\top K\mathbf{c} \geq \mathbf{1} - A^\top\xi$$

optim: $E(\mathbf{c}^*, \xi^*)$

partially constrained primal

$$\min \tfrac{1}{2}\mathbf{c}^\top K\mathbf{c} + \tfrac{C}{N}\mathbf{1}^\top\xi$$
$$\xi \geq 0,$$
$$[E^\top K\mathbf{c}]_S \geq [\mathbf{1} - A^\top\xi]_S$$

active set

optim: $E(\mathbf{c}^*_S, \xi^*_S)$

Consider smallest ε such that

$$(\mathbf{c}, \xi) = (\mathbf{c}^*_S, \xi^*_S + \epsilon\mathbf{1})$$

is feasible for full primal. Then:

$$E(\mathbf{c}, \xi) = E(\mathbf{c}^*_S, \xi^*_S) + \epsilon C$$

ε-satisfaction ⇒ ε-optimality

full primal

$$\min \tfrac{1}{2}\mathbf{c}^\top K\mathbf{c} + \tfrac{C}{N}\mathbf{1}^\top\xi$$
$$\xi \geq 0,$$
$$E^\top K\mathbf{c} \geq \mathbf{1} - A^\top\xi$$

optim: $E(\mathbf{c}^*, \xi^*)$

partially constrained primal

$$\min \tfrac{1}{2}\mathbf{c}^\top K\mathbf{c} + \tfrac{C}{N}\mathbf{1}^\top\xi$$
$$\xi \geq 0,$$
$$[E^\top K\mathbf{c}]_S \geq [\mathbf{1} - A^\top\xi]_S$$

active set

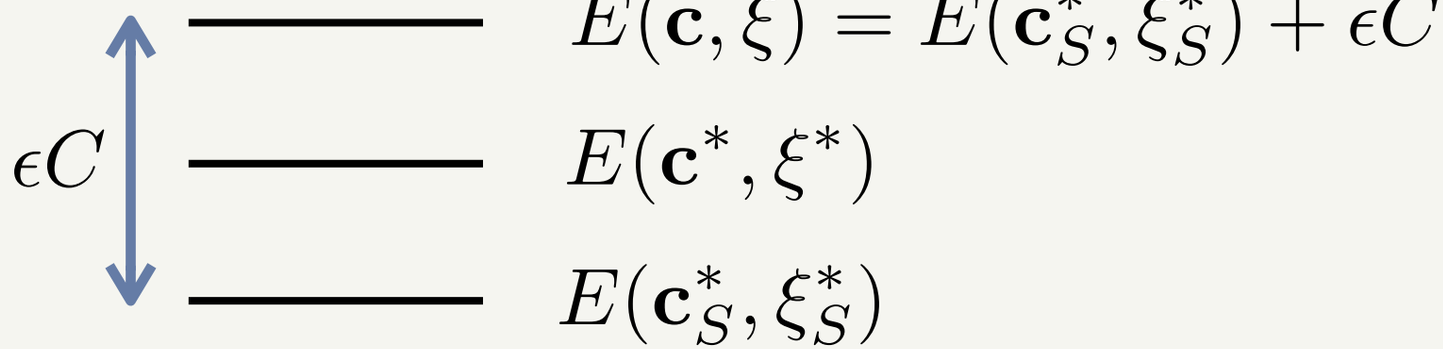
optim: $E(\mathbf{c}^*_S, \xi^*_S)$

Consider smallest ε such that

$$(\mathbf{c}, \xi) = (\mathbf{c}^*_S, \xi^*_S + \epsilon\mathbf{1})$$

is feasible for full primal. Then:

$$E(\mathbf{c}, \xi) = E(\mathbf{c}^*_S, \xi^*_S) + \epsilon C$$



Hunting ε-violations

Hunting ε-violations

- Start with $S = \{\}$, a precision ε

Hunting ε-violations

- Start with $S = \{\}$, a precision ε
  - Solve the *partial* dual problem (one var for each ***active*** constraint)

$$\max \mathbf{1}^\top [\alpha]_S - \tfrac{1}{2}[\alpha]_S^\top [E^\top K E]_S [\alpha]_S$$
$$[\alpha]_S \geq 0,$$
$$[A]_S [\alpha]_S \leq \tfrac{C}{N}\mathbf{1}$$

Hunting ε-violations

- Start with $S = \{\}$, a precision ε
  - Solve the *partial* dual problem (one var for each ***active*** constraint)
  - Strong duality: get solution of the *partial primal* problem:

$$\max \mathbf{1}^\top [\alpha]_S - \tfrac{1}{2}[\alpha]_S^\top [E^\top K E]_S [\alpha]_S$$
$$[\alpha]_S \geq 0,$$
$$[A]_S [\alpha]_S \leq \tfrac{C}{N}\mathbf{1}$$

$$\mathbf{c}_S^* = [E]_S [\alpha]_S$$

$$[\xi_S^*]_i = \max\{\Delta(\mathbf{y}_i, \mathbf{y})[1 - E^\top K \mathbf{c}_S^*]_{i\mathbf{y}} : i\mathbf{y} \in S\} \cup \{0\}$$

Hunting ε-violations

- Start with $S = \{\}$, a precision ε
  - Solve the *partial* dual problem (one var for each ***active*** constraint)

$$\max \mathbf{1}^\top [\alpha]_S - \tfrac{1}{2}[\alpha]_S^\top [E^\top K E]_S [\alpha]_S$$
$$[\alpha]_S \geq 0,$$
$$[A]_S [\alpha]_S \leq \tfrac{C}{N}\mathbf{1}$$

  - Strong duality: get solution of the *partial primal* problem:

$$\mathbf{c}_S^* = [E]_S [\alpha]_S$$

$$[\xi_S^*]_i = \max\{\Delta(\mathbf{y}_i, \mathbf{y})[1 - E^\top K \mathbf{c}_S^*]_{i\mathbf{y}} : i\mathbf{y} \in S\} \cup \{0\}$$

- **Remark.** ξ_S^* satisfies only *active* constraints.

Hunting ε-violations

- Start with $S = \{\}$, a precision ε
  - Solve the *partial* dual problem (one var for each ***active*** constraint)

$$\max \mathbf{1}^\top [\alpha]_S - \tfrac{1}{2}[\alpha]_S^\top [E^\top K E]_S [\alpha]_S$$
$$[\alpha]_S \geq 0,$$
$$[A]_S [\alpha]_S \leq \tfrac{C}{N}\mathbf{1}$$

  - Strong duality: get solution of the *partial primal* problem:

$$\mathbf{c}_S^* = [E]_S [\alpha]_S$$

$$[\xi_S^*]_i = \max\{\Delta(\mathbf{y}_i, \mathbf{y})[1 - E^\top K \mathbf{c}_S^*]_{i\mathbf{y}} : i\mathbf{y} \in S\} \cup \{0\}$$

    - **Remark.** ξ_S^* satisfies only *active* constraints.
- Find ε' such that $\xi_S^* + \epsilon'\mathbf{1}$ satisfies *all* constraints:

$$[\xi]_i = \max\{\Delta(\mathbf{y}_i, \mathbf{y})[1 - E^\top K \mathbf{c}_S^*]_{i\mathbf{y}} \ \forall i\mathbf{y}\} \cup \{0\}$$

$$\epsilon' = \max_i [\xi - \xi_S^*]_i$$

Hunting ε-violations

- Start with $S = \{\}$, a precision ε
  - Solve the *partial* dual problem (one var for each ***active*** constraint)

$$\max \mathbf{1}^\top [\alpha]_S - \tfrac{1}{2}[\alpha]_S^\top [E^\top K E]_S [\alpha]_S$$
$$[\alpha]_S \geq 0,$$
$$[A]_S [\alpha]_S \leq \tfrac{C}{N}\mathbf{1}$$

  - Strong duality: get solution of the *partial primal* problem:

$$\mathbf{c}_S^* = [E]_S [\alpha]_S$$

$$[\xi_S^*]_i = \max\{\Delta(\mathbf{y}_i, \mathbf{y})[1 - E^\top K \mathbf{c}_S^*]_{i\mathbf{y}} : i\mathbf{y} \in S\} \cup \{0\}$$

    - **Remark.** ξ_S^* satisfies only *active* constraints.
- Find ε' such that $\xi_S^* + \epsilon'\mathbf{1}$ satisfies *all* constraints:

$$[\xi]_i = \max\{\Delta(\mathbf{y}_i, \mathbf{y})[1 - E^\top K \mathbf{c}_S^*]_{i\mathbf{y}} \ \forall i\mathbf{y}\} \cup \{0\}$$

$$\epsilon' = \max_i [\xi - \xi_S^*]_i$$

- Stop if $\varepsilon' < \varepsilon$

Finiteness

Finiteness

- When the algorithm stops, it returns an ε-optimal solution.

Finiteness

- When the algorithm stops, it returns an ε-optimal solution.
- But ... does it stop?

Finiteness

- When the algorithm stops, it returns an ε-optimal solution.
- But ... does it stop?
  - ***Theorem*** (convergence). Let

$$\bar{R} = \max_{i,\mathbf{y}} \|\delta\Psi_i(\mathbf{y})\| \qquad \bar{\Delta} = \max_{i,\mathbf{y}} \Delta(\mathbf{y}_i, \mathbf{y})$$

Then the algorithm stops within

$$\frac{C\bar{\Delta}^2\bar{R}^2 + N\bar{\Delta}}{\epsilon^2}$$

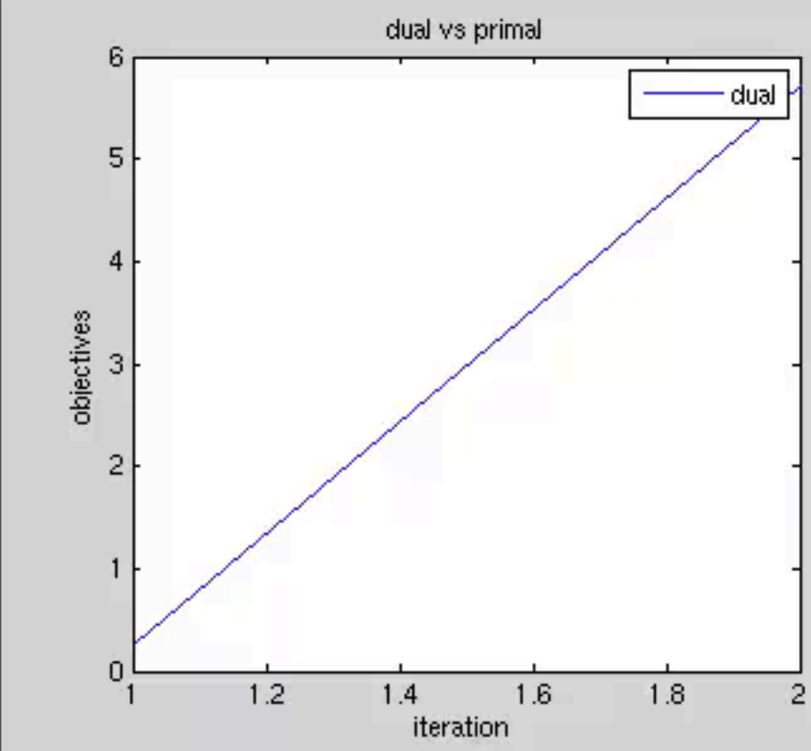
iterations.

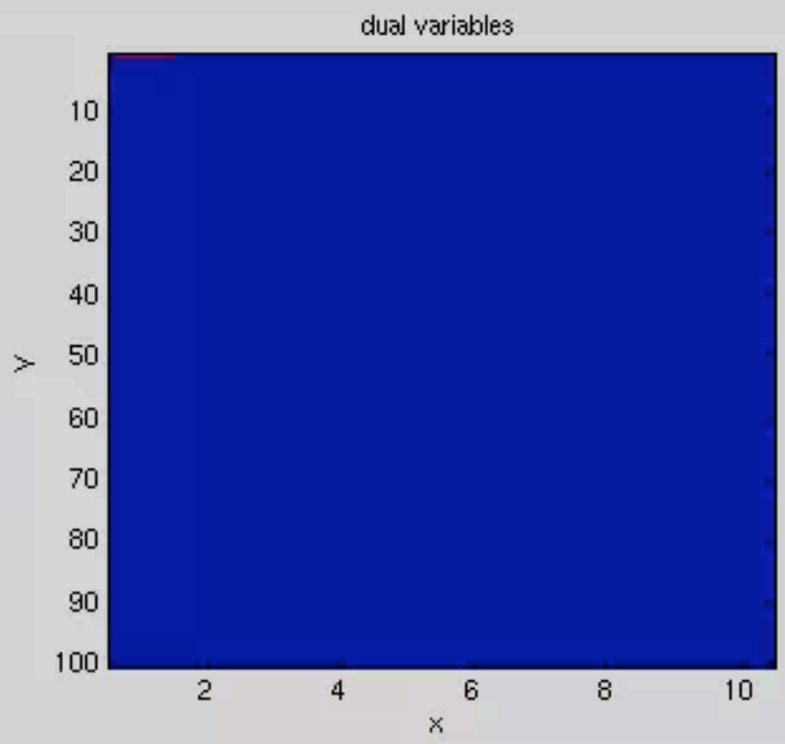
Demo: Fitting a real function

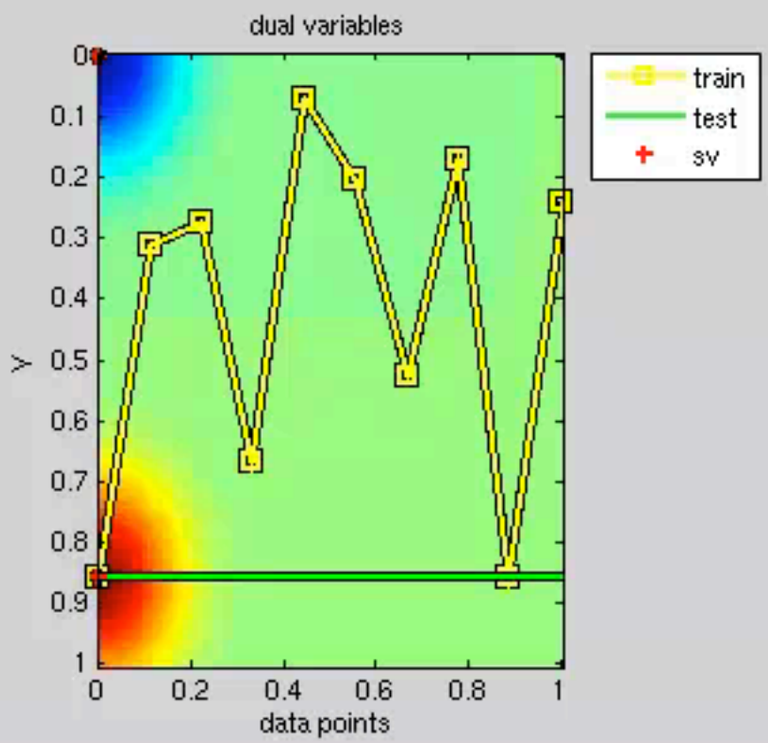
- Hunting constraints to fit a real function

Demo: Fitting a real function

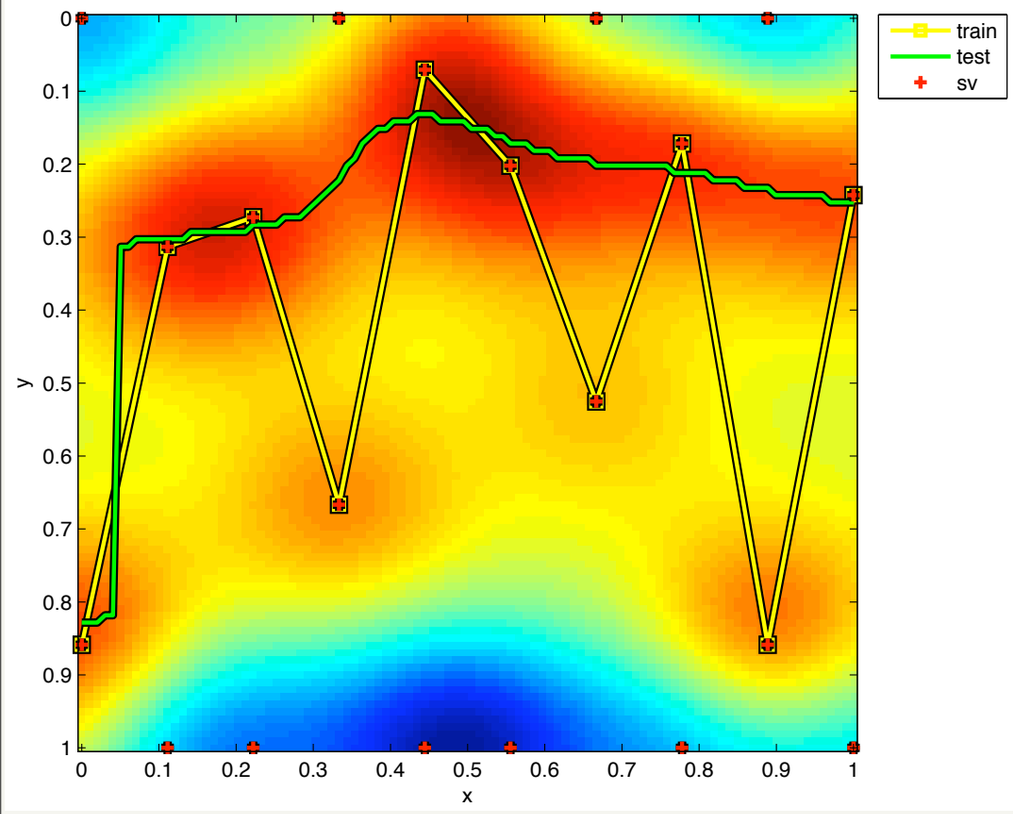
- Hunting constraints to fit a real function





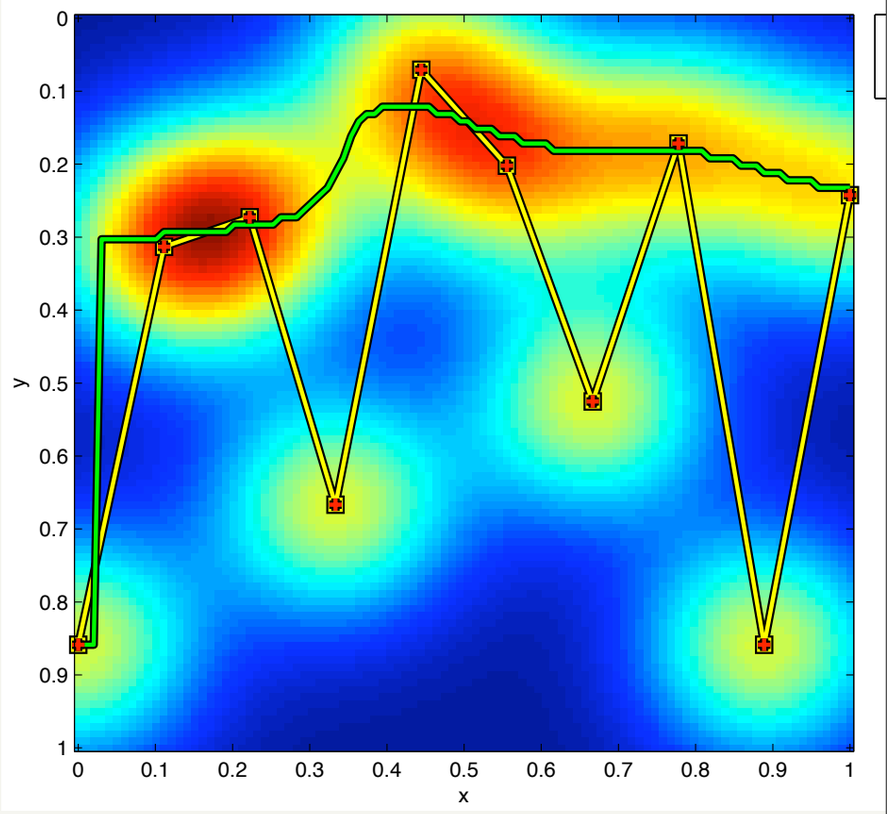


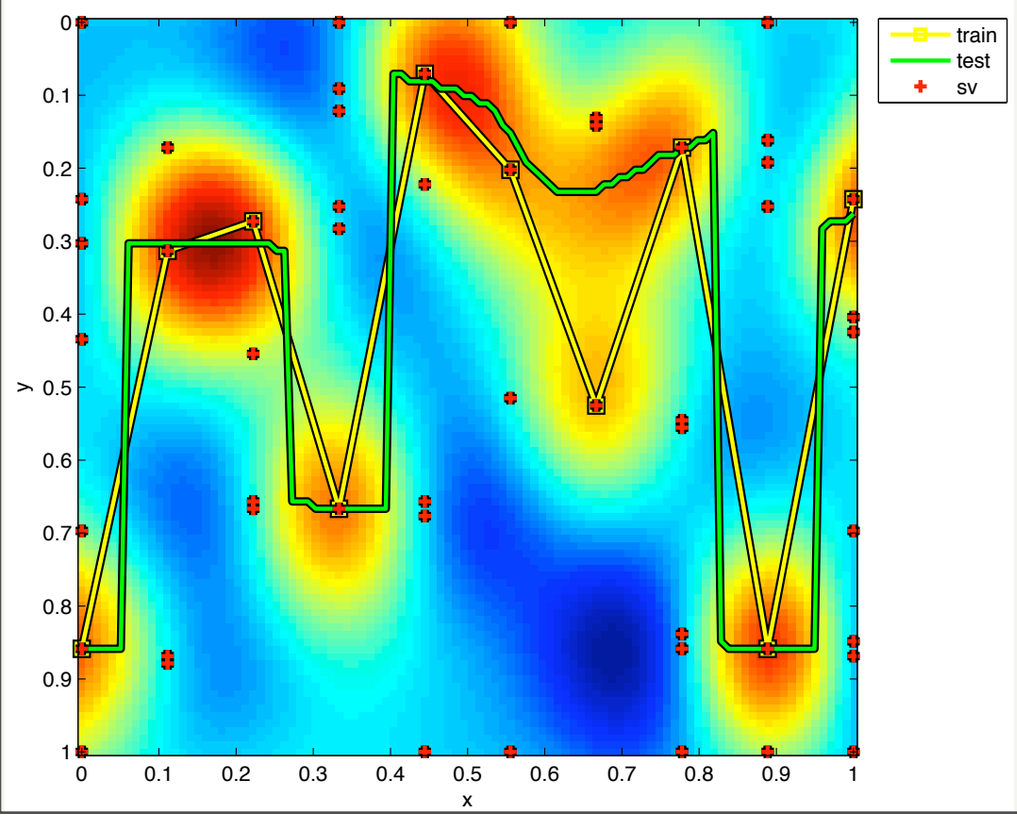
Demo: Effect of C



C=1

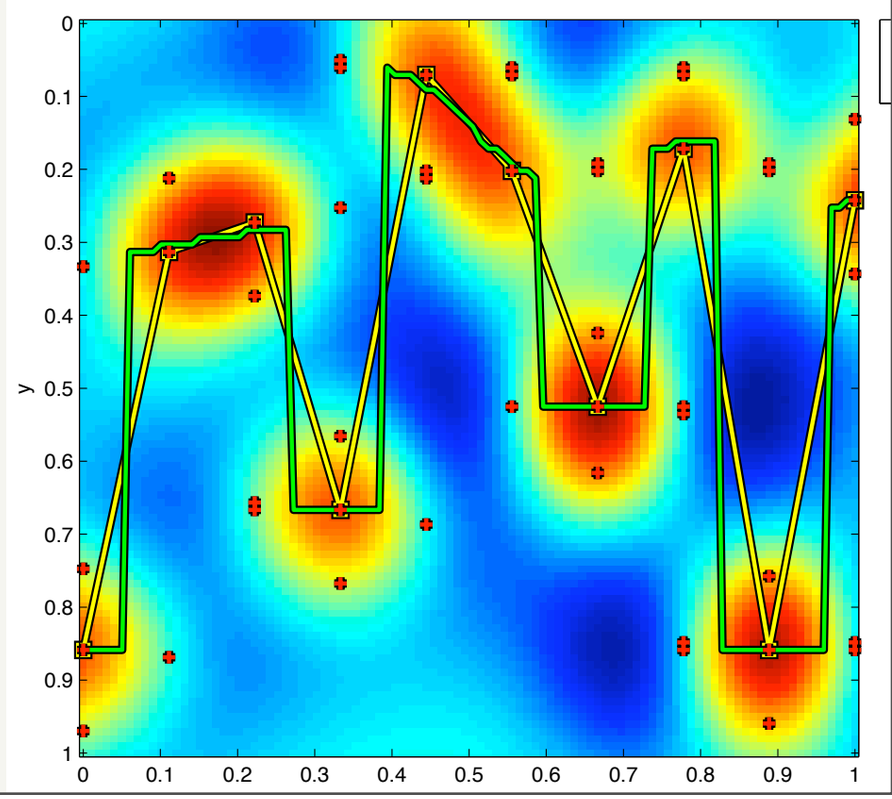
C=10





C=100

C=1000



An Application: Localization

Blaschko Lampert 08

The problem

- **Goal.** Learn a function from an image to a bounding box and label.
  - ***input*** x = image
  - ***output*** y = (label, bounding box)

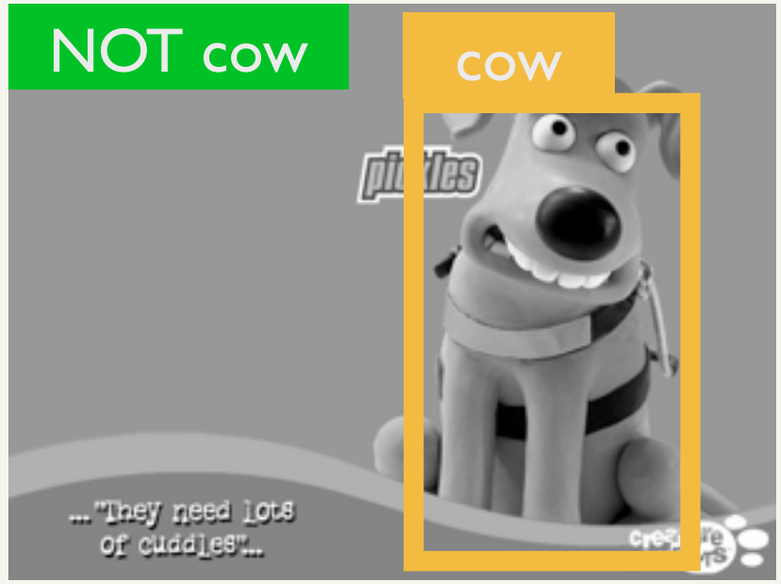
Training data

- Pairs (x_1, y_1), ..., (x_N, y_N) :

  (image, label + bounding box)

Loss

$\Delta(\mathbf{y}, \hat{\mathbf{y}})$ = overlap err.

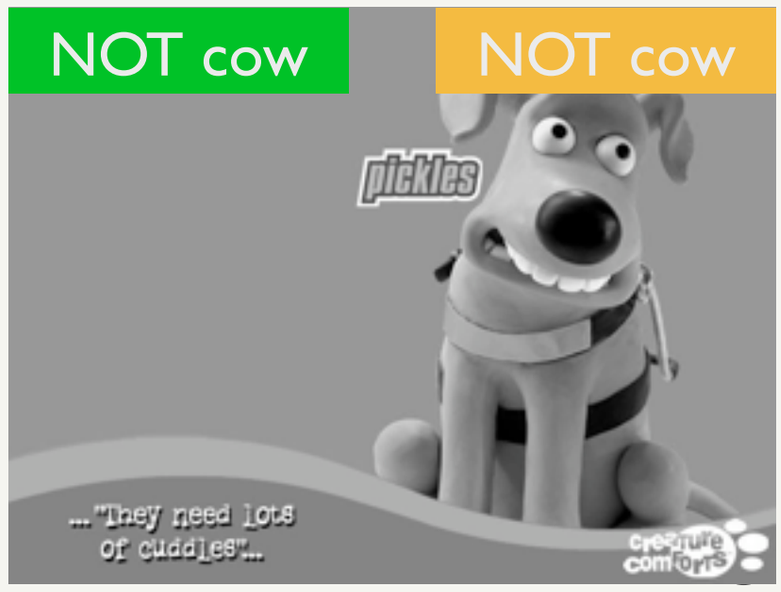


$\Delta(\mathbf{y}, \hat{\mathbf{y}}) = 1$

$$\Delta(\mathbf{y}, \hat{\mathbf{y}})$$

$\Delta(\mathbf{y}, \hat{\mathbf{y}}) = 1$

$\Delta(\mathbf{y}, \hat{\mathbf{y}}) = 0$



Kernel

Kernel

- Intuition:
  $x \approx x'$ and $y \approx y'$ then $k((x,y), (x', y'))$ large

Kernel

- Intuition:
  $x \approx x'$ and $y \approx y'$ then $k((x,y),(x',y'))$ large

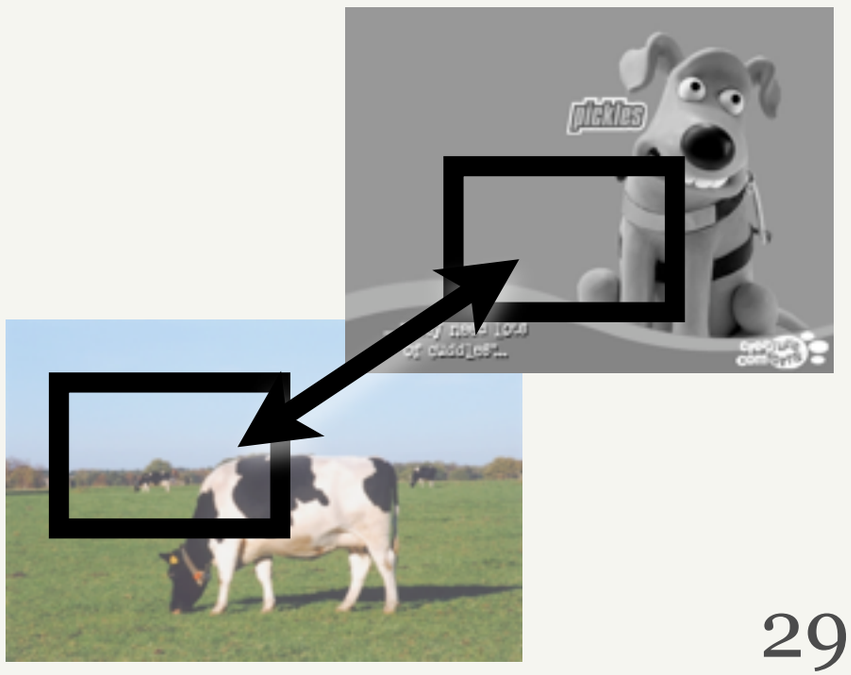
$$k((x,y),(x',y')) = \begin{cases} 0 & \text{if } \mathrm{lab}(y) = \mathrm{NOT} \\ & \text{or } \mathrm{lab}(y') = \mathrm{NOT} \end{cases}$$

Kernel

- Intuition:
  $x \approx x'$ and $y \approx y'$ then $k((x,y), (x', y'))$ large

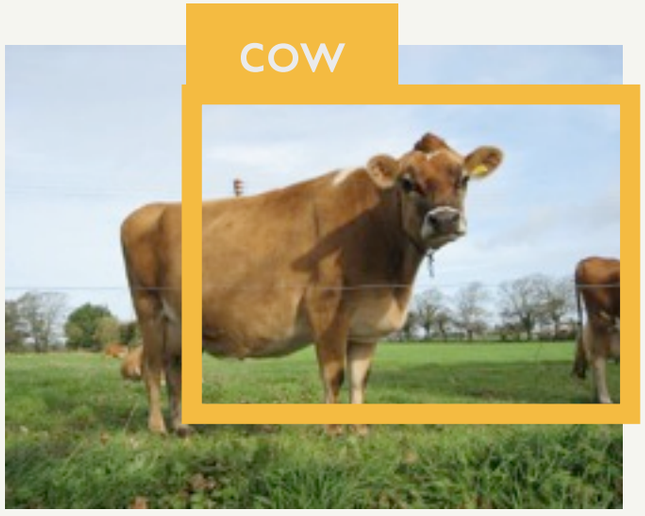
$$k((x,y), (x', y')) = \begin{cases} 0 & \text{if } \mathrm{lab}(y) = \mathrm{NOT} \text{ or } \mathrm{lab}(y') = \mathrm{NOT} \\ k(crop(x|y), crop(x'|y')) & \text{if } \mathrm{lab}(y) = \mathrm{lab}(y') = \mathrm{COW} \end{cases}$$

"image patch similarity"



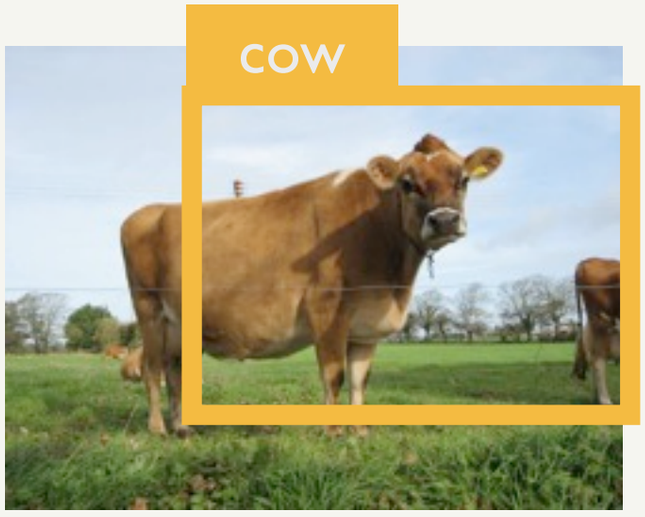
SVM evaluation

Hypothesis



SVM evaluation

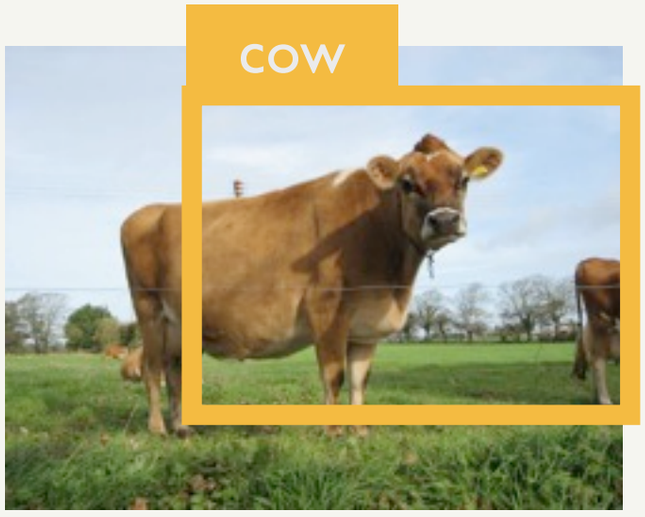
Hypothesis



Scoring

SVM evaluation

Hypothesis



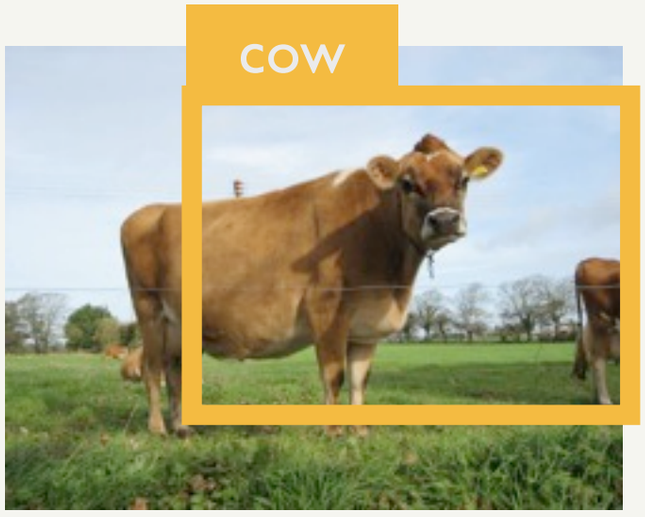
Scoring

$$F(\boldsymbol{x},\boldsymbol{y}|w) = 0$$

"do nothing"

SVM evaluation

Hypothesis



Scoring

$F(\boldsymbol{x},\boldsymbol{y}|w) = 0$

"do nothing"

"match SVs"

SVM evaluation

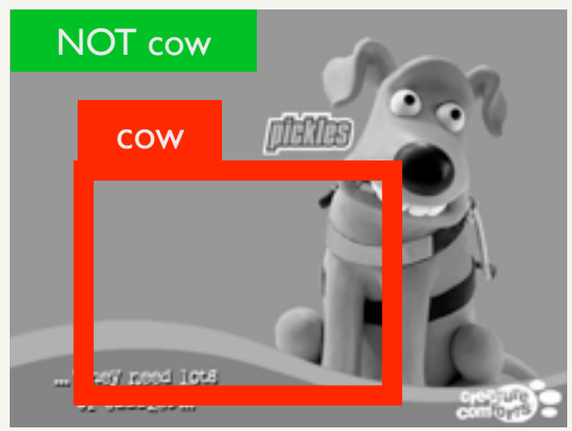
Hypothesis

Scoring

$F(\boldsymbol{x},\boldsymbol{y}|w) = 0$

"do nothing"

"match SVs"



SVM evaluation

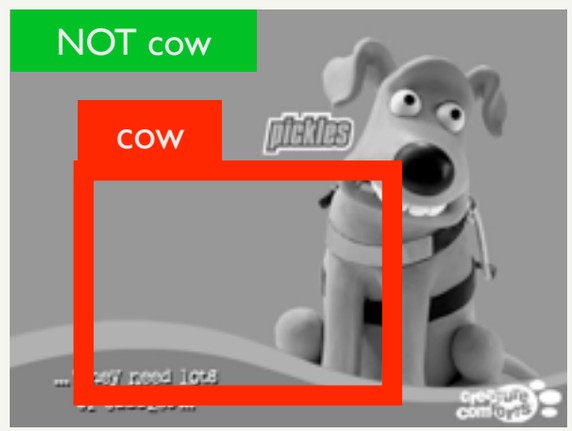
Hypothesis

Scoring

$F(\boldsymbol{x},\boldsymbol{y}|w) = 0$

"do nothing"

"match SVs"



Return best hypothesis

Most violated constraints

$$max\{slack \times \Delta(y,y')\}$$

Most violated constraints

$$max\{slack \times \Delta(y,y')\}$$

new constraint

Most violated constraints

$$max\{slack \times \Delta(y,y')\}$$

new constraint

- Finding most violated constraints
  - similar to evaluating the SVM
  - scoring weighed by $\Delta(y,y')$
  - repeated many times
    - needs a fast search method (e.g. branch and bound)

M^3 networks

Taskar Guestrin Koller 03

M³ network

M³ network

- Structured SVM entails a large number of constraints
  - So far, handled by adding one constraint per time

M^3 network

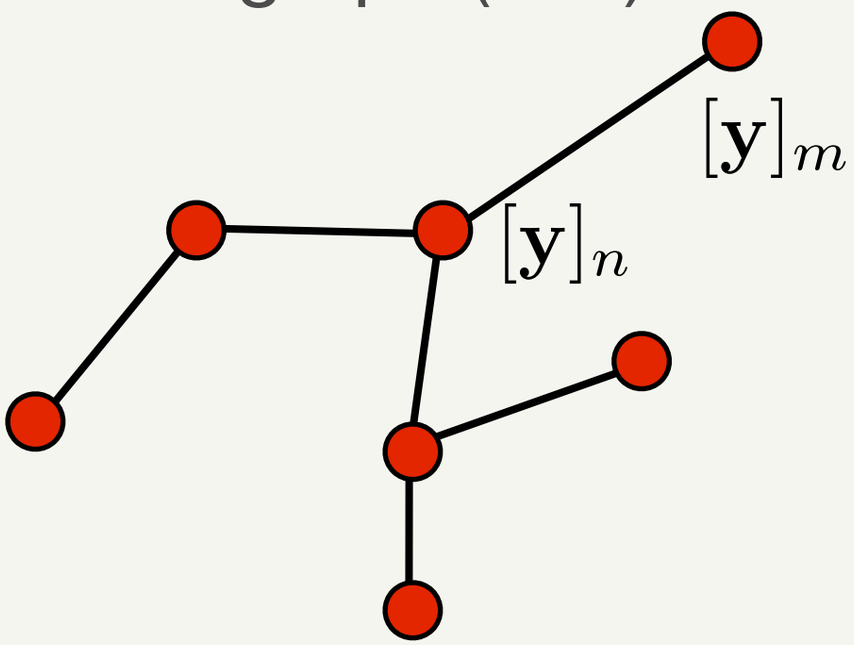
- Structured SVM entails a large number of constraints
  - So far, handled by adding one constraint per time
- ***M^3 network*** is a graphical model
  - Markov network (encodes indep. relations)
  - Inference in structured output framework
  - Reduction of exp to poly number of constraints

M^3 network

- Structured SVM entails a large number of constraints
  - So far, handled by adding one constraint per time
- ***M^3 network*** is a graphical model
  - Markov network (encodes indep. relations)
  - Inference in structured output framework
  - Reduction of exp to poly number of constraints
- *M^3 network* = Max-Margin Markov network

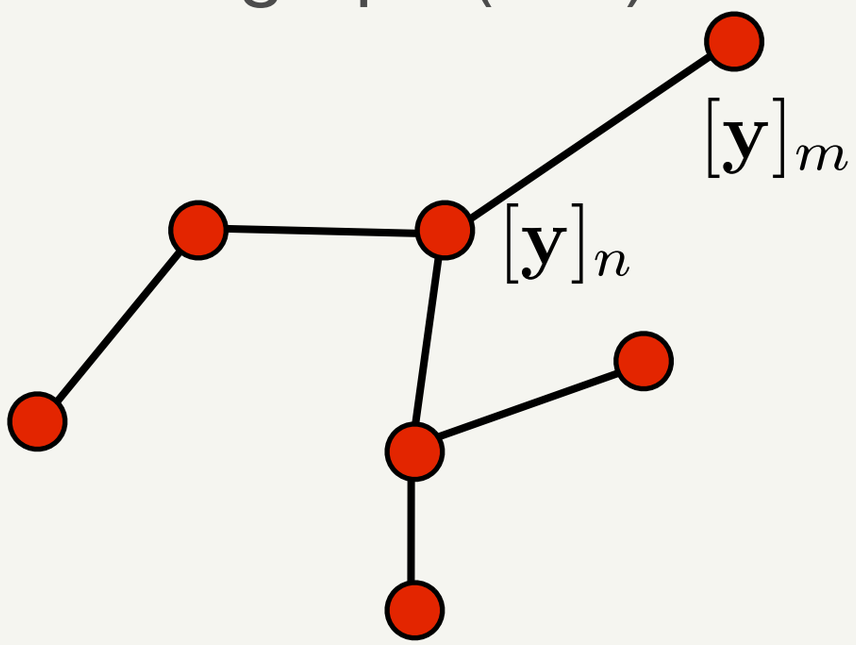
M^3 Model

graph (tree)



M^3 Model

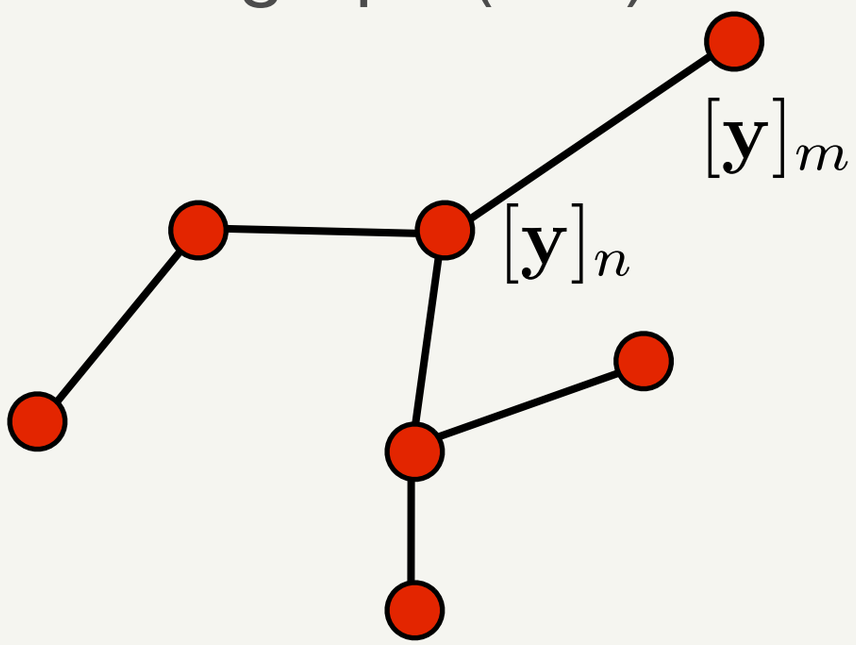
graph (tree)



State $\mathbf{y} = \begin{bmatrix} 1 & 1 & 0 & 1 & \dots & 0 \end{bmatrix}$

M^3 Model

graph (tree)

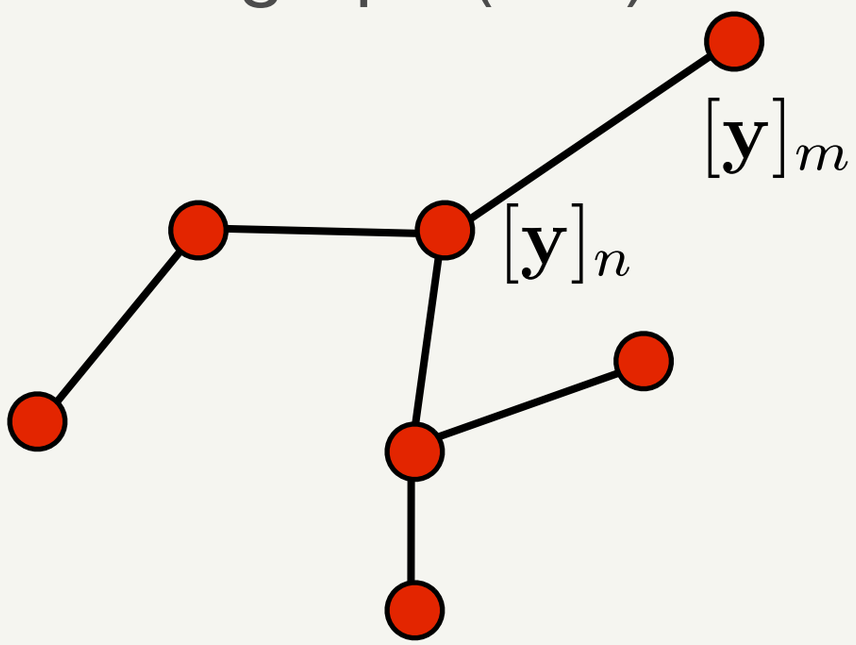


State $\mathbf{y} = \begin{bmatrix} 1 & 1 & 0 & 1 & \dots & 0 \end{bmatrix}$

Discriminative model $F(\mathbf{x}, \mathbf{y}|w) \propto P(\mathbf{y}|\mathbf{x}, w)$

M^3 Model

graph (tree)



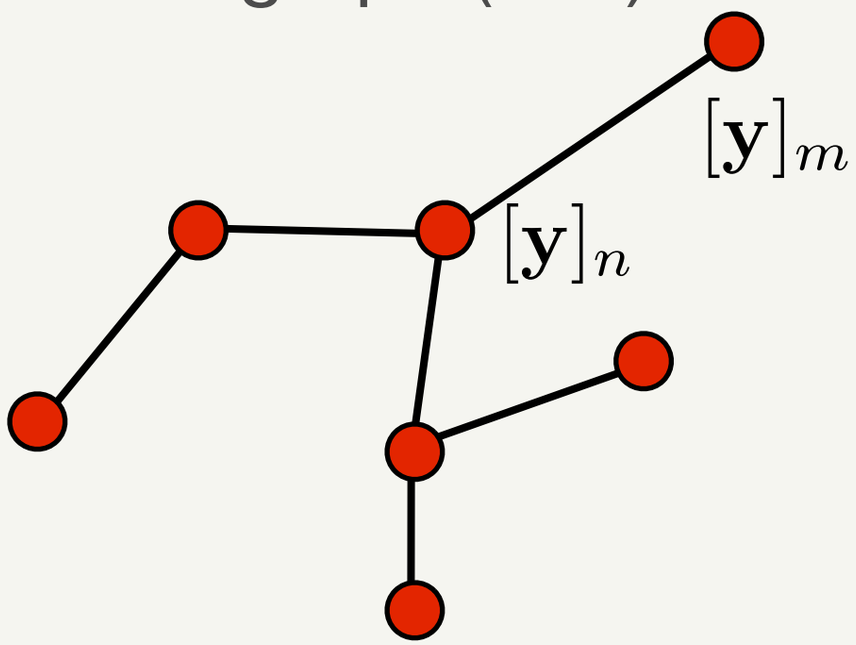
State $\mathbf{y} = \begin{bmatrix} 1 & 1 & 0 & 1 & \dots & 0 \end{bmatrix}$

Discriminative model $F(\mathbf{x}, \mathbf{y}|w) \propto P(\mathbf{y}|\mathbf{x}, w)$

Inference $f(\mathbf{x}|w) = \underset{\mathbf{y} \in \{0,1\}^N}{\operatorname{argmax}} F(\mathbf{x}, \mathbf{y}|w)$

M^3 Model

graph (tree)



State $\mathbf{y} = \begin{bmatrix} 1 & 1 & 0 & 1 & \dots & 0 \end{bmatrix}$

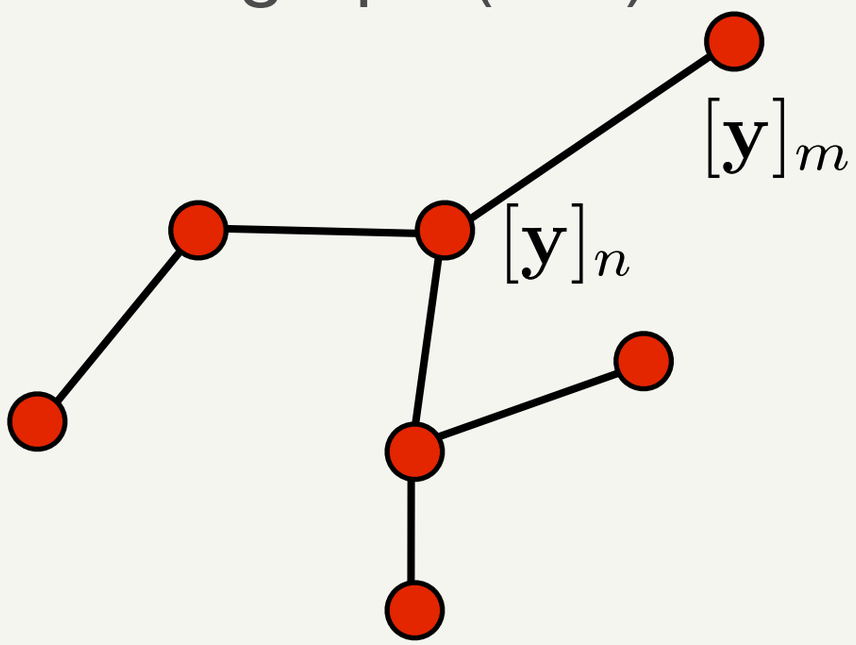
Discriminative model $F(\mathbf{x}, \mathbf{y}|w) \propto P(\mathbf{y}|\mathbf{x}, w)$

Inference $f(\mathbf{x}|w) = \underset{\mathbf{y} \in \{0,1\}^N}{\operatorname{argmax}} F(\mathbf{x}, \mathbf{y}|w)$

Risk $R(w) = \frac{1}{N} \sum_{i=1}^{N} \Delta(\mathbf{y}_i, f(\mathbf{x}_i|w))$

M^3 Model

graph (tree)



State $\mathbf{y} = \begin{bmatrix} 1 & 1 & 0 & 1 & \dots & 0 \end{bmatrix}$

Discriminative model $F(\mathbf{x}, \mathbf{y}|w) \propto P(\mathbf{y}|\mathbf{x}, w)$

Inference $f(\mathbf{x}|w) = \underset{\mathbf{y} \in \{0,1\}^N}{\operatorname{argmax}} F(\mathbf{x}, \mathbf{y}|w)$

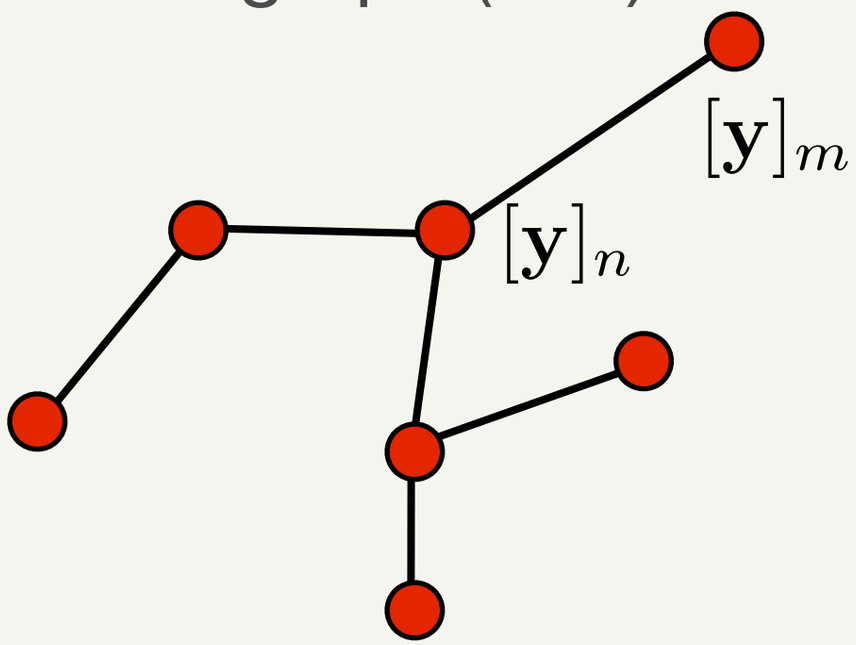
Risk $R(w) = \frac{1}{N} \sum_{i=1}^{N} \Delta(\mathbf{y}_i, f(\mathbf{x}_i|w))$

(1) SVM Property

$F(\mathbf{x}, \mathbf{y}|w) = \langle w, \Psi(\mathbf{x}, \mathbf{y}) \rangle$

M^3 Model

graph (tree)



State $\mathbf{y} = \begin{bmatrix} 1 & 1 & 0 & 1 & \dots & 0 \end{bmatrix}$

Discriminative model $F(\mathbf{x}, \mathbf{y}|w) \propto P(\mathbf{y}|\mathbf{x}, w)$

Inference $f(\mathbf{x}|w) = \underset{\mathbf{y} \in \{0,1\}^N}{\operatorname{argmax}} F(\mathbf{x}, \mathbf{y}|w)$

Risk $R(w) = \frac{1}{N} \sum_{i=1}^{N} \Delta(\mathbf{y}_i, f(\mathbf{x}_i|w))$

(1) SVM Property

$$F(\mathbf{x}, \mathbf{y}|w) = \langle w, \Psi(\mathbf{x}, \mathbf{y}) \rangle$$

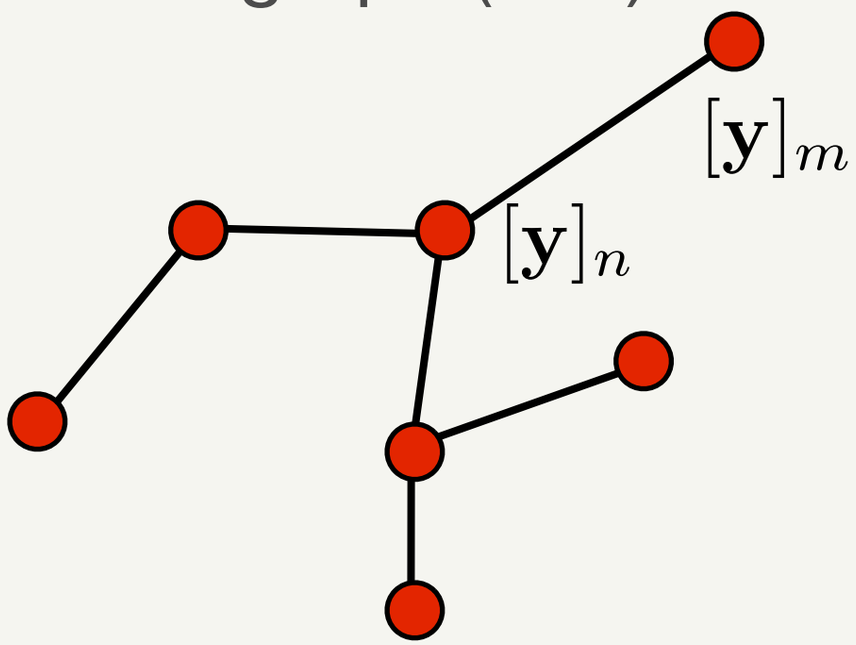
(2) Markov Property

$$\Delta(\mathbf{y}, \hat{\mathbf{y}}) = \sum_n \Delta_1([\mathbf{y}]_n, [\hat{\mathbf{y}}]_n)$$

$$\Psi(\mathbf{x}, \mathbf{y}) = \sum_{mn} \Psi_{mn}([\mathbf{y}]_n, [\mathbf{y}]_m)$$

M^3 Model

graph (tree)



State $\mathbf{y} = \begin{bmatrix} 1 & 1 & 0 & 1 & \dots & 0 \end{bmatrix}$

Discriminative model $F(\mathbf{x}, \mathbf{y}|w) \propto P(\mathbf{y}|\mathbf{x}, w)$

Inference $f(\mathbf{x}|w) = \underset{\mathbf{y} \in \{0,1\}^N}{\operatorname{argmax}} F(\mathbf{x}, \mathbf{y}|w)$

Risk $R(w) = \frac{1}{N} \sum_{i=1}^{N} \Delta(\mathbf{y}_i, f(\mathbf{x}_i|w))$

Node n state $[\mathbf{y}]_n$

(1) SVM Property

$$F(\mathbf{x}, \mathbf{y}|w) = \langle w, \Psi(\mathbf{x}, \mathbf{y}) \rangle$$

(2) Markov Property

$$\Delta(\mathbf{y}, \hat{\mathbf{y}}) = \sum_n \Delta_1([\mathbf{y}]_n, [\hat{\mathbf{y}}]_n)$$

$$\Psi(\mathbf{x}, \mathbf{y}) = \sum_{mn} \Psi_{mn}([\mathbf{y}]_n, [\mathbf{y}]_m)$$

M^3 as Structured SVM

M^3 as Structured SVM

- M^3 is a structured SVM

M^3 as Structured SVM

- M^3 is a structured SVM
  - usual dual problem:

$$\begin{aligned} &\max \mathbf{1}^\top \alpha - \tfrac{1}{2}\alpha^\top E^\top K E \alpha \\ &\alpha \geq 0, \\ &A\alpha \leq \tfrac{C}{N}\mathbf{1} \end{aligned}$$

M^3 as Structured SVM

- M^3 is a structured SVM
  - usual dual problem:

$$\begin{aligned} &\max \mathbf{1}^\top \alpha - \tfrac{1}{2}\alpha^\top E^\top K E \alpha \\ &\alpha \geq 0, \\ &A\alpha \leq \tfrac{C}{N}\mathbf{1} \end{aligned}$$

- How many dual variables?

M^3 as Structured SVM

- M^3 is a structured SVM
  - usual dual problem:

$$\begin{aligned} &\max \mathbf{1}^\top \alpha - \tfrac{1}{2}\alpha^\top E^\top K E \alpha \\ &\alpha \geq 0, \\ &A\alpha \leq \tfrac{C}{N}\mathbf{1} \end{aligned}$$

- How many dual variables?
  - $\alpha = \begin{bmatrix} \dots & \alpha_{i\mathbf{y}} & \dots \end{bmatrix}$

M^3 as Structured SVM

- M^3 is a structured SVM
  - usual dual problem:

$$\begin{aligned} &\max \mathbf{1}^\top \alpha - \tfrac{1}{2}\alpha^\top E^\top K E \alpha \\ &\alpha \geq 0, \\ &A\alpha \leq \tfrac{C}{N}\mathbf{1} \end{aligned}$$

- How many dual variables?
  - $\alpha = \begin{bmatrix} \dots & \alpha_{i\mathbf{y}} & \dots \end{bmatrix}$
  - *i* ranges over number of examples (*N*)

M^3 as Structured SVM

- M^3 is a structured SVM
  - usual dual problem:

$$\begin{aligned} &\max \mathbf{1}^\top \alpha - \tfrac{1}{2}\alpha^\top E^\top K E \alpha \\ &\alpha \geq 0, \\ &A\alpha \leq \tfrac{C}{N}\mathbf{1} \end{aligned}$$

- How many dual variables?
  - $\alpha = \begin{bmatrix} \dots & \alpha_{i\mathbf{y}} & \dots \end{bmatrix}$
  - *i* ranges over number of examples (*N*)
  - ***y*** ranges over number of settings (2^M)

M^3 as Structured SVM

- M^3 is a structured SVM
  - usual dual problem:

$$\begin{aligned} &\max \mathbf{1}^\top \alpha - \tfrac{1}{2}\alpha^\top E^\top K E \alpha \\ &\alpha \geq 0, \\ &A\alpha \leq \tfrac{C}{N}\mathbf{1} \end{aligned}$$

- How many dual variables?
  - $\alpha = \begin{bmatrix} \dots & \alpha_{i\mathbf{y}} & \dots \end{bmatrix}$
  - i ranges over number of examples (N)
  - $\mathbf{y}$ ranges over number of settings (2^M)
  - total: $N \times 2^M$

Marginalizing dual variables

$$\max \mathbf{1}^\top \alpha - \tfrac{1}{2}\alpha^\top E^\top K E \alpha$$
$$\alpha \geq 0,$$
$$A\alpha \leq \tfrac{C}{N}\mathbf{1}$$

Marginalizing dual variables

$$\max \mathbf{1}^\top \alpha - \tfrac{1}{2}\alpha^\top E^\top K E \alpha$$
$$\alpha \geq 0,$$
$$A\alpha \leq \tfrac{C}{N}\mathbf{1}$$

$$A = \begin{bmatrix} \dots & \Delta(\mathbf{y}_1\mathbf{y})^{-1} & \dots & & \dots 0 \dots \\ & & \ddots & & \\ & \dots 0 \dots & & \dots & \Delta(\mathbf{y}_N\mathbf{y})^{-1} & \dots \end{bmatrix}$$

Marginalizing dual variables

$$\max \mathbf{1}^\top \alpha - \tfrac{1}{2}\alpha^\top E^\top K E \alpha$$
$$\alpha \geq 0,$$
$$A\alpha \leq \tfrac{C}{N}\mathbf{1}$$

$$A = \begin{bmatrix} \dots & \Delta(\mathbf{y}_1\mathbf{y})^{-1} & \dots & \dots 0 \dots \\ & & \ddots & \\ \dots 0 \dots & & \dots & \Delta(\mathbf{y}_N\mathbf{y})^{-1} & \dots \end{bmatrix}$$

$$\beta_{i\mathbf{y}} = \begin{cases} \frac{\alpha_{i\mathbf{y}}}{\Delta(\mathbf{y}_i,\mathbf{y})}, & \mathbf{y} \neq \mathbf{y}_i, \\ \frac{C}{N} - \sum_{\mathbf{y}' \neq \mathbf{y}_i} \frac{\alpha_{i\mathbf{y}'}}{\Delta(\mathbf{y}_i,\mathbf{y}')}, & \mathbf{y} = \mathbf{y}_i. \end{cases}$$

Marginalizing dual variables

$$\begin{aligned}\max\ & \mathbf{1}^\top\alpha - \tfrac{1}{2}\alpha^\top E^\top K E\alpha \\ \alpha &\geq 0, \\ A\alpha &\leq \tfrac{C}{N}\mathbf{1}\end{aligned}$$

$$A = \begin{bmatrix} \dots\ \Delta(\mathbf{y}_1\mathbf{y})^{-1}\ \dots & & \dots 0 \dots \\ & \ddots & \\ \dots 0 \dots & & \dots\ \Delta(\mathbf{y}_N\mathbf{y})^{-1}\ \dots \end{bmatrix}$$

$$\beta_{i\mathbf{y}} = \begin{cases} \frac{\alpha_{i\mathbf{y}}}{\Delta(\mathbf{y}_i,\mathbf{y})}, & \mathbf{y} \neq \mathbf{y}_i, \\ \frac{C}{N} - \sum_{\mathbf{y}'\neq\mathbf{y}_i} \frac{\alpha_{i\mathbf{y}'}}{\Delta(\mathbf{y}_i,\mathbf{y}')}, & \mathbf{y} = \mathbf{y}_i. \end{cases}$$

$$\begin{aligned}\mathbf{1}^\top\alpha &= \sum_i \sum_{\mathbf{y}} \Delta(\mathbf{y}_i,\mathbf{y})\beta_{i\mathbf{y}} \\ &= \sum_i \sum_{\mathbf{y}} \sum_n \Delta([\mathbf{y}_i]_n,[\mathbf{y}]_n)\beta_{i\mathbf{y}} \\ &= \sum_i \sum_n \sum_{\mathbf{y}} \Delta([\mathbf{y}_i]_n,[\mathbf{y}]_n)\beta_{i\mathbf{y}} \\ &= \sum_i \sum_n \sum_{[\mathbf{y}]_n} \Delta([\mathbf{y}_i]_n,[\mathbf{y}]_n)\mu_{i[\mathbf{y}]_n}\end{aligned}$$

Marginalizing dual variables

$$\max \mathbf{1}^\top \alpha - \tfrac{1}{2}\alpha^\top E^\top K E \alpha$$
$$\alpha \geq 0,$$
$$A\alpha \leq \tfrac{C}{N}\mathbf{1}$$

$$A = \begin{bmatrix} \dots & \Delta(\mathbf{y}_1\mathbf{y})^{-1} & \dots & \dots 0 \dots \\ & & \ddots & \\ \dots 0 \dots & & \dots & \Delta(\mathbf{y}_N\mathbf{y})^{-1} & \dots \end{bmatrix}$$

$$\beta_{i\mathbf{y}} = \begin{cases} \frac{\alpha_{i\mathbf{y}}}{\Delta(\mathbf{y}_i,\mathbf{y})}, & \mathbf{y} \neq \mathbf{y}_i, \\ \frac{C}{N} - \sum_{\mathbf{y}' \neq \mathbf{y}_i} \frac{\alpha_{i\mathbf{y}'}}{\Delta(\mathbf{y}_i,\mathbf{y}')}, & \mathbf{y} = \mathbf{y}_i. \end{cases}$$

$$\begin{aligned} \mathbf{1}^\top \alpha &= \sum_i \sum_{\mathbf{y}} \Delta(\mathbf{y}_i, \mathbf{y}) \beta_{i\mathbf{y}} \\ &= \sum_i \sum_{\mathbf{y}} \sum_n \Delta([\mathbf{y}_i]_n, [\mathbf{y}]_n) \beta_{i\mathbf{y}} \\ &= \sum_i \sum_n \sum_{\mathbf{y}} \Delta([\mathbf{y}_i]_n, [\mathbf{y}]_n) \beta_{i\mathbf{y}} \\ &= \sum_i \sum_n \sum_{[\mathbf{y}]_n} \Delta([\mathbf{y}_i]_n, [\mathbf{y}]_n) \mu_{i[\mathbf{y}]_n} \end{aligned}$$

$$\mu_{i[\mathbf{y}]_n} = \sum_{\mathbf{y} \sim [\mathbf{y}]_n} \beta_{i\mathbf{y}}$$

$\beta \rightarrow$ marginals

$\beta \rightarrow$ marginals

- Given β

$\beta \rightarrow$ marginals

- Given $\boldsymbol{\beta}$
  - compute $1\times$ and $2\times$ marginal variables

$$\mu_{i[\mathbf{y}]_n} = \sum_{\mathbf{y}\sim[\mathbf{y}]_n} \beta_{i\mathbf{y}} \qquad \mu_{i[\mathbf{y}]_{mn}} = \sum_{\mathbf{y}\sim[\mathbf{y}]_{mn}} \beta_{i\mathbf{y}}$$

$\beta \rightarrow$ marginals

- Given $\boldsymbol{\beta}$
  - compute $1\times$ and $2\times$ marginal variables

$$\mu_{i[\mathbf{y}]_n} = \sum_{\mathbf{y} \sim [\mathbf{y}]_n} \beta_{i\mathbf{y}} \qquad \mu_{i[\mathbf{y}]_{mn}} = \sum_{\mathbf{y} \sim [\mathbf{y}]_{mn}} \beta_{i\mathbf{y}}$$

  - compute objective as function of marginals

$$\max \mathbf{1}^\top \alpha - \tfrac{1}{2}\alpha^\top E^\top K E \alpha$$
$$\alpha \geq 0,$$
$$A\alpha \leq \tfrac{C}{N}\mathbf{1}$$

$$\sum_i \sum_n \sum_{[\mathbf{y}]_n} \Delta([\mathbf{y}_i]_n, [\mathbf{y}]_n) \mu_{i[\mathbf{y}]_n}$$

$$\sum_{ij} \sum_{mnpq} \sum_{[\mathbf{y}]_{mn}[\mathbf{y}']_{pq}} B_{i[\mathbf{y}]_{mn}j[\mathbf{y}]_{pq}} \mu_{i[\mathbf{y}]_{mn}} \mu_{j[\mathbf{y}']_{pq}}$$

Marginals $\to \beta$

$$\mu_{i[\mathbf{y}]_n} = \sum_{\mathbf{y} \sim [\mathbf{y}]_n} \beta_{i\mathbf{y}} \qquad \mu_{i[\mathbf{y}]_{mn}} = \sum_{\mathbf{y} \sim [\mathbf{y}]_{mn}} \beta_{i\mathbf{y}}$$

Marginals $\rightarrow \beta$

- Given *1* $\times$ and *2* $\times$ marginal variables

$$\mu_{i[\mathbf{y}]_n} = \sum_{\mathbf{y} \sim [\mathbf{y}]_n} \beta_{i\mathbf{y}} \qquad \mu_{i[\mathbf{y}]_{mn}} = \sum_{\mathbf{y} \sim [\mathbf{y}]_{mn}} \beta_{i\mathbf{y}}$$

  is there a corresponding $\boldsymbol{\beta}$?

Marginals $\rightarrow \beta$

- Given *1* $\times$ and *2* $\times$ marginal variables

$$\mu_{i[\mathbf{y}]_n} = \sum_{\mathbf{y} \sim [\mathbf{y}]_n} \beta_{i\mathbf{y}} \qquad \mu_{i[\mathbf{y}]_{mn}} = \sum_{\mathbf{y} \sim [\mathbf{y}]_{mn}} \beta_{i\mathbf{y}}$$

  is there a corresponding $\boldsymbol{\beta}$?

- We also know

$$\sum_{\mathbf{y}} \beta_{i\mathbf{y}} = \frac{C}{N} \qquad \beta_{i\mathbf{y}} \geq 0$$

Marginals $\rightarrow \beta$

- Given *1* $\times$ and *2* $\times$ marginal variables

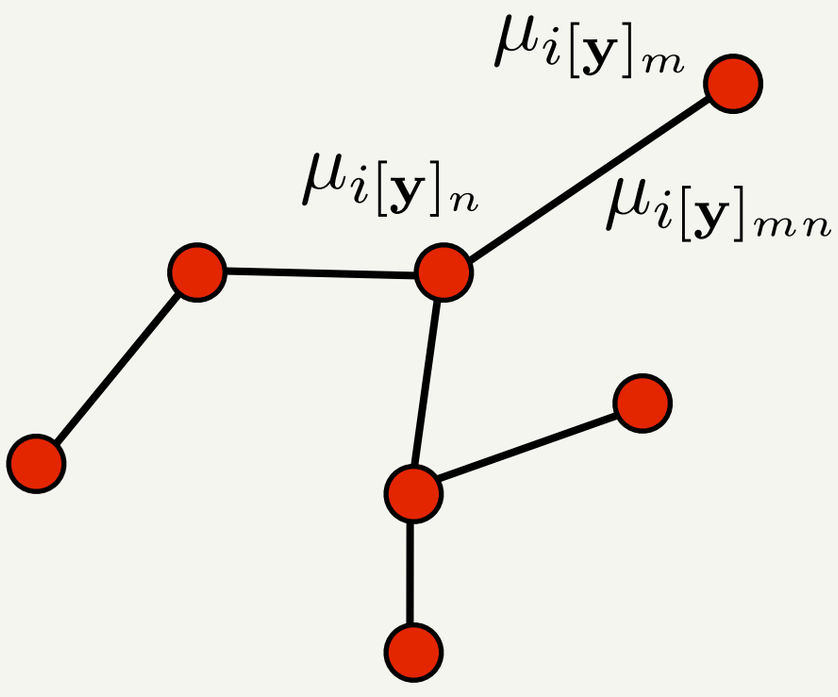
$$\mu_{i[\mathbf{y}]_n} = \sum_{\mathbf{y} \sim [\mathbf{y}]_n} \beta_{i\mathbf{y}} \qquad \mu_{i[\mathbf{y}]_{mn}} = \sum_{\mathbf{y} \sim [\mathbf{y}]_{mn}} \beta_{i\mathbf{y}}$$

  is there a corresponding $\boldsymbol{\beta}$?

- We also know

$$\sum_{\mathbf{y}} \beta_{i\mathbf{y}} = \frac{C}{N} \qquad \beta_{i\mathbf{y}} \geq 0$$

- A sufficient condition



$$\sum_{[\mathbf{y}]_m} \mu_{i[\mathbf{y}]_{mn}} = \mu_{i[\mathbf{y}]_n}$$

$$\sum_{[\mathbf{y}]_n} \mu_{i[\mathbf{y}]_n} = \frac{C}{N}$$

$$\mu_{i[\mathbf{y}]_{mn}} \geq 0$$

Recap

$$\max \mathbf{1}^\top \alpha - \tfrac{1}{2}\alpha^\top E^\top K E \alpha$$
$$\alpha \geq 0,$$
$$A\alpha \leq \tfrac{C}{N}\mathbf{1}$$

$$\sum_i \sum_n \sum_{[\mathbf{y}]_n} \Delta([\mathbf{y}_i]_n, [\mathbf{y}]_n) \mu_{i[\mathbf{y}]_n}$$

$$\sum_{ij} \sum_{mnpq} \sum_{[\mathbf{y}]_{mn}[\mathbf{y}']_{pq}} B_{i[\mathbf{y}]_{mn}j[\mathbf{y}]_{pq}} \mu_{i[\mathbf{y}]_{mn}} \mu_{j[\mathbf{y}']_{pq}}$$

$$\sum_{[\mathbf{y}]_m} \mu_{i[\mathbf{y}]_{mn}} = \mu_{i[\mathbf{y}]_n} \qquad \sum_{[\mathbf{y}]_n} \mu_{i[\mathbf{y}]_n} = \frac{C}{N}$$

$$\mu_{i[\mathbf{y}]_{mn}} \geq 0$$

Recap

$$\max \mathbf{1}^\top \alpha - \tfrac{1}{2}\alpha^\top E^\top K E \alpha$$
$$\alpha \geq 0,$$
$$A\alpha \leq \tfrac{C}{N}\mathbf{1}$$

$$\sum_i \sum_n \sum_{[\mathbf{y}]_n} \Delta([\mathbf{y}_i]_n, [\mathbf{y}]_n) \mu_{i[\mathbf{y}]_n}$$

$$\sum_{ij} \sum_{mnpq} \sum_{[\mathbf{y}]_{mn}[\mathbf{y}']_{pq}} B_{i[\mathbf{y}]_{mn}j[\mathbf{y}]_{pq}} \mu_{i[\mathbf{y}]_{mn}} \mu_{j[\mathbf{y}']_{pq}}$$

$$\sum_{[\mathbf{y}]_m} \mu_{i[\mathbf{y}]_{mn}} = \mu_{i[\mathbf{y}]_n} \qquad \sum_{[\mathbf{y}]_n} \mu_{i[\mathbf{y}]_n} = \frac{C}{N}$$

$$\mu_{i[\mathbf{y}]_{mn}} \geq 0$$

- Simplification results:

Recap

$$\max \mathbf{1}^\top \alpha - \tfrac{1}{2}\alpha^\top E^\top K E \alpha$$
$$\alpha \geq 0,$$
$$A\alpha \leq \tfrac{C}{N}\mathbf{1}$$

$$\sum_i \sum_n \sum_{[\mathbf{y}]_n} \Delta([\mathbf{y}_i]_n, [\mathbf{y}]_n) \mu_{i[\mathbf{y}]_n}$$

$$\sum_{ij} \sum_{mnpq} \sum_{[\mathbf{y}]_{mn}[\mathbf{y}']_{pq}} B_{i[\mathbf{y}]_{mn}j[\mathbf{y}]_{pq}} \mu_{i[\mathbf{y}]_{mn}} \mu_{j[\mathbf{y}']_{pq}}$$

$$\sum_{[\mathbf{y}]_m} \mu_{i[\mathbf{y}]_{mn}} = \mu_{i[\mathbf{y}]_n} \qquad \sum_{[\mathbf{y}]_n} \mu_{i[\mathbf{y}]_n} = \frac{C}{N}$$

$$\mu_{i[\mathbf{y}]_{mn}} \geq 0$$

- Simplification results:
  - Potentials decompose on edges

Recap

$$\max \mathbf{1}^\top \alpha - \tfrac{1}{2}\alpha^\top E^\top K E \alpha$$
$$\alpha \geq 0,$$
$$A\alpha \leq \tfrac{C}{N}\mathbf{1}$$

$$\sum_i \sum_n \sum_{[\mathbf{y}]_n} \Delta([\mathbf{y}_i]_n, [\mathbf{y}]_n) \mu_{i[\mathbf{y}]_n}$$

$$\sum_{ij} \sum_{mnpq} \sum_{[\mathbf{y}]_{mn}[\mathbf{y}']_{pq}} B_{i[\mathbf{y}]_{mn}j[\mathbf{y}]_{pq}} \mu_{i[\mathbf{y}]_{mn}} \mu_{j[\mathbf{y}']_{pq}}$$

$$\sum_{[\mathbf{y}]_m} \mu_{i[\mathbf{y}]_{mn}} = \mu_{i[\mathbf{y}]_n} \qquad \sum_{[\mathbf{y}]_n} \mu_{i[\mathbf{y}]_n} = \frac{C}{N}$$

$$\mu_{i[\mathbf{y}]_{mn}} \geq 0$$

- Simplification results:
  - Potentials decompose on edges
  - Risk decomposes on nodes

Recap

$$\max \mathbf{1}^\top \alpha - \tfrac{1}{2}\alpha^\top E^\top K E \alpha$$
$$\alpha \geq 0,$$
$$A\alpha \leq \tfrac{C}{N}\mathbf{1}$$

$$\sum_i \sum_n \sum_{[\mathbf{y}]_n} \Delta([\mathbf{y}_i]_n, [\mathbf{y}]_n) \mu_{i[\mathbf{y}]_n}$$

$$\sum_{ij} \sum_{mnpq} \sum_{[\mathbf{y}]_{mn}[\mathbf{y}']_{pq}} B_{i[\mathbf{y}]_{mn}j[\mathbf{y}]_{pq}} \mu_{i[\mathbf{y}]_{mn}} \mu_{j[\mathbf{y}']_{pq}}$$

$$\sum_{[\mathbf{y}]_m} \mu_{i[\mathbf{y}]_{mn}} = \mu_{i[\mathbf{y}]_n} \qquad \sum_{[\mathbf{y}]_n} \mu_{i[\mathbf{y}]_n} = \frac{C}{N}$$

$$\mu_{i[\mathbf{y}]_{mn}} \geq 0$$

- Simplification results:
  - Potentials decompose on edges
  - Risk decomposes on nodes
  - variables: $N2^M$ down to $N(M^2 + M)$

Recap

$$\max \mathbf{1}^\top \alpha - \tfrac{1}{2}\alpha^\top E^\top K E \alpha$$
$$\alpha \geq 0,$$
$$A\alpha \leq \tfrac{C}{N}\mathbf{1}$$

$$\sum_i \sum_n \sum_{[\mathbf{y}]_n} \Delta([\mathbf{y}_i]_n, [\mathbf{y}]_n) \mu_{i[\mathbf{y}]_n}$$

$$\sum_{ij} \sum_{mnpq} \sum_{[\mathbf{y}]_{mn}[\mathbf{y}']_{pq}} B_{i[\mathbf{y}]_{mn}j[\mathbf{y}]_{pq}} \mu_{i[\mathbf{y}]_{mn}} \mu_{j[\mathbf{y}']_{pq}}$$

$$\sum_{[\mathbf{y}]_m} \mu_{i[\mathbf{y}]_{mn}} = \mu_{i[\mathbf{y}]_n} \qquad \sum_{[\mathbf{y}]_n} \mu_{i[\mathbf{y}]_n} = \frac{C}{N}$$

$$\mu_{i[\mathbf{y}]_{mn}} \geq 0$$

- Simplification results:
  - Potentials decompose on edges
  - Risk decomposes on nodes
  - variables: $N2^M$ down to $N(M^2 + M)$
  - constraints: $N2^M$ down to NM^2

Recap

$$\max \mathbf{1}^\top \alpha - \tfrac{1}{2}\alpha^\top E^\top K E \alpha$$
$$\alpha \geq 0,$$
$$A\alpha \leq \tfrac{C}{N}\mathbf{1}$$

$$\sum_i \sum_n \sum_{[\mathbf{y}]_n} \Delta([\mathbf{y}_i]_n, [\mathbf{y}]_n) \mu_{i[\mathbf{y}]_n}$$

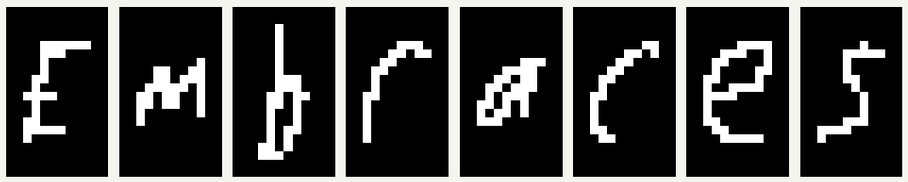
$$\sum_{ij} \sum_{mnpq} \sum_{[\mathbf{y}]_{mn}[\mathbf{y}']_{pq}} B_{i[\mathbf{y}]_{mn}j[\mathbf{y}]_{pq}} \mu_{i[\mathbf{y}]_{mn}} \mu_{j[\mathbf{y}']_{pq}}$$

$$\sum_{[\mathbf{y}]_m} \mu_{i[\mathbf{y}]_{mn}} = \mu_{i[\mathbf{y}]_n} \qquad \sum_{[\mathbf{y}]_n} \mu_{i[\mathbf{y}]_n} = \frac{C}{N}$$

$$\mu_{i[\mathbf{y}]_{mn}} \geq 0$$

- Simplification results:
  - Potentials decompose on edges
  - Risk decomposes on nodes
  - variables: $N2^M$ down to $N(M^2 + M)$
  - constraints: $N2^M$ down to NM^2
- If not a tree the simplified dual is a relaxation

Example application

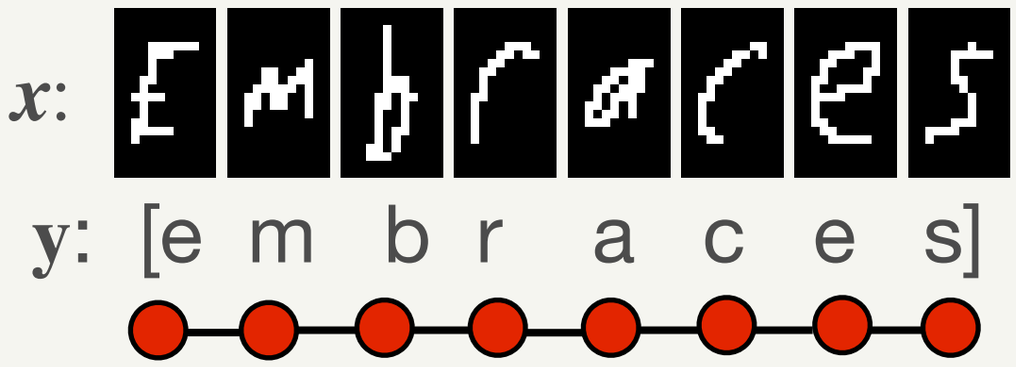
x: 

y: [e m b r a c e s]

$[x]_n$: pixel data for one characters

$[y]_n$: one of {a,b,c, ..., z}

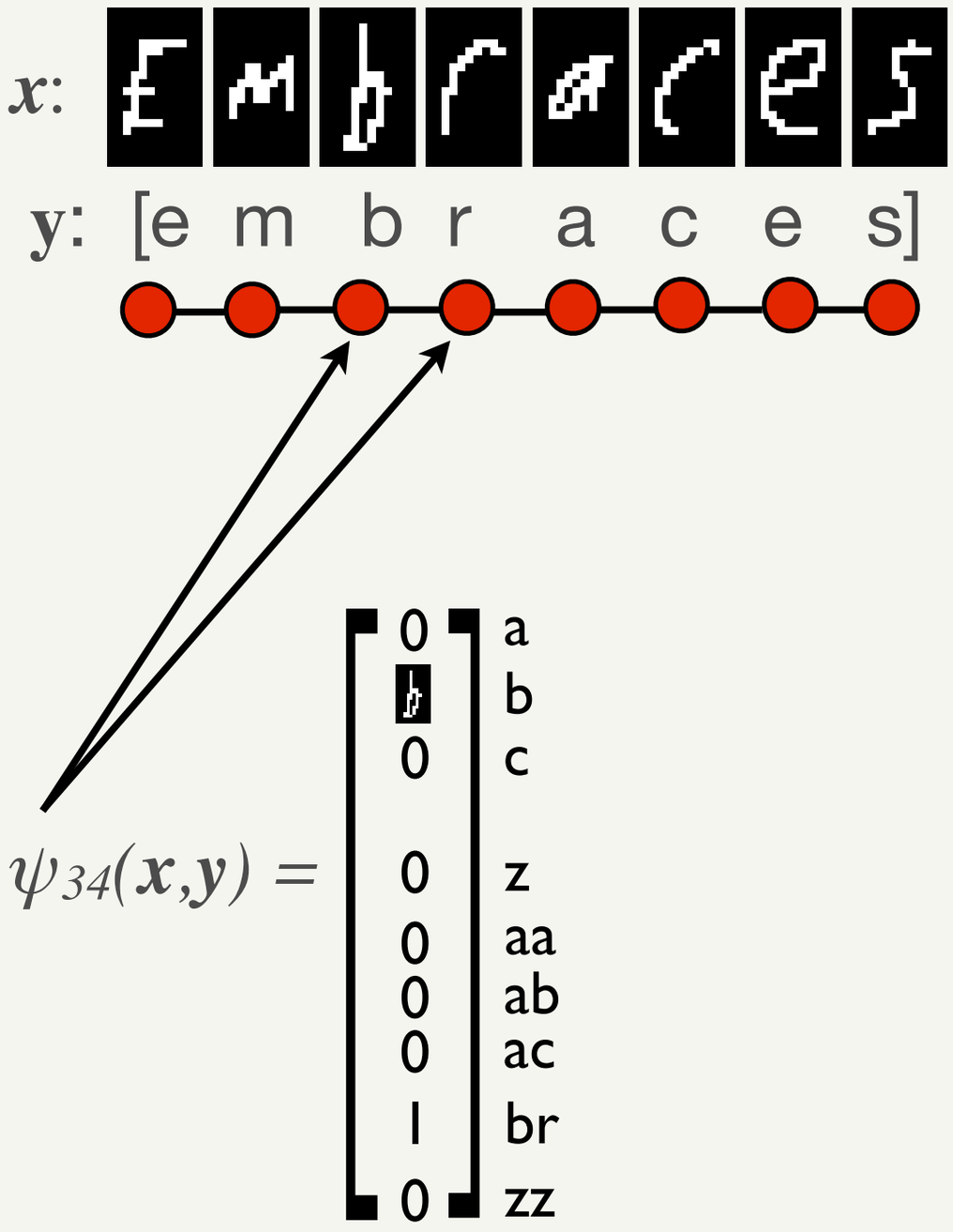
Example application



$[x]_n$: pixel data for one characters

$[y]_n$: one of {a,b,c, ..., z}

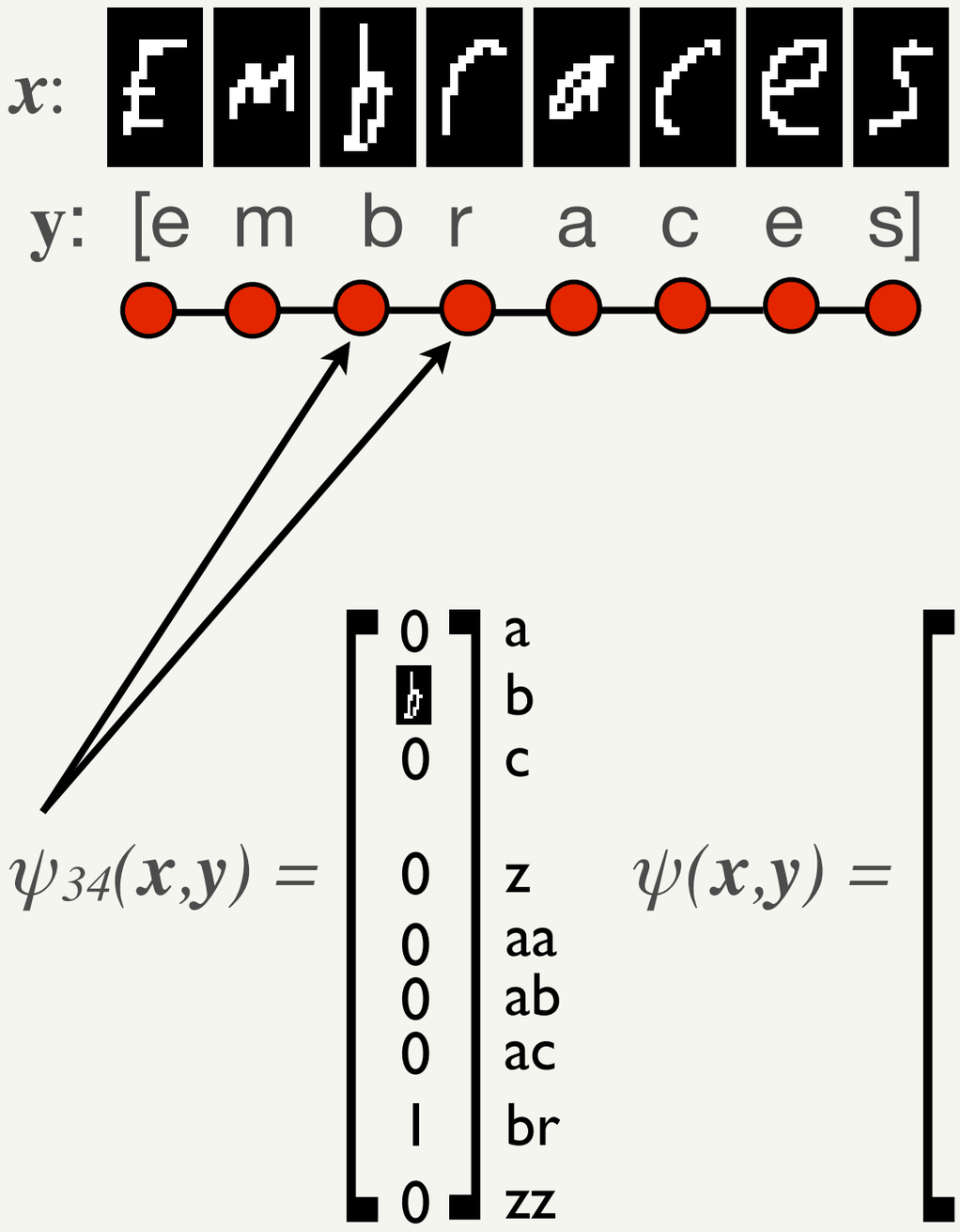
Example application



$[x]_n$: pixel data for one characters

$[y]_n$: one of {a,b,c, ..., z}

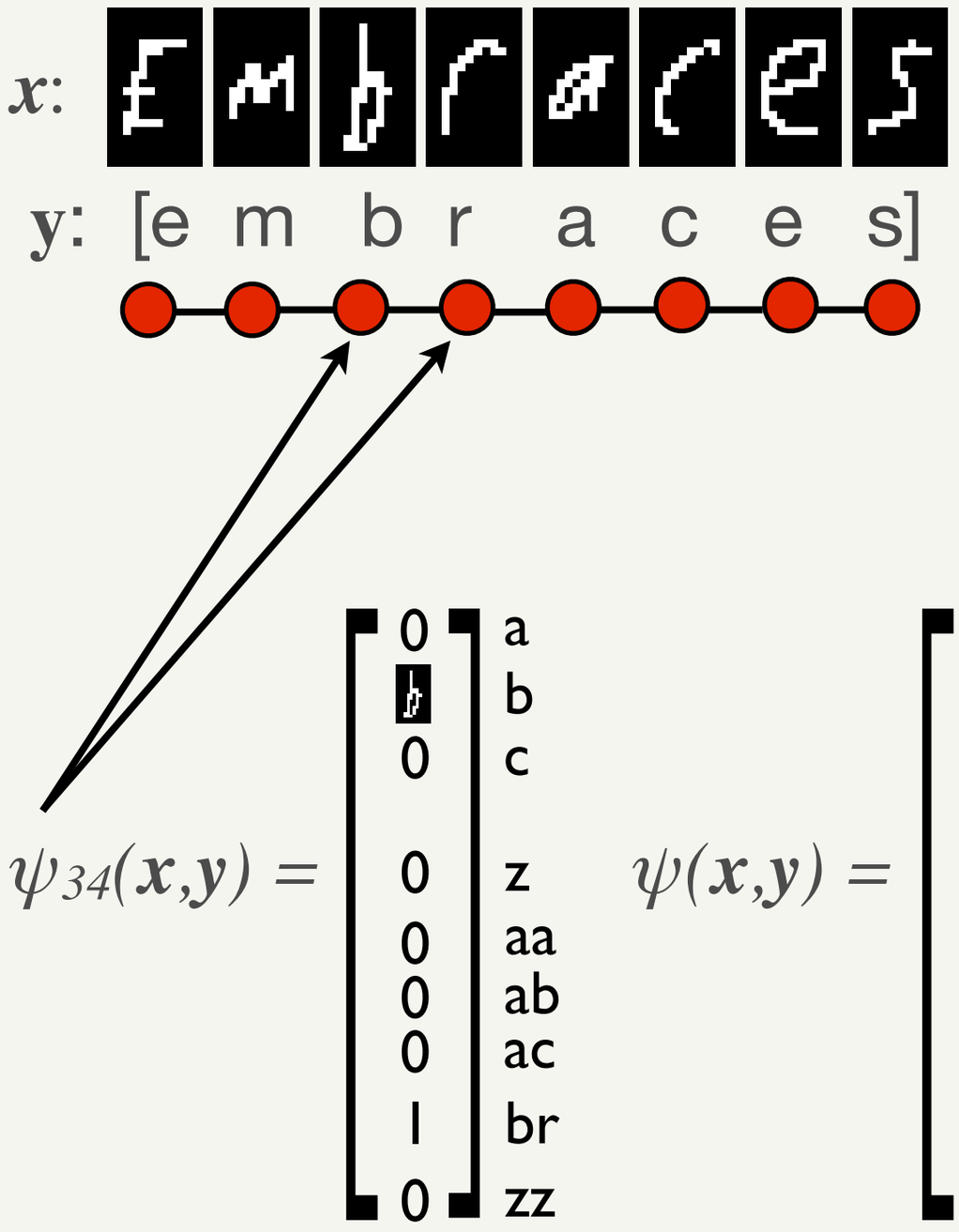
Example application



$[x]_n$: pixel data for one characters

$[y]_n$: one of {a,b,c, ..., z}

Example application



$[x]_n$: pixel data for one characters

$[y]_n$: one of {a,b,c, ..., z}

Comparing words

$\langle \psi(x_1,y_1), \psi(x_2,y_2) \rangle$

correl. homogeneous chars +
correl. char co-occurence matrix

Conclusions

Conclusions

- SVM can be used for arbitrary output
  - very general

Conclusions

- SVM can be used for arbitrary output
  - very general
- Function evaluation entails a maximization (inference) step

Conclusions

- SVM can be used for arbitrary output
  - very general
- Function evaluation entails a maximization (inference) step
- Hinge loss adapted to bound arbitrary loss

Conclusions

- SVM can be used for arbitrary output
  - very general
- Function evaluation entails a maximization (inference) step
- Hinge loss adapted to bound arbitrary loss
- Risk minimization is a large convex program.
  Handled by:
- Iteratively adding one constraint
  - also requires inference step
- M^3 marginalization