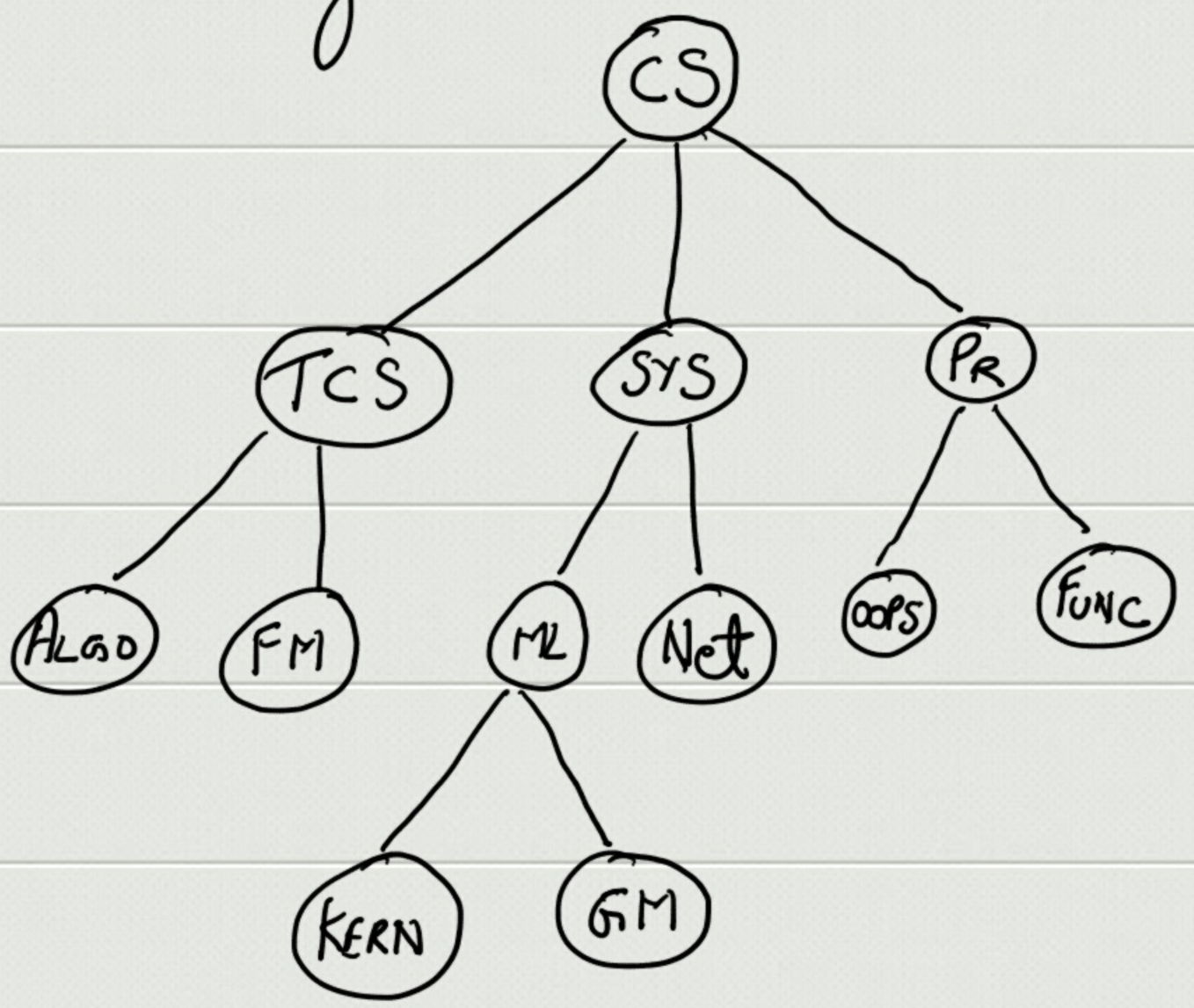
If classes have more `complex' relations than total ordering, then ?

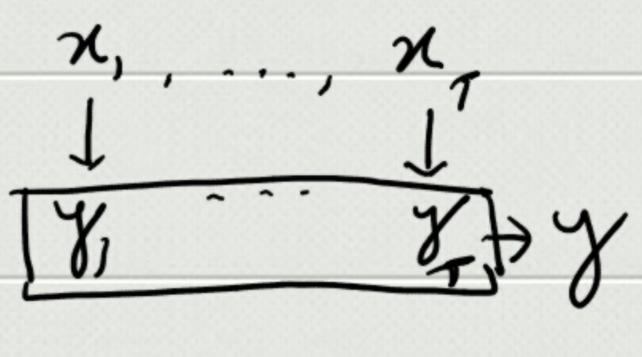
→ eg. classes have a partial order

eg.1 Taxonomy



→ eg. classes have sub-classes

eg.2 Sequence labeling



(Exponential no. of classes)

STRUCTURED PREDICTION

Y has complex struc. other than R.

→ Cumbersome to (linearly) parameterize $g: X \to Y$.

→ Alternative: pose as regression problem

find $g: \underline{X} \to \mathbb{R}$, where $\underline{X} = X \times Y$.

to learn compatibility of pair (x, y).

→ Now linear parametrization is known to us.

(i) → How to get training set?

(ii) → How to get kernel over $X \times Y$

→ k_1 is kernel over X, k_2 is kernel over Y,

Any analytic function (≥0 coeff.) is a kernel over $X \times Y$

(iii) → How to infer label given x using 'g'.

→ find most compatible:

$$\hat{y} = \underset{y \in Y}{\operatorname{argmax}} \; g(x, y)$$

Answer (i) insist on,

$$w^T \varphi(x_i, y_i) - \max_{y \in Y \setminus \{y_i\}} w^T \varphi(x_i, y) \geq 0$$

Again, this situation of one free parameter.

→ Maximize margin, i.e. min. $\frac{1}{2}\|w\|^2$ while imposing:

$$w^T \varphi(x_i, y_i) - \max_{y \in Y \setminus \{y_i\}} w^T \varphi(x_i, y) \geq 1 \quad \forall i$$

→ Allowing for hinge-loss, we have:

$$\min_{\substack{w \in \mathbb{R}^n, \\ \xi \in \mathbb{R}^m}} \frac{1}{2}\|w\|^2 + \frac{C}{m} \sum_{i=1}^{m} \xi_i$$

$$\text{s.t.} \quad \left.\begin{aligned} w^T \varphi(x_i, y_i) - \max_{y \in Y \setminus \{y_i\}} w^T \varphi(x_i, y) &\geq 1 - \xi_i, \\ \xi_i &\geq 0 \end{aligned}\right\} \forall i$$

$$l\left(y_i, g(x_i)\right) \equiv \max\left(0, 1 - \left(w^T \varphi(x_i, y_i) - \max_{y \in Y \setminus \{y_i\}} w^T \varphi(x_i, y)\right)\right)$$

Equivalently,

$$\min_{w \in R^n,\ \xi \in R^m} \frac{1}{2}\|w\|^2 + \frac{C}{m}\sum_{i=1}^{m} \xi_i ,$$

$$\text{s.t.} \quad w^T \psi_i(y) \geq 1 - \xi_i \ , \ \forall\, y \in Y - \{y_i\} \Big\}\ \forall i$$

$$\xi_i \geq 0$$

where, $\psi_i(y) \equiv \varphi(x_i, y_i) - \varphi(x_i, y)$

Let $\alpha_{i,y}$ be the dual variables $[m(|Y|-1)$ in number$]$

→ Since $|Y|$ itself may be large, co-ordinate descent ✗

→ alternative is active set method.

Initialize (A)

→ repeat until convergence

→ for every example

→ Solve using (A)

→ Find KKT violator and append to (A)

Dual is:

$$\max_{\alpha \in R^{m(|Y|-1)}} \sum_{i=1}^{m} \sum_{y \neq y_i} \alpha_{iy} - \frac{1}{2} \sum_{i=1}^{m} \sum_{y \neq y_i} \sum_{s=1}^{m} \sum_{\bar{y} \neq y_s} \alpha_{iy} \alpha_{s\bar{y}} \psi_i(y)^T \psi_s(\bar{y})$$

s.t. $\quad 0 \leq \alpha_{iy} \;\forall\, y \neq y_i, \forall i \;;\; \sum_{y \neq y_i} \alpha_{iy} = \frac{C}{m} \;\forall i$

KKT conditions:

(i) $\omega^T \psi_i(y) \geq 1 - \xi_i$, $\alpha_{iy} \geq 0 \;\forall\, y \neq y_i, \forall i$, $\xi_i \geq 0 \;\forall i$

(ii) $\omega = \sum_{i=1}^{m} \sum_{y \neq y_i} \alpha_{iy} \psi_i(y)$, $\sum_{y \neq y_i} \alpha_{iy} = \frac{C}{m} \;\forall i$,

$\alpha_{iy} = 0 \Rightarrow \xi_i \geq 1 - \omega^T \psi_i(y)$, $\xi_i \begin{cases} \geq 0 & \text{if } \sum_{y \neq y_i} \alpha_{iy} = C/m \\ = 0 & \text{otherwise} \end{cases}$

$0 < \alpha_{iy} < \frac{C}{m} \Rightarrow \xi_i = 1 - \omega^T \psi_i(y)$, $\xi_i \begin{cases} \geq 0 & \text{if } \sum_{y \neq y_i} \alpha_{iy} = C/m \\ = 0 & \text{otherwise} \end{cases}$

$\alpha_{iy} = C/m \Rightarrow \xi_i = 1 - \omega^T \psi_i(y) \geq 0$

$\therefore$ No KKT violation if $\xi_i \geq \max_{y \neq y_i} 1 - w^T \psi_i(y)$,

else the worst violator is given by:

$$\hat{y}_i = \operatorname*{argmax}_{y \neq y_i} 1 - w^T \psi_i(y) = \operatorname*{argmax}_{y \neq y_i} w^T \varphi(x_i, y)$$

$\hat{y}_i$ can be appended to active set (A).

$\rightarrow$ Training, Inference tractable if $\hat{y}_i$ is tractable.