Example with very large numbers of constraints

STRUCTURE LEARNING

- We have a problem with data $\underline{x}$ and discrete labels y
- examples $\underline{x}_i, y_i$
- We believe that a soln can be obtained from

$$\hat{y} = \arg\min_y F(y, \underline{x})$$

where F is some cost function. We assume that F is such that optimization for y is tractable.

- Q: given some data $\underline{x}_i\, y_i$, what is a good F?
- A: The F that gives the right answer (or almost) for each e.g.

→ This reasoning leads to an inequality constrained problem w/ many ineqs

Concrete case:

- sequences of data and labels

$\underline{x}_{1:N,i}$ $y_{1:N,i}$

(the $\underline{x}$ could be - say - sounds and the y phoneme labels; the $\underline{x}$ are - say - words and the y are part of speech tags; the $\underline{x}$ are - say - image measurements and the y are activity labels)

- Cost function has the form

$$F(y_{1:N}, \underline{x}_{1:N}) = \sum_{i=1}^{N} m(y_i, x_i) + \sum_{i=1}^{N-1} b(y_i, y_{i+1}; x_i, x_{i+1})$$

Ex: y can be recovered by dynamic programming

- But what m, b ?
- Traditional: handcraft
- More recent: learn

learning m and b

write $$m(y;x) = \sum_i \theta_i \, m_i(y,x)$$
$$b(y,y';x,x') = \sum_i \theta_i \, b_i(y,y';x,x').$$

(where m_i, b_i might be handcrafted!)

then $$F(y,x) = \theta^T \phi(y,x)$$

Now we have example sequences
$$y^*_{1:N,i}, \; x^*_{1:N,i}$$

Ideally, choose θ such that
$$F(y^*_{1:N,i}, x^*_{1:N,i}) \leq F(y_{1:N,i}, x^*_{1:N,i})$$

for <u>ANY</u> sequence $y_{1:N,i}$
(so this inequality is actually an awful lot of inequalities — one per sequence!)

- in fact, we would like $F(y, x^*)$ to be larger if y is further from y^*

- so write

$$F(y^*_{1:N,i}, \underline{x}^*_{1:N,i}) \leqslant \underbrace{F(y_{1:N,i}, x^*_{1:N,i}) - \tfrac{1}{2}\|y_{1:N} - y^*_{1:N}\|^2}$$

this could be a variety of norms

- Now we won't necessarily get a feasible set, so introduce a slack var, and rearrange, to get

$$\xi_i \geqslant F^* - F + d$$

the norm

$$\xi_i \geqslant 0$$

recall I had $F = \Theta^T \varphi(y, x)$

so these are

$$\xi_i \geq \Theta^T \left[\varphi(y^*_{1:N,i}, x^*_{1:N,i}) - \varphi(y_{1:N,i}, x^*_{1:N,i}) \right] + d(y^*_{1:N,i}, y_{1:N,i})$$

$$\xi_i \geq 0$$

AND these are inequalities are

~~one~~
- in Θ
- one per sequence per example

(so LOTS)

- learning problem

$$\min \frac{\sum_i \xi_i}{N} + \frac{\lambda}{2} \Theta^T \Theta$$

s.t. ineqs

lots but linear!

INTERMISSION

[working set methods]

Q: a working set method requires that we FIND violated inequalities

- For each i, ineq. is:

$$\xi_i \geq \theta^T(\varphi(y_i^*, x_i) - \varphi(y, x_i)) + d(y_i^*, y)$$

known.

so we are interested in

$$\min_y \left[\theta^T \varphi(y, \underline{x}_i) + d(y^*, y)\right]$$

which will give most violated ineq for example i.