Example with very large numbers of constraints

STRUCTURE LEARNING

We have a problem with data $x$ and discrete labels $y$

egatives $x_i, y_i$

We believe that a solution can be obtained from

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

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

Q: given some data $x_i, y_i$,

What is a good $F$?

A: This $F$ that gives the right answer (or almost) for each eg.

→ This reasoning leads to an inequality constrained problem w/ many ineq.
Concrete case:

sequences of data and labels

\[ x_i : N, \quad y_i : N \]

(the \( x \) could be - say - sounds and the \( y \) phoneeme labels; the \( x \) are - say - words and the \( y \) are part of speech tags; the \( x \) are - say - image measurements and the \( y \) are activity labels)

Cost function has the form

\[
F(y_{1:N}, x_{1:N}) = \sum_{i=1}^{N} m(y_i, x_i) + \sum_{i=1}^{N-1} b(y_i, y_{i+1}; x_i, x_{i+1})
\]

\[ \text{Ex: } y \text{ can be recovered by dynamic programming} \]

- But what \( m, b \)?
- **Traditional:** handcraft
- More recent: learn
Learning $m$ and $b$

Write

\[ m \left( y; x \right) = \sum_{i} \Theta_{i} b_{i} \left( y, x \right) \]
\[ b \left( y, y'; x, x' \right) = \sum_{i} \Theta_{i} b_{i} \left( y, y'; x, x' \right) \]

(Where $m_{i}$, $b_{i}$ might be handcrafted.)

Then

\[ F(y, x) = \Theta^{T} \phi \left( y, x \right) \]

Now we have example sequences

\[ y_{i}; n_{i}, x_{i}; n_{i}, i \]

Ideally, choose $\Theta$ such that

\[ F(y^{*}; n_{i}, x^{*}; n_{i}, i) \leq F(y; n_{i}, x; n_{i}, i) \]

for any sequence $y_{i}; n_{i}$

(\text{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:

$$F(y_{1:N,i}^*, x_{1:N,i}^*) \ll F(y_{1:N,i}, x_{1:N,i})$$

$$- \frac{1}{2} \| y_{1:N} - y_{1:N,i}^* \|_2^2$$

Thus could be a variety of norms.

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

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

$$\xi_i \geq 0$$
Recall I had \( F = \Theta^T \varphi(y, x) \)

so these are

\[
\xi_i \geq \Theta^T \left[ \varphi(y_{i:n-1}^*, x_{i:n-1}^*) - \varphi(y_{i:n-1}^*, x_{i:n-1}^*) \right] \\
+ d(y_{i:n-i}, y_{i:n-i})
\]

\( \xi_i \geq 0 \)

AND these are inequalities are

one \( \in \Theta \)

one per sequence per example

(so LOTS)

learning problem

\[
\min \sum_{i} \frac{\xi_i}{N} + \frac{\lambda}{2} \Theta^T \Theta
\]

s.t. \( \text{meas} \)

\( \rightarrow \)

LOTS but linear!
INTERMISSION

[ working set methods ]

Q: A working set method requires that we find violated inequalities.

- For each $i$, in eq. is:

$$\xi_i > \Theta^T(\phi(c_i^*, x_i) - \phi(y, x_i)) + d(c_i^*, y)$$

known.

So we are interested in

$$\min_y [\Theta^T \phi(y, x_i) - d(c_i^*, y)]$$

which will give most violated ineq for example $i$. 