Idea:
- we have data $D$, params $\theta$, and missing data $\Delta$
- $P(D, \Delta | \theta)$ would be easy to work with
- $P(D | \theta) = \int P(D, \Delta | \theta) d\Delta$

is usually hard
- $\log \log$ is sum

Algorithm: given $\theta^{(n)}$

$E$ step: form $Q(\theta; \theta^{(n)}) = \mathbb{E}_{\Delta | \theta^{(n)}}[P(D, \Delta | \theta^{(n)})]$

$M$ step: form $\theta^{(n+1)} = \arg \max_{\theta} Q(\theta; \theta^{(n)})$
This takes a usefully simple form when \( \log P(D, \Delta | \theta) \) is linear in \( \theta \).

**Example:** Dynamic model with only one clock interval.

\[
D = Y^{(i)}_0
\]

\[
\Delta = \delta^{(i)}_{o j}
\]

\[
= \begin{cases} 
1 & (x_o = x_j) \\
0 & \text{(otherwise)}
\end{cases}
\]

\[
\log P(D, \Delta | \theta) = \log P(D | \Delta, \theta) + \log P(\Delta | \theta)
\]

This term determined by emission model, this is known prior.
\[
P(D_{ij} \mid \Delta \mid \theta) P(Y_o \mid x_o, \theta) \\
= \sum_{i,j} \left( \frac{(Y_o - \mu(x_j, \theta))^2}{\sigma^2} \right)^{-1} \\
= \sum_{i,j} \left[ (Y_o - \mu(x_j, \theta))^2 \right]^{-1} \\
= (Y_o - \mu(x_j, \theta))^2 \Sigma^{-1} (Y_o - \mu(x_j, \theta)) \\
+ \log K_n \\
\frac{\text{but this is not an}}{\text{fit of } \theta, \sigma}
\]
\[
\log P(D, \Delta | \theta) = \sum_{ij} S_{ij} \left[ (y_0 - \mu(x_j, \theta))^T \sum_{i} \frac{1}{2} (y_0 - \mu(x_j, \theta)) \right] + \text{constant terms}
\]

This acts like a switch.

Now

\[
E[\delta_j^i | S_j^i = 1, D, \theta] = 1 \cdot P(\delta_j^i = 1 | D, \theta) + 0.
\]

\[
P(\delta_j^i = 1 | D, \theta) = \frac{P(y_0^i | S_j^i, \theta) P(S_j^i = 1 | D)}{\sum_{u} P(y_0^i | S_u^i, \theta) P(S_u^i = 1 | D)}
\]
In this case, we get

\[ P(\mathcal{Y}_j | D, \Theta) = \exp \left[ \sum_{i=1}^{\frac{1}{2}} \left( \frac{y_{0}^{(i)} - \mu(x_j^i, \Theta)(\Sigma^{-1})_{ij} - \mu(x_j^i, \Theta)}{\Sigma} \right) \right] \times \prod_{n} \text{terms as above} \]

so the step is straightforward.

\underline{M-step}

- depends on \( \mu(x_j^i, \Theta) \) (form of function)

\[ \text{e.g., } \mu(x_j^i, \Theta) = \Theta \cdot x_j^i \]

and this has a 1 in j'th location and zeros elsewhere
this case is one mean per state

- Now look at LLH as fn of j'th mean

\[ \sum_i P(S_i^j | D, \Theta^{(u)}) \cdot \left( \frac{(Y_i^j - M_j) \sum (Y_i^j - M_j)}{2} \right) \]

+ other terms that don't depend on \( m_j \)

But this is just a weighted mean.
Case 2: \( P(Y_{o^*} | X_o) \) is a table because \( Y \) is discrete. Maximization is by weighted counts.

Example 2: Sequences, multiple examples

\[
P(Y_0^{(i)} \ldots Y_n^{(i)} , S_{oj}^{i} \ldots S_{n_j}^{i} | \theta) = P(Y_0^{(i)} \ldots Y_n^{(i)} | S_{oj}^{i} \ldots S_{n_j}^{i}, \theta) \times P(S_{oj}^{i} \ldots S_{n_j}^{i}, \theta)
\]

- We are assuming that dynamics are known, so second term is fixed.
\[
\log P(Y_0^{(i)} \cdot Y_n^{(i)} \mid s_{o}^i \cdot s_{n}^i, \Theta) = \sum_j \left[ \log P(Y_0^{(i)} \mid x_0 = x_j, \Theta) \right] s_{o}^j \\
+ \sum_j \left[ \log P(Y_i^{(i)} \mid x_i = x_j, \Theta) \right] s_{n}^j \\
+ \ldots
\]

Switch

1 per clock tick.

Now consider the E step

\[ P(S_{eji}^i = 1 \mid Y_0^{(i)} \cdots Y_n^{(i)}, \Theta) = \frac{P(X_e = x_j \mid Y_0^{(i)}, Y_n^{(i)}, \Theta)}{P(Y_0^{(i)} \cdots Y_n^{(i)}, \Theta)} \]
Constrained optimization:

\[ \min f(x) \quad \text{st} \quad c_i(x) = 0 \]
\[ g_i(x) > 0 \]

Lagrangian

\[ L(x, \lambda) = f(x) - \lambda^T c - \lambda^T g \]

(\( \lambda \) is a vector of constraints whose elements correspond to equality constraints (\( \lambda^{(e)} \)) or inequality constraints (\( \lambda^{(i)} \))

Necessary conditions (KKT conditions)

\[ \nabla_x L = 0 \]
\[ g_i(x) > 0 \]
\[ c_i(x) = 0 \]
\[ (e) \]
\[ \lambda_i c_i = 0 \]
\[ (e) \]
\[ \lambda_i g_i = 0 \]
Assume inequality constraints only.

\[
\min f(x) \quad s.t. \quad g_i(x) \geq 0
\]

Example:

Assume \( \min -\frac{\partial^T}{\partial x} \) is strongly convex \( x = 6 \)

\[
L(x, \lambda) = x^T x - \lambda^T (Ax - b)
\]

\[
J(x, \lambda) = -\frac{\partial}{\partial x} - \lambda^T g(x)
\]

From first condition objective \( f \) to be

\[
x^T \lambda = 0
\]

i.e. \( \bar{x} = \text{arg}\max_{x} J(x, \lambda) \)

Substitute on domain such that \( g(x) \geq 0 \)

\[
\frac{1}{2} \lambda^T AA \lambda - \lambda^T (AAx - b)
\]

dual problem:

\[
\lambda \rightarrow x
\]

Knowledge of \( \lambda \) with values \( \lambda \geq 0 \) is powerful!
Then: \( q \) is concave, domain is convex (straightforward)

Then: for feasible \( x \), any \( \lambda \)

\[ q(\lambda) \leq f(x) \] (straightforward)

Then: suppose \( x \) is soln of primal, \( f \) and \( -g_i \) are convex; then \( \lambda \) such that \((x, \lambda)\) satisfies KKT is a soln of dual

Then: with other way round requires stronger technical (unds)

Then: Value of dual \( \leq \) Value of primal
Common application: in important cases, one may be able to write the dual directly.

SVM

$$\min \frac{w^T w}{2} \quad \text{Primal form, Separable}$$

$$\begin{align*}
st & \quad y_i (w^T x_i + b) \geq 1 \\
L(w, \lambda) &= \frac{w^T w}{2} - \sum_i \lambda_i \left[ y_i (w^T x_i + b) - 1 \right] \\
\nabla_w L &= 0 = w - \sum_i \lambda_i \left\{ [y_i x_i]^T \right\} \\
\nabla_b L &= 0 = -\sum_i \lambda_i y_i
\end{align*}$$
\[
I = \sum_{i} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j [y_i y_j \langle x_i, x_j \rangle]
\]

Notice constraints

\[\sum \lambda_i y_i = 0\]
\[\lambda_i \geq 0\]

and we must \[\text{max}\] this in \[\lambda\].

If there is an \(fp\) for primal, the \[\text{max}\] is \[\text{soln}\] to primal

i.e. \[\text{Value (Dual)} = \text{Value (Primal)}\]
What if data is not separable?

\[
\begin{align*}
\text{min} & \quad \frac{1}{2} w' w + c \sum \xi_i \\
\text{st} & \quad y_i (w' x_i + b) \geq 1 - \xi_i \\
\xi_i & \geq 0
\end{align*}
\]

\(\xi_i\) are slack variables

\[
\begin{align*}
L_p &= \frac{1}{2} w' w + c \sum \xi_i - \sum \lambda_i [y_i (w' x_i + b) - 1 + \xi_i] \\
&\quad - \sum \mu_i \xi_i
\end{align*}
\]

\[
\begin{align*}
\nabla_w L_p &= w - \sum \lambda_i y_i x_i = 0 \\
\nabla_b L_p &= 0 = -\sum \lambda_i y_i \\
\nabla_{\xi_i} L_p &= c - \lambda_i - \mu_i = 0 \quad \Rightarrow \quad \text{this gets rid of } \xi_i
\end{align*}
\]
So, we have

\[ I_0 = \sum \lambda_i - \frac{1}{2} \sum y_i y_j \lambda_i \lambda_j x_i x_j \]

subject to

\[ \sum \lambda_i y_i = 0 \]

\[ 0 \leq \lambda_i \leq C \]

Notice that \( \xi_i \) can be interpreted as a **loss**

hinge loss \((y_i y_P) = \max(0, 1 - y_i y_P)\)
Methods:

**Quadratic penalty method**

(assume equalities)

\[
\min_x f(x) + \mu \sum_i c_i^2(x) = Q_\mu(x)
\]

and drive \( \mu \to \infty \), resolve

Notice at soln

\[
\nabla_x Q_\mu = 0 = \nabla f + \sum_i (\mu_k c_i(x)) \nabla c_i(x)
\]

By inspection, this would match

\[
\nabla Z = 0, \quad \text{if} \quad -\mu_k c_i = \lambda_i^*
\]

Which suggests that at opt \( c_i = \frac{-\lambda_i^*}{\mu_k} \)}
This looks OK, because $\mu_k \to 0$, but not exact. Also $\mu_k \to \infty$ creates major probs w/ Hessian.

Augmented Lagrangian method

Consider

$$L_A(x, \lambda; \mu) = f - \sum_i \lambda_i c_i + \frac{\mu}{2} \sum_i c_i^2$$

- have an set of $\lambda^k$, $\mu_k$, get $x^*$

- at $x^*$, $\nabla f = \sum_i (\lambda_i^k - \mu_k c_i) \nabla c_i$

- This suggests $\lambda_i^* \approx (\lambda_i^k - \mu_k c_i)$

and $c_i \approx -\frac{1}{\mu_k} [\lambda_i^* - \lambda_i^k]$

Which suggests moving $\lambda_i \to \lambda_i^*$
but we have a good est:

\[ x_i^* \approx (x_i^k - \mu_k c_i) \]

so update ests, go again.

1) Method converges w/o increasing \( \mu_k \) indefinitely
Conjugate gradient

We now have:

Start: $x_0, \quad r_0 = Ax_0 - b, \quad p_0 = -r_0$

Step:

\[
x_{k+1} = x_k + \alpha_k p_k
\]
\[
\alpha_k = -\frac{r_k' A p_k}{p_k' A p_k}
\]
\[
p_{k+1} = -r_{k+1} + \beta_{k+1} p_k
\]
\[
\beta_{k+1} = \frac{p_{k+1}' A r_{k+1}}{p_k' A p_k}
\]

We can make this more efficient
This gives

\[
\frac{1}{2} \left[ (x_k + \alpha_k p_k)' A (x_k + \alpha_k p_k) \right] - b_k (x_k + \alpha_k p_k)
\]

new is at:

\[
-(Ax_k - b)' p_k
\]

\[
\frac{p_k' A p_k}{p_k' A p_k}
\]

write

\[
r_k = Ax_k - b
\]

so \( \alpha_k = -r_k' p_k \)

\[
\frac{p_k' A p_k}{p_k' A p_k}
\]
gate gradient (simple form)

Start: \( x_0, \ p_0 = Ax_0 - 6, \ p_0 = -r_0 \)

Step:

\[
\begin{align*}
x_{k+1} &= x_k + \alpha_k p_k \\
\alpha_k &= \frac{-r_k^T p_k}{p_k^T A p_k} \\
p_{k+1} &= -r_{k+1} + \beta_{k+1} p_k \\
\beta_{k+1} &= \frac{r_{k+1}^T A p_k}{p_k^T A p_k} \\
r_{k+1} &= r_k + \alpha_k A p_k
\end{align*}
\]
Conjugate gradient,

Cleaner form:

By properties, we have

\[ \alpha_{k+1} = \frac{\tilde{r}_k' \tilde{r}_k}{p_k' A p_k} \]

Now \( \alpha_k A p_k = r_{k+1} - r_k \)

So \[ \beta_{k+1} = \frac{r_{k+1}' (r_{k+1} - r_k)}{\alpha_k} \cdot \frac{1}{p_k' A p_k} \]

\[ = \frac{r_{k+1}' (r_{k+1} - r_k)}{\tilde{r}_k' \tilde{r}_k} \]

\[ = \frac{r_{k+1}' r_{k+1}}{\tilde{r}_k' \tilde{r}_k} \quad \text{(By properties)} \]
Properties of conj. direction

\[ r_k' p_i = 0, \quad \forall i \leq k \]

(Show this by induction)

\[ r_k' r_i = 0 \quad \forall i \leq k \]

(thus 5.3 at end)
Conjugate directions in incremental form

Start with $x_0$, $P_0$

$$x_1 = x_0 + \alpha_0 P_0$$

now min wrt $\alpha_0$

to get

$$\frac{(Ax_0 - b)' P_0}{P_0' A P_0} = \alpha_0$$

write

$$r_K = (Ax_K - b)$$

and get

$$x_{K+1} = x_K + \alpha_K P_K$$

$$\alpha_K = \frac{r_K' P_K}{P_K' A P_K}$$

$$r_{K+1} = r_K + \alpha_K A P_K$$
Conjugate direction methods:

- a set of vectors, $P_0 \ldots P_n$ is conjugate for A positive definite if

$$P_i^TAP_j = 0 \quad \text{if} \quad i \neq j$$

- Assume we wish to min

$$\frac{x^TAX - b^Tx}{2}$$

- Useful because:

a) Solution to $AX = b$ for $A$ p.d.

b) $\min_x \|Ax - b\|^2$ is like this

- Now write

$$x = \alpha_0 P_0 + \alpha_1 P_1 + \ldots$$