Idea:

- we have data D, params θ
  and missing data Δ
- $P(D, \Delta | \theta)$ would be easy to
  work with
- $P(D|\theta) = \int P(D, \Delta | \theta) \, d\Delta$
  is usually hard
  - eg log of sum

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

E step · form $Q(\theta; \theta^{(n)}) = E_{\Delta|\theta^{(n)}} [P(D, \Delta|$

M step · form $\theta^{(n+1)} = \underset{\theta}{\operatorname{argmax}} \; Q(\theta; \theta^{(n)}$

~~This takes a usefully simple form when $\log P(D, \Delta | \theta)$ is linear in~~

<u>Example</u>: dynamic model with only one clock interval

$D = Y_0^{(i)}$ ← lots of different obs of 0'th emission

$$\Delta = \delta_{0j}^{(i)} = \begin{cases} 1 & (X_0 = x_j) \\ 0 & (\text{otherwise}) \end{cases}$$

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

this term determined by emission model

this is known prior

~~Range~~ $\underline{\log P(D|;\Delta|\theta)}$ $P(Y_o \mid x_o, \theta)$

$$= \sum_{i,j} \delta_j^i \left[\cancel{(Y_o^{(i)} \text{ is } \mu(X_j, \theta))' \Sigma^{-1} (x_o, \theta) (Y_o^{(i)} \boxed{\mu}(x), \theta} \right.$$

$$+ \text{ constant fixed and}$$

here δ^i operates like ~~kanoswitch~~

so $\log_j P(Y_o \mid \delta_{o_j}, \theta)$

$$= (Y_o - \mu(x_j, \theta))' \Sigma^{-1} (Y_o - \mu(x_j, \theta))$$

$$+ \underbrace{\log K_n}$$

but this is not an fn of θ, δ

$\log P(D, \Delta | \theta)$

$$= \sum_{i,j} \delta_i^i \left[(Y_o^i - \mu(x_j, \theta))^T \underline{\Sigma}^{-1} (Y_o^i - \mu(x_j, \theta\right.$$

$$+ \text{constant terms}$$

This acts like a switch

Now $E[\delta_j^i]_{\delta_j^i | D, \theta}$

$$= 1 . P(\delta_j^i = 1 | D, \Theta) + 0 \cdot \times$$

$$P(\delta_j^i = 1 | D, \theta) = \frac{P(Y_o^{(i)} | \delta_j^i, \theta) \, P(\delta_j^i | \theta}{\sum_u P(Y_o^{(i)} | \delta_u^i, \theta) P(\delta_u^i | \theta}$$

In this case, we get

$$P(\delta_j^i \mid D, \theta) = \frac{\exp\left[(Y_0^{(i)} - \mu(x_j^i, \theta)) \Sigma^{-1} \times (Y_0^{(i)} - \mu(x_j^i, \theta)) \right] \times \pi_j}{\sum_u (\text{terms as above})}$$

So E step is straightforward.

<u>M-step</u>

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

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 <u>mean</u>

$$\sum_i P(S_i^j \mid D, \theta^{(n)}) \cdot \left[(Y_o^{(i)} - \mu_j)' \Sigma^{-1} (Y_o^{(i)} - \mu_j) \right]$$

+ <u>other terms that don't depend on</u> μ_j

But this is just a <u>weighted</u> mean.

<u>case 2:</u> $P(Y_0^{*} \mid X_0)$ is a table because Y is discrete

Maximization is by weighted counts

<u>Example 2:</u> sequences, multiple examples

$$P(Y_0^{(i)} \cdots Y_n^{(i)}, \delta_{0j}^{i} \cdots \delta_{nj}^{i} \mid \theta)$$

$$= P(Y_0^{(i)} \cdots Y_n^{(i)} \mid \delta_{0j}^{i} \cdots \delta_{nj}^{i}, \theta) \times P(\delta_{0j}^{i} \cdots \delta_{nj}^{i}, \theta)$$

- we are assuming that dynamics are known, so second term is fixed

$$\log P(Y_0^{(i)} \cdots Y_n^{(i)} \mid \delta_{0j}^{i} \cdots \delta_{nj}^{i}, \theta)$$

Switch

1 per clock tick.

$$= \sum_j \left[\log P(Y_0^{(i)} \mid X_0 = x_j, \theta)\right] \cdot \delta_{0j}^{(i)}$$

$$+ \sum_j \left[\log P(Y_1^{(i)} \mid X_1 = x_j, \theta)\right] \delta_{1j}^{i}$$

$$+ \quad \vdots$$

Now consider the E step

$$P(\delta_{ej}^{i} = 1 \mid Y_0^{(i)} \cdots Y_n^{(i)}, \theta)$$

$$= P(X_e^{i} = x_j \mid Y_0^{(i)} \cdots Y_n^{(i)}, \theta)$$

$$= \frac{P(X_e^{i} = x_j, Y_0^{(i)} \cdots Y_n^{(i)}, \theta)}{P(Y_0^{(i)} \cdots Y_n^{(i)}, \theta)}$$

we kn
thes

Constrained optimization:

$$\min \ f(x) \qquad \text{st} \qquad c_i(x) = 0$$
$$g_i(x) \geq 0$$

Lagrangian

$$\mathcal{L}(x, \lambda) = f(x) - \lambda^{(e)^T} c - \lambda^{(i)^T} g$$

(here λ is a vector of constraints whose elements csp ~~torq~~ ineq ($\lambda^{(i)}$) or eq constraints ($\lambda^{(e)}$)

Necessary conditions (KKT conds)

$$\nabla_x \mathcal{L} = 0 \qquad g_i(x) \geq 0$$

$$c_i(x) = 0 \qquad \lambda^{(i)} \geq 0$$

$$\lambda_i^{(e)} c_i = 0 \qquad \lambda_i^{(i)} g_i = 0$$

- Assume inequality constraints only.

Duality

$$\min f(x) \quad \text{st} \quad g_i(x) \geq 0$$

example:

- Assume $\min \frac{x^T x}{2}$ is st $\cancel{\text{convex}}$ $= b$

$$\mathcal{L}(x,\lambda) = x^T x - \lambda^T (Ax - b)$$

$$\mathcal{L}(x,\lambda) = \frac{f(x)}{2} - \lambda^T g(x)$$

from first condition

Define dual objective fn to be

$$x^T - \lambda^T A = 0$$

$$\text{i.e. } q(\lambda) = \max_x \mathcal{L}(x,\lambda)$$

substitute

on domain such that $q(\lambda) > -\infty$

$$\frac{\lambda^T A A^T \lambda}{2} - \lambda^T (A A^T \lambda - b)$$

dual problem:

$\longrightarrow \lambda \longrightarrow x$

$$\max q(\lambda) \quad \lambda_i \geq 0$$

Knowledge of λ wrt values is powerful!

<u>Thm</u>: q is concave, domain is convex
(straightforward)

<u>Thm</u>: for feasible x, any λ

$$q(\lambda) \leq f(x)$$

(straightforward)

<u>Thm</u>: suppose x is soln of primal, f and $-g_i$ are convex; then λ such that (x, λ) satisfies KKT is a soln of dual

~~Thm: with~~ other way round requires stronger technical conds

<u>Thm</u>: value of dual $\leq$ value of primal.

Common application: in important cases, one may be able to write the dual directly.

SVM

$$\left.\begin{array}{l} \min \quad \dfrac{w'w}{2} \\[2ex] \text{st} \;\; y_i\,(w'x_i + b) \geq 1 \end{array}\right] \text{Primal form, Separable}$$

$$\mathcal{L}(w,\lambda) = \frac{w'w}{2} - \sum_i \lambda_i \left\{ \left[y_i\,(w'x_i + b) \right] - 1 \right\}$$

$$\nabla_w \mathcal{L} = 0 = w - \sum_i \lambda_i \left\{ \left[y_i x_i \right] \right\}$$

$$\nabla_b \mathcal{L} = 0 = -\sum_i \lambda_i\, y_i$$

Subst :

~~Alin~~

P

<u>subst</u>

$$\mathcal{L}_D = \sum_i \lambda_i - \frac{1}{2} \sum_{ij} \lambda_i \lambda_j \left[y_i y_j (x_i^T x_j) \right]$$

Notice constraints

$$\sum_i \lambda_i y_i = 0$$

$$\lambda_i \geq 0$$

and we must <u>max</u> this in λ

<u>If</u> there is an fp for primal, the max is soln to primal

i.e. Value (Dual) = Value(Primal

What if data is not separable?

$$\left.\begin{aligned} \min \quad & \frac{w'w}{2} + C\sum_i \xi_i \\ \text{st} \quad & y_i(w'x_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned}\right]$$ Primal prob

ξ_i are slack variables

$$\mathcal{L}_p = \frac{w'w}{2} + C\sum_i \xi_i - \sum_i \lambda_i [y_i(w'x_i + b) - 1 + \xi_i] - \sum_i \mu_i \xi_i$$

$$\nabla_w \mathcal{L}_p = w - \sum_i \lambda_i y_i x_i = 0$$

$$\nabla_b \mathcal{L}_p = 0 = -\sum_i \lambda_i y_i$$

$$\nabla_{\xi_i} \mathcal{L}_p = C - \lambda_i - \mu_i = 0 \quad \Big] \rightarrow$$ this gets rid of ξ_i

Step 1 we have

$$\mathcal{L}_D = \sum_i \lambda_i - \frac{1}{2} \sum_{ij} y_i y_j \lambda_i \lambda_j x_i' x_j$$

subject to

$$\sum_i \lambda_i y_i = 0$$

$$0 \leq \lambda_i \leq C$$

Notice that ξ_i can be interpreted as a <u>loss</u>

$$\text{hinge loss}(y, y_p) = \max(0, 1 - y_i y_p)$$

Methods:

Quadratic penalty method

(assume equalities)

$$\min_x f(x) + \frac{\mu}{2} \sum_i c_i^2(x) = Q_k(x)$$

and drive $\mu \to \infty$, resolve

Notice at soln

$$\nabla_x Q_k \approx 0 = \nabla f + \sum_i \left(\mu_k c_i(x_k)\right) \nabla c_i(x)$$

By inspection, this would match $\nabla_x \mathcal{L} = 0$, if

$$-\mu_k c_i = \lambda_i^*$$

Which suggests that at conv $c_i = \frac{-\lambda_i^*}{\mu_k}$

This looks OK, because $\mu_k \to \infty$, but not exact. Also $\mu_k \to \infty$ creates major probs w/ Hessian

Augmented Lagrangian method

Consider

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

- have an est of λ^k, μ_k, get x^*

- at x^* $\quad \nabla_x \mathcal{L}_A = 0 = \nabla f - \sum_i (\lambda_i^k - \mu_k c_i) \nabla c_i$

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

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

which suggests moving $\lambda_i \to \lambda_i^*$

But we have a good est:

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

so update ests, go again.

1) Method converges w/o increasing μ_k indefinitely

Conjugate gradient

We now have:

Start: x_0 , $r_0 = A x_0 - b$, $p_0 = -r_0$

Step:

$$x_{k+1} = x_k + \alpha_k p_k$$

$$\alpha_k = \frac{-r_k' A p_k}{p_k' A p_k}$$

$$p_{k+1} = -r_{k+1} + \beta_{k+1} p_k$$

$$\beta_{k+1} = \frac{p_k' A r_{k+1}}{p_k' A p_k}$$

We can make this more efficient

Again ~~This~~ gives

$$\frac{1}{2}\left[(x_K + \alpha_K p_K)' A (x_K + \alpha_K p_K) \right]$$
$$- b_K (x_K + \alpha_K p_K)$$

min is at:

$$- \frac{(A x_K - b)' p_K}{p_K' A p_K}$$

write

$$r_K = A x_K - b$$

so $$\alpha_K = - \frac{r_K' p_K}{p_K' A p_K}$$

gate gradient (simple form)

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

Step:

$$x_{k+1} = x_k + \alpha_k p_k$$

$$\alpha_k = \frac{-r_k' p_k}{p_k' A p_k}$$

$$p_{k+1} = -r_{k+1} + \beta_{k+1} p_k$$

$$\beta_{k+1} = \frac{r_{k+1}' A p_k}{p_k^T A p_k}$$

$$r_{k+1} = r_k + \alpha_k A p_k$$

Conjugate gradient.

<u>Cleaner form</u>:

- By properties, we have

$$\alpha_{K\cancel{+1}} = \frac{r_K' r_K}{p_K' A p_K}$$

- Now $\alpha_K A p_K = r_{K+1} - r_K$

So
$$\beta_{K+1} = r_{K+1}' \frac{(r_{K+1} - r_K)}{\alpha_K} \cdot \frac{1}{p_K' A p_K}$$

$$= \frac{r_{K+1}' (r_{K+1} - r_K)}{r_K' r_K}$$

$$= \frac{r_{K+1}' r_{K+1}}{r_K' r_K}$$ (By properties)

Properties of conj. direction

$$r_k' p_i = 0, \quad \forall_i < k$$

(Show this by induction)

$$r_k' r_i = 0 \quad \forall_i < k$$

(thm 5.3 at end).

Conjugate direction in incremental form

Start with x_0 , p_0

$$x_1 = x_0 + \alpha_0 p_0$$

now min wrt α_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, $p_0 \cdots p_n$ is conjugate for A positive definite if

$$p_i' A p_j = 0 \qquad i \neq j$$

- Assume we wish to min

$$\frac{x' A x}{2} - b' x$$

- useful because:

  a) solution to $Ax = b$ for A p.d.

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

- now write

$$x = \alpha_0 p_0 + \alpha_1 p_1 + \cdots$$