Exponential Families

the maximum entropy problem

X is a random variable with some distribution $p(x)$.

assume $X_{(i)} \overset{iid}{\sim} p$, $i = 1, 2, \cdots, n$;

given a function $\phi_\alpha : \mathcal{X} \to \mathbb{R}$ $(X \in \mathcal{X})$

we ~~won~~ can compute the empirical expectation

$$\hat{\mu}_\alpha = \frac{1}{n} \sum_{i=1}^{n} \phi_\alpha(X_{(i)}) .$$

Now assume $\alpha \in I$ where I is some index set, and based on the $|I|$-dim. empirical expectation vector

$$\hat{\mu} = (\hat{\mu}_\alpha, \alpha \in I),$$

we want to solve the following maximum entropy problem

$$p^* = \arg\max_{p \in \mathcal{P}} H(p) \quad \text{s.t.} \quad E_p[\phi_\alpha(X)] = \hat{\mu}_\alpha, \forall \alpha \in I$$

where $\mathcal{P}$ is the set of all probability distributions over X.

Solution

We define Lagrange multipliers $\{\theta_\alpha\}$ λ and modify the objective:

$$p^*, \{\theta^* = \arg\max_{p \in \mathcal{P}, \theta \in \mathbb{R}^d} H(p) + \Sigma_\alpha \theta_\alpha \{E_p[\phi_\alpha(X)] - \hat{\mu}_\alpha\} + \lambda(\int p - 1)$$

where we denote $\theta = (\theta_\alpha, \alpha \in I)$ and the objective as $L(p, \theta, \lambda)$

Then we want $\nabla_p L(p, \theta, \lambda) = 0$

$$\Leftrightarrow \quad p \propto \exp[\Sigma_\alpha \theta_\alpha \phi_\alpha(X)] .$$

λ is to make p a valid distribution, and by solving θ we add in the constraints specified by the empirical expectations $\hat{\mu}$.

Formal Definition

Given a random variable $X = (X_1, X_2, \cdots, X_m) \in \mathcal{X}^m$,
Let $\phi = (\phi_\alpha, \alpha \in I)$ be a collection of functions $\phi_\alpha : \mathcal{X}^m \to \mathbb{R}$,
and $\theta = (\theta_\alpha, \alpha \in I)$ be an associated vector. We define the
<u>exponential family</u> associated with ϕ ~~is~~ as the set of the
following parameterized density functions

$$p_\theta(x_1, x_2, \cdots, x_m) = \exp\{\langle\theta, \phi(x)\rangle - A(\theta)\},$$

$$\theta \in \Lambda := \{\theta \in \mathbb{R}^d \mid A(\theta) = \log \int_{\mathcal{X}^m} \exp\{\langle\theta, \phi(x)\rangle\} \nu dx < +\infty\}.$$

Short-hands: $\phi(x) = (\phi_\alpha(x), \alpha \in I) \in \mathbb{R}^d$, $p_\theta(x) = p_\theta(x_1, \cdots, x_m)$, $x = (x_1, \cdots, x_m)$

names of quantities:

ϕ: (log) potential functions, sufficient statistics

θ: canonical / exponential / natural parameters

$\mu = E_p[\phi(x)]$: mean parameters / moments

Other notions:

regular families: Λ is an open set

minimal: there is a unique parameter θ associated with each distribution.

Overcomplete: $\exists\, \theta \in \mathbb{R}^d$ s.t $\langle\theta, \phi(x)\rangle$ is a constant.

Properties of $A(\theta)$

Proposition 3.1 If $A(\theta)$ is associated with some regular family, then:

(a) $\dfrac{\partial A}{\partial \theta_\alpha} = E_\theta[\phi_\alpha(X)]$ (we short-hand $E_{p_\theta(x)}[\cdot]$ as $E_\theta[\cdot]$)

$\dfrac{\partial^2 A}{\partial\theta_\alpha \partial\theta_\beta} = \mathrm{cov}(\phi_\alpha(X), \phi_\beta(X))$

(b) A is a convex function of θ on Λ, and strictly so if the representation is minimal.

<u>Proof</u>. (a) is straight forward by calculus.

(b): from (a) we have the full Hessian $\nabla^2 A$ the covariance matrix, so $\nabla^2 A \succeq 0$; also if the family is minimal, then $\mathrm{Var}_\theta[\langle a, \phi(x)\rangle] = a^T \nabla^2 A(\theta)\, a > 0$, $\forall a \in \mathbb{R}^d, \theta \in \Lambda \Rightarrow \nabla^2 A \succ 0$. $\square$

Marginal Polytope

Recall the mean parameters' definition

$\mu = (\mu_\alpha, \alpha \in I)$, $\mu_\alpha = E_p[\phi_\alpha(X)]$ (for any valid distribution p)

We define the marginal polytope as those μ's realizable:

$$\mathcal{M} = \{\mu \in \mathbb{R}^d \mid \exists p, \text{ s.t } E_p[\phi(X)] = \mu\}$$

$\boxed{\text{properties}}$

(a) $\mathcal{M} = \text{conv}\{\phi(x), x \in \mathcal{X}^m\}$ (convex hull)

(proof by definition)

(b) $\exists \{(a_j, b_j) \in \mathbb{R}^d \times \mathbb{R} \mid j \in J\}$ with $|J| < +\infty$, s.t

$$\mathcal{M} = \{\mu \in \mathbb{R}^d \mid \langle a_j, \mu \rangle \geq b_j, \forall j \in J\}$$

example: consider an Ising model

$$p(x_1, x_2) \propto \exp\{b_1 x_1 + b_2 x_2 + W x_1 x_2\}, \quad x_1, x_2 \in \{0, 1\}$$

then (a) $\mu = \{E_p[x_1], E_p[x_2], E_p[x_1 x_2]\}$

$= \{p(x_1=1), p(x_2=1), p(x_1=1, x_2=1)\}$

$= p(x_1=0, x_2=0)\cdot(0,0,0) + p(x_1=1, x_2=0)\cdot(1,0,0)$
$+ p(x_1=0, x_2=1)\cdot(0,1,0) + p(x_1=1, x_2=1)\cdot(1,1,1)$

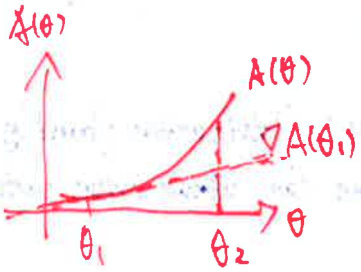
(b) also we need to constraint μ by

$0 \leq p(x_1=1, x_2=1) \leq p(x_i=1)$, $i = 1, 2$

and $1 + p(x_1=1, x_2=1) - p(x_1=1) - p(x_2=1) \geq 0$

(c) $\boxed{\text{Proposition 3.2}}$ from proposition 3.1 we have the gradient $\nabla A = E_\theta[\phi(x)]$ defines a mapping $\nabla A: \Lambda \to \mathcal{M}$; this mapping is one-to-one iff. the family is minimal.

Proof. 1) If the family is not minimal, then $\exists \gamma \in \mathbb{R}^d \setminus \{0\}$, s.t. $\langle \gamma, \phi(x) \rangle$ is a constant. We can choose some t s.t for a given θ_1, $\theta_2 = \theta_1 + t\gamma \in \Lambda$. Then we have p_{θ_1} and p_{θ_2} induce the same distribution, and so $\nabla A(\theta_1) = \nabla A(\theta_2)$.



2) Conversely if the family is minimal, then from proposition 3.1 A is strictly convex, so we have $\forall \theta_1 \neq \theta_2$,

$A(\theta_2) > A(\theta_1) + \langle \nabla A(\theta_1), \theta_2 - \theta_1 \rangle$, and by symmetricity,

$\langle \nabla A(\theta_1) - \nabla A(\theta_2), \theta_1 - \theta_2 \rangle > 0 \Rightarrow \nabla A(\theta_1) \neq \nabla A(\theta_2)$ $\square$

d) $\boxed{\text{Theorem 3.3}}$ If the family is minimal, then the gradient map ∇A is onto the interior $\mathcal{M}^o$, i.e. $\forall \mu \in \mathcal{M}^o$, $\exists\ \theta = \theta(\mu) \in \Lambda$ s.t. $E_\theta[\phi(X)] = \mu$.

Conjugate Duality

From Theorem 3.3 we know that ∇A is invertible on $\mathcal{M}^o$, let's consider the form of this inverse mapping.

Before that some more notations ~~are~~ is needed:

Conjugate dual of $A(\theta)$:

$$A^*(\mu) = \sup_{\theta \in \Lambda} \{ \langle \mu, \theta \rangle - A(\theta) \}, \quad \mu \in \mathbb{R}^d$$

connections to the maximum likelihood problem on exponential families: consider $D = \{ X_{(1)}, X_{(2)}, \cdots, X_{(N)} \}$, which induces $\mu = \sum_{i=1}^N \phi(X_{(i)})$, then the ML problem is

$$\sup_{\theta \in \Lambda} \sum_{i=1}^N \log p_\theta(X_{(i)}) = N \sup_{\theta \in \Lambda} \{ \langle \mu, \theta \rangle - A(\theta) \}.$$

Solving the ML problem returns optimum $A^*(\mu)$, and the optimizer θ satisfies

$$E_\theta[\phi(X)] = \nabla A(\theta) = \mu$$

Writing $\theta = \theta(\mu)$ for the corresponded optimizer as Thm. 3.3 suggested,

$\boxed{\text{Theorem 3.4}}$ (a) $\forall \mu \in \mathcal{M}^o$ and the corresponded $\theta(\mu)$

$$A^*(\mu) = -H(p_{\theta(\mu)})$$

(b) $$A(\theta) = \sup_{\mu \in \mathcal{M}} \{ \langle \theta, \mu \rangle - A^*(\mu) \}$$

(c) the supremum in (b) is attained uniquely at $\mu = E_\theta[\phi(X)] \in \mathcal{M}^o$

<u>Remark</u> ① We can solve the ML problem with any distribution family, but the solution in exponential family gives the maximum entropy.

② When the family is minimal & regular,

$$(\nabla A)^{-1} = \nabla A^*$$

Mean Field Methods

Consider an exponential family with $\phi = (\phi_\alpha, \alpha \in \mathcal{I})$ on graph $G=(V,E)$. In ~~learning~~ Inference we are interested in getting ~~$\theta = \theta_{opt}$~~ μ where θ ~~opt~~ is ~~calculated from data~~ given

Difficulties: ① from Thm. 3.4 we have $\theta = \nabla A^*(\mu)$, $\mu = \nabla A(\theta)$ but we generally don't know A^* and ∇A^* (also A)

② $A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle\theta, \mu\rangle - A^*(\mu)\}$, but we have little idea on the nature of $\mathcal{M}$.

Mean field methods approximates the distribution with tractable distributions: let F a subgraph of G, $\mathcal{I}(F) \subseteq \mathcal{I}$ the subset of sufficient statistics evaluated on F, then we can define an exponential family with ϕ on graph F, where the parameters belong to

$$\Lambda(F) = \{\theta \in \Lambda \mid \theta_\alpha = 0, \forall \alpha \in \mathcal{I} \backslash \mathcal{I}(F)\};$$

and the new "marginal" polytope defined on F:

$$\mathcal{M}_F(G;\phi) = \{\mu \in \mathbb{R}^d \mid \mu = E_\theta[\phi(x)], \forall \theta \in \Lambda(F)\}.$$

* $\mathcal{M}_F^\circ(G;\phi) \subseteq \mathcal{M}^\circ(G;\phi)$ since $\begin{cases} \nabla A(\Lambda) = \mathcal{M}^\circ(G;\phi) \\ \nabla A(\Lambda(F)) = \mathcal{M}_F^\circ(G;\phi) \end{cases}$

$\boxed{\text{Proposition 5.1}}$ (Mean field lower bound)

$A(\theta) \geq \langle\theta, \mu\rangle - A^*(\mu)$ for $\forall \mu \in \mathcal{M}^\circ$; equality holds iff. $\mu = E_\theta[\phi(x)]$.

Proof.
$$\begin{aligned} A(\theta) &= \log \int_{\mathcal{X}^m} \exp\{\langle\theta, \phi(x)\rangle\} dx \\ &\geq \int_{\mathcal{X}^m} q(x)[\langle\theta, \phi(x)\rangle - \log q(x)] dx && \text{(Jensen's)} \\ &= \langle\theta, \mu\rangle + H(q) && (\mu = E_q[\phi(x)]) \\ &= \langle\theta, \mu\rangle - A^*(\mu) && \text{(Thm. 3.4)} \end{aligned}$$ □

It says the log-partition function is obtained by optimizing the RHS objective. However as we don't know $A^*(\mu)$ where $\mu \in \mathcal{M}^\circ$ (also the structure of $\mathcal{M}^\circ$), we restrict the searching space to $\mathcal{M}_F(G;\phi)$ (short-handed as $\mathcal{M}_F(G)$). Denote $A_F^* = A^*|_{\mathcal{M}_F(G)}$, we get the best lower bound within $\mathcal{M}_F(G)$:

$$\max_{\mu \in \mathcal{M}_F(G)} \{\langle\mu, \theta\rangle - A_F^*(\mu)\}$$

KL-divergence

We write $D(\theta_1 \| \theta_2) = D(p_{\theta^1} \| p_{\theta^2})$

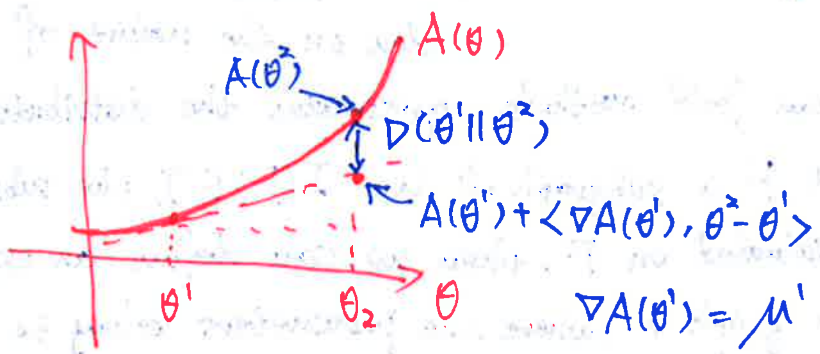
then little calculation reveals that: $\quad (\mu^i = E_{p_{\theta^i}}[\phi(X)],\ i=1,2)$

$$D(\theta^1 \| \theta^2) = A(\theta^2) - A(\theta^1) - \langle \mu^1, \theta^2 - \theta^1 \rangle \quad \text{(primal form)}$$

$$= A(\theta^2) + A^*(\mu^1) - \langle \mu^1, \theta^2 \rangle \quad \text{(mixed form)}$$

$$= A^*(\mu^1) - A^*(\mu^2) - \langle \theta^2, \mu^1 - \mu^2 \rangle \quad \text{(dual form)}$$

· primal form:



mixed form: as $D(\theta_1 \| \theta_2) \geq 0$, this also verifies that

$$A(\theta) \geq \langle \theta, \mu \rangle - A^*(\mu), \quad \forall \mu \in \mathcal{M}^\circ$$

dual form: change A to A^* in the figure of primal form.

back to mean field

$$\max.\ \langle \mu, \theta \rangle - A^*_F(\mu) \iff \min D(\mu \| \theta) = A(\theta) + A^*_F(\mu) - \langle \mu, \theta \rangle$$

all subject to $\mu \in \mathcal{M}_F(G)$

example: Naive mean field for Ising model

short-hand $\mu_s = E[X_s] = P(X_s = 1)$, $\mu_{st} = E[X_s X_t] = P[X_s=1, X_t=1]$

then $\mathcal{M}_{F_0}(G) = \{ \mu \in \mathbb{R}^{|V|+|E|} \mid 0 \leq \mu_s \leq 1, \forall s \in V$, and $\mu_{st} = \mu_s \mu_t, \forall (s,t) \in E \}$

$(F_0 = (V, \phi))$

and now $-A^*_{F_0}(\mu) = H(p_{\theta(\mu)}) = \sum_{s\in V} H_s(\mu_s)$.

the optimization problem reduces to

$$\max_{\mu} \left\{ \sum_{s\in V} \theta_s \mu_s + \sum_{(s,t)\in E} \theta_{st} \mu_s \mu_t + \sum_{s \in V} H_s(\mu_s) \right\}$$

solution by coordinate descent:

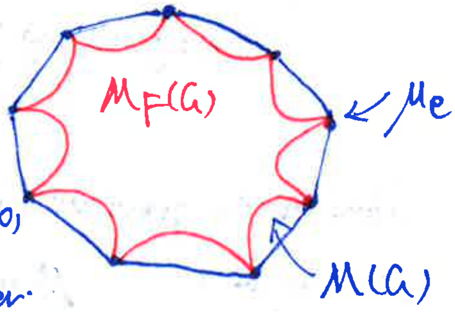
$$\mu_s \leftarrow \sigma\Big(\theta_s + \underbrace{\sum_{t\in N(s)}}_{\text{neighbour set}} \theta_{st} \mu_t\Big)$$

Non-convexity of mean field

from definition of the marginal polytope

$\mathcal{M}(G) = \mathrm{conv}\{\phi(e), e \in \mathcal{X}^m\}$

where e is the basis vector $(0,\cdots,0,1,0\cdots0)$



we know that μ_e is the mean parameter that the corresponded p_θ put all mass on a single element x_m.

Such a point belongs to $\overline{\mathcal{M}_F(G)}$ (easy to prove)

So if $\mathcal{M}_F(G) \subsetneq \mathcal{M}(G)$, then $\mathcal{M}_F(G)$ should be non-convex.

Variational Methods

maximum likelihood estimation

Given observation $D = \{X_{(1)}, \cdots, X_{(N)}\}$, we can compute the empirical expectation

$$\hat{\mu} = \hat{E}[\phi(X)] = \frac{1}{N}\sum_{i=1}^{N}\phi(X_{(i)})$$

and the MLE problem reduces to

$$l(\theta; D) = \langle\theta, \hat{\mu}\rangle - A(\theta).$$

solution: $\hat{\theta}$ satisfies $E_{\hat{\theta}}[\phi(X)] = \hat{\mu}$

we know that $\lim_{N\to\infty}\hat{\mu} = \mu$ (Law of Large Numbers)

and MLE is an unbiased estimator:

$$\lim_{N\to\infty}\hat{\theta} = \theta \quad \text{where } \theta \text{ satisfies } E_\theta[\phi(X)] = \mu.$$

exact EM

define a joint exponential family distribution for (X, Y):

$$p_\theta(x,y) = \exp\{\langle\theta, \phi(x,y)\rangle - A(\theta)\}$$

where Y is observed and X is some hidden/latent variable:

given an observation $Y=y$, we get the conditional distribution

$$p_\theta(x|y) = \exp\{\langle\theta, \phi(x,y)\rangle - A_y(\theta)\}$$

$$A_y(\theta) = \log\int_{\mathcal{X}^m}\exp\{\langle\theta, \phi(x,y)\rangle\}dx$$

MLE problem on models with hidden variables:

$$l(\theta;y) = \log p_\theta(y) = \log \int_X \exp\{\langle\theta,\phi(x,y)\rangle - A(\theta)\}dx$$
$$= A_y(\theta) - A(\theta).$$

Similar to previous definitions of mean parameter, marginal polytope, etc.

We define: $\mathcal{M}_y = \{\mu \in \mathbb{R}^d \mid \mu = E_p[\phi(X,y)] \text{ for some } p\}$

$$\mu_y = E_{p_\theta}[\phi(X,y)]$$

$\Rightarrow$ (dual) $\quad A_y(\theta) = \sup_{\mu_y \in \mathcal{M}_y} \{\langle\theta, \mu_y\rangle - A_y^*(\mu_y)\}$

$$A_y^*(\theta) = \sup_{\theta \in dom(A_y)} \{\langle\mu_y, \theta\rangle - A_y(\theta)\}$$

$\Rightarrow$ extension of proposition 5.1:

$$A_y(\theta) \geq \langle\mu_y,\theta\rangle - A_y^*(\mu_y), \quad \forall \mu_y \in \mathcal{M}_y$$

so we put a lower bound to the MLE problem as

$$l(\theta;y) \geq \langle\mu_y,\theta\rangle - A_y^*(\mu_y) - A(\theta) \triangleq L(\theta;y),$$

and update μ_y, θ with EM:

E-step: $\mu_y^{(t+1)} = \arg\max_{\mu_y \in \mathcal{M}_y} L(\mu_y, \theta^{(t)})$

$\left(\max_{\mu_y \in \mathcal{M}_y} \{\langle\mu_y,\theta^{(t)}\rangle - A_y^*(\mu_y)\}\right)$ solution: $\mu_y^{(t+1)} = \underline{E_{\theta^{(t)}}[\phi(X,y)]}$ on the conditional

which makes the bound tight: $l(\theta^{(t)};y) = L(\theta^{(t)};y)$

M-step: $\theta^{(t+1)} = \arg\max_{\theta \in \Lambda} L(\mu_y^{(t+1)}, \theta)$

$\left(\max_{\theta\in\Lambda} \{\langle\mu_y^{(t+1)},\theta\rangle - A(\theta)\}\right)$ solution: $\theta^{(t+1)} = \theta(\mu_y^{(t+1)})$, i.e. $\underline{\mu_y^{(t+1)} = E_{\theta^{(t+1)}}[\phi(X,Y)]}$

increase L but also make $l \geq L$ again $\leftarrow$ important: expectation on the joint distribution

Variational EM: what if we cannot compute μ_y?

Recall the lower bound $L(\theta;y)$, where we search optimizers in $\mathcal{M}_y$ in E-step. Now $\mathcal{M}_y$ is generally unknown, but we can use the subset $\mathcal{M}_{F,y}(G) \subsetneqq \mathcal{M}_y$ to approximately solve E-step.

Since the problem is solved in a subset, we generally have

$$L(\mu_y^{(t+1)}, \theta^{(t)}) \leq l(\theta^{(t)};y) \quad \text{where } \mu_y^{(t+1)} = \arg\max_{\mu_y \in \mathcal{M}_{F,y}(G)} L(\mu_y, \theta^{(t)}).$$

Variational Bayes

In previous notes we assume θ an unknown parameter to estimate, now we introduce Bayesian inference and consider it as a random variable.

Model: let Y be the observed random variable and X be some hidden variable. Assume the joint distribution / likelihood is

$$p(x,y|\theta) = \exp\{\langle \eta(\theta), \phi(x,y)\rangle - A(\eta(\theta))\} \qquad (\eta: \mathbb{R}^d \to \mathbb{R}^d)$$

and the prior distribution is

$$p_{\xi,\lambda}(\theta) = \exp\{\langle \xi, \eta(\theta)\rangle - \lambda A(\eta(\theta)) - B(\xi,\lambda)\}.$$

sufficient statistics   log-partition function

Given a parameterization of the prior (ξ^*, λ^*), Bayesian inference compute the marginal likelihood

$$\begin{aligned}
\log p_{\xi^*,\lambda^*}(y) &= \log \int p_{\xi^*,\lambda^*}(\theta)\, p(y|\theta)\, d\theta \\
&= \log \int p_{\xi^*,\lambda^*}(\theta)\, p(y|\theta)\, \frac{p_{\xi,\lambda}(\theta)}{p_{\xi,\lambda}(\theta)}\, d\theta \\
&\geq E_{\xi,\lambda}[\log p(y|\theta)] + E_{\xi,\lambda}\left[\log \frac{p_{\xi^*,\lambda^*}(\theta)}{p_{\xi,\lambda}(\theta)}\right] \\
&= E_{\xi,\lambda}[A_y(\eta(\theta)) - A(\eta(\theta))] + E_{\xi,\lambda}\left[\log \frac{p_{\xi^*,\lambda^*}(\theta)}{p_{\xi,\lambda}(\theta)}\right] \\
\text{(proposition 5.1)} \quad &\geq E_{\xi,\lambda}[\langle \mu(\theta), \eta(\theta)\rangle - A_y^*(\mu(\theta)) - A(\eta(\theta))] +
\end{aligned}$$

Now we short-hand: $\bar{\eta} = E_{\xi,\lambda}[\eta(\theta)]$, $\bar{A} = E_{\xi,\lambda}[A(\eta(\theta))]$, $\mu = \mu(\theta)$

then the lower bound reduces to:

$$[\langle \mu, \bar{\eta}\rangle - A_y^*(\mu) - \bar{A}] + \langle \bar{\eta}, \xi^* - \xi\rangle + \langle -\bar{A}, \lambda^* - \lambda\rangle - B(\xi^*, \lambda) + B(\xi, \lambda),$$

also since $B^*(\bar{\eta}, \bar{A}) = \langle \bar{\eta}, \xi\rangle + \langle -\bar{A}, \lambda\rangle - B(\xi, \lambda)$, we have the lower bound changes to

$$\langle \mu + \xi^*, \bar{\eta}\rangle - A_y^*(\mu) + \langle \lambda^* + 1, -\bar{A}\rangle - B^*(\bar{\eta}, \bar{A}) \triangleq L(\mu, \bar{\eta}, \bar{A})$$

and now we can do EM on this objective.

VB-E-step: $\mu^{(t+1)} = \underset{\mu \in M_y}{\arg\max}\, L(\mu, \bar{\eta}^{(t)}, \bar{A}^{(t)}) = \underset{\mu \in M_y}{\arg\max}\, \langle \mu, \bar{\eta}^{(t)}\rangle - A_y^*(\mu)$

solution: $\mu^{(t+1)} = E_{\bar{\eta}^{(t)}}[\phi(X, y)]$

VB-M-step: $(\bar{\eta}^{(t+1)}, \bar{A}^{(t+1)}) = \underset{(\bar{\eta}, \bar{A})}{\arg\max}\ L(\mu^{(t+1)}, \bar{\eta}, \bar{A})$

$$= \underset{(\bar{\eta}, \bar{A})}{\arg\max} \{ \langle \mu^{(t+1)} + \xi^*, \bar{\eta} \rangle - (1+\lambda^*)\bar{A} - B^*(\bar{\eta}, \bar{A}) \}$$

$\Rightarrow$ then $(\bar{\eta}^{(t+1)}, \bar{A}^{(t+1)}) = E_{\xi^{(t+1)}, \lambda^{(t+1)}}[(\bar{\eta}, \bar{A})]$

where we can get $(\xi^{(t+1)}, \lambda^{(t+1)}) = (\mu^{(t+1)} + \xi^*, \lambda^* + 1)$

<u>In practice</u>: $p_{\xi,\lambda}(\theta)$ is interpreted as an approximation of $p(\theta|y)$, and we often assume family $\{p_{\xi,\lambda}(\theta)\}$ is minimal, hence the VB-M-step is valid (i.e. one-to-one correspondence between (ξ, λ) and $(\bar{\eta}, \bar{A})$).

However the solution for $\lambda \equiv \lambda^* + 1$, which means if we choose $\lambda^* = 0$ (as we often do), the posterior approximation is also intractable.

Also in computations of μ, we need a tractable $\bar{\eta} = E_{\xi,\lambda}[\eta(\theta)]$, i.e. we may want $\lambda^* + 1 = 0$, i.e. set $\lambda^* = -1$. But this is impossible: as we extend VB to N data points, update will change to $\lambda^{(t+1)} \leftarrow \lambda^* + N$, i.e. λ^* set as $-N$ $\Rightarrow$ the prior ~~is~~ depends on the observations, which is counter intuitive!