Machine learning problem

Data $D = \{ (x_1, y_1), \ldots, (x_n, y_n) \}$ — input data, label

Decision theory: Minimize $\mathbb{E}[L(f(x), y)]$ — prediction function; $\hat{y}$ output

Optimize decision when expected difference between prediction function $f(x)$ and actual value $\hat{y}$ is minimized

Key quantity $p(\hat{y} | x)$ vs. y for x

Generative model assumption:

Data is generated from a process

=> Model the data generating process itself → Richer model

Joint distribution: $p(x, y)$

Factorization:

$$p(x, y) = p(y|x)\,p(x) \Rightarrow \boxed{p(y|x) = \frac{p(x,y)}{p(x)}}$$ Predictive distribution

or

$$\boxed{p(x, y) = p(x|y)\,p(y)}$$ Generative distribution

② Sample x given y ① Sample y

Marginal distribution (w.r.t θ) for predictive distribution

$$p(y|x, D) = \int p(y, \theta | x, D)\, d\theta$$

Posterior distr. GIVEN a parameter

Assuming parameters independent of input x

$$= \int p(y \mid \theta, x, D)\, p(\theta \mid D)\, d\theta$$

Hard — Easy — Hard — Posterior distr. of (all) parameters

$\int \ldots d\theta$ intractable

⇒ Approx. integral w. "Deterministic approximation" approach

- Laplace approx. (Distr. of θ ⇒ Gaussian approx.)
- Variational methods (Find approx. using optimization)

Estimating parameter distribution using Bayesian inference

$$p(\theta \mid D) = \frac{P(\theta, D)}{P(D)} = \frac{P(D \mid \theta) P(\theta)}{P(D)}$$

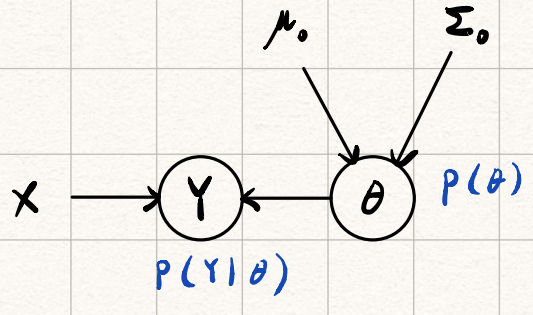
i.e. Data $\{x_0, \ldots, x_n\}$ generated by modeled generative process

Likelihood of data: $P(D \mid \theta) = P(x_0, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} [P(x_i \mid \theta)]$

$$p(\theta \mid D) = \frac{\prod_{i=1}^{n} [P(x_i \mid \theta)]\, P(\theta)}{P(D)}$$

← Prior parameter distribution

Bayesian Logistic Regression



- Want the posterior distribution for θ which maximizes the likelihood of the data $D = \{(x_0, y_0), \ldots, (x_n, y_n)\}$

$$p(\theta \mid D) = \frac{P(\theta, D)}{P(D)} = \frac{P(D \mid \theta) P(\theta)}{P(D)} \propto P(D \mid \theta) P(\theta)$$

"Brave" step to skip the normalization term

(Not important as $p(D|\theta)p(\theta)$ eventually approx. as Gaussian)

- Goal: Get an expression for $p(\theta|D) \propto \underbrace{p(D|\theta)p(\theta)}_{?}$

① Assume a prior parameter distribution that is Gaussian

Parameter Vector ↓ Hyperparameters ("Known") ↓

$$p(\theta) = \mathcal{N}(\theta | \mu_0, \Sigma_0)$$

② Assume logistic regression formulation for likelihood function

$$p(y|\theta) = \prod_{i=1}^{n} p(C_1 = 1 | x_i, \theta)^{y_i} \, p(C_2 = 1 | x_i, \theta)^{1-y_i}$$

$$= \prod_{i=1}^{n} f(x_i, \theta)^{y_i} \left(1 - f(x_i, \theta)\right)^{1-y_i}$$

$$= \prod_{i=1}^{n} (\hat{y}_i)^{y_i} (1 - \hat{y}_i)^{1-y_i}$$

Data 'x' is subsumed by output '$\hat{y}$'

↳ $\hat{y}_i = \sigma(\theta^T x_i)$

$$p(\theta|D) \propto \prod_{i=1}^{n} (\hat{y}_i)^{y_i} (1 - \hat{y}_i)^{1-y_i} \cdot \mathcal{N}(\theta | \mu_0, \Sigma_0)$$

- Approximation approach: Laplace approximation (i.e. Gaussian approx.)
  - Gaussian w. mean = mode
    variance = maximize likelihood of data

1. Mode of $P(\theta|D)$ → MAP solution θ_{MAP}

$$\log p(\theta|D) \propto \sum_{i=1}^{n} [y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i)]$$

$$-\frac{1}{2}(\theta-\mu_0)^T \Sigma_0^{-1}(\theta-\mu_0) + \text{const.}$$

$$\Rightarrow \boxed{\theta_{MAP} = f(\mu_0, \Sigma_0, D)}$$

2. Variance of $P(\theta|D)$ → Inverse of negative log likelihood Hessian matrix (second derivative)

$$\Rightarrow \boxed{\Sigma_n = -\nabla^2 \log p(\theta|D) = \Sigma_0^{-1} + \sum_{i=1}^{n} \hat{y}_i(1-\hat{y}_i) x_i x_i^T}$$

- Finish: Gaussian approximation to posterior distr. $p(\theta|D)$

$$* \quad \boxed{q(\theta) = \mathcal{N}(\theta \mid \theta_{MAP}, \Sigma_n)}$$

Predictive distribution for Bayesian Logistic Regression

$$p(Y=C_1 \mid x^{new}, D) = \int p(C_1, \theta \mid x^{new}, D)\, d\theta$$

$$= \int p(C_1 \mid x^{new}, \theta)\, p(\theta|D)\, d\theta$$

$$\simeq \int \sigma(\theta^T x^{new})\, q(\theta)\, d\theta$$

Integrating this expression using Dirac function substitution

$$\sigma(\theta^T x^{new}) = \int \delta(a - \theta^T x^{new})\, \sigma(a)\, da$$

⟨ Confusing part ... ⟩

$$\int \sigma(\theta^T x^{new})\, q(\theta)\, d\theta = \int \sigma(a)\, p(a)\, da$$

$$p(a) = \int \delta\left(a - \theta^T x^{new}\right) q(\theta)\, d\theta \equiv \int q(\theta)\, d\theta \quad \text{when} \quad a = \theta^T x^{new}$$

$$\mathbb{E}[a] = \int p(a)\, a\, da = \int q(\theta)\, \theta^T x^{new}\, d\theta = \boxed{\theta_{MAP}^T x^{new} = \mu_a}$$

$$\text{var}[a] = \int p(a)\left(a^2 - \mathbb{E}[a]^2\right) da = \int q(\theta)\left[(\theta^T x^{new})^2 - (\theta_{MAP}^T x^{new})^2\right] d\theta$$

$$= \boxed{(x^{new})^T \Sigma_n\, x^{new} = \sigma_a^2}$$

Variational approximation to the predictive distribution:

$$p(Y = C_1 \mid x^{new}, D) \simeq \int \sigma(a)\, \mathcal{N}(a \mid \mu_a, \sigma_a^2)\, da$$

Sigmoid function CANNOT be integrated analytically w. Gaussian

=> Approximate $\sigma(a)$ by Probit function $\Phi(\lambda a)$ (which CAN be integrated analytically!)

- What is λ? Coefficient for adjusting approximation

Requiring $\sigma(a)$, $\Phi(\lambda a)$ to have same derivative (i.e. slope) at origin

$$\Rightarrow \boxed{\lambda^2 = \frac{1}{8}\pi}$$

- Integration

$$\lambda \cdot \frac{\mu_a}{(\lambda^2/\lambda^{-2} + \lambda^2\sigma_a)^{1/2}}$$

$$\int \underbrace{\Phi(\lambda \cdot a)}_{\simeq \sigma(a)} \mathcal{N}(a|\mu_a, \sigma_a^2)\, da = \underbrace{\Phi\left(\frac{\mu_a}{(\lambda^{-2} + \sigma_a^2)^{1/2}}\right)}_{\simeq \sigma(a^*)}$$

Again approximating

$$\sigma(a) \simeq \Phi(\lambda a)$$

$$\Rightarrow \int \sigma(a)\, \mathcal{N}(a|\mu_a, \sigma_a^2) = \sigma(a^*)$$

$$a^* = \frac{\mu_a}{(1 + \lambda^2\sigma_a^2)^{1/2}} = \underbrace{\mu_a\left(1 + \frac{1}{8}\pi\sigma_a^2\right)^{-1/2}}_{K(\sigma_a^2)}, \quad \lambda^2 = \frac{1}{8}\pi$$

$\Rightarrow$ Final approx. predictive distribution

$*$

$$\boxed{p(Y = C_1 | x^{new}, D) = \sigma\left(K(\sigma_a^2) \cdot \mu_a\right)}$$

Variational Logistic Regression w. set priors

- Improvement on regular Bayesian Log. Reg:
  - Better Gaussian approx. of posterior distr. $p(\theta)$ than Laplace approx.
    - Optimize lower bound of approx. and true distr.

      (NOT restricting mode = mean, etc.!)

  => Greater flexibility, greater accuracy

  (param. distr. can become more optimal)

=> Approx. posterior normal param. distr. $q(\theta)$ incorporating prior $p(\theta)$ and evidence $p(D|\theta)$

Goal: Approx. logistic regression likelihood func. by "Exponential of quadratic term" ~ Gaussian-like func.

$$e^{[ax^2+bx+c]} = e^{[a(x^2+\frac{b}{a}x)+c]}$$

$$= e^{[a(x^2+2\frac{b}{2a}x+\frac{b^2}{4a^2})-\frac{b^2}{4a^2}+c]}$$

$$= e^{[a(x+\frac{b}{2a})^2-\frac{b^2}{4a^2}+c]}$$

$$= e^{(c-\frac{b^2}{4a^2})} e^{a(x+\frac{b}{2a})^2}$$

$$\equiv \underbrace{(2\pi\sigma^2)^{-\frac{1}{2}}}_{\eta} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$\eta = e^{(c-\frac{b^2}{4a^2})}$

$-\frac{1}{2\sigma^2} = a$

$\mu = -\frac{b}{2a}$

$$\Rightarrow \sigma(x) \simeq e^{(ax^2+bx+c)} \equiv \mathcal{N}(\mu, \sigma^2)$$

* Known analytical integration

=> Lower bound to optimize

Marginal likelihood of Bayesian logistic regression model:

* Likelihood of data over all models

Likelihood of data conditioned on a model θ

$$p(D) = \int p(D,\theta)\,d\theta = \int p(D|\theta)p(\theta)\,d\theta = \int \left[\prod_{i=1}^{N} p(\hat{y}_i|\theta)\right] p(\theta)\,d\theta$$

Rewriting $p(\hat{y}|\theta)$

Known "family of models" ~ Parametric function

$$= \sigma(\underbrace{\theta^T x}_{a})^y (1-\sigma(\underbrace{\theta^T x}_{a}))^{(1-y)}$$

Substitution: $a = \theta^T x$

$$= \sigma(a)^y (1-\sigma(a))^{(1-y)}$$

$$= \left(\frac{1}{1+e^{-a}}\right)\left(1 - \frac{1}{1+e^{-a}}\right)$$

$$\vdots \quad ?$$

$$\boxed{p(\hat{y}|\theta) = e^{ay}\,\sigma(-a)}$$

- Obtaining a lower bound on $p(D)$ by using known lower bound on $\sigma(\cdot)$ $\qquad q(x, \xi)$

*

$$\boxed{\sigma(z) \geq \sigma(\xi)\exp\left[\frac{z-\xi}{2} - \lambda(\xi)(z^2 - \xi^2)\right]}$$

Optimal variational parameter λ which gives optimal lower bound approximation $q(x, \xi)$ for $\sigma(x)$

$$\lambda(\xi) = \frac{1}{2\xi}\left[\sigma(\xi) - \frac{1}{2}\right]$$

Expressing likelihood of data $p(\hat{y}|\theta)$ in terms of the lower bound $q(\cdot)$ for $\sigma(\cdot)$

$$p(\hat{y}_i|\theta) = e^{a_i y_i}\sigma(-a_i) \geq e^{a_i y_i}\sigma(\xi_i)\exp\left(-\frac{a_i - \xi_i}{2} - \lambda(\xi_i)(a_i^2 - \xi_i^2)\right)$$

Sample-specific parameters: $x_i \rightarrow a_i, y_i, \xi_i$

Expressing the joint distribution (of all data) using the lower bound $q(\cdot)$ for $\sigma(\cdot)$

$$p(D, \theta) = p(D|\theta)\,p(\theta) = \left[\prod_{i=1}^{N} p(\hat{y}_i|\theta)\right] p(\theta)$$

$$\geq \left[\prod_{i=1}^{N} \exp(\theta^T x_i \cdot y)\, \sigma(\xi_i) \exp\left(-\tfrac{1}{2}(\theta^T x_i - \xi_i) - \lambda(\xi_i)\left([\theta^T x_i]^2 - \xi_i^2\right)\right)\right] p(\theta)$$

"Label y" term "x = ξ" term "Variational param. λ" term

$$= \underbrace{\left[\prod_{i=1}^{N} \sigma(\xi_i) \exp\left(\theta^T x_i \cdot y_i - \tfrac{1}{2}(\theta^T x_i - \xi_i) - \lambda(\xi_i)\left([\theta^T x_i]^2 - \xi_i^2\right)\right)\right]}_{h(\theta, \xi)} p(\theta)$$

← Parameters that vary in the functional?

Log of the joint distribution $p(D, \theta)$

Why? – Because it is easier to manipulate the expressions

$\log(a \cdot b) = \log a + \log b$

$$\log p(D, \theta) \geq$$

$$\log \left[\prod_{i=1}^{N} \sigma(\xi_i) \exp\left(\theta^T x_i \cdot y_i - \tfrac{1}{2}(\theta^T x_i - \xi_i) - \lambda(\xi_i)\left([\theta^T x_i]^2 - \xi_i^2\right)\right)\right] p(\theta)$$

$$= \log p(\theta) + \sum_{i=1}^{N}\left[\log \sigma(\xi_i) + \theta^T x_i y_i - \tfrac{1}{2}(\theta^T x_i - \xi_i) - \lambda(\xi_i)\left([\theta^T x_i]^2 - \xi_i^2\right)\right]$$

Assuming a Gaussian prior for parameters (like in Bayesian log. reg.)

$$p(\theta) = \eta(\Sigma_0)\, e^{-\frac{1}{2}(\theta - \mu_0)^T \Sigma_0^{-1} (\theta - \mu_0)}$$

$$\log p(\theta) = \log \eta(\Sigma_0) - \tfrac{1}{2}(\theta - \mu_0)^2 \Sigma_0^{-1} (\theta - \mu_0)$$

Inserting into the joint prob. expression

$$\log p(D, \theta) \geq \log \eta(\Sigma_0) - \tfrac{1}{2}(\theta - \mu_0)^2 \Sigma_0^{-1} (\theta - \mu_0)$$
$$+ \sum_{i=1}^{N}\left[\log \sigma(\xi_i) + \theta^T x_i y_i - \tfrac{1}{2}(\theta^T x_i - \xi_i) - \lambda(\xi_i)\left([\theta^T x_i]^2 - \xi_i^2\right)\right]$$

Optimizing θ => Discard terms NOT affecting θ

$$\log p(D,\theta) \geq -\frac{1}{2}(\theta-\mu_0)^T \Sigma_0^{-1}(\theta-\mu_0) + \sum_{i=1}^{N}\left[\theta^T x_i y_i - \frac{1}{2}(\theta^T x_i - \xi_i) - \lambda(\xi_i)\,\theta^T(x_i x_i^T)\theta\right] + \text{const}$$

POINT: The above joint distribution $p(D,\theta)$ encompass the prior $p(\theta)$ and evidence $p(D|\theta)$

=> Approx. posterior $q(\theta)$ incorporating both

$$\exp(\log p(D,\theta)) \geq q(\theta) = \exp(\ldots)$$

$$e^{\left(c-\frac{b^2}{4a}\right)} e^{a\left(x+\frac{b}{2a}\right)^2} \qquad (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$$\Rightarrow \boxed{q(\theta) = \mathcal{N}(\theta|\mu_N, \Sigma_N)}$$

$$\mu_N = \Sigma_N\left(\Sigma_0^{-1}\mu_0 + \sum_{i=1}^{N}\left(y_i - \frac{1}{2}\right)x_i\right)$$

$$\Sigma_N = \Sigma_0^{-1} + 2\sum_{i=1}^{N}\lambda(\xi_i)\,x_i x_i^T$$

* Need to read original paper to see how to connect with normal form ... ?

POINT: Because $q(\theta)$ has a Gaussian form => Already normalized

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \quad \text{(posterior; Normalization)} \quad \Leftrightarrow \quad p(D)\,p(\theta|D) = p(D|\theta)p(\theta)$$

Evidence, prior; $p(D)$: "negligible"; $p(\theta|D) \geq q(\theta) = \mathcal{N}(\,)$

$$\Rightarrow q(\theta) \leq p(D|\theta)p(\theta)$$

Determine variational parameters ξ_i

How? Maximize the marginal log likelihood lower bound

$$\log p(D) = \log \int p(D|\theta)p(\theta)\,d\theta \geq \log \int h(\theta, \xi)\, p(\theta)\, d\theta = \mathcal{L}(\xi)$$

↑ Use the EM algorithm to optimize

- Latent variable is θ

⓪ Initialize $\xi^{old} = \{\xi_1, \ldots, \xi_N\}$ to some value

① E step : Use ξ^{old} to find "best-estimate" posterior distribution $q(\theta)$ (i.e. the latent params.)

⇒ Compute μ_N, Σ_N using $(D, \mu_0, \Sigma_0, \xi^{old})$

⇒ Current "best-estimate" $q(\theta) = \mathcal{N}(\theta|\mu_N, \Sigma_N)$

② M-step : Maximize expected complete-data log likelihood w.r.t. current "best estimate" $q(\theta)$

"Standard variational method approach" REPEAT !

$$Q(\xi, \xi^{old}) = \mathbb{E}_\xi[\mathcal{L}(\xi)] = \mathbb{E}_\xi[\log h(\theta, \xi)\, p(\theta)]$$

$$= \mathbb{E}_\xi[\log h(\theta, \xi) + \log p(\theta)]$$

→ const. as NOT depending on ξ

$$= \mathbb{E}_\xi[\log h(\theta, \xi)] + \text{const.}$$

$$= \mathbb{E}_\xi \sum_{i=1}^{N} \left[\log \sigma(\xi_i) + \theta^T x_i y_i - \tfrac{1}{2}(\theta^T x_i - \xi_i) - \lambda(\xi_i)\left([\theta^T x_i]^2 - \xi_i^2\right)\right] + \text{const.}$$

→ const → const

$$= E_\xi \sum_{i=1}^{N} \left[\log \sigma(\xi_i) - \tfrac{1}{2}\xi_i - \lambda(\xi_i)\left(x_i^T E[\theta\theta^T] x_i - \xi_i^2\right)\right] + \text{const.}$$

Taking the derivative to find maximum of expression

Applying dd. of $\sigma(\xi)$, $\lambda(\xi)$

$$\frac{d}{d\xi_i}\left[Q(\xi, \xi^{old})\right] = \ldots = \lambda'(\xi_i)\left(x_i^T E[\theta\theta^T] x_i - \xi_i^2\right) = 0$$

- As $\lambda'(\xi) \neq 0$ as it is
  - A monotonic function
  - Symmetric around $\xi = 0$

$$\Rightarrow x_i^T E[\theta\theta^T] x_i - \xi_i^2 = 0$$

Expectation identity:

$$E[x^2] = E[x]^2 + \text{var}[x]$$

Re-estimation equation:

$$\left(\xi_i^{new}\right)^2 = x_i^T E[\theta\theta^T] x_i = x_i^T\left(\Sigma_N + \mu_N \mu_N^T\right) x_i$$

Evaluating the lower bound $\mathcal{L}(\xi)$:

- Possible to do analytically because
  ① $p(\theta)$ is Gaussian
  ② $h(\theta, \xi)$ is "exponential of a quadratic function of θ"
  $\Rightarrow$ Can be formulated as a Gaussian

How? Details in original paper?

$$\mathcal{L}(\xi) = \int h(\theta, \xi)\, p(\theta)\, d\theta$$

$$= \frac{1}{2}\log\frac{|\Sigma_N|}{|\Sigma_0|} - \frac{1}{2}\mu_N^T \Sigma_N^{-1} \mu_N + \frac{1}{2}\mu_0^T \Sigma_0^{-1} \mu_0 + \sum_{i=1}^{N}\left[\log\sigma(\xi_i) - \frac{1}{2}\xi_i - \lambda(\xi_i)\xi_i^2\right]$$

Extension: Inferring hyperparameters from data

- Approach: Combining 'Global' and 'Local' variational approximation

Extendable to prior with nonzero mean?

Parameter prior assumption: $p(\theta) = \mathcal{N}(\theta \mid 0, \alpha^{-1} I)$

$\mu_0 = 0, \Sigma_0 = \alpha^{-1} I$

Hyperprior assumption: $p(\alpha) = \text{Gam}(\alpha \mid a_0, b_0)$

Joint distribution: $p(y, \theta, \alpha) = p(y \mid \theta) p(\theta \mid \alpha) p(\alpha)$

Marginal likelihood: $p(D) = \iint p(y, \theta, \alpha) \, d\theta \, d\alpha$

Approximate using both 'Global' and 'Local' variational approaches in same model

Introducing variational distribution $q(\theta, \alpha)$

Repeat...

- General formulation:

Lower bound ↓ Diff. between lower bound and true distr. ~ Remainder

$$\log p(D) = \mathcal{L}(q) + KL(q \| p)$$

$$\mathcal{L}(q) = \iint q(\theta, \alpha) \log \left[\frac{p(\theta, \alpha, y)}{q(\theta, \alpha)} \right] d\theta \, d\alpha$$

$$KL(q \| p) = -\iint q(\theta, \alpha) \log \left[\frac{p(\theta, \alpha \mid y)}{q(\theta, \alpha)} \right] d\theta \, d\alpha$$

POINT: Intractability of $\mathcal{L}(q)$ stems from $p(y \mid \theta)$ w. the logistic sigmoid ($p(\theta \mid \alpha) p(\alpha)$ analytically OK)

=> Apply the Local variational bound

$$p(y, \theta, \alpha) = p(y|\theta)\, p(\theta|\alpha)\, p(\alpha)$$
$$\geq h(\theta, \xi)\, p(\theta|\alpha)\, p(\alpha)$$

w. local approx. of sigmoid

$$\Rightarrow \log p(D) \geq \mathcal{L}(q) \geq \tilde{\mathcal{L}}(q, \xi)$$

$$= \iint q(\theta, \alpha) \log \left[\frac{h(\theta, \xi)\, p(\theta|\alpha)\, p(\alpha)}{q(\theta, \alpha)} \right] d\theta\, d\alpha$$

Mean Field Approximation

$$q(\theta, \alpha) = q(\theta)\, q(\alpha)$$

$$\log q_j^*(z_j) = \mathbb{E}_{-j}[\log p(X, Z)] + \text{const}$$

Applying the "general result"

=> Optimal factors for $q(\theta)$, $q(\alpha)$

① $q(\theta)$:

$$\log q(\theta) = \mathbb{E}_\alpha [\log p(y, \theta, \alpha)] + \text{const}$$

$$\simeq \mathbb{E}_\alpha [\log (h(\theta, \xi)\, p(\theta|\alpha)\, p(\alpha))] + \text{const}$$

$$= \mathbb{E}_\alpha [\log h(\theta, \xi) + \log p(\theta|\alpha) + \log p(\alpha)] + \text{const}$$

Independent of $\theta \rightarrow$ const

$$= \log h(\theta, \xi) + \mathbb{E}_\alpha[\log p(\theta|\alpha)] + \text{const}$$

Known Known

$$= \sum_{i=1}^{N}\Big[\underbrace{\log \sigma(\xi_i)}_{\text{Independent of } \theta \rightarrow \text{const}} + \theta^T x_i y_i - \frac{1}{2}(\theta^T x_i - \xi_i) - \lambda(\xi_i)([\theta^T x_i]^2 - \xi_i^2)\Big]$$

(ξ_i: Independent of $\theta \rightarrow$ const)

$$+ \mathbb{E}_\alpha\Big[\underbrace{\log \mathcal{Z}(\alpha)}_{\text{Independent of } \theta \rightarrow \text{const}} + \log \exp\left(-\frac{1}{2}(\theta - \bar{0})^T \alpha^{-1} I (\theta - \bar{0})\right)\Big] + \text{const.}$$

$$= \sum_{i=1}^{N}\left[\theta^T x_i y_i - \frac{1}{2}\theta^T x_i - \lambda(\xi_i)\theta^T x_i x_i^T \theta\right] + \mathbb{E}_\alpha\left[-\frac{1}{2}\alpha \theta^T I \theta\right] + \text{const}$$

$$= \sum_{i=1}^{N}\left[(y_i - \frac{1}{2})\theta^T x_i - \lambda(\xi_i)\theta^T x_i x_i^T \theta\right] - \frac{1}{2}\mathbb{E}[\alpha]\theta^T\theta + \text{const}$$

POINT : Quadratic function of θ => Express $q(\theta)$ in Gaussian form

* $$\boxed{q(\theta) = \mathcal{N}(\theta|\mu_N, \Sigma_N)}$$

$$\mu_N = \Sigma_N\left(\sum_{i=1}^{N}(y_i - \frac{1}{2})x_i\right)$$

$$\Sigma_N^{-1} = \mathbb{E}[\alpha] I + 2\sum_{i=1}^{N}\lambda(\xi_i) x_i x_i^T$$

Expectation over Gamma distr. param.

$$\mathbb{E}[\alpha] = \frac{a_N}{b_N}$$

② $q(\alpha)$:

$$\log q(\alpha) = \mathbb{E}_\theta[\log p(y, \theta, \alpha)] + \text{const}$$

$$= E_\theta \left[\underbrace{\log h(\theta, \xi)}_{\text{Independent of } \alpha \rightarrow \text{const}} + \log p(\theta|\alpha) + \log p(\alpha) \right] + \text{const}$$

$$= E_\theta \left[\log p(\theta|\alpha) \right] + \log p(\alpha) + \text{const}$$

$$= E_\theta \left[\log \frac{1}{(2\pi)^D} \frac{1}{|\alpha^{-1} I|^{1/2}} \exp\left(-\tfrac{1}{2} (\theta - \bar{0})^T \alpha I (\theta - \bar{0}) \right) \right]$$

$$+ \log \frac{1}{\Gamma(a_0)} b_0^{a_0} \alpha^{a_0 - 1} e^{-b_0 \alpha} + \text{const}$$

$$|\alpha^{-1} I| = \alpha^{-1} \cdot \alpha^{-1} \cdots = \alpha^{-D}$$

$$= E_\theta \left[\underbrace{\log (2\pi)^{-D}}_{\rightarrow \text{const}} + \underbrace{\log |\alpha^{-1} I|^{-1/2}}_{\text{Dim. of covariance matrix } \alpha^{-1} I} - \tfrac{1}{2} \alpha \, \theta^T \theta \right] + \underbrace{\log \Gamma(a_0)^{-1} b_0^{a_0}}_{\rightarrow \text{const}} + \log \alpha^{a_0 - 1} + \log e^{-b_0 \alpha} + \text{const}$$

$$= E_\theta \left[\frac{D}{2} \log \alpha - \tfrac{1}{2} \alpha \, \theta^T \theta \right] + (a_0 - 1) \log \alpha - b_0 \alpha + \text{const}$$

$$= \frac{D}{2} \log \alpha - \tfrac{1}{2} \alpha \, E_\theta [\theta^T \theta] + (a_0 - 1) \log \alpha - b_0 \alpha + \text{const}$$

POINT: With ingenuity, see that this expression has the form of a Gamma distribution

*

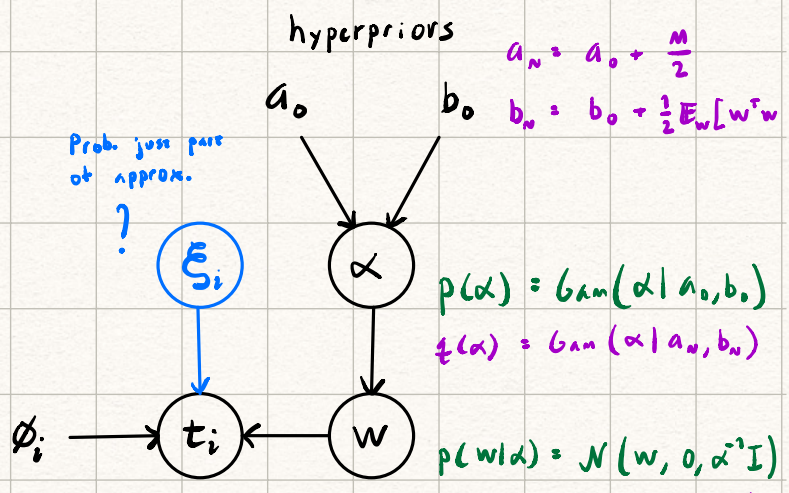
$$q(\alpha) = \text{Gam}(\alpha | a_N, b_N) = \frac{1}{\Gamma(a_N)} a_N^{b_N} \alpha^{a_0 - 1} e^{-b_0 \alpha}$$

$$a_N = a_0 + \frac{D}{2}$$

$$b_N = b_0 + \tfrac{1}{2} E_\theta [\theta^T \theta]$$

Optimizing ξ has same form as previously

$\Rightarrow$ $q(\alpha)$, $q(w)$



$a_N = a_0 + \frac{M}{2}$

$b_N = b_0 + \frac{1}{2}E_w[w^T w]$

$p(\alpha) = \text{Gam}(\alpha | a_0, b_0)$

$q(\alpha) = \text{Gam}(\alpha | a_N, b_N)$

$p(w|\alpha) = \mathcal{N}(w, 0, \alpha^{-1}I)$

$q(w) = \mathcal{N}(w | \mu_N, \Sigma_N)$

$(\xi_i^{new})^2 = \phi_i^T (\Sigma_N + \mu_N \mu_N^T) \phi_i$

$\mathbb{E}[\alpha] = \frac{a_N}{b_N}$ (Gamma)

$\mathbb{E}[w^T w] = \underbrace{\Sigma_N}_{\text{matrix}} + \underbrace{\mu_N^T \mu_N}_{\text{scalar}}$

$(\mathbb{E}[ww^T] = \Sigma_N + \mu_N \mu_N^T)$ matrix

$\Sigma_N^{-1} \mu_N = \sum_{i=1}^{N} (t_i - \frac{1}{2}) \phi_i$

$\Sigma_N^{-1} = \underbrace{\mathbb{E}[\alpha] I}_{S_0 = \Sigma_0} + 2 \sum_{i=1}^{N} \lambda(\xi_i) \phi_i \phi_i^T$

Initialize parameters

- a_0, b_0, ξ_i

$$\lambda(\xi) = \frac{1}{2\xi}\left(\sigma(\xi) - \frac{1}{2}\right)$$

$\to 0$ as $\xi \to 0$

* add numerical ϵ

Expectation

$$\mathbb{E}[\alpha] = \frac{a_N}{b_N}$$

$$\mathbb{E}_w[w^T w] = (\Sigma_N +) \mu_N^T \mu_N$$

→ Removed ?

Expectation step

1. $\mathbb{E}[\alpha]$
2. $\mathbb{E}[w^T w]$

Maximization (Re-estimation equations)

$$q(w) = \mathcal{N}(w | \mu_N, \Sigma_N)$$

- $\mu_N = \Sigma_N \sum_{i=1}^{N} (t_i - \frac{1}{2}) \phi_i$

Maximization step

(1. a_N)

2. b_N

3. ξ

- $\Sigma_N^{-1} = \mathbb{E}[\alpha] I + 2 \sum_{i=1}^{N} \lambda(\xi_i) \phi_i \phi_i^T$

4. Σ_N^{-1}
5. μ_N

$q(\alpha) = \text{Gam}(\alpha \mid a_N, b_N)$

- $a_N = a_0 + \frac{M}{2}$ ← Feature dim.

- $b_N = b_0 + \frac{1}{2} \mathbb{E}_w[w^T w]$

Remember!

$(\xi_i^{new})^2 = \phi_i^T (\Sigma_N + \mu_N \mu_N^T) \phi_i$

Predictive distr. $p(t \mid \phi, D)$

- Assumption : "Same form"

=> Same approximate predictive distribution ?

$p(C_1 \mid \phi, D) = \sigma(K(\sigma_a^2) \cdot \mu_a)$

$K(\sigma_a^2) = (1 + \frac{\pi}{8} \sigma_a^2)^{-1/2}$

$\mu_a = \mu_N^T \phi$

$\Sigma_N = S_N$ in $\mathcal{N}(w_{MAP}, S_N)$

$\sigma_a^2 = \phi^T \Sigma_N \phi$