$$p(\text{Data}, \theta) = p(\text{Data}|\theta)\, p(\theta)$$

$$p(\theta|\text{Data}) = \frac{p(\text{Data}|\theta)\, p(\theta)}{\int p(\text{Data}|\theta)\, p(\theta)\, d\theta}$$

$p(\text{Data})$

Evidence

$$p_j(\text{Data}, \theta) \qquad j \in \{1, \dots, K\}$$

$$p(\text{Data}|\theta)\, p_j(\theta)$$

$$j^* = \arg\max_j \; p_j(\text{Data}) = \arg\max_j \int p_j(\text{Data}|\theta)\, p_j(\theta)\, d\theta \qquad j \in \mathbb{R}$$

$$\log p(\text{Data}) = \int q(\theta) \log p(\text{Data})\, d\theta = \left\{ p(\text{Data}) = \frac{p(\text{Data}, \theta)}{p(\theta|\text{Data})} \right\} =$$

$$= \int q(\theta) \log \frac{p(\text{Data}, \theta)\, q(\theta)}{p(\theta|\text{Data})\, q(\theta)}\, d\theta =$$

$$= \int q(\theta) \log \frac{p(\text{Data}, \theta)}{q(\theta)}\, d\theta + \int q(\theta) \log \frac{q(\theta)}{p(\theta|\text{Data})}\, d\theta \geq \mathcal{L}(q, \theta)$$

≥ 0

$\square + \square$

$\mathcal{L}(q, \theta)$

$$\underline{\log p_j(\text{Data})} \geq \underbrace{\int \underline{q(\theta)} \log \frac{p_j(\text{Data}, \theta)}{q(\theta)} \, d\theta} \longrightarrow \max_{q, j} \qquad \forall q$$

1) Ideal case: $q^*(\theta) = p(\theta \mid \text{Data}) \iff p(\text{Data})$ - tractable

2) <u>Mean-field</u>: $q(\theta) = \prod_{j=1}^{m} \underline{q_j}(\theta_j)$

3) Parametric VI: $q(\theta) = q(\theta \mid \eta)$

$$\mathcal{L}(q, j) = \mathcal{L}(\eta, j) = \underbrace{\int q(\theta \mid \eta) \log \frac{p_j(\text{Data}, \theta)}{q(\theta \mid \eta)} \, d\theta} \longrightarrow \max_{\eta, j}$$

↳ we don't need to compute it

Function	Stochastic gradient	Randomness
$\sum_{i=1}^{n} f_i(x)$	$\nabla f_j(x)$	$j \sim U\{1,\dots,n\}$
$\int p(y) f(x,y)\,dy$	$\frac{\partial}{\partial x} f(x,\hat{y})$	$\hat{y} \sim p(y)$
$\int p(y) \left[\sum_{i=1}^{n} f_i(x,y)\right] dy$	$\frac{\partial}{\partial x} f_j(x,\hat{y})$	$j \sim U\{1,\dots,n\}$ $\hat{y} \sim p(y)$
$\int p(y\mid x) f(x,y)\,dy$	$\cancel{\frac{\partial}{\partial x} f(x,\hat{y})}$	$\hat{y} \sim p(y\mid x)$

$$\sum_{i=1}^{n} \int p(y\mid x)\, f_i(y)\,dy$$

ξ

$$p(t, \theta \mid x) = p(t \mid \theta, x)\, p(\theta) = \frac{1}{1 + \exp(-t\theta^T x)} \mathcal{N}(\theta \mid 0, \underline{\Lambda})$$

$t \in \{-1, +1\}$ $\quad x, \theta \in \mathbb{R}^d$

$$\boxed{\Lambda = \begin{pmatrix} \lambda_1^2 & & 0 \\ & \ddots & \\ 0 & & \lambda_d^2 \end{pmatrix}}$$

$\approx$

$$(X_{tr}, T_{tr}) = \{(x_i, t_i)\}_{i=1}^n$$

$$\theta_{MP} = \arg\max_\theta p(\theta \mid X_{tr}, T_{tr}) = \arg\max_\theta \underline{p(T_{tr} \mid X_{tr}, \theta)}\, p(\theta)$$

1) $\lim\limits_{\Lambda \to 0} \theta_{MP} = 0$

2) $\lim\limits_{\Lambda \to +\infty} \theta_{MP} = \theta_{ML}$

$\cancel{n \gg d}$

Фиксируем все $\lambda_j = \underline{\lambda_0}$, вспомним Байеса для бедных $\Longrightarrow$

$\Longrightarrow$ L2-регуляр. логистич. регрессия

Классический RVM имеет сложность $\underline{O(d^3)}$

$$p(t,\theta|x) = \underline{p(t|x,\theta)}\, p(\theta|\Lambda)$$

$$p(t|x,\theta) = \frac{1}{1+\exp(-t\theta^T x)} \qquad p(\theta|\Lambda) = \mathcal{N}(\theta|0,\Lambda)$$

В идеале $\Lambda^* = \arg\max_\Lambda p(T_{tr}|X_{tr},\Lambda) = \arg\max_\Lambda \int \underline{p(T_{tr}|X_{tr},\theta)}\, \underline{p(\theta|\Lambda)}\, d\theta$

1) $n, d \sim 10 - 10^3$ — Classical RVM

2) $n \gg d$ — Logistic regression with no regularization

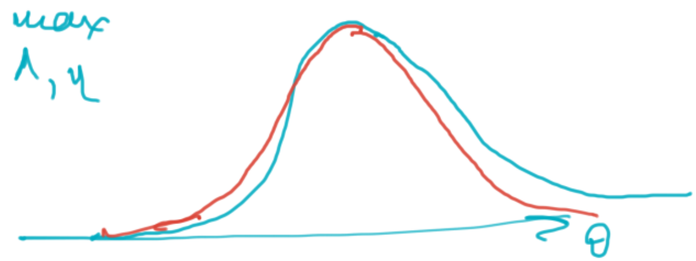
3) Пусть $n, d \gg 1$

$$\log p(T_{tr}|X_{tr},\Lambda) \geq \int q(\theta|\eta) \log \frac{p(T_{tr},\theta|X_{tr},\Lambda)}{q(\theta|\eta)} d\theta =$$

$$= \int q(\theta|\eta) \log \frac{p(T_{tr}|X_{tr},\theta)\, p(\theta|\Lambda)}{q(\theta|\eta)} d\theta \to \max_{\Lambda,\eta}$$

$\underline{p(\theta|X_{tr},T_{tr},\Lambda)}$ — нон-гауссовое

$$q(\theta|\eta) \approx p(\theta|X_{tr},T_{tr})$$



$$q(\theta|\eta) = \prod_{j=1}^{d} \mathcal{N}(\theta_j | \mu_j, \sigma_j^2) \qquad p(\theta|\Lambda) = \prod_{j=1}^{d} \mathcal{N}(\theta | 0, \lambda_j^2)$$

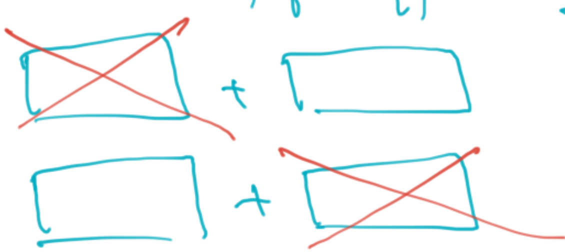
$$\mathcal{L}(\eta, \Lambda) = \int q(\theta|\eta) \log \frac{p(T_{tr}|X_{tr}, \theta)\, p(\theta|\Lambda)}{q(\theta|\eta)} d\theta =$$

$$= \int q(\theta|\eta) \log p(T_{tr}|X_{tr}, \theta)\, d\theta + \int q(\theta|\eta) \log \frac{p(\theta|\Lambda)}{q(\theta|\eta)} d\theta =$$

$$= \underbrace{\int q(\theta|\eta) \log p(T_{tr}|X_{tr}, \theta)\, d\theta}_{\text{Data term}} - \underbrace{KL(q(\theta|\eta) \| p(\theta|\Lambda))}_{\text{Regularizator}} \to \max_{\eta, \Lambda}$$

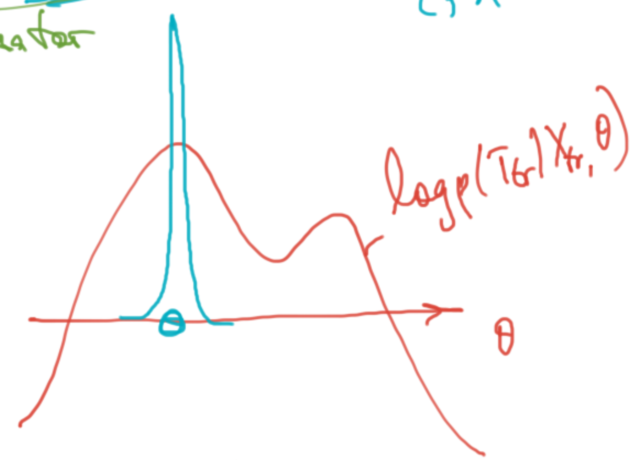
Пусть 1) $p(\theta|\Lambda)$ – фиксировано

2) $q(\theta|\eta)$ – содержит вырожд. распр-ия



$$q(\theta|\eta) = p(\theta|\Lambda)$$

$$q(\theta|\eta) = \delta(\theta - \theta_{ML})$$



$$p(\theta|\Lambda) = \prod_{j=1}^{d} \mathcal{N}(\theta_j|0, \lambda_j^2) \qquad q(\theta|\eta) = \prod_{j=1}^{d} \mathcal{N}(\theta_j|\mu_j, \sigma_j^2)$$

$\max_{\eta, \Lambda}$

$\eta = \{\vec{\mu}, \vec{\sigma}\}$

tractable

$$\mathcal{L}(\eta, \Lambda) = \int q(\theta|\eta) \log p(T_{tr}|X_{tr}, \theta)\, d\theta - \int q(\theta|\eta) \log \frac{\prod_j q(\theta|\eta)}{\prod_j p(\theta|\Lambda)}\, d\theta =$$

$$= \sum_{i=1}^{n} \int q(\theta|\eta) \log p(t_i|x_i, \theta)\, d\theta - \sum_{j=1}^{d} KL\left(\mathcal{N}(\theta_j|\mu_j, \sigma_j^2) \,\|\, \mathcal{N}(\theta_j|0, \lambda_j^2)\right)$$

$$KL(q(\theta_j|\eta_j)\|p(\theta_j|\lambda_j)) = \log \frac{\lambda_j}{\sigma_j} + \frac{\sigma_j^2 + \mu_j^2}{2\lambda_j^2} \qquad \Big| \frac{\partial}{\partial \lambda_j}, = 0$$

$$\frac{1}{\lambda_j} - \frac{\sigma_j^2 + \mu_j^2}{\lambda_j^3} = 0 \iff \boxed{\lambda_{j*}^2 = \sigma_j^2 + \mu_j^2}$$

$$\sum_{j=1}^{d} \log \frac{\sigma_j^2}{\sigma_j^2 + \mu_j^2} + \text{Const}$$

$$KL(q(\theta_j|\eta_j)\|p(\theta_j|\lambda_{j*})) = \log \frac{\sigma_j^2 + \mu_j^2}{\sigma_j^2} + \text{Const}$$

$$\frac{\partial}{\partial \mu}, \frac{\partial}{\partial \sigma}$$

Рассм. 1ое слагаемое в $\mathcal{L}(\eta, \Lambda)$

$$\sum_{i=1}^{n} \int q(\theta|\eta) \log p(t_i|x_i, \theta)\, d\theta \;\ominus$$

Reparametrization trick

$$q(\theta|\eta) \longrightarrow r(\varepsilon)$$

$$\theta = \underline{g}(\varepsilon, \eta)$$

Если $q(\theta|\eta) = \prod_{j=1}^{d} \mathcal{N}(\theta_j|\mu_j, \sigma_j^2)$,

тогда $\theta_j = g(\varepsilon_j, \eta_j) = \underline{\mu_j + \varepsilon\sigma_j}$, где $\varepsilon \sim \mathcal{N}(\varepsilon|0, 1)$

$$\ominus \sum_{i=1}^{n} \int r(\varepsilon) \log p(t_i|x_i, g(\varepsilon, \eta))\, d\varepsilon \quad \Big|\ \text{stoch. grad}$$

не зависит от η

$j \sim U\{1, \dots, n\} \quad \hat{\varepsilon} \sim \mathcal{N}(\varepsilon|0, I)$

$$\frac{\partial}{\partial \eta} \int q(\theta|\eta) \log p(T_{tr}|X_{tr}, \theta)\, d\theta \approx \underline{\frac{\partial}{\partial \eta} \log p(t_j|x_j, g(\hat{\varepsilon}, \eta))}$$

Нужно $\exists \frac{\partial g}{\partial \eta}$

$$\mathcal{L}(\eta) = \sum_{i=1}^{n} \int q(\theta|\eta) \log p(t_i|x_i,\theta)\, d\theta + \frac{1}{2}\sum_{j=1}^{d} \log \frac{\sigma_j^2}{\sigma_j^2+\mu_j^2} \to \max_{\eta}$$

$$\lambda_{j*} = \sigma_j^2 + \mu_j^2$$

Doubly Stochastic Variational Inference

1) Mini-batching
2) Monte-Carlo estimation of intractable integrals

Значит, если $\lambda_{j*} \to 0$, т.е. j-ый признак можно исключить