$$p(t, \theta \mid x, \lambda) = p(t \mid x, \theta)\, p(\theta \mid \lambda)$$

4 4

✓ 1) Model Selection, λ - ?

✓ 2) Bayesian inference

i) $\lambda^* = \arg\max_{\lambda} p(T_{tr} \mid X_{tr}, \lambda) = \arg\max_{\lambda} \int p(T_{tr} \mid X_{tr}, \theta)\, p(\theta \mid \lambda)\, d\theta$

ii) $p(\theta \mid X_{tr}, T_{tr}, \lambda) = \dfrac{p(T_{tr} \mid X_{tr}, \theta)\, p(\theta \mid \lambda)}{\int p(T_{tr} \mid X_{tr}, \theta)\, p(\theta \mid \lambda)\, d\theta}$

$p(T_{tr} \mid X_{tr}, \theta)\, p(\theta \mid \lambda)$

Variational bayes

$$\log p(T_{tr} \mid X_{tr}, \lambda) \geq \mathcal{L}(q, \lambda) = \int q(\theta \mid \eta) \log \frac{p(T_{tr}, \theta \mid X_{tr}, \lambda)}{q(\theta \mid \eta)}\, d\theta \to \max_{\eta, \lambda}$$

η

$$\int q(\theta \mid \eta) \log p(T_{tr} \mid X_{tr}, \theta)\, d\theta - KL(q(\theta \mid \eta) \| p(\theta \mid \lambda)) =$$

MC estimation

$$= \sum_{i=1}^{n} \int q(\theta \mid \eta) \log p(t_i \mid x_i, \theta)\, d\theta - KL(q(\theta \mid \eta) \| p(\theta \mid \lambda))$$

$\lambda^* = \arg\max \mathcal{L}$

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

$p(t, \theta | x, \lambda) = p(t|x,\theta)\, p(\theta|\lambda)$

$p(t|x,\theta)$ ~ DNN

✓ $p(\theta|\lambda) = \prod_{i,j,l} \mathcal{N}(\theta_{ijl} | 0, \lambda_{ijl}^2)$

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

✓ $q(\theta|\eta) = \prod_{i,j,l} \mathcal{N}(\theta_{ijl} | \mu_{ijl}, \sigma_{ijl}^2)$

$\eta = \{(\mu_{ijl}, \sigma_{ijl})\}$

Не зависит от λ

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

$$= \sum_{i=1}^{n} \int q(\theta|\eta) \log p(t_i | x_i, \theta)\, d\theta - \sum_{i,j,l} KL\left(\mathcal{N}(\theta_{ijl} | \mu_{ijl}, \sigma_{ijl}^2) \| \mathcal{N}(\theta_{ijl} | 0, \lambda_{ijl}^2)\right)$$

$$KL(\mathcal{N}(\theta|\mu,\sigma^2) \| \mathcal{N}(\theta|0,\lambda^2)) = \log \frac{\lambda}{\sigma} + \frac{\sigma^2 + \mu^2}{2\lambda^2} = \left\{(\lambda^*)^2 = \mu^2 + \sigma^2\right\} =$$

$$= \text{Const} + \frac{1}{2} \log \frac{\mu^2 + \sigma^2}{\sigma^2}$$

$$\sum_{i=1}^{n} \int q(\theta|\eta) \log p(t_i|x_i,\theta)\, d\theta = \left\{ \begin{array}{l} RT \\ \theta = g(\varepsilon,\eta) \\ \varepsilon \sim r(\varepsilon) \end{array} \right. \left. \begin{array}{l} \theta_{ijl} = \mu_{ijl} + \varepsilon_{ijl}\sigma_{ijl} \\ \varepsilon_{ijl} \sim N(\varepsilon_{ijl}|0,1) \\ g(\varepsilon,\eta) \end{array} \right\} =$$

$$= \sum_{i=1}^{n} \int r(\varepsilon) \log p(t_i|x_i, g(\varepsilon,\eta))\, d\theta$$

$$\text{stoch grad}_\eta \left[\sum_{i=1}^{n} \int r(\varepsilon) \log p(t_i|x_i,g(\varepsilon,\eta))\, d\varepsilon \right] = \left\{ j \sim U\{1,\dots,n\} \right\}$$

$$= n \cdot \text{stoch grad}_\eta \int r(\varepsilon) \log p(t_j|x_j, g(\varepsilon,\eta))\, d\varepsilon =$$

$$= n \int r(\varepsilon)\, \text{stochgrad}_\eta \log p(t_j|x_j,g(\varepsilon,\eta))\, d\varepsilon = \left\{ \hat{\varepsilon} \sim r(\varepsilon) \right\} =$$

$$= n \cdot \frac{\partial}{\partial \eta} \log p(t_j|x_j, g(\hat{\varepsilon},\eta) = n \frac{\partial}{\partial \theta} \log p(t_j|x_j,\theta) \cdot \frac{\partial g(\hat{\varepsilon},\eta)}{\partial \eta}$$

$$\mathcal{L}(\eta) = \sum_{i=1}^{n} \int q(\theta|\eta) \log p(T_{tr}|X_{tr},\theta)\, d\theta - \frac{1}{2} \sum_{i,j,l} \log \frac{\mu_{ijl}^2 + \sigma_{ijl}^2}{\sigma_{ijl}^2} \to \max_{\eta}$$

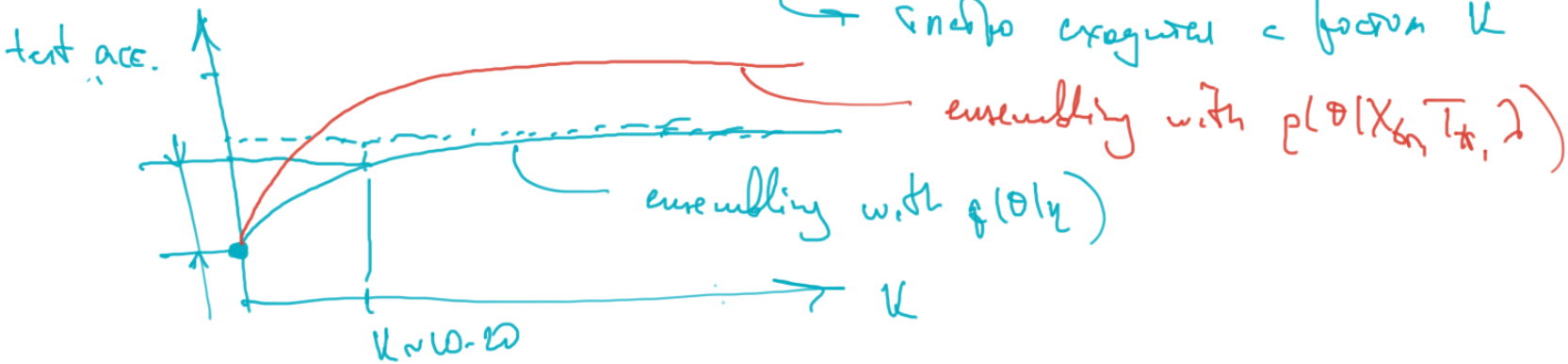
Вырожденность μ_{ijl} и $\sigma_{ijl} \to 0$, соответственно $\lambda_{ijl} \to 0$

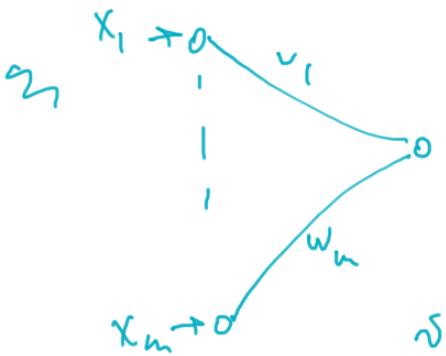
$$q(\theta|\eta) \approx p(\theta|X_{tr}, T_{tr}, \lambda)$$

3) Ensembling $p(t|x, X_{tr}, T_{tr}) = \int p(t|x,\theta)\, p(\theta|X_{tr}, T_{tr})\, d\theta \approx$

$\approx \int p(t|x,\theta)\, q(\theta|\eta)\, d\theta \approx \frac{1}{K} \sum_{k=1}^{K} p(t|x,\theta_k)$, где $\theta_k \sim q(\theta|\eta)$

↳ сильно зависит с ростом K





$$y = f\left(\sum_{i=1}^{m} w_i x_i\right)$$

$$w_i = \tilde{w}_i \cdot z_i$$

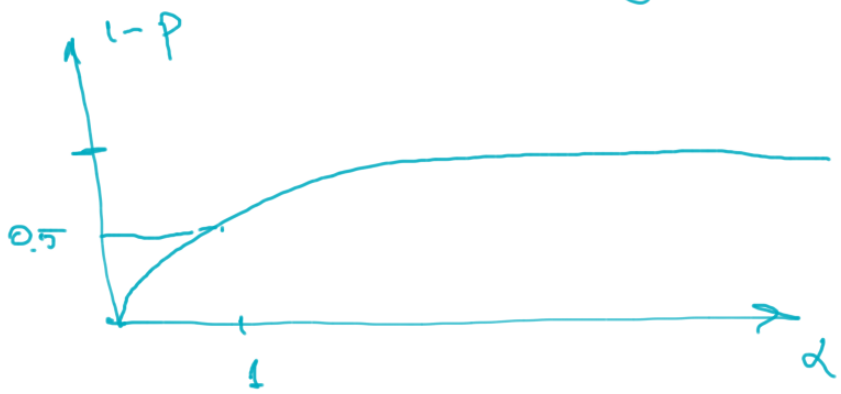
$z_i \sim \mathrm{Bar}(p)$ — число двигателей на нормальном ходу

$$v = \sum_{i=1}^{m} w_i x_i \xrightarrow{m \gg 1} \mathcal{N}(v \mid \mu, \sigma^2)$$

$$w_i = \tilde{w}_i \cdot \delta \qquad \delta \sim \mathcal{N}(\delta \mid 1, \alpha)$$

Существует взаимно-однозначное соответствие между α и p

$$\alpha^2 \quad \frac{1-p}{p}$$



$$\mathcal{L}(\eta) = \int q(\theta|\eta) \log p(T_{tr}|X_{tr},\theta)\,d\theta - \frac{1}{2}\sum_{i,j,l}\log\frac{\mu_{ijl}^2+\sigma_{ijl}^2}{\sigma_{ijl}^2} =$$

$$= \int \prod_{i,j,l} \mathcal{N}(\theta_{ijl}|\mu_{ijl},\sigma_{ijl}^2)\log p(T_{tr}|X_{tr},\theta)\,d\theta - \frac{1}{2}\sum_{i,j,l}\log\frac{\mu_{ijl}^2+\sigma_{ijl}^2}{\sigma_{ijl}^2} =$$

$$= \left\{\begin{array}{l}\sigma_{ijl}^2 = \alpha_{ijl}\mu_{ijl}^2 \\ \alpha_{ijl} = \dfrac{\sigma_{ijl}^2}{\mu_{ijl}^2}\end{array}\right\} = \int \prod_{i,j,l}\mathcal{N}(\theta_{ijl}|\mu_{ijl},\alpha_{ijl}\mu_{ijl}^2)\log p(T_{tr}|X_{tr},\theta)\,d\theta - \frac{1}{2}\sum_{i,j,l}\log\frac{1+\alpha_{ijl}}{\alpha_{ijl}}$$

↳ мы забыли от μ_{ijl} !

Пусть все $\alpha_{ijl} = \alpha$ - фиксированы

$$\max_{\mu}\left[\int \prod \mathcal{N}(\theta_{ijl}|\mu_{ijl},\alpha\mu_{ijl}^2)\log p(T_{tr}|X_{tr},\theta)\,d\theta - \frac{1}{2}\sum_{i,j,l}\log\frac{1+\alpha_{ijl}}{\alpha_{ijl}}\right] =$$

$$= \max_{\mu}\left[\int \prod \mathcal{N}(\theta_{ijl}|\mu_{ijl},\alpha\mu_{ijl}^2)\log p(T_{tr}|X_{tr},\theta)\,d\theta\right] = \left\{\begin{array}{l}\theta_{ijl} = \mu_{ijl}\delta_{ijl} \\ \delta_{ijl} \sim \mathcal{N}(\delta_{ijl}|1,\alpha)\end{array}\right\}$$

$$= \max_{\mu}\left[\int \prod \mathcal{N}(\delta_{ijl}|1,\alpha)\log p(T_{tr}|X_{tr},\mu\cdot\delta)\,d\delta\right]$$

$D(\mu) = \int \prod_{i,j,l} \mathcal{N}(\delta_{ijl} \mid 1, \alpha) \log p(T_{tr} \mid X_{tr}, \mu \cdot \delta)\, d\delta \to \max_{\mu}$

↳ не зависит от μ

| stoch. grad$_\mu$

$j \sim U\{1, \dots, n_j\}$

$\hat{\delta}_{ijl} \sim \mathcal{N}(\delta_{ijl} \mid 1, \alpha)$

$$\text{stoch grad}_\mu D(\mu) = n \frac{\partial}{\partial \mu} \log p(t_j \mid x_j, \mu \cdot \hat{\delta}) =$$

$$= n \frac{\partial}{\partial \theta} \log p(t_j \mid x_j, \theta) \cdot \frac{\partial \theta}{\partial \mu} \leftarrow \text{совпадает со стох. градиентом в гауссовской границе}$$

Гауссовский граница соотв. вариан Байесовскому выводу в том, где

$q(\theta \mid \eta) = \prod_{i,j,l} \mathcal{N}(\theta_{ijl} \mid \mu_{ijl}, \alpha \mu_{ijl}^2)$ где $\eta = \vec{\mu}$, а α - фиксирован

можем расширить модель введя разные α_{ijl} для каждого θ_{ijl}

$\mathcal{L}(\mu, \alpha) \to \max_{\mu}$ $\Longleftrightarrow$ Гауссовский фронаут с фиксир. α

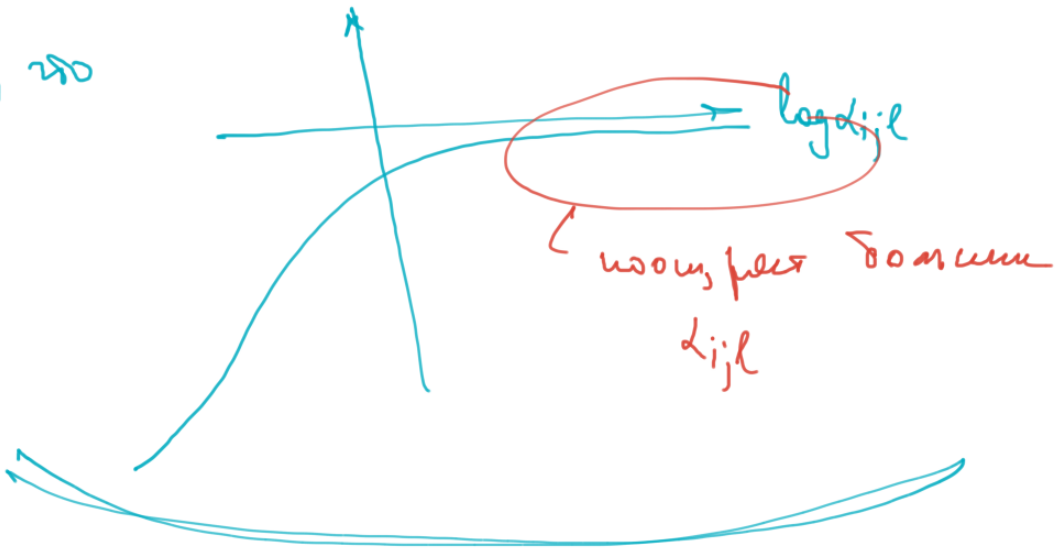
$\mathcal{L}(\mu, \alpha) \to \max_{\mu, \alpha}$ $\Longleftrightarrow$ Sparse Variational Dropout

Замеч. часть $\alpha_{ij} \to +\infty$

Вторая слагаемое в $\mathcal{L}(\mu, \alpha)$ имеет вид $-\frac{1}{2} \sum_{ij} \log \frac{1+\alpha_{ij}}{\alpha_{ij}}$

$\Rightarrow$ при этом $\mu_{ij} \to 0$ так, что

$\sigma_{ij}^2 = \alpha_{ij} \mu_{ij}^2 \to 0$



Новая постановка регуляризации

$$\log p(T_{tr} | X_{tr}, \theta) + \lambda R(\theta) \longrightarrow \max_{\theta}$$

$$\theta_{MP} = \arg\max_{\theta} \frac{p(T_{tr} | X_{tr}, \theta)\, p(\theta)}{\int p(T_{tr} | X_{tr}, \theta)\, p(\theta)\, d\theta} = \arg\max_{\theta} p(T_{tr} | X_{tr}, \theta)\, p(\theta)$$

где $p(\theta) = \frac{1}{A} \exp(\lambda R(\theta))$

Классика

Современная трактовка

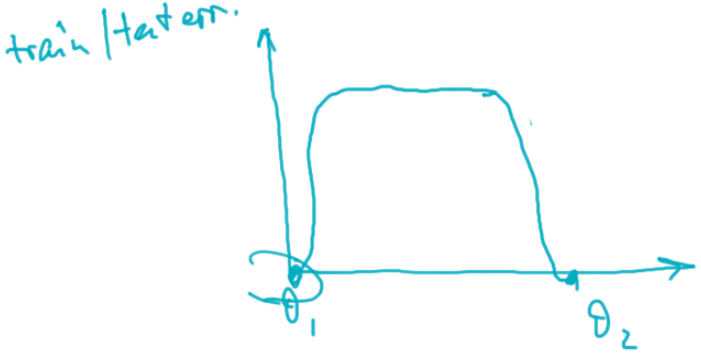
$$\int r(\varepsilon) \log p(T_{tr} | X_{tr}, g(\varepsilon, \theta))\, d\varepsilon \longrightarrow \max$$, где

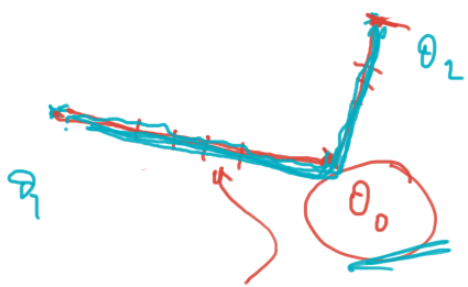
$g(\varepsilon, \theta)$ – зашумленная версия весов θ из шума ε

Вид $r(\varepsilon)$ – задается пользователем в байесовских моделях

Введение Байесовской модели позволяет подбирая не только шум

θ_1, θ_2 – два подбираемых параметра $-\log p(T_{tr} | X_{tr}, \theta)$





хотим искать ошибку вдоль всего пути

$$\theta(t, \theta_0) = \begin{cases} (1-t)\theta_1 + t\theta_0, & t \in [0,1] \\ (2-t)\theta_0 + (t-1)\theta_2, & t \in [1,2] \end{cases}$$

$$\Phi(\theta_0) = \int_0^2 \underline{\tau(t)} \log p(T_{tr} | X_{tr}, \theta(t, \theta_0))\, dt \to \max_{\theta_0} \qquad \tau(t) = U[0,2]$$

$$\text{stoch grad } \underline{\Phi(\theta_0)} = n \frac{\partial}{\partial \theta} \log p(t_j | x_j, \theta) \cdot \frac{\partial \theta(\hat{t}, \theta_0)}{\partial \theta_0} \qquad \hat{t} \sim U[0,2], \quad j \sim U\{1, \dots, n\}$$