Bayesian Linear Models for Classification

E. Burnaev, Skoltech

Logistic Regression

$$p(\mathcal{C}_1|\phi) = y(\phi) = \sigma\left(\mathbf{w}^{\mathrm{T}}\phi\right) \qquad p(\mathcal{C}_2|\phi) = 1 - p(\mathcal{C}_1|\phi)$$

Vector of basis functions $\phi(\mathbf{x})$

Gaussian class conditional densities together with the class prior has a total of M (M + 5)/2 + 1 parameters

For an M-dimensional feature space ϕ, this model has M adjustable parameters

A data set $\{\phi_n, t_n\}$**, where** $t_n \in \{0, 1\}$ **and** $\phi_n = \phi(\mathbf{x}_n)$

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n} \left\{1 - y_n\right\}^{1-t_n}$$

$$\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}} \quad y_n = p(\mathcal{C}_1|\phi_n)$$

Cross-entropy error function

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}$$

where $y_n = \sigma(a_n)$ **and** $a_n = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}_n$

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n) \boldsymbol{\phi}_n$$

Iterative Reweighted Least Squares

Newton-Raphson method

$$\mathbf{w}^{(\text{new})} = \mathbf{w}^{(\text{old})} - \mathbf{H}^{-1}\nabla E(\mathbf{w})$$

In case of linear regression with sum-of-squares error

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}(\mathbf{w}^{\mathrm{T}}\phi_n - t_n)\phi_n = \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{w} - \mathbf{\Phi}^{\mathrm{T}}\mathbf{t}$$

$$\mathbf{H} = \nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N}\phi_n\phi_n^{\mathrm{T}} = \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$$

$\mathbf{\Phi}$ **is the** $N \times M$ **design matrix**

$$\begin{aligned}
\mathbf{w}^{(\text{new})} &= \mathbf{w}^{(\text{old})} - (\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}\left\{\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{w}^{(\text{old})} - \mathbf{\Phi}^{\mathrm{T}}\mathbf{t}\right\} \\
&= (\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}
\end{aligned}$$

Newton-Raphson method for logistic regression

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}(y_n - t_n)\phi_n = \mathbf{\Phi}^{\mathrm{T}}(\mathbf{y} - \mathbf{t})$$

$$\mathbf{H} = \nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N} y_n(1 - y_n)\phi_n\phi_n^{\mathrm{T}} = \mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{\Phi}$$

$\mathbf{R}$ **is a diagonal matrix N*N with elements** $R_{nn} = y_n(1 - y_n).$

Since $0 < y_n < 1$ **then** $\mathbf{u}^{\mathrm{T}}\mathbf{H}\mathbf{u} > 0$**, i.e.** $\mathbf{H}$ **Is positive definite. Thus the error function is a concave function**

$$\begin{aligned}
\mathbf{w}^{(\mathrm{new})} &= \mathbf{w}^{(\mathrm{old})} - (\mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathrm{T}}(\mathbf{y} - \mathbf{t}) \\
&= (\mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{\Phi})^{-1}\left\{\mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{\Phi}\mathbf{w}^{(\mathrm{old})} - \mathbf{\Phi}^{\mathrm{T}}(\mathbf{y} - \mathbf{t})\right\} \\
&= (\mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{z}
\end{aligned}$$

$$\begin{aligned}
\mathbf{w}^{(\text{new})} &= \mathbf{w}^{(\text{old})} - (\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{y}-\mathbf{t}) \\
&= (\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi})^{-1}\left\{\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi}\mathbf{w}^{(\text{old})} - \boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{y}-\mathbf{t})\right\} \\
&= (\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{z}
\end{aligned}$$

$$\mathbf{z} = \boldsymbol{\Phi}\mathbf{w}^{(\text{old})} - \mathbf{R}^{-1}(\mathbf{y}-\mathbf{t}).$$

$\mathbf{R}$ can be interpreted as a variance, since

$$\begin{aligned}
\mathbb{E}[t] &= \sigma(\mathbf{x}) = y \\
\operatorname{var}[t] &= \mathbb{E}[t^2] - \mathbb{E}[t]^2 = \sigma(\mathbf{x}) - \sigma(\mathbf{x})^2 = y(1-y)
\end{aligned}$$

$$
\begin{aligned}
\mathbf{w}^{(\mathrm{new})} &= \mathbf{w}^{(\mathrm{old})} - (\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{y}-\mathbf{t}) \\
&= (\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi})^{-1}\left\{\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi}\mathbf{w}^{(\mathrm{old})} - \boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{y}-\mathbf{t})\right\} \\
&= (\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{z}
\end{aligned}
$$

$$
\mathbf{z} = \boldsymbol{\Phi}\mathbf{w}^{(\mathrm{old})} - \mathbf{R}^{-1}(\mathbf{y}-\mathbf{t}).
$$

Iterative Reweighted Least Squares (IRLS) ~ solution of the linearized problem in the space of the variable $a = \mathbf{w}^{\mathrm{T}}\phi$

$$
\begin{aligned}
a_n(\mathbf{w}) &\simeq a_n(\mathbf{w}^{(\mathrm{old})}) + \left.\frac{\mathrm{d}a_n}{\mathrm{d}y_n}\right|_{\mathbf{w}^{(\mathrm{old})}} (t_n - y_n) \\
&= \phi_n^{\mathrm{T}}\mathbf{w}^{(\mathrm{old})} - \frac{(y_n - t_n)}{y_n(1-y_n)} = z_n.
\end{aligned}
$$

Multiclass logistic regression

$$p(\mathcal{C}_k|\boldsymbol{\phi}) = y_k(\boldsymbol{\phi}) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

$$a_k = \mathbf{w}_k^{\mathrm{T}}\boldsymbol{\phi}.$$

$$\frac{\partial y_k}{\partial a_j} = y_k(I_{kj} - y_j)$$

Likelihood??? => 1-of-K coding scheme!!!

$$p(\mathbf{T}|\mathbf{w}_1,\ldots,\mathbf{w}_K) = \prod_{n=1}^{N}\prod_{k=1}^{K} p(\mathcal{C}_k|\boldsymbol{\phi}_n)^{t_{nk}} = \prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}^{t_{nk}}$$

$$y_{nk} = y_k(\boldsymbol{\phi}_n)$$

$\mathbf{T}$ is an N*K matrix of target variables with elements t_{nk}.

$$E(\mathbf{w}_1,\ldots,\mathbf{w}_K) = -\ln p(\mathbf{T}|\mathbf{w}_1,\ldots,\mathbf{w}_K) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\ln y_{nk}$$

$$\nabla_{\mathbf{w}_j} E(\mathbf{w}_1,\ldots,\mathbf{w}_K) = \sum_{n=1}^{N} (y_{nj} - t_{nj})\,\boldsymbol{\phi}_n$$

Hessian couples blocks of size M*M in which block (j,k) is given by

$$\nabla_{\mathbf{w}_k}\nabla_{\mathbf{w}_j} E(\mathbf{w}_1,\ldots,\mathbf{w}_K) = -\sum_{n=1}^{N} y_{nk}(I_{kj} - y_{nj})\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}.$$

Probit regression

$$p(t = 1|a) = f(a)$$

$a = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}$, $\quad f(\cdot)$ **is the activation function**

$$\begin{cases} t_n = 1 & \text{if } a_n \geqslant \theta \\ t_n = 0 & \text{otherwise.} \end{cases}$$

$$\theta \sim p(\theta) \qquad f(a) = \int_{-\infty}^{a} p(\theta)\,\mathrm{d}\theta$$

$$\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta|0,1)\,\mathrm{d}\theta \qquad \mathrm{erf}(a) = \frac{2}{\sqrt{\pi}} \int_{0}^{a} \exp(-\theta^2/2)\,\mathrm{d}\theta$$

$$\boldsymbol{\Phi}(a) = \frac{1}{2}\left\{1 + \frac{1}{\sqrt{2}}\mathrm{erf}(a)\right\}.$$

The Laplace Approximation

$$p(z) = \frac{1}{Z} f(z) \qquad Z = \int f(z)\,\mathrm{d}z$$

$$p'(z_0) = 0 \;\Leftrightarrow\; \left.\frac{df(z)}{dz}\right|_{z=z_0} = 0.$$

$$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2} A (z - z_0)^2$$

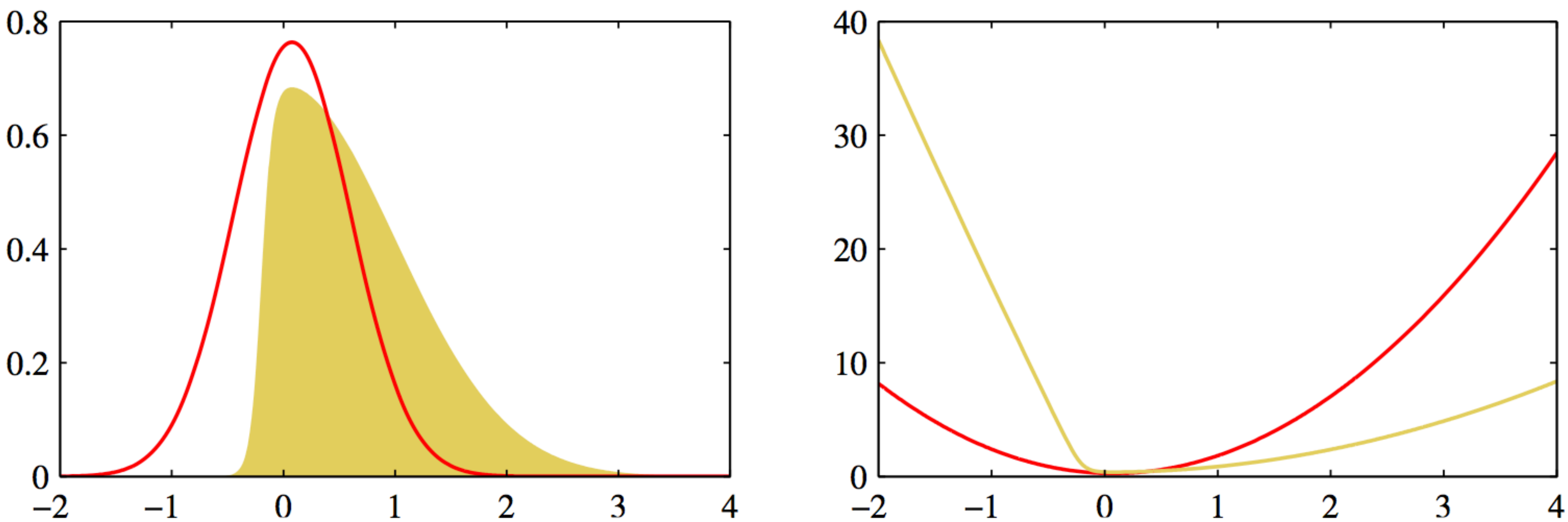
$$A = - \left.\frac{d^2}{dz^2} \ln f(z)\right|_{z=z_0}.$$

$$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2} A (z - z_0)^2$$

$$A = - \left. \frac{d^2}{dz^2} \ln f(z) \right|_{z=z_0} .$$

$$q(z) = \left(\frac{A}{2\pi} \right)^{1/2} \exp \left\{ -\frac{A}{2} (z - z_0)^2 \right\} .$$

$$q(z) = \left(\frac{A}{2\pi}\right)^{1/2} \exp\left\{-\frac{A}{2}(z - z_0)^2\right\}.$$



$p(z) \propto \exp(-z^2/2)\sigma(20z + 4)$

$$p(\mathbf{z}) = f(\mathbf{z})/Z$$

$$\ln f(\mathbf{z}) \simeq \ln f(\mathbf{z}_0) - \frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{z}-\mathbf{z}_0)$$

$$\mathbf{A} = -\nabla\nabla \ln f(\mathbf{z})\big|_{\mathbf{z}=\mathbf{z}_0}$$

$$f(\mathbf{z}) \simeq f(\mathbf{z}_0)\exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{z}-\mathbf{z}_0)\right\}$$

$$q(\mathbf{z}) = \frac{|\mathbf{A}|^{1/2}}{(2\pi)^{M/2}}\exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{z}-\mathbf{z}_0)\right\} = \mathcal{N}(\mathbf{z}|\mathbf{z}_0, \mathbf{A}^{-1})$$

Model Comparison and BIC

$$\begin{aligned} Z &= \int f(\mathbf{z})\,\mathrm{d}\mathbf{z} \\ &\simeq f(\mathbf{z}_0)\int \exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{z}-\mathbf{z}_0)\right\}\mathrm{d}\mathbf{z} \\ &= f(\mathbf{z}_0)\frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}} \end{aligned}$$

data set $\mathcal{D}$

set of models $\{\mathcal{M}_i\}$ **with parameters** $\{\boldsymbol{\theta}_i\}$.

$p(\boldsymbol{\theta}_i|\mathcal{M}_i)$ **is a prior over parameters**

$p(\mathcal{D}|\mathcal{M}_i)$ **is a model evidence**

$$p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}.$$

$$f(\boldsymbol{\theta}) = p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \qquad Z = p(\mathcal{D})$$

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) + \underbrace{\ln p(\boldsymbol{\theta}_{\mathrm{MAP}}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}|}_{\text{Occam factor}}$$

$$\mathbf{A} = -\nabla\nabla \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}})p(\boldsymbol{\theta}_{\mathrm{MAP}}) = -\nabla\nabla \ln p(\boldsymbol{\theta}_{\mathrm{MAP}}|\mathcal{D}).$$

If we assume that the Gaussian prior distribution over parameters is broad, and that the Hessian has full rank, then we can approximate

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) - \frac{1}{2}M \ln N$$

Bayesian Logistic Regression

$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0)$ **is a prior over parameters**

$\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$ $\qquad p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{w})p(\mathbf{t}|\mathbf{w})$

$$
\begin{aligned}
\ln p(\mathbf{w}|\mathbf{t}) &= -\frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0) \\
&\quad + \sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} + \mathrm{const}
\end{aligned}
$$

$$y_n = \sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n)$$

$$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0) + \sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} + \text{const}$$

$$y_n = \sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n)$$

To obtain a Gaussian approximation to the posterior distribution, we first maximize the posterior distribution to give the MAP

$$\mathbf{S}_N = -\nabla\nabla \ln p(\mathbf{w}|\mathbf{t}) = \mathbf{S}_0^{-1} + \sum_{n=1}^{N} y_n(1-y_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}.$$

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{\mathrm{MAP}}, \mathbf{S}_N).$$

Predictive Distribution

$p(\mathbf{w}|\mathbf{t})$ is approximated by a Gaussian $q(\mathbf{w})$

$$p(\mathcal{C}_1|\boldsymbol{\phi},\mathbf{t}) = \int p(\mathcal{C}_1|\boldsymbol{\phi},\mathbf{w})p(\mathbf{w}|\mathbf{t})\,\mathrm{d}\mathbf{w} \simeq \int \sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi})q(\mathbf{w})\,\mathrm{d}\mathbf{w}$$

$$p(\mathcal{C}_2|\boldsymbol{\phi},\mathbf{t}) = 1 - p(\mathcal{C}_1|\boldsymbol{\phi},\mathbf{t}).$$

$$\sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}) = \int \delta(a - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi})\sigma(a)\,\mathrm{d}a$$ **with** $a = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}$

$$\int \sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi})q(\mathbf{w})\,\mathrm{d}\mathbf{w} = \int \sigma(a)p(a)\,\mathrm{d}a$$ **with**

$$p(a) = \int \delta(a - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi})q(\mathbf{w})\,\mathrm{d}\mathbf{w}.$$

$$p(a) = \int \delta(a - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi})q(\mathbf{w})\,\mathrm{d}\mathbf{w}.$$

delta function imposes a linear constraint on w and so forms a marginal distribution by integrating out all directions orthogonal to ϕ. Thus we will get a 1d Gaussian distribution with

$$\mu_a = \mathbb{E}[a] = \int p(a)a\,\mathrm{d}a = \int q(\mathbf{w})\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}\,\mathrm{d}\mathbf{w} = \mathbf{w}_{\mathrm{MAP}}^{\mathrm{T}}\boldsymbol{\phi}$$

$$\begin{aligned}\sigma_a^2 &= \mathrm{var}[a] = \int p(a)\left\{a^2 - \mathbb{E}[a]^2\right\}\mathrm{d}a \\ &= \int q(\mathbf{w})\left\{(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi})^2 - (\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi})^2\right\}\mathrm{d}\mathbf{w} = \boldsymbol{\phi}^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}.\end{aligned}$$

$$p(\mathcal{C}_1|\mathbf{t}) = \int \sigma(a)p(a)\,\mathrm{d}a = \int \sigma(a)\mathcal{N}(a|\mu_a, \sigma_a^2)\,\mathrm{d}a.$$

We approximate $\sigma(a)$ **by** $\Phi(\lambda a)$. **Here** $\lambda^2 = \pi/8$. **(same slope at the origin)**

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

$$\int \Phi(\lambda a)\mathcal{N}(a|\mu, \sigma^2)\,\mathrm{d}a = \Phi\left(\frac{\mu}{(\lambda^{-2} + \sigma^2)^{1/2}}\right).$$

$$\int \sigma(a)\mathcal{N}(a|\mu, \sigma^2)\,\mathrm{d}a \simeq \sigma\left(\kappa(\sigma^2)\mu\right)$$

$$\kappa(\sigma^2) = (1 + \pi\sigma^2/8)^{-1/2}.$$

$$p(\mathcal{C}_1|\boldsymbol{\phi}, \mathbf{t}) = \sigma\left(\kappa(\sigma_a^2)\mu_a\right)$$