Bayesian Models in Machine Learning

Lukáš Burget

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

BAYa lectures, November 2024

Some images taken from: C. M. Bishop: Pattern Recognition and Machine Learning, Springer, 2006.

Frequentist vs. Bayesian

- Frequentist point of view:
  - Probability is the frequency of an event occurring in a large (infinite) number of trials
  - E.g. When flipping a coin many times, what is the proportion of heads?
- Bayesian
  - Inferring probabilities for events that have never occurred or believes which are not directly observed
  - Prior believes are specified in terms of prior probabilities
  - Taking into account uncertainty (posterior distribution) of the estimated parameters or hidden variables in our probabilistic model.

Coin flipping example

$$P(head|\mu) = \mu \qquad P(tail|\mu) = 1 - \mu$$

$$\mathbf{x} = [x_1, x_2, x_3, \dots x_N] = [taill, head, head, \dots tail]$$

- Lets flip the coin N = 1000 times getting H = 750 heads and T = 250 tails.
- What is μ? Intuitive (and also ML) estimate is 750 / 1000 = 0.75.
- Given some μ, we can calculate probability (likelihood) of X

$$P(\mathbf{x}|\mu) = \prod_i P(x_i|\mu) = \mu^H (1-\mu)^T$$

- Now lets express our *ignorant* prior belief about μ as:

$$p(\mu) = \mathcal{U}(0,1)$$

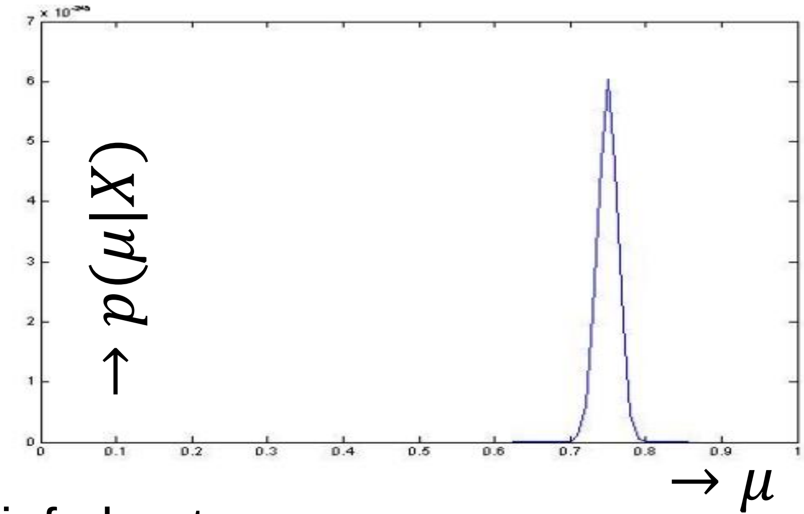
Then using Bayes rule, we obtain probability density function for μ :

$$p(\mu|\mathbf{x}) = \frac{P(\mathbf{X}|\mu)p(\mu)}{P(\mathbf{x})} = \frac{\prod_i P(x_i|\mu)\cdot 1}{P(\mathbf{x})} \propto \mu^H (1-\mu)^T$$

Coin flipping example (cont.)

N = 1000, H = 750, T = 250

$$p(\mu|\mathbf{x}) \propto \mu^H (1-\mu)^T$$



- Posterior distribution is our *new* belief about μ
- Flipping the coin once more, what is the probability of head?

$$p(head|\mathbf{x}) = \int p(head, \mu|\mathbf{x})\mathrm{d}\mu = \int P(head|\mu, \mathbf{x})p(\mu|\mathbf{x})\mathrm{d}\mu$$

$$= (H+1)/(N+2) \ = 751/1002 \ = 0.7495$$

- Note that we never computed value of μ
- Rule of succession used by Pierre-Simon Laplace to estimate that the probability of sun rising tomorrow is $(5000*365.25+1)/(5000*365.25+2)$

Distributions from our example

- Likelihood of observed data $P(X|\mu)$ given a parametric model of probability distribution
  - Bernoulli distribution with parameter μ
- Prior on the parameters of the model $p(\mu)$
  - Uniform prior as a special case of Beta distribution

- Posterior distribution of model parameters given an observed data

$$p(\mu|X) = \frac{P(X|\mu)p(\mu)}{P(X)}$$

- Posterior predictive distribution of a new observation give prior (training) observations

$$p(head|X) = \int P(head|\mu)p(\mu|X)\mathrm{d}\mu$$

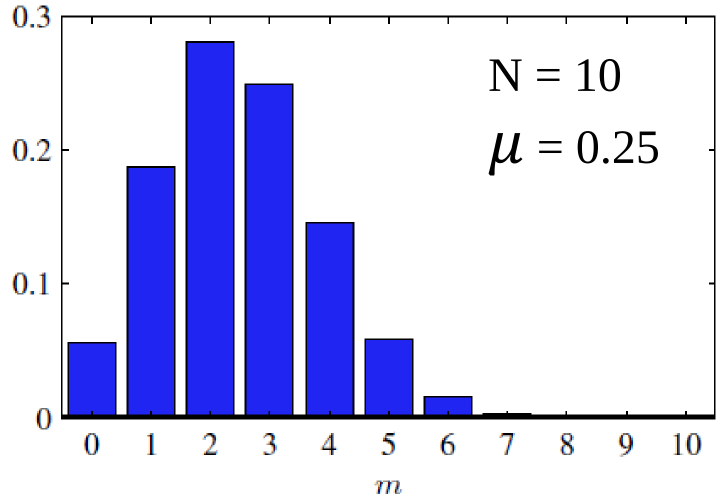
Bernoulli and Binomial distributions

$$\mathrm{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x}$$

- The “coin flipping” distribution is **Bernoulli distribution**
- Flipping the coin once, what is the probability of $x = 1$ (head) or $x = 0$ (tail)

$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

- Related **binomial distribution** is also described by single probability μ
- How many heads do I get if I flip the coin N times?

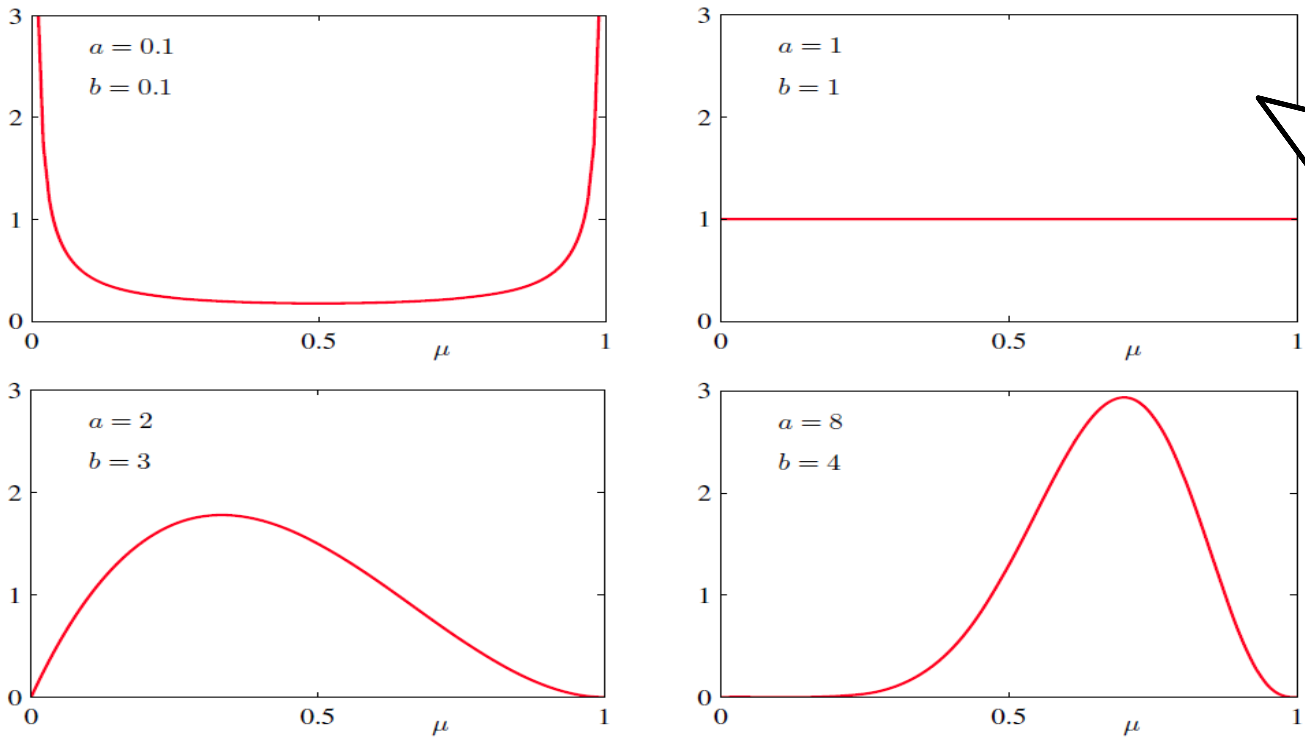


Beta distribution

Normalizing constant

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

- **Beta distribution** has "similar" form as Bern or Bin, but it is now function of μ
- Continuous distribution for μ over the interval (0,1)
- Can be used to express our prior beliefs about the Bernoulli dist. parameter μ



Uniform distribution over μ as was the prior in our coin flipping example

Beta as a conjugate prior

$$\mathbf{x} = [x_1, x_2, x_3, \dots x_N] = [1,0,0,1, \dots, 0]$$

$$P(\mathbf{x}|\mu) = \prod_i Bern(x_i|\mu) = \prod_i \mu^{x_i}(1-\mu)^{1-x_i} = \mu^H(1-\mu)^T$$

$$\text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

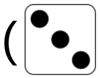
$$p(\mu|\mathbf{x}) = \frac{P(\mathbf{x}|\mu)p(\mu)}{P(\mathbf{x})} \propto \mu^H(1-\mu)^T\,\mu^{a-1}(1-\mu)^{b-1}$$

$$= \mu^{H+a-1}(1-\mu)^{T+b-1} \propto \text{Beta}(\mu|H+a, T+b)$$

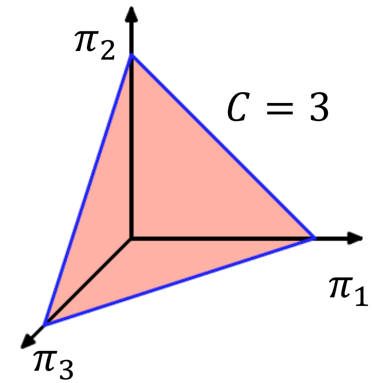
Sufficient statistics

- Using **Beta as a prior for Bernoulli parameter** μ results in **Beta posterior** distribution ➔ **Beta is conjugate prior to Bernoulli**
- $a-1$ and $b-1$ can be seen as a prior counts of heads and tails.
- Continuous distribution of μ over the interval (0,1)
- Beta distribution can be used to express our prior beliefs about the Bernoulli distributions parameter μ

Categorical and Multinomial distribution

$\mathbf{x} = [0, 0, 1, 0, 0, 0\ ]$ One-hot encoding of a discrete event ( on a dice)

$\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_C]$ Probabilities of the events (eg. $\left[\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right]$ for fair dice)



$$\mathrm{Cat}(\mathbf{x}|\boldsymbol{\pi}) = \prod_c \pi_c^{x_c}$$

$\sum_c \pi_c = 1$ ➔ $\boldsymbol{\pi}$ is a point on a simplex

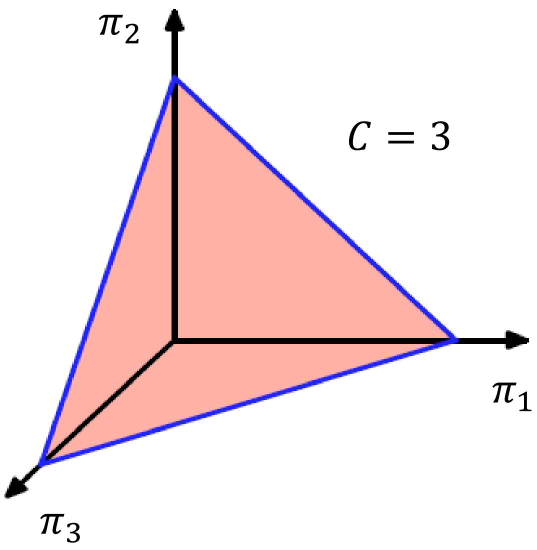
- **Categorical distribution** simply "returns" the probability of a given event $\mathbf{x}$
- Sample from the distribution is the event (or its one-hot encoding)

$$\mathrm{Mult}(m_1, m_2, \ldots, m_C|\boldsymbol{\pi}, N) = \binom{N}{m_1 m_2 \ldots m_C} \prod_c \pi_c^{m_c}$$

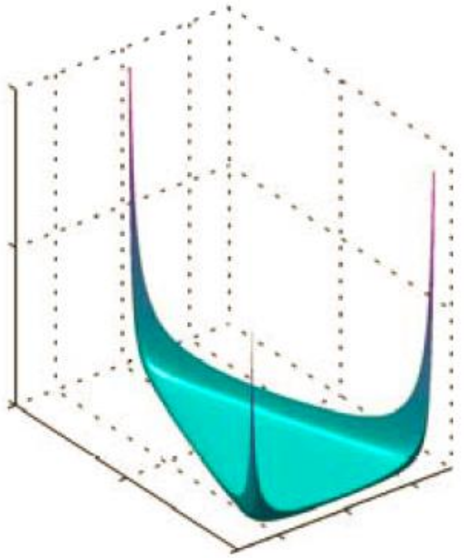
- **Multinomial distribution** is also described by single probability vector $\boldsymbol{\pi}$
- How many ones, twos, threes, … do I get if I throw the dice N times?
- Sample from the distribution is vector of numbers (e.g. 11x one, 8x two, …)

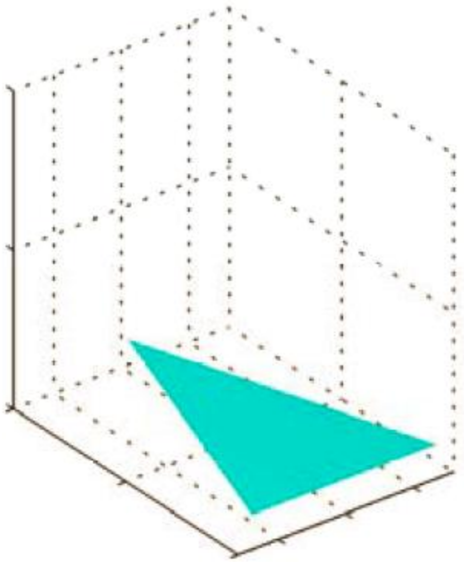
Dirichlet distribution

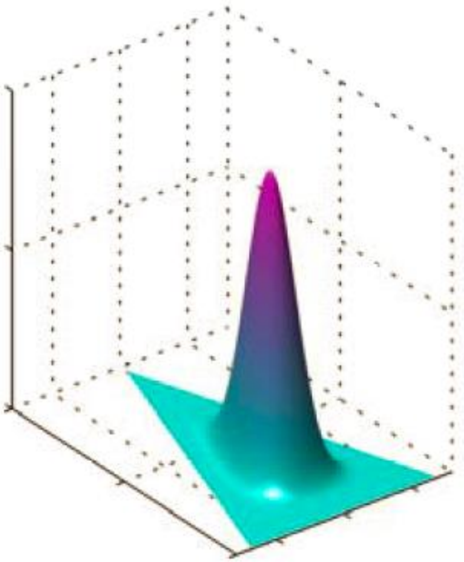
$$\mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_c \alpha_c)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_C)} \prod_{c=1} \pi_c^{\alpha_c - 1}$$



- **Dirichlet distribution** is continuous distribution over the points $\boldsymbol{\pi}$ on a K dimensional simplex.
- Can be used to express our prior beliefs about the categorical distribution parameter $\boldsymbol{\pi}$







Dirichlet as a conjugate prior

$$P(\mathbf{X}|\boldsymbol{\pi}) = \prod_n \mathrm{Cat}(\mathbf{x}_n|\boldsymbol{\pi}) = \prod_n \prod_c \pi_c^{x_{cn}} = \prod_c \pi_c^{m_c}$$

number of training observations of category c

$$\mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_c \alpha_c)}{\Gamma(\alpha_1)\dots\Gamma(\alpha_C)} \prod_{c=1} \pi_c^{\alpha_c - 1}$$

$$p(\boldsymbol{\pi}|\mathbf{X}) = \frac{P(\mathbf{X}|\boldsymbol{\pi})p(\boldsymbol{\pi})}{P(\mathbf{X})} \propto \prod_c \pi_c^{m_c} \prod_c \pi_c^{\alpha_c - 1}$$

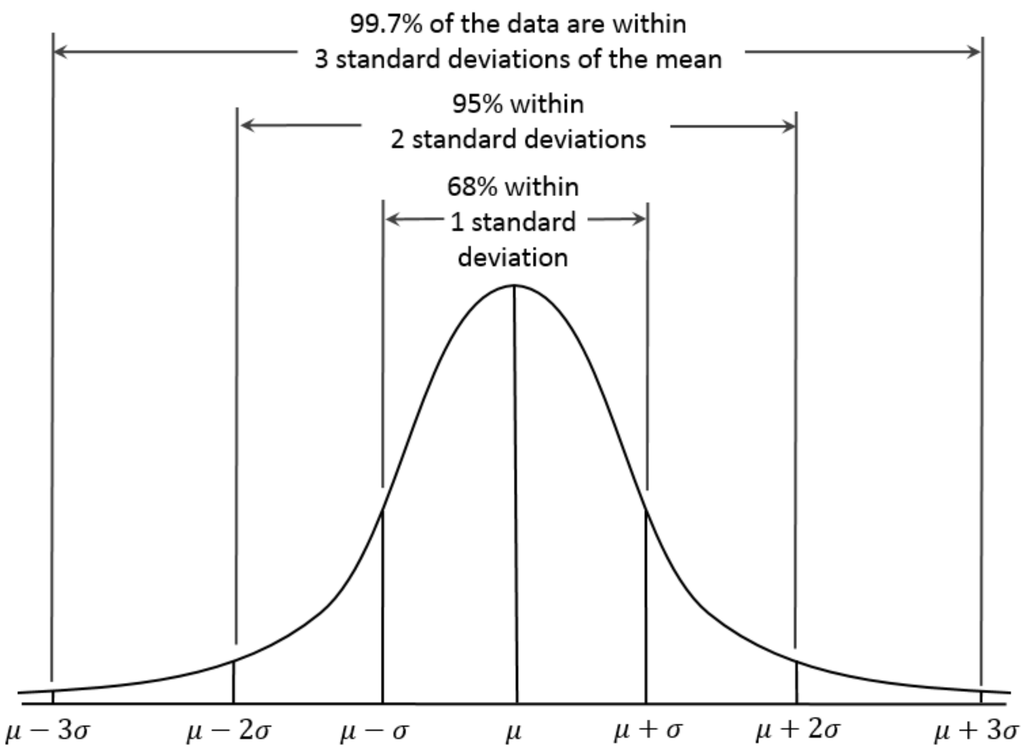
$$= \prod_{c=1} \pi_c^{m_c + \alpha_c - 1} \propto \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha} + \mathbf{m})$$

Sufficient statistics $\mathbf{m} = [m_1, \dots, m_C]$,

- Using **Dirichlet as a prior for Categorical parameter** $\boldsymbol{\pi}$ results in **Dirichlet posterior** distribution ➔ **Dirichlet is conjugate prior to Categorical dist.**
- $\alpha_c - 1$ can be seen as a prior count for the individual events.

Gaussian distribution (univariate)

$$p(x) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



ML estimates of parameters

$$\mu = \frac{1}{N} \sum_n x_n$$

$$\sigma^2 = \frac{1}{N} \sum_n (x_n - \mu)^2$$

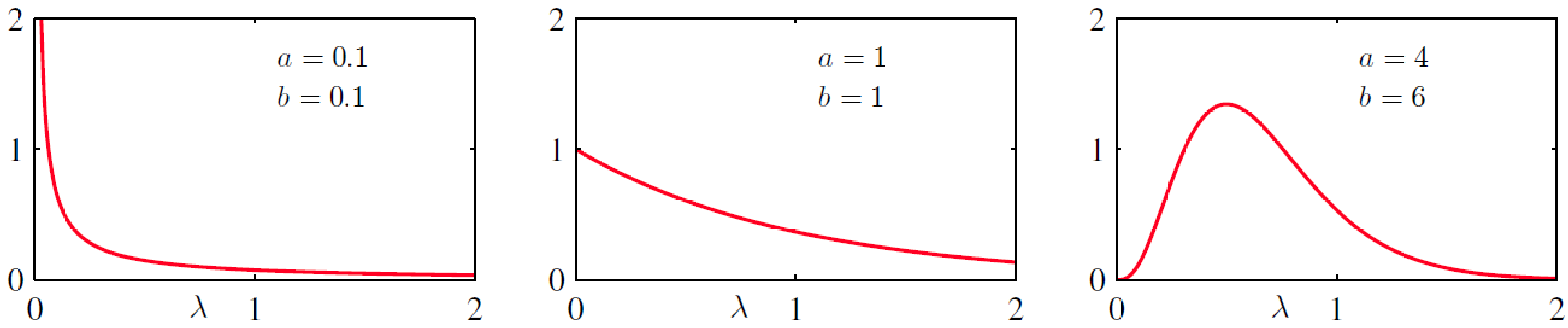
Gamma distribution

Normal distribution can be expressed in terms of precision $\lambda = \frac{1}{\sigma^2}$

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \sqrt{\frac{\lambda}{2\pi}} e^{-\frac{\lambda}{2}(x-\mu)^2}$$

$$\text{Gam}(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} e^{-b\lambda}$$

Gamma distribution defined for $\lambda > 0$ can be used as a prior over the precision

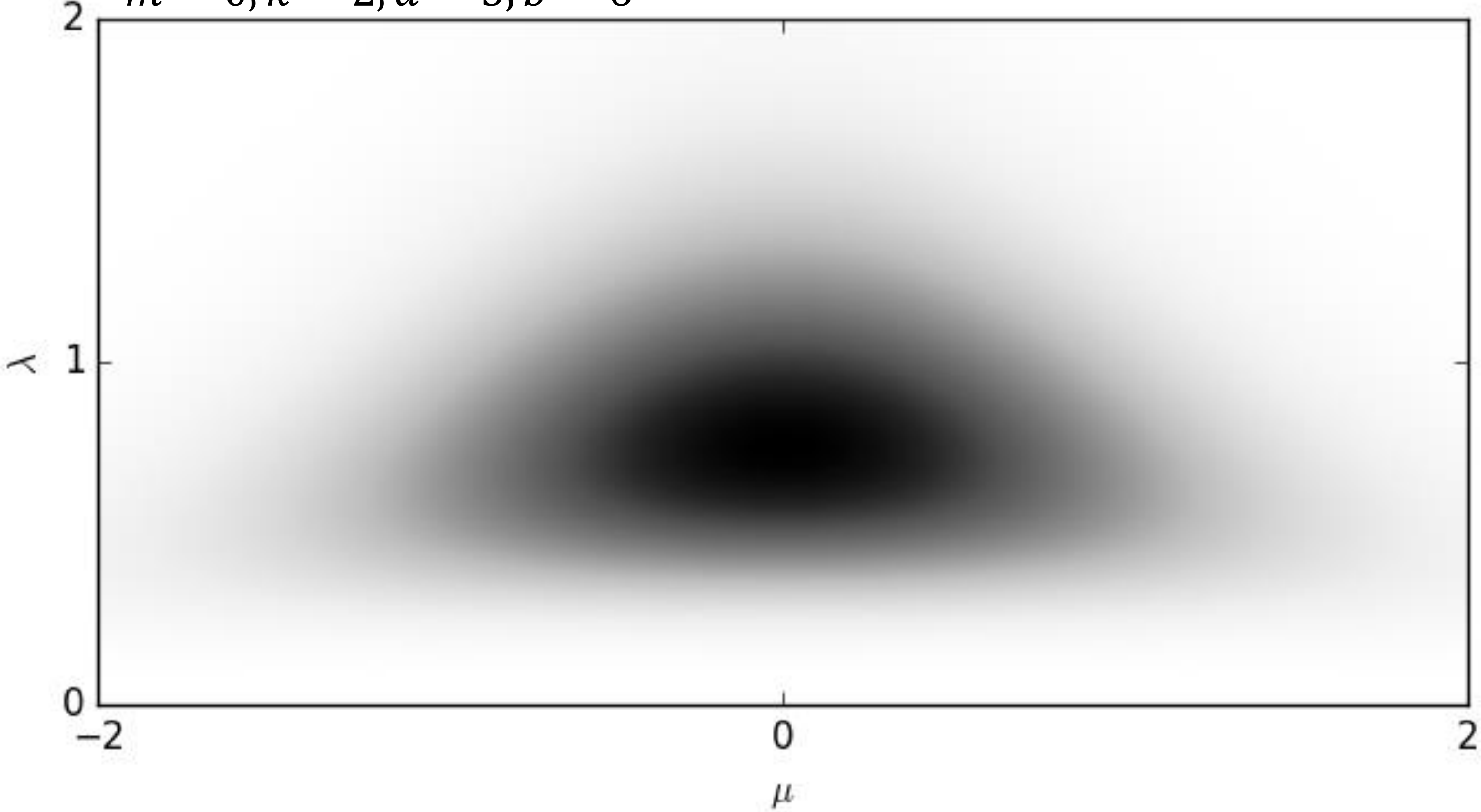


NormalGamma distribution

$$\text{NormalGama}(\mu, \lambda | m, \kappa, a, b) = \mathcal{N}(\mu | m, (\kappa\lambda)^{-1}) \text{Gam}(\lambda | a, b)$$

Joint distribution over μ and λ. Note that μ and λ are not independent.

$m = 0, \kappa = 2, a = 5, b = 6$



NormalGamma distribution

- **NormalGamma distribution** is the conjugate prior for Gaussian dist.
- Given observations $\mathbf{x} = [x_1, x_2, x_3, \dots x_N]$, the posterior distribution

$$p(\mu, \lambda|\mathbf{x}) = \frac{p(\mathbf{x}|\mu, \lambda)p(\mu, \lambda)}{p(\mathbf{x})}$$

$$\propto \prod_i \mathcal{N}(x_i; \mu, \sigma^2)\ \mathrm{NormalGamma}(\mu, \lambda | m, \kappa, a, b)$$

$$\propto \mathrm{NormalGamma}\left(\mu, \lambda \middle| \frac{\kappa m + N\bar{x}}{\kappa + N}, \kappa + N, a + \frac{N}{2}, b + \frac{N}{2}\left(s + \frac{\kappa(\bar{x} - m)^2}{\kappa + N}\right)\right)$$

Defined in terms of sufficient statistics N and $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$ $\quad s = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2$

Note that the prior parameters can be interpreted as follows:

$2a$ - prior number of observation for precision (or variance)

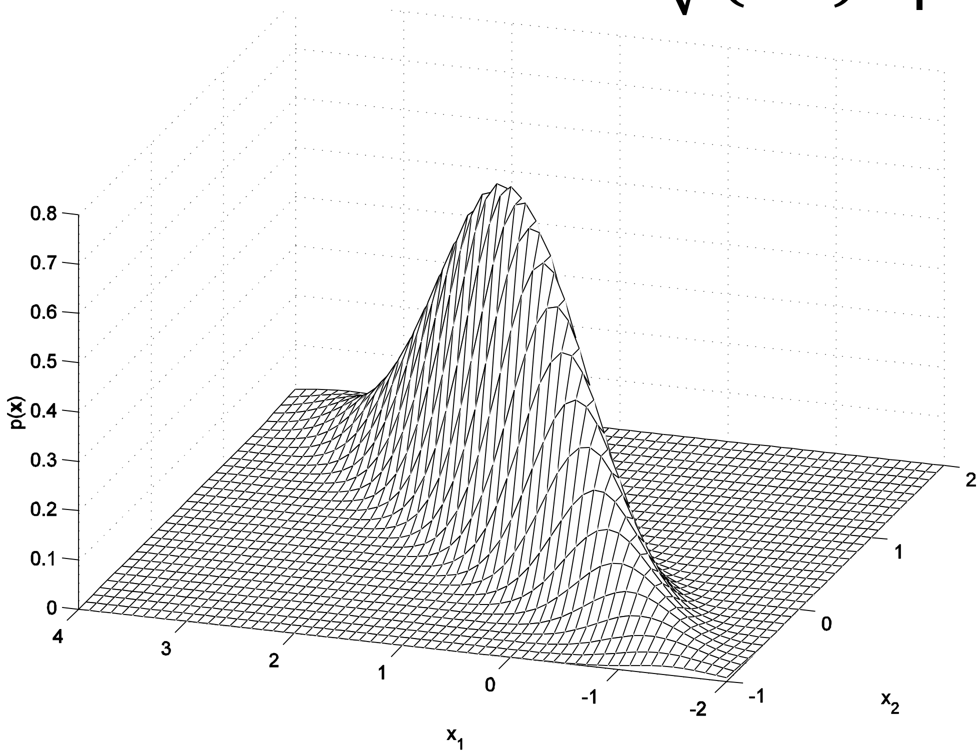
b/a - prior variance (around m)

κ - number of prior observations for mean

m - prior mean

Gaussian distribution (multivariate)

$$p(x_1, \dots, x_D) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$



ML estimates of parameters

$$\boldsymbol{\mu} = \frac{1}{N}\sum_n \mathbf{x}_n$$

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_n (\mathbf{x}_\mathrm{n} - \boldsymbol{\mu})(\mathbf{x}_\mathrm{n} - \boldsymbol{\mu})^T$$

Gaussian distribution (multivariate)

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

Conjugate prior is **Normal-Wishart**

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda} | \boldsymbol{\mu}_0, \beta, \mathbf{W}, \nu) = \mathcal{N}(\boldsymbol{\mu} | \boldsymbol{\mu}_0, (\beta\boldsymbol{\Lambda})^{-1}) \, \mathcal{W}(\boldsymbol{\Lambda} | \mathbf{W}, \nu)$$

where

$$\mathcal{W}(\boldsymbol{\Lambda} | \mathbf{W}, \nu) = B |\boldsymbol{\Lambda}|^{(\nu - D - 1)/2} \exp\left(-\frac{1}{2} \mathrm{Tr}(\mathbf{W}^{-1} \boldsymbol{\Lambda})\right)$$

is **Wishart distribution** and

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$$

Exponential family

- All the distributions described so far are distributions from the **exponential family**, which can be expressed in the following form
$$p(\mathbf{x}|\boldsymbol{\eta}) = \mathrm{h}(\mathbf{x})\, g(\boldsymbol{\eta}) \exp\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\}$$

- For example, for Gaussian distribution:
$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2} x^2 + \frac{\mu}{\sigma^2} x - \frac{\mu^2}{2\sigma^2}\right\}$$
$$\boldsymbol{\eta} = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix} \quad \mathbf{u}(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \quad g(\boldsymbol{\eta}) = \sqrt{-\frac{2\eta_2}{2\pi}} \exp\left(\frac{\eta_1^2}{4\eta_2}\right) \qquad \mathrm{h}(x) = 1$$

- To evaluate likelihood of set of observations:
$$\prod_n \mathcal{N}(x_n; \mu, \sigma^2) = \exp\left\{-\frac{1}{2\sigma^2} \sum_n x_n^2 + \frac{\mu}{\sigma^2} \sum_n x_n - N\left(\frac{\mu^2}{2\sigma^2} + \frac{\log(2\pi\sigma^2)}{2}\right)\right\}$$
$$= g(\boldsymbol{\eta})^{\mathrm{N}} \exp\left\{\boldsymbol{\eta}^T \sum_{n=1}^{N} \mathbf{u}(x_n)\right\} \prod_n \mathrm{h}(x_n)$$

Exponential family

For any distributions from exponential family

$$p(\mathbf{x}|\boldsymbol{\eta}) = \mathrm{h}(\mathbf{x})\, g(\boldsymbol{\eta})\, \exp\{\boldsymbol{\eta}^T \mathbf{u}(\mathbf{x})\}$$

- Likelihood $p(\mathbf{X}|\boldsymbol{\eta})$ of observed data $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]$ can be evaluated using the sufficient statistics N and $\sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)$:

$$p(\mathbf{X}|\boldsymbol{\eta}) = g(\boldsymbol{\eta})^{\mathrm{N}} \exp\left\{\boldsymbol{\eta}^T \sum_{n=1}^{N} \mathbf{u}(x_n)\right\} \prod_n \mathrm{h}(x_n)$$

- Conjugate prior distribution over parameter $\boldsymbol{\eta}$ exists in form:

$$p(\boldsymbol{\eta}|\boldsymbol{\theta}, \nu) = \mathrm{f}(\boldsymbol{\theta}, \nu)\, g(\boldsymbol{\eta})^{\nu} \exp\{\boldsymbol{\eta}^T \boldsymbol{\theta}\} = \mathrm{f}(\boldsymbol{\theta}, \nu) \exp\left\{\begin{bmatrix}\nu \\ \boldsymbol{\theta}\end{bmatrix}^T \begin{bmatrix}\ln g(\boldsymbol{\eta}) \\ \boldsymbol{\eta}\end{bmatrix}\right\} = \mathrm{f}(\boldsymbol{\theta}, \nu) \exp\{\, \widehat{\boldsymbol{\theta}}^T \mathbf{v}(\boldsymbol{\eta})\}$$

- Posterior distribution takes the same form as the conjugate prior, and we need only the prior parameters and the sufficient stats to evaluate it:

$$p(\boldsymbol{\eta}|\mathbf{X}) = p\left(\boldsymbol{\eta}\middle|\boldsymbol{\theta} + \textstyle\sum_{n=1}^{N} \mathbf{u}(x_n), \nu + N\right) \propto g(\boldsymbol{\eta})^{\mathrm{N}+\nu} \exp\left\{\boldsymbol{\eta}^T \left(\boldsymbol{\theta} + \sum_{n=1}^{N} \mathbf{u}(x_n)\right)\right\}$$

- $\boldsymbol{\theta}/\nu$ can be seen as a prior observation and ν as a prior count of observations

Parameter estimation revisited

- Let's estimate again parameters $\boldsymbol{\eta}$ of a chosen $p(\mathbf{x}|\boldsymbol{\eta})$ distribution given some of observed data $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]$
- Using the Bayes rule, we get the posterior distribution

$$p(\boldsymbol{\eta}|\mathbf{X}) = \frac{P(\mathbf{X}|\boldsymbol{\eta})p(\boldsymbol{\eta})}{P(\mathbf{X})}$$

- We can choose the most likelihood parameters: **Maximum a-posteriori (MAP)** estimate

$$\hat{\boldsymbol{\eta}}^{MAP} = \arg\max_{\boldsymbol{\eta}} p(\boldsymbol{\eta}|\mathbf{X}) = \arg\max_{\boldsymbol{\eta}} p(\mathbf{X}|\boldsymbol{\eta})p(\boldsymbol{\eta})$$

- Assuming flat (constant) prior $p(\boldsymbol{\eta}) = const$, we obtain **Maximum likelihood (ML)** estimate as a special case:

$$\hat{\boldsymbol{\eta}}^{ML} = \arg\max_{\boldsymbol{\eta}} P(\mathbf{X}|\boldsymbol{\eta})$$

Posterior predictive distribution

- We do not need to obtain a point estimate of the parameters $\hat{\boldsymbol{\eta}}$
- It is always good to postpone making hard decisions
- Instead, we can take into account the uncertainty encoded in the posterior distribution $p(\boldsymbol{\eta}|\mathbf{X})$ when evaluating **posterior predictive probability** for a new data point $\boldsymbol{x}'$ (as we did in our coin flipping example)

$$p(x'|\mathbf{X}) = \int p(x', \boldsymbol{\eta}|\mathbf{X}) \mathrm{d}\,\boldsymbol{\eta} = \int p(x'|\boldsymbol{\eta}) p(\boldsymbol{\eta}|\mathbf{X}) \mathrm{d}\,\boldsymbol{\eta}$$

- Rather than using one most likely setting of parameters $\hat{\boldsymbol{\eta}}$, we average over their different setting, which could possibly generate the observed data $\mathbf{X}$
  ➔ this approach is robust to overfitting

Posterior predictive for Bernoulli

- Beta prior on parameters of Bernoulli distribution leads to Beta posterior

$$p(\mu|\mathbf{x}) \propto \prod_n \mathrm{Bern}(x_n|\mu)\,\mathrm{Beta}(\mu|a_0, b_0) \propto \mathrm{Beta}(\mu|a_0 + H, b_0 + T)$$
$$= \mathrm{Beta}(\mu|a_N, b_N)$$

- The posterior predictive distribution is again Bernoulli

$$p(x'|\mathbf{x}) = \int p(x'|\mu)p(\mu|\mathbf{x})\,\mathrm{d}\mu = \int \mathrm{Bern}(x'|\mu)\mathrm{Beta}(\mu|a_N, b_N)\,\mathrm{d}\mu$$
$$= \mathrm{Bern}\left(x' \middle| \frac{a_N}{a_N + b_N}\right) = \mathrm{Bern}\left(x' \middle| \frac{a_0 + H}{a_0 + b_0 + N}\right)$$

- In our coin flipping example:

$$p(\mu) \quad = \mathcal{U}(0,1) = \mathrm{Beta}(\mu|a_0, b_0) = \mathrm{Beta}(\mu|1,1)$$
$$p(\mu|\mathbf{x}) \quad = \mathrm{Beta}(\mu|a_N, b_N) = \mathrm{Beta}(\mu|a_0 + H, b_0 + T) = \mathrm{Beta}(\mu|1 + 750, 1 + 250)$$
$$p(x'|\mathbf{x}) = \mathrm{Bern}\left(x' \middle| \frac{a_N}{a_N + b_N}\right) = 751/1002 \quad = 0.7495$$

Posterior predictive for Categorical

- Dirichlet prior on parameters of Categorical distribution leads to Dirichlet posterior

$$p(\boldsymbol{\pi}|\mathbf{X}) \propto \prod_n \mathrm{Cat}(\mathbf{x}_n|\boldsymbol{\pi})\,\mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) \propto \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0 + \mathbf{m}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_N)$$

- The posterior predictive distribution is again Categorical

$$p(\mathbf{x}'|\mathbf{X}) = \int p(\mathbf{x}'|\boldsymbol{\pi})p(\boldsymbol{\pi}|\mathbf{X})\,\mathrm{d}\boldsymbol{\pi} = \int \mathrm{Cat}(\mathbf{x}'|\boldsymbol{\pi})\mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_N)\,\mathrm{d}\boldsymbol{\pi}$$

$$= \mathrm{Cat}\left(\mathbf{x}' \middle| \frac{\boldsymbol{\alpha}_N}{\sum_c \alpha_{Nc}}\right) = \mathrm{Cat}\left(\mathbf{x}' \middle| \frac{\boldsymbol{\alpha}_0 + \mathbf{m}}{\sum_c \alpha_{0c} + m_c}\right)$$

Student's t-distribution

- NormalGamma prior on parameters of Gaussian distribution leads to NormalGamma posterior

$$p(\mu, \lambda|\mathbf{x}) \propto \prod_i \mathcal{N}(x_i; \mu, \sigma^2)\ \text{NormalGamma}(\mu, \lambda|m_0, \kappa_0, a_0, b_0)$$

$$\propto \text{NormalGamma}\left(\mu, \lambda \middle| \frac{\kappa_0 m_0 + N\bar{x}}{\kappa_0 + N}, \kappa_0 + N, a_0 + \frac{N}{2}, b_0 + \frac{N}{2}\left(s + \frac{\kappa_0(\bar{x} - m_0)^2}{\kappa_0 + N}\right)\right)$$

$$= \text{NormalGamma}(\mu, \lambda|m_N, \kappa_N, a_N, b_N)$$

- The posterior predictive distribution is Student's t-distribution

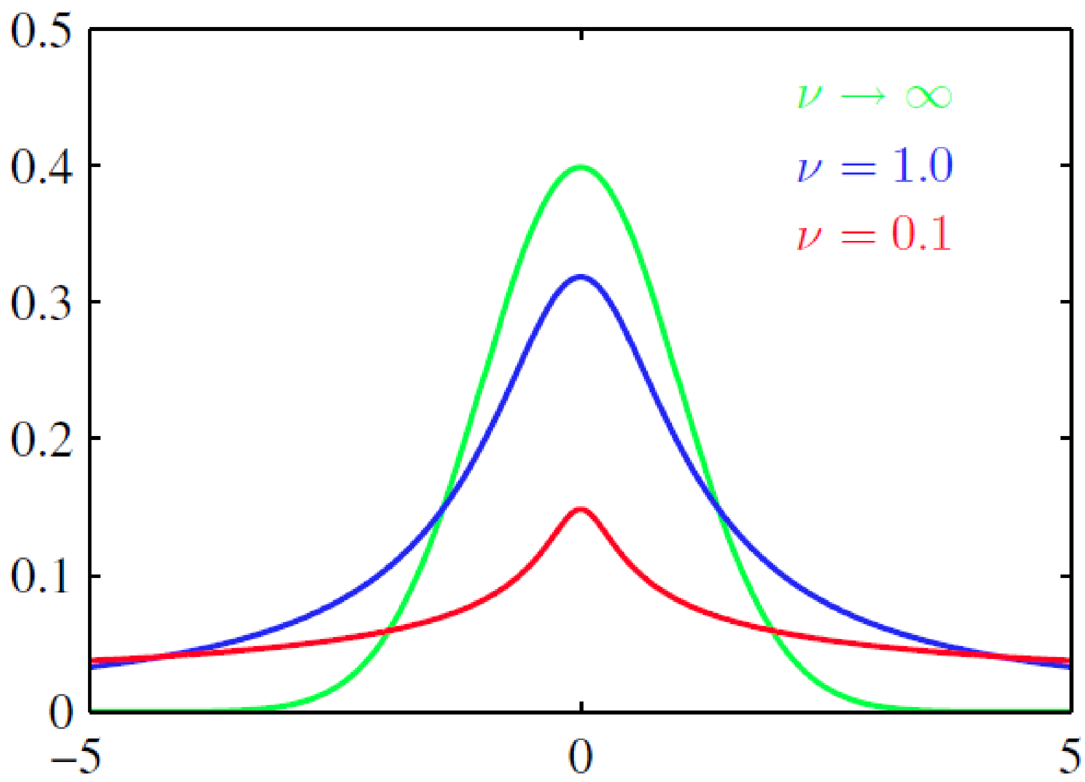
$$\mathrm{p}(\mathrm{x}'|\mathbf{x}) = \iint p(x'|\mu, \lambda)\mathrm{p}(\mu, \lambda|\mathbf{x})\ \mathrm{d}\mu\ \mathrm{d}\lambda$$

$$= \iint \mathcal{N}(x'|\mu, \lambda^{-1})\text{NormalGamma}(\mu, \lambda|m_N, \kappa_N, a_N, b_N)\ \mathrm{d}\mu\ \mathrm{d}\lambda$$

$$= \mathrm{St}\left(x'|m_N, 2a_N, \frac{a_N \kappa_N}{b_N(\kappa_N + 1)}\right)$$

Student's t-distribution

$$\mathrm{St}(x \mid \mu, \nu, \gamma) = \frac{\Gamma\left(\frac{\nu}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(\frac{\gamma}{\pi\nu}\right)^{\frac{1}{2}} \left[1 + \frac{\gamma(x - \mu)^2}{\nu}\right]^{-\frac{\nu}{2}-\frac{1}{2}}$$



- Gaussian distribution is a special case of Student's with degree of freedom $\nu \to \infty$
- For the posterior $p(\mu, \lambda|\mathbf{x})$, $\nu = 2a_N = 2a_0 + N$