Class Activity 2 [20 minutes]: ~~Beta-Binomial~~ *Binomial-Discrete (prior)*

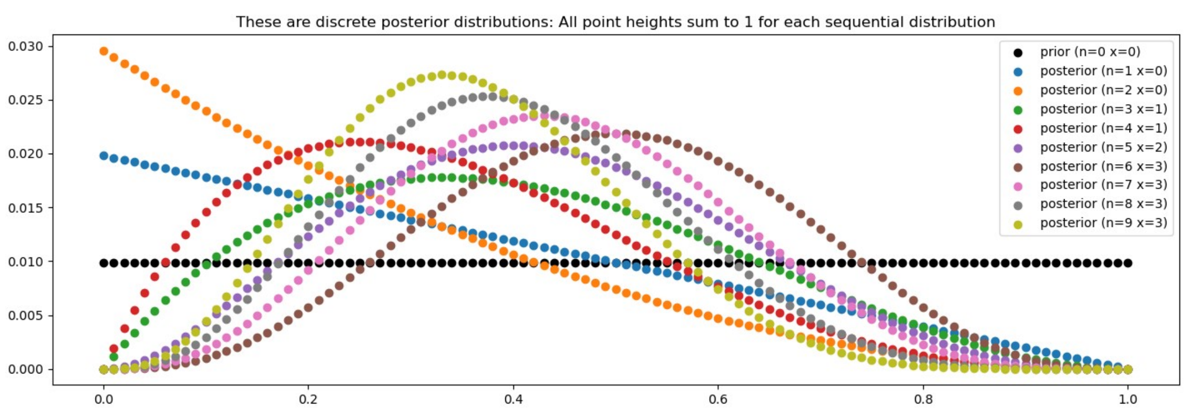
For **Bernoulli data** $n_1 = 1$ and $x_1 = 0$ and **prior**

$$p(\theta = p) = \begin{cases} \frac{1}{101} & \text{for } p = \frac{i}{101} \text{ for } i = 0, 1, \cdots, 100 \\ 0 & \text{otherwise} \end{cases}$$

1. What is $p\left(\theta = \frac{i}{101} | x\right)$ for $i = 0, 1, \cdots, 100$?
2. Keep going... what is the sequence of **priors** for the observations $0, 0, 1, 0, 1, 1, 0, 0, 0$ up to the final **binomial data** of $n = 9$ and $x_9 = 3$?
3. Bonus: is **sequential** or **batch** updating more computationally efficient?

Class Activity 2: Batch Solution [15 minutes]

In [21]:
```
import numpy as np; from scipy import stats; import matplotlib.pyplot as plt
observations = 0,0,1,0,1,1,0,0,0; parameter_domain = np.linspace(0,1,101) # pa
n,x,prior = 0,0,np.ones(101)/101; plt.figure(figsize=(16,5))
plt.scatter(parameter_domain, prior, label="prior (n="+str(n)+" x="+str(x)+")"
for observation in observations:
    n += 1; x += observation; likelihood = stats.binom.pmf(n=n, k=x, p=paramet
    joint = likelihood*prior # elementwise multiplication
    normalized_joint = joint/joint.sum()
    plt.scatter(parameter_domain, normalized_joint, label="posterior (n="+str(
plt.legend(); plt.title("These are discrete posterior distributions: All point
```



Class Activity 2: Sequential Solution [10 minutes]

Authored by Leo Watson

```
import numpy as np; from scipy import stats; import matplotlib.pyplot as plt
# Sequential Learning <-- Using initial prior and data, want to update posterior.
# Then, keep doing this recursively with posterior of previous iteration becoming
# prior of new iteration
# Assumptions:
#   Assuming data comes from a Bernoulli distribution.
data = [0, 0, 1, 0, 1, 1, 0, 0, 0] # n = 9, x (num "successes") = 3
n = 0 # num datapoints so far
s = 0 # num successes so far
xk = np.arange(101) / 101
pk = (1/101,) * 101
parameter_domain = np.linspace(0,1,101)
initial_prior = stats.rv_discrete(name='initial_prior', values=(xk, pk))
current_prior = initial_prior.pmf(xk)
plt.figure(figsize=(10,4))
#plt.legend(loc='best')
# pdf for discrete uniform
for l in data:
  n+= 1; s += l
  liklihood = stats.bernoulli.pmf(k = l, p = parameter_domain)
  # normalizing constant
  joint = current_prior * liklihood # Note that joint is simply product of likelihoods b/c indep.
  updated_posterior = liklihood * initial_prior / joint.sum()
  current_prior = updated_posterior # new_prior_is_now_last_posterior = stats.rv_discrete(name='new_prior', values=(xk, updated_posterior))
  plt.scatter(parameter_domain, initial_prior, label="posterior (n="+str(n)+" x="+str(x)+")")
```

The results and figures of by the Sequential and Batch methods match

Class Activity 2: Bonus Timing [5 minutes]

- Authored by Leo Watson -- Note: "Timing implementation isn't perfect; approx" -- Leo

> "I think the runtime is quite a bit slower for batch though. I believe sequential is O(n) as each observation is only considered once when updating the posterior, while batch is O(n^2) because for n observations, use first observation in batch update n times, second n-1 times, and so forth => (1+2+ .. + n) = n(n+1)/2. I empirically found batch to be significantly slower than sequential for 10000 iterations in my code as well (approx 50% slower), which was a good sanity check." -- Leo

Batch	Sequential

```
import time
import numpy as np; from scipy impor
t stats; import matplotlib.pyplot as
```

```
# Sequential Learning <-- Using ini
tial prior and data, want to update
posterior.
# Then, keep doing this recursively
with posterior of previous iteratio
n becoming
# prior of new iteration
```

The first (toy) model: beta-binomial [5 minutes]

$$p(x|\theta, n) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \qquad [n \text{ given}]$$

$$p(\theta|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad \alpha \text{ and } \beta \text{ are called } \textbf{hyperparameters}$$

$$p(\theta|x) = ?$$
$$\text{Hint: } p(\theta|x) \propto p(\theta, x)$$

$$p(x|\theta)p(\theta)$$
$$\binom{n}{k} \theta^x (1-\theta)^{n-x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$
$$\propto \theta^x (1-\theta)^{n-x} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$
$$\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \Rightarrow \text{What distribution?}$$

The first (toy) model: beta-binomial [4 minutes]

$$p(x|\theta, n) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \qquad [n \text{ given}]$$

$$p(\theta|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad \alpha \text{ and } \beta \text{ are called } \textbf{hyperparameters}$$

$$p(\theta|x) = ?$$
$$\text{Hint: } p(\theta|x) \propto p(\theta, x)$$

$$\begin{aligned}
&= p(x|\theta)p(\theta) \\
&= \binom{n}{k} \theta^x (1-\theta)^{n-x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&\propto \theta^x (1-\theta)^{n-x} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&= \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \Rightarrow \text{What distribution?}
\end{aligned}$$

The second (toy) model: binomial-discrete [6 minutes]

$$p(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \qquad\qquad [n \text{ given}]$$

$$p(\theta = p) = \begin{cases} \frac{1}{101} & \text{for } p = \frac{i}{101} \text{ for } i=0,1,\cdots,100 \\ 0 & \text{otherwise} \end{cases} \qquad \begin{array}{l} \text{The 'uniform' choice is} \\ \textbf{the 'hyperparameter'} \end{array}$$

$$p(\theta|x) = \,?$$
$$\text{Hint: } p(\theta|x) \propto p(\theta, x) = p(x|\theta)p(\theta)$$

$$\binom{n}{k} \theta^x (1-\theta)^{n-x} \genfrac{}{}{0pt}{}{\sum_{i=0}^{101} \delta_i(\theta)}{101} \begin{array}{l} \leftarrow \text{Dirac delta functions} \\ \quad\;\text{Point mass functions} \end{array}$$
$$\Rightarrow \text{What distribution?}$$

The second (toy) model: binomial-discrete [5 minutes]

$$p(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \qquad\qquad [n \text{ given}]$$

$$p(\theta = p) = \begin{cases} \frac{1}{101} & \text{for } p = \frac{i}{101} \text{ for } i=0,1,\cdots,100 \\ 0 & \text{otherwise} \end{cases} \qquad \begin{array}{l} \text{The 'uniform' choice is} \\ \textbf{the 'hyperparameter'} \end{array}$$

$$p(\theta|x) = \,?$$
$$\text{Hint: } p(\theta|x) \propto p(\theta, x) = p(x|\theta)p(\theta)$$

$$\binom{n}{k} \theta^x (1-\theta)^{n-x} \frac{\sum_{i=0}^{101} \delta_i(\theta)}{101} \begin{array}{l} \leftarrow \text{Dirac delta functions} \\ \quad\;\text{Point mass functions} \end{array}$$
$$\Rightarrow \text{What distribution?}$$

A third (toy) model: normal-normal [12 minutes]

$$p(x|\theta, \phi = \sigma^{-2}) = \prod_{i=1}^{n} \sqrt{\frac{\phi}{2\pi}} e^{-\frac{\phi(x_i-\theta)^2}{2}} \quad [\textbf{precision } \phi = \sigma^{-2} \text{ given}]$$

$$p(\theta|\theta_0, \tau) = \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau(\theta-\theta_0)^2}{2}} \quad \begin{array}{l} \theta_0 \text{ and the } \textbf{precision } \tau \text{ are} \\ \text{called } \textbf{hyperparameters} \end{array}$$

$$p(\theta|x) = ?$$
$$\text{Hint: } p(\theta|x) \propto p(\theta, x) = p(x|\theta)p(\theta)$$

$$\left[\prod_{i=1}^{n} \sqrt{\frac{\phi}{2\pi}} e^{-\frac{\phi(x_i-\theta)^2}{2}}\right] \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau(\theta-\theta_0)^2}{2}}$$
$$\propto \exp\left[-\frac{1}{2}\left(\left(\sum_{i=1}^{n} \phi(x_i-\theta)^2\right) + \tau(\theta-\theta_0)^2\right)\right]$$

A third (toy) model: normal-normal [3 minutes]

$$p(x|\theta, \phi = \sigma^{-2}) = \prod_{i=1}^{n} \sqrt{\frac{\phi}{2\pi}} e^{-\frac{\phi(x_i-\theta)^2}{2}} \quad [\textbf{precision } \phi = \sigma^{-2} \text{ given}]$$

$$p(\theta|\theta_0, \tau) = \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau(\theta-\theta_0)^2}{2}} \quad \begin{array}{l} \theta_0 \text{ and the } \textbf{precision } \tau \text{ are} \\ \text{called } \textbf{hyperparameters} \end{array}$$

$$p(\theta|x) = ?$$
$$\text{Hint: } p(\theta|x) \propto p(\theta, x) = p(x|\theta)p(\theta)$$

$$= \left[\prod_{i=1}^{n} \sqrt{\frac{\phi}{2\pi}} e^{-\frac{\phi(x_i-\theta)^2}{2}}\right] \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau(\theta-\theta_0)^2}{2}}$$
$$\propto \exp\left[-\frac{1}{2}\left(\left(\sum_{i=1}^{n} \phi(x_i-\theta)^2\right) + \tau(\theta-\theta_0)^2\right)\right]$$
$$\propto \exp\left(-\frac{1}{2}(\tau + n\phi)\left(\theta - \frac{1}{\tau + n\phi}\left(\tau\theta_0 + \phi\sum_{i=1}^{n} x_i\right)\right)^2\right)$$

A third (toy) model: normal-normal [5 minutes]

$$p(x|\theta, \phi = \sigma^{-2}) = \prod_{i=1}^{n} \sqrt{\frac{\phi}{2\pi}} e^{-\frac{\phi(x_i-\theta)^2}{2}} \quad [\textbf{precision } \phi = \sigma^{-2} \text{ given}]$$

$$p(\theta|\theta_0, \tau) = \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau(\theta-\theta_0)^2}{2}} \quad \begin{array}{l}\theta_0 \text{ and the } \textbf{precision } \tau \text{ are} \\ \text{called } \textbf{hyperparameters}\end{array}$$

$$\begin{aligned} p(\theta|x) &= ? \\ \text{Hint: } p(\theta|x) &\propto p(\theta, x) = p(x|\theta)p(\theta) \\ &= \left[\prod_{i=1}^{n} \sqrt{\frac{\phi}{2\pi}} e^{-\frac{\phi(x_i-\theta)^2}{2}}\right] \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau(\theta-\theta_0)^2}{2}} \\ \begin{array}{r}\text{What distribution} \\ \text{is this?} \longrightarrow\end{array} &\propto \exp\left(-\frac{1}{2}(\tau + n\phi)\left(\theta - \frac{1}{\tau + n\phi}\left(\tau\theta_0 + \phi\sum_{i=1}^{n} x_i\right)\right)^2\right) \end{aligned}$$

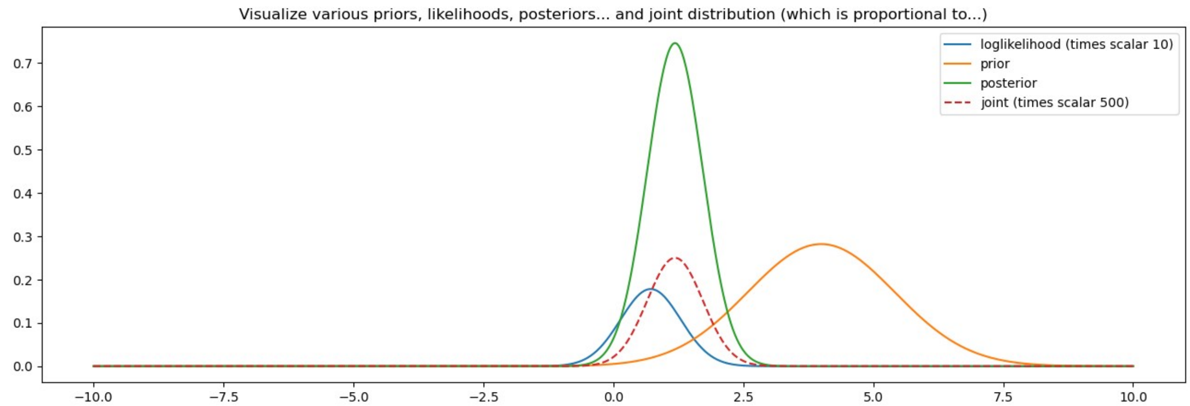
Homework #2 Part I:
Prove that as a function of θ...

$$\begin{aligned} &\propto \exp\left[-\frac{1}{2}\left(\left(\sum_{i=1}^{n} \phi(x_i - \theta)^2\right) + \tau(\theta - \theta_0)^2\right)\right] \\ &\propto \exp\left(-\frac{1}{2}(\tau + n\phi)\left(\theta - \frac{1}{\tau + n\phi}\left(\tau\theta_0 + \phi\sum_{i=1}^{n} x_i\right)\right)^2\right) \end{aligned}$$

Hint: expand the squares, and rebuild the quadratic in terms of θ...

The "Classic Bayes" Figure [15 minutes]

In [27]:
```
import numpy as np; from scipy import stats; import matplotlib.pyplot as plt
n = 3; x = stats.norm(loc=0,scale=1).rvs(n); theta_0,tau = 4,1/2; parameter_do
likelihood = stats.norm.pdf((x[:,np.newaxis])*np.ones(parameter_domain.shape),
                            loc=parameter_domain, scale=1).prod(axis=0) # sigm
prior = stats.norm(loc=theta_0, scale=1/tau**0.5).pdf(parameter_domain) # What
c,c2 = 500,10;
joint = c*likelihood*prior # elementwise multiplication
posterior = stats.norm(loc=(tau*theta_0+x.sum())/(tau+n), scale=1/(tau+n)**0.5
plt.figure(figsize=(16,5)); plt.plot(parameter_domain, c2*likelihood, label="l
```



Homework #2 Part II: Complete the following...

1. Make the "Classic Bayes" figure, but instead of a single likelihood, plot the individual likelihoods of each observation. *Hint: Consider the code...*

```
for xi in range(x):
    print(xi)
```

2. Make the "Classic Bayes" figure for the normal-normal model on a log scale. *Hint: plot loglikehood, prior logpdf, and the log of the joint distribution using* `stats.norm.logpdf()`
3. What is the ***posterior distribution*** of a **normal likelihood** with $p(\theta) \propto 1$?
4. How do ***log likelihoods*** and ***posteriors*** differ from each other as a result of ***improper priors*** pdiscussed below] and ***marginal likelihood constants***?

Types of Priors [7 minutes]

- **Conjugate Prior**
  - The **posterior** parametric family same as **prior**
  - **Beta-binomial** and **normal-normal** specifications are examples of this
- **Flat Prior** $p(\theta) \propto 1$
  - A $\text{Beta}(\alpha = 1, \beta = 1)$ **prior** for the **beta-binomial** model is a **flat prior**
- **Improper Prior**
  - **Improper priors** don't integrate to 1 or guanatee **proper posteriors**
  - A $p(\mu) \propto 1$ **prior** for a **normal likelihood** is an **improper prior** but it does in fact

produce a **proper normal posterior** in this specific case

Types of Priors [8 minutes]

- **Weakly Informative** and **Informative priors**
  - For the **beta-binomial** this depends on the relative sizes of x, n, α, β

$$p(\theta|x,n,\alpha,\beta) = \text{Beta}(n-x+\alpha, x+\beta)$$
$$\propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$

  - For **normal-normal** it depends on the relative values of ϕ, n, σ / x, θ

$$p(\theta|x,\theta_0,\tau,\phi) = \text{N}\left(\overbrace{\frac{\left(\tau\theta_0 + \phi\sum_{i=1}^n x_i\right)}{(\tau+n\phi)}}^{\textbf{mean}}, \overbrace{\tau+n\phi}^{\textbf{precision}}\right)$$
$$\propto \exp\left[-\frac{1}{2}(\tau+n\phi)\left(\theta - \frac{1}{\tau+n\phi}\left(\tau\theta_0 + \phi\sum_{i=1}^n x_i\right)\right)^2\right]$$

Types of Priors [5 minutes]

- **Diffuse priors**
  - A $p(\theta) = \text{Beta}(\alpha = \epsilon_1, \beta = \epsilon_2)$ **prior** specification with very small ϵ_1 and ϵ_2 for the **beta-binomial** model is **uninformative**
  - While a $p(\mu) = N(\mu_0, c)$ for some very large constant c for a **normal-normal** model specification is **diffuse** and **uninformative**
- There are **other kinds of priors** that can be encountered as well...
  - Some examples are **Jeffreys** (https://en.wikipedia.org/wiki/Jeffreys_prior) (***Fisher information*** based $p(\theta) \propto \sqrt{\det \mathcal{I}(\theta)}$) **priors** , **Ojective** (https://en.wikipedia.org/wiki/Prior_probability#Uninformative_priors)

(often misnomered as **Uninformative** (https://en.wikipedia.org/wiki/Prior_probability#Uninformative_priors)) **priors**, and **Empirical Bayes** (https://en.wikipedia.org/wiki/Empirical_Bayes_method#:~:text=Empirical%20Bayes%20methods%20are%20procedures. (data estimated) **priors**

A fourth (toy) model: normal-gamma [ϕ unknown] Extending the third (toy) model [15 minutes]

$$
\begin{aligned}
p(x|\theta, \phi = \sigma^{-2}) &= \prod_{i=1}^{n} \sqrt{\frac{\phi}{2\pi}} e^{-\frac{\phi(x_i-\theta)^2}{2}} && [\textbf{precision } \phi = \sigma^{-2} \text{ unknown}]\\
p(\theta|\theta_0, \tau) &= \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau(\theta-\theta_0)^2}{2}} && \theta_0 \text{ and the } \textbf{precision } \tau \text{ are called } \textbf{hyperparameters}\\
p(\phi|\alpha, \beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \phi^{\alpha-1} e^{-\beta\phi} && \text{Gamma hyperparameters shape } \alpha \text{ rate } \beta \text{ (scale } 1/\beta)
\end{aligned}
$$

A fourth (toy) model: normal-gamma $[\phi \text{ unknown}]$ Extending the third (toy) model [5 minutes]

$$
\begin{aligned}
p(x|\theta, \phi = \sigma^{-2}) &= \prod_{i=1}^{n} \sqrt{\frac{\phi}{2\pi}} e^{-\frac{\phi(x_i-\theta)^2}{2}} && [\textbf{precision } \phi = \sigma^{-2} \text{ unknown}]\\
p(\theta|\theta_0, \tau) &= \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau(\theta-\theta_0)^2}{2}} && \theta_0 \text{ and the } \textbf{precision } \tau \text{ are called } \textbf{hyperparameters}\\
p(\phi|\alpha, \beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \phi^{\alpha-1} e^{-\beta\phi} && \text{Gamma hyperparameters shape } \alpha \text{ rate } \beta \text{ (scale } 1/\beta)
\end{aligned}
$$

$$
\begin{aligned}
p(\theta|x) &= ?\\
\text{Hint: } p(\theta, \phi|x) &\propto p(\theta, \phi, x) = p(x|\theta)p(\theta)p(\phi) \quad (\theta \perp\!\!\!\perp \phi) \leftarrow \text{priors}\\
&= \left[\prod_{i=1}^{n} \sqrt{\frac{\phi}{2\pi}} e^{-\frac{\phi(x_i-\theta)^2}{2}}\right] \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau(\theta-\theta_0)^2}{2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \phi^{\alpha-1} e^{-\beta\phi}
\end{aligned}
$$

A fourth (toy) model: normal-gamma $[\phi \text{ unknown}]$ Extending the third (toy) model [5 minutes]

$$
\begin{aligned}
p(\theta, \phi|x) &\propto p(\theta, \phi, x) = p(x|\theta)p(\theta)p(\phi) \quad (\theta \perp\!\!\!\perp \phi) \leftarrow \text{priors}\\
&= \left[\prod_{i=1}^{n} \sqrt{\frac{\phi}{2\pi}} e^{-\frac{\phi(x_i-\theta)^2}{2}}\right] \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau(\theta-\theta_0)^2}{2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \phi^{\alpha-1} e^{-\beta\phi}
\end{aligned}
$$

$$
\Downarrow
$$

$$
p(\theta|x, \theta_0, \tau, \phi) = \text{N}\left(\overbrace{\frac{\left(\tau\theta_0 + \phi\sum_{i=1}^{n} x_i\right)}{(\tau + n\phi)}}^{\textbf{mean}}, \overbrace{\tau + n\phi}^{\textbf{precision}}\right) \quad \longleftarrow \text{The first homework problem works on the mean}
$$

$\longleftarrow$ Do you see why this precision specification is true?

$$
p(\phi|x, \alpha, \beta, \theta) = \text{Gamma}\left(\overbrace{\alpha + \frac{n}{2}}^{\textbf{shape}}, \overbrace{\beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2}^{\textbf{pratep}}\right) \quad \longleftarrow \text{And do you see why this is true?}
$$

A fourth (toy) model: normal-gamma [5 minutes]

Full (joint) posterior

$$p(\theta, \phi|x) = p(\theta, \phi|x, \theta_0, \tau, \alpha, \beta)$$

- How do we sample from $p(\theta, \phi|x)$? We have **full conditional distributions**!!

full conditional distributions

$$p(\theta|x, \theta_0, \tau, \phi) = \text{N}\left(\frac{\left(\tau\theta_0 + \phi \sum_{i=1}^n x_i\right)}{(\tau + n\phi)}, \tau + n\phi\right)$$

$$p(\phi|x, \alpha, \beta, \theta) = \text{Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^n (x_i - \theta)^2\right)$$