Bayesian Inference for Some Basic Models

Piyush Rai

Topics in Probabilistic Modeling and Inference (CS698X)

Jan 12, 2019
Recap: Bayesian Inference

- Given data $X$ from a model $m$ with parameters $\theta$, the posterior over the parameters $\theta$

$$p(\theta|X, m) = \frac{p(X, \theta|m)}{p(X|m)}$$
Recap: Bayesian Inference

Given data $X$ from a model $m$ with parameters $\theta$, the posterior over the parameters $\theta$

$$p(\theta|X, m) = \frac{p(X, \theta|m)}{p(X|m)} = \frac{p(X|\theta, m)p(\theta|m)}{\int p(X|\theta, m)p(\theta|m)d\theta}$$

Can use the posterior for various purposes, e.g.,

- Getting point estimates e.g., mode (though, for this, directly doing point estimation is often easier)
- Uncertainty in our estimates of $\theta$ (variance, credible intervals, etc)
- Computing the posterior predictive distribution (PPD) for new data, e.g.,

$$p(x^*|X, m) = \int p(x^*|\theta, m)p(\theta|m)d\theta$$

Caveat: Computing the posterior/PPD is in general hard (due to the intractable integrals involved)
Recap: Bayesian Inference

Given data $X$ from a model $m$ with parameters $\theta$, the posterior over the parameters $\theta$

$$p(\theta|X, m) = \frac{p(X, \theta|m)}{p(X|m)} = \frac{p(X|\theta, m)p(\theta|m)}{\int p(X|\theta, m)p(\theta|m)d\theta} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}}$$

Can use the posterior for various purposes, e.g.,
- Getting point estimates e.g., mode (though, for this, directly doing point estimation is often easier)
- Uncertaintly in our estimates of $\theta$ (variance, credible intervals, etc)
- Computing the posterior predictive distribution (PPD) for new data, e.g.,

$$p(x^*|X, m) = \int p(x^*|\theta, m)p(\theta|X, m)d\theta$$

Caveat: Computing the posterior/PPD is in general hard (due to the intractable integrals involved)
Recap: Bayesian Inference

- Given data $X$ from a model $m$ with parameters $\theta$, the posterior over the parameters $\theta$

$$p(\theta|X, m) = \frac{p(X, \theta|m)}{p(X|m)} = \frac{p(X|\theta, m)p(\theta|m)}{\int p(X|\theta, m)p(\theta|m)d\theta} = \frac{\text{Likelihood } \times \text{ Prior}}{\text{Marginal likelihood}}$$
Recap: Bayesian Inference

- Given data $X$ from a model $m$ with parameters $\theta$, the posterior over the parameters $\theta$

$$p(\theta|X, m) = \frac{p(X, \theta|m)}{p(X|m)} = \frac{p(X|\theta, m)p(\theta|m)}{\int p(X|\theta, m)p(\theta|m)d\theta} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}}$$

- Can use the posterior for various purposes
Recap: Bayesian Inference

Given data $\mathbf{X}$ from a model $m$ with parameters $\theta$, the posterior over the parameters $\theta$

$$p(\theta|\mathbf{X}, m) = \frac{p(\mathbf{X}, \theta|m)}{p(\mathbf{X}|m)} = \frac{p(\mathbf{X}|\theta, m)p(\theta|m)}{\int p(\mathbf{X}|\theta, m)p(\theta|m)d\theta} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}}$$

Can use the posterior for various purposes, e.g.,

- Getting point estimates e.g., mode (though, for this, directly doing point estimation is often easier)
Recap: Bayesian Inference

Given data $X$ from a model $m$ with parameters $\theta$, the posterior over the parameters $\theta$

$$p(\theta|X, m) = \frac{p(X, \theta|m)}{p(X|m)} = \frac{p(X|\theta, m)p(\theta|m)}{\int p(X|\theta, m)p(\theta|m)d\theta} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}}$$

Can use the posterior for various purposes, e.g.,
- Getting point estimates e.g., mode (though, for this, directly doing point estimation is often easier)
- Uncertainty in our estimates of $\theta$ (variance, credible intervals, etc)
Recap: Bayesian Inference

- Given data $X$ from a model $m$ with parameters $\theta$, the posterior over the parameters $\theta$

$$p(\theta|X, m) = \frac{p(X, \theta|m)}{p(X|m)} = \frac{p(X|\theta, m)p(\theta|m)}{\int p(X|\theta, m)p(\theta|m)d\theta} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}}$$

- Can use the posterior for various purposes, e.g.,
  - Getting point estimates e.g., mode (though, for this, directly doing point estimation is often easier)
  - Uncertainty in our estimates of $\theta$ (variance, credible intervals, etc)
  - Computing the posterior predictive distribution (PPD) for new data, e.g.,

$$p(x_*|X, m) = \int p(x_*|\theta, m)p(\theta|X, m)d\theta$$
Recap: Bayesian Inference

- Given data $X$ from a model $m$ with parameters $\theta$, the posterior over the parameters $\theta$

$$p(\theta|X, m) = \frac{p(X, \theta|m)}{p(X|m)} = \frac{p(X|\theta, m)p(\theta|m)}{\int p(X|\theta, m)p(\theta|m)d\theta} = \text{Likelihood} \times \text{Prior} \over \text{Marginal likelihood}$$

- Can use the posterior for various purposes, e.g.,
  - Getting point estimates e.g., mode (though, for this, directly doing point estimation is often easier)
  - Uncertainty in our estimates of $\theta$ (variance, credible intervals, etc)
  - Computing the posterior predictive distribution (PPD) for new data, e.g.,

$$p(x_*|X, m) = \int p(x_*|\theta, m)p(\theta|X, m)d\theta$$

- Caveat: Computing the posterior/PPD is in general hard (due to the intractable integrals involved)
Recap: Marginal Likelihood and Its Usefulness

- Likelihood vs Marginal Likelihood: $p(X|\theta, m)$ vs $p(X|m)$

Can use marginal likelihood $p(X|m)$ to select the best model from a finite set of models

$$\hat{m} = \arg \max_m p(m|X) = \arg \max_m p(X|m)p(m)$$

If $p(m)$ is uniform

Also useful for estimating hyperparameters of the assumed model (if we consider $m$ as the hyperparameters)

Suppose hyperparameters of likelihood are $\alpha_\ell$ and that of prior are $\alpha_p$ (so here $m = \{\alpha_\ell, \alpha_p\}$)

Assuming $p(\alpha_\ell, \alpha_p)$ is uniform, hyperparameters can be estimated via MLE-II (a.k.a. empirical Bayes)

$$\{\hat{\alpha}_\ell, \hat{\alpha}_p\} = \arg \max_{\alpha_\ell, \alpha_p} p(X|\alpha_\ell, \alpha_p)$$

Again, note that the integral here may be intractable and may need to be approximated

Can also compute $p(m|X)$ and do Bayesian Model Averaging:

$$p(x^*|X) = \sum_{M} p(x^*|X, m)p(m|X)$$
Recap: Marginal Likelihood and Its Usefulness

- Likelihood vs Marginal Likelihood: $p(X|\theta, m)$ vs $p(X|m)$
  - Prob. of $X$ for a single $\theta$ under model $m$ vs prob. of $X$ averaged over all $\theta$'s under model $m$

- Can use marginal likelihood $p(X|m)$ to select the best model from a finite set of models
  $$\hat{m} = \arg \max_m p(m|X) = \arg \max_m p(X|m)$$

- Also useful for estimating hyperparam of the assumed model (if we consider $m$ as the hyperparams)

  Suppose hyperparams of likelihood are $\alpha_\ell$ and that of prior are $\alpha_p$ (so here $m = \{\alpha_\ell, \alpha_p\}$)

  Assuming $p(\alpha_\ell, \alpha_p)$ is uniform, hyperparams can be estimated via MLE-II (a.k.a. empirical Bayes)
  $$\hat{\alpha}_\ell, \hat{\alpha}_p = \arg \max_{\alpha_\ell, \alpha_p} p(X|\alpha_\ell, \alpha_p) = \arg \max_{\alpha_\ell, \alpha_p} \int p(X|\theta, \alpha_\ell) p(\theta|\alpha_p) d\theta$$

  Again, note that the integral here may be intractable and may need to be approximated

- Can also compute $p(m|X)$ and do Bayesian Model Averaging:
  $$p(x^*|X) = \sum_{m=1}^M p(x^*|X, m) p(m|X)$$
Recap: Marginal Likelihood and Its Usefulness

- Likelihood vs Marginal Likelihood: \( p(X|\theta, m) \) vs \( p(X|m) \)
  - Prob. of \( X \) for a single \( \theta \) under model \( m \) vs prob. of \( X \) averaged over all \( \theta \)'s under model \( m \)
- Can use marginal likelihood \( p(X|m) \) to select the best model from a finite set of models

\[
\hat{m} = \arg \max_m p(m|X)
\]
Recap: Marginal Likelihood and Its Usefulness

- Likelihood vs Marginal Likelihood: $p(X|\theta, m)$ vs $p(X|m)$
  - Prob. of $X$ for a single $\theta$ under model $m$ vs prob. of $X$ averaged over all $\theta$’s under model $m$
- Can use marginal likelihood $p(X|m)$ to select the best model from a finite set of models

$$\hat{m} = \arg \max_m p(m|X) = \arg \max_m p(X|m)p(m)$$
Recap: Marginal Likelihood and Its Usefulness

- Likelihood vs Marginal Likelihood: \( p(X|\theta, m) \ vs \ p(X|m) \)
  - Prob. of \( X \) for a single \( \theta \) under model \( m \) vs prob. of \( X \) averaged over all \( \theta \)'s under model \( m \)
- Can use marginal likelihood \( p(X|m) \) to select the best model from a finite set of models

\[
\hat{m} = \arg \max_m p(m|X) = \arg \max_m p(X|m)p(m) = \arg \max_m p(X|m), \text{ if } p(m) \text{ is uniform}
\]
Recap: Marginal Likelihood and Its Usefulness

- Likelihood vs Marginal Likelihood: \( p(X|\theta, m) \) vs \( p(X|m) \)
  - Prob. of \( X \) for a single \( \theta \) under model \( m \) vs prob. of \( X \) averaged over all \( \theta \)'s under model \( m \)
- Can use marginal likelihood \( p(X|m) \) to select the best model from a finite set of models
  \[
  \hat{m} = \arg \max_m p(m|X) = \arg \max_m p(X|m)p(m) = \arg \max_m p(X|m), \text{ if } p(m) \text{ is uniform}
  \]
- Also useful for estimating hyperparam of the assumed model (if we consider \( m \) as the hyperparams)
Recap: Marginal Likelihood and Its Usefulness

- Likelihood vs Marginal Likelihood: $p(X|\theta, m)$ vs $p(X|m)$
  - Prob. of $X$ for a single $\theta$ under model $m$ vs prob. of $X$ averaged over all $\theta$’s under model $m$
- Can use marginal likelihood $p(X|m)$ to select the best model from a finite set of models
  \[
  \hat{m} = \arg \max_m p(m|X) = \arg \max_m p(X|m)p(m) = \arg \max_m p(X|m), \text{if } p(m) \text{ is uniform}
  \]
- Also useful for estimating hyperparam of the assumed model (if we consider $m$ as the hyperparams)
  - Suppose hyperparams of likelihood are $\alpha_\ell$ and that of prior are $\alpha_p$ (so here $m = \{\alpha_\ell, \alpha_p\}$)
Recap: Marginal Likelihood and Its Usefulness

- Likelihood vs Marginal Likelihood: \( p(X|\theta, m) \) vs \( p(X|m) \)
  - Prob. of \( X \) for a single \( \theta \) under model \( m \) vs prob. of \( X \) averaged over all \( \theta \)'s under model \( m \)
- Can use marginal likelihood \( p(X|m) \) to select the best model from a finite set of models
  \[
  \hat{m} = \arg \max_m p(m|X) = \arg \max_m p(X|m)p(m) = \arg \max_m p(X|m), \text{if } p(m) \text{ is uniform}
  \]
- Also useful for estimating hyperparam of the assumed model (if we consider \( m \) as the hyperparams)
  - Suppose hyperparams of likelihood are \( \alpha_\ell \) and that of prior are \( \alpha_p \) (so here \( m = \{\alpha_\ell, \alpha_p\} \))
  - Assuming \( p(\alpha_\ell, \alpha_p) \) is uniform, hyperparams can be estimated via MLE-II (a.k.a. empirical Bayes)

\[\text{Prob. Mod. & Inference - CS698X (Piyush Rai, IITK)}\]

Bayesian Inference for Some Basic Models
Recap: Marginal Likelihood and Its Usefulness

- Likelihood vs Marginal Likelihood: \( p(\mathbf{X}|\theta, m) \) vs \( p(\mathbf{X}|m) \)
  - Prob. of \( \mathbf{X} \) for a single \( \theta \) under model \( m \) vs prob. of \( \mathbf{X} \) averaged over all \( \theta \)'s under model \( m \)
- Can use marginal likelihood \( p(\mathbf{X}|m) \) to select the best model from a finite set of models
  \[
  \hat{m} = \arg \max_m p(m|\mathbf{X}) = \arg \max_m p(\mathbf{X}|m)p(m) = \arg \max_m p(\mathbf{X}|m), \text{if } p(m) \text{ is uniform}
  \]
- Also useful for estimating hyperparam of the assumed model (if we consider \( m \) as the hyperparams)
  - Suppose hyperparams of likelihood are \( \alpha_\ell \) and that of prior are \( \alpha_p \) (so here \( m = \{\alpha_\ell, \alpha_p\} \))
  - Assuming \( p(\alpha_\ell, \alpha_p) \) is uniform, hyperparams can be estimated via MLE-II (a.k.a. empirical Bayes)
    \[
    \{\hat{\alpha}_\ell, \hat{\alpha}_p\} = \arg \max_{\alpha_\ell, \alpha_p} p(\mathbf{X}|\alpha_\ell, \alpha_p)
    \]
Recap: Marginal Likelihood and Its Usefulness

- Likelihood vs Marginal Likelihood: $p(X|\theta, m) \ vs \ p(X|m)$
  - Prob. of $X$ for a single $\theta$ under model $m$ vs prob. of $X$ averaged over all $\theta$'s under model $m$

- Can use marginal likelihood $p(X|m)$ to select the best model from a finite set of models
  \[
  \hat{m} = \arg\max_m p(m|X) = \arg\max_m p(X|m)p(m) = \arg\max_m p(X|m), \text{if } p(m) \text{ is uniform}
  \]

- Also useful for estimating hyperparam of the assumed model (if we consider $m$ as the hyperparams)
  - Suppose hyperparams of likelihood are $\alpha_\ell$ and that of prior are $\alpha_p$ (so here $m = \{\alpha_\ell, \alpha_p\}$)
  - Assuming $p(\alpha_\ell, \alpha_p)$ is uniform, hyperparams can be estimated via MLE-II (a.k.a. empirical Bayes)
    \[
    \{\hat{\alpha_\ell}, \hat{\alpha_p}\} = \arg\max_{\alpha_\ell, \alpha_p} p(X|\alpha_\ell, \alpha_p) = \arg\max_{\alpha_\ell, \alpha_p} \int p(X|\theta, \alpha_\ell)p(\theta|\alpha_p)d\theta
    \]
Recap: Marginal Likelihood and Its Usefulness

- Likelihood vs Marginal Likelihood: $p(X|\theta, m) \text{ vs } p(X|m)$
  - Prob. of $X$ for a single $\theta$ under model $m$ vs prob. of $X$ averaged over all $\theta$’s under model $m$
- Can use marginal likelihood $p(X|m)$ to select the best model from a finite set of models
  $$\hat{m} = \arg \max_m p(m|X) = \arg \max_m p(X|m)p(m) = \arg \max_m p(X|m), \text{ if } p(m) \text{ is uniform}$$
- Also useful for estimating hyperparam of the assumed model (if we consider $m$ as the hyperparams)
  - Suppose hyperparams of likelihood are $\alpha_\ell$ and that of prior are $\alpha_p$ (so here $m = \{\alpha_\ell, \alpha_p\}$)
  - Assuming $p(\alpha_\ell, \alpha_p)$ is uniform, hyperparams can be estimated via MLE-II (a.k.a. empirical Bayes)
    $$\{\hat{\alpha}_\ell, \hat{\alpha}_p\} = \arg \max_{\alpha_\ell, \alpha_p} p(X|\alpha_\ell, \alpha_p) = \arg \max_{\alpha_\ell, \alpha_p} \int p(X|\theta, \alpha_\ell)p(\theta|\alpha_p)d\theta$$
  - Again, note that the integral here may be intractable and may need to be approximated
Recap: Marginal Likelihood and Its Usefulness

- Likelihood vs Marginal Likelihood: $p(X|\theta, m)$ vs $p(X|m)$
  - Prob. of $X$ for a single $\theta$ under model $m$ vs prob. of $X$ averaged over all $\theta$'s under model $m$
- Can use marginal likelihood $p(X|m)$ to select the best model from a finite set of models
  \[
  \hat{m} = \arg \max_m p(m|X) = \arg \max_m p(X|m)p(m) = \arg \max_m p(X|m), \text{if } p(m) \text{ is uniform}
  \]
- Also useful for estimating hyperparam of the assumed model (if we consider $m$ as the hyperparams)
  - Suppose hyperparams of likelihood are $\alpha_\ell$ and that of prior are $\alpha_p$ (so here $m = \{\alpha_\ell, \alpha_p\}$)
  - Assuming $p(\alpha_\ell, \alpha_p)$ is uniform, hyperparams can be estimated via MLE-II (a.k.a. empirical Bayes)
    \[
    \{\hat{\alpha}_\ell, \hat{\alpha}_p\} = \arg \max_{\alpha_\ell, \alpha_p} p(X|\alpha_\ell, \alpha_p) = \arg \max_{\alpha_\ell, \alpha_p} \int p(X|\theta, \alpha_\ell)p(\theta|\alpha_p)d\theta
    \]
    - Again, note that the integral here may be intractable and may need to be approximated
- Can also compute $p(m|X)$ and do Bayesian Model Averaging: $p(x_*|X) = \sum_{m=1}^M p(x_*|X, m)p(m|X)$
Recap: Bayesian Inference for a Beta-Bernoulli Model

- Saw the example of estimating the bias \( \theta \in (0, 1) \) of a coin using Bayesian inference
- Chose a Bernoulli likelihood for each coin toss and a conjugate Beta prior for \( \theta \)

\[
p(x_n|\theta) = \text{Bernoulli}(x_n|\theta) = \theta^{x_n}(1 - \theta)^{1-x_n}
\]
\[
p(\theta|\alpha, \beta) = \text{Beta}(\theta|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha - 1}(1 - \theta)^{\beta - 1}
\]
Recap: Bayesian Inference for a Beta-Bernoulli Model

- Saw the example of estimating the bias $\theta \in (0, 1)$ of a coin using Bayesian inference
- Chose a Bernoulli likelihood for each coin toss and a **conjugate** Beta prior for $\theta$

\[
p(x_n|\theta) = \text{Bernoulli}(x_n|\theta) = \theta^{x_n}(1 - \theta)^{1-x_n}
\]
\[
p(\theta|\alpha, \beta) = \text{Beta}(\theta|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1 - \theta)^{\beta-1}
\]

- Here, prior’s hyperparams (assumed fixed here) control its shape; also act as pseudo-observations
Recap: Bayesian Inference for a Beta-Bernoulli Model

• Saw the example of estimating the bias $\theta \in (0, 1)$ of a coin using Bayesian inference
• Chose a Bernoulli likelihood for each coin toss and a conjugate Beta prior for $\theta$

\[
p(x_n|\theta) = \text{Bernoulli}(x_n|\theta) = \theta^{x_n}(1 - \theta)^{1-x_n}
\]
\[
p(\theta|\alpha, \beta) = \text{Beta}(\theta|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1 - \theta)^{\beta-1}
\]

• Here, prior’s hyperparams (assumed fixed here) control its shape; also act as pseudo-observations
• Assuming $x_n$’s as i.i.d. given $\theta$, posterior $p(\theta|X, \alpha, \beta) \propto p(X|\theta)p(\theta|\alpha, \beta)$ turned out to be Beta
Recap: Bayesian Inference for a Beta-Bernoulli Model

- Saw the example of estimating the bias \( \theta \in (0, 1) \) of a coin using Bayesian inference
- Chose a Bernoulli likelihood for each coin toss and a conjugate Beta prior for \( \theta \)

\[
p(x_n|\theta) = \text{Bernoulli}(x_n|\theta) = \theta^{x_n}(1-\theta)^{1-x_n}
\]

\[
p(\theta|\alpha, \beta) = \text{Beta}(\theta|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}
\]

- Here, prior’s hyperparams (assumed fixed here) control its shape; also act as pseudo-observations
- Assuming \( x_n \)'s as i.i.d. given \( \theta \), posterior \( p(\theta|X, \alpha, \beta) \propto p(X|\theta)p(\theta|\alpha, \beta) \) turned out to be Beta

\[
p(\theta|X, \alpha, \beta) = \text{Beta}(\theta|\alpha + \sum_{n=1}^{N} x_n, \beta + N - \sum_{n=1}^{N} x_n)
\]
Recap: Bayesian Inference for a Beta-Bernoulli Model

- Saw the example of estimating the bias $\theta \in (0, 1)$ of a coin using Bayesian inference
- Chose a Bernoulli likelihood for each coin toss and a conjugate Beta prior for $\theta$

\[
p(x_n|\theta) = \text{Bernoulli}(x_n|\theta) = \theta^{x_n}(1 - \theta)^{1-x_n}
\]
\[
p(\theta|\alpha, \beta) = \text{Beta}(\theta|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1 - \theta)^{\beta-1}
\]

- Here, prior’s hyperparams (assumed fixed here) control its shape; also act as pseudo-observations
- Assuming $x_n$’s as i.i.d. given $\theta$, posterior $p(\theta|X, \alpha, \beta) \propto p(X|\theta)p(\theta|\alpha, \beta)$ turned out to be Beta

\[
p(\theta|X, \alpha, \beta) = \text{Beta}(\theta|\alpha + \sum_{n=1}^{N} x_n, \beta + N - \sum_{n=1}^{N} x_n) = \text{Beta}(\theta|\alpha + N_1, \beta + N_0)
\]
Recap: Bayesian Inference for a Beta-Bernoulli Model

- Saw the example of estimating the bias $\theta \in (0, 1)$ of a coin using Bayesian inference
- Chose a Bernoulli likelihood for each coin toss and a conjugate Beta prior for $\theta$
  
  $$p(x_n|\theta) = \text{Bernoulli}(x_n|\theta) = \theta^{x_n}(1 - \theta)^{1-x_n}$$
  $$p(\theta|\alpha, \beta) = \text{Beta}(\theta|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1 - \theta)^{\beta-1}$$

- Here, prior’s hyperparams (assumed fixed here) control its shape; also act as pseudo-observations
- Assuming $x_n$’s as i.i.d. given $\theta$, posterior $p(\theta|\mathbf{X}, \alpha, \beta) \propto p(\mathbf{X}|\theta)p(\theta|\alpha, \beta)$ turned out to be Beta
  
  $$p(\theta|\mathbf{X}, \alpha, \beta) = \text{Beta}(\theta|\alpha + \sum_{n=1}^{N} x_n, \beta + N - \sum_{n=1}^{N} x_n) = \text{Beta}(\theta|\alpha + N_1, \beta + N_0)$$

- Note: Here posterior only depends on data $\mathbf{X} = \{x_1, \ldots, x_N\}$ via sufficient statistics $N_1$ and $N_0$
  
  $$p(\theta|\mathbf{X}, \alpha, \beta) = p(\theta|s(\mathbf{X}))$$
Recap: Bayesian Inference for a Beta-Bernoulli Model

- Saw the example of estimating the bias $\theta \in (0, 1)$ of a coin using Bayesian inference
- Chose a Bernoulli likelihood for each coin toss and a conjugate Beta prior for $\theta$

$$
p(x_n|\theta) = \text{Bernoulli}(x_n|\theta) = \theta^{x_n}(1 - \theta)^{1-x_n}
$$

$$
p(\theta|\alpha, \beta) = \text{Beta}(\theta|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1 - \theta)^{\beta-1}
$$

- Here, prior’s hyperparams (assumed fixed here) control its shape; also act as pseudo-observations
- Assuming $x_n$’s as i.i.d. given $\theta$, posterior $p(\theta|X, \alpha, \beta) \propto p(X|\theta)p(\theta|\alpha, \beta)$ turned out to be Beta

$$
p(\theta|X, \alpha, \beta) = \text{Beta}(\theta|\alpha + \sum_{n=1}^{N} x_n, \beta + N - \sum_{n=1}^{N} x_n) = \text{Beta}(\theta|\alpha + N_1, \beta + N_0)
$$

- Note: Here posterior only depends on data $X = \{x_1, \ldots, x_N\}$ via sufficient statistics $N_1$ and $N_0$

$$
p(\theta|X, \alpha, \beta) = p(\theta|s(X))
$$

- We will see many other cases where the posterior depends on data only via some sufficient statistics
Recap: Making Predictions in the Beta-Bernoulli Model

- The posterior predictive distribution (averaging over all $\theta$ weighted by their posterior probabilities):

$$p(x_{N+1} = 1|X, \alpha, \beta) = \int_0^1 p(x_{N+1} = 1|\theta)p(\theta|X, \alpha, \beta)d\theta$$
Recap: Making Predictions in the Beta-Bernoulli Model

- The posterior predictive distribution (averaging over all $\theta$ weighted by their posterior probabilities):

$$p(x_{N+1} = 1|X, \alpha, \beta) = \int_0^1 p(x_{N+1} = 1|\theta)p(\theta|X, \alpha, \beta)d\theta$$

$$= \int_0^1 \theta \times \text{Beta}(\theta|\alpha + N_1, \beta + N_0)d\theta$$
Recap: Making Predictions in the Beta-Bernoulli Model

- The posterior predictive distribution (averaging over all $\theta$ weighted by their posterior probabilities):

$$
p(x_{N+1} = 1|X, \alpha, \beta) = \int_0^1 p(x_{N+1} = 1|\theta) p(\theta|X, \alpha, \beta) d\theta
$$

$$
= \int_0^1 \theta \times \text{Beta}(\theta|\alpha + N_1, \beta + N_0) d\theta
$$

$$
= \mathbb{E}[\theta|X]
$$
Recap: Making Predictions in the Beta-Bernoulli Model

- The **posterior predictive distribution** (averaging over all $\theta$ weighted by their posterior probabilities):

$$p(x_{N+1} = 1|\mathbf{X}, \alpha, \beta) = \int_0^1 p(x_{N+1} = 1|\theta)p(\theta|\mathbf{X}, \alpha, \beta)d\theta$$

$$= \int_0^1 \theta \times \text{Beta}(\theta|\alpha + N_1, \beta + N_0)d\theta$$

$$= \mathbb{E}[\theta|\mathbf{X}]$$

$$= \frac{\alpha + N_1}{\alpha + \beta + N}$$
Recap: Making Predictions in the Beta-Bernoulli Model

- The posterior predictive distribution (averaging over all $\theta$ weighted by their posterior probabilities):

$$p(x_{N+1} = 1|X, \alpha, \beta) = \int_0^1 p(x_{N+1} = 1|\theta) p(\theta|X, \alpha, \beta) d\theta$$

$$= \int_0^1 \theta \times \text{Beta}(\theta|\alpha + N_1, \beta + N_0) d\theta$$

$$= \mathbb{E}[\theta|X]$$

$$= \frac{\alpha + N_1}{\alpha + \beta + N}$$

- Therefore the posterior predictive distribution: $p(x_{N+1}|X) = \text{Bernoulli}(x_{N+1} | \mathbb{E}[\theta|X])$
Recap: Making Predictions in the Beta-Bernoulli Model

- The posterior predictive distribution (averaging over all $\theta$ weighted by their posterior probabilities):

$$p(x_{N+1} = 1|\mathbf{X}, \alpha, \beta) = \int_0^1 p(x_{N+1} = 1|\theta)p(\theta|\mathbf{X}, \alpha, \beta)d\theta$$

$$= \int_0^1 \theta \times \text{Beta}(\theta|\alpha + N_1, \beta + N_0)d\theta$$

$$= \mathbb{E}[\theta|\mathbf{X}]$$

$$= \frac{\alpha + N_1}{\alpha + \beta + N}$$

Therefore the posterior predictive distribution: $p(x_{N+1}|\mathbf{X}) = \text{Bernoulli}(x_{N+1} | \mathbb{E}[\theta|\mathbf{X}])$

- In contrast, the plug-in predictive distribution using a point estimate $\hat{\theta}$ (e.g., using MLE/MAP)

$$p(x_{N+1} = 1|\mathbf{X}, \alpha, \beta) \approx p(x_{N+1} = 1|\hat{\theta}) = \hat{\theta}$$
Recap: Making Predictions in the Beta-Bernoulli Model

- The **posterior predictive distribution** (averaging over all $\theta$ weighted by their posterior probabilities):

  \[
p(x_{N+1} = 1|X, \alpha, \beta) = \int_0^1 p(x_{N+1} = 1|\theta)p(\theta|X, \alpha, \beta)d\theta
  \]
  \[
  = \int_0^1 \theta \times \text{Beta}(\theta|\alpha + N_1, \beta + N_0)d\theta
  \]
  \[
  = \mathbb{E}[\theta|X]
  \]
  \[
  = \frac{\alpha + N_1}{\alpha + \beta + N}
  \]

- Therefore the posterior predictive distribution: $p(x_{N+1}|X) = \text{Bernoulli}(x_{N+1} | \mathbb{E}[\theta|X])$

- In contrast, the **plug-in predictive** distribution using a point estimate $\hat{\theta}$ (e.g., using MLE/MAP)

  \[
p(x_{N+1} = 1|X, \alpha, \beta) \approx p(x_{N+1} = 1|\hat{\theta}) = \hat{\theta}
  \]
  \[
  \text{or equivalently} \quad p(x_{N+1}|X) \approx \text{Bernoulli}(x_{N+1} | \hat{\theta})
  \]
More Examples..
Bayesian Inference for Multinoulli/Multinomial

Assume $N$ discrete-valued observations $\{x_1, \ldots, x_N\}$ with each $x_n \in \{1, \ldots, K\}$, e.g., $x_n$ represents the outcome of a dice roll with $K$ faces, $x_n$ represents the class label of the $n$-th example (total $K$ classes), $x_n$ represents the identity of the $n$-th word in a sequence of words.

Assume likelihood to be multinoulli with unknown parameters $\pi = [\pi_1, \ldots, \pi_K]$ such that $\sum_{k=1}^{K} \pi_k = 1$.

$$p(x_n | \pi) = \text{multinoulli}(x_n | \pi) = K \prod_{k=1}^{K} \pi_k \mathbb{1}_{[x_n = k]}$$

$\pi$ is a vector of probabilities ("probability vector"), e.g., biases of the $K$ sides of the dice, prior class probabilities in multi-class classification, probabilities of observing each word in the vocabulary.

Assume a conjugate Dirichlet prior on $\pi$ with hyperparameters $\alpha = [\alpha_1, \ldots, \alpha_K]$ (also, $\alpha_k \geq 0, \forall k$).

$$p(\pi | \alpha) = \text{Dirichlet}(\pi | \alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} K \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

Probability mass function (PMF) of a multinomial distribution.
Bayesian Inference for Multinoulli/Multinomial

- Assume $N$ discrete-valued observations $\{x_1, \ldots, x_N\}$ with each $x_n \in \{1, \ldots, K\}$, e.g.,
  - $x_n$ represents the outcome of a dice roll with $K$ faces
  - $x_n$ represents the class label of the $n$-th example (total $K$ classes)
  - $x_n$ represents the identity of the $n$-th word in a sequence of words
Bayesian Inference for Multinoulli/Multinomial

- Assume $N$ discrete-valued observations $\{x_1, \ldots, x_N\}$ with each $x_n \in \{1, \ldots, K\}$, e.g.,
  - $x_n$ represents the outcome of a dice roll with $K$ faces
  - $x_n$ represents the class label of the $n$-th example (total $K$ classes)
  - $x_n$ represents the identity of the $n$-th word in a sequence of words
- Assume likelihood to be multinoulli with unknown params $\pi = [\pi_1, \ldots, \pi_K]$ s.t. $\sum_{k=1}^{K} \pi_k = 1$
Assume $N$ discrete-valued observations $\{x_1, \ldots, x_N\}$ with each $x_n \in \{1, \ldots, K\}$, e.g.,

- $x_n$ represents the outcome of a dice roll with $K$ faces
- $x_n$ represents the class label of the $n$-th example (total $K$ classes)
- $x_n$ represents the identity of the $n$-th word in a sequence of words

Assume likelihood to be multinoulli with unknown params $\pi = [\pi_1, \ldots, \pi_K]$ s.t. $\sum_{k=1}^{K} \pi_k = 1$

$$p(x_n|\pi) = \text{multinoulli}(x_n|\pi)$$
Bayesian Inference for Multinoulli/Multinomial

- Assume $N$ discrete-valued observations $\{x_1, \ldots, x_N\}$ with each $x_n \in \{1, \ldots, K\}$, e.g.,
  - $x_n$ represents the outcome of a dice roll with $K$ faces
  - $x_n$ represents the class label of the $n$-th example (total $K$ classes)
  - $x_n$ represents the identity of the $n$-th word in a sequence of words

- Assume likelihood to be multinoulli with unknown params $\pi = [\pi_1, \ldots, \pi_K]$ s.t. $\sum_{k=1}^{K} \pi_k = 1$
  
  $$p(x_n|\pi) = \text{multinoulli}(x_n|\pi) = \prod_{k=1}^{K} \pi_k I[x_n=k]$$

$\pi$ is a vector of probabilities ("probability vector"), e.g.,

- Biases of the $K$ sides of the dice
- Prior class probabilities in multi-class classification
- Probabilities of observing each words in the vocabulary

- Assume a conjugate Dirichlet prior on $\pi$ with hyperparams $\alpha = [\alpha_1, \ldots, \alpha_K]$ (also, $\alpha_k \geq 0$, $\forall k$)
  
  $$p(\pi|\alpha) = \text{Dirichlet}(\pi|\alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k-1}$$

$\alpha_k$ is the bias of the $k$-th side of the dice.
Assume $N$ discrete-valued observations $\{x_1, \ldots, x_N\}$ with each $x_n \in \{1, \ldots, K\}$, e.g.,

- $x_n$ represents the outcome of a dice roll with $K$ faces
- $x_n$ represents the class label of the $n$-th example (total $K$ classes)
- $x_n$ represents the identity of the $n$-th word in a sequence of words

Assume likelihood to be multinoulli with unknown params $\pi = [\pi_1, \ldots, \pi_K]$ s.t. $\sum_{k=1}^{K} \pi_k = 1$

$$p(x_n | \pi) = \text{multinoulli}(x_n | \pi) = \prod_{k=1}^{K} \pi_k [x_n = k]$$

- $\pi$ is a vector of probabilities ("probability vector")
Bayesian Inference for Multinoulli/Multinomial

- Assume $N$ discrete-valued observations $\{x_1, \ldots, x_N\}$ with each $x_n \in \{1, \ldots, K\}$, e.g.,
  - $x_n$ represents the outcome of a dice roll with $K$ faces
  - $x_n$ represents the class label of the $n$-th example (total $K$ classes)
  - $x_n$ represents the identity of the $n$-th word in a sequence of words

- Assume likelihood to be multinoulli with unknown params $\pi = [\pi_1, \ldots, \pi_K]$ s.t. $\sum_{k=1}^{K} \pi_k = 1$
  $$p(x_n|\pi) = \text{multinoulli}(x_n|\pi) = \prod_{k=1}^{K} \pi_k^{I[x_n=k]}$$

- $\pi$ is a vector of probabilities ("probability vector"), e.g.,
  - Biases of the $K$ sides of the dice
  - Prior class probabilities in multi-class classification
  - Probabilities of observing each words in the vocabulary
Assume $N$ discrete-valued observations $\{x_1, \ldots, x_N\}$ with each $x_n \in \{1, \ldots, K\}$, e.g.,

- $x_n$ represents the outcome of a dice roll with $K$ faces
- $x_n$ represents the class label of the $n$-th example (total $K$ classes)
- $x_n$ represents the identity of the $n$-th word in a sequence of words

Assume likelihood to be multinoulli with unknown params $\pi = [\pi_1, \ldots, \pi_K]$ s.t. $\sum_{k=1}^{K} \pi_k = 1$

$$p(x_n|\pi) = \text{multinoulli}(x_n|\pi) = \prod_{k=1}^{K} \pi_k^{I[x_n=k]}$$

$\pi$ is a vector of probabilities ("probability vector"), e.g.,

- Biases of the $K$ sides of the dice
- Prior class probabilities in multi-class classification
- Probabilities of observing each words in the vocabulary

Assume a conjugate Dirichlet prior on $\pi$ with hyperparams $\alpha = [\alpha_1, \ldots, \alpha_K]$ (also, $\alpha_k \geq 0$, $\forall k$)

$$p(\pi|\alpha) = \text{Dirichlet}(\pi|\alpha_1, \ldots, \alpha_K)$$
Bayesian Inference for Multinoulli/Multinomial

- Assume $N$ discrete-valued observations $\{x_1, \ldots, x_N\}$ with each $x_n \in \{1, \ldots, K\}$, e.g.,
  - $x_n$ represents the outcome of a dice roll with $K$ faces
  - $x_n$ represents the class label of the $n$-th example (total $K$ classes)
  - $x_n$ represents the identity of the $n$-th word in a sequence of words

- Assume likelihood to be multinoulli with unknown params $\pi = [\pi_1, \ldots, \pi_K]$ s.t. $\sum_{k=1}^{K} \pi_k = 1$
  $$p(x_n|\pi) = \text{multinoulli}(x_n|\pi) = \prod_{k=1}^{K} \pi_k^{I[x_n=k]}$$

- $\pi$ is a vector of probabilities ("probability vector"), e.g.,
  - Biases of the $K$ sides of the dice
  - Prior class probabilities in multi-class classification
  - Probabilities of observing each words in the vocabulary

- Assume a conjugate Dirichlet prior on $\pi$ with hyperparams $\alpha = [\alpha_1, \ldots, \alpha_K]$ (also, $\alpha_k \geq 0, \forall k$)
  $$p(\pi|\alpha) = \text{Dirichlet}(\pi|\alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k-1}$$

Prob. Mod. & Inference - CS698X (Piyush Rai, IITK)
Bayesian Inference for Multinoulli/Multinomial

- Assume \( N \) discrete-valued observations \( \{x_1, \ldots, x_N\} \) with each \( x_n \in \{1, \ldots, K\} \), e.g.,
  - \( x_n \) represents the outcome of a dice roll with \( K \) faces
  - \( x_n \) represents the class label of the \( n \)-th example (total \( K \) classes)
  - \( x_n \) represents the identity of the \( n \)-th word in a sequence of words

- Assume likelihood to be multinoulli with unknown params \( \pi = [\pi_1, \ldots, \pi_K] \) s.t. \( \sum_{k=1}^{K} \pi_k = 1 \)
  \[
p(x_n | \pi) = \text{multinoulli}(x_n | \pi) = \prod_{k=1}^{K} \pi_k^{[x_n = k]}\]

- \( \pi \) is a vector of probabilities ("probability vector"), e.g.,
  - Biases of the \( K \) sides of the dice
  - Prior class probabilities in multi-class classification
  - Probabilities of observing each words in the vocabulary

- Assume a conjugate Dirichlet prior on \( \pi \) with hyperparams \( \alpha = [\alpha_1, \ldots, \alpha_K] \) (also, \( \alpha_k \geq 0, \forall k \))
  \[
p(\pi | \alpha) = \text{Dirichlet}(\pi | \alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k-1} = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \pi_k^{\alpha_k-1}\]
Brief Detour: Dirichlet Distribution

• Very important distribution: Models non-neg. vectors $\pi$ that sum to one (e.g., probability vectors)
Brief Detour: Dirichlet Distribution

- Very important distribution: Models non-neg. vectors $\pi$ that sum to one (e.g., probability vectors)
- A random draw from Dirichlet will be a point under the probability simplex

Hyperparams $\alpha = [\alpha_1, \ldots, \alpha_K]$ control the shape of Dirichlet (akin to Beta’s hyperparams)
Brief Detour: Dirichlet Distribution

- Very important distribution: Models non-neg. vectors $\pi$ that sum to one (e.g., probability vectors)
- A random draw from Dirichlet will be a point under the probability simplex

$$p(\pi|\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

Hyperparams $\alpha = [\alpha_1, \ldots, \alpha_K]$ control the shape of Dirichlet (akin to Beta’s hyperparams)
- Can also be thought of as a multi-dimensional Beta distribution
Brief Detour: Dirichlet Distribution

- Very important distribution: Models non-neg. vectors $\pi$ that sum to one (e.g., probability vectors)
- A random draw from Dirichlet will be a point under the probability simplex

$$p(\pi | \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

- Hyperparams $\alpha = [\alpha_1, \ldots, \alpha_K]$ control the shape of Dirichlet (akin to Beta’s hyperparams)
- Can also be thought of as a multi-dimensional Beta distribution
- Note: Can also be seen as normalized version of $K$ independent gamma random variables
The posterior over $\pi$ is easy to compute in this case due to conjugacy b/w multinoulli and Dirichlet

$$p(\pi|X, \alpha) = \frac{p(X|\pi, \alpha)p(\pi|\alpha)}{p(X|\alpha)} = \frac{p(X|\pi)p(\pi|\alpha)}{p(X|\alpha)}$$
Bayesian Inference for Multinoulli/Multinomial

- The posterior over \( \pi \) is easy to compute in this case due to conjugacy b/w multinoulli and Dirichlet

\[
p(\pi|X, \alpha) = \frac{p(X|\pi, \alpha)p(\pi|\alpha)}{p(X|\alpha)} = \frac{p(X|\pi)p(\pi|\alpha)}{p(X|\alpha)}
\]

- Assuming \( x_n \)'s are i.i.d. given \( \pi \), \( p(X|\pi) = \prod_{n=1}^{N} p(x_n|\pi) \), therefore

\[
p(\pi|X, \alpha) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{I[x_n=k]} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}
\]

Note: \( N_1, \ldots, N_K \) are the sufficient statistics in this case

Note: If we want, we can also get the MAP estimate of \( \pi \) (mode of the above Dirichlet)

MAP estimation via standard way will require solving a constraint opt. problem (via Lagrangian)
Bayesian Inference for Multinoulli/Multinomial

- The posterior over $\pi$ is easy to compute in this case due to conjugacy b/w multinoulli and Dirichlet

$$p(\pi|X, \alpha) = \frac{p(X|\pi, \alpha)p(\pi|\alpha)}{p(X|\alpha)} = \frac{p(X|\pi)p(\pi|\alpha)}{p(X|\alpha)}$$

- Assuming $x_n$'s are i.i.d. given $\pi$, $p(X|\pi) = \prod_{n=1}^{N} p(x_n|\pi)$, therefore

$$p(\pi|X, \alpha) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{I[x_n=k]} \prod_{k=1}^{K} \pi_k^{\alpha_k-1} = \prod_{k=1}^{K} \pi_k^{\alpha_k + \sum_{n=1}^{N} I[x_n=k]-1}$$
Bayesian Inference for Multinoulli/Multinomial

- The posterior over $\pi$ is easy to compute in this case due to conjugacy b/w multinoulli and Dirichlet

$$p(\pi|X, \alpha) = \frac{p(X|\pi, \alpha)p(\pi|\alpha)}{p(X|\alpha)} = \frac{p(X|\pi)p(\pi|\alpha)}{p(X|\alpha)}$$

- Assuming $x_n$'s are i.i.d. given $\pi$, $p(X|\pi) = \prod_{n=1}^{N} p(x_n|\pi)$, therefore

$$p(\pi|X, \alpha) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_{k}^{[x_n=k]} \prod_{k=1}^{K} \pi_{k}^{\alpha_{k}-1} = \prod_{k=1}^{K} \pi_{k}^{\alpha_{k}+\sum_{n=1}^{N} [x_n=k]-1}$$

- Even without computing the normalization constant $p(X|\alpha)$, we can see that it's a Dirichlet! :-)
The posterior over $\pi$ is easy to compute in this case due to conjugacy b/w multinoulli and Dirichlet

$$p(\pi|X, \alpha) = \frac{p(X|\pi, \alpha)p(\pi|\alpha)}{p(X|\alpha)} = \frac{p(X|\pi)p(\pi|\alpha)}{p(X|\alpha)}$$

Assuming $x_n$'s are i.i.d. given $\pi$, $p(X|\pi) = \prod_{n=1}^{N} p(x_n|\pi)$, therefore

$$p(\pi|X, \alpha) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{I[x_n=k]} \prod_{k=1}^{K} \pi_k^{\alpha_k-1} = \prod_{k=1}^{K} \pi_k^{\alpha_k + \sum_{n=1}^{N} I[x_n=k] - 1}$$

Even without computing the normalization constant $p(X|\alpha)$, we can see that it's a Dirichlet! :-)

Denoting $N_k = \sum_{n=1}^{N} I[x_n = k]$, i.e., number of observations with value $k$, the posterior will be

$$p(\pi|X, \alpha) = \text{Dirichlet}(\pi|\alpha_1 + N_1, \ldots, \alpha_K + N_K)$$
Bayesian Inference for Multinoulli/Multinomial

- The posterior over $\pi$ is easy to compute in this case due to conjugacy b/w multinoulli and Dirichlet

$$p(\pi|X, \alpha) = \frac{p(X|\pi, \alpha)p(\pi|\alpha)}{p(X|\alpha)} = \frac{p(X|\pi)p(\pi|\alpha)}{p(X|\alpha)}$$

- Assuming $x_n$’s are i.i.d. given $\pi$, $p(X|\pi) = \prod_{n=1}^{N} p(x_n|\pi)$, therefore

$$p(\pi|X, \alpha) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_{k}^{[x_n=k]} \prod_{k=1}^{K} \pi_{k}^{\alpha_k-1} = \prod_{k=1}^{K} \pi_{k}^{\alpha_k+\sum_{n=1}^{N} [x_n=k]-1}$$

- Even without computing the normalization constant $p(X|\alpha)$, we can see that it’s a Dirichlet! :-)

- Denoting $N_k = \sum_{n=1}^{N} [x_n = k]$, i.e., number of observations with value $k$, the posterior will be

$$p(\pi|X, \alpha) = \text{Dirichlet}(\pi|\alpha_1 + N_1, \ldots, \alpha_K + N_K)$$

- Note: $N_1, \ldots, N_K$ are the sufficient statistics in this case
The posterior over \( \pi \) is easy to compute in this case due to conjugacy b/w multinoulli and Dirichlet

\[
p(\pi|X, \alpha) = \frac{p(X|\pi, \alpha)p(\pi|\alpha)}{p(X|\alpha)} = \frac{p(X|\pi)p(\pi|\alpha)}{p(X|\alpha)}
\]

Assuming \( x_n \)'s are i.i.d. given \( \pi \),
\[
p(X|\pi) = \prod_{n=1}^{N} p(x_n|\pi),
\]

therefore

\[
p(\pi|X, \alpha) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{[x_n=k]} \prod_{k=1}^{K} \pi_k^{\alpha_k-1} = \prod_{k=1}^{K} \pi_k^{\alpha_k + \sum_{n=1}^{N} [x_n=k] - 1}
\]

Even without computing the normalization constant \( p(X|\alpha) \), we can see that it's a Dirichlet! :-)

Denoting \( N_k = \sum_{n=1}^{N} [x_n = k] \), i.e., number of observations with value \( k \), the posterior will be

\[
p(\pi|X, \alpha) = \text{Dirichlet}(\pi|\alpha_1 + N_1, \ldots, \alpha_K + N_K)
\]

Note: \( N_1, \ldots, N_K \) are the sufficient statistics in this case

Note: If we want, we can also get the MAP estimate of \( \pi \) (mode of the above Dirichlet)
Bayesian Inference for Multinoulli/Multinomial

- The posterior over $\pi$ is easy to compute in this case due to conjugacy b/w multinoulli and Dirichlet

$$p(\pi|X, \alpha) = \frac{p(X|\pi, \alpha)p(\pi|\alpha)}{p(X|\alpha)} = \frac{p(X|\pi)p(\pi|\alpha)}{p(X|\alpha)}$$

- Assuming $x_n$'s are i.i.d. given $\pi$, $p(X|\pi) = \prod_{n=1}^{N} p(x_n|\pi)$, therefore

$$p(\pi|X, \alpha) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_{k}^{I[x_n=k]} \prod_{k=1}^{K} \pi_{k}^{\alpha_k-1} = \prod_{k=1}^{K} \pi_{k}^{\alpha_k+N \sum_{n=1}^{N} I[x_n=k]-1}$$

- Even without computing the normalization constant $p(X|\alpha)$, we can see that it's a Dirichlet! :-)

- Denoting $N_k = \sum_{n=1}^{N} I[x_n = k]$, i.e., number of observations with value $k$, the posterior will be

$$p(\pi|X, \alpha) = \text{Dirichlet}(\pi|\alpha_1 + N_1, \ldots, \alpha_K + N_K)$$

- Note: $N_1, \ldots, N_K$ are the sufficient statistics in this case

- Note: If we want, we can also get the MAP estimate of $\pi$ (mode of the above Dirichlet)

  - MAP estimation via standard way will require solving a constraint opt. problem (via Lagrangian)
Finally, let’s also look at the posterior predictive distribution (i.e., the probability distribution of a new observation $x_\ast \in \{1, \ldots, K\}$ given the previous observations $X = \{x_1, \ldots, x_N\}$)

$$p(x_\ast | X, \alpha) = \int p(x_\ast | \pi)p(\pi | X, \alpha) d\pi$$
Finally, let’s also look at the posterior predictive distribution (i.e., the probability distribution of a new observation $x_* \in \{1, \ldots, K\}$ given the previous observations $X = \{x_1, \ldots, x_N\}$)

$$p(x_* | X, \alpha) = \int p(x_* | \pi) p(\pi | X, \alpha) d\pi$$

Note that $p(x_* | \pi) = \text{multinoulli}(x_* | \pi)$ and $p(\pi | X, \alpha) = \text{Dirichlet}(\pi | \alpha_1 + N_1, \ldots, \alpha_K + N_K)$
Bayesian Inference for Multinoulli/Multinomial

- Finally, let’s also look at the **posterior predictive distribution** (i.e., the probability distribution of a new observation \( x_* \in \{1, \ldots, K\} \) given the previous observations \( X = \{x_1, \ldots, x_N\} \))

\[
p(x_*|X, \alpha) = \int p(x_*|\pi)p(\pi|X, \alpha)d\pi
\]

- Note that \( p(x_*|\pi) = \text{multinoulli}(x_*|\pi) \) and \( p(\pi|X, \alpha) = \text{Dirichlet}(\pi|\alpha_1 + N_1, \ldots, \alpha_K + N_K) \)

- We can compute the posterior predictive for each possible outcome (\( K \) possibilities)

\[
p(x_* = k|X, \alpha) = \int p(x_* = k|\pi)p(\pi|X, \alpha)d\pi
\]

Note that the predicted probabilities are smoothed (the effect of averaging over all possible \( \pi \)'s)

Recall that the PPD for the Beta-Bernoulli model also had a similar form!
Finally, let’s also look at the posterior predictive distribution (i.e., the probability distribution of a new observation $x_* \in \{1, \ldots, K\}$ given the previous observations $X = \{x_1, \ldots, x_N\}$)

$$p(x_*|X, \alpha) = \int p(x_*|\pi)p(\pi|X, \alpha)d\pi$$

Note that $p(x_*|\pi) = \text{multinoulli}(x_*|\pi)$ and $p(\pi|X, \alpha) = \text{Dirichlet}(\pi|\alpha_1 + N_1, \ldots, \alpha_K + N_K)$.

We can compute the posterior predictive for each possible outcome ($K$ possibilities)

$$p(x_* = k|X, \alpha) = \int p(x_* = k|\pi)p(\pi|X, \alpha)d\pi$$

$$= \int \pi_k \times \text{Dirichlet}(\pi|\alpha_1 + N_1, \ldots, \alpha_K + N_K)d\pi$$
Finally, let’s also look at the **posterior predictive distribution** (i.e., the probability distribution of a new observation $x^* \in \{1, \ldots, K\}$ given the previous observations $X = \{x_1, \ldots, x_N\}$)

$$p(x^*|X, \alpha) = \int p(x^*|\pi)p(\pi|X, \alpha)d\pi$$

Note that $p(x^*|\pi) = \text{multinoulli}(x^*|\pi)$ and $p(\pi|X, \alpha) = \text{Dirichlet}(\pi|\alpha_1 + N_1, \ldots, \alpha_K + N_K)$

We can compute the posterior predictive for each possible outcome ($K$ possibilities)

$$p(x^* = k|X, \alpha) = \int p(x^* = k|\pi)p(\pi|X, \alpha)d\pi$$

$$= \int \pi_k \times \text{Dirichlet}(\pi|\alpha_1 + N_1, \ldots, \alpha_K + N_K)d\pi$$

$$= \frac{\alpha_k + N_k}{\sum_{k=1}^{K} \alpha_k + N} \quad \text{(expectation of } \pi_k \text{ under the Dirichlet posterior)}$$
Finally, let’s also look at the posterior predictive distribution (i.e., the probability distribution of a new observation $x^* \in \{1, \ldots, K\}$ given the previous observations $X = \{x_1, \ldots, x_N\}$)

$$p(x^* | X, \alpha) = \int p(x^* | \pi)p(\pi | X, \alpha)d\pi$$

Note that $p(x^* | \pi) = \text{multinoulli}(x^* | \pi)$ and $p(\pi | X, \alpha) = \text{Dirichlet}(\pi | \alpha_1 + N_1, \ldots, \alpha_K + N_K)$

We can compute the posterior predictive for each possible outcome ($K$ possibilities)

$$p(x^* = k | X, \alpha) = \int p(x^* = k | \pi)p(\pi | X, \alpha)d\pi$$

$$= \int \pi_k \times \text{Dirichlet}(\pi | \alpha_1 + N_1, \ldots, \alpha_K + N_K)d\pi$$

$$= \frac{\alpha_k + N_k}{\sum_{k=1}^K \alpha_k + N} \quad \text{(expectation of } \pi_k \text{ under the Dirichlet posterior)}$$

Therefore the posterior predictive distribution is multinoulli with posterior mean given as above
Bayesian Inference for Multinoulli/Multinomial

Finally, let’s also look at the posterior predictive distribution (i.e., the probability distribution of a new observation \( x_\ast \in \{1, \ldots, K\} \) given the previous observations \( X = \{x_1, \ldots, x_N\} \))

\[
p(x_\ast | X, \alpha) = \int p(x_\ast | \pi)p(\pi | X, \alpha) d\pi
\]

Note that \( p(x_\ast | \pi) = \text{multinoulli}(x_\ast | \pi) \) and \( p(\pi | X, \alpha) = \text{Dirichlet}(\pi | \alpha_1 + N_1, \ldots, \alpha_K + N_K) \)

We can compute the posterior predictive for each possible outcome (\( K \) possibilities)

\[
p(x_\ast = k | X, \alpha) = \int p(x_\ast = k | \pi)p(\pi | X, \alpha) d\pi
\]

\[
= \int \pi_k \times \text{Dirichlet}(\pi | \alpha_1 + N_1, \ldots, \alpha_K + N_K) d\pi
\]

\[
= \frac{\alpha_k + N_k}{\sum_{k=1}^{K} \alpha_k + N} \quad \text{(expectation of } \pi_k \text{ under the Dirichlet posterior)}
\]

Therefore the posterior predictive distribution is multinoulli with posterior mean given as above

Note that the predicted probabilities are smoothed (the effect of averaging over all possible \( \pi \)’s).
Finally, let’s also look at the posterior predictive distribution (i.e., the probability distribution of a new observation $x_* \in \{1, \ldots, K\}$ given the previous observations $X = \{x_1, \ldots, x_N\}$)

$$p(x_*|X, \alpha) = \int p(x_*|\pi)p(\pi|X, \alpha)d\pi$$

Note that $p(x_*|\pi) = \text{multinoulli}(x_*|\pi)$ and $p(\pi|X, \alpha) = \text{Dirichlet}(\pi|\alpha_1 + N_1, \ldots, \alpha_K + N_K)$

We can compute the posterior predictive for each possible outcome ($K$ possibilities)

$$p(x_* = k|X, \alpha) = \int p(x_* = k|\pi)p(\pi|X, \alpha)d\pi$$

$$= \int \pi_k \times \text{Dirichlet}(\pi|\alpha_1 + N_1, \ldots, \alpha_K + N_K)d\pi$$

$$= \frac{\alpha_k + N_k}{\sum_{k=1}^{K} \alpha_k + N}$$

(expectation of $\pi_k$ under the Dirichlet posterior)

Therefore the posterior predictive distribution is multinoulli with posterior mean given as above

Note that the predicted probabilities are smoothed (the effect of averaging over all possible $\pi$’s)

Recall that the PPD for the Beta-Bernoulli model also had a similar form!
Applications?

- Both, Beta-Bernoulli and Dirichlet-Multinoulli/Multinomial models are widely used
Applications?

- Both, Beta-Bernoulli and Dirichlet-Multinoulli/Multinomial models are widely used.
- We now know how to do fully Bayesian inference if parts of our model have such components.
Applications?

- Both, Beta-Bernoulli and Dirichlet-Multinoulli/Multinomial models are widely used.
- We now know how to do fully Bayesian inference if parts of our model have such components.

Some popular examples are:

- Models for text data: Each document can be modeled as a bag-of-words (Beta-Bernoulli) or a sequence of token (Dirichlet-Multinoulli).
Applications?

- Both, Beta-Bernoulli and Dirichlet-Multinoulli/Multinomial models are widely used.
- We now know how to do fully Bayesian inference if parts of our model have such components.

Some popular examples are

- Models for text data: Each document can be modeled as a bag-of-words (Beta-Bernoulli) or a sequence of tokens (Dirichlet-Multinomial).
- Bayesian inference for class probabilities in classification models: Class labels of training examples are observations and class probabilities are to be estimated.
Applications?

- Both, Beta-Bernoulli and Dirichlet-Multinoulli/Multinomial models are widely used.
- We now know how to do fully Bayesian inference if parts of our model have such components.

Some popular examples are:

- Models for text data: Each document can be modeled as a bag-of-words (Beta-Bernoulli) or a sequence of tokens (Dirichlet-Multinoulli).
- Bayesian inference for class probabilities in classification models: Class labels of training examples are observations and class probabilities are to be estimated.
- Bayesian inference for mixture models: Cluster ids are our (latent) “observations” of Dir-Mult model and mixing proportions are to be estimated.
Applications?

- Both, Beta-Bernoulli and Dirichlet-Multinoulli/Multinomial models are widely used
- We now know how to do fully Bayesian inference if parts of our model have such components

Some popular examples are

- Models for text data: Each document can be modeled as a bag-of-words (Beta-Bernoulli) or a sequence of token (Dirichlet-Multinoulli)
- Bayesian inference for class probabilities in classification models: Class labels of training examples are observations and class probabilities are to be estimated
- Bayesian inference for mixture models: Cluster ids are our (latent) “observations” of Dir-Mult model and mixing proportions are to be estimated
- .. and several others, which we will see later..
Some More Examples..
Bayesian Inference for Mean of a Gaussian

Consider $N$ i.i.d. observations $\mathbf{X} = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \sigma^2)$

$$p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \propto \exp\left[ -\frac{(x_n - \mu)^2}{2\sigma^2} \right]$$

$$p(\mathbf{X}|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)$$
Bayesian Inference for Mean of a Gaussian

- Consider $N$ i.i.d. observations $X = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \sigma^2)$
  
  \[
p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \propto \exp\left[-\frac{(x_n - \mu)^2}{2\sigma^2}\right]
  \]

  \[
p(X|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)
  \]

- Assume the mean $\mu \in \mathbb{R}$ of the Gaussian is unknown and assume variance $\sigma^2$ to be known/fixed
Bayesian Inference for Mean of a Gaussian

- Consider $N$ i.i.d. observations $X = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \sigma^2)$

\[
p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \propto \exp\left[-\frac{(x_n - \mu)^2}{2\sigma^2}\right]
\]

\[
p(X|\mu, \sigma^2) = \prod_{n=1}^N p(x_n|\mu, \sigma^2)
\]

- Assume the mean $\mu \in \mathbb{R}$ of the Gaussian is unknown and assume variance $\sigma^2$ to be known/fixed

- We wish to estimate the unknown $\mu$ given the data $X$
Bayesian Inference for Mean of a Gaussian

Consider \( N \) i.i.d. observations \( X = \{x_1, \ldots, x_N\} \) drawn from a one-dim Gaussian \( \mathcal{N}(x|\mu, \sigma^2) \)

\[
p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \propto \exp \left(-\frac{(x_n - \mu)^2}{2\sigma^2}\right)
\]

\[
p(X|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)
\]

Assume the mean \( \mu \in \mathbb{R} \) of the Gaussian is unknown and assume variance \( \sigma^2 \) to be known/fixed

We wish to estimate the unknown \( \mu \) given the data \( X \)

Let's do fully Bayesian inference for \( \mu \) (not MLE/MAP)
Bayesian Inference for Mean of a Gaussian

- Consider $N$ i.i.d. observations $\mathbf{X} = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \sigma^2)$

$$p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \propto \exp \left[-\frac{(x_n - \mu)^2}{2\sigma^2}\right]$$

$$p(\mathbf{X}|\mu, \sigma^2) = \prod_{n=1}^N p(x_n|\mu, \sigma^2)$$

- Assume the mean $\mu \in \mathbb{R}$ of the Gaussian is unknown and assume variance $\sigma^2$ to be known/fixed

- We wish to estimate the unknown $\mu$ given the data $\mathbf{X}$

- Let's do fully Bayesian inference for $\mu$ (not MLE/MAP)

- We first need a prior distribution for the unknown param. $\mu$
Bayesian Inference for Mean of a Gaussian

Consider $N$ i.i.d. observations $\mathbf{X} = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \sigma^2)$

$$p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \propto \exp\left[\frac{-(x_n - \mu)^2}{2\sigma^2}\right]$$

$$p(\mathbf{X}|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)$$

Assume the mean $\mu \in \mathbb{R}$ of the Gaussian is unknown and assume variance $\sigma^2$ to be known/fixed.

We wish to estimate the unknown $\mu$ given the data $\mathbf{X}$.

Let’s do fully Bayesian inference for $\mu$ (not MLE/MAP).

We first need a prior distribution for the unknown param. $\mu$.

Let’s choose a Gaussian prior on $\mu$, i.e., $p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$ with $\mu_0, \sigma_0^2$ as fixed.
Bayesian Inference for Mean of a Gaussian

- Consider $N$ i.i.d. observations $X = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \sigma^2)$

$$p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \propto \exp \left[ -\frac{(x_n - \mu)^2}{2\sigma^2} \right]$$

$$p(X|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)$$

- Assume the mean $\mu \in \mathbb{R}$ of the Gaussian is unknown and assume variance $\sigma^2$ to be known/fixed
- We wish to estimate the unknown $\mu$ given the data $X$
- Let’s do fully Bayesian inference for $\mu$ (not MLE/MAP)
- We first need a prior distribution for the unknown param. $\mu$
- Let’s choose a Gaussian prior on $\mu$, i.e., $p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma^2_0)$ with $\mu_0, \sigma^2_0$ as fixed
- The prior basically says that the mean $\mu$ is close to $\mu_0$ (with some uncertainty depending on $\sigma^2_0$)
Bayesian Inference for Mean of a Gaussian

The posterior distribution for the unknown mean parameter $\mu$

$$p(\mu|X) = \frac{p(X|\mu)p(\mu)}{p(X)} \propto \prod_{n=1}^{N} \exp \left[ -\frac{(x_n - \mu)^2}{2\sigma^2} \right] \times \exp \left[ -\frac{(\mu - \mu_0)^2}{2\sigma_0^2} \right]$$
Bayesian Inference for Mean of a Gaussian

The posterior distribution for the unknown mean parameter $\mu$

$$p(\mu|X) = \frac{p(X|\mu)p(\mu)}{p(X)} \propto \prod_{n=1}^{N} \exp \left[-\frac{(x_n - \mu)^2}{2\sigma^2}\right] \times \exp \left[-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right]$$

Simplifying the above (using completing the squares trick) gives $p(\mu|X) \propto \exp \left[-\frac{(\mu - \bar{x})^2}{2\sigma_N^2}\right]$ with

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

Posterior and prior have the same form (not surprising; the prior was conjugate to the likelihood)

Consider what happens as $N$ (number of observations) grows very large?

The posterior's variance $\sigma_N^2$ approaches $\sigma^2/N$ (and goes to 0 as $N \to \infty$)

The posterior's mean $\mu_N$ approaches $\bar{x}$ (which is also the MLE solution)
Bayesian Inference for Mean of a Gaussian

The posterior distribution for the unknown mean parameter $\mu$

$$p(\mu|X) = \frac{p(X|\mu)p(\mu)}{p(X)} \propto \prod_{n=1}^{N} \exp \left[ - \frac{(x_n - \mu)^2}{2\sigma^2} \right] \times \exp \left[ - \frac{(\mu - \mu_0)^2}{2\sigma_0^2} \right]$$

Simplifying the above (using completing the squares trick) gives $p(\mu|X) \propto \exp \left[ - \frac{(\mu - \mu_N)^2}{2\sigma_N^2} \right]$ with

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \bar{x} \quad \text{(where } \bar{x} = \frac{\sum_{n=1}^{N} x_n}{N})$$
Bayesian Inference for Mean of a Gaussian

- The posterior distribution for the unknown mean parameter $\mu$

$$p(\mu|X) = \frac{p(X|\mu)p(\mu)}{p(X)} \propto \prod_{n=1}^{N} \exp \left[ -\frac{(x_n - \mu)^2}{2\sigma^2} \right] \times \exp \left[ -\frac{(\mu - \mu_0)^2}{2\sigma_0^2} \right]$$

- Simplifying the above (using completing the squares trick) gives $p(\mu|X) \propto \exp \left[ -\frac{(\mu - \mu_N)^2}{2\sigma_N^2} \right]$ with

$$\frac{1}{\sigma^2_N} = \frac{1}{\sigma^2_0} + \frac{N}{\sigma^2}$$

$$\mu_N = \frac{\sigma^2}{N\sigma^2_0 + \sigma^2} \mu_0 + \frac{N\sigma^2_0}{N\sigma^2_0 + \sigma^2} \bar{x} \quad \text{(where $\bar{x} = \frac{\sum_{n=1}^{N} x_n}{N}$)}$$

- Posterior and prior have the same form (not surprising; the prior was conjugate to the likelihood)
Bayesian Inference for Mean of a Gaussian

- The posterior distribution for the unknown mean parameter $\mu$

$$p(\mu|X) = \frac{p(X|\mu)p(\mu)}{p(X)} \propto \prod_{n=1}^{N} \exp \left[ -\frac{(x_n - \mu)^2}{2\sigma^2} \right] \times \exp \left[ -\frac{(\mu - \mu_0)^2}{2\sigma_0^2} \right]$$

- Simplifying the above (using completing the squares trick) gives $p(\mu|X) \propto \exp \left[ -\frac{(\mu - \mu_N)^2}{2\sigma_N^2} \right]$ with

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \bar{x}$$

(where $\bar{x} = \frac{\sum_{n=1}^{N} x_n}{N}$)

- Posterior and prior have the same form (not surprising; the prior was conjugate to the likelihood)

- Consider what happens as $N$ (number of observations) grows very large?
Bayesian Inference for Mean of a Gaussian

- The posterior distribution for the unknown mean parameter $\mu$

$$p(\mu|X) = \frac{p(X|\mu)p(\mu)}{p(X)} \propto \prod_{n=1}^{N} \exp \left[-\frac{(x_n - \mu)^2}{2\sigma^2}\right] \times \exp \left[-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right]$$

- Simplifying the above (using completing the squares trick) gives $p(\mu|X) \propto \exp \left[-\frac{(\mu - \mu_N)^2}{2\sigma_N^2}\right]$ with

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \bar{x} \quad (\text{where } \bar{x} = \frac{\sum_{n=1}^{N} x_n}{N})$$

- Posterior and prior have the same form (not surprising; the prior was conjugate to the likelihood)

- Consider what happens as $N$ (number of observations) grows very large?

  - The posterior’s variance $\sigma_N^2$ approaches $\sigma^2/N$ (and goes to 0 as $N \to \infty$)
Bayesian Inference for Mean of a Gaussian

- The posterior distribution for the unknown mean parameter $\mu$
  \[
p(\mu|X) = \frac{p(X|\mu)p(\mu)}{p(X)} \propto \prod_{n=1}^{N} \exp \left[ -\frac{(x_n - \mu)^2}{2\sigma^2} \right] \times \exp \left[ -\frac{(\mu - \mu_0)^2}{2\sigma_0^2} \right]
  \]

- Simplifying the above (using completing the squares trick) gives $p(\mu|X) \propto \exp \left[ -\frac{(\mu - \mu_N)^2}{2\sigma_N^2} \right]$ with
  \[
  \frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}
  \]
  \[
  \mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2 \mu_0} + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \bar{x}
  \]
  (where $\bar{x} = \frac{\sum_{n=1}^{N} x_n}{N}$)

- Posterior and prior have the same form (not surprising; the prior was conjugate to the likelihood)

- Consider what happens as $N$ (number of observations) grows very large?
  - The posterior’s variance $\sigma_N^2$ approaches $\sigma^2/N$ (and goes to 0 as $N \to \infty$)
  - The posterior’s mean $\mu_N$ approaches $\bar{x}$ (which is also the MLE solution)
What is the posterior predictive distribution \( p(x_\star | X) \) of a new observation \( x_\star \)?

Using the inferred posterior \( p(\mu | X) \), we can find the posterior predictive distribution

\[
p(x_\star | X) = \int p(x_\star | \mu, \sigma^2) p(\mu | X) d\mu
\]
Bayesian Inference for Mean of a Gaussian

- What is the posterior predictive distribution \( p(x_*|X) \) of a new observation \( x_* \)?
- Using the inferred posterior \( p(\mu|X) \), we can find the posterior predictive distribution

\[
p(x_*|X) = \int p(x_*|\mu, \sigma^2)p(\mu|X)d\mu = \int \mathcal{N}(x_*|\mu, \sigma^2)\mathcal{N}(\mu|\mu_N, \sigma^2_N)d\mu
\]
Bayesian Inference for Mean of a Gaussian

- What is the posterior predictive distribution \( p(x_*|X) \) of a new observation \( x_* \)?
- Using the inferred posterior \( p(\mu|X) \), we can find the posterior predictive distribution

\[
p(x_*|X) = \int p(x_*|\mu, \sigma^2)p(\mu|X)d\mu = \int \mathcal{N}(x_*|\mu, \sigma^2)\mathcal{N}(\mu|\mu_N, \sigma^2_N)d\mu = \mathcal{N}(x_*|\mu_N, \sigma^2 + \sigma^2_N)
\]
Bayesian Inference for Mean of a Gaussian

- What is the posterior predictive distribution $p(x_*|X)$ of a new observation $x_*$?
- Using the inferred posterior $p(\mu|X)$, we can find the posterior predictive distribution

$$p(x_*|X) = \int p(x_*|\mu, \sigma^2)p(\mu|X)d\mu = \int \mathcal{N}(x_*|\mu, \sigma^2)\mathcal{N}(\mu|\mu_N, \sigma^2_N)d\mu = \mathcal{N}(x_*|\mu_N, \sigma^2 + \sigma^2_N)$$

- Note; Can also get the above result by thinking of $x_*$ as $x_* = \mu + \epsilon$ where $\mu \sim \mathcal{N}(\mu_N, \sigma^2_N)$, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is independently added observation noise.
Bayesian Inference for Mean of a Gaussian

- What is the **posterior predictive distribution** \( p(x_*|X) \) of a new observation \( x_* \)?

- Using the inferred posterior \( p(\mu|X) \), we can find the posterior predictive distribution

  \[
p(x_*|X) = \int p(x_*|\mu, \sigma^2)p(\mu|X)d\mu = \int \mathcal{N}(x_*|\mu, \sigma^2)\mathcal{N}(\mu|\mu_N, \sigma_N^2)d\mu = \mathcal{N}(x_*|\mu_N, \sigma^2 + \sigma_N^2)
  \]

- Note; Can also get the above result by thinking of \( x_* \) as \( x_* = \mu + \epsilon \) where \( \mu \sim \mathcal{N}(\mu_N, \sigma_N^2) \), and \( \epsilon \sim \mathcal{N}(0, \sigma^2) \) is independently added observation noise

- Note that, as per the above, the uncertainty in distribution of \( x_* \) now has two components
What is the posterior predictive distribution $p(x_*|X)$ of a new observation $x_*$?

Using the inferred posterior $p(\mu|X)$, we can find the posterior predictive distribution

$$p(x_*|X) = \int p(x_*|\mu, \sigma^2)p(\mu|X)d\mu = \int \mathcal{N}(x_*|\mu, \sigma^2)\mathcal{N}(\mu|\mu_N, \sigma^2_N)d\mu = \mathcal{N}(x_*|\mu_N, \sigma^2 + \sigma^2_N)$$

Note; Can also get the above result by thinking of $x_*$ as $x_* = \mu + \epsilon$ where $\mu \sim \mathcal{N}(\mu_N, \sigma^2_N)$, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is independently added observation noise

Note that, as per the above, the uncertainty in distribution of $x_*$ now has two components

- $\sigma^2$: Due to the noisy observation model, $\sigma^2_N$: Due to the uncertainty in $\mu$
Bayesian Inference for Mean of a Gaussian

- What is the posterior predictive distribution $p(x_*|X)$ of a new observation $x_*$?
- Using the inferred posterior $p(\mu|X)$, we can find the posterior predictive distribution

$$p(x_*|X) = \int p(x_*|\mu, \sigma^2)p(\mu|X)d\mu = \int \mathcal{N}(x_*|\mu, \sigma^2)\mathcal{N}(\mu|\mu_N, \sigma^2_N)d\mu = \mathcal{N}(x_*|\mu_N, \sigma^2 + \sigma^2_N)$$

- Note: Can also get the above result by thinking of $x_*$ as $x_* = \mu + \epsilon$ where $\mu \sim \mathcal{N}(\mu_N, \sigma^2_N)$, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is independently added observation noise.
- Note that, as per the above, the uncertainty in distribution of $x_*$ now has two components
  - $\sigma^2$: Due to the noisy observation model, $\sigma^2_N$: Due to the uncertainty in $\mu$
- In contrast, the plug-in predictive posterior, given a point estimate $\hat{\mu}$ (e.g., MLE/MAP) would be

$$p(x_*|X) \approx p(x_*|\hat{\mu}, \sigma^2)$$
Bayesian Inference for Mean of a Gaussian

- What is the posterior predictive distribution $p(x_*|X)$ of a new observation $x_*$?
- Using the inferred posterior $p(\mu|X)$, we can find the posterior predictive distribution

$$p(x_*|X) = \int p(x_*|\mu, \sigma^2)p(\mu|X)d\mu = \int \mathcal{N}(x_*|\mu, \sigma^2)\mathcal{N}(\mu|\mu_N, \sigma^2_N)d\mu = \mathcal{N}(x_*|\mu_N, \sigma^2 + \sigma^2_N)$$

- Note; Can also get the above result by thinking of $x_*$ as $x_* = \mu + \epsilon$ where $\mu \sim \mathcal{N}(\mu_N, \sigma^2_N)$, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is independently added observation noise.

- Note that, as per the above, the uncertainty in distribution of $x_*$ now has two components
  - $\sigma^2$: Due to the noisy observation model, $\sigma^2_N$: Due to the uncertainty in $\mu$

- In contrast, the plug-in predictive posterior, given a point estimate $\hat{\mu}$ (e.g., MLE/MAP) would be

$$p(x_*|X) = \int p(x_*|\mu, \sigma^2)p(\mu|X)d\mu \approx p(x_*|\hat{\mu}, \sigma^2) = \mathcal{N}(x_*|\hat{\mu}, \sigma^2)$$
Bayesian Inference for Mean of a Gaussian

- What is the posterior predictive distribution $p(x_\ast|X)$ of a new observation $x_\ast$?

- Using the inferred posterior $p(\mu|X)$, we can find the posterior predictive distribution

$$p(x_\ast|X) = \int p(x_\ast|\mu, \sigma^2)p(\mu|X)d\mu = \int \mathcal{N}(x_\ast|\mu, \sigma^2)\mathcal{N}(\mu|\mu_N, \sigma^2_N)d\mu = \mathcal{N}(x_\ast|\mu_N, \sigma^2 + \sigma^2_N)$$

- Note: Can also get the above result by thinking of $x_\ast$ as $x_\ast = \mu + \epsilon$ where $\mu \sim \mathcal{N}(\mu_N, \sigma^2_N)$, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is independently added observation noise.

- Note that, as per the above, the uncertainty in distribution of $x_\ast$ now has two components
  - $\sigma^2$: Due to the noisy observation model,
  - $\sigma^2_N$: Due to the uncertainty in $\mu$.

- In contrast, the \textbf{plug-in predictive posterior}, given a point estimate $\hat{\mu}$ (e.g., MLE/MAP) would be

$$p(x_\ast|X) = \int p(x_\ast|\mu, \sigma^2)p(\mu|X)d\mu \approx p(x_\ast|\hat{\mu}, \sigma^2) = \mathcal{N}(x_\ast|\hat{\mu}, \sigma^2)$$

.. which doesn’t incorporate the uncertainty in our estimate of $\mu$ (since we used a point estimate).
Bayesian Inference for Mean of a Gaussian

What is the posterior predictive distribution $p(x_*|X)$ of a new observation $x_*$?

Using the inferred posterior $p(\mu|X)$, we can find the posterior predictive distribution

$$p(x_*|X) = \int p(x_*|\mu, \sigma^2)p(\mu|X)d\mu = \int \mathcal{N}(x_*|\mu, \sigma^2)\mathcal{N}(\mu|\mu_N, \sigma_N^2)d\mu = \mathcal{N}(x_*|\mu_N, \sigma^2 + \sigma_N^2)$$

Note; Can also get the above result by thinking of $x_*$ as $x_* = \mu + \epsilon$ where $\mu \sim \mathcal{N}(\mu_N, \sigma_N^2)$, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is independently added observation noise

Note that, as per the above, the uncertainty in distribution of $x_*$ now has two components

- $\sigma^2$: Due to the noisy observation model, $\sigma_N^2$: Due to the uncertainty in $\mu$

In contrast, the plug-in predictive posterior, given a point estimate $\hat{\mu}$ (e.g., MLE/MAP) would be

$$p(x_*|X) = \int p(x_*|\mu, \sigma^2)p(\mu|X)d\mu \approx p(x_*|\hat{\mu}, \sigma^2) = \mathcal{N}(x_*|\hat{\mu}, \sigma^2)$$

.. which doesn’t incorporate the uncertainty in our estimate of $\mu$ (since we used a point estimate)

Note that as $N \to \infty$, both approaches would give the same $p(x_*|X)$ since $\sigma_N^2 \to 0$
Again consider $N$ i.i.d. observations $X = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \sigma^2)$.

$$p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \quad \text{and} \quad p(X|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)$$
Bayesian Inference for Variance of a Gaussian

Again consider $N$ i.i.d. observations $X = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \sigma^2)$

$$p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \quad \text{and} \quad p(X|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)$$

Assume the variance $\sigma^2 \in \mathbb{R}_+$ of the Gaussian is unknown and assume mean $\mu$ to be known/fixed
Again consider $N$ i.i.d. observations $X = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \sigma^2)$.

\[
p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \quad \text{and} \quad p(X|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)
\]

Assume the variance $\sigma^2 \in \mathbb{R}_+$ of the Gaussian is unknown and assume mean $\mu$ to be known/fixed.

Let's estimate $\sigma^2$ given the data $X$ using fully Bayesian inference (not MLE/MAP).
Bayesian Inference for Variance of a Gaussian

Again consider $N$ i.i.d. observations $X = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \sigma^2)$

$$p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \quad \text{and} \quad p(X|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)$$

Assume the variance $\sigma^2 \in \mathbb{R}_+$ of the Gaussian is unknown and assume mean $\mu$ to be known/fixed

Let’s estimate $\sigma^2$ given the data $X$ using fully Bayesian inference (not MLE/MAP)

We first need a prior distribution for $\sigma^2$. What prior $p(\sigma^2)$ to choose in this case?
Bayesian Inference for Variance of a Gaussian

Again consider $N$ i.i.d. observations $X = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \sigma^2)$

$$p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \quad \text{and} \quad p(X|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)$$

Assume the variance $\sigma^2 \in \mathbb{R}_+$ of the Gaussian is unknown and assume mean $\mu$ to be known/fixed

Let’s estimate $\sigma^2$ given the data $X$ using fully Bayesian inference (not MLE/MAP)

We first need a prior distribution for $\sigma^2$. What prior $p(\sigma^2)$ to choose in this case?

If we want a conjugate prior, it should have the same form as the likelihood

$$p(x_n|\mu, \sigma^2) \propto (\sigma^2)^{-1/2} \exp \left[ -\frac{(x_n - \mu)^2}{2\sigma^2} \right]$$
Bayesian Inference for Variance of a Gaussian

Again consider $N$ i.i.d. observations $X = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \sigma^2)$

$$p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \quad \text{and} \quad p(X|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)$$

Assume the variance $\sigma^2 \in \mathbb{R}_+$ of the Gaussian is unknown and assume mean $\mu$ to be known/fixe

Let’s estimate $\sigma^2$ given the data $X$ using fully Bayesian inference (not MLE/MAP)

We first need a prior distribution for $\sigma^2$. What prior $p(\sigma^2)$ to choose in this case?

If we want a conjugate prior, it should have the same form as the likelihood

$$p(x_n|\mu, \sigma^2) \propto (\sigma^2)^{-1/2} \exp\left[-\frac{(x_n - \mu)^2}{2\sigma^2}\right]$$

An inverse-gamma prior $\text{IG}(\alpha, \beta)$ has this form ($\alpha, \beta$ are shape and scale hyperparams, resp)

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} \exp\left[-\frac{\beta}{\sigma^2}\right]$$
Bayesian Inference for Variance of a Gaussian

- Again consider $N$ i.i.d. observations $X = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \sigma^2)$

$$p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \quad \text{and} \quad p(X|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)$$

- Assume the variance $\sigma^2 \in \mathbb{R}_+$ of the Gaussian is unknown and assume mean $\mu$ to be known/fixed

- Let's estimate $\sigma^2$ given the data $X$ using fully Bayesian inference (not MLE/MAP)

- We first need a prior distribution for $\sigma^2$. What prior $p(\sigma^2)$ to choose in this case?

- If we want a conjugate prior, it should have the same form as the likelihood

$$p(x_n|\mu, \sigma^2) \propto (\sigma^2)^{-1/2} \exp\left[-\frac{(x_n - \mu)^2}{2\sigma^2}\right]$$

- An inverse-gamma prior $\text{IG}(\alpha, \beta)$ has this form ($\alpha$, $\beta$ are shape and scale hyperparams, resp)

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} \exp\left[-\frac{\beta}{\sigma^2}\right] \quad \text{(note: mean of } \text{IG}(\alpha, \beta) = \frac{\beta}{\alpha - 1})$$
Bayesian Inference for Variance of a Gaussian

Again consider \( N \) i.i.d. observations \( \mathbf{X} = \{x_1, \ldots, x_N\} \) drawn from a one-dim Gaussian \( \mathcal{N}(x|\mu, \sigma^2) \)

\[
p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \quad \text{and} \quad p(\mathbf{X}|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)
\]

Assume the variance \( \sigma^2 \in \mathbb{R}_+ \) of the Gaussian is unknown and assume mean \( \mu \) to be known/fixed

Let’s estimate \( \sigma^2 \) given the data \( \mathbf{X} \) using fully Bayesian inference (not MLE/MAP)

We first need a prior distribution for \( \sigma^2 \). What prior \( p(\sigma^2) \) to choose in this case?

If we want a conjugate prior, it should have the same form as the likelihood

\[
p(x_n|\mu, \sigma^2) \propto (\sigma^2)^{-1/2} \exp \left[ -\frac{(x_n - \mu)^2}{2\sigma^2} \right]
\]

An inverse-gamma prior \( \text{IG}(\alpha, \beta) \) has this form (\( \alpha, \beta \) are shape and scale hyperparams, resp)

\[
p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} \exp \left[ -\frac{\beta}{\sigma^2} \right] \quad \text{(note: mean of } \text{IG}(\alpha, \beta) = \frac{\beta}{\alpha - 1})
\]

(Verify) The posterior \( p(\sigma^2|\mathbf{X}) = \text{IG}(\alpha + \frac{N}{2}, \beta + \sum_{n=1}^{N} \frac{(x_n - \mu)^2}{2}) \). Again IG due to conjugacy.
Often, it is easier to work with the precision ($=1/$variance) rather than variance

$$p(x_n|\mu, \tau) = \mathcal{N}(x|\mu, \tau) = \sqrt{\frac{\tau}{2\pi}} \exp \left[ -\frac{\tau}{2} (x_n - \mu)^2 \right]$$
Working with Gaussians: Variance vs Precision

- Often, it is easier to work with the precision (=1/variance) rather than variance

\[ p(x_n|\mu, \tau) = \mathcal{N}(x|\mu, \tau) = \sqrt{\frac{\tau}{2\pi}} \exp \left[ -\frac{\tau}{2}(x_n - \mu)^2 \right] \]

- If mean is known, for precision Gamma($\alpha, \beta$) is a conjugate prior to Gaussian likelihood

\[ p(\tau) \propto (\tau)^{\alpha-1} \exp [-\beta \tau] \]
Working with Gaussians: Variance vs Precision

- Often, it is easier to work with the precision (=1/variance) rather than variance

\[ p(x_n | \mu, \tau) = \mathcal{N}(x | \mu, \tau) = \sqrt{\frac{\tau}{2\pi}} \exp \left[ -\frac{\tau}{2} (x_n - \mu)^2 \right] \]

- If mean is known, for precision Gamma(\(\alpha, \beta\)) is a conjugate prior to Gaussian likelihood

\[ p(\tau) \propto (\tau)^{\alpha-1} \exp [-\beta \tau] \quad \text{(note: mean of Gamma(\(\alpha, \beta\)) = \(\frac{\alpha}{\beta}\))} \]

.. where \(\alpha\) and \(\beta\) are the shape and rate hyperparameters, respectively, for the Gamma
Working with Gaussians: Variance vs Precision

Often, it is easier to work with the precision (=1/variance) rather than variance

\[
p(x_n|\mu, \tau) = \mathcal{N}(x|\mu, \tau) = \sqrt{\frac{\tau}{2\pi}} \exp\left[ -\frac{\tau}{2}(x_n - \mu)^2 \right]
\]

If mean is known, for precision \( \text{Gamma}(\alpha, \beta) \) is a conjugate prior to Gaussian likelihood

\[
p(\tau) \propto (\tau)^{\alpha-1} \exp[-\beta \tau] \quad \text{(note: mean of Gamma}(\alpha, \beta) = \frac{\alpha}{\beta})
\]

.. where \( \alpha \) and \( \beta \) are the shape and rate hyperparamers, respectively, for the Gamma

(Verify) The posterior \( p(\tau|X) \) will also be \( \text{Gamma}(\alpha + \frac{N}{2}, \beta + \sum_{n=1}^{N}(x_n-\mu)^2) \)
Working with Gaussians: Variance vs Precision

Often, it is easier to work with the precision (=1/variance) rather than variance

\[ p(x_n|\mu, \tau) = \mathcal{N}(x|\mu, \tau) = \sqrt{\frac{\tau}{2\pi}} \exp \left[ -\frac{\tau}{2}(x_n - \mu)^2 \right] \]

If mean is known, for precision Gamma(\(\alpha, \beta\)) is a conjugate prior to Gaussian likelihood

\[ p(\tau) \propto (\tau)^{\alpha-1} \exp[-\beta\tau] \quad \text{(note: mean of Gamma(\(\alpha, \beta\)) = } \frac{\alpha}{\beta} \text{)} \]

.. where \(\alpha\) and \(\beta\) are the shape and rate hyperparamers, respectively, for the Gamma

(Verify) The posterior \(p(\tau|X)\) will also be Gamma(\(\alpha + \frac{N}{2}, \beta + \sum_{n=1}^{N}(x_n - \mu)^2\))

Note: Gamma distribution can be defined in terms of shape and scale or shape and rate parametrization (scale = 1/rate). Likewise, inverse Gamma can also be defined both shape and scale (which we saw) as well as shape and rate parametrizations.
Bayesian Inference for Both Parameters of a Gaussian!

- Gaussian with unknown scalar mean and unknown scalar precision (two parameters)
Bayesian Inference for Both Parameters of a Gaussian!

- Gaussian with unknown scalar mean and unknown scalar precision (two parameters)
- Consider $N$ i.i.d. observations $X = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \lambda^{-1})$
Bayesian Inference for Both Parameters of a Gaussian!

- Gaussian with unknown scalar mean and unknown scalar precision (two parameters)
- Consider $N$ i.i.d. observations $X = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \lambda^{-1})$
- Assume both mean $\mu$ and precision $\lambda$ to be unknown. The likelihood will be

$$p(X|\mu, \lambda) = \prod_{n=1}^{N} \sqrt{\frac{\lambda}{2\pi}} \exp \left[-\frac{\lambda}{2}(x_n - \mu)^2\right]$$
Bayesian Inference for Both Parameters of a Gaussian!

- Gaussian with unknown scalar mean and unknown scalar precision (two parameters)
- Consider $N$ i.i.d. observations $X = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \lambda^{-1})$
- Assume both mean $\mu$ and precision $\lambda$ to be unknown. The likelihood will be
  \[
p(X|\mu, \lambda) = \prod_{n=1}^{N} \sqrt{\frac{\lambda}{2\pi}} \exp \left[ -\frac{\lambda}{2} (x_n - \mu)^2 \right]
\]

  \[
  \propto \left[ \lambda^{1/2} \exp \left( -\frac{\lambda \mu^2}{2} \right) \right]^N \exp \left[ \lambda \mu \sum_{n=1}^{N} x_n - \frac{\lambda}{2} \sum_{n=1}^{N} x_n^2 \right]
\]
Bayesian Inference for Both Parameters of a Gaussian!

- Gaussian with unknown scalar mean and unknown scalar precision (two parameters)
- Consider $N$ i.i.d. observations $X = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \lambda^{-1})$
- Assume both mean $\mu$ and precision $\lambda$ to be unknown. The likelihood will be
  \[
p(X|\mu, \lambda) = \prod_{n=1}^{N} \sqrt{\frac{\lambda}{2\pi}} \exp\left[-\frac{\lambda}{2}(x_n - \mu)^2\right]
\]
  \[
  \propto \left[\lambda^{1/2} \exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left[\lambda\mu \sum_{n=1}^{N} x_n - \frac{\lambda}{2} \sum_{n=1}^{N} x_n^2\right]
\]
- If we want a conjugate joint prior $p(\mu, \lambda)$, it must have the same form as likelihood.

Note: Its multivariate version is the Normal-Wishart (for multivariate mean and precision matrix)
Bayesian Inference for Both Parameters of a Gaussian!

- Gaussian with unknown scalar mean and unknown scalar precision (two parameters)
- Consider $N$ i.i.d. observations $X = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \lambda^{-1})$
- Assume both mean $\mu$ and precision $\lambda$ to be unknown. The likelihood will be

$$
p(X|\mu, \lambda) = \prod_{n=1}^{N} \sqrt{\frac{\lambda}{2\pi}} \exp \left[ -\frac{\lambda}{2} (x_n - \mu)^2 \right]$$

$$
\propto \left[ \lambda^{1/2} \exp \left( -\frac{\lambda \mu^2}{2} \right) \right]^N \exp \left[ \lambda \mu \sum_{n=1}^{N} x_n - \frac{\lambda}{2} \sum_{n=1}^{N} x_n^2 \right]
$$

- If we want a conjugate joint prior $p(\mu, \lambda)$, it must have the same form as likelihood. Suppose

$$
p(\mu, \lambda) \propto \left[ \lambda^{1/2} \exp \left( -\frac{\lambda \mu^2}{2} \right) \right]^\kappa_0 \exp [\lambda \mu c - \lambda d]
$$


Bayesian Inference for Both Parameters of a Gaussian!

- Gaussian with unknown scalar mean and unknown scalar precision (two parameters)
- Consider \( N \) i.i.d. observations \( \mathbf{X} = \{x_1, \ldots, x_N\} \) drawn from a one-dim Gaussian \( \mathcal{N}(x|\mu, \lambda^{-1}) \)
- Assume both mean \( \mu \) and precision \( \lambda \) to be unknown. The likelihood will be

\[
p(\mathbf{X}|\mu, \lambda) = \prod_{n=1}^{N} \sqrt{\frac{\lambda}{2\pi}} \exp \left[ -\frac{\lambda}{2} (x_n - \mu)^2 \right]
\]

\[
\propto \left[ \lambda^{1/2} \exp \left( -\frac{\lambda \mu^2}{2} \right) \right]^N \exp \left[ \lambda \mu \sum_{n=1}^{N} x_n - \frac{\lambda}{2} \sum_{n=1}^{N} x_n^2 \right]
\]

- If we want a conjugate joint prior \( p(\mu, \lambda) \), it must have the same form as likelihood. Suppose

\[
p(\mu, \lambda) \propto \left[ \lambda^{1/2} \exp \left( -\frac{\lambda \mu^2}{2} \right) \right]^{\kappa_0} \exp [\lambda \mu c - \lambda d]
\]

- What’s this prior? A normal-gamma (Gaussian-gamma) distribution! (will see its form shortly)
Bayesian Inference for Both Parameters of a Gaussian!

- Gaussian with unknown scalar mean and unknown scalar precision (two parameters)
- Consider $N$ i.i.d. observations $X = \{x_1, \ldots, x_N\}$ drawn from a one-dimensional Gaussian $\mathcal{N}(x|\mu, \lambda^{-1})$
- Assume both mean $\mu$ and precision $\lambda$ to be unknown. The likelihood will be
  \[
p(X|\mu, \lambda) = \prod_{n=1}^{N} \sqrt{\frac{\lambda}{2\pi}} \exp \left[ -\frac{\lambda}{2} (x_n - \mu)^2 \right]
  \propto \left[ \lambda^{1/2} \exp \left( -\frac{\lambda \mu^2}{2} \right) \right]^N \exp \left[ \lambda \mu \frac{N}{2} \sum_{n=1}^{N} x_n - \frac{\lambda}{2} \sum_{n=1}^{N} x_n^2 \right]
\]
- If we want a conjugate joint prior $p(\mu, \lambda)$, it must have the same form as likelihood. Suppose
  \[
p(\mu, \lambda) \propto \left[ \lambda^{1/2} \exp \left( -\frac{\lambda \mu^2}{2} \right) \right]^{\kappa_0} \exp [\lambda \mu c - \lambda d]
\]
  - What’s this prior? A normal-gamma (Gaussian-gamma) distribution! (will see its form shortly)
  - Can be used when we wish to estimate the unknown mean and unknown precision of a Gaussian
Bayesian Inference for Both Parameters of a Gaussian!

- Gaussian with unknown scalar mean and unknown scalar precision (two parameters)
- Consider $N$ i.i.d. observations $X = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \lambda^{-1})$
- Assume both mean $\mu$ and precision $\lambda$ to be unknown. The likelihood will be

$$p(X|\mu, \lambda) = \prod_{n=1}^{N} \sqrt{\frac{\lambda}{2\pi}} \exp\left[-\frac{\lambda}{2}(x_n - \mu)^2\right]$$

$$\propto \left[\lambda^{1/2} \exp\left(-\frac{\lambda \mu^2}{2}\right)\right]^N \exp\left[\lambda \mu \sum_{n=1}^{N} x_n - \frac{\lambda}{2} \sum_{n=1}^{N} x_n^2\right]$$

- If we want a conjugate joint prior $p(\mu, \lambda)$, it must have the same form as likelihood. Suppose

$$p(\mu, \lambda) \propto \left[\lambda^{1/2} \exp\left(-\frac{\lambda \mu^2}{2}\right)\right]^\kappa_0 \exp[\lambda \mu c - \lambda d]$$

- What’s this prior? A normal-gamma (Gaussian-gamma) distribution! (will see its form shortly)
  - Can be used when we wish to estimate the unknown mean and unknown precision of a Gaussian
  - Note: Its multivariate version is the Normal-Wishart (for multivariate mean and precision matrix)
Normal-gamma (Gaussian-gamma) Distribution

- We saw that the conjugate prior needed to have the form

  \[ p(\mu, \lambda) \propto \left[ \lambda^{1/2} \exp \left( -\frac{\lambda \mu^2}{2} \right) \right]^{\kappa_0} \exp \left[ \lambda \mu c - \lambda d \right] \]

1 shape-rate parametrization assumed for the gamma
Normal-gamma (Gaussian-gamma) Distribution

We saw that the conjugate prior needed to have the form

\[ p(\mu, \lambda) \propto \left[ \lambda^{1/2} \exp \left( -\frac{\lambda \mu^2}{2} \right) \right]^{\kappa_0} \exp [\lambda \mu c - \lambda d] \]

\[ = \exp \left[ -\frac{\kappa_0 \lambda}{2} (\mu - c/\kappa_0)^2 \right] \lambda^{\kappa_0/2} \exp \left[ - \left( d - \frac{c^2}{2\kappa_0} \right) \lambda \right] \]

(re-arranging terms)

\[ \propto \text{prop. to a Gaussian} \]

\[ \propto \text{prop. to a gamma} \]

---

1 shape-rate parametrization assumed for the gamma
Normal-gamma (Gaussian-gamma) Distribution

We saw that the conjugate prior needed to have the form

\[ p(\mu, \lambda) \propto \left[ \lambda^{1/2} \exp \left( -\frac{\lambda \mu^2}{2} \right) \right]^{\kappa_0} \exp [\lambda \mu c - \lambda d] \]

\begin{align*}
= \exp \left[ -\frac{\kappa_0 \lambda}{2} \left( \mu - c/\kappa_0 \right)^2 \right] \lambda^{\kappa_0/2} \exp \left[ - \left( d - \frac{c^2}{2\kappa_0} \right) \lambda \right]
\end{align*}

(prop. to a Gaussian) (prop. to a gamma)

The above is product of a normal and a gamma distribution\(^1\)

\[ p(\mu, \lambda) = \mathcal{N}(\mu | \mu_0, (\kappa_0 \lambda)^{-1}) \Gamma(\lambda | \alpha_0, \beta_0) = \text{NG}(\mu_0, \kappa_0, \alpha_0, \beta_0) \]

where \(\mu_0 = c/\kappa_0\), \(\alpha_0 = 1 + \kappa_0/2\), \(\beta_0 = d - c^2/2\kappa_0\) are prior's hyperparameters

---

\(^1\) shape-rate parametrization assumed for the gamma
Normal-gamma (Gaussian-gamma) Distribution

- We saw that the conjugate prior needed to have the form

\[ p(\mu, \lambda) \propto \left[ \lambda^{1/2} \exp \left( -\frac{\lambda \mu^2}{2} \right) \right]^{\kappa_0} \exp [\lambda \mu c - \lambda d] \]

\[ = \exp \left[ -\frac{\kappa_0 \lambda}{2} (\mu - c/\kappa_0)^2 \right] \lambda^{\kappa_0/2} \exp \left[ -\left( d - \frac{c^2}{2\kappa_0} \right) \lambda \right] \] (re-arranging terms)

- The above is product of a normal and a gamma distribution

\[ p(\mu, \lambda) = \mathcal{N}(\mu | \mu_0, (\kappa_0 \lambda)^{-1}) \Gamma(\lambda | \alpha_0, \beta_0) = \text{NG}(\mu_0, \kappa_0, \alpha_0, \beta_0) \]

where \( \mu_0 = c/\kappa_0, \alpha_0 = 1 + \kappa_0/2, \beta_0 = d - c^2/2\kappa_0 \) are prior's hyperparameters

- \( p(\mu, \lambda) = \text{NG}(\mu_0, \kappa_0, \alpha_0, \beta_0) \) is a conjugate for the mean-precision pair \((\mu, \lambda)\)
Normal-gamma (Gaussian-gamma) Distribution

- We saw that the conjugate prior needed to have the form

\[ p(\mu, \lambda) \propto \left[ \lambda^{1/2} \exp \left( -\frac{\lambda \mu^2}{2} \right) \right]^{\kappa_0} \exp [\lambda \mu c - \lambda d] \]

\[ = \exp \left[ -\frac{\kappa_0 \lambda}{2} (\mu - c/\kappa_0)^2 \right] \lambda^{\kappa_0/2} \exp \left[ -\left( d - \frac{c^2}{2 \kappa_0} \right) \lambda \right] \]

(prop. to a Gaussian \[ \text{prop. to a gamma} \])

(re-arranging terms)

The above is product of a normal and a gamma distribution\(^1\)

\[ p(\mu, \lambda) = \mathcal{N}(\mu | \mu_0, (\kappa_0 \lambda)^{-1}) \Gamma(\lambda | \alpha_0, \beta_0) = \text{NG}(\mu_0, \kappa_0, \alpha_0, \beta_0) \]

where \( \mu_0 = c/\kappa_0, \alpha_0 = 1 + \kappa_0/2, \beta_0 = d - c^2/2\kappa_0 \) are prior’s hyperparameters

- \( p(\mu, \lambda) = \text{NG}(\mu_0, \kappa_0, \alpha_0, \beta_0) \) is a conjugate for the mean-precision pair \((\mu, \lambda)\)

- A useful prior in many problems involving Gaussians with unknown mean and precision

\(^1\) shape-rate parametrization assumed for the gamma
Joint Posterior

- Due to conjugacy, the joint posterior $p(\mu, \lambda | X)$ will also be normal-gamma

$$p(\mu, \lambda | X) \propto p(X | \mu, \lambda) p(\mu, \lambda)$$

\[\sum_{n=1}^{N} (x_n - \bar{x})^2 + \kappa_0 N (\bar{x} - \mu_0)^2 \]

For full derivation, refer to “Conjugate Bayesian analysis of the Gaussian distribution” - Murphy (2007)
Joint Posterior

- Due to conjugacy, the joint posterior $p(\mu, \lambda|X)$ will also be normal-gamma

$$p(\mu, \lambda|X) \propto p(X|\mu, \lambda)p(\mu, \lambda)$$

- Plugging in the expressions for $p(X|\mu, \lambda)$ and $p(\mu, \lambda)$, we get

$$p(\mu, \lambda|X) = \text{NG}(\mu_N, \kappa_N, \alpha_N, \beta_N) = \mathcal{N}(\mu|\mu_N, (\kappa_N\lambda)^{-1})\Gamma(\lambda|\alpha_N, \beta_N)$$

\[\text{For full derivation, refer to “Conjugate Bayesian analysis of the Gaussian distribution” - Murphy (2007)}\]
Joint Posterior

- Due to conjugacy, the joint posterior $p(\mu, \lambda | X)$ will also be normal-gamma

  $$p(\mu, \lambda | X) \propto p(X | \mu, \lambda) p(\mu, \lambda)$$

- Plugging in the expressions for $p(X | \mu, \lambda)$ and $p(\mu, \lambda)$, we get

  $$p(\mu, \lambda | X) = \text{NG}(\mu_N, \kappa_N, \alpha_N, \beta_N) = \mathcal{N}(\mu | \mu_N, (\kappa_N \lambda)^{-1}) \text{Gamma}(\lambda | \alpha_N, \beta_N)$$

  where the updated posterior hyperparameters are given by\(^2\)

  $$\begin{align*}
  \mu_N &= \frac{\kappa_0 \mu_0 + N \bar{x}}{\kappa_0 + N} \\
  \kappa_N &= \kappa_0 + N \\
  \alpha_N &= \alpha_0 + N/2 \\
  \beta_N &= \beta_0 + \frac{1}{2} \sum_{n=1}^{N} (x_n - \bar{x})^2 + \frac{\kappa_0 N (\bar{x} - \mu_0)^2}{2(\kappa_0 + N)}
  \end{align*}$$

\(^2\)For full derivation, refer to “Conjugate Bayesian analysis of the Gaussian distribution” - Murphy (2007)
Other Quantities of Interest

- Already saw that joint post. \( p(\mu, \lambda | X) = \text{NG}(\mu_N, \kappa_N, \alpha_N, \beta_N) = \mathcal{N}(\mu | \mu_N, (\kappa_N \lambda)^{-1}) \Gamma(\lambda | \alpha_N, \beta_N) \)

---

3 For full derivations, refer to "Conjugate Bayesian analysis of the Gaussian distribution" - Murphy (2007)
Other Quantities of Interest

- Already saw that joint post. $p(\mu, \lambda | X) = \text{NG}(\mu_N, \kappa_N, \alpha_N, \beta_N) = \mathcal{N}(\mu | \mu_N, (\kappa_N \lambda)^{-1}) \text{Gamma}(\lambda | \alpha_N, \beta_N)$

- Marginal posteriors for $\mu$ and $\lambda$

$$p(\lambda | X) = \int p(\mu, \lambda | X) d\mu = \text{Gamma}(\lambda | \alpha_N, \beta_N)$$
Other Quantities of Interest

- Already saw that joint post. \( p(\mu, \lambda | X) = \text{NG}(\mu_N, \kappa_N, \alpha_N, \beta_N) = N(\mu | \mu_N, (\kappa_N \lambda)^{-1}) \Gamma(\lambda | \alpha_N, \beta_N) \)

- Marginal posteriors for \( \mu \) and \( \lambda \)
  \[
  p(\lambda | X) = \int p(\mu, \lambda | X) d\mu = \Gamma(\lambda | \alpha_N, \beta_N) \\
  p(\mu | X) = \int p(\mu, \lambda | X) d\lambda = \int p(\mu | \lambda, X) p(\lambda | X) d\lambda = t_{2\alpha_N} (\mu | \mu_N, \beta_N / (\alpha_N \kappa_N)) \]

For full derivations, refer to "Conjugate Bayesian analysis of the Gaussian distribution" - Murphy (2007)
Other Quantities of Interest\(^3\)

- Already saw that joint post. \(p(\mu, \lambda|X) = \text{NG}(\mu_N, \kappa_N, \alpha_N, \beta_N) = \mathcal{N}(\mu|\mu_N, (\kappa_N \lambda)^{-1})\Gamma(\lambda|\alpha_N, \beta_N)\)

- Marginal posteriors for \(\mu\) and \(\lambda\)
  
  \[
  p(\lambda|X) = \int p(\mu, \lambda|X) d\mu = \Gamma(\lambda|\alpha_N, \beta_N)
  \]
  
  \[
  p(\mu|X) = \int p(\mu, \lambda|X) d\lambda = \int p(\mu|\lambda, X)p(\lambda|X) d\lambda = \frac{t_{2\alpha_N}(\mu|\mu_N, \beta_N/(\alpha_N \kappa_N))}{t \text{ distribution}}
  \]

- Exercise: What will be the conditional posteriors \(p(\mu|\lambda, X)\) and \(p(\lambda|\mu, X)\)?

\(^3\) For full derivations, refer to "Conjugate Bayesian analysis of the Gaussian distribution" - Murphy (2007)
Other Quantities of Interest

- Already saw that joint post. $p(\mu, \lambda | X) = \text{NG}(\mu_N, \kappa_N, \alpha_N, \beta_N) = \mathcal{N}(\mu | \mu_N, (\kappa_N \lambda)^{-1}) \text{Gamma}(\lambda | \alpha_N, \beta_N)$
- Marginal posteriors for $\mu$ and $\lambda$
  
  $$p(\lambda | X) = \int p(\mu, \lambda | X) d\mu = \text{Gamma}(\lambda | \alpha_N, \beta_N)$$
  
  $$p(\mu | X) = \int p(\mu, \lambda | X) d\lambda = \int p(\mu | \lambda, X)p(\lambda | X) d\lambda = \frac{t_{2\alpha_N}(\mu | \mu_N, \beta_N/(\alpha_N \kappa_N))}{t_{2\alpha_N}(\mu | \mu_N, \beta_N/(\alpha_N \kappa_N))}$$

- Exercise: What will be the conditional posteriors $p(\mu | \lambda, X)$ and $p(\lambda | \mu, X)$?
- Marginal likelihood of the model

$$p(X) = \frac{\Gamma(\alpha_N)}{\Gamma(\alpha_0)} \frac{\beta_0^{\alpha_0}}{\beta_N^{\alpha_N}} \left( \frac{\kappa_0}{\kappa_N} \right)^{\frac{1}{2}} (2\pi)^{-N/2}$$

---

3 For full derivations, refer to "Conjugate Bayesian analysis of the Gaussian distribution" - Murphy (2007)
Other Quantities of Interest\(^3\)

- Already saw that joint post.  \( p(\mu, \lambda | X) = \text{NG}(\mu_N, \kappa_N, \alpha_N, \beta_N) = \mathcal{N}(\mu | \mu_N, (\kappa_N \lambda)^{-1}) \text{Gamma}(\lambda | \alpha_N, \beta_N) \)

- Marginal posteriors for \( \mu \) and \( \lambda \)

\[
\begin{align*}
p(\lambda | X) &= \int p(\mu, \lambda | X) d\mu = \text{Gamma}(\lambda | \alpha_N, \beta_N) \\
p(\mu | X) &= \int p(\mu, \lambda | X) d\lambda = \int p(\mu | \lambda, X)p(\lambda | X) d\lambda = \frac{\text{t}_2}{\pi} \left( \frac{\alpha_N}{\kappa_N + 1} \right)
\end{align*}
\]

- Exercise: What will be the conditional posteriors \( p(\mu | \lambda, X) \) and \( p(\lambda | \mu, X) \)?

- Marginal likelihood of the model

\[
p(X) = \frac{\Gamma(\alpha_N)}{\Gamma(\alpha_0)} \frac{\beta_0^{\alpha_0}}{\beta_0^{\alpha_N}} \left( \frac{\kappa_0}{\kappa_N} \right)^{-\frac{1}{2}} (2\pi)^{-N/2}
\]

- Posterior predictive distribution of a new observation \( x_* \)

\[
p(x_* | X) = \int p(x_* | \mu, \lambda) p(\mu, \lambda | X) d\mu d\lambda = \frac{\beta_N (\kappa_N + 1)}{\alpha_N \kappa_N} \text{Normal-Gamma}
\]

\(^3\) For full derivations, refer to "Conjugate Bayesian analysis of the Gaussian distribution" - Murphy (2007)
An Aside: general-t and Student-t distribution

- Equivalent to an infinite sum of Gaussian distributions, with same means but different precisions

\[
p(x|\mu, a, b) = \int \mathcal{N}(x|\mu, \lambda^{-1}) \text{Gamma}(\lambda|a, b) d\lambda
\]

\[
= t_{2a}(x|\mu, b/a) = t_{\nu}(x|\mu, \sigma^2) \quad \text{(general-t distribution)}
\]

\[
\mu = 0, \sigma^2 = 1 \text{ gives the Student-t distribution} \quad (t_{\nu}(x|\mu, \sigma^2))
\]

Also a useful prior for sparse modeling

An illustration of student-t distribution has a "fatter" tail than a Gaussian and also sharper around the mean
An Aside: general-t and Student-t distribution

- Equivalent to an infinite sum of Gaussian distributions, with same means but different precisions:

\[ p(x|\mu, a, b) = \int \mathcal{N}(x|\mu, \lambda^{-1})\text{Gamma}(\lambda|a, b)d\lambda \]

\[ = t_{2a}(x|\mu, b/a) = t_\nu(x|\mu, \sigma^2) \quad \text{(general-t distribution)} \]

- \( \mu = 0, \sigma^2 = 1 \) gives the Student-t distribution (\( t_\nu \)). Note: If \( x \sim t_\nu(\mu, \sigma^2) \) then \( \frac{x-\mu}{\sigma} \sim t_\nu \)
An Aside: general-t and Student-t distribution

- Equivalent to an infinite sum of Gaussian distributions, with same means but different precisions

\[
p(x|\mu, a, b) = \int \mathcal{N}(x|\mu, \lambda^{-1})\text{Gamma}(\lambda|a, b)d\lambda
\]

\[
= t_{2a}(x|\mu, b/a) = t_\nu(x|\mu, \sigma^2) \quad \text{(general-t distribution)}
\]

- \(\mu = 0, \sigma^2 = 1\) gives the Student-t distribution \((t_\nu)\). Note: If \(x \sim t_\nu(\mu, \sigma^2)\) then \(\frac{x-\mu}{\sigma} \sim t_\nu\)

- An illustration of student-t

![An illustration of student-t distribution](image_url)
An Aside: general-t and Student-t distribution

- Equivalent to an infinite sum of Gaussian distributions, with same means but different precisions

\[
p(x|\mu, a, b) = \int \mathcal{N}(x|\mu, \lambda^{-1}) \text{Gamma}(\lambda|a, b) d\lambda
\]

\[
= \ t_2a(x|\mu, b/a) = t_{\nu}(x|\mu, \sigma^2) \quad \text{(general-t distribution)}
\]

- \(\mu = 0, \sigma^2 = 1\) gives the Student-t distribution (\(t_\nu\)). Note: If \(x \sim t_\nu(\mu, \sigma^2)\) then \(\frac{x - \mu}{\sigma} \sim t_\nu\)

- An illustration of student-t

\[\nu \rightarrow \infty, \quad \nu = 1.0, \quad \nu = 0.1\]

- \(t\) distribution has a “fatter” tail than a Gaussian and also sharper around the mean
An Aside: general-t and Student-t distribution

- Equivalent to an infinite sum of Gaussian distributions, with same means but different precisions

\[ p(x|\mu, a, b) = \int \mathcal{N}(x|\mu, \lambda^{-1}) \Gamma(\lambda|a, b) d\lambda \]

\[ = t_{2a}(x|\mu, b/a) = t_{\nu}(x|\mu, \sigma^2) \quad \text{(general-t distribution)} \]

- \( \mu = 0, \sigma^2 = 1 \) gives the Student-t distribution \( (t_{\nu}) \). Note: If \( x \sim t_{\nu}(\mu, \sigma^2) \) then \( \frac{x-\mu}{\sigma} \sim t_{\nu} \)

- An illustration of student-t

\[ X \sim t_{\nu}(\mu, \sigma^2) \]

- \( t \) distribution has a “fatter” tail than a Gaussian and also sharper around the mean

- Also a useful prior for sparse modeling
We only considered the simple 1-D Gaussian distribution.

The approach also extends to inferring parameters of a multivariate Gaussian. For the unknown mean and precision matrix, normal-Wishart distribution can be used as prior. Posterior updates have forms similar to that in the 1-D case. When working with mean-variance, we can use normal-inverse gamma as conjugate prior (or normal-inverse Wishart when working with mean-covariance matrix in case of multivariate Gaussian distribution). Other priors can also be used as well when inferring parameters of Gaussians, e.g., normal-Inverse $\chi^2$ distribution is commonly used in Statistics community for scalar mean-variance. Uniform priors can also be used. Look at BDA Chapter 3 for such examples. Also refer to “Conjugate Bayesian analysis of the Gaussian distribution” - Murphy (2007) for various examples and more detailed derivations.
We only considered the simple 1-D Gaussian distribution

The approach also extends to inferring parameters of a multivariate Gaussian
Inferring Parameters of Gaussian: Some Other Cases

- We only considered the simple 1-D Gaussian distribution.
- The approach also extends to inferring parameters of a multivariate Gaussian.
  - For the unknown mean and precision matrix, normal-Wishart distribution can be used as prior.
Inferring Parameters of Gaussian: Some Other Cases

- We only considered the simple 1-D Gaussian distribution.
- The approach also extends to inferring parameters of a multivariate Gaussian.
  - For the unknown mean and precision matrix, normal-Wishart distribution can be used as prior.
- Posterior updates have forms similar to that in the 1-D case.
Inferring Parameters of Gaussian: Some Other Cases

- We only considered the simple 1-D Gaussian distribution.
- The approach also extends to inferring parameters of a multivariate Gaussian.
  - For the unknown mean and precision matrix, normal-Wishart distribution can be used as prior.
- Posterior updates have forms similar to that in the 1-D case.
- When working with mean-variance, we can use normal-inverse gamma as conjugate prior (or normal-inverse Wishart when working with mean-covariance matrix in case of multivariate Gaussian distribution).

Other priors can also be used as well when inferring parameters of Gaussians, e.g., normal-Inverse $\chi^2$ distribution is commonly used in Statistics community for scalar mean-variance. Uniform priors can also be used.

Look at BDA Chapter 3 for such examples. Also refer to "Conjugate Bayesian analysis of the Gaussian distribution" - Murphy (2007) for various examples and more detailed derivations.
Inferring Parameters of Gaussian: Some Other Cases

- We only considered the simple 1-D Gaussian distribution
- The approach also extends to inferring parameters of a multivariate Gaussian
  - For the unknown mean and precision matrix, normal-Wishart distribution can be used as prior
- Posterior updates have forms similar to that in the 1-D case
- When working with mean-variance, we can use normal-inverse gamma as conjugate prior (or normal-inverse Wishart when working with mean-covariance matrix in case of multivariate Gaussian distribution)
- Other priors can also be used as well when inferring parameters of Gaussians
Inferring Parameters of Gaussian: Some Other Cases

- We only considered the simple 1-D Gaussian distribution.
- The approach also extends to inferring parameters of a multivariate Gaussian.
  - For the unknown mean and precision matrix, **normal-Wishart** distribution can be used as prior.
- Posterior updates have forms similar to that in the 1-D case.
- When working with mean-variance, we can use **normal-inverse gamma** as conjugate prior (or **normal-inverse Wishart** when working with mean-covariance matrix in case of multivariate Gaussian distribution).
- Other priors can also be used as well when inferring parameters of Gaussians, e.g.,
  - normal-Inverse $\chi^2$ distribution is commonly used in Statistics community for scalar mean-variance.
Inferring Parameters of Gaussian: Some Other Cases

- We only considered the simple 1-D Gaussian distribution.
- The approach also extends to inferring parameters of a multivariate Gaussian.
  - For the unknown mean and precision matrix, normal-Wishart distribution can be used as prior.
- Posterior updates have forms similar to that in the 1-D case.
- When working with mean-variance, we can use normal-inverse gamma as conjugate prior (or normal-inverse Wishart when working with mean-covariance matrix in case of multivariate Gaussian distribution).
- Other priors can also be used as well when inferring parameters of Gaussians, e.g.,
  - normal-Inverse $\chi^2$ distribution is commonly used in Statistics community for scalar mean-variance.
  - Uniform priors can also be used.
Inferring Parameters of Gaussian: Some Other Cases

- We only considered the simple 1-D Gaussian distribution
- The approach also extends to inferring parameters of a multivariate Gaussian
  - For the unknown mean and precision matrix, normal-Wishart distribution can be used as prior
- Posterior updates have forms similar to that in the 1-D case
- When working with mean-variance, we can use normal-inverse gamma as conjugate prior (or normal-inverse Wishart when working with mean-covariance matrix in case of multivariate Gaussian distribution)
- Other priors can also be used as well when inferring parameters of Gaussians, e.g.,
  - normal-Inverse $\chi^2$ distribution is commonly used in Statistics community for scalar mean-variance
  - Uniform priors can also be used
  - Look at BDA Chapter 3 for such examples
Inferring Parameters of Gaussian: Some Other Cases

- We only considered the simple 1-D Gaussian distribution
- The approach also extends to inferring parameters of a multivariate Gaussian
  - For the unknown mean and precision matrix, normal-Wishart distribution can be used as prior
- Posterior updates have forms similar to that in the 1-D case
- When working with mean-variance, we can use normal-inverse gamma as conjugate prior (or normal-inverse Wishart when working with mean-covariance matrix in case of multivariate Gaussian distribution)
  - Other priors can also be used as well when inferring parameters of Gaussians, e.g.,
    - normal-Inverse $\chi^2$ distribution is commonly used in Statistics community for scalar mean-variance
    - Uniform priors can also be used
    - Look at BDA Chapter 3 for such examples
  - Also refer to “Conjugate Bayesian analysis of the Gaussian distribution” - Murphy (2007) for various examples and more detailed derivations
Next Class: More examples of Bayesian inference with Gaussians