# Estimating Parameters and Predictive Distributions: Some Simple Cases

CS772A: Probabilistic Machine Learning

Piyush Rai

# Plan today

- Parameter estimation (point est. and posterior) and predictive distribution for
  - Bernoulli observation model (binary-valued observations)
  - Multinoulli observation model (discrete-valued observations)
- Focus today on cases with conjugate prior on parameters (easy to compute posterior)
- Gaussian distribution and some of its important properties
- Parameter estimation and predictive distribution for Gaussian observation models

# Bernoulli Observation Model

# Estimating a Coin's Bias

- Consider a sequence of $N$ coin toss outcomes (observations)

Probability of a head

- Each observation $y_n$ is a binary random variable. Head: $y_n = 1$, Tail: $y_n = 0$
- Each $y_n$ is assumed generated by a **Bernoulli distribution** with param $\theta \in (0,1)$

Likelihood or observation model

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

- Here $\theta$ the unknown param (probability of head). Let's do MLE

assuming i.i.d. data

- Log-likelihood: $\sum_{n=1}^{N} \log p(y_n|\theta) = \sum_{n=1}^{N} [y_n \log \theta + (1-y_n)\log(1-\theta)]$
- Maximizing log-lik, or minimizing neg. log-lik (NLL) w.r.t. $\theta$ gives

I tossed a coin 5 times – gave 1 head and 4 tails. Does it means $\theta$ = 0.2?? The MLE approach says so. What is I see 0 head and 5 tails. Does it mean $\theta$ = 0?

$$\theta_{MLE} = \frac{\sum_{n=1}^{N} y_n}{N}$$

Thus MLE solution is simply the fraction of heads! ☺ Makes intuitive sense!

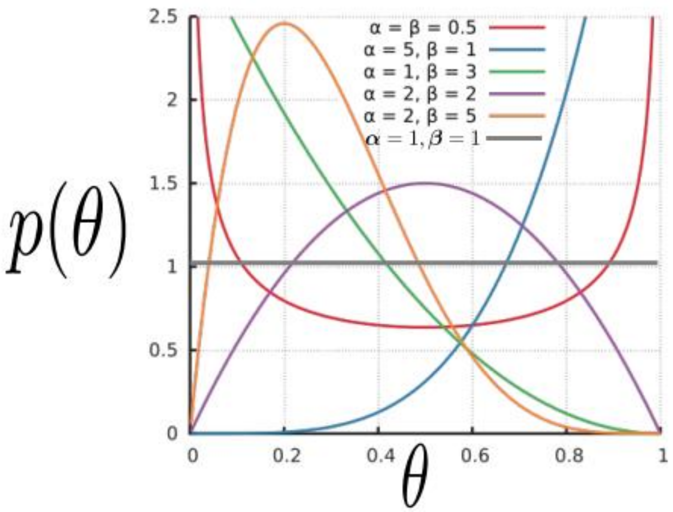
Indeed, with a small number of training observations, MLE may overfit and may not be reliable. An alternative is MAP estimation which can incorporate a prior distribution over $\theta$

# Estimating a Coin's Bias

- Let's do MAP estimation for the bias of the coin
- Each likelihood term is Bernoulli

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

- Also need a prior since we want to do MAP estimation
- Since $\theta \in (0,1)$, a reasonable choice of prior for $\theta$ would be Beta distribution




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

The gamma function

$\alpha$ and $\beta$ (both non-negative reals) are the two hyperparameters of this Beta prior

Using $\alpha = 1$ and $\beta = 1$ will make the Beta prior a uniform prior

Can set these based on intuition, cross-validation, or even learn them

# Estimating a Coin's Bias

- The log posterior for the coin-toss model is log-lik + log-prior

$$LP(\theta) = \sum_{n=1}^{N} \log p(y_n|\theta) + \log p(\theta|\alpha,\beta)$$

- Plugging in the expressions for Bernoulli and Beta and ignoring any terms that don't depend on $\boldsymbol{\theta}$, the log posterior simplifies to

$$LP(\theta) = \sum_{n=1}^{N} [y_n \log\theta + (1-y_n)\log(1-\theta)] + (\alpha-1)\log\theta + (\beta-1)\log(1-\theta)$$

- Maximizing the above log post. (or min. of its negative) w.r.t. $\boldsymbol{\theta}$ gives

$$\theta_{MAP} = \frac{\sum_{n=1}^{N} y_n + \alpha - 1}{N + \alpha + \beta - 2}$$

Using $\boldsymbol{\alpha} = 1$ and $\boldsymbol{\beta} = 1$ gives us the same solution as MLE

Recall that $\boldsymbol{\alpha} = 1$ and $\boldsymbol{\beta} = 1$ for Beta distribution is in fact equivalent to a uniform prior (hence making MAP equivalent to MLE)

Prior's hyperparameters have an interesting interpretation. Can think of $\alpha - 1$ and $\beta - 1$ as the number of heads and tails, respectively, before starting the coin-toss experiment (akin to "pseudo-observations")

Such interpretations of prior's hyperparameters as being "pseudo-observations" exist for various other prior distributions as well (in particular, distributions belonging to "exponential family" of distributions

# The Posterior Distribution

- Let's do fully Bayesian inference and compute the posterior distribution
- Bernoulli likelihood: $p(y_n|\theta) = \text{Bernoulli}(y_n|\theta) = \theta^{y_n}(1-\theta)^{1-y_n}$
- Beta prior: $p(\theta) = \text{Beta}(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
- The posterior can be computed as

Hyperparams $\alpha, \beta$ not shown for brevity

$$p(\theta|\boldsymbol{y}) = \frac{p(\theta)p(\boldsymbol{y}|\theta)}{p(\boldsymbol{y})} = \frac{p(\theta)\prod_{n=1}^{N} p(y_n|\theta)}{p(\boldsymbol{y})} = \frac{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\prod_{n=1}^{N}\theta^{y_n}(1-\theta)^{1-y_n}}{\int \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\prod_{n=1}^{N}\theta^{y_n}(1-\theta)^{1-y_n}d\theta}$$

$\theta^{\sum_{n=1}^{N} y_n}(1-\theta)^{N-\sum_{n=1}^{N} y_n}$

Number of heads $(N_1)$

Number of tails $(N_0)$

- Here, even without computing the denominator (marg lik), we can identify the posterior
  - It is Beta distribution since $p(\theta|\boldsymbol{y}) \propto \theta^{\alpha+N_1-1}(1-\theta)^{\beta+N_0-1}$
  - Thus $p(\theta|\boldsymbol{y}) = \text{Beta}(\theta|\alpha+N_1, \beta+N_0)$

Hint: Use the fact that the posterior must integrate to 1 $\int p(\theta|\boldsymbol{y})d\theta = 1$

Exercise: Show that the normalization constant equals $\frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+\sum_{n=1}^{N} y_n)\Gamma(\beta+N-\sum_{n=1}^{N} y_n)}$

- Here, finding the posterior boiled down to simply "multiply, add stuff, and identify"
- Here, posterior has the same form as prior (both Beta): property of conjugate priors

# Conjugacy and Conjugate Priors

- Many pairs of distributions are conjugate to each other
  - Bernoulli (likelihood) + Beta (prior) ⇒ Beta posterior
  - Binomial (likelihood) + Beta (prior) ⇒ Beta posterior
  - Multinomial (likelihood) + Dirichlet (prior) ⇒ Dirichlet posterior
  - Poisson (likelihood) + Gamma (prior) ⇒ Gamma posterior
  - Gaussian (likelihood) + Gaussian (prior) ⇒ Gaussian posterior
  - and many other such pairs ..

Not true in general, but in some cases (e.g., the variance of the Gaussian likelihood is fixed)

- Tip: If two distr are conjugate to each other, their functional forms are similar
  - Example: Bernoulli and Beta have the forms

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

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

This is why, when we multiply them while computing the posterior, the exponents get added and we get the same form for the posterior as the prior but with just updated hyperparameter. Also, we can identify the posterior and its hyperparameters simply by inspection

- More on conjugate priors when we look at exponential family distributions

# Predictive Distribution

- Suppose we want to compute the prob that the next outcome $y_{N+1}$ will be head (=1)
- The posterior predictive distribution (averaging over all $\theta$'s weighted by their respective posterior probabilities)

$$p(y_{N+1}=1|\boldsymbol{y}) = \int_0^1 p(y_{N+1}=1,\theta|\boldsymbol{y})\,d\theta = \int_0^1 p(y_{N+1}=1|\theta)p(\theta|\boldsymbol{y})\,d\theta$$

$$= \int_0^1 \theta \times p(\theta|\boldsymbol{y})\,d\theta$$

$$= \mathbb{E}_{p(\theta|\boldsymbol{y})}[\theta]$$

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

Expectation of $\theta$ w.r.t. the Beta posterior distribution $p(\theta|\boldsymbol{y}) = \text{Beta}(\theta|\alpha+N_1,\beta+N_0)$

- Therefore the PPD will be

$$p(y_{N+1}|\boldsymbol{y}) = \text{Bernoulli}(y_{N+1}|\mathbb{E}_{p(\theta|\boldsymbol{y})}[\theta])$$

For models where likelihood and prior are conjugate to each other, the PPD can be computed easily in closed form (more on this when we talk about exponential family distributions)

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

$$p(y_{N+1}=1|\boldsymbol{y}) \approx p(y_{N+1}=1|\hat{\theta}) = \hat{\theta} \quad \Longrightarrow \quad p(y_{N+1}|\boldsymbol{y}) = \text{Bernoulli}(y_{N+1}|\hat{\theta})$$

# Multinoulli Observation Model

# The Posterior Distribution

MLE/MAP left as an exercise

- Assume $N$ discrete obs $\boldsymbol{y} = \{y_1, y_2, \ldots, y_N\}$ with each $y_n \in \{1,2,\ldots,K\}$, e.g.,
  - $y_n$ represents the outcome of a dice roll with $K$ faces
  - $y_n$ represents the class label of the $n^{th}$ example in a classification problem (total $K$ classes)
  - $y_n$ represents the identity of the $n^{th}$ word in a sequence of words
- Assume likelihood to be multinoulli with unknown params $\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_K]$

These sum to 1

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

Generalization of Bernoulli to $K > 2$ discrete outcomes

- $\boldsymbol{\pi}$ is a vector of probabilities ("probability vector"), e.g.,
  - Biases of the $K$ sides of the dice
  - Prior class probabilities in multi-class classification $(p(y_n = k) = \pi_k)$
  - Probabilities of observing each word of the $K$ words in a vocabulary
- Assume a conjugate prior (Dirichlet) on $\boldsymbol{\pi}$ with hyperparams $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_K]$

Called the concentration parameter of the Dirichlet (assumed known for now)

Large values of $\alpha$ will give a Dirichlet peaked around its mean (next slides illustrates this)

Each $\alpha_k \geq 0$

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

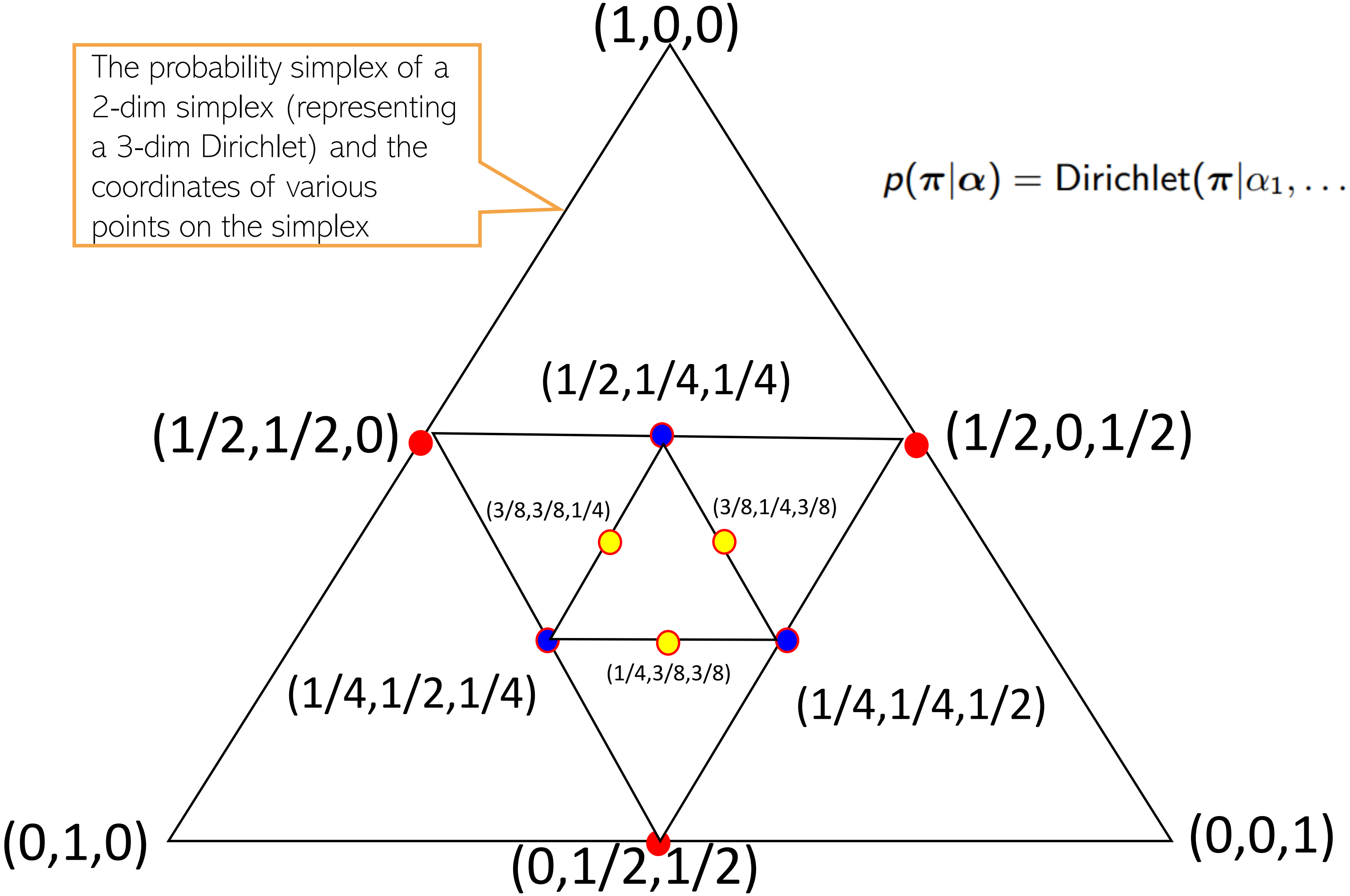
Generalization of Beta to $K$-dimensional probability vectors

# Brief Detour: Dirichlet Distribution

Basically, probability vectors

- An important distribution. Models non-neg. vectors $\boldsymbol{\pi}$ that also sum to one
- A random draw from $K$-dim Dirich. will be a point under $(K\text{-}1)$-dim probability simplex

The probability simplex of a 2-dim simplex (representing a 3-dim Dirichlet) and the coordinates of various points on the simplex




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

$$\text{Mean} = \left[\frac{\alpha_1}{\sum_{k=1}^K \alpha_k}, \ldots, \frac{\alpha_K}{\sum_{k=1}^K \alpha_k}\right]$$

$$\text{Mode} = \left[\frac{\alpha_1 - 1}{\sum_{k=1}^K \alpha_k - K}, \ldots, \frac{\alpha_K - 1}{\sum_{k=1}^K \alpha_k - K}\right] \quad (\alpha_k > 1)$$

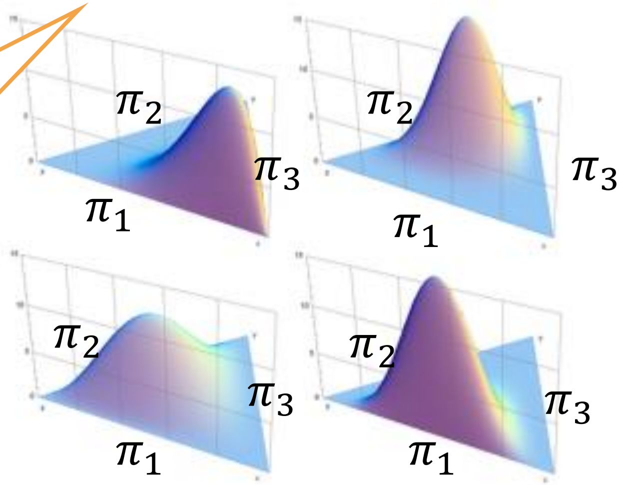
$$\text{var}(\pi_k) = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)} \qquad \alpha_0 = \sum_{k=1}^K \alpha_k$$

# Brief Detour: Dirichlet Distribution

- A visualization of Dirichlet distribution for different values of concentration param

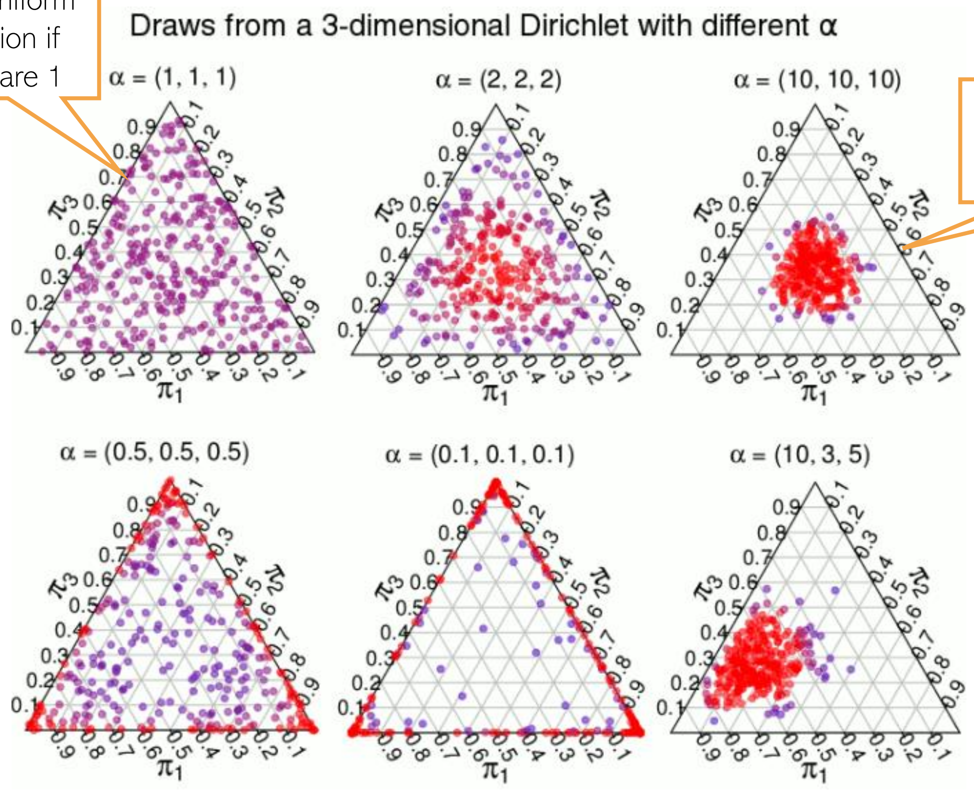
Visualizations of PDFs of some 3-dim Dirichlet distributions (each generated using a different conc. Param vector $\boldsymbol{\alpha}$)

$\boldsymbol{\alpha}$ controls the shape of the Dirichlet (just like Beta distribution's hyperparameters)




Like a uniform distribution if all $\alpha_k$'s are 1

All $\alpha_k$'s large results in peak around the center of the simplex




- Interesting fact: Can generate a $K$-dim Dirichlet random variable by independently generating $K$ gamma random variables and normalizing them to sum to 1

# The Posterior Distribution

- Posterior $p(\boldsymbol{\pi}|\boldsymbol{y})$ is easy to compute due to conjugacy b/w multinoulli and Dir.

Likelihood

Prior

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

Don't need to compute for this case because of conjugacy

Marg-lik = $\int p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\boldsymbol{y}|\boldsymbol{\pi})d\boldsymbol{\pi}$

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

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

- Even without computing marg-lik, $p(\boldsymbol{y}|\boldsymbol{\alpha})$, we can see that the posterior is Dirichlet
- Denoting $N_k = \sum_{n=1}^{N} \mathbb{I}[y_n = k]$, number of observations with with value $k$

$$p(\boldsymbol{\pi}|\boldsymbol{y},\boldsymbol{\alpha}) = \text{Dirichlet}(\boldsymbol{\pi}|\alpha_1 + N_1, \alpha_2 + N_2, \ldots, \alpha_K + N_K)$$

Similar to number of heads and tails for the coin bias estimation problem

- Note: $N_1, , N_2 \ldots, N_K$ are the sufficient statistics for this estimation problem
  - We only need the suff-stats to estimate the parameters and values of individual observations aren't needed (another property from exponential family of distributions – more on this later)

# The Predictive Distribution

- Finally, let's also look at the posterior predictive distribution for this model
- PPD is the prob distr of a new $y_* \in \{1,2,\dots,K\}$, given training data $\boldsymbol{y} = \{y_1, y_2, \dots, y_N\}$

Will be a multinoulli. Just need to estimate the probabilities of each of the $K$ outcomes

$$p(y_*|\boldsymbol{y},\boldsymbol{\alpha}) = \int p(y_*|\boldsymbol{\pi})p(\boldsymbol{\pi}|\boldsymbol{y},\boldsymbol{\alpha})d\boldsymbol{\pi}$$

- $p(y_*|\boldsymbol{\pi}) = \text{multinoulli}(y_*|\boldsymbol{\pi})$, $p(\boldsymbol{\pi}|\boldsymbol{y},\boldsymbol{\alpha}) = \text{Dirichlet}(\boldsymbol{\pi}|\alpha_1+N_1,\alpha_2+N_2,\dots,\alpha_K+N_K)$
- Can compute the posterior predictive probability for each of the $K$ possible outcomes

$$p(y_* = k|\boldsymbol{y},\boldsymbol{\alpha}) = \int p(y_* = k|\boldsymbol{\pi})p(\boldsymbol{\pi}|\boldsymbol{y},\boldsymbol{\alpha})d\boldsymbol{\pi}$$

$$= \int \pi_k \times \text{Dirichlet}(\boldsymbol{\pi}|\alpha_1+N_1,\alpha_2+N_2,\dots,\alpha_K+N_K)d\pi$$

$$= \frac{\alpha_k + N_k}{\sum_{k=1}^K \alpha_k + N}$$ (Expectation of $\pi_k$ w.r.t the Dirichlet posterior)

- Thus PPD is multinoulli with probability vector $\left\{\frac{\alpha_k+N_k}{\sum_{k=1}^K \alpha_k+N}\right\}_{k=1}^K$

Note how these probabilities have been "smoothened" due to the use of the prior + the averaging over the posterior

A similar effect was achieved in the Beta-Bernoulli model, too

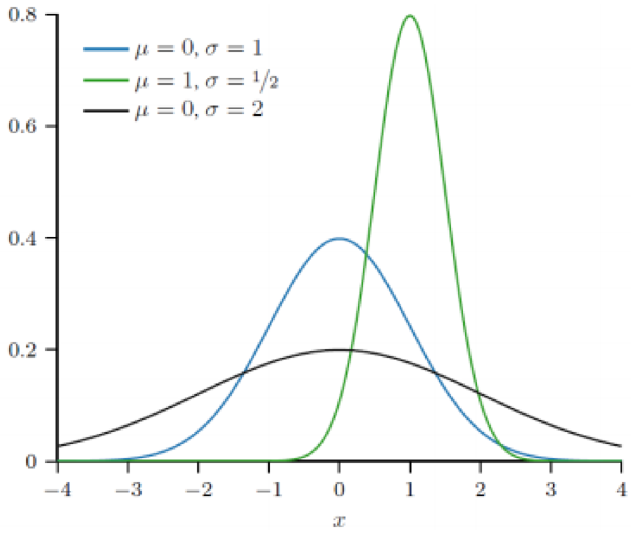
- Plug-in predictive will also be multinoulli but with prob vector given by the point estimate of $\boldsymbol{\pi}$

# Gaussian Observation Model

# Gaussian Distribution (Univariate)

- Distribution over real-valued scalar random variables $Y \in \mathbb{R}$, e.g., height of students in a class
- Defined by a scalar mean $\mu$ and a scalar variance $\sigma^2$

$$\mathcal{N}(Y = y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$$




- Mean: $\mathbb{E}[Y] = \mu$
- Variance: $\text{var}[Y] = \sigma^2$
- Inverse of variance is called precision: $\beta = \frac{1}{\sigma^2}$.

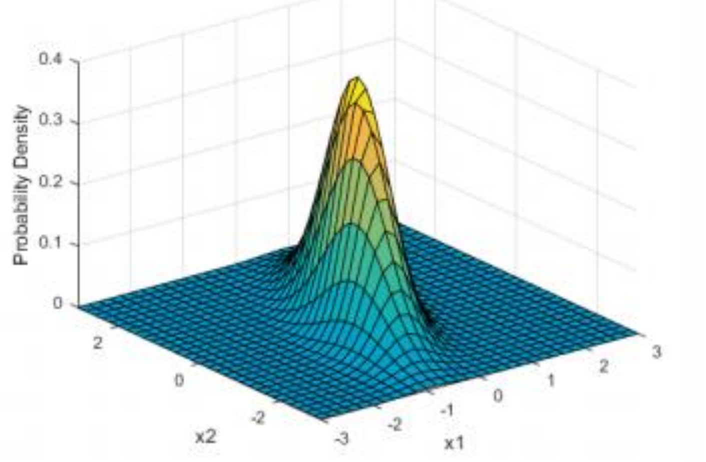
Gaussian PDF in terms of precision

$$\mathcal{N}(Y = y|\mu, \beta) = \sqrt{\frac{\beta}{2\pi}} \exp\left[-\frac{\beta}{2}(y-\mu)^2\right]$$

# Gaussian Distribution (Multivariate)

- Distribution over real-valued vector random variables $\boldsymbol{Y} \in \mathbb{R}^D$
- Defined by a mean vector $\boldsymbol{\mu} \in \mathbb{R}^D$ and a covariance matrix $\boldsymbol{\Sigma}$

A two-dimensional Gaussian

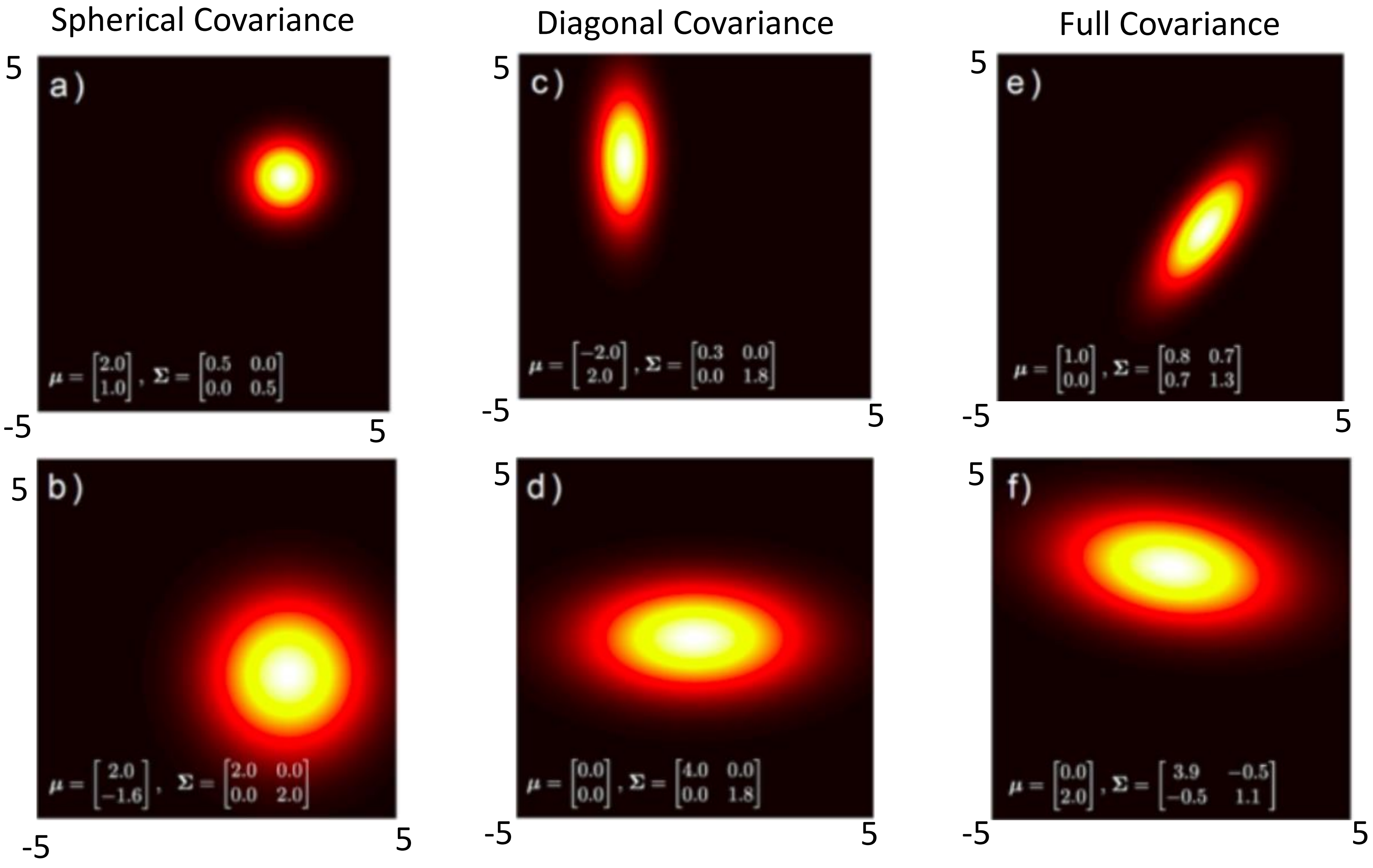



$$\mathcal{N}(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp[-(\boldsymbol{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu})]$$

- Note: The cov. matrix $\boldsymbol{\Sigma}$ must be symmetric and PSD
  - All eigenvalues are positive
  - $\boldsymbol{z}^\top \boldsymbol{\Sigma} \boldsymbol{z} \geq 0$ for any real vector $\boldsymbol{z}$

- The covariance matrix also controls the shape of the Gaussian
- Sometimes we work with precision matrix (inverse of covariance matrix) $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$

# Covariance Matrix for Multivariate Gaussian




Spherical: Equal spreads (variances) along all dimensions

Diagonal: Unequal spreads (variances) along all directions but still axis-parallel

Full: Unequal spreads (variances) along all directions and also spreads along oblique directions

# Multivariate Gaussian: Marginals and Conditionals

- Given $\boldsymbol{x}$ having multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\Sigma)$ with $\Lambda = \Sigma^{-1}$. Suppose

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}, \quad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

- The marginal distribution is simply

$$p(\boldsymbol{x}_a) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a,\Sigma_{aa})$$

- The conditional distribution is given by

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{a|b},\boldsymbol{\Lambda}_{aa}^{-1})$$

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)$$

**Thus marginals and conditionals of Gaussians are Gaussians**

# Transformation of Random Variables

- Suppose $Y = f(X) = AX + b$ be a linear function of a vector-valued r.v. $X$ ($A$ is a matrix and $b$ is a vector, both constants)
- Suppose $\mathbb{E}[X] = \mu$ and $\mathbf{cov}[X] = \Sigma$, then for the vector-valued r.v. $Y$

$$\mathbb{E}[Y] = \mathbb{E}[AX + b] = A\mu + b$$

$$\text{cov}[Y] = \text{cov}[AX + b] = A\Sigma A^\top$$

- Likewise, if $Y = f(X) = a^\top X + b$ be a linear function of a vector-valued r.v. $X$ ($a$ is a vector and $b$ is a scalar, both constants)
- Suppose $\mathbb{E}[X] = \mu$ and $\mathbf{cov}[X] = \Sigma$, then for the scalar-valued r.v. $Y$

$$\mathbb{E}[Y] = \mathbb{E}[a^\top X + b] = a^\top \mu + b$$

$$\text{var}[Y] = \text{var}[a^\top X + b] = a^\top \Sigma a$$

# Linear Gaussian Model (LGM)

- LGM defines a noisy lin. transform of a Gaussian r.v. $\boldsymbol{\theta}$ with $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$

Both $\boldsymbol{\theta}$ and $\boldsymbol{y}$ are vectors (can be of different sizes)

Also assume $\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{\Lambda}, \boldsymbol{L}$ to be known; only $\boldsymbol{\theta}$ is unknown

$$\boldsymbol{y} = \boldsymbol{A\theta} + \boldsymbol{b} + \boldsymbol{\epsilon}$$

Noise vector - independently and drawn from $\mathcal{N}(\boldsymbol{\epsilon}|\boldsymbol{0}, \boldsymbol{L}^{-1})$

- Easy to see that, conditioned on $\boldsymbol{\theta}$, $\boldsymbol{y}$ too has a Gaussian distribution

Conditional distribution

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{A\theta} + \boldsymbol{b}, \boldsymbol{L}^{-1})$$

- Assume $p(\boldsymbol{\theta})$ as prior and $p(\boldsymbol{y}|\boldsymbol{\theta})$ as the likelihood, and defining $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \boldsymbol{A}^\top \boldsymbol{L}\boldsymbol{A})^{-1}$

Posterior of $\boldsymbol{\theta}$

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})} = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\Sigma}(\boldsymbol{A}^\top \boldsymbol{L}(\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Lambda\mu}), \boldsymbol{\Sigma})$$

Marginal distribution

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} = \mathcal{N}(\boldsymbol{y}|\boldsymbol{A\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{A}^\top + \boldsymbol{L}^{-1})$$

- Many probabilistic ML models are LGMs
- These results are very widely used (PRML Chap. 2 contains a proof)

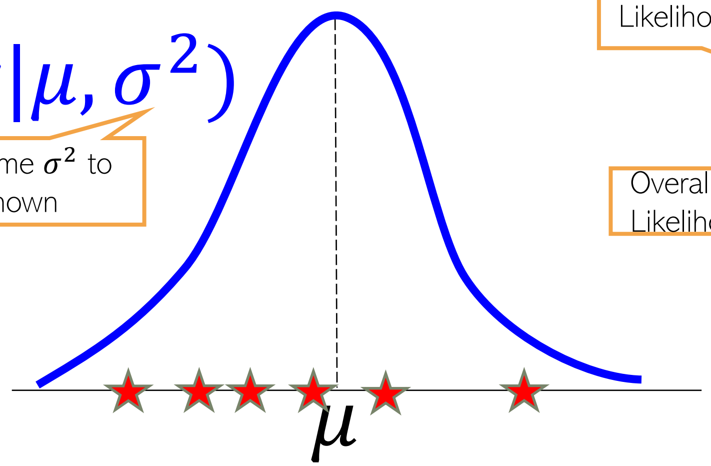
# Posterior Distribution for Gaussian’s Mean

Its MLE/MAP estimation left as an exercise

- Given: $N$ i.i.d. scalar observations $\boldsymbol{y} = \{y_1, y_2, \dots, y_N\}$ assumed drawn from $\mathcal{N}(y|\mu, \sigma^2)$

$\mathcal{N}(y|\mu, \sigma^2)$

Assume $\sigma^2$ to be known




Likelihood

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

Overall Likelihood

$$p(\boldsymbol{y}|\mu, \sigma^2) = \prod_{n=1}^{N} p(y_n|\mu, \sigma^2)$$

- Note: Easy to see that each $y_n$ drawn from $\mathcal{N}(y|\mu, \sigma^2)$ is equivalent to the following

Thus $y_n$ is like a noisy version of $\mu$ with zero mean Gaussian noise added to it

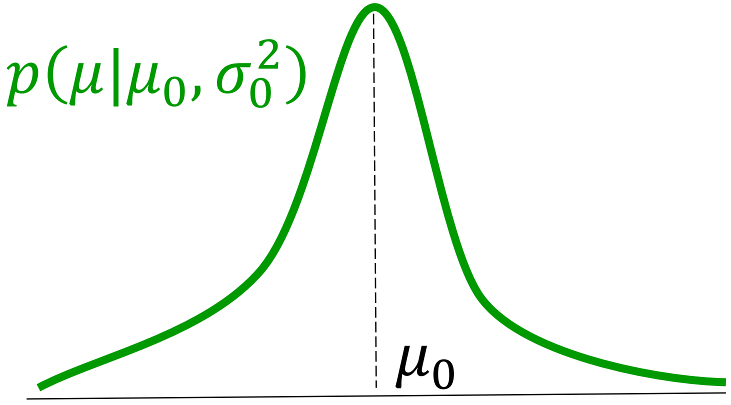
$$y_n = \mu + \epsilon_n \quad \text{where } \epsilon_n \sim \mathcal{N}(0, \sigma^2)$$

- Let’s estimate mean $\mu$ given $\boldsymbol{y}$ using fully Bayesian inference (not point estimation)

# A prior distribution for the mean

- To computer posterior, need a prior over $\boldsymbol{\mu}$
- Let's choose a Gaussian prior

$$p(\mu|\mu_0,\sigma_0^2) = \mathcal{N}(\mu|\mu_0,\sigma_0^2)$$
$$\propto \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right]$$




- The prior basically says that *a priori* we believe $\boldsymbol{\mu}$ is close to $\boldsymbol{\mu_0}$
- The prior's variance $\boldsymbol{\sigma_0^2}$ denotes how certain we are about our belief
- We will assume that the prior's hyperparameters $(\boldsymbol{\mu_0}, \boldsymbol{\sigma_0^2})$ are known
- Since $\boldsymbol{\sigma^2}$ in the likelihood $\mathcal{N}(y|\mu,\sigma^2)$ is known, Gaussian prior $\mathcal{N}(\mu|\mu_0,\sigma_0^2)$ on $\boldsymbol{\mu}$ is also conjugate to the likelihood (thus posterior of $\boldsymbol{\mu}$ will also be Gaussian)

# The posterior distribution for the mean

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

On conditioning side, skipping all fixed params and hyperparams from the notation

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

- Easy to see that the above will be prop. to exp of a quadratic function of $\boldsymbol{\mu}$. Simplifying:

$$p(\mu|\boldsymbol{y}) \propto \exp\left[-\frac{(\mu-\mu_N)^2}{2\sigma_N^2}\right]$$

Gaussian posterior (not a surprise since the chosen prior was conjugate to the likelihood)

Gaussian posterior's precision is the sum of the prior's precision and sum of the noise precisions of all the observations

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

Contribution from the prior

Contribution from the data

Also the MLE solution for $\mu$

Gaussian posterior's mean is a convex combination of prior's mean and the MLE solution

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\bar{y} \quad \left(\text{where } \bar{y} = \frac{\sum_{n=1}^{N} y_n}{N}\right)$$

- What happens to the posterior as $N$ (number of observations) grows very large?
  - Data (likelihood part) overwhelms the prior
  - Posterior's variance $\boldsymbol{\sigma_N^2}$ will approximately be $\boldsymbol{\sigma^2/N}$ (and goes to 0 as $N \to \infty$)
  - The posterior's mean $\boldsymbol{\mu_N}$ approaches $\bar{\boldsymbol{y}}$ (which is also the MLE solution)

Meaning, we become very-very certain about the estimate of $\mu$

# The Predictive Distribution

- If given a point estimate $\hat{\mu}$, the plug-in predictive distribution for a test $y_*$ would be

$$p(y_*|\hat{\mu}, \sigma^2) = \mathcal{N}(y_*|\hat{\mu}, \sigma^2)$$

The best point estimate

This is an approximation of the true PPD $p(y_*|\boldsymbol{y})$

- On the other hand, the posterior predictive distribution of $\boldsymbol{x}_*$ would be

$$p(y_*|\boldsymbol{y}) = \int p(y_*|\mu, \sigma^2) p(\mu|\boldsymbol{y}) d\mu$$

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

$$= \mathcal{N}(y_*|\mu_N, \sigma^2 + \sigma_N^2)$$

This "extra" variance $\sigma_N^2$ in PPD is due to the averaging over the posterior's uncertainty

If conditional is Gaussian then marginal is also Gaussian

A useful fact: When we have conjugacy, the posterior predictive distribution also has a closed form (will see this result more formally when talking about exponential family distributions)

PRML [Bis 06], 2.115, and also mentioned in prob-stats refresher slides

- For an alternative way to get the above result, note that, for test data

$$y_* = \mu + \epsilon \qquad \mu \sim \mathcal{N}(\mu_N, \sigma_N^2) \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Using the <u>posterior</u> of $\mu$ since we are at test stage now

$$\Rightarrow \quad p(y_*|\boldsymbol{y}) = \mathcal{N}(y_*|\mu_N, \sigma^2 + \sigma_N^2)$$

Since both $\mu$ and $\epsilon$ are Gaussian r.v., and are independent, $y_*$ also has a Gaussian posterior predictive, and the respective means and variances of $\mu$ and $\epsilon$ get added up

# Gaussian Observation Model: Some Other Facts

- MLE/MAP for $\mu, \sigma^2$ (or both) is straightforward in Gaussian observation models.
- Posterior also straightforward in most situations for such models
  - (As we saw) computing posterior of $\mu$ is easy (using Gaussian prior) if variance $\sigma^2$ is known
  - Likewise, computing posterior of $\sigma^2$ is easy (using gamma prior on $\sigma^2$) if mean $\mu$ is known
- If $\mu, \sigma^2$ both are unknown, posterior computation requires computing $p(\mu, \sigma^2|\boldsymbol{y})$
  - Computing joint posterior $p(\mu, \sigma^2|\boldsymbol{y})$ exactly requires a jointly conjuage prior $p(\mu, \sigma^2)$
  - "Gaussian-gamma" ("Normal-gamma") is such a conjugate prior – a product of normal and gamma
  - Note: Computing joint posteriors exactly is possible only in rare cases such this one
- If each observation $\boldsymbol{y}_n \in \mathbb{R}^D$, can assume a likelihood/observation model $\mathcal{N}(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - Need to estimate a vector-valued mean $\boldsymbol{\mu} \in \mathbb{R}^D$. Can use a multivariate Gaussian prior
  - Need to estimate a $D \times D$ positive definite covariance matrix $\boldsymbol{\Sigma}$. Can use a Wishart prior
  - If $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ both are unknown, can use Normal-Wishart as a conjugate prior