Gaussian and Linear Gaussian Observation Models

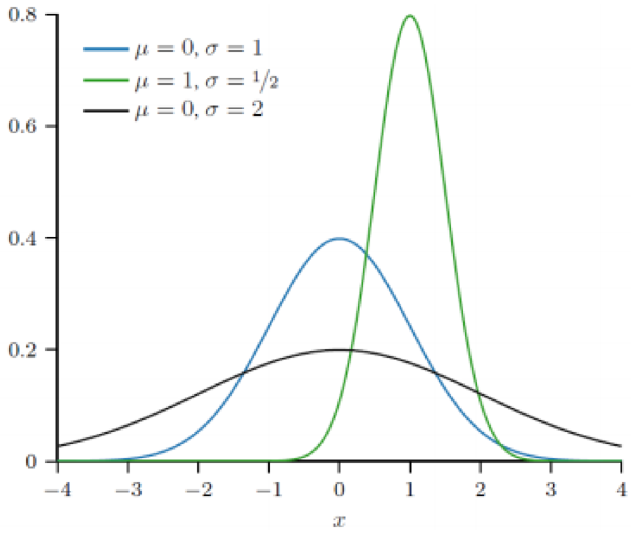
CS772A: Probabilistic Machine Learning

Piyush Rai

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 μ and a scalar variance σ^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

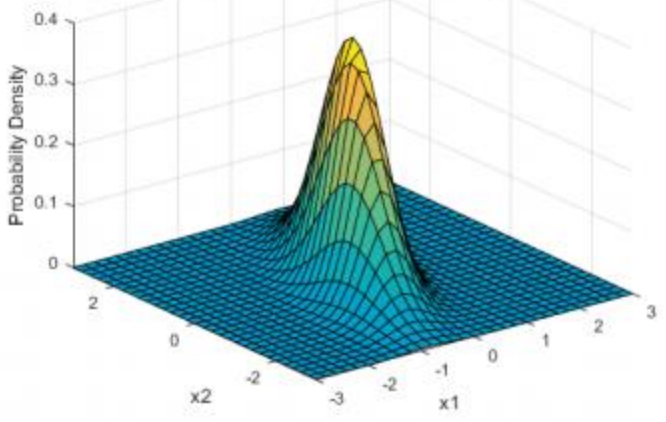
$$\mathcal{N}(Y = y|\mu, \beta) = \sqrt{\frac{\beta}{2\pi}} \exp\left[-\frac{\beta}{2}(y-\mu)^2\right]$$

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 $\boldsymbol{\Sigma}$

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

A two-dimensional Gaussian

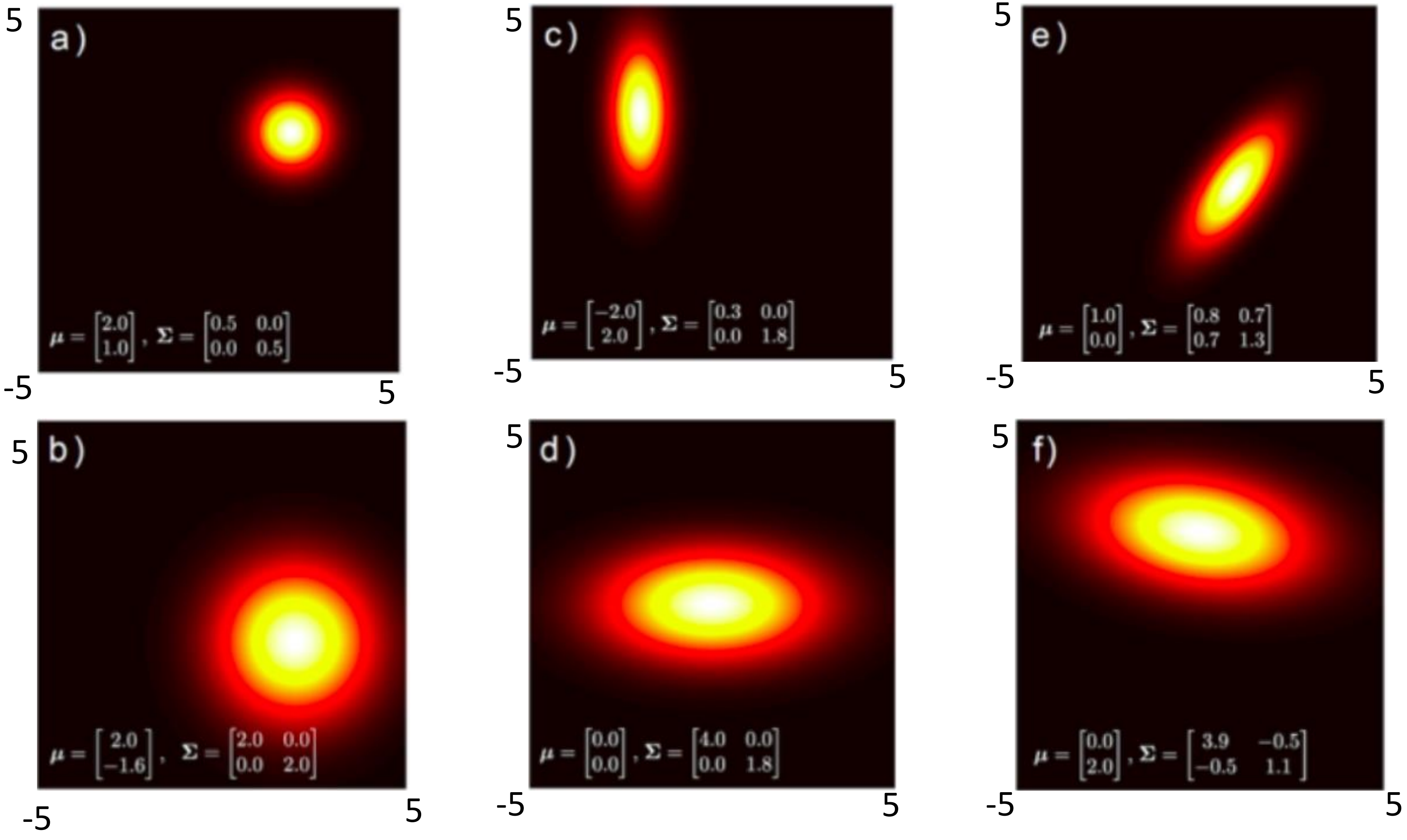


- Note: The cov. matrix $\boldsymbol{\Sigma}$ must be symmetric and PSD
  - All eigenvalues are positive
  - $\boldsymbol{z}^\top \boldsymbol{\Sigma} \boldsymbol{z} \geq \mathbf{0}$ for any real vector $\boldsymbol{z}$

- 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 Covariance | Diagonal Covariance | Full Covariance



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

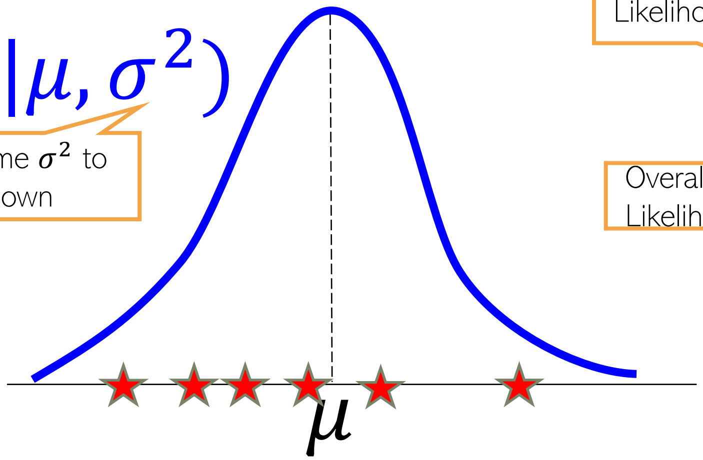
Posterior Distribution for Gaussian's Mean

Its MLE/MAP estimation left as an exercise

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

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

Assume σ^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 μ with zero mean Gaussian noise added to it

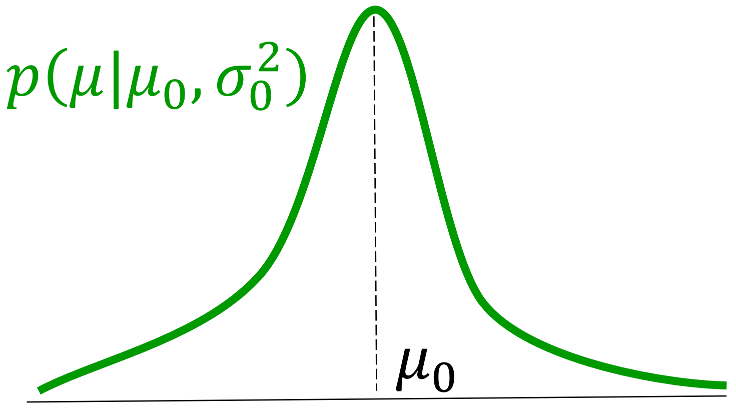
$$y_n = \mu + \epsilon_n \quad \text{where } \epsilon_n \sim \mathcal{N}(0,\sigma^2)$$

- Let's estimate mean μ 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 μ

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

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

- 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 μ

The Predictive Distribution

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

The best point estimate

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

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

- 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 σ_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 μ 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 μ and ϵ are Gaussian r.v., and are independent, y_* also has a Gaussian posterior predictive, and the respective means and variances of μ and ϵ get added up

Gaussian Observation Model: Some Other Facts

- MLE/MAP for μ, σ^2 (or both) is straightforward in Gaussian observation models.
- Posterior also straightforward in most situations for such models
  - (As we saw) computing posterior of μ is easy (using Gaussian prior) if variance σ^2 is known
  - Likewise, computing posterior of σ^2 is easy (using gamma prior on σ^2) if mean μ is known
- If μ, σ^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

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

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

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

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

- 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 $\boldsymbol{\theta}$

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

Marginal likelihood

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

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

Closed form expressions for posterior and marginal likelihood (and both Gaussian)