The Gaussian Distribution – MLE Estimators and Introduction to Bayesian Estimation

Prof. Nicholas Zabaras
School of Engineering
University of Warwick
Coventry CV4 7AL
United Kingdom

Email: nzabaras@gmail.com
URL: http://www.zabaras.com/

August 7, 2014

Contents

- Following closely Chris Bishops' PRML book, Chapter 2
- Kevin Murphy's, Machine Learning: A probablistic perspective, Chapter 2

The Gaussian Distribution

- A random variable $X \in \mathbb{R}$ is Gaussian or normally distributed $X \sim \mathcal{N}(\mu, \sigma^2)$ if:

$$P\{X \le t\} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{t} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx$$

- The following can be shown easily with direct integration:

$$\mathbb{E}[X] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) x dx = \mu,$$

$$\mathbb{E}\left[X^2\right] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) x^2 dx = \mu^2 + \sigma^2, \quad \mathrm{var}[X] = \mathbb{E}\left[\left(X-\mu\right)^2\right] = \sigma^2$$

- The following integrals are useful in these derivations :

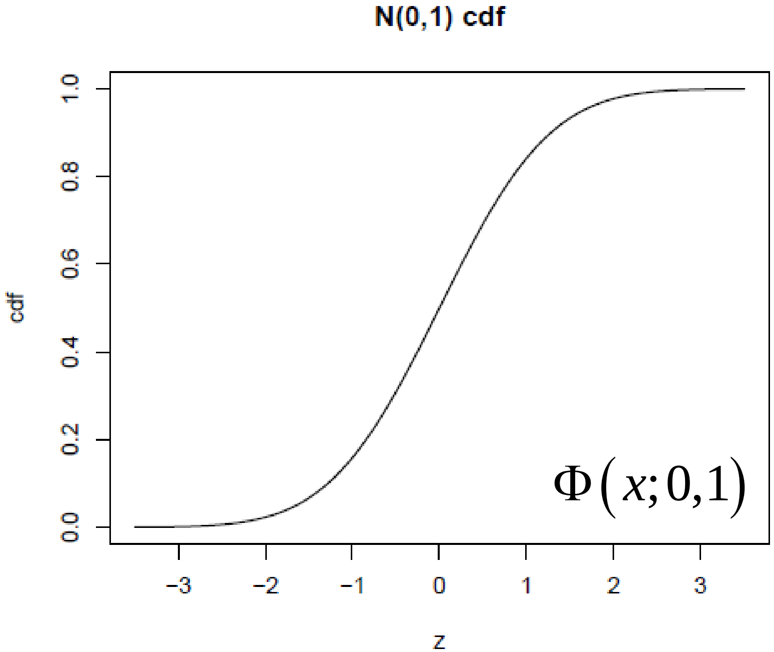
$$\int_{-\infty}^{+\infty} \exp\left(-u^2\right) du = \sqrt{\pi}, \quad \int_{-\infty}^{+\infty} u \exp\left(-u^2\right) du = 0, \quad \int_{-\infty}^{+\infty} u^2 \exp\left(-u^2\right) du = \frac{\sqrt{\pi}}{2}$$

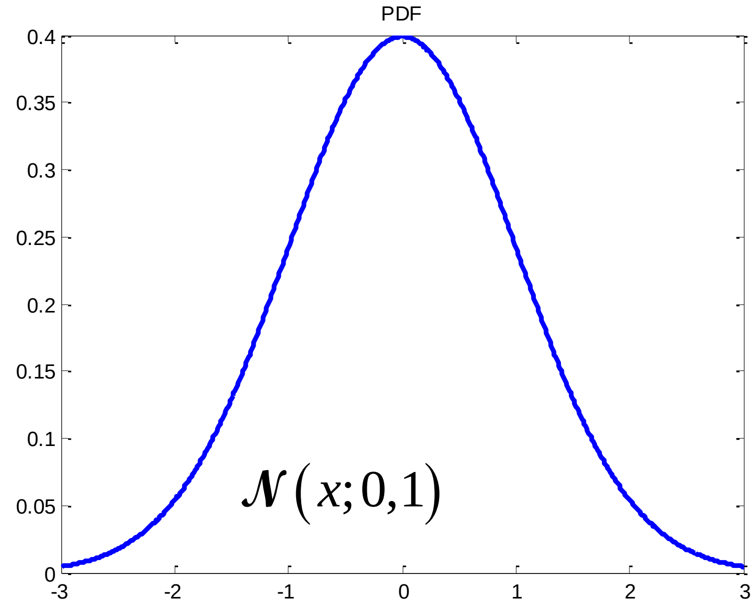
- *We often work with the precision of a Gaussian $\lambda = 1/\sigma^2$. The higher λ the narrower the distribution is.*

Standard Normal, CDF, Error Function

➢ Plot of the Standard Normal $\mathcal{N}(0,1)$ and CDF. Let $\Phi(x;0,1)$ [the corresponding CDF]().

Run *gaussPlotDemo* from PMTK





$$\int_{-\infty}^{x} \mathcal{N}(z \mid \mu,\sigma^2)dz = \Phi(z;0,1),\ z=(x-\mu)/\sigma$$

$$\Phi(z;0,1) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-t^2/2}dt = \frac{1}{2}\left[1+erf\left(z/\sqrt{2}\right)\right]$$

$$erf(x) = \frac{2}{\sqrt{\pi}}\int_{0}^{x} e^{-t^2}dt$$

Degenerate Gaussian Distribution

- Note that as $\sigma^2 \rightarrow 0$, the Gaussian becomes a delta function centered at the mean μ:

$$\lim_{\sigma^2 \to 0} \mathcal{N}\left(x \mid \mu, \sigma^2\right) = \delta(x - \mu)$$

Multivariate Gaussian

- A multivariate $X \in \mathbb{R}^D$ is Gaussian if its probability density is

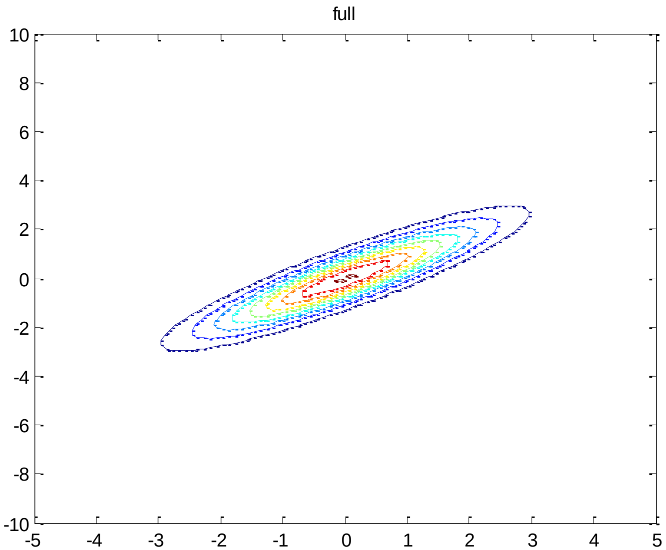
$$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left(\frac{1}{(2\pi)^D \det \boldsymbol{\Sigma}} \right)^{1/2} \exp\left(-\frac{1}{2}(x-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu}) \right)$$

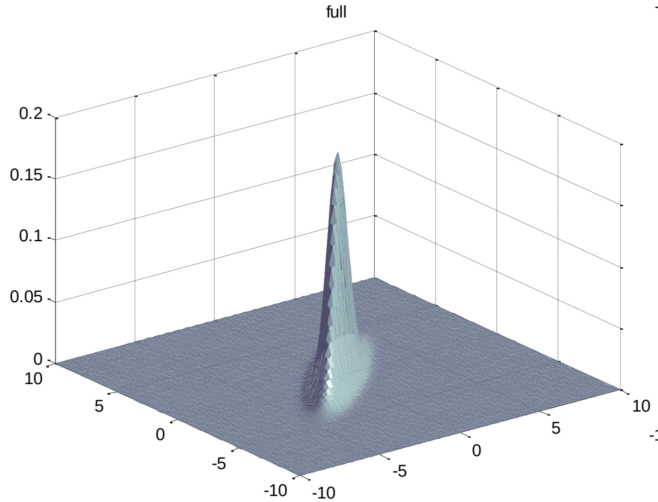
where $\boldsymbol{\mu} \in \mathbb{R}^D, \boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$ is symmetric positive definite matrix (*covariance matrix*).

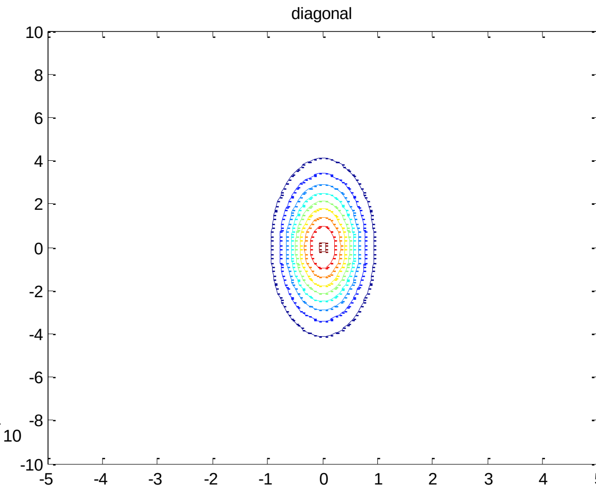
- We often work with *the precision matrix* $\boldsymbol{\Lambda}=\boldsymbol{\Sigma}^{-1}$

2D Gaussian

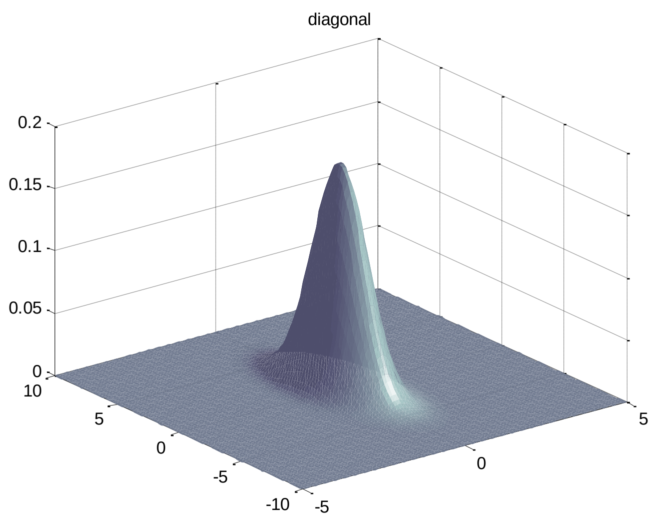
- Level sets of 2D Gaussians (full, diagonal and spherical covariance matrix)

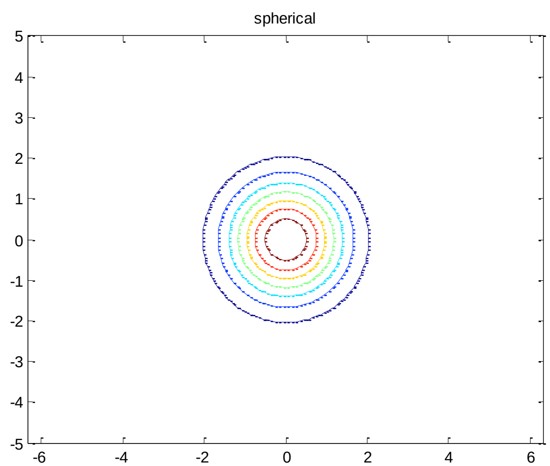


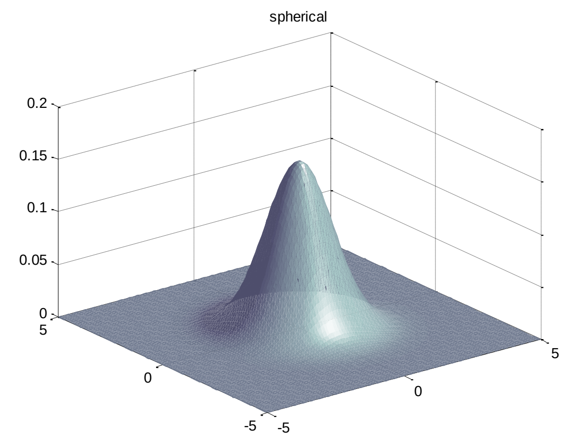




gaussPlot2DDemo from PMTK







Multivariate Gaussian: Maximum Entropy

- We can *show that the multivariate Gaussian maximizes the entropy H with the constraints of normalization with given mean μ and given variance Σ:*

$$\max_{p(\boldsymbol{x}),\lambda,\boldsymbol{m},\boldsymbol{L}} = -\int p(\boldsymbol{x})\ln p(\boldsymbol{x})d\boldsymbol{x} + \lambda\left(\int p(\boldsymbol{x})d\boldsymbol{x}-1\right)+\boldsymbol{m}^T\left(\int \boldsymbol{x}p(\boldsymbol{x})d\boldsymbol{x}-\boldsymbol{\mu}\right)$$
$$+Tr\left(\boldsymbol{L}\left(\int p(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T d\boldsymbol{x}-\boldsymbol{\Sigma}\right)\right)$$

- Setting the derivative wrt $p(\boldsymbol{x})$ to zero gives:

$$0=-1-\ln p(\boldsymbol{x})+\lambda+\boldsymbol{m}^T\boldsymbol{x}+Tr\left(\boldsymbol{L}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\right)$$

$$p(\boldsymbol{x})=e^{-1+\lambda+\boldsymbol{m}^T\boldsymbol{x}+(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{L}(\boldsymbol{x}-\boldsymbol{\mu})}$$

- The coefficients can be found by satisfying the constraints. We start by completing the square.

Multivariate Gaussian: Maximum Entropy

$$p(\boldsymbol{x}) = e^{-1+\lambda+\boldsymbol{m}^T\boldsymbol{x}+(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{L}(\boldsymbol{x}-\boldsymbol{\mu})} =$$

$$= e^{-1+\lambda+\boldsymbol{\mu}^T\boldsymbol{m}-\frac{1}{4}\boldsymbol{m}^T\boldsymbol{L}^{-1}\boldsymbol{m}+(\boldsymbol{x}-\boldsymbol{\mu}+\frac{1}{2}\boldsymbol{L}^{-1}\boldsymbol{m})^T\boldsymbol{L}\overbrace{(\boldsymbol{x}-\boldsymbol{\mu}+\frac{1}{2}\boldsymbol{L}^{-1}\boldsymbol{m})}^{\boldsymbol{y}}}$$

- Satisfying the mean constraint:

$$\int e^{-1+\lambda+\boldsymbol{\mu}^T\boldsymbol{m}-\frac{1}{4}\boldsymbol{m}^T\boldsymbol{L}^{-1}\boldsymbol{m}+\boldsymbol{y}^T\boldsymbol{L}\boldsymbol{y}}\left(\boldsymbol{y}+\boldsymbol{\mu}-\frac{1}{2}\boldsymbol{L}^{-1}\boldsymbol{m}\right)d\boldsymbol{y} = \boldsymbol{\mu}$$

- The 1st term drops from symmetry, the 2nd gives μ from normalization, thus we need to have:

$$-\frac{1}{2}\boldsymbol{L}^{-1}\boldsymbol{m} = \boldsymbol{0} \Rightarrow \boldsymbol{m} = \boldsymbol{0}$$

Multivariate Gaussian: Maximum Entropy

$\boldsymbol{z}$

$$p(\boldsymbol{x}) = e^{-1+\lambda+(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{L}(\boldsymbol{x}-\boldsymbol{\mu})}$$

- Satisfying the variance constraint:

$$\int e^{-1+\lambda+\boldsymbol{z}^T\boldsymbol{L}\boldsymbol{z}}\,\mathbf{z}\mathbf{z}^T d\mathbf{z} = \boldsymbol{\Sigma}$$

- Note that with $\boldsymbol{L} = -\boldsymbol{\Sigma}/2$, the 3rd term from the exponential when integrated gives:

$$\int e^{\boldsymbol{z}^T\boldsymbol{L}\boldsymbol{z}}\,\mathbf{z}\mathbf{z}^T d\mathbf{z} = \boldsymbol{\Sigma}(2\pi)^{D/2}\left|\boldsymbol{\Sigma}\right|^{1/2}$$

- It remains to select λ such that:

$$e^{-1+\lambda} = (2\pi)^{-D/2}\left|\boldsymbol{\Sigma}\right|^{-1/2} \Rightarrow \lambda - 1 = \ln\left\{\frac{1}{(2\pi)^{D/2}}\frac{1}{\left|\boldsymbol{\Sigma}\right|^{1/2}}\right\}$$

- The optimizing p(**x**) is now clearly the Gaussian.

Multivariate Gaussian: Maximum Entropy

- The entropy of the multivariate Gaussian is now computed as follows:

$$
\begin{aligned}
\boldsymbol{H}\left[\boldsymbol{x}\right] &= -\int \mathcal{N}(\boldsymbol{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) \ln \mathcal{N}(\boldsymbol{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) d\boldsymbol{x} \\
&= \int \mathcal{N}(\boldsymbol{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) \frac{1}{2}\left(D\ln(2\pi) + \ln\left|\boldsymbol{\Sigma}\right| + (\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) d\boldsymbol{x} \\
&= \frac{1}{2}\left(D\ln(2\pi) + \ln\left|\boldsymbol{\Sigma}\right|\right) + \int \mathcal{N}(\boldsymbol{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) \frac{1}{2} tr\left((\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}\right) d\boldsymbol{x} \\
&= \frac{1}{2}\left(D\ln(2\pi) + \ln\left|\boldsymbol{\Sigma}\right|\right) + \frac{1}{2} tr\left(\left(\int \mathcal{N}(\boldsymbol{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma})(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T d\boldsymbol{x}\right)\boldsymbol{\Sigma}^{-1}\right) \\
&= \frac{1}{2}\left(D\ln(2\pi) + \ln\left|\boldsymbol{\Sigma}\right|\right) + \frac{1}{2} tr\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\right) \\
&= \frac{1}{2}\left(D\ln(2\pi) + \ln\left|\boldsymbol{\Sigma}\right| + tr\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\right)\right) \\
&= \frac{1}{2}\left(D\ln(2\pi) + \ln\left|\boldsymbol{\Sigma}\right| + D\right)
\end{aligned}
$$

Multivariate Gaussian: Maximum Entropy

- Using also the KL distance definition, one can show that *the Gaussian has the largest entropy from any other distribution satisfying the mean and 2nd moment constraints.* To make the presentation simple, consider

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{0}, \Sigma), \int q(\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^T d\boldsymbol{x} = \boldsymbol{\Sigma}$$

- Then:

$$\begin{aligned}
0 \leq KL(q \| p) &= -\int q(\boldsymbol{x}) \ln \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x} = -\int q(\boldsymbol{x}) \ln p(\boldsymbol{x}) d\boldsymbol{x} + \int q(\boldsymbol{x}) \ln q(\boldsymbol{x}) d\boldsymbol{x} \\
&= -\int q(\boldsymbol{x}) \ln p(\boldsymbol{x}) d\boldsymbol{x} - H[q] = -\int p(\boldsymbol{x}) \ln p(\boldsymbol{x}) d\boldsymbol{x} - H[q] \\
&= H[p] - H[q] \Rightarrow H[p] \geq H[q]
\end{aligned}$$

- The intermediate step in the proof above accounts for the moment constraint on q and the fact that log(p) is quadratic in ***x***!

The CLT and the Gaussian Distribution

- Let $(X_1, X_2, \ldots X_n)$ be independent and identically distributed (i.i.d.) continuous random variables each with expectation μ and variance σ^2.

- Define: $$Z_n = \frac{1}{\sigma\sqrt{N}}\left(X_1 + X_2 + \ldots + X_n - N\mu\right)$$

- As $N \to \infty$, the distribution of Z_n converges to the distribution of a standard normal random variable

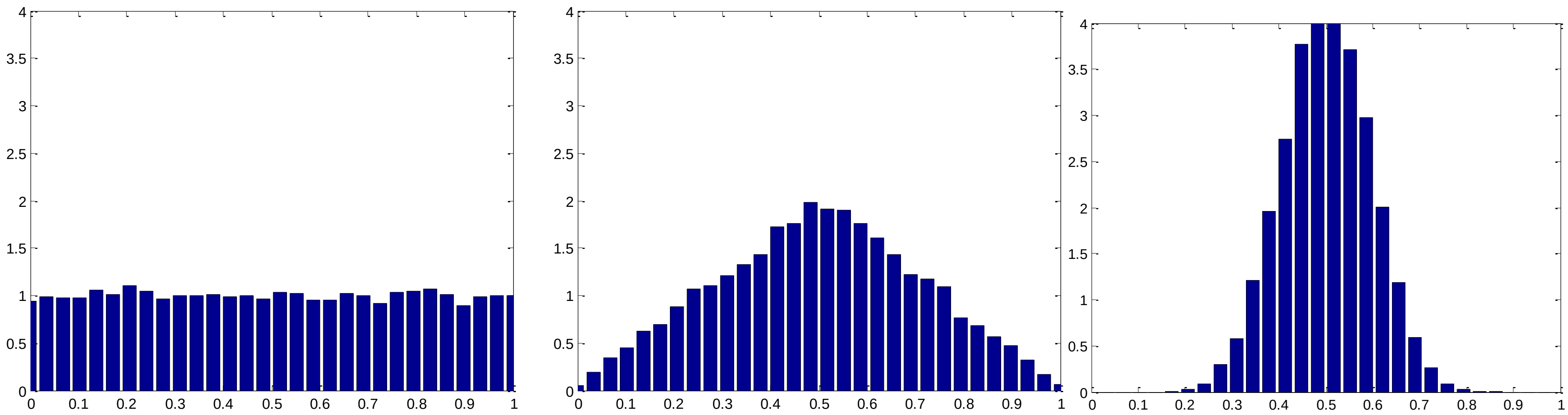
$$\lim_{N\to\infty} P\left\{Z_n \le x\right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$$

- If $\overline{X_n} = \frac{1}{N}\sum_{j=1}^{N} X_j$, for N large, $\overline{X_n} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right) \text{ as } N \to \infty$

- Somewhat of a justification for assuming that Gaussian noise is common

The CLT and the Gaussian Distribution

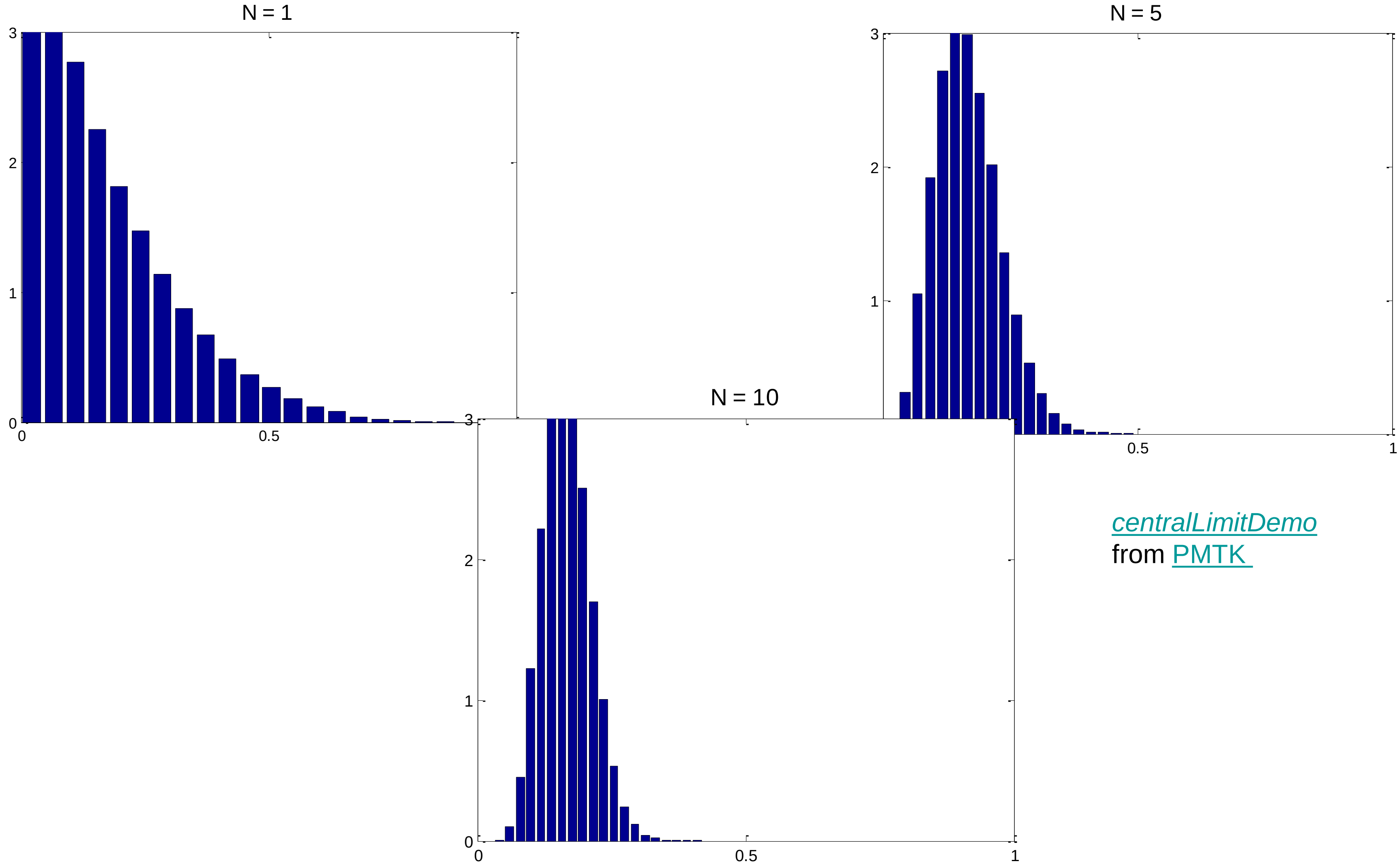
- As an example, assume N variables $(X_1, X_2, \ldots X_n)$ each of which has a uniform distribution over [0, 1] and then consider the distribution of the mean $(X_1+X_2+ \ldots +X_n)/N$. For large N, this distribution tends to a Gaussian. The convergence as N increases can be rapid.



MatLab Code

The CLT and the Gaussian Distribution

- Histogram of $\frac{1}{N}\sum_{j=1}^{10000} x_{ij}$ where $x_{ij} \sim \mathcal{Beta}(1,5)$



centralLimitDemo
from PMTK

The CLT and the Gaussian Distribution

- One consequence of this result is that the binomial distribution which is a distribution over *m* *defined by the sum of N observations of the random binary variable x*, will tend to a Gaussian as $N \to \infty$.

Example of the Convolution of Gaussians

- Consider 2 Gaussians $x_1 \sim \mathcal{N}(\mu_1, \tau_1^{-1}),\, x_2 \sim \mathcal{N}(\mu_2, \tau_2^{-1})$. We want to compute the entropy of the distribution of $x=x_1+x_2$.
- $p(x)$ can be computed from the convolution of two Gaussians

$$p(x) = \int \underbrace{p(x\,|\,x_2)}_{\mathcal{N}(\mu_1+x_2,\tau_1^{-1})} \;\underbrace{p(x_2)}_{\mathcal{N}(\mu_2,\tau_2^{-1})}\; dx_2$$

- We need to *complete the square in the exponential in x_2*:

$$-\frac{1}{2}\tau_1\left(x-(\mu_1+x_2)\right)^2 - \frac{1}{2}\tau_2\left(x_2-\mu_2\right)^2 =$$

$$-\frac{1}{2}\left(\tau_1+\tau_2\right)\left(x_2 - \frac{\tau_1(x-\mu_1)+\tau_2\mu_2}{\tau_1+\tau_2}\right)^2 - \frac{1}{2}\tau_1\left(x-\mu_1\right)^2 + \frac{1}{2}\frac{\left(\tau_1(x-\mu_1)+\tau_2\mu_2\right)^2}{\tau_1+\tau_2}$$

- The 1st term is integrated out and the precision of x is:

$$\tau_1 - \frac{\tau_1^2}{\tau_1+\tau_2} = \frac{\tau_1\tau_2}{\tau_1+\tau_2}$$

- Thus the entropy of x is: $H[x] = \frac{1}{2}\ln\left(2\pi e\sigma^2\right) = \frac{1}{2}\ln\left(2\pi e\frac{\tau_1+\tau_2}{\tau_1\tau_2}\right)$

Maximum Likelihood for a Gaussian

- Suppose that we have a data set of observations $\mathcal{D} = (x_1, \ldots, x_N)^T$, representing N observations of the scalar random variable X. The observations are drawn independently from a Gaussian distribution whose mean μ and variance σ^2 are unknown.

- We would like to determine these parameters from the data set.

- *Data points that are drawn independently from the same distribution are said to be independent and identically distributed, which is often abbreviated to i.i.d.*

Maximum Likelihood for a Gaussian

- Because our data set $\mathcal{D}$ is i.i.d., we can write the probability of the data set, given μ *and* σ^2, *in the form*

$$\text{Likelihood function}: p(\boldsymbol{x} \mid \mu, \sigma^2) = \prod_{i=1}^{N} \mathcal{N}(x_i \mid \mu, \sigma^2)$$

This is seen as a function of μ, σ^2

Max Likelihood for a Gaussian Distribution

$$\textit{Likelihood function}: p(\mathbf{x}\,|\,\mu,\sigma^2) = \prod_{i=1}^{N} \mathcal{N}(x_i\,|\mu,\sigma^2)$$

- One common criterion for determining the parameters in a probability distribution using an observed data set is to find the parameter values that *maximize the likelihood function, i.e. maximizing the probability of the data given the parameters* (contrast this with maximizing the probability of the parameters given the data).

- We can equivalently maximize the log-likelihood:

$$\max_{\mu,\sigma^2} \ln p(\mathbf{x}\,|\,\mu,\sigma^2) = \max_{\mu,\sigma^2}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)\right) \Rightarrow$$

$$\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i,\ \sigma^2_{ML} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{ML})^2$$

Maximum Likelihood for a Gaussian Distribution

$$\mu_{ML} = \underbrace{\frac{1}{N}\sum_{i=1}^{N} x_i}_{\textit{Sample mean}}, \quad \sigma^2_{ML} = \underbrace{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{ML})^2}_{\substack{\textit{Sample variance wrt ML} \\ \textit{mean (not the exact mean)}}}$$

- The MLE underestimates the variance (*bias* due to overfitting) because μ_{ML} fitted some of the noise in the data.
- The maximum likelihood solutions μ_{ML}, σ^2_{ML} are functions of the data set values $x_1, \ldots, x_N$. Consider the expectations of these quantities with respect to the data set values, which come from a Gaussian.
- Using the equations above you can show that :

$$\mathbb{E}\left[\mu_{ML}\right] = \mu, \quad \mathbb{E}\left[\sigma^2_{ML}\right] = \frac{N-1}{N}\sigma^2$$

In this derivation you need to use:

$$E\left[x_i x_j\right] = \sigma^2 \;\; \textit{for } i \neq j$$

$$E\left[x_i^2\right] = \sigma^2 + \mu^2$$

Maximum Likelihood for a Gaussian Distribution

$$\boxed{\sigma_{ML}^2 = \frac{N-1}{N}\sigma^2}$$

We use:

$$E\left[x_i x_j\right] = \sigma^2 \quad \text{for } i \neq j$$

$$E\left[x_i^2\right] = \sigma^2 + \mu^2$$

$$\mathbb{E}\left[\sigma_{ML}^2\right] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2\right] = \frac{1}{N}\mathbb{E}\left[\sum_{n=1}^{N}(x_n - \frac{1}{N}\sum_{m=1}^{N}x_m)^2\right]$$

$$= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[x_n^2 - \frac{2}{N}x_n\sum_{m=1}^{N}x_m + \frac{1}{N^2}\sum_{m=1}^{N}\sum_{l=1}^{N}x_m x_l\right]$$

$$= \frac{1}{N}\left\{N(\mu^2+\sigma^2) - N\frac{2}{N}\Big((N-1)\mu^2 + (\mu^2+\sigma^2)\Big) + N\frac{1}{N^2}N\Big((N-1)\mu^2 + (\mu^2+\sigma^2)\Big)\right\}$$

$$= \frac{1}{N}\left\{N(\mu^2+\sigma^2) - \Big(N\mu^2 + \sigma^2\Big)\right\}$$

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

Maximum Likelihood for a Gaussian Distribution

$$\mathbb{E}\left[\mu_{ML}\right]=\mu,\ \sigma_{ML}^{2}=\frac{N-1}{N}\sigma^{2}$$

- On average the MLE estimate obtains the correct mean but will *underestimate the true variance by a factor (N − 1)/N.*
- An unbiased estimate of the variance is given as:

$$\sigma^{2}=\frac{N}{N-1}\sigma_{ML}^{2}=\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu_{ML})^{2}$$

For large *N*, the bias is not a problem

- This result can be obtained from a Bayesian treatment in which we *marginalize over the unknown mean*.
- The *N-1* factor takes account the fact that *1 degree of freedom has been used in fitting the mean and removes the bias of MLE*.

MLE for the Multivariate Gaussian

- We can easily generalize the earlier results for a multivariate Gaussian. The log-likelihood takes the form:

$$\ln p(\boldsymbol{X} \mid \mathscr{D}, \boldsymbol{\mu}, \Sigma) = -\frac{ND}{2}\ln\left(2\pi\right) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu})$$

- Setting the derivatives wrt $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ equal to zero gives the following:

$$\boldsymbol{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n, \; \boldsymbol{\Sigma}_{ML} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})^T$$

- We provide a proof of the calculation of Σ_{ML} next.

MLE for the Multivariate Guassian

$$\ln p(\boldsymbol{X}|\mathscr{D},\boldsymbol{\mu},\Sigma) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu})$$

- We differentiate the log likelihood wrt Σ^{-1}. Each contributing term is:

$$-\frac{N}{2}\frac{\partial}{\partial\boldsymbol{\Sigma}^{-1}}\ln|\boldsymbol{\Sigma}| = \frac{N}{2}\frac{\partial}{\partial\boldsymbol{\Sigma}^{-1}}\ln|\boldsymbol{\Sigma}^{-1}| = \frac{N}{2}\boldsymbol{\Sigma}^T = \boxed{\frac{N}{2}\boldsymbol{\Sigma}}$$

A useful trick!

$$-\frac{1}{2}\frac{\partial}{\partial\boldsymbol{\Sigma}^{-1}}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}) = -\frac{1}{2}N\frac{\partial}{\partial\boldsymbol{\Sigma}^{-1}}Tr\left(\boldsymbol{\Sigma}^{-1}\sum_{n=1}^{N}\frac{1}{N}(\boldsymbol{x}_n-\boldsymbol{\mu})(\boldsymbol{x}_n-\boldsymbol{\mu})^T\right)$$

$$= -\frac{1}{2}N\frac{\partial}{\partial\boldsymbol{\Sigma}^{-1}}Tr\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\right) = -\frac{1}{2}N\boldsymbol{S} =$$

S symmetric

$$\boxed{-\frac{1}{2}N\boldsymbol{S},}\quad where\ \boldsymbol{S} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})(\boldsymbol{x}_n-\boldsymbol{\mu})^T$$

- So finally $\boldsymbol{\Sigma}_{ML} = \boldsymbol{S}$

- Here we used:

$$\frac{\partial}{\partial\mathbf{A}}Tr(\mathbf{AB}) = \mathbf{B}^T,\ \frac{\partial}{\partial\boldsymbol{A}}\ln|\boldsymbol{A}| = \left(\boldsymbol{A}^{-1}\right)^T,$$

$$|\boldsymbol{A}^{-1}| = |\boldsymbol{A}|^{-1},\ tr(\boldsymbol{AB}) = tr(\boldsymbol{BA})$$

Appendix: Some Useful Matrix Operations

- Show that

$$\frac{\partial}{\partial \mathbf{A}} Tr\left(\mathbf{AB}\right) = \mathbf{B}^T \ and \ \frac{\partial}{\partial \mathbf{B}} Tr\left(\mathbf{AB}\right) = \mathbf{A}^T$$

Indeed

$$\frac{\partial}{\partial A_{mn}} Tr\left(\mathbf{AB}\right) = \frac{\partial}{\partial A_{mn}}\left(A_{ik}B_{ki}\right) = B_{nm} \Rightarrow \frac{\partial}{\partial \mathbf{A}} Tr\left(\mathbf{AB}\right) = \mathbf{B}^T$$

- Show that

$$\frac{\partial}{\partial \boldsymbol{A}} \ln|\boldsymbol{A}| = \left(\boldsymbol{A}^{-1}\right)^T$$

Using the cofactor expansion of the det:

$$\frac{\partial}{\partial A_{mn}} \ln|\boldsymbol{A}| = \frac{1}{|\boldsymbol{A}|}\frac{\partial}{\partial A_{mn}}|\boldsymbol{A}| = \frac{1}{|\boldsymbol{A}|}\frac{\partial}{\partial A_{mn}}\sum_j (-1)^{i+j} A_{ij} M_{ij} = \frac{1}{|\boldsymbol{A}|}(-1)^{m+n} M_{mn} = \left(\boldsymbol{A}^{-1}\right)_{nm}$$

where in the last step we used Cramer’s rule.

MLE for a Multivariate Gaussian

$$\boldsymbol{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n \equiv \overline{\boldsymbol{x}},\ \boldsymbol{\Sigma}_{ML} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})^T = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^T - \overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^T$$

- Note that *the unconstrained maximization of the log-likelihood gives a symmetric Σ.*

- As for the univariate case, we can define an unbiased covariance as:

$$\boldsymbol{\Sigma}_{ML} = \frac{1}{N-1}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})^T,\ \mathbb{E}\left[\boldsymbol{\Sigma}_{ML}\right] = \boldsymbol{\Sigma}$$

- To prove this, you will need to use that:

$$\mathbb{E}\left[\boldsymbol{x}_n\boldsymbol{x}_m^T\right] = \boldsymbol{\mu}\boldsymbol{\mu}^T + \delta_{mn}\boldsymbol{\Sigma}$$

Sequential MLE Estimation for Gaussians

- Often we are interested to compute sequentially an estimate of μ_{ML} as more data arrive. This can easily be done:

$$\mu_{ML}^{(N)} = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{x_N}{N} + \frac{1}{N}\sum_{n=1}^{N-1} x_n =$$

$$= \frac{x_N}{N} + \frac{N-1}{N}\frac{1}{N-1}\sum_{n=1}^{N-1} x_n =$$

$$= \frac{x_N}{N} + \frac{N-1}{N}\mu_{ML}^{(N-1)} = \mu_{ML}^{(N-1)} + \underbrace{\frac{1}{N}}_{\text{Learning rate}} \underbrace{\left(x_N - \mu_{ML}^{(N-1)}\right)}_{\text{Error signal}}$$

- This sequential approach cannot easily be generalized to other cases (non-Gaussians, etc.)

Robbins-Monro Algorithm

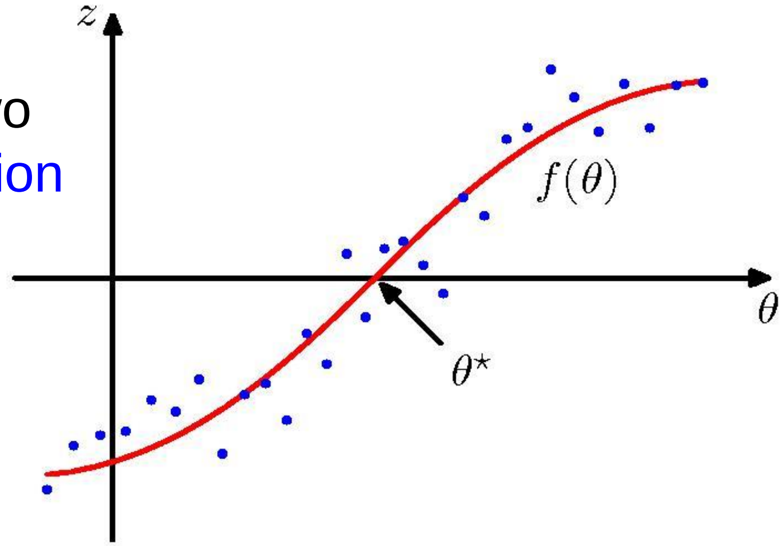
- A more powerful approach to computing sequentially the MLE estimates is via the Robbins-Monro algorithm.

- We review the algorithm by considering the calculation of the zero of a regression function.*

- Consider the joint distribution p(z,θ) of two random variables and define the regression function as:

$$f(\theta) = \mathbb{E}(z \mid \theta) = \int z p(z \mid \theta) dz$$

- Assume we are given samples from *p(z,θ)* one at a time.



* Effectively, we don't know the regression function f(θ) but we have data of a noisy version z of that. We take the regression function to be the expectation $\mathbb{E}(z \mid \theta)$.

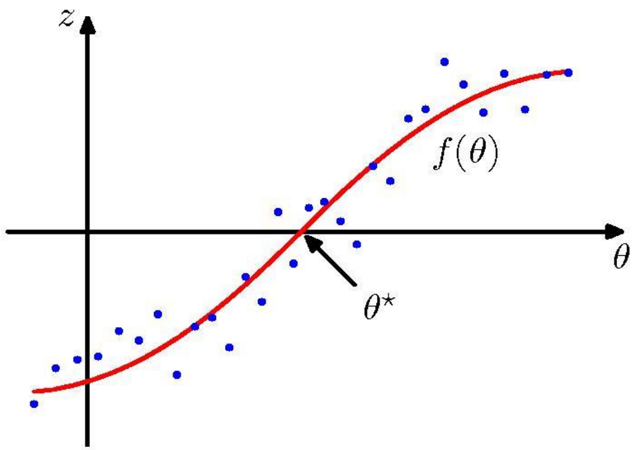
- Robbins, H. and S. Monro (1951). A stochastic approximation method. *Annals of Mathematical Statistics* **22**, 400–407.
- Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition* (Second ed.). Academic Press.

Robbins-Monro Algorithm

$$f(\theta) = \mathbb{E}(z \mid \theta) = \int z p(z \mid \theta) dz$$

- We want to find the root $f(\theta^*) = 0$ in a sequential manner: The Robbins-Monro algorithm proceeds as:

$$\theta^{(N)} = \theta^{(N-1)} - a_{N-1} z\left(\theta^{(N-1)}\right)$$



- The *learning coefficients* $\{a_N\}$ should satisfy:

$$\lim_{N\to\infty} a_N = 0, \sum_{n=1}^{\infty} a_N = \infty, \sum_{n=1}^{\infty} a_N^2 < \infty$$

Robbins-Monro Algorithm

- We can state the MLE calculation μ_{ML} for our Gaussian example as finding the root of a regression function:

$$-\frac{\partial}{\partial\mu}\left\{\frac{1}{N}\sum_{n=1}^{N}\ln p(x_n\,|\,\mu)\right\}|_{ML}=0 \Rightarrow -\sum_{n=1}^{N}\frac{1}{N}\frac{\partial}{\partial\mu}\ln p(x_n\,|\,\mu)|_{ML}=0 \underset{N\to\infty}{\overset{CLT}{\Rightarrow}} \mathbb{E}\left[\underbrace{-\frac{\partial}{\partial\mu}\ln p(x\,|\,\mu)}_{z|\mu}\right]_{ML}=0$$

- In the context of the Robbins-Monro algorithm,

$$z=\frac{\partial}{\partial\mu}\left[-\ln p(x\,|\,\mu)\right]|_{\mu=\mu_{ML}}=-\frac{x-\mu_{ML}}{\sigma^2},\ z \textit{ is a Gaussian},\ f(\mu)=\mathbb{E}(z\,|\,\mu)=-\frac{\mu-\mu_{ML}}{\sigma^2}$$

- The algorithm takes the form:

$$\mu_{ML}^{(N)}=\mu_{ML}^{(N-1)}+a_{N-1}\frac{x^{(N)}-\mu_{ML}^{(N-1)}}{\sigma^2}$$

- Substituting $a_{N-1}=\frac{\sigma^2}{N}$ gives the estimate discussed earlier.

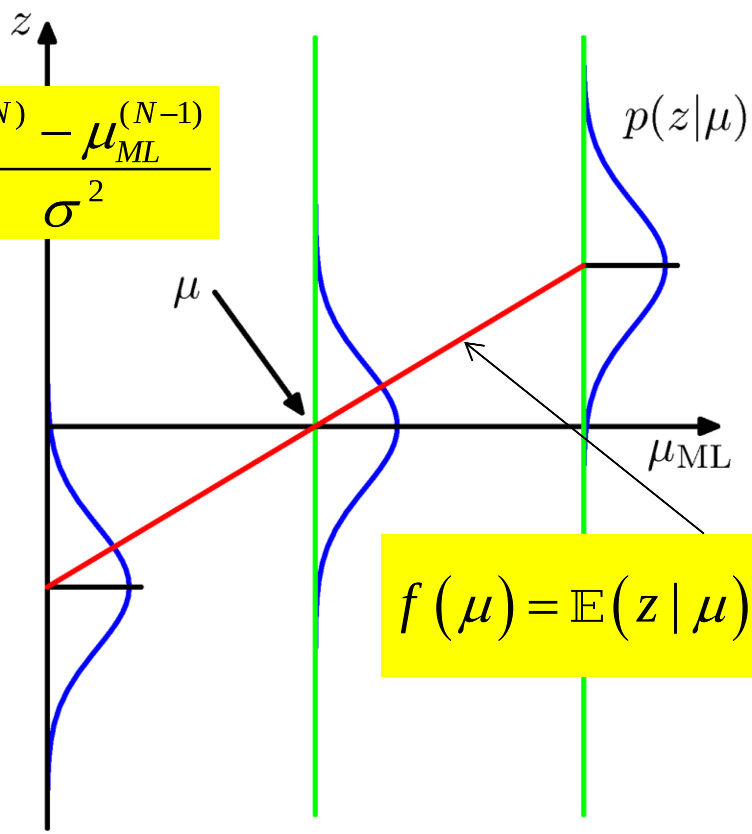
Robbins-Monro Algorithm

- A graphical interpretation of the algorithm is shown here.

$$\mu_{ML}^{(N)} = \mu_{ML}^{(N-1)} + a_{N-1}\frac{x^{(N)} - \mu_{ML}^{(N-1)}}{\sigma^2}$$

$$z = -\frac{x-\mu_{ML}}{\sigma^2}$$

$p(z \mid \mu)$ *is a Gaussian*



$$f(\mu) = \mathbb{E}(z \mid \mu) = -\frac{\mu-\mu_{ML}}{\sigma^2}$$

- The Robbin-Monro algorithm computes the zero of the regression function.

- Blum, J. A. (1965). Multidimensional stochastic approximation methods. *Annals of Mathematical Statistics* **25**, 737–744.

Sequential MLE Estimation for Gaussians

- Let us now repeat the same calculations but for the MLE estimate of σ^2:

$$\sigma^2_{(N)} = \frac{1}{N}\sum_{n=1}^{N}\left(x_n-\mu\right)^2 = \frac{1}{N}\sum_{n=1}^{N-1}\left(x_n-\mu\right)^2 + \frac{\left(x_N-\mu\right)^2}{N}$$

$$= \frac{N-1}{N}\sigma^2_{(N-1)} + \frac{\left(x_N-\mu\right)^2}{N} =$$

$$= \sigma^2_{(N-1)} + \frac{1}{N}\left\{\left(x_N-\mu\right)^2 - \sigma^2_{(N-1)}\right\}$$

- If we substitute the expression for the Gaussian likelihood into the Robbins-Monro procedure for maximizing likelihood:

$$\sigma^2_{(N)} = \sigma^2_{(N-1)} + a_{N-1}\frac{\partial}{\partial\sigma^2_{(N-1)}}\left\{-\frac{1}{2}\ln\sigma^2_{(N-1)} - \frac{\left(x_N-\mu\right)^2}{2\sigma^2_{(N-1)}}\right\} = \sigma^2_{(N-1)} + a_{N-1}\frac{1}{2\sigma^4_{(N-1)}}\left\{\left(x_N-\mu\right)^2 - \sigma^2_{(N-1)}\right\}$$

- The 2 formulas are identical for: $a_{N-1} = 2\sigma^4_{(N-1)}/N.$

Sequential MLE: Multivariate Gaussian

- To simplify things, *assume that* $\boldsymbol{\mu}_{ML}=\boldsymbol{\mu}$ and thus:

$$\boldsymbol{\Sigma}_{ML}^{(N)}=\frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n-\boldsymbol{\mu})(\boldsymbol{x}_n-\boldsymbol{\mu})^T$$

- From this equation we can derive:

$$\boldsymbol{\Sigma}_{ML}^{(N)}=\boldsymbol{\Sigma}_{ML}^{(N-1)}+\frac{1}{N}\left((\boldsymbol{x}_N-\boldsymbol{\mu})(\boldsymbol{x}_N-\boldsymbol{\mu})^T-\boldsymbol{\Sigma}_{ML}^{(N-1)}\right)$$

- To apply the Robbins-Monro algorithm, *assume that Σ is diagonal* and as before compute the derivative

$$\frac{\partial}{\partial\boldsymbol{\Sigma}_{ML}^{(N-1)}}\left(-\ln p\left(\boldsymbol{x}_N\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}_{ML}^{(N-1)}\right)\right)=-\frac{1}{2}\left(\boldsymbol{\Sigma}_{ML}^{(N-1)}\right)^{-2}\left((\boldsymbol{x}_N-\boldsymbol{\mu})(\boldsymbol{x}_N-\boldsymbol{\mu})^T-\boldsymbol{\Sigma}_{ML}^{(N-1)}\right)$$

- Substituting into the RM algorithm:

$$\boldsymbol{\Sigma}_{ML}^{(N)}=\boldsymbol{\Sigma}_{ML}^{(N-1)}+\boldsymbol{A}_{N-1}\frac{1}{2}\left(\boldsymbol{\Sigma}_{ML}^{(N-1)}\right)^{-2}\left((\boldsymbol{x}_N-\boldsymbol{\mu})(\boldsymbol{x}_N-\boldsymbol{\mu})^T-\boldsymbol{\Sigma}_{ML}^{(N-1)}\right)$$

- Thus from the RM algorithm, we can obtain the exact update by selecting

$$\boldsymbol{A}_{N-1}=\frac{2}{N}\left(\boldsymbol{\Sigma}_{ML}^{(N-1)}\right)^{2}$$

Bayesian Inference for the Gaussian: Known Variance

- Consider $X_1 \mid \mu \sim \mathcal{N}(\mu, \sigma^2)$, *with prior* $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$. We want to infer μ with the variance σ^2 taken as known. The case with multiple data points will be considered later on.

- Then we can derive the following:

$$\pi(\mu \mid x_1) \propto f(x_1 \mid \mu)\pi(\mu) \propto \exp\left(-\frac{(x_1-\mu)^2}{2\sigma^2} - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right) \Rightarrow$$

$$\pi(\mu \mid x_1) \propto \exp\left(-\frac{\mu^2}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}\right) + \mu\left(\frac{x_1}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)\right) \propto \exp\left(-\frac{1}{2\sigma_1^2}(\mu-\mu_1)^2\right) \Rightarrow$$

$$\mu \mid x_1 \sim \mathcal{N}(\mu_1, \sigma_1^2) \text{ with}$$

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{1}{\sigma^2} \Rightarrow \sigma_1^2 = \frac{\sigma_0^2\sigma^2}{\sigma_0^2+\sigma^2}, \text{ and}$$

$$\mu_1 = \sigma_1^2\left(\frac{x_1}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)$$

Bayesian Inference: Predictive distribution

- To predict the distribution of a new observation $X \mid \mu \sim \mathcal{N}(\mu, \sigma^2)$ in light of x_1, we use the predictive distribution as follows:

$$f(x \mid x_1) = \int \underbrace{f(x \mid \mu)}_{Likelihood} \underbrace{\pi(\mu \mid x_1)}_{Posterior} d\mu \propto \int e^{-\frac{(x-\mu)^2}{2\sigma^2}} e^{-\frac{(\mu-\mu_1)^2}{2\sigma_1^2}} d\mu = \int e^{-\frac{1}{2}\left(\frac{(x-\mu)^2}{\sigma^2}+\frac{(\mu-\mu_1)^2}{\sigma_1^2}\right)} d\mu$$

- We can complete the square by treating the integrand above as a bivariate Gaussian in (x,μ). One can verify that:

$$\frac{1}{2}\left(\frac{(x-\mu)^2}{\sigma^2}+\frac{(\mu-\mu_1)^2}{\sigma_1^2}\right) = \frac{1}{2}\begin{pmatrix} x-\mu_1 & \mu-\mu_1 \end{pmatrix} \underbrace{\begin{pmatrix} \frac{1}{\sigma^2} & -\frac{1}{\sigma^2} \\ -\frac{1}{\sigma^2} & \frac{1}{\sigma^2}+\frac{1}{\sigma_1^2} \end{pmatrix}}_{\Sigma^{-1}} \begin{pmatrix} x-\mu_1 \\ \mu-\mu_1 \end{pmatrix} + const.$$

- From the above expression note that: $\Sigma = \begin{pmatrix} \sigma^2+\sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 \end{pmatrix}$

Bayesian Inference: Predictive distribution

- We will see at a follow up lecture that if we partition the mean and variance of a multivariate Gaussian as:

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$$

then, the marginal

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

- In our predictive distribution we need to integrate out μ. Thus based on the above result and $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_1 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 + \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 \end{pmatrix}$, we have:

$$f(x \mid x_1) = \int \underbrace{f(x \mid \mu)}_{Likelihood} \underbrace{\pi(\mu \mid x_1)}_{Posterior} d\mu = \mathcal{N}\left(x \mid \mu_1, \sigma^2 + \sigma_1^2\right)$$

- Note *the variance is the sum of model variance + variance of posterior uncertainty in μ.*

Bayesian Inference for the Gaussian

- Consider $\boldsymbol{X} = \{x_1, x_2, ..., x_N\} \sim \mathcal{N}(\mu, \sigma^2)$, *with prior* $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$.

- The likelihood takes the form:

$$p(\boldsymbol{X} \mid \mu) = \prod_{n=1}^{N} f(x_n \mid \mu) = \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left(-\frac{\sum_{n=1}^{N} (x_n - \mu)^2}{2\sigma^2} \right)$$

- Note that in terms of μ this is not a probability density and is not normalized. Introducing the conjugate (Gaussian) prior on μ leads to:

$$\pi(\mu \mid \boldsymbol{X}) = \prod_{n=1}^{N} f(x_n \mid \mu) \pi(\mu) \propto \exp\left(-\frac{\sum_{n=1}^{N} (x_n - \mu)^2}{2\sigma^2} - \frac{(\mu - \mu_0)^2}{2\sigma_0^2} \right) \Rightarrow$$

$$\pi(\mu \mid \boldsymbol{X}) \propto \exp\left(-\frac{\mu^2}{2}\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \right) + \mu \left(\frac{\sum_{n=1}^{N} x_n}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) \right) \propto \exp\left(-\frac{1}{2\sigma_1^2} (\mu - \mu_N)^2 \right)$$

Bayesian Inference for the Gaussian

$$\pi(\mu \mid \boldsymbol{X}) \propto \exp\left(-\frac{\mu^2}{2}\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right) + \mu\left(\frac{\sum_{n=1}^{N} x_n}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)\right) \propto \exp\left(-\frac{1}{2\sigma_N^2}(\mu-\mu_N)^2\right)$$

- So the posterior is a Gaussian as before with

$$\mu \mid \boldsymbol{X} \sim \mathcal{N}(\mu_N, \sigma_N^2) \ with$$

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

$$\mu_N = \sigma_N^2\left(\frac{\sum_{n=1}^{N} x_n}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right) = \sigma_N^2\left(\frac{N\mu_{ML}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right) = \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{ML} + \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0$$

Bayesian Inference for the Gaussian

$$\mu \mid \boldsymbol{X} \sim \mathcal{N}(\mu_N, \sigma_N^2) \; with$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \Rightarrow \sigma_N^2 = \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2}, \; and \; \mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0$$

- Observe the posterior mean for $N \to \infty$ and $N \to 0$.

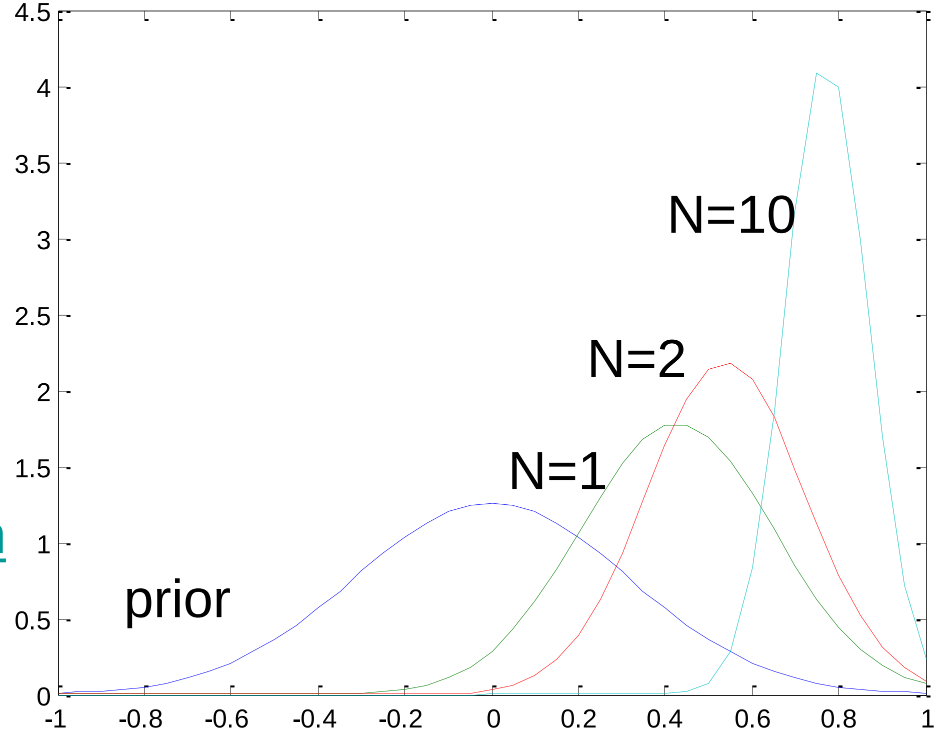
- *The posterior precision is the sum of the precision of the prior plus one contribution of the data precision for each observed data point*. As we have seen before for $N \to \infty$ the posterior peaks around the μ_{ML} and the posterior variance goes to zero, i.e. the point MLE estimate is recovered within the Bayesian paradigm for infinite data.

- How about when $\sigma_0^2 \to \infty$? In this case note that $\sigma_N^2 \to \dfrac{\sigma^2}{N}$ *and* $\mu_N \to \mu_{ML}$

Bayesian Inference for the Gaussian

$$\mu \mid \boldsymbol{X} \sim \mathcal{N}(\mu_N, \sigma_N^2) \text{ with } \sigma_N^2 = \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2}, \text{ and } \mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0$$

MatLab implementation



$$\boldsymbol{X} = \{x_1, x_2, ..., x_N\} \sim \mathcal{N}(0.8, 0.1), \text{ with prior } \mu \sim \mathcal{N}(0, 0.1).$$

Sequential Bayesian Inference

$$\mu \mid \boldsymbol{X} \sim \mathcal{N}(\mu_N, \sigma_N^2) \; with$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \Rightarrow \sigma_N^2 = \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2}, \; and \; \mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0$$

- We can easily derive sequential estimates of the MLE. They are as follows:

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_{N-1}^2} + \frac{1}{\sigma^2}, \; and \; \mu_N = \frac{\sigma_N^2}{\sigma_{N-1}^2}\mu_{N-1} + \frac{\sigma_N^2}{\sigma^2} x_N$$