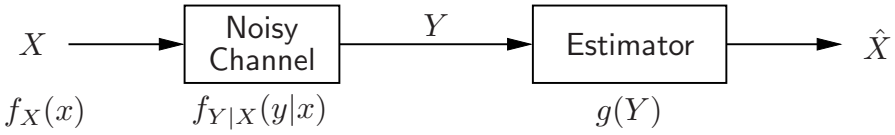
Lecture Notes 5
Mean Square Error Estimation

- Minimum MSE Estimation
- Linear Estimation
- Jointly Gaussian Random Variables

Minimum MSE Estimation

- Consider the following signal processing problem:



- X is a signal with known statistics, i.e., known pdf $f_X(x)$
- The signal is transmitted (or stored) over a noisy channel with known statistics, i.e., conditional pdf $f_{Y|X}(y|x)$
- We observe the signal Y and wish to find the *estimate* $\hat{X} = g(Y)$ of X that minimizes the *mean square error*

$$\text{MSE} = \text{E}\left[(X - \hat{X})^2\right] = \text{E}\left[(X - g(Y))^2\right]$$

- The $\hat{X}$ that achieves the minimum MSE is called the *minimum MSE estimate* (MMSE) of X (given Y)

MMSE Estimate

- Theorem: The MMSE estimate of X given the observation Y and complete knowledge of the joint pdf $f_{X,Y}(x,y)$ is

$$\hat{X} = \mathrm{E}(X \,|\, Y)\,,$$

  and the MSE of $\hat{X}$, i.e., the minimum MSE, is

$$\mathrm{MMSE} = \mathrm{E}_Y(\mathrm{Var}(X \,|\, Y)) = \mathrm{E}(X^2) - \mathrm{E}\left[(\mathrm{E}(X \,|\, Y))^2\right]$$

- Properties of the minimum MSE estimator:
  - Since $\mathrm{E}(\hat{X}) = \mathrm{E}_Y[\mathrm{E}(X \,|\, Y)] = \mathrm{E}(X)$, the best MSE estimate is *unbiased*
  - If X and Y are independent, then the best MSE estimate is $\mathrm{E}(X)$
  - The conditional expectation of the estimation error, $\mathrm{E}\left[(X - \hat{X}) \,|\, Y = y\right]$, is 0 for all y, i.e., the error is unbiased for every $Y = y$

  - The estimation error and the estimate are orthogonal

$$\begin{aligned}
\mathrm{E}\left[(X - \hat{X})\hat{X}\right] &= \mathrm{E}_Y\left[\mathrm{E}\left((X - \hat{X})\hat{X} \,|\, Y\right)\right] \\
&= \mathrm{E}_Y\left[\hat{X}\,\mathrm{E}((X - \hat{X}) \,|\, Y)\right] \\
&= \mathrm{E}_Y\left[\hat{X}(\mathrm{E}(X \,|\, Y) - \hat{X}) \,|\, Y)\right] \\
&= 0
\end{aligned}$$

    In fact, the estimation error is orthogonal to *any* function $g(Y)$ of Y

  - From the law of conditional variance

$$\mathrm{Var}(X) = \mathrm{Var}(\hat{X}) + \mathrm{E}(\mathrm{Var}(X \,|\, Y))\,,$$

    i.e., the sum of the variance of the estimate and the minimum MSE is equal to the variance of the signal

- Proof of Theorem: We first show that $\min_a \mathrm{E}\left((X-a)^2\right) = \mathrm{Var}(X)$ and that the minimum is achieved for $a = \mathrm{E}(X)$, i.e., in the absence of any observations, the mean of X is its minimum MSE estimate

  To show this, consider

$$\begin{aligned}\mathrm{E}\left[(X-a)^2\right] &= \mathrm{E}\left[(X-\mathrm{E}(X)\ +\ \mathrm{E}(X)-a)^2\right] \\ &= \mathrm{E}\left[(X-\mathrm{E}(X))^2\right]\ +\ \big(\mathrm{E}(X)-a\big)^2 + \\ &\qquad 2\,\mathrm{E}(X-\mathrm{E}(X))(\mathrm{E}(X)-a) \\ &= \mathrm{E}\left[(X-\mathrm{E}(X))^2\right]\ +\ \big(\mathrm{E}(X)-a\big)^2 \\ &\geq \mathrm{E}\left[(X-\mathrm{E}(X))^2\right]\end{aligned}$$

  Equality holds if and only if $a = \mathrm{E}(X)$

  We use this result to show that $\mathrm{E}(X \,|\, Y)$ is the MMSE estimate of X given Y

  First write

$$\mathrm{E}\left[(X-g(Y))^2\right]\ =\ \mathrm{E}_Y\left[\mathrm{E}_X((X-g(Y))^2 \,|\, Y)\right]$$

  From the previous result we know that for each $Y = y$ the minimum value for $\mathrm{E}_X\left[(X-g(y))^2 \,|\, Y = y\right]$ is obtained when $g(y) = \mathrm{E}(X \,|\, Y = y)$

  Therefore the overall MSE is minimized for $g(Y) = \mathrm{E}(X \,|\, Y)$

  In fact, $\mathrm{E}(X \,|\, Y)$ minimizes the MSE conditioned on every $Y = y$ and not just its average over Y

  To find the minimum MSE, consider

$$\begin{aligned}\mathrm{E}\left[(X-\mathrm{E}(X \,|\, Y))^2\right] &= \mathrm{E}_Y\left(\mathrm{E}_X\left[(X-\mathrm{E}(X \,|\, Y))^2 \,|\, Y\right]\right) \\ &= \mathrm{E}_Y\left(\mathrm{Var}(X \,|\, Y)\right)\end{aligned}$$

Example

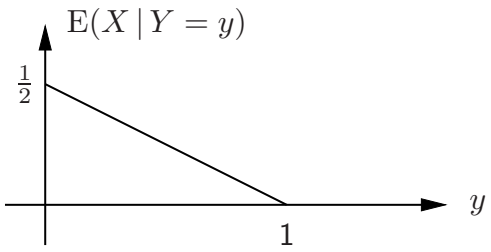
- Again let

$$f_{X,Y}(x,y) = \begin{cases} 2 & x \geq 0,\ y \geq 0,\ x+y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the MMSE estimate of X given Y and its MSE

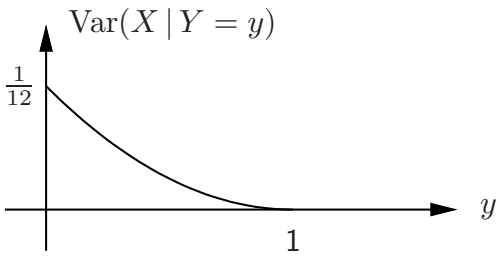
Solution: We aleady know that the MMSE estimate is given by

$$\mathrm{E}(X \,|\, Y) = \frac{1-Y}{2}, \quad 0 \leq Y \leq 1$$



And, for $Y = y$, the minimum MSE is given by

$$\mathrm{Var}(X \,|\, Y = y) = \frac{(1-y)^2}{12},\ 0 \leq y < 1$$



Thus the minimum MSE is $\mathrm{E}_Y(\mathrm{Var}(X \,|\, Y)) = \frac{1}{24}$, compared to $\mathrm{Var}(X) = \frac{1}{18}$

The difference is $\mathrm{Var}(\mathrm{E}(X \,|\, Y)) = \frac{1}{72}$, which is the variance of the estimate

The Additive Gaussian Noise Channel

- Consider a communication channel with input $X \sim \mathcal{N}(\mu, P)$, noise $Z \sim \mathcal{N}(0, N)$, and output $Y = X + Z$. X and Z are independent

  Find the MMSE estimate of X given Y and its MSE, i.e., $\mathrm{E}(X \,|\, Y)$ and $\mathrm{E}(\mathrm{Var}(X \,|\, Y))$

- To find $f_{X|Y}(x|y)$ we use Bayes rule:

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)}{f_Y(y)} f_X(x)$$

  We know that $X \sim \mathcal{N}(\mu, P)$, and since X and Z are independent and Gaussian, $Y = X + Z \sim \mathcal{N}(\mu,\, P + N)$ (to be proved later)

  To find $f_{Y|X}(y|x)$, we use "my" favorite trick: since Y is the sum of two independent r.v.s

$$f_{Y|X}(y|x) = f_{Z|X}(y - x|x) = f_Z(y - x) = \frac{1}{\sqrt{2\pi N}} e^{-\frac{(y-x)^2}{2N}}$$

  In other words, $Y \,|\, \{X = x\} \sim \mathcal{N}(x, N)$ (See Section 3.9 of G&D for details of trick.)

- Substituting in the Bayes rule formula, we finally obtain

$$f_{X|Y}(x|y) = \frac{1}{\sqrt{2\pi \frac{PN}{P+N}}} e^{-\frac{\left(x - (\frac{P}{P+N}y + \frac{N}{P+N}\mu)\right)^2}{2\frac{PN}{P+N}}}, \text{ that is,}$$

$$X \,|\, \{Y = y\} \sim \mathcal{N}\left(\frac{P}{P+N}y + \frac{N}{P+N}\mu,\, \frac{PN}{P+N}\right)$$

  Thus

$$\mathrm{E}(X \,|\, Y) = \frac{P}{P+N}Y + \frac{N}{P+N}\mu$$

$$\mathrm{E}(\mathrm{Var}(X \,|\, Y)) = \frac{PN}{P+N}$$

- Note: In the above two examples, the MMSE estimate turned out to be an affine function of Y (i.e., of the form $aY + b$)

  This is not always the case; for example, let

$$f(x|y) = \begin{cases} ye^{-yx} & x \geq 0,\, y > 0 \\ 0 & \text{otherwise} \end{cases}$$

  In this case $\mathrm{E}(X \,|\, Y) = 1/Y$

Linear Estimation

- To find the MMSE estimate one needs to know the statistics of the signal and the channel — $f_{X,Y}(x,y)$ — which is rarely the case in practice
- We typically have estimates only of the first and second moments of the signal and the observation, i.e., means, variances, and covariance of X and Y
- This is not, in general, sufficient information for computing the MMSE estimate, but as we shall see is enough to compute the MMSE linear (or affine) estimate of the signal X given the observation Y, i.e., the estimate of the form

$$\hat{X} = aY + b$$

that minimizes the mean square error

$$\text{MSE} = \text{E}\left[(X - \hat{X})^2\right]$$

The MMSE Linear Estimate

- Theorem: The MMSE linear estimate of X given Y is

$$\begin{aligned}\hat{X} &= \frac{\text{Cov}(X,Y)}{\sigma_Y^2}\big(Y - \text{E}(Y)\big) \; + \; \text{E}(X) \\ &= \rho_{X,Y}\sigma_X\left(\frac{Y - \text{E}(Y)}{\sigma_Y}\right) \; + \; \text{E}(X)\end{aligned}$$

and its MSE is given by

$$\text{MSE} \; = \; \sigma_X^2 - \frac{\text{Cov}^2(X,Y)}{\sigma_Y^2} \; = \; (1 - \rho_{X,Y}^2)\sigma_X^2$$

- Properties of best linear MSE estimate:
  - $\text{E}(\hat{X}) = \text{E}(X)$, i.e., estimate is unbiased (also true for best MSE estimate)
  - If $\rho_{X,Y} = 0$, i.e., X and Y are uncorrelated, then $\hat{X} = \text{E}(X)$ — the observation Y is ignored!
  - If $\rho_{X,Y} = \pm 1$, i.e., $(X - \text{E}(X))$ and $(Y - \text{E}(Y))$ are linearly dependent, then the linear estimate is perfect

- First proof (use calculus): To find the coefficients a and b we take derivatives and set them to 0

$$\text{MSE} = \text{E}\left[(X-\hat{X})^2\right] = \text{E}\left[(X-(aY+b))^2\right]$$

$$\frac{\partial}{\partial b}\text{MSE} = 0 \Rightarrow \text{E}(X-\hat{X}) = 0 \Rightarrow \text{E}(\hat{X}) = \text{E}(X)$$

$$\frac{\partial}{\partial a}\text{MSE} = 0 \Rightarrow \text{E}\left[(X-\hat{X})Y\right] = 0$$

Thus

$$\text{E}\left[[(X-\text{E}(X)) - (\hat{X}-\text{E}(\hat{X}))]\cdot[Y-\text{E}(Y)]\right] = 0$$

or

$$\text{E}\left[[(X-\text{E}(X)) - a(Y-\text{E}(Y))]\cdot[Y-\text{E}(Y)]\right] = 0$$

hence

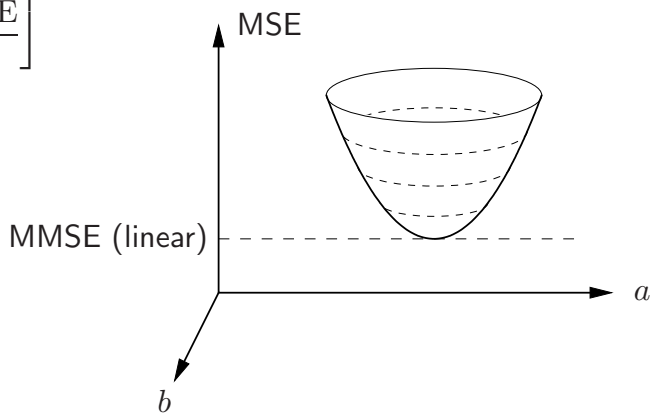
$$\text{Cov}(X,Y) - a\sigma_Y^2 = 0$$

Therefore

$$a = \frac{\text{Cov}(X,Y)}{\sigma_Y^2} \quad \text{and} \quad b = \text{E}(X) - \frac{\text{Cov}(X,Y)}{\sigma_Y^2}\text{E}(Y)$$

Note: These a and b *globally* minimize the MSE since the MSE is *convex* in a and b, which can be established by showing that the *Hessian* matrix is *nonnegative definite* (check it)

$$\mathcal{H} = \begin{bmatrix} \dfrac{\partial^2 \text{MSE}}{\partial a^2} & \dfrac{\partial^2 \text{MSE}}{\partial a \partial b} \\ \dfrac{\partial^2 \text{MSE}}{\partial a \partial b} & \dfrac{\partial^2 \text{MSE}}{\partial b^2} \end{bmatrix}$$



To find the MMSE, substitute a and b into the MSE expression $\text{E}\left[(X-(aY+b))^2\right]$

Alternative proof: The theorem can be proved directly without calculus.

Find a and b which minimize

$$E[(X - \hat{X}(Y))^2] = E[(X - aY - b)^2].$$

Rewrite with means removed:

$$\begin{aligned} & E\left([X - (aY + b)]^2\right) \\ &= E\left([(X - EX) - a(Y - EY) - (b - EX + aEY)]^2\right) \\ &= \sigma_X^2 + a^2\sigma_Y^2 + (b - EX + aEY)^2 - 2a\mathrm{Cov}(X, Y) \end{aligned}$$

(remaining cross-products are all zero).

First term does't depend on a or b, minimizing MSE $\Leftrightarrow$ minimizing

$$a^2\sigma_Y^2 + (b - EX + aEY)^2 - 2a\mathrm{Cov}(X, Y)$$

Middle term is nonnegative. Given a, minimizing b is $b = EY - aEX$.

Therefore best a must minimize

$$\begin{aligned} a^2\sigma_Y^2 - 2a\mathrm{Cov}(X, Y) &= (a\sigma_Y - \frac{\mathrm{Cov}(X, Y)}{\sigma_Y})^2 - (\frac{\mathrm{Cov}(X, Y)}{\sigma_Y})^2 \\ &\geq -(\frac{\mathrm{Cov}(X, Y)}{\sigma_Y})^2 \end{aligned}$$

Hence global minimum achieved by

$$a = \frac{\mathrm{Cov}(X, Y)}{\sigma_Y^2}$$

$\Rightarrow$ best b is

$$b = EX - \frac{\mathrm{Cov}(X, Y)}{\sigma_Y^2}EY.$$

$$\mathrm{MMSE} = \sigma_X^2 - (\frac{\mathrm{Cov}(X, Y)}{\sigma_Y})^2$$

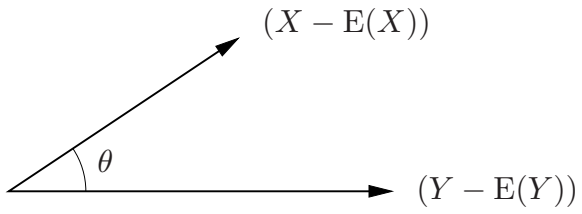
Geometric Formulation of Linear Estimation

- First we introduce some needed background
- A *vector space* $\mathcal{V}$, e.g., Euclidean space, consists of a set of vectors that are closed under two operations:
  - *vector addition*: if $v_1, v_2 \in \mathcal{V}$ then $v_1 + v_2 \in \mathcal{V}$
  - *scalar multiplication*: if $a \in \mathbf{R}$ and $v \in \mathcal{V}$, then $av \in \mathcal{V}$
- An *inner product*, e.g., dot product in Euclidean space, is a real-valued operation $u \cdot v$ satisfying these three conditions:
  - commutativity: $u \cdot v = v \cdot u$
  - linearity: $(au + bv) \cdot w = a(u \cdot w) + b(v \cdot w)$
  - nonnegativity: $u \cdot u \geq 0$ and $u \cdot u = 0$ iff $u = 0$
- The *norm* of u is defined as $\|u\| = \sqrt{u \cdot u}$
- u and v are *orthogonal* (written $u \perp v$) if $u \cdot v = 0$
- A vector space with an inner product is called an *inner product space*
  Example: Euclidean space with dot product

Back to Linear Estimation

- View $(X - \mathrm{E}(X))$ and $(Y - \mathrm{E}(Y))$ as vectors in an inner product space $\mathcal{V}$ that consists of all zero mean random variables defined over the same probability space, with
  - vector addition: $V_1 + V_2 \in \mathcal{V}$
    adding two zero mean r.v.s yields a zero mean r.v.
  - scalar multiplication: $aV \in \mathcal{V}$
    multiplying a zero mean r.v. by a constant yields a zero mean r.v.
  - inner product: $\mathrm{E}(V_1 V_2)$
    exercise: check that this is a legitimate inner product
  - norm of V: $\|V\| = \sqrt{\mathrm{E}(V^2)} = \sigma_V$

- So we have the following picture for the r.v.s $(X - \mathrm{E}(X))$ and $(Y - \mathrm{E}(Y))$:

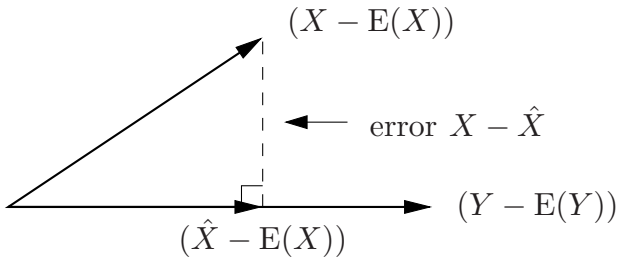


inner product	$\Leftrightarrow$	$\mathrm{Cov}(X, Y)$
norm of $(X - \mathrm{E}(X))$	$\Leftrightarrow$	σ_X
norm of $(Y - \mathrm{E}(Y))$	$\Leftrightarrow$	σ_Y
$\cos\theta$	$\Leftrightarrow$	$\rho_{X,Y}$

Note that $(X - \mathrm{E}(X))$ and $(Y - \mathrm{E}(Y))$ can live in a vector space of very high dimension. We are concerned only with the two dimensional subspace spanned by these two vectors

Orthogonality Principle

- The linear estimation problem can now be recast as a geometry problem



Find a vector $(\hat{X} - \mathrm{E}(X)) = a(Y - \mathrm{E}(Y))$ that minimizes $\|X - \hat{X}\|$

- Clearly $(X - \hat{X}) \perp (Y - \mathrm{E}(Y))$ minimizes $\|X - \hat{X}\|$, i.e.,

$$\mathrm{E}\left((X - \hat{X})(Y - \mathrm{E}(Y)\right) = 0 \Rightarrow a = \frac{\mathrm{Cov}(X, Y)}{\sigma_Y^2}$$

- This argument is called the *orthogonality principle*. Later we will see that it is key to deriving the minimum MSE linear estimate in more complex settings

Linear vs. MMSE (Nonlinear) Estimate

- The linear estimate is not, in general, as good as the MMSE estimate
- Example: Let $Y \sim \mathrm{U}[-1,1]$ and $X = Y^2$

  The MMSE estimate of X given Y is Y^2 — perfect!

  To find the MMSE linear estimate we compute

$$\begin{aligned} \mathrm{E}(Y) &= 0 \\ \mathrm{E}(X) &= \int_{-1}^{1} \tfrac{1}{2} y^2 \, dy = \tfrac{1}{3} \\ \mathrm{Cov}(X,Y) &= \mathrm{E}(XY) - 0 = \mathrm{E}(Y^3) = 0 \end{aligned}$$

  Thus the MMSE linear estimate $\hat{X} = \mathrm{E}(X) = \frac{1}{3}$, i.e., the observation Y is totally ignored, even though it completely determines X!
- There is a very important class of r.v.s for which the MMSE estimate is linear, the *jointly Gaussian* random variables

Jointly Gaussian Random Variables

- Two r.v.s are *jointly Gaussian* if their joint pdf is of the form

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}} \, e^{-\frac{1}{2(1-\rho_{X,Y}^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - 2\rho_{X,Y}\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right)}$$

- The pdf is a function only of μ_X, μ_Y, σ_X^2, σ_Y^2, and $\rho_{X,Y}$
- Note: In Lecture Notes 6 we shall define this in a more general way
- Example: For the additive Gaussian noise channel, where $X \sim \mathcal{N}(\mu, P)$ and $Z \sim \mathcal{N}(0, N)$ are independent and $Y = X + Z$, show that (a) X and Z are jointly Gaussian, and (b) X and Y are jointly Gaussian

  Solution: (a) It is easy to show that if two Gaussian r.v.s are independent, their joint pdf has the above form with $\rho_{X,Y} = 0$. (b) Now consider

$$\begin{aligned} f(x,y) &= f_X(x) f_{Y|X}(y|x) \\ &= f_X(x) f_{Z|X}(y-x|x) = f_X(x) f_Z(y-x) \end{aligned}$$

  Now we can write $f(x,y)$ in the form of a jointly Gaussian pdf

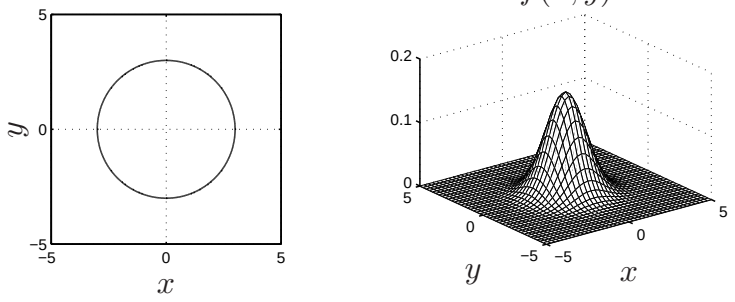
- If X and Y are jointly Gaussian, contours of equal joint pdf are ellipses defined by the quadratic equation

$$\frac{(x-\mu_X)^2}{\sigma_X^2}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}-2\rho_{X,Y}\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}=c\geq 0$$

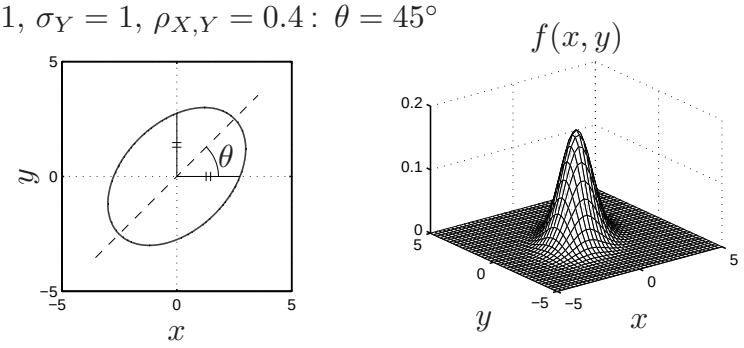
- Examples: In the following examples we plot contours of equal joint pdf $f(x,y)$ for zero mean jointly Gaussian r.v.s for different values of σ_X, σ_Y, and $\rho_{X,Y}$

  The orientation of the major axis of the ellipse is $\theta=\frac{1}{2}\arctan\left(\frac{2\rho_{X,Y}\sigma_X\sigma_Y}{\sigma_X^2-\sigma_Y^2}\right)$

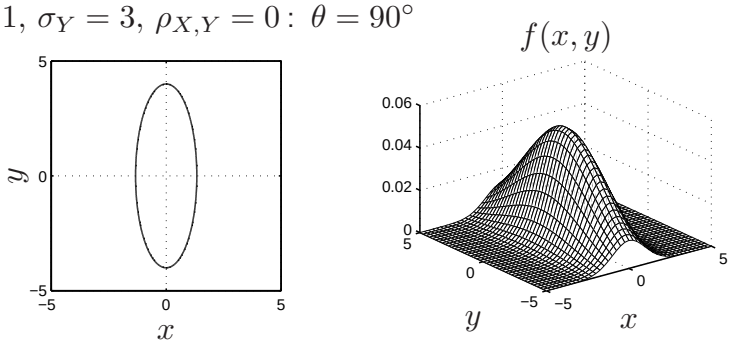
$\sigma_X=1,\ \sigma_Y=1,\ \rho_{X,Y}=0$



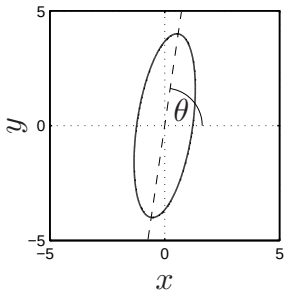
$\sigma_X=1,\ \sigma_Y=1,\ \rho_{X,Y}=0.4:\ \theta=45°$

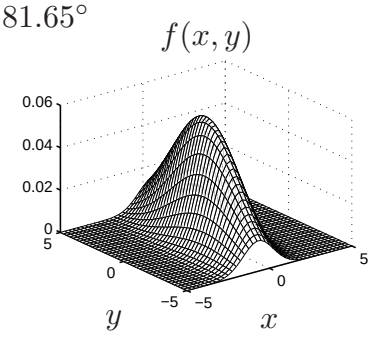


$\sigma_X=1,\ \sigma_Y=3,\ \rho_{X,Y}=0:\ \theta=90°$

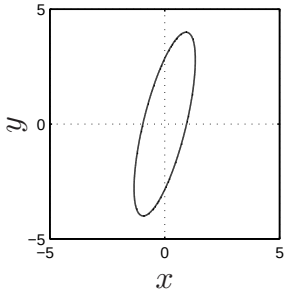


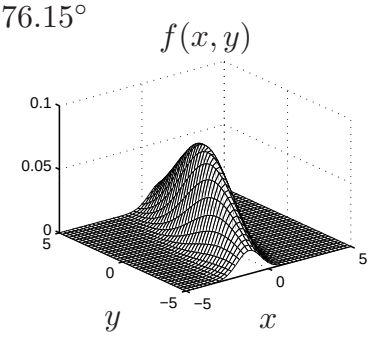
$\sigma_X = 1,\ \sigma_Y = 3,\ \rho_{X,Y} = 0.4 :\ \theta = 81.65°$



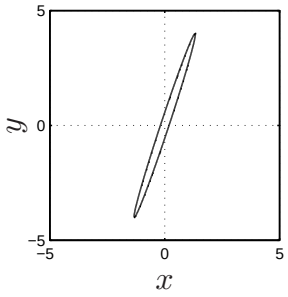


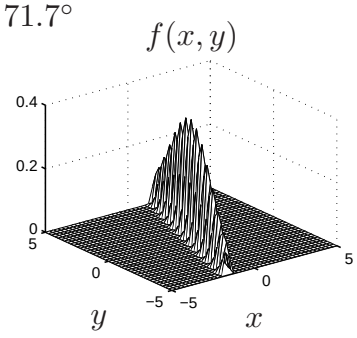
$\sigma_X = 1,\ \sigma_Y = 3,\ \rho_{X,Y} = 0.7 :\ \theta = 76.15°$



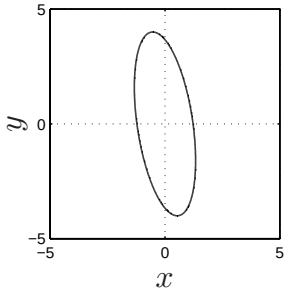


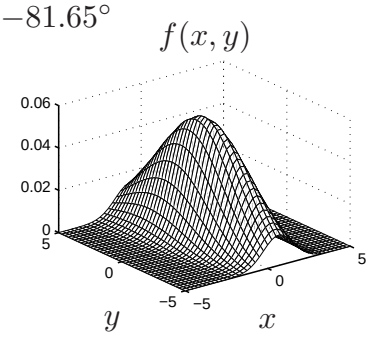
$\sigma_X = 1,\ \sigma_Y = 3,\ \rho_{X,Y} = 0.99 :\ \theta = 71.7°$





$\sigma_X = 1,\ \sigma_Y = 3,\ \rho_{X,Y} = -0.4 :\ \theta = -81.65°$





Properties of Jointly Gaussian Random Variables

- If X and Y are jointly Gaussian, they are individually Gaussian, i.e., the marginals of $f_{X,Y}(x,y)$ are Gaussian, i.e.,

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2), \quad Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$

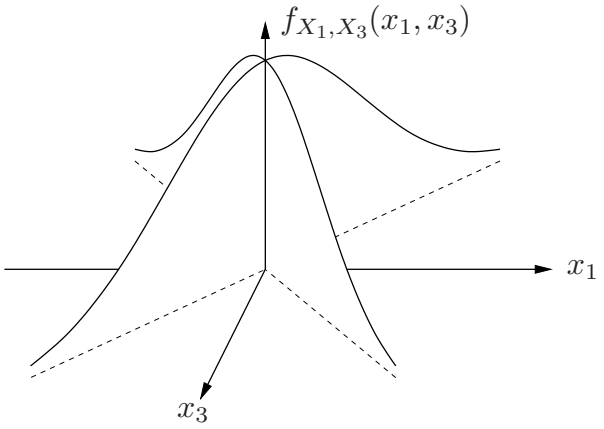
- The converse is not necessarily true, i.e., Gaussian marginals do not necessarily mean that the r.v.s are jointly Gaussian

  Example: Let $X_1 \sim \mathcal{N}(0,1)$ and

$$X_2 = \begin{cases} +1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2} \end{cases}$$

  be independent r.v.s, and let $X_3 = X_1 X_2$

  - Clearly, $X_3 \sim \mathcal{N}(0,1)$
  - However, X_1, X_3 do not have a joint pdf. Using delta funcions, "$f_{X_1,X_3}(x_1,x_3)$" has the form shown in the following figure



- If X and Y are jointly Gaussian, the conditional pdf is Gaussian:

$$X \,|\, \{Y = y\} \sim \mathcal{N}\Big(\rho_{X,Y}\sigma_X \frac{(y-\mu_Y)}{\sigma_Y} + \mu_X,\ (1-\rho_{X,Y}^2)\sigma_X^2\Big),$$

  which shows that the MMSE estimate is linear

- If X and Y are jointly Gaussian and uncorrelated, i.e., $\rho_{X,Y} = 0$, then they are also independent