Lecture Notes 2
Expectation

- Definition and Properties
- Mean and Variance
- Markov and Chebychev Inequalites
- Expectations involving two random variables
- Scalar MSE Estimation
- Scalar Linear Estimation
- Jointly Gaussian random variables

Expectation

- Let $X \in \mathcal{X}$ be a discrete r.v. with pmf $p_X(x)$ and let $g(x)$ be a function of x. The expectation (or expected value or mean) of $g(X)$ can be defined as

$$\mathrm{E}(g(X)) = \sum_{x \in \mathcal{X}} g(x) p_X(x)$$

- For a continuous r.v. $X \sim f_X(x)$, the expected value of $g(X)$ can be defined as

$$\mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$$

- Expectation is linear, i.e., for any constant a

$$\mathrm{E}[a g_1(X) + g_2(X)] = a\,\mathrm{E}(g_1(X)) + \mathrm{E}(g_2(X))$$

  In particular, $\mathrm{E}(a) = a$

- Remark: We know that a r.v. is completely specified by its cdf (pdf, pmf), so why do we need expectation?
  - Expectation provides a summary or an estimate of the r.v. — a single number — instead of specifying the entire distribution
  - It is far easier to estimate the expectation of a r.v. from data than to estimate its distribution
  - Expectation can be used to bound or estimate probabilities of interesting events (as we shall see)

Mean and Variance

- The first moment (or mean) of $X \sim f_X(x)$ is
$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx$$
- The second moment (or mean squared or average power) of X is
$$\mathrm{E}(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\,dx$$
- The variance of X is
$$\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}(X))^2\right] = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$$
Hence $\mathrm{E}(X^2) \geq (\mathrm{E}(X))^2$
- The standard deviation of X is defined as $\sigma_X = \sqrt{\mathrm{Var}(X)}$, i.e., $\mathrm{Var}(X) = \sigma_X^2$
- In general, the kth moment (k a positive integer) is
$$\mathrm{E}(X^k) = \int_{-\infty}^{\infty} x^k f_X(x)\,dx$$

- Mean and Variance for Famous RVs:

Random Variable	Mean	Variance
$\mathrm{Bern}(p)$	p	$p(1-p)$
$\mathrm{Geom}(p)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$\mathrm{Binom}(n,p)$	np	$np(1-p)$
$\mathrm{Poisson}(\lambda)$	λ	λ
$\mathrm{U}[\,a,b\,]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$\mathrm{Exp}(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
$\mathrm{Laplace}(\lambda)$	0	$\frac{2}{\lambda^2}$
$\mathcal{N}(\mu,\sigma^2)$	μ	σ^2

Expectation Can Be Infinite or May Not Exist

- Expectation can be infinite. For example

$$f_X(x) = \begin{cases} 1/x^2 & 1 \le x < \infty \\ 0 & \text{otherwise} \end{cases} \Rightarrow \mathrm{E}(X) = \int_1^\infty x/x^2 \, dx = \infty$$

- Expectation may not exist. To find conditions for expectation to exist, consider

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x) \, dx = -\int_{-\infty}^{0} |x| f_X(x) \, dx + \int_0^\infty |x| f_X(x) \, dx \,,$$

so either $\int_{-\infty}^{0} |x| f_X(x) \, dx$ or $\int_0^\infty |x| f_X(x) \, dx$ must be finite

- Example: The standard Cauchy r.v. has the pdf

$$f(x) = \frac{1}{\pi(1 + x^2)}$$

Since both $\int_{-\infty}^{0} |x| f(x) \, dx$ and $\int_0^\infty |x| f(x) \, dx$ are infinite, its mean does not exist! (The second moment of the Cauchy is $\mathrm{E}(X^2) = \infty$, so it exists)

Bounding Probability Using Expectation

- In many cases we do not know the distribution of a r.v. X but want to find the probability of an event such as $\{X > a\}$ or $\{|X - \mathrm{E}(X)| > a\}$
- The Markov and Chebyshev inequalities give upper bounds on the probabilities of such events in terms of the mean and variance of the random variable
- Example: Let $X \geq 0$ represent the age of a person in the Bay Area. If we know that $\mathrm{E}(X) = 35$ years, what fraction of the population is ≥ 70 years old?

  Clearly we cannot answer this question knowing only the mean, but we can say that $\mathrm{P}\{X \geq 70\} \leq 0.5$, since otherwise the mean would be larger than 35
- This is an application of the Markov inequality

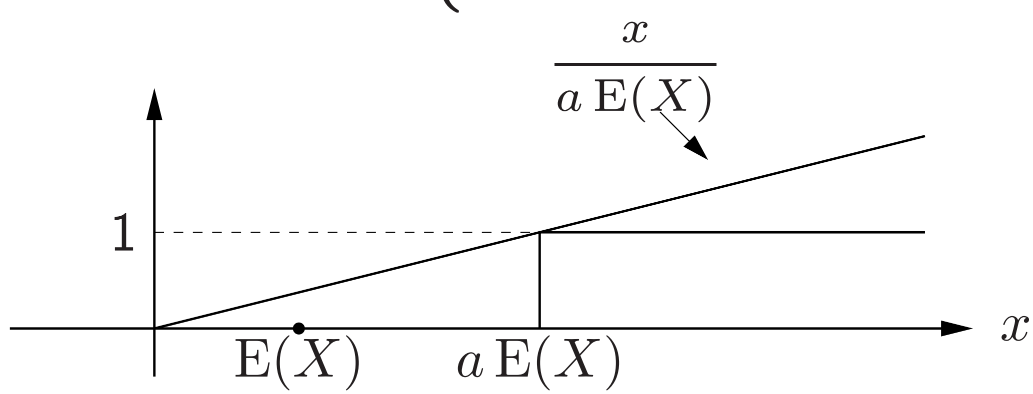
Markov Inequality

- For any r.v. $X \geq 0$ with finite mean $\mathrm{E}(X)$ and any $a > 1$,

$$\mathrm{P}\{X \geq a\,\mathrm{E}(X)\} \leq \frac{1}{a}$$

Proof: Define the indicator function of the set $A = \{x \geq a\,\mathrm{E}(X)\}$:

$$\mathrm{I}_A(x) = \begin{cases} 1 & x \geq a\,\mathrm{E}(X) \\ 0 & \text{otherwise} \end{cases}$$



Clearly, $\mathrm{I}_A \leq \dfrac{X}{a\,\mathrm{E}(X)}$

Since $\mathrm{E}(\mathrm{I}_A) = \mathrm{P}(A) = \mathrm{P}\{X \geq a\,\mathrm{E}(X)\}$, taking the expectations of both sides we obtain the Markov Inequality

Chebyshev Inequality

- Let X be a device parameter in an integrated circuit (IC) with known mean and variance. The IC is out-of-spec if X is more than, say, $3\sigma_X$ away from its mean. We wish to find the fraction of out-of-spec ICs, namely, $\mathrm{P}\{|X - \mathrm{E}(X)| \geq 3\sigma_X\}$

  The Chebyshev inequality gives us an upper bound on this fraction in terms the mean and variance of X

- Let X be a r.v. with known $\mathrm{E}(X)$ and $\mathrm{Var}(X) = \sigma_X^2$. The Chebyshev inequality states that for every $a > 1$,

$$\mathrm{P}\{|X - \mathrm{E}(X)| \geq a\sigma_X\} \leq \frac{1}{a^2}$$

  Proof: We use the Markov inequality. Define the r.v. $Y = (X - \mathrm{E}(X))^2 \geq 0$. Since $\mathrm{E}(Y) = \sigma_X^2$, the Markov inequality gives

$$\mathrm{P}\{Y \geq a^2\sigma_X^2\} \leq \frac{1}{a^2}$$

  But $\{|X - \mathrm{E}(X)| \geq a\sigma_X\}$ occurs iff $\{Y \geq a^2\sigma_X^2\}$. Thus

$$\mathrm{P}\{|X - \mathrm{E}(X)| \geq a\sigma_X\} \leq \frac{1}{a^2}$$

Expectation Involving Two RVs

- Let $(X, Y) \sim f_{X,Y}(x, y)$ and let $g(x, y)$ be a function of x and y. The expectation of $g(X, Y)$ is given by

$$\mathrm{E}(g(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y)\, dx\, dy$$

  The function $g(X, Y)$ may be X, Y, X^2, $X + Y$, etc.

- The correlation of X and Y is defined as $\mathrm{E}(XY)$

  X and Y are said to be orthogonal if $\mathrm{E}(XY) = 0$

- The covariance of X and Y is defined as

$$\mathrm{Cov}(X, Y) = \mathrm{E}\left[(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))\right] = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)$$

  X and Y are said to be uncorrelated if $\mathrm{Cov}(X, Y) = 0$

- Note that $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$

- If X and Y are independent then they are uncorrelated

- X and Y uncorrelated does not necessarily imply that they are independent

- The correlation coefficient of X and Y is defined as

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

  Fact: $|\rho_{X,Y}| \leq 1$ with equality iff $(X - \mathrm{E}(X))$ is a linear function of $(Y - \mathrm{E}(Y))$

  The correlation coefficient is a measure of how closely $(X - \mathrm{E}(x))$ can be approximated by a linear function of $(Y - \mathrm{E}(Y))$ (more on this soon)

Conditional Expectation

- Let $(X, Y) \sim f_{X,Y}(x, y)$. If $f_Y(y) \neq 0$, the conditional pdf of X given $Y = y$ is given by
$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$
- We know that $f_{X|Y}(x|y)$ is a pdf for X (function of y), so we can define the expectation of any function $g(X, Y)$ w.r.t. $f_{X|Y}(x|y)$ as
$$\mathrm{E}(g(X, Y) \,|\, Y = y) = \int_{-\infty}^{\infty} g(x, y) f_{X|Y}(x|y)\, dx$$
- Example: If $g(X, Y) = X$, then the conditional expectation of X given $Y = y$ is
$$\mathrm{E}(X \,|\, Y = y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, dx$$
- Example: If $g(X, Y) = XY$, then $\mathrm{E}(XY \,|\, Y = y) = y\, \mathrm{E}(X \,|\, Y = y)$
- We define the conditional expectation of $g(X, Y)$ given Y as the random variable $\mathrm{E}(g(X, Y) \,|\, Y)$, which is a function of the random variable Y

- In particular, $\mathrm{E}(X \mid Y)$ is the conditional expectation of X given Y, a r.v. that is a function of Y

- Iterated expectation: In general we can find $\mathrm{E}(g(X,Y))$ as

$$\mathrm{E}(g(X,Y)) = \mathrm{E}_Y\left[\mathrm{E}_X(g(X,Y) \mid Y)\right],$$

  where E_X means expectation w.r.t. $f_{X|Y}(x|y)$ and E_Y means expectation w.r.t. $f_Y(y)$

- Example: Coin with random bias. A coin with random bias P such that $\mathrm{E}(P) = 1/3$ is flipped n times independently. Let X be the number of heads. Find $\mathrm{E}(X)$

Conditional Variance

- Let X and Y be two r.v.s. We define the conditional variance of X given $Y = y$ to be the variance of X using $f_{X|Y}(x|y)$, i.e.,

$$\begin{aligned} \mathrm{Var}(X \mid Y = y) &= \mathrm{E}\left[(X - \mathrm{E}(X \mid Y = y))^2 \mid Y = y\right] \\ &= \mathrm{E}(X^2 \mid Y = y) - \left[\mathrm{E}(X \mid Y = y)\right]^2 \end{aligned}$$

- The r.v. $\mathrm{Var}(X \mid Y)$ is simply a function of Y that takes on the values $\mathrm{Var}(X \mid Y = y)$. Its expected value is

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

- Since $\mathrm{E}(X \mid Y)$ is a r.v., it has a variance

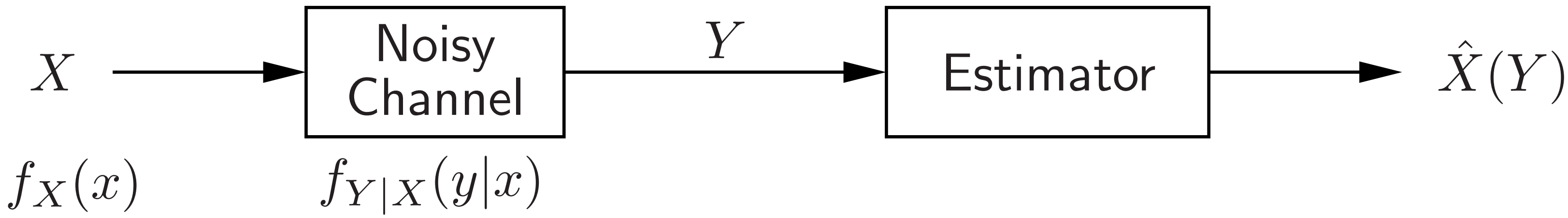
$$\mathrm{Var}(\mathrm{E}(X \mid Y)) = \mathrm{E}_Y\left[\left(\mathrm{E}(X \mid Y) - \mathrm{E}[\mathrm{E}(X \mid Y)]\right)^2\right] = \mathrm{E}\left[(\mathrm{E}(X \mid Y))^2\right] - (\mathrm{E}(X))^2$$

- Law of Conditional Variances: Adding the above expressions, we obtain

$$\mathrm{Var}(X) = \mathrm{E}\left(\mathrm{Var}(X \mid Y)\right) + \mathrm{Var}\left(\mathrm{E}(X \mid Y)\right)$$

Scalar 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 channel output Y and wish to find the estimate $\hat{X}(Y)$ of X that minimizes the mean squared error

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

- The $\hat{X}$ that achieves the minimum MSE is called the MMSE estimate 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

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

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

$$\mathrm{MMSE} = \mathrm{E}_Y(\mathrm{Var}(X \mid Y)) = \mathrm{Var}(X) - \mathrm{Var}\left(\mathrm{E}(X \mid Y)\right)$$

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

- The estimation error and the estimate are orthogonal

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

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

- MMSE estimate is linear: Let $X = aU + V$ and $\hat{U}$ and $\hat{V}$ be the MMSE estimates of U and V, respectively

  Then, the MMSE estimate of X is

$$
\hat{X} = a\hat{U} + \hat{V}
$$

- 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 MMSE 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 $\mathrm{E}(X \mid Y)$ is the MMSE estimate of X given Y

First write

$$\mathrm{E}\left[(X-\hat{X}(Y))^2\right] \;=\; \mathrm{E}_Y\left[\mathrm{E}_X((X-\hat{X}(Y))^2 \mid Y)\right]$$

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

Therefore the overall MSE is minimized for $\hat{X}(Y) = \mathrm{E}(X \mid Y)$

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

To find the minimum MSE, consider

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

- Finally, by the law of conditional variance,

$$\mathrm{E}(\mathrm{Var}(X \mid Y)) = \mathrm{Var}(X) - \mathrm{Var}(\mathrm{E}(X \mid Y)),$$

i.e., the minimum MSE is the difference between the variance of the signal and the variance of the MMSE estimate

The Additive Gaussian Noise Channel

- Consider a noisy 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., $\mathrm{E}(X \mid Y)$ and $\mathrm{E}(\mathrm{Var}(X \mid Y))$

- 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 $f_{Y|X}(y|x)$, since Y is the sum of two independent r.v.s, we have

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

  In other words, $Y \mid \{X = x\} \sim \mathcal{N}(x, N)$

- Substituting in the Bayes rule formula, we finally obtain

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

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

Thus

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

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

Scalar 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 squared error

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

The MMSE Linear Estimate

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

$$\hat{X} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}\big(Y - \mathrm{E}(Y)\big) \;+\; \mathrm{E}(X)$$
$$= \rho_{X,Y}\sigma_X \left(\frac{Y - \mathrm{E}(Y)}{\sigma_Y}\right) \;+\; \mathrm{E}(X)$$

  and its MSE is

$$\mathrm{MSE} \;=\; \mathrm{Var}(X) - \frac{\mathrm{Cov}^2(X,Y)}{\mathrm{Var}(Y)} \;=\; (1 - \rho^2_{X,Y})\mathrm{Var}(X)$$

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

- MMSE linear estimate is linear: Let $X = \alpha U + V$ and $\hat{U}$ and $\hat{V}$ be the MMSE linear estimates of U and V, respectively

  Then, the MMSE linear estimate of X is

$$\hat{X} = \alpha\hat{U} + \hat{V}$$

- Proof of the Theorem:
  - Write the minimization problem as:

$$\min_a \min_b \mathrm{E}[((X - aY) - b)^2]$$

    From MMSE estimate derivation with no observation, we know that the MMSE estimate of $(X - aY)$ is its mean $\mathrm{E}(X) - a\,\mathrm{E}(Y)$

  - Hence we can replace b with $\mathrm{E}(X) - a\,\mathrm{E}(Y)$, which reduces the linear estimation problem to finding the coefficient a that minimizes

$$\mathrm{E}[((X - \mathrm{E}(X)) - a(Y - \mathrm{E}(Y))]^2 = \mathrm{E}[((X - \mathrm{E}(X)) - (\hat{X} - \mathrm{E}(X))]^2,$$

    i.e., the problem reduces to finding $(\hat{X} - \mathrm{E}(X)) = a(Y - \mathrm{E}(Y))$ that minimizes the MSE

  - This problem can be solved using calculus. Instead we use a geometric argument that will help us solve more involved linear estimation problems

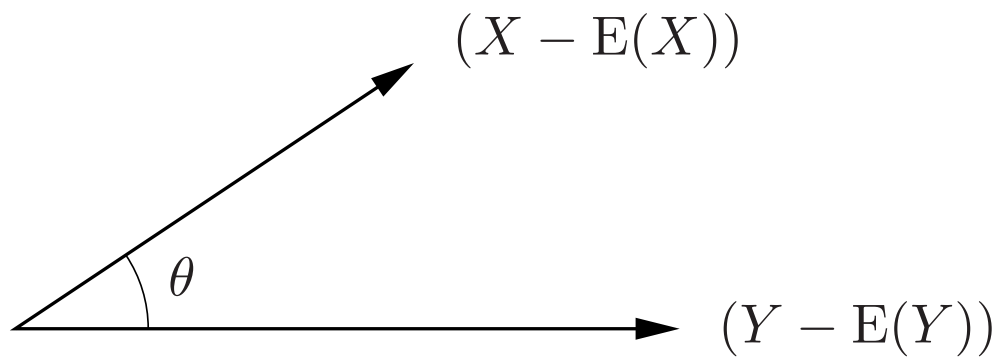
Inner product space

- 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 the three conditions:
  - commutativity: $u \cdot v = v \cdot u$
  - linearity: $(au + v) \cdot w = a(u \cdot w) + v \cdot w$
  - nonnegativity: $u \cdot u \geq 0$ and $u \cdot u = 0$ iff $u = 0$
- A vector space with an inner product is called an inner product space
  Example: Euclidean space with dot product
- The norm of u is defined as $\|u\| = \sqrt{u \cdot u}$
- Angle between vectors u and v: $\theta = \arccos(u \cdot v / \|u\| \times \|v\|)$
- u and v are orthogonal (written $u \perp v$) if $u \cdot v = 0$

Geometric Formulation of 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 $(Y - \mathrm{E}(Y))$:

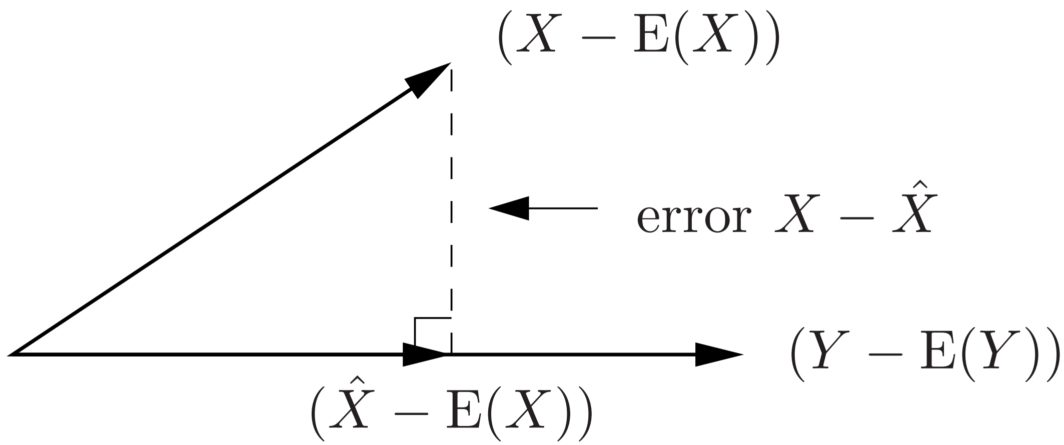


inner product	$\Leftrightarrow$	$\mathrm{Cov}(X, Y)$
norm of $(X - \mathrm{E}(X))$	$\Leftrightarrow$	σ_X
norm of $(Y - \mathrm{E}(Y))$	$\Leftrightarrow$	σ_Y
$\cos\theta$	$\Leftrightarrow$	$\rho_{X,Y}$

Note that $(X - \mathrm{E}(X))$ and $(Y - \mathrm{E}(Y))$ can live in a vector space of very high dimension. We are interested only in the 2-dimensional subspace spanned by these two vectors

Orthogonality Principle

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



  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.,

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

- Minimum MSE $= \sigma_X^2 \sin^2\theta = \sigma_X^2(1 - \cos^2\theta) = \sigma_X^2(1 - \rho_{X,Y}^2)$

- This argument is called the orthogonality principle. Later we will see that it is key to deriving the MMSE 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 $\hat{X} = \mathrm{E}(X) = 1/3$, 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 3 we will 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

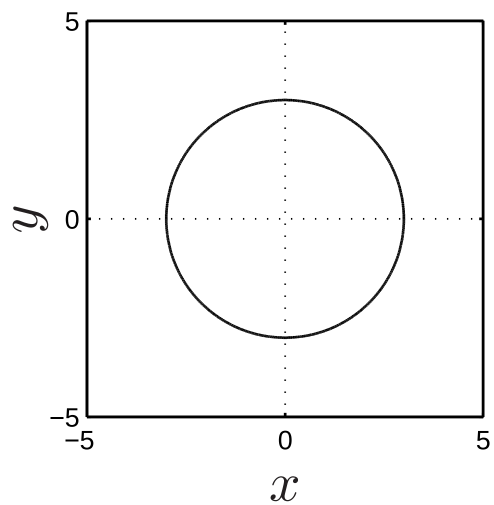
- If X and Y are jointly Gaussian, contours of equal joint pdf are ellipses defined by the quadratic equation

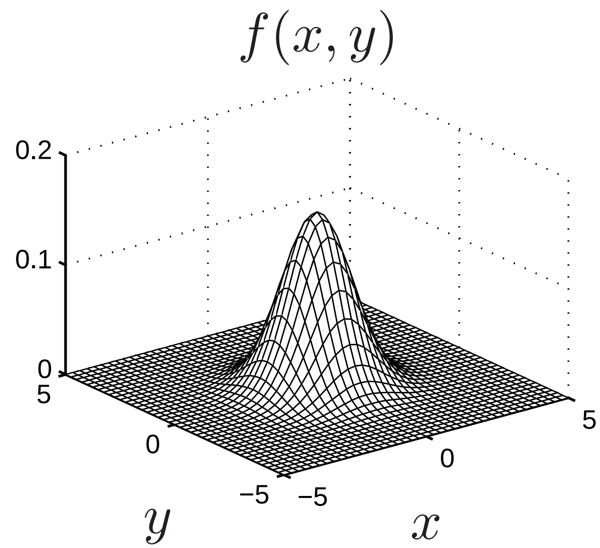
$$\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - 2\rho_{X,Y}\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} = c \geq 0$$

- 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 $\rho_{X,Y}$

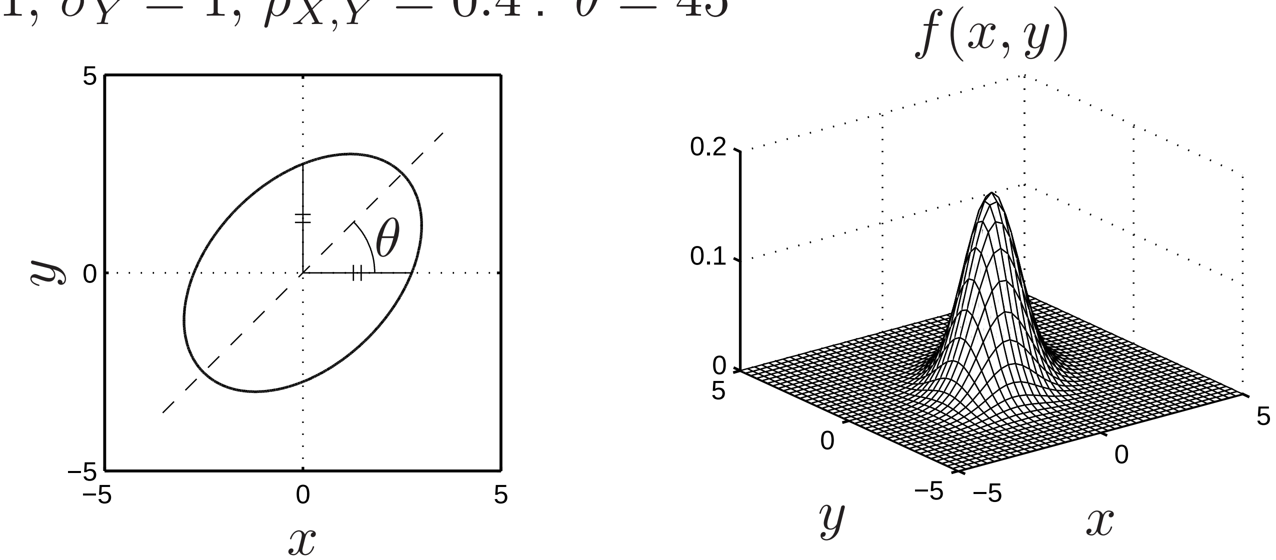
  The orientation of the major axis of the ellipse is $\theta = \frac{1}{2}\arctan\left(\frac{2\rho_{X,Y}\sigma_X\sigma_Y}{\sigma_X^2 - \sigma_Y^2}\right)$

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

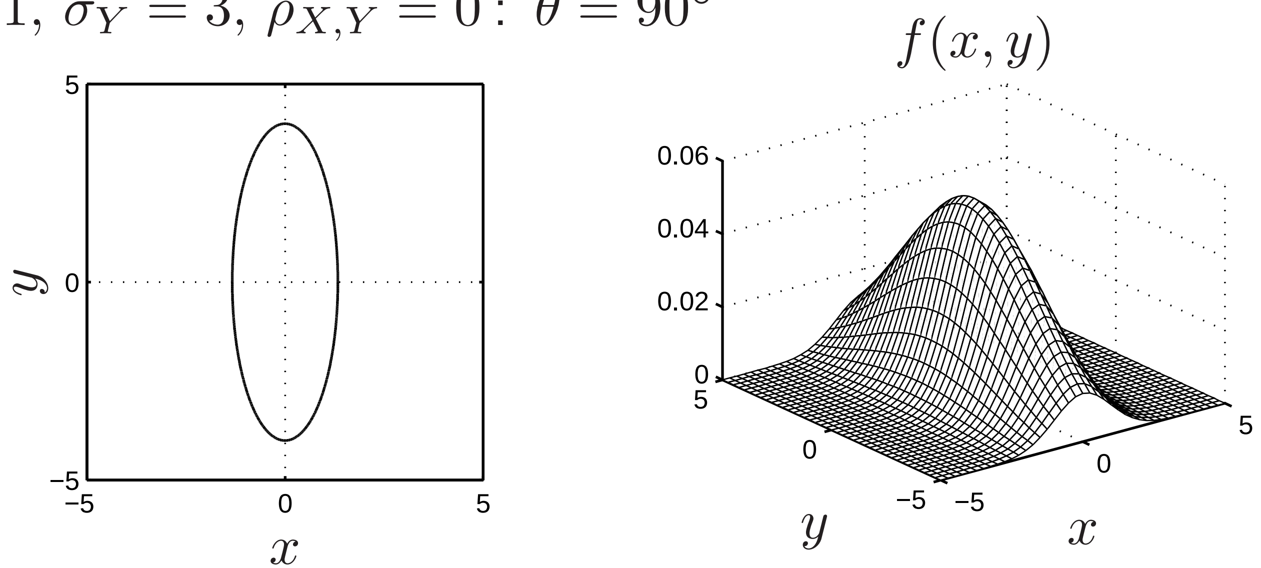




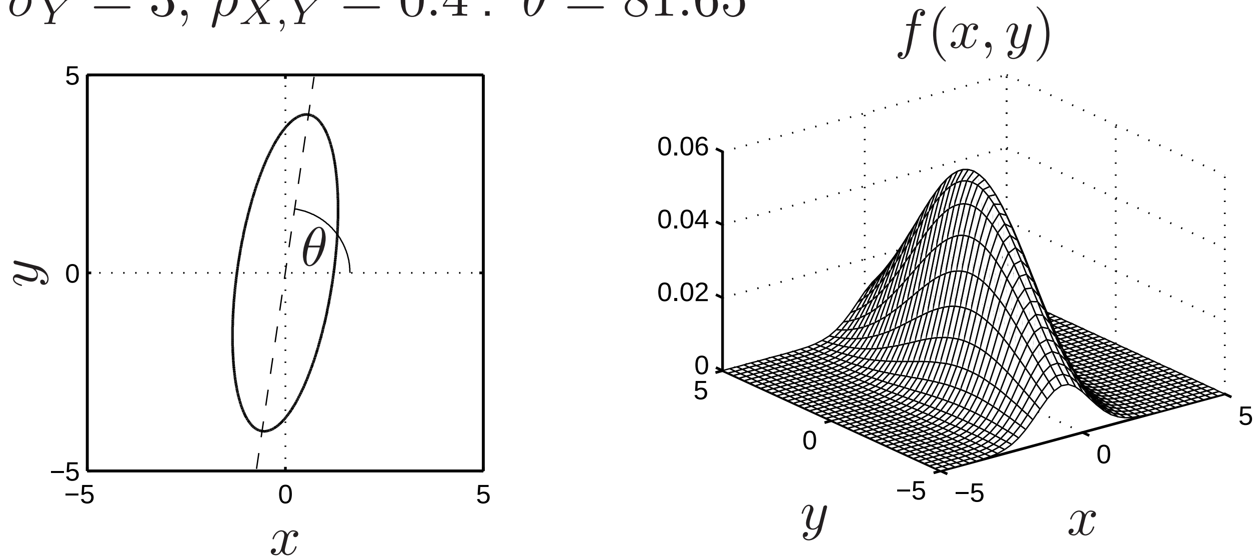
$\sigma_X = 1,\ \sigma_Y = 1,\ \rho_{X,Y} = 0.4:\ \theta = 45^\circ$



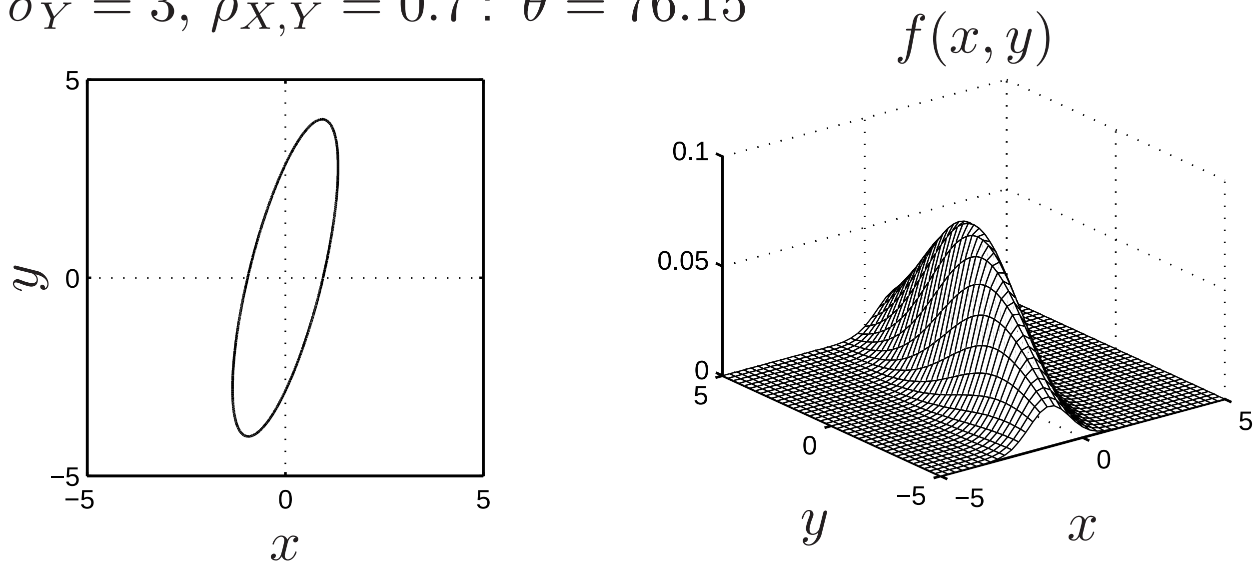
$\sigma_X = 1,\ \sigma_Y = 3,\ \rho_{X,Y} = 0:\ \theta = 90^\circ$



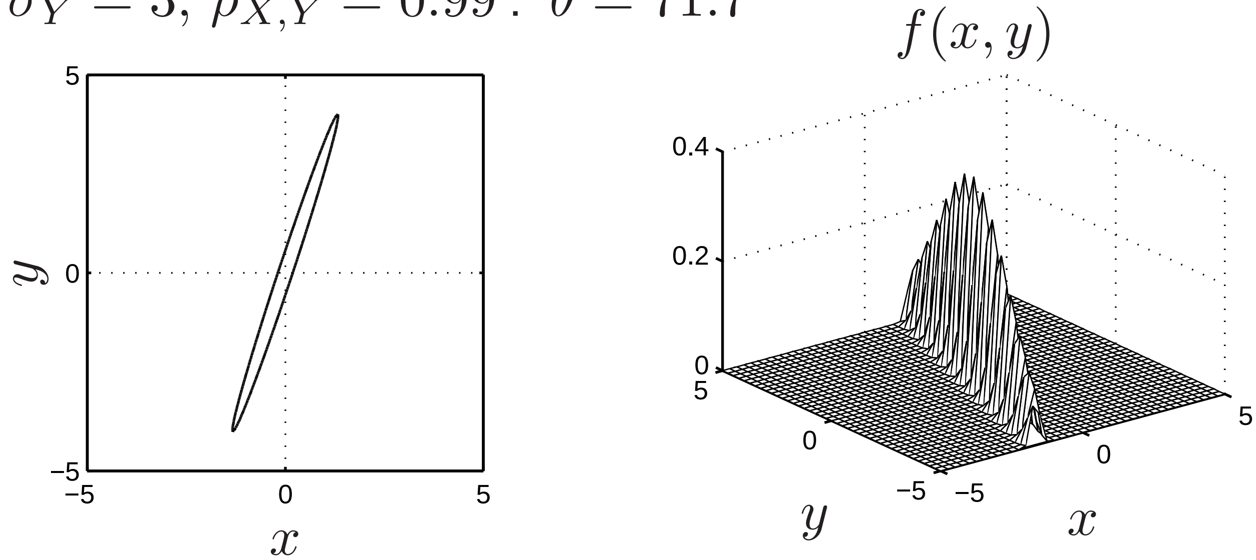
$\sigma_X = 1,\ \sigma_Y = 3,\ \rho_{X,Y} = 0.4:\ \theta = 81.65^\circ$



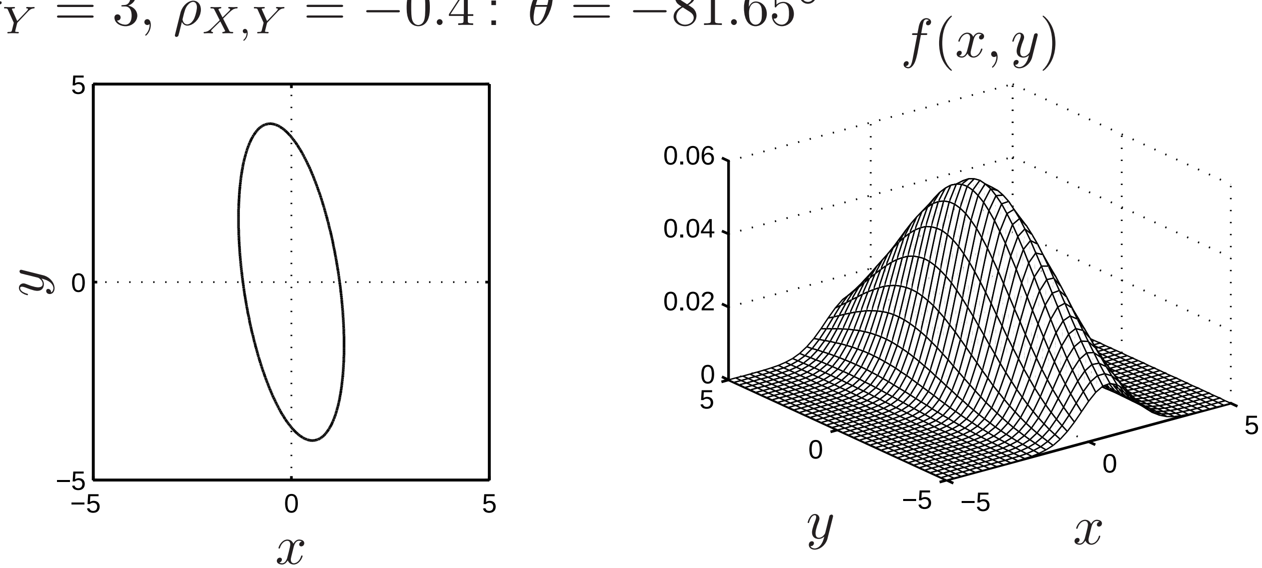
$\sigma_X = 1,\ \sigma_Y = 3,\ \rho_{X,Y} = 0.7:\ \theta = 76.15^\circ$



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



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



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.,

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

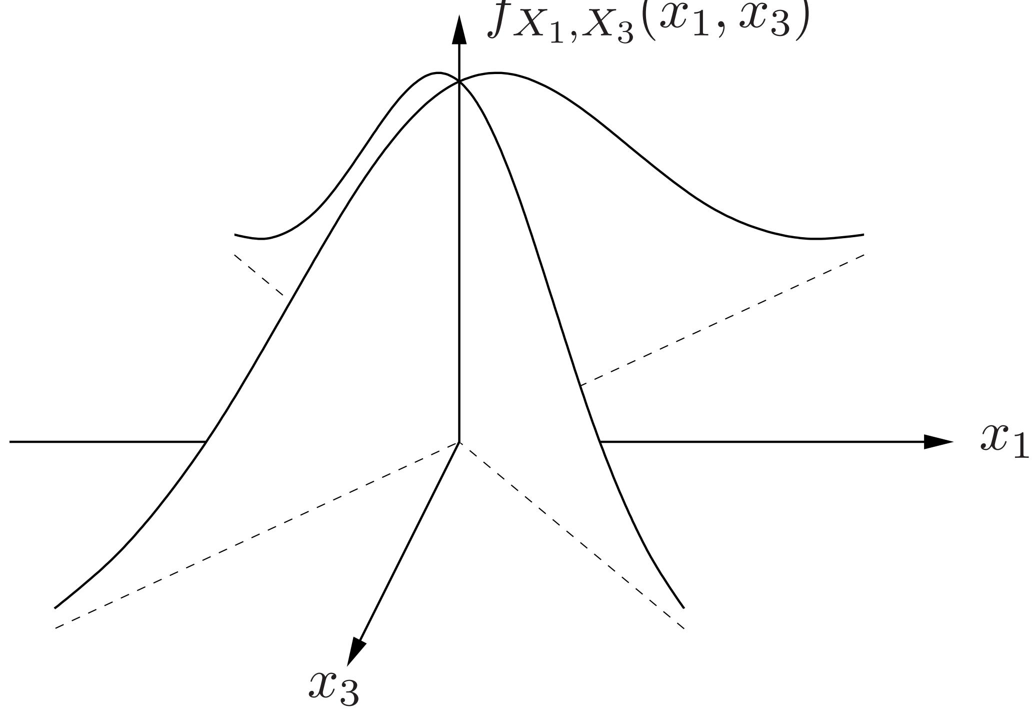
- The converse is not necessarily true, i.e., Gaussian marginals do not necessarily mean that the r.v.s are jointly Gaussian

  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 functions, "$f_{X_1,X_3}(x_1,x_3)$" has the form shown in the following figure



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

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

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