Chapter 1. Basic Concepts (continued)

♠ Sufficient Statistics

• Introduction:

The concept of a sufficient statistic is introduced in order to simplify statistical problems. A sufficient statistic is a function of the data which summarizes all the available information concerning θ.

• Sufficient Statistic

Let $\boldsymbol{X}$ be a random variable (or vector) whose distribution depends on the unknown parameter θ, but is otherwise unknown. A function T of X is said to be sufficient statistic for θ if the conditional distribution of $\boldsymbol{X}$, given $T(\boldsymbol{X}) = t$, is independent of θ (with probability one).

- **Partition of the Sample Space Induced by T**

If $T(\boldsymbol{X})$ is a statistic with range $|calT$ (i.e., $\mathcal{T} = \{T(\boldsymbol{x}) : \ \boldsymbol{x} \in \mathcal{X}\}$), the partition of $\mathcal{X}$ induced by T is defined by

$$\mathcal{X}_t = \{\boldsymbol{x} \in \mathcal{X} : \ T(\boldsymbol{x}) = t\}$$

for $t \in \mathcal{T}$.

If $t_1 \neq t_2$, then $\mathcal{X}_{t_1} \cap \mathcal{X}_{t_2} = \emptyset$, and $\cup_{t \in \mathcal{T}} \mathcal{X}_t = \mathcal{X}$.

A *sufficient partition* of $\mathcal{X}$ is a partition induced by a sufficient statistic T.

• **Factorization Theorem**:

Let $f(\boldsymbol{x}|\theta)$ denote the density function of $\boldsymbol{X}$. Then, T is a sufficient statistic iff (if and only if)

$$l(\theta) = f(\boldsymbol{x}|\theta) = h(\boldsymbol{x})g(T(x)|\theta),$$

where the function $h(\boldsymbol{x})$ does not depend on θ.

• **An Illustration**:

For a random sample of size n from $\mathcal{U}(0, \theta)$, the j.p.d.f. of $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is

$$f(\boldsymbol{x}|\theta) = (1/\theta)^n, \text{ if } 0 < x_i < \theta \text{ for } i = 1, 2, \ldots, n,$$

and zero elsewhere. Since θ exceeds all of observations iff it exceeds the largest x_i, the likelihood function can be written as

$$l(\theta) = f(\boldsymbol{x}|\theta) = \begin{cases} 1/\theta^n, & \theta > X_{(n)}, \\ 0, & \theta \leq X_{(n)}, \end{cases}$$

where $X_{(n)} = \max\{X_1, X_2, \ldots, X_n)'$. Thus, $T = X_{(n)}$ is sufficient by the factorization theorem.

- **Another Illustration**
- **Example 4.17**:

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)'$ denote a random sample from $N(\mu, \sigma^2)$. Also let $\boldsymbol{\theta} = (\mu, \sigma^2)'$. Then, the likelihood function is

$$
\begin{aligned}
l(\boldsymbol{\theta}) = f(\boldsymbol{x}|\boldsymbol{\theta}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \\
&=(2\pi)^{-\frac{n}{2}}\sigma^{-n} \exp\left\{-\frac{\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2}{2\sigma^2}\right\} \\
&=(2\pi)^{-\frac{n}{2}}\sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^{n} x_i - n\frac{\mu^2}{2\sigma^2}\right\}.
\end{aligned}
$$

Let

$$\boldsymbol{T} = (T_1, T_2)' = (\sum_{i=1}^{n} X_i^2, \sum_{i=1}^{n} X_i)',$$

$$h(\boldsymbol{x}) = (2\pi)^{-\frac{n}{2}},$$

and

$$g(\boldsymbol{t}|\boldsymbol{\theta}) = \exp\left\{\frac{1}{2\sigma^2}t_1 + \frac{\mu}{\sigma^2}t_2 - n\frac{\mu^2}{2\sigma^2}\right\}.$$

Factorization gives
$\boldsymbol{T} = (T_1, T_2)' = (\sum_{i=1}^{n} X_i^2, \sum_{i=1}^{n} X_i)'$ is sufficient.

- **Conditional Expectation**

The conditional distribution of $\boldsymbol{X}$, given $T(\boldsymbol{X}) = t$, is a distribution giving probability one to the set $\mathcal{X}_t$. Let $f_t(\boldsymbol{x})$ denote this conditional density on $\mathcal{X}_t$.

The conditional expectation of a function $h(\boldsymbol{x})$, given $T = t$, is given by

$$E^{\mathbf{X}|t}[h(\boldsymbol{X})] = \begin{cases} \int_{\mathcal{X}_t} h(\boldsymbol{x}) f_t(\boldsymbol{x}) & \text{continuous case,} \\ \sum_{\boldsymbol{x} \in \mathcal{X}_t} h(\boldsymbol{x}) f_t(\boldsymbol{x}) & \text{discrete case.} \end{cases}$$

- **A Useful Result**: The expectation of the conditional expectation is equal to the unconditional expectation. That is,

$$E^{\boldsymbol{X}}[h(\boldsymbol{X})] = E^{\boldsymbol{T}} E^{\boldsymbol{X}|\boldsymbol{T}}[h(\boldsymbol{X})].$$

- **Randomized Decision Rules Based on T**

For any statistic $T(\boldsymbol{X})$, randomized decision rules $\delta^*(t, \cdot)$ based on T are defined to be the usual randomized decision rules with $\mathcal{T}$ being considered the sample space. The risk function of such a rule is

$$R(\theta, \delta^*) = E^{\boldsymbol{T}}[L(\theta, \delta^*(T, \cdot)].$$

Theorem 1: *Assume that T is a sufficient statistic for θ, and let $\delta_0^*(\boldsymbol{x}, \cdot)$ be any randomized rule in $\mathcal{D}^*$. Then (subject to measurability conditions) there exists a randomized rule $\delta_1^*(t, \cdot)$, depending only on $T(\boldsymbol{x})$, which is R-equivalent to $\delta_0^*(\boldsymbol{x}, \cdot)$.*

Proof: Let

$$\delta_1^*(t, A) = E^{\boldsymbol{X}|t}[\delta_0^*(\boldsymbol{X}, A)].$$

Then $\delta_1^*(t, A)$ does not depend on θ, since T is a sufficient statistic. Now

$$R(\theta, \delta_1^*) = E^{\boldsymbol{T}}[L(\theta, \delta_1^*(T, \cdot)].$$

Note

$$L(\theta, \delta_1^*(t, \cdot)) = E^{\delta_1^*(t,\cdot)}[L(\theta, a)] = E^{\boldsymbol{X}|t} E^{\delta_0^*(\boldsymbol{X},\cdot)}[L(\theta, a)].$$

Therefore,

$$\begin{aligned} R(\theta, \delta_1^*) =& E^{\boldsymbol{T}}[L(\theta, \delta_1^*(T, \cdot))] = E^{\boldsymbol{T}} E^{\boldsymbol{X}|T} E^{\delta_0^*(\boldsymbol{X},\cdot)}[L(\theta, a)] \\ =& E^{\boldsymbol{X}} E^{\delta_0^*(\boldsymbol{X},\cdot)}[L(\theta, a)] = R(\theta, \delta_0^*), \end{aligned}$$

which completes the proof. $\square$

Comment: If δ_0^* is nonrandomized, the R-equivalent decision rule δ_1^* may be randomized, because

$$\delta_1^*(t, A) = E^{\boldsymbol{X}|t}[<\delta_0>(X, A)]$$
$$=E^{\boldsymbol{X}|t}[I_A(\delta_0(X))] = P^{\boldsymbol{X}|t}(\delta_0(X) \in A).$$

♠ Convexity

• Convex Set

A set $\Omega \subset R^m$ is *convex* if for any two points $\boldsymbol{x}$ and $\boldsymbol{y}$ in Ω, the point, $\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}$, is in Ω for $0 \leq \alpha \leq 1$.

Note that $\{\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y} : \ 0 \leq \alpha \leq 1\}$ is the line segment joining $\boldsymbol{x}$ and $\boldsymbol{y}$. Hence Ω is convex if the line segment between any two points in Ω is a subset of Ω.

- **Convex Hull**

If $\{\boldsymbol{x}^1, \boldsymbol{x}^2, \dots\}$ is a sequence of points in R^m, and $0 \le \alpha_i \le 1$ are numbers such that $\sum_{i=1}^{\infty} \alpha_i = 1$, then $\sum_{i=1}^{\infty} \alpha_i \boldsymbol{x}^i$ (providing it is finite) is called a *convex combination* of the $\{\boldsymbol{x}^i\}$. The *convex hull* of a set Ω is the set of all points which are convex combinations of points in Ω.

Note 1: It is more standard to define the convex hull only in terms of combinations of a finite number of $\{\boldsymbol{x}^i\}$, but the definitions turn out to be equivalent for R^m.

Note 2: The convex hull of a set is formed by connecting all points of the set by lines, and then filling in the interiors of the surfaces and solids so formed. It is easy to check that the convex hull of a set Ω is itself convex, and that it is the smallest convex set containing Ω.

Examples of Convex Sets

Ellipses and regular polygons in R^2; and solid pyramids, cubes, and balls in R^3 all are convex sets.

- **Convex and Concave Functions**

A real valued function $g(\boldsymbol{x})$ defined on a convex set Ω is *convex* if

$$g(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \leq \alpha g(\boldsymbol{x}) + (1-\alpha)g(\boldsymbol{y})$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \Omega$, and $0 < \alpha < 1$. If the inequality is strict for $\boldsymbol{x} \neq \boldsymbol{y}$, then g is *strictly convex.* If

$$g(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \geq \alpha g(\boldsymbol{x}) + (1-\alpha)g(\boldsymbol{y})$$

then g is *concave.* If the inequality is strict for $\boldsymbol{x} \neq \boldsymbol{y}$, then g is *strictly concave.*

Examples of Convex Functions

Functions x^2 e^x, and e^{-x} on R^1 are strictly convex, while $|x|$ on R^1 is convex, but not strictly convex (why?).

Examples of Concave Functions

Functions $-x^2$ and $-e^x$ on R^1, and $\log(x)$ on $(0, \infty)$ are strictly concave.

Lemma 1: *Let $g(\boldsymbol{x})$ be a function defined on an open convex subset Ω of R^m for which all second-order partial derivatives*

$$g^{ij}(\boldsymbol{x}) = \frac{\partial^2}{\partial x_i \partial x_j} g(\boldsymbol{x})$$

exist and are finite. Let

$$G = \begin{pmatrix} g^{11}(\boldsymbol{x}) & g^{12}(\boldsymbol{x}) & \cdots & g^{1m}(\boldsymbol{x}) \\ g^{21}(\boldsymbol{x}) & g^{22}(\boldsymbol{x}) & \cdots & g^{2m}(\boldsymbol{x}) \\ \vdots & \vdots & \ddots & \vdots \\ g^{m1}(\boldsymbol{x}) & g^{m2}(\boldsymbol{x}) & \cdots & g^{mm}(\boldsymbol{x}) \end{pmatrix}.$$

(Note that G is symmetric.) Then $g(\boldsymbol{x})$ is convex iff the matrix G is nonnegative definite for all $\boldsymbol{x} \in \Omega$ (i.e., $\boldsymbol{z}^t G \boldsymbol{z} \geq 0$ for all $\boldsymbol{z} \in R^m$ and $\boldsymbol{x} \in \Omega$). Likewise, $g(\boldsymbol{x})$ is concave if $-G$ is nonnegative definite. If G is positive (negative) definite, then $g(\boldsymbol{x})$ is strictly convex (concave).

Note: For $m = 1$, Lemma 1 says that g is convex (concave) if $g''(x) \geq 0$ ($g''(x) \leq 0$), where g'' is the second derivative. Moreover, if $g''(x) > 0$ ($g''(x) < 0$), then g is strictly convex (concave).

Using Lemma 1, we can easily show that (i) $g(x) = x^2$; (ii) $g(x) = e^x$; and (iii) $g(x_1, x_2) = x_1^2 + x_1 x_2 + x_2^2$ are strictly convex.

In fact, for (i) $g''(x) = 2 > 0$; for (ii) $g''(x) = e^x > 0$; and for (iii)

$$G = \begin{pmatrix} \frac{\partial^2}{\partial x_1 \partial x_1} g(x_1, x_2) & \frac{\partial^2}{\partial x_1 \partial x_2} g(x_1, x_2) \\ \frac{\partial^2}{\partial x_2 \partial x_1} g(x_1, x_2) & \frac{\partial^2}{\partial x_2 \partial x_2} g(x_1, x_2) \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

is positive definite.

Lemma 2: *Let $\boldsymbol{X}$ be an m-variate random vector such that $E[|\boldsymbol{X}|] < \infty$, where $|\boldsymbol{X}|$ denotes the m-dimensional Euclidean norm, and $P(\boldsymbol{X} \in \Omega) = 1$, where Ω is a convex subset of R^m. Then $E[\boldsymbol{X}] \in \Omega$.*

Theorem 2 (Jensen's Inequality) : *Let $g(\boldsymbol{x})$ be a convex real-valued function defined on a convex set Ω of R^m, and let $\boldsymbol{X}$ be an m-variate random vector for which $E[|\boldsymbol{X}|] < \infty$. Suppose also that $P(\boldsymbol{X} \in \Omega) = 1$. Then*

$$g(E[\boldsymbol{X}]) \leq E[g(\boldsymbol{X})],$$

with strict inequality if g is strictly convex and $\boldsymbol{X}$ is not concentrated at a point. (Note from Lemma 2 that $E[\boldsymbol{X}] \in \Omega$, so that $g(E[\boldsymbol{X}])$ is well defined.)

The formal proofs of Lemma 2 and Theorem 2 are postponed to Chapter 5.

However, the proof is extremely simple for the cases in which $m = 1$ and $g''(x)$ exists. Since g is convex, then $g''(x) \geq 0$. Let $\mu = E[X]$. Taylor Expansion gives

$$g(x) = g(\mu) + g'(u)(x - \mu) + \frac{g''(\xi)}{2}(x - \mu)^2,$$

where ξ is between μ and x. Since $g(x)$ is convex,

$$g(x) \geq g(\mu) + g'(\mu)(x - \mu),$$

and

$$\begin{aligned} E[g(X)] \geq & E[g(\mu)] + E[g'(\mu)(X - \mu)] \\ = & g(\mu) + g'(\mu)E[X - \mu] \\ = & g(E[X]). \end{aligned}$$

Preview: If the loss is convex, only nonrandomized rules need to be considered.

Theorem 3: *Assume that $\mathcal{A}$ is a convex subset of R^m, and that for each $\theta \in \Theta$ the loss function $L(\theta, \boldsymbol{a})$ is a convex function of $\boldsymbol{a}$. Let δ^* be a randomized decision rule in $\mathcal{D}^*$ for which $E^{\delta^*(\boldsymbol{x},\cdot)}[|\boldsymbol{a}|] < \infty$. Then the nonrandomized rule*

$$\delta(\boldsymbol{x}) = E^{\delta^*(\boldsymbol{x},\cdot)}[\boldsymbol{a}]$$

has

$$L(\theta, \delta(\boldsymbol{x})) \leq L(\theta, \delta^*(\boldsymbol{x}, \cdot))$$

for all $\boldsymbol{x}$ and θ.

Proof: From Lemma 2, it is clear that $\delta(\boldsymbol{x}) \in \mathcal{A}$. Jensen's inequality then gives that

$$\begin{aligned} L(\theta, \delta(\boldsymbol{x})) &= L(\theta, E^{\delta^*(\boldsymbol{x},\cdot)}[\boldsymbol{a}]) \\ &\leq E^{\delta^*(\boldsymbol{x},\cdot)}[L(\theta, \boldsymbol{a})] = L(\theta, \delta^*(\boldsymbol{x}, \cdot)). \end{aligned}$$

Theorem 4 (Rao-Blackwell): *Assume that $\mathcal{A}$ is a convex subset of R^m and that $L(\theta, \boldsymbol{a})$ is a convex function of $\boldsymbol{a}$ for all $\theta \in \Theta$. Suppose also that T is a sufficient statistic for θ, and that $\delta^0(\boldsymbol{x})$ is a nonrandomized decision rule in $\mathcal{D}$. Then the decision rule, based on $T(\boldsymbol{x}) = t$, defined by*

$$\delta^1(t) = E^{\boldsymbol{X}|t}[\delta^0(\boldsymbol{X})],$$

is R-equivalent to or R-better than $\delta^0(\boldsymbol{x})$, provided the expectation exists.

Proof: By the definition of a sufficient statistic, $\delta^1(t)$ does not depend on θ. By Jensen's inequality

$$L(\theta, \delta^1(t)) = L(\theta, E^{\boldsymbol{X}|t}[\delta^0(\boldsymbol{X})]) \leq E^{\boldsymbol{X}|t}[L(\theta, \delta^0(\boldsymbol{X}))].$$

Hence

$$R(\theta, \delta^1) = E^{\boldsymbol{T}}[L(\theta, \delta^1(t))]$$

$$\leq E^{\boldsymbol{T}} E^{\boldsymbol{X}|t}[L(\theta, \delta^0(\boldsymbol{X}))] = E^{\boldsymbol{X}}[L(\theta, \delta^0(\boldsymbol{X}))] = R(\theta, \delta^0).$$

- **Example: Exercise 19 (page 45)**

Assume a random variable $X \sim \mathcal{B}(n, \theta)$ is observed. It is desired to estimate θ under squared-error loss. Find a nonrandomized decision rule δ which is R-better than the randomized decision rule

$$\delta^* = \frac{1}{2}\langle\delta_1\rangle + \frac{1}{2}\langle\delta_2\rangle,$$

where $\delta_1(x) = x/n$ and $\delta_2 = \frac{1}{2}$. Also, calculate $R(\theta, \delta_1)$, $R(\theta, \delta_2)$, and $R(\theta, \delta^*)$.

Solution: Using Theorem 3, a nonrandomized decision rule δ which is R-better than δ^* is given by

$$\delta(x) = E^{\delta^*(x,\cdot)}[a] = \frac{1}{2}\frac{x}{n} + \frac{1}{2}\frac{1}{2} = \frac{x}{2n} + \frac{1}{4}$$

and the corresponding risk function is

$$\begin{aligned} R(\theta, \delta) =& E^{\boldsymbol{X}}\left[(\theta - \delta(X))^2\right] = E^{\boldsymbol{X}}\left[\left(\theta - \frac{X}{2n} - \frac{1}{4}\right)^2\right] \\ =& E^{\boldsymbol{X}}\left[\left(\frac{n\theta}{2n} - \frac{X}{2n} + \frac{\theta}{2} - \frac{1}{4}\right)^2\right] \\ =& \frac{\theta(1-\theta)}{4n} + \left(\frac{\theta}{2} - \frac{1}{4}\right)^2 = \frac{\theta(1-\theta)}{4n} + \frac{(2\theta-1)^2}{16}. \end{aligned}$$

Now

$$R(\theta, \delta_1) = E^{\boldsymbol{X}}[(\frac{x}{n} - \theta)^2] = \frac{\theta(1-\theta)}{n},$$

$$R(\theta, \delta_2) = E^{\boldsymbol{X}}[(\frac{1}{2} - \theta)^2] = \frac{(2\theta-1)^2}{4},$$

and

$$\begin{aligned} R(\theta, \delta^*) =& E^{\boldsymbol{X}} E^{\delta^*(X,\cdot)}[L(\theta, \delta^*)] = \frac{1}{2}R(\theta, \delta_1) + \frac{1}{2}R(\theta, \delta_2). \\ =& \frac{\theta(1-\theta)}{2n} + \frac{(2\theta-1)^2}{8}. \end{aligned}$$

Clearly, $R(\theta, \delta) < R(\theta, \delta^*)$ for all $0 \le \theta \le 1$.

- **Example: Exercise 25 (page 45)**

If $y_i \geq 0$, $i = 1, \ldots, m$, and $\alpha_i \geq 0$, $i = 1, 2, \ldots, m$, with $\sum_{i=1}^{m} \alpha_i = 1$, then prove that

$$\prod_{i=1}^{m} y_i^{\alpha_i} \leq \sum_{i=1}^{m} \alpha_i y_i.$$

Proof: Assume that if $y_i = 0$, then $\alpha_i > 0$. Under the above assumption, it is clear that if one of y_i's is equal to 0, the inequality is true. Now it is sufficient to show that the inequality holds when all $y_i > 0$. Note that when $y_i > 0$ for $i = 1, 2, \ldots, m$,

$$\log\left(\prod_{i=1}^{m} y_i^{\alpha_i}\right) = \sum_{i=1}^{m} \alpha_i \log(y_i).$$

Let Y be a random variable such that

$$P(Y = y_i) = \alpha_i \ \text{ for } i = 1, 2, \ldots, m.$$

Since $\log(y)$ is concave on $(0, \infty)$, then Jensen's Inequality gives

$$E[-\log(Y)] \geq -\log(E(Y)) \Longrightarrow E[\log(Y)] \leq \log(E(Y)).$$

Therefore,

$$E[\log(Y)] = \sum_{i=1}^{m} \alpha_i \log(y_i) \leq \log(E(Y)) = \log\left(\sum_{i=1}^{m} \alpha_i y_i\right),$$

which implies

$$\prod_{i=1}^{m} y_i^{\alpha_i} \leq \sum_{i=1}^{m} \alpha_i y_i.$$

- **An Example:**

Show that

$$f(x_1, x_2) = -x_1^p x_2^{1-p}, \quad 0 < p < 1,$$

is convex on $\{x_1 > 0,\ x_2 > 0\}$.

Proof: First, we compute

$$\frac{\partial}{\partial x_1} f(x_1, x_2) = -p x_1^{p-1} x_2^{1-p},$$

$$\frac{\partial^2}{\partial x_1^2} f(x_1, x_2) = -p(p-1) x_1^{p-2} x_2^{1-p},$$

$$\frac{\partial}{\partial x_2} f(x_1, x_2) = -(1-p) x_1^p x_2^{-p},$$

$$\frac{\partial^2}{\partial x_2^2} f(x_1, x_2) = -(1-p)(-p) x_1^p x_2^{-p-1},$$

and

$$\frac{\partial^2}{\partial x_1 \partial x_2} f(x_1, x_2) = -p(1-p) x_1^{p-1} x_2^{-p}.$$

Then, the G matrix in Lemma 1 is given by

$$G = \begin{pmatrix} \frac{\partial^2}{\partial x_1^2} f(x_1, x_2) & \frac{\partial^2}{\partial x_1 \partial x_2} f(x_1, x_2) \\ \frac{\partial^2}{\partial x_1 \partial x_2} f(x_1, x_2) & \frac{\partial^2}{\partial x_2^2} f(x_1, x_2) \end{pmatrix}$$

$$= \begin{pmatrix} -p(p-1)x_1^{p-2}x_2^{1-p} & -p(1-p)x_1^{p-1}x_2^{-p} \\ -p(1-p)x_1^{p-1}x_2^{-p} & -(1-p)(-p)x_1^p x_2^{-p-1} \end{pmatrix}.$$

Now, we need to show that G is nonnegative definite. To do so, we need to show that (i) two diagonal elements are nonnegative; and (ii) the determinant of G is nonnegative.

It is straightforward to see that

$$-p(p-1)x_1^{p-2}x_2^{1-p} > 0$$

and

$$-(1-p)(-p)x_1^p x_2^{-p-1} > 0,$$

since $0 < p < 1$, $x_1 > 0$, and $x_2 > 0$. Finally,

$$\begin{aligned} |G| =& [-p(p-1)x_1^{p-2}x_2^{1-p}][-(1-p)(-p)x_1^p x_2^{-p-1}] \\ & - [-p(1-p)x_1^{p-1}x_2^{-p}]^2 \\ =& p^2(1-p)^2[x_1^{2p-2}x_2^{-2p} - x_1^{2p-2}x_2^{-2p}] = 0. \end{aligned}$$

Thus, $f(x_1, x_2)$ is convex; but not strictly convex. $\square$