COS 302 Precept 3

Spring 2020

Princeton University

Outline

Motivation for Today's Precept

- Vectors that are at right angles (**orthogonal**) have many important mathematical properties that are heavily used in machine learning.
- Today we will see how we can use right angles to construct special bases and vector subspaces, and even create high-dimensional shadows!

Outline

Angles in High Dimensions

First, before we even define orthogonality, we need to understand what exactly an angle between vectors is:

Definition: Angle between two vectors

Assume $\mathbf{x} \neq 0$, and $\mathbf{y} \neq 0$ be two vectors that live in an inner product space. The unique angle $\omega \in [0, \pi]$ between two vectors is given by:

$$\omega = \cos^{-1}\left(\frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \, \|\mathbf{y}\|}\right)$$

Example 1

- Let's calculate the angle between:
  - $\mathbf{x} = [1, 0]^T$ and $\mathbf{y} = [\sqrt{2}/2, \sqrt{2}/2]^T$.
- $\langle \mathbf{x}, \mathbf{y} \rangle = 1 \cdot \sqrt{2}/2 + 0 \cdot \sqrt{2}/2 = \sqrt{2}/2$
- $\|\mathbf{x}\| = \sqrt{\mathbf{x}^\top \mathbf{x}} = \sqrt{1 + 0} = 1$
- $\|\mathbf{y}\| = \sqrt{\mathbf{y}^\top \mathbf{y}} = \sqrt{2/4 + 2/4} = 1$
- $\omega = \cos^{-1}\left(\frac{\sqrt{2}}{2}\right) = 45^\circ$

Orthogonal Vectors

Definition: Orthogonality and Orthonormality

Two vectors $\mathbf{x}$ and $\mathbf{y}$ are **orthogonal** if and only if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. Moreover, if both vectors **also** have unit length, i.e. $\|\mathbf{x}\| = 1 = \|\mathbf{y}\|$, we say that $\mathbf{x}$ and $\mathbf{y}$ are **orthonormal**.

You can check that this definition coincides with the usual "two vectors are orthogonal if they are separated by 90°", but is much more general and useful.

Matrices and Right Angles

- We can already get a sense of why orthonormal vectors are useful: adding/multiplying by 0/1 keeps things the same.

Definition: Orthogonal Matrix

A square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is an **orthogonal matrix** if and only if its columns are orthonormal so that

$$\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{I} = \boldsymbol{A}^\top \boldsymbol{A}.$$

This implies the special property that $\boldsymbol{A}^{-1} = \boldsymbol{A}^\top$.

Orthogonal Matrices Preserve Distances

Let $\boldsymbol{A}$ be an orthogonal matrix and let $\mathbf{v}_x = \boldsymbol{A}\mathbf{x}$. It turns out that $\mathbf{v}_x$ and $\mathbf{x}$ have the same length:

$$\begin{aligned}
\|\mathbf{v}_x\|^2 &= \|\boldsymbol{A}\mathbf{x}\|^2 \\
&= (\boldsymbol{A}\mathbf{x})^\top(\boldsymbol{A}\mathbf{x}) \\
&= \mathbf{x}^\top \boldsymbol{A}^\top \boldsymbol{A}\mathbf{x} \\
&= \mathbf{x}^\top \boldsymbol{I}\mathbf{x} \\
&= \mathbf{x}^\top \mathbf{x} \\
\|\mathbf{v}_x\|^2 &= \|\mathbf{x}\|^2
\end{aligned}$$

Orthogonal Matrices Preserve Angles

Let $\boldsymbol{A}$ be an orthogonal matrix, and let $\mathbf{v}_x = \boldsymbol{A}\mathbf{x}$ and $\mathbf{v}_y = \boldsymbol{A}\mathbf{y}$ be two vectors. Then the angle between $\mathbf{v}_x$ and $\mathbf{v}_y$ is the same as the angle between $\mathbf{x}$ and $\mathbf{y}$:

$$\begin{aligned}
\cos\omega = \frac{(\mathbf{v}_x)^\top(\mathbf{v}_y)}{\|\mathbf{v}_x\|\,\|\mathbf{x}_v\|} &= \frac{(\boldsymbol{A}\mathbf{x})^\top(\boldsymbol{A}\mathbf{y})}{\|\boldsymbol{A}\mathbf{x}\|\,\|\boldsymbol{A}\mathbf{y}\|} \\
&= \frac{\mathbf{x}\top\boldsymbol{A}^\top\boldsymbol{A}\mathbf{y}}{\sqrt{\mathbf{x}^\top\boldsymbol{A}^\top\boldsymbol{A}\mathbf{x}\mathbf{y}^\top\boldsymbol{A}^\top\boldsymbol{A}\mathbf{y}}} \\
&= \frac{\mathbf{x}^\top\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|}
\end{aligned}$$

Outline

Orthogonal and Orthonormal Bases

Definition: Orthonormal Basis

Consider an n-dimensional vector space V and a basis $\{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ of V. If

$$\langle \mathbf{b}_i, \mathbf{b}_j \rangle = 0 \text{ for } i \neq j \tag{1}$$

$$\langle \mathbf{b}_i, \mathbf{b}_i \rangle = 1 \tag{2}$$

for all $i, j = 1, \ldots, n$ then the basis is called an **orthonormal basis**. If only (1) is satisfied, the basis is instead called an **orthogonal basis**. Moreover, (2) implies that every basis vector has length 1.

Example 2

Examples of orthonormal bases:

- The standard basis for $\mathbb{R}^n$ is orthonormal.
- In $\mathbb{R}^2$, the vectors

$$\mathbf{b}_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{b}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

form an orthonormal basis as $\mathbf{b}_1^\top \mathbf{b}_2 = 0$ and $\|\mathbf{b}_1\| = 1 = \|\mathbf{b}_2\|$.

Outline

Orthogonal Vector Spaces

It turns out that *even* vector spaces can be orthogonal to each other:

Definition: Orthogonal Complement

Let V be a D-dimensional vector space, and $U \subseteq V$ an M-dimensional subspace. Then U's **orthogonal complement**, denoted as $U^{\perp}$, is a $(D-M)$-dimensional subspace of V that contains all vectors in V that are orthogonal to every vector in U.

Example 3: 2D Plane and Normal Vector

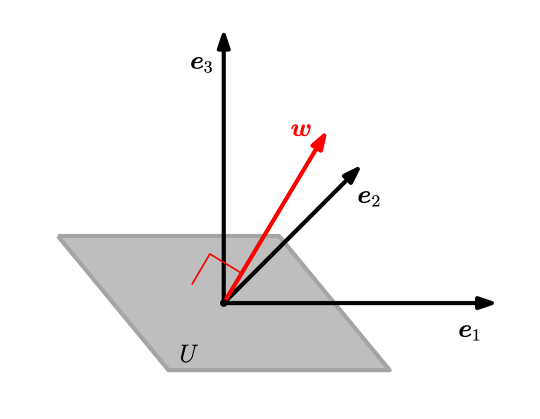


Figure 1: A plane U in a three-dimensional vector space can be described by its normal vector, $\mathbf{w}$, which spans its orthogonal complement $U^{\top}$.

Vector Decomposition

Since $U \cap U^{\perp} = \{\mathbf{0}\}$, we can write any vector $\mathbf{x} \in V$ into two separate sums involving vectors from U and $U^{\perp}$ respectively:

$$\mathbf{x} = \sum_{m=1}^{M} \lambda_m \mathbf{b}_m + \sum_{j=1}^{D-M} \psi_j \mathbf{b}_j^{\perp}, \quad \lambda_m, \psi_j \in \mathbb{R}$$

where $(\mathbf{b}_1, \ldots, \mathbf{b}_M)$ is a basis of U and $(\mathbf{b}_1^{\perp}, \ldots, \mathbf{b}_{D-M}^{\perp})$ is a basis of $U^{\perp}$.

Why should we care?

- Planes, and their higher-dimensional analogs known as hyperplanes can be described using the orthogonal complement as shown in the previous example.
- One reason hyperplanes are so important is that many machine learning algorithms such as support vector machines, heavily depend on the notion of hyperplanes.

Outline

Projection

Definition

Let V be a vector space and $U \subseteq V$ a subspace of V . A linear mapping $\pi : V \to U$ is called a **projection** if $\pi^2 = \pi \circ \pi = \pi$.

- Linear mappings can be defined using transformation matrices. (Recall that every linear map corresponds to a matrix).
- Projection matrix P_π has the property that $P_\pi = P_\pi \cdot P_\pi$

Orthogonal Projection

Definition

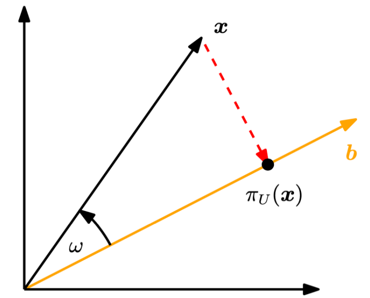
Let V be a vector space and $U \subseteq V$ a subspace of V. A linear mapping $\pi : V \to U$ is called an **orthogonal projection** if $\forall v \in V$, $u = \pi(v)$ is the closest to v for all vectors in U.

- Orthogonal projection is a type of projection
- Easy to check that $\pi^2 = \pi$, as $\pi(u) = u$.

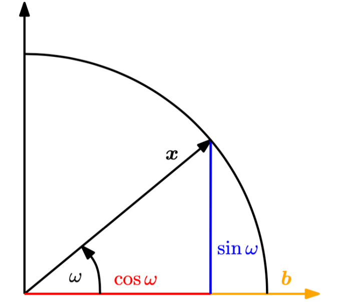
Outline

Projection onto 1D Subspaces (Lines)

Project $\mathbf{x} \in \mathbb{R}^n$ onto one-dimensional subspace $U \subseteq \mathbb{R}^n$ spanned by $\mathbf{b} \in \mathbb{R}^n$, where $\pi_U(\mathbf{x}) \in \mathbb{R}^n$ is the closest to $\mathbf{x}$ on U. Below is an illustration on $\mathbb{R}^2$:



(a) Projection of $\boldsymbol{x} \in \mathbb{R}^2$ onto a subspace U with basis vector $\boldsymbol{b}$.



(b) Projection of a two-dimensional vector $\boldsymbol{x}$ with $\|\boldsymbol{x}\| = 1$ onto a one-dimensional subspace spanned by $\boldsymbol{b}$.

Projection onto 1D Subspaces (Lines)

What we know about $\pi_U(\mathbf{x})$:

- The projection $\pi_U(\mathbf{x})$ is closest to $\mathbf{x}$ in U
  $\implies$ vector $\pi_U(\mathbf{x}) - \mathbf{x}$ is orthogonal to U
- The projection $\pi_U(\mathbf{x})$ belongs to $U = \text{span}(\mathbf{b})$
  $\implies$ $\pi_U(\mathbf{x}) = \lambda\mathbf{b}$ for some $\lambda \in \mathbb{R}$

Projection onto 1D Subspaces

3 Steps for computing the projection:

1. Find the coordinate λ
2. Compute the projection $\pi_U(\mathbf{x})$
3. Compute the projection matrix $\mathbf{P}_\pi$

Projection onto 1D Subspaces (Step 1/3)

How do we find $\pi_U(\mathbf{x})$?

1. **Find the coordinate λ:**

$$\langle \boldsymbol{x} - \pi_U(\boldsymbol{x}), \boldsymbol{b} \rangle = 0 \overset{\pi_U(\boldsymbol{x}) = \lambda \boldsymbol{b}}{\iff} \langle \boldsymbol{x} - \lambda \boldsymbol{b}, \boldsymbol{b} \rangle = 0$$

$$\langle \boldsymbol{x}, \boldsymbol{b} \rangle - \lambda \langle \boldsymbol{b}, \boldsymbol{b} \rangle = 0 \iff \lambda = \frac{\langle \boldsymbol{x}, \boldsymbol{b} \rangle}{\langle \boldsymbol{b}, \boldsymbol{b} \rangle} = \frac{\langle \boldsymbol{b}, \boldsymbol{x} \rangle}{\|\boldsymbol{b}\|^2}$$

Projection onto 1D Subspaces (Step 2/3)

2. **Compute the projection $\pi_U(\mathbf{x}) \in U$:**

$$\pi_U(\mathbf{x}) = \lambda \mathbf{b} = \frac{\langle \boldsymbol{b}, \boldsymbol{x} \rangle}{\|\boldsymbol{b}\|^2} \boldsymbol{b} = \frac{\boldsymbol{b}^\top \boldsymbol{x}}{\|\boldsymbol{b}\|^2} \boldsymbol{b}$$

   Note: Let ω be the angle between $\boldsymbol{b}$ and $\boldsymbol{x}$, we have

$$\begin{aligned} \|\pi_U(\boldsymbol{x})\| = \frac{\boldsymbol{b}^\top \boldsymbol{x}}{\|\boldsymbol{b}\|^2} \|\boldsymbol{b}\| &= |\cos \omega| \|\boldsymbol{x}\| \|\boldsymbol{b}\| \frac{\|\boldsymbol{b}\|}{\|\boldsymbol{b}\|^2} \\ &= |\cos \omega| \|\boldsymbol{x}\|, \end{aligned}$$

   length of $\pi_U(\boldsymbol{x})$ is scaled by $|\cos \omega|$.

Projection onto 1D Subspaces (Step 3/3)

3. **Compute the projection matrix $\mathbf{P}_\pi$:**

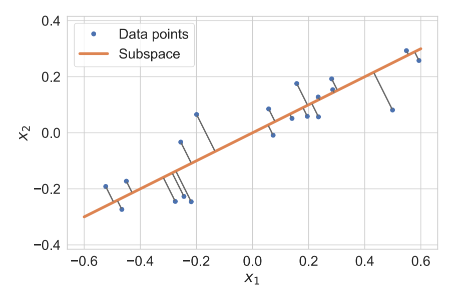
- We know π_U is a linear mapping, there exists a projection matrix $\mathbf{P}_\pi$ such that $\pi_U(\mathbf{x}) = \mathbf{P}_\pi \mathbf{x}$
- We have $\mathbf{P}_\pi = \frac{\boldsymbol{b}\boldsymbol{b}^\top}{\|\boldsymbol{b}\|^2}$, since

$$\pi_U(\boldsymbol{x}) = \lambda \boldsymbol{b} = \boldsymbol{b}\lambda = \boldsymbol{b}\frac{\boldsymbol{b}^\top \boldsymbol{x}}{\|\boldsymbol{b}\|^2} = \frac{\boldsymbol{b}\boldsymbol{b}^\top}{\|\boldsymbol{b}\|^2}\boldsymbol{x}$$

Note that $\mathbf{P}_\pi$ is of rank 1. It projects any vector $\mathbf{x} \in \mathbb{R}^n$ onto the line through origin with the direction $\mathbf{b}$.

Projection onto 1D Subspaces: Example

Example in $\mathbb{R}^2$: Go to www.tinyurl.com/cos302-precept3



Subspace U spanned by $\mathbf{b} = [1, 0.5]^\top \in \mathbb{R}^2$, different data points (blue dots) are projected onto subspace U.

Outline

Projection onto General Subspaces

Project $\mathbf{x} \in \mathbb{R}^n$ onto subspace $U \subseteq \mathbb{R}^n$ with $\dim(U) = m \leq n$, usually U is low-dimensional and m can be much smaller than n.

- Assume U has a basis $(\mathbf{b}_1, \cdots, \mathbf{b}_m)$, then $\pi_U(\mathbf{x})$ can be represented by a linear combination of the basis such that $\pi_U(\mathbf{x}) = \sum_{i=1}^{m} \lambda_i \mathbf{b}_i$
- Similar to 1D subspace case: we could first find the coordinates λ_i's and then find $\pi_U(\mathbf{x})$ and its corresponding $\mathbf{P}_\pi$.

Projection onto General Subspaces

3 Steps for computing the projection:

1. Find the coordinates $\lambda_1 \cdots \lambda_m$
2. Compute the projection $\pi_U(\mathbf{x})$
3. Compute the projection matrix $\mathbf{P}_\pi$

Projection onto General Subspaces (Step 1/3)

1. **Find the coordinates $\lambda_1 \cdots \lambda_m$:**

   First, we could write $\pi_U(\boldsymbol{x}) = \sum_{i=1}^{m} \lambda_i \boldsymbol{b}_i = \boldsymbol{B}\boldsymbol{\lambda}$,
   where $\boldsymbol{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m] \in \mathbb{R}^{n \times m}$, and $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_m]^\top \in \mathbb{R}^m$.
   Since $\pi_U(\boldsymbol{x})$ is closest to $\mathbf{x}$, we know that $\boldsymbol{x} - \pi_U(\boldsymbol{x})$ must be orthogonal to all vectors of U, which is equivalent to being orthogonal to all the basis vectors of U, i.e. $\mathbf{b}_1, \cdots, \mathbf{b}_m$. Therefore, we have

$$
\begin{gathered}
\langle \boldsymbol{b}_1, \boldsymbol{x} - \pi_U(\boldsymbol{x}) \rangle = \boldsymbol{b}_1^\top (\boldsymbol{x} - \pi_U(\boldsymbol{x})) = 0 \\
\vdots \\
\langle \boldsymbol{b}_m, \boldsymbol{x} - \pi_U(\boldsymbol{x}) \rangle = \boldsymbol{b}_m^\top (\boldsymbol{x} - \pi_U(\boldsymbol{x})) = 0
\end{gathered}
$$

   Rewrite in matrix form,

$$
\begin{aligned}
\begin{bmatrix} \boldsymbol{b}_1^\top \\ \vdots \\ \boldsymbol{b}_m^\top \end{bmatrix} [\boldsymbol{x} - \boldsymbol{B}\boldsymbol{\lambda}] = \mathbf{0} &\Longleftrightarrow \boldsymbol{B}^\top (\boldsymbol{x} - \boldsymbol{B}\boldsymbol{\lambda}) = \mathbf{0} \\
&\Longleftrightarrow \boldsymbol{B}^\top \boldsymbol{B} \boldsymbol{\lambda} = \boldsymbol{B}^\top \boldsymbol{x} \qquad (3)
\end{aligned}
$$

   Equation (3) is called the normal equation.

Projection onto General Subspaces (Step 1/3)

1. Cont'd:

Solve the normal equation: since $(\mathbf{b}_1, \cdots, \mathbf{b}_m)$ are linearly independent, $\boldsymbol{B}^\top \boldsymbol{B} \in \mathbb{R}^{m \times m}$ is regular and can be inverted.[1]
Coordinates are solved by:

$$\boldsymbol{\lambda} = \left(\boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} \boldsymbol{B}^\top \boldsymbol{x}$$

- Matrix $\left(\boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} \boldsymbol{B}^\top$ is called the pseudo-inverse of $\boldsymbol{B}$. In the case when $\boldsymbol{B}$ is full rank, $\boldsymbol{B}^\top \boldsymbol{B}$ is positive definite and invertible.
- In general, such as when solving the normal equation in ordinary least squares (discussed later), $\boldsymbol{B}^\top \boldsymbol{B}$ is only guaranteed to be positive semi-definite, a "jitter term" $\epsilon \mathbf{I}$ is added to it so that it becomes positive definite and invertible.

[1] https://math.stackexchange.com/questions/1840801/why-is-ata-invertible-if-a-has-independent-columns

Projection onto General Subspaces (Step 2&3/3)

2. **Compute the projection $\pi_U(\mathbf{x})$:**

$$\pi_U(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{\lambda} = \boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} \boldsymbol{B}^\top \boldsymbol{x}$$

3. **Compute the projection matrix $\mathbf{P}_\pi$:**

$$\mathbf{P}_\pi = \boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} \boldsymbol{B}^\top$$

Projection onto General Subspaces: Remarks

- The projections $\pi_U(\boldsymbol{x})$ are still vectors in $\mathbb{R}^n$, although they lie in an m-dimensional subspace $U \subseteq \mathbb{R}^n$.
- However, to represent a projected vector we only need the m coordinates $\lambda_1, \cdots, \lambda_m$ with respect to the basis vectors $\mathbf{b}_1, \cdots, \mathbf{b}_m$ of U.

Projection onto General Subspaces: Connection with Ordinary Least Squares

Suppose we have n observations $\{\mathbf{x}_i, y_i\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. We would like to find a weight vector $\mathbf{w} \in \mathbb{R}^d$ such that $y_i \approx w^\top \mathbf{x}_i$ for all i.

In other words, we want

$$\mathbf{y} \approx \mathbf{X}^T \mathbf{w},$$

$\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ and $\mathbf{y} = [y_1, \cdots, y_n]^\top \in \mathbb{R}^n$.

Projection onto General Subspaces: Connection with Ordinary Least Squares

Normally, $d < n$, the linear system $\mathbf{X}^T\mathbf{w} = \mathbf{y}$ is over-determined and usually does not have a solution. We could get an approximate solution by minimizing the squared errors:

$$\arg\min_{\mathbf{w}} S(\mathbf{w})$$

where

$$S(\mathbf{w}) = \sum_{i=1}^{n} |y_i - \mathbf{x}_i^\top \mathbf{w}|^2 = \|\mathbf{y} - \mathbf{X}^T\mathbf{w}\|_2^2$$

Projection onto General Subspaces: Connection with Ordinary Least Squares

A projection perspective:

- minimizing the squared error is equivalent to finding the vector within the subspace ($\mathbf{X}^T\mathbf{w}$) that is closest to $\mathbf{y}$, where $\mathbf{w}$ is a vector of the coordinates.
- we could find $\mathbf{w}$ by computing the orthogonal projection of $\mathbf{y}$ onto the subspace spanned by the columns of $\mathbf{X}^\top$.

Projection onto General Subspaces with Orthonormal Basis

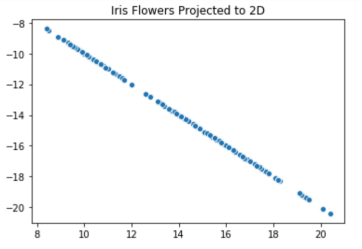
If the basis is an ONB, we have $\boldsymbol{B}^\top \boldsymbol{B} = \mathbf{I}$.

$$\pi_U(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{\lambda} = \boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{B}\right)^{-1} \boldsymbol{B}^\top \boldsymbol{x} = \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}$$

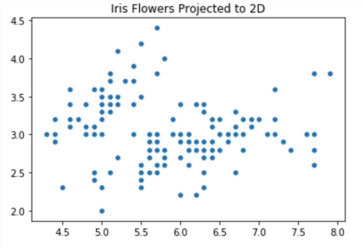
- Coordinate $\lambda_i = \mathbf{b}_i^\top \mathbf{x}$: project $\mathbf{x}$ onto $\mathbf{b}_i$ and get the coordinate by taking the inner product.
- $\pi_U(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{\lambda} = \sum_{i=1}^m \lambda_i \mathbf{b}_i$: linearly combine the basis using the coordinates.
- No inverse needed, computationally efficient: a reason why we like orthonormal basis.

Projection onto General Subspaces: Iris Dataset Example

Open the colab notebook at www.tinyurl.com/cos302-precept3



(a) Non-Orthogonal Projection



(b) Orthogonal Projection

Figure 2: Projecting the four dimensional iris dataset onto two dimensions