


# Intensity-based image registration

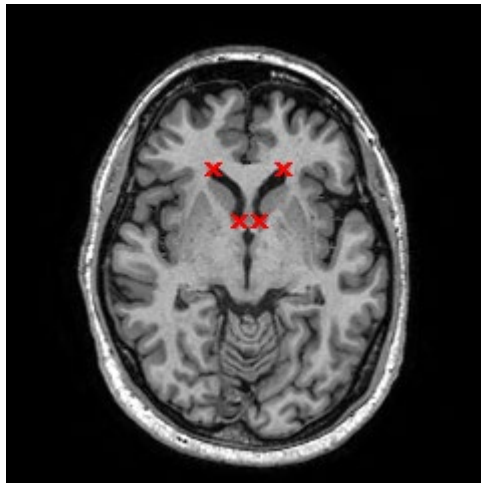
Ruisheng Su,  
Maureen van Eijnatten

## Today:

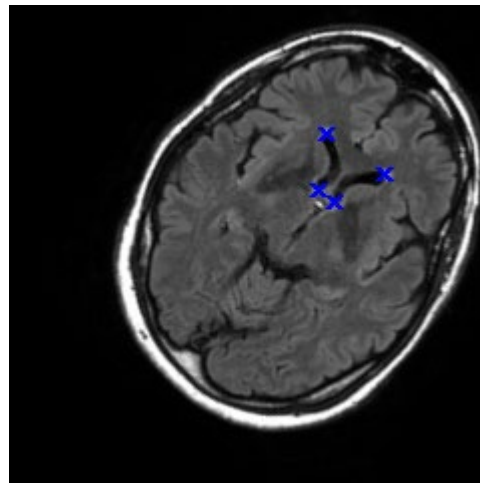
- Intensity-based similarity metrics
- Probability theory
- Optimization
- Intensity-based image registration 

## Recap (previous lecture)

- Point-based registration requires some manual input
- Can it be used with automatic keypoint selection?



Fixed



Moving



## Classification of image registration:

- Image dimensionality: 2D, 3D, 3D + time...
- Registration basis: point sets, intensity ..
- Geometrical transformations: rigid, affine, nonlinear...
- Degree of interaction: automatic, manual, semi-automatic
- Optimization procedure: closed-form solution, iterative
- Modalities: multi-modal, intra-modal
- Subject: inter-patient, intra-patient, atlas
- Object: brain, head, vertebra, liver...



## Learning outcomes

The student can:

- explain and implement three important intensity-based image similarity metrics, namely sum of square differences, cross correlation, and mutual information.
- select the correct image similarity metric for an image registration task based on the assumptions of these metrics.
- explain how the joint probability mass function (p.m.f.) can be used to measure the similarity between two images if we consider the image intensities as random variables.
- interpret joint histograms to judge whether two images are well aligned
- describe the numerical procedure to register two images by maximizing a similarity function (gradient ascent/descent)
- explain the effect of the learning rate and the initialization of the parameters on the optimization process when using gradient ascent/descent



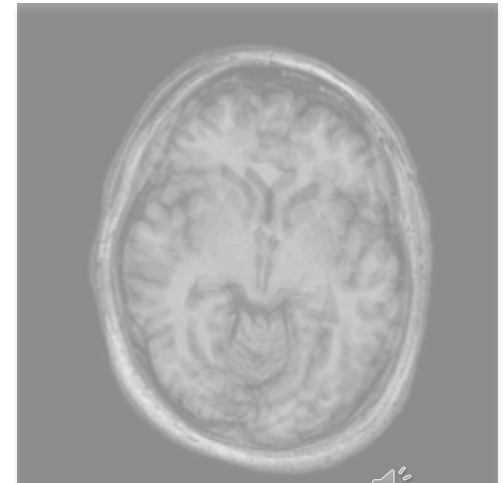
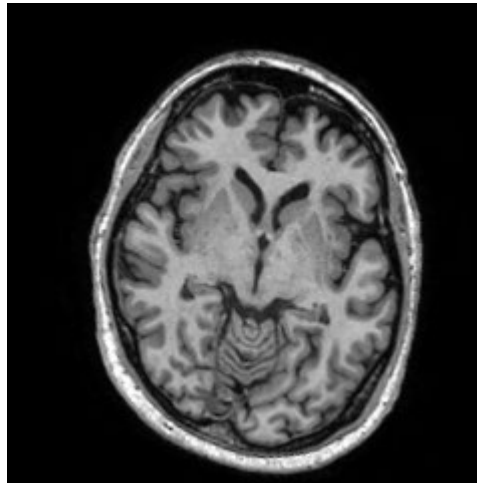
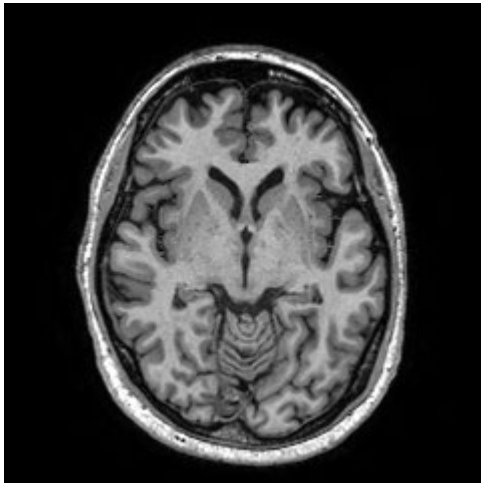
# Intensity-based similarity metrics



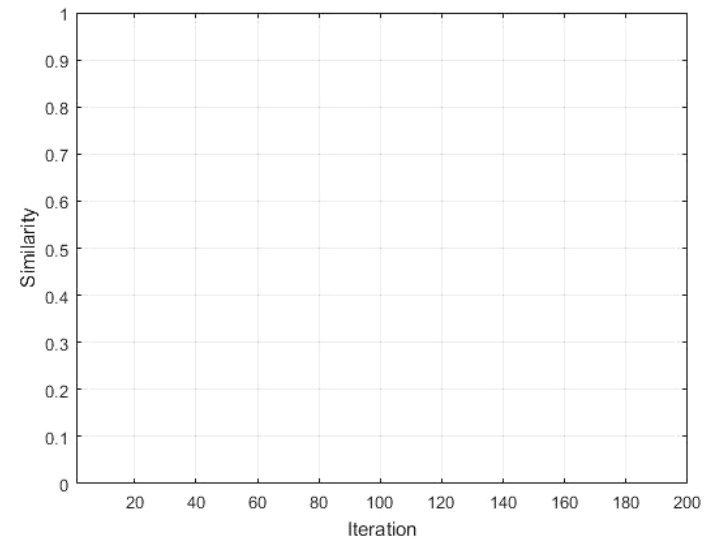
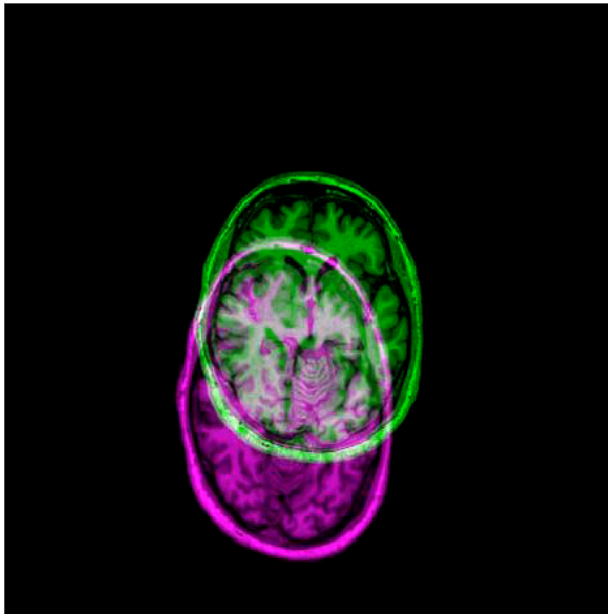
Image intensity is an alternative registration basis to points.

It is the most widely used registration basis.

Compared to point-based registration, requires less user interaction.



Intensity-based image registration works by iterative optimization of an intensity-based similarity measure.



## Outline:

### Intensity-based similarity measures:

- Sum of square differences
- Cross-correlation
- Mutual information

### Optimization for intensity-based registration:

- Gradient ascent (descent)





Let  $I$  and  $J$  be two images and  $i$  the pixel locations.

A simple and intuitive intensity-based measure of the similarity of  $I$  and  $J$  is the sum of squared differences (SSD):

$$\text{SSD}(I, J) = \sum_{i=1}^n (I(i) - J(i))^2$$

<b>10</b>	<b>12</b>	<b>16</b>
<b>8</b>	<b>12</b>	<b>18</b>
<b>4</b>	<b>8</b>	<b>10</b>

<b>8</b>	<b>14</b>	<b>16</b>
<b>12</b>	<b>12</b>	<b>20</b>
<b>2</b>	<b>6</b>	<b>12</b>

$(10-8)^2$	$(12-14)^2$	$(16-16)^2$
$(8-12)^2$	$(12-12)^2$	$(18-20)^2$
$(4-2)^2$	$(8-6)^2$	$(10-12)^2$

If  $I$  is the fixed image in a registration problem, and  $J$  is the moving image transformed with a transformation  $\mathbf{T}$  the similarity measure will be a function of the transformation :

$$\text{SSD}(I, J, \mathbf{T}) = \sum_{i=1}^n (I(i) - J_{\mathbf{T}}(i))^2$$

The SSD will be lowest when the images are perfectly aligned and will increase with misalignment.

When lowest == zero?



It can be shown that this measure is optimal when two images differ only by Gaussian noise. This is an implicit assumption of this measure.

- Not true for inter-modality registration
- Rarely true for intra-modality registration (e.g. MRI noise is not Gaussian, there will be changes between acquisitions etc.)

Nevertheless, SSD can be still used with success in intra-modality registration.

Drawback: It can be very sensitive to a few “outlier” intensity differences.



Another measure that makes slightly less assumptions is normalized cross correlation  $C$ :

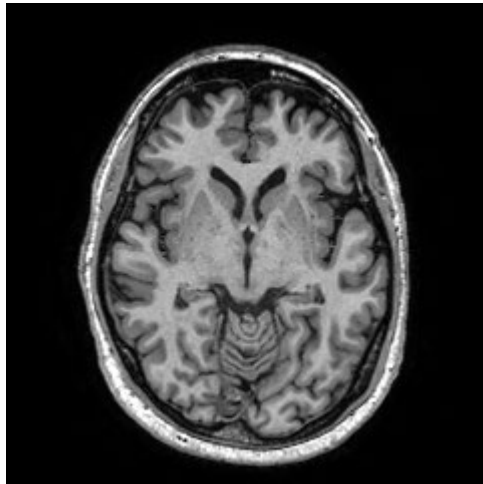
$$C(I, J) = \frac{\sum_{i=1}^n (I(i) - \bar{I})(J(i) - \bar{J})}{\sqrt{\sum_{i=1}^n (I(i) - \bar{I})^2 \sum_{j=1}^n (J(j) - \bar{J})^2}}$$

The main assumption of normalized cross-correlation is that there is a linear relationship between the pixel intensities in the two images.

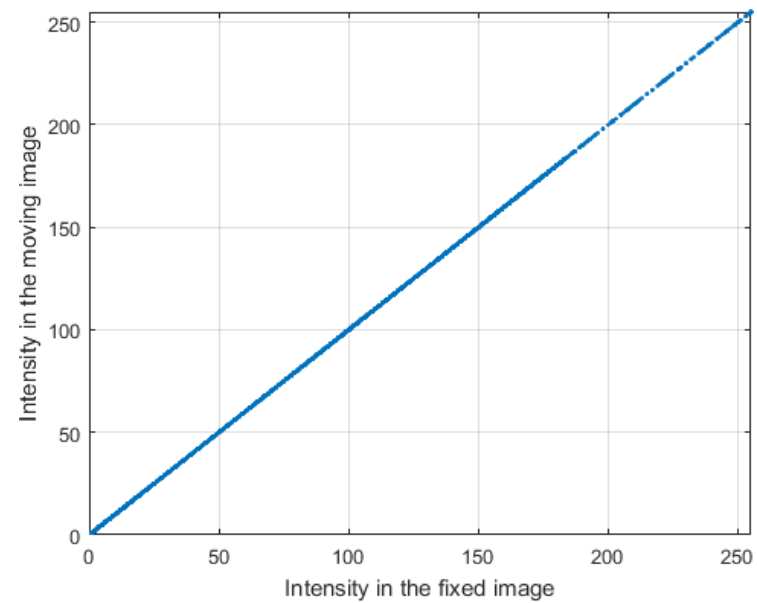
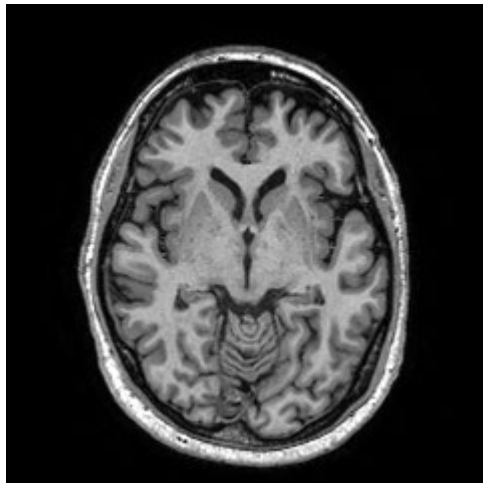
- Mostly true for intra-modality registration
- Usually not true for inter-modality registration



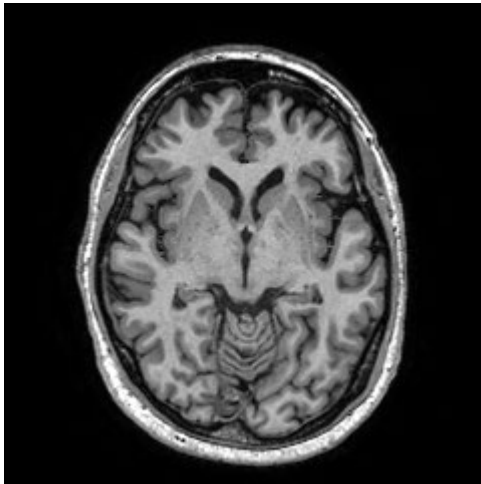
T1



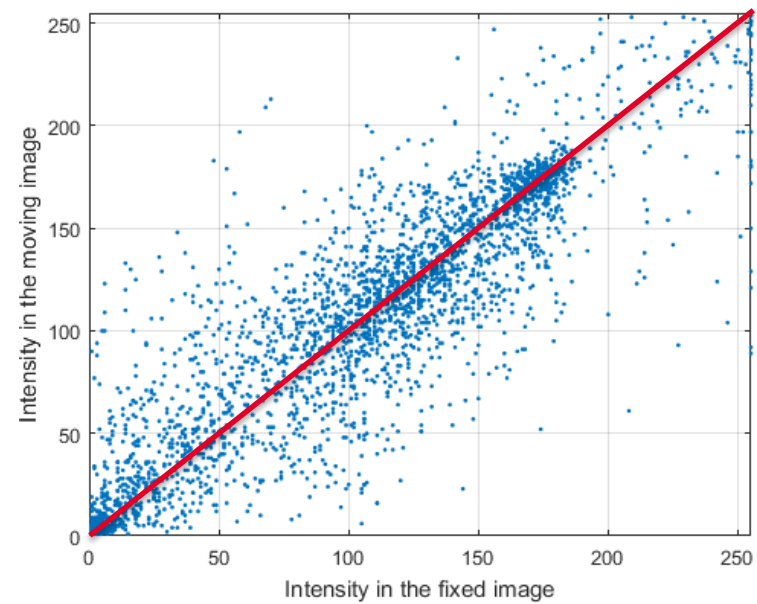
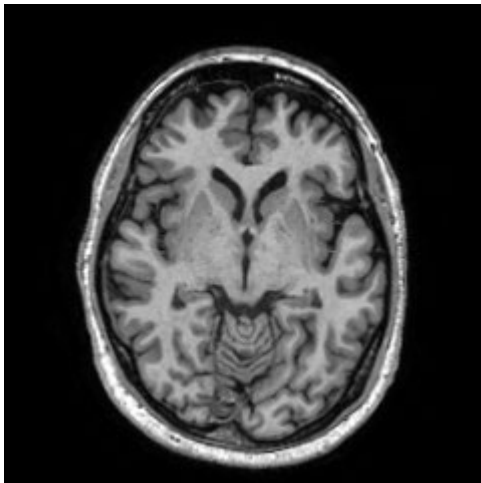
T1



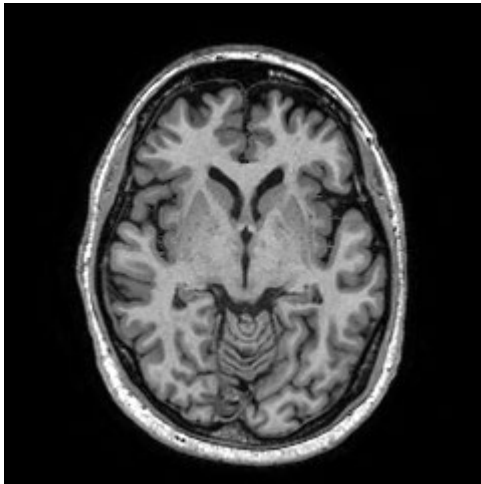
T1



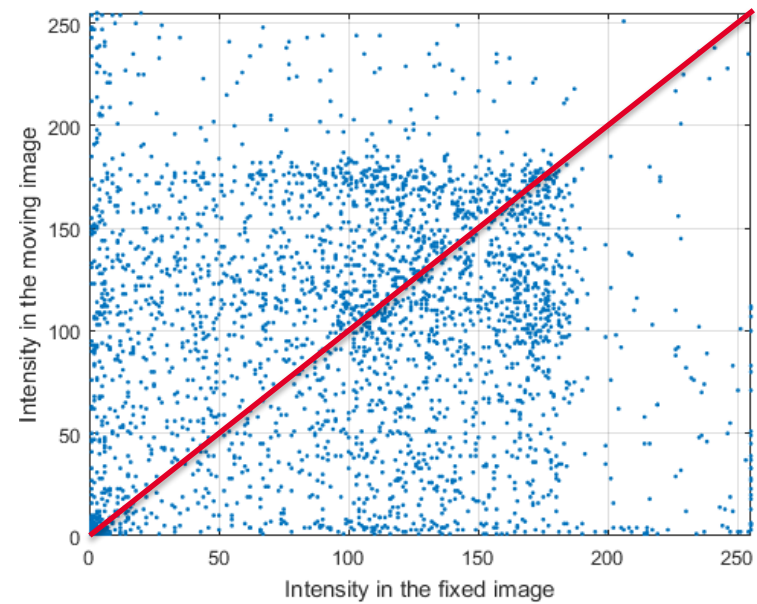
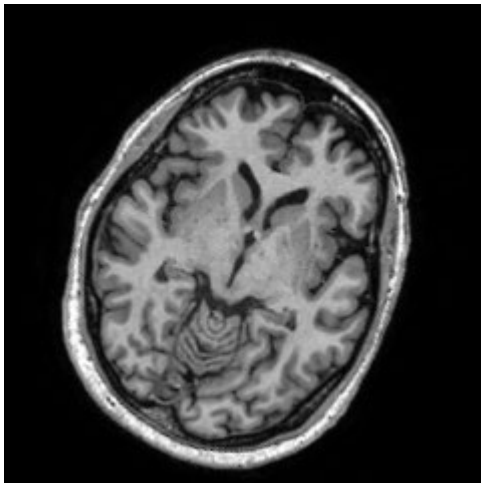
T1



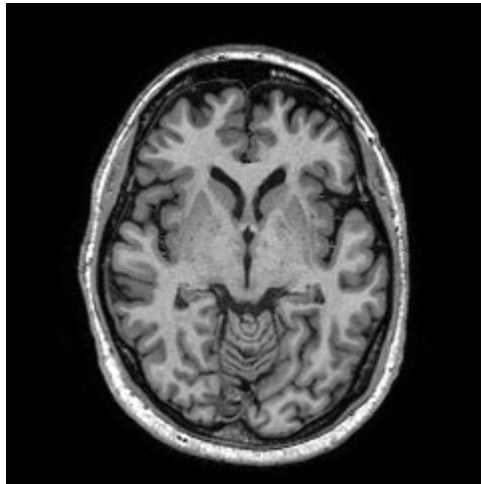
T1



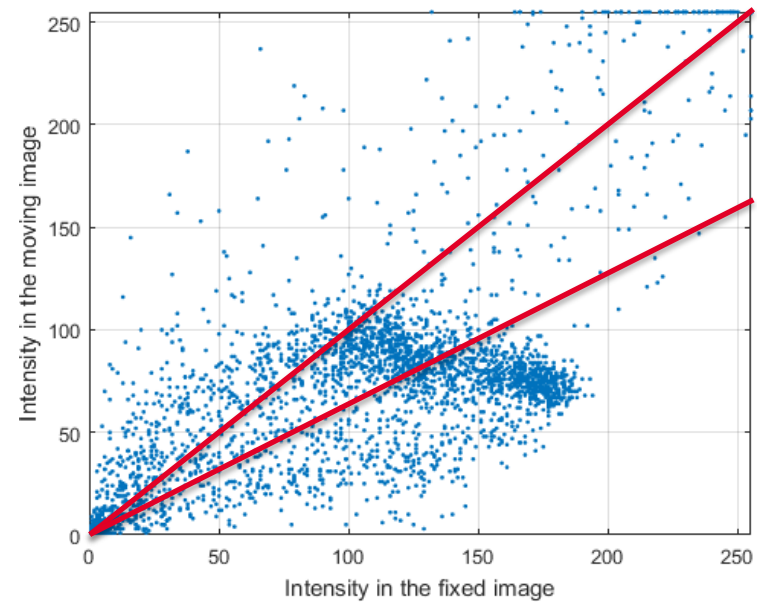
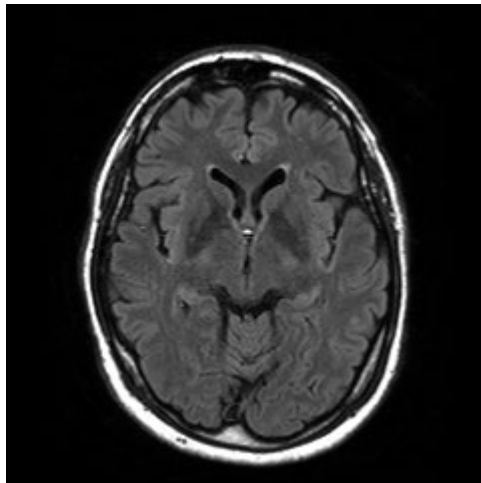
T1



T1



T2

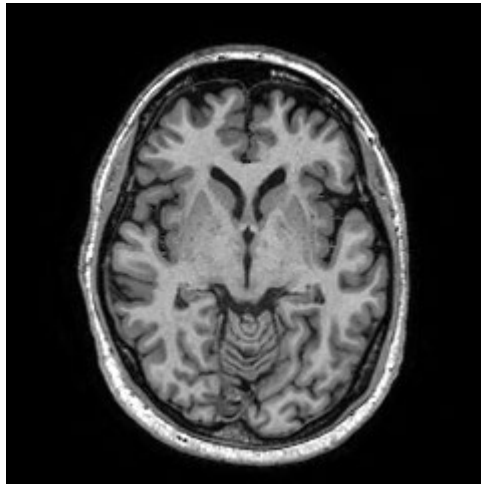


**Violates the assumption of linear  
relationship between the intensities –  
not optimal**

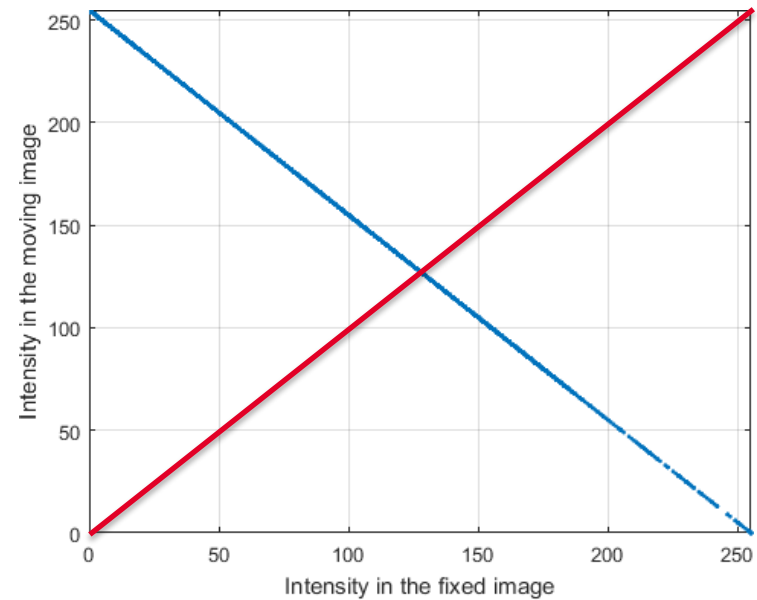
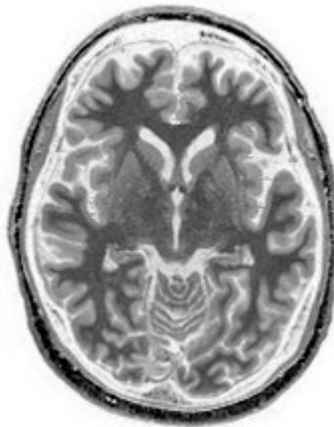




T1



Simulated  
modality



**Completely opposite of the  
assumption (inverse linear  
relationship)**



# Probability theory



Random variables map the outcomes of random phenomena to numbers.

Example random phenomenon: coin toss.

Random variable  $X$ : the outcome of the coin toss.

Another random variable  $Y$ : the number of heads in a series of 3 tosses.

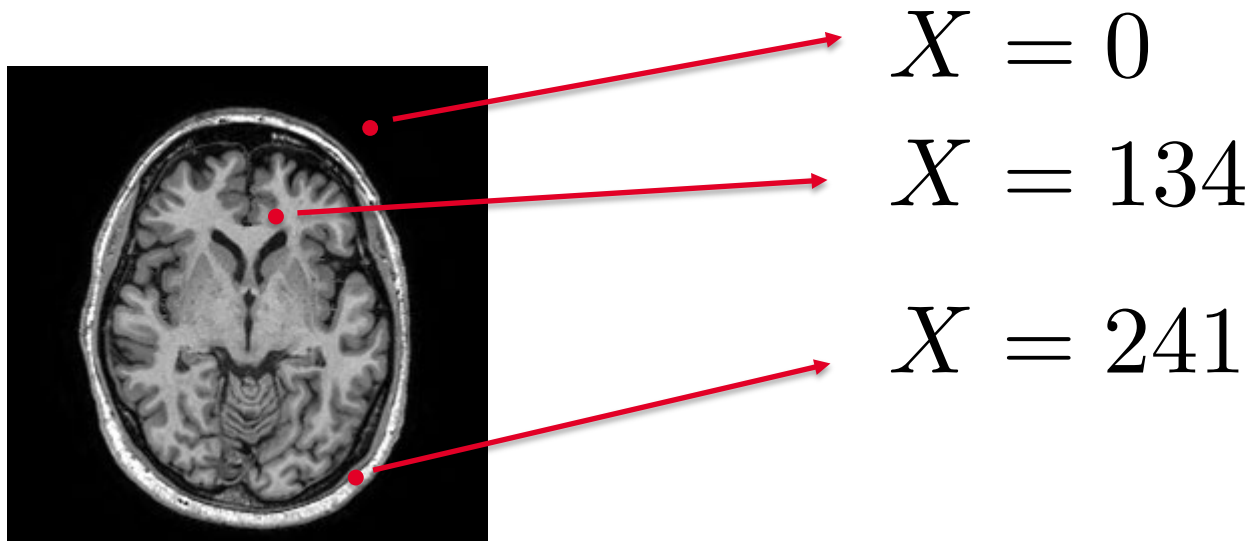
$$X = \begin{cases} 1, & \text{if heads} \\ 0, & \text{if tails} \end{cases}$$



## Image intensities as random variables:

Random phenomenon: pick a random pixel location.

In this case, the pixel intensity can be treated as a random variable.



Each outcome from the random phenomenon we are studying can be associated with a probability.

If a random variable  $X$  can have a finite set of possible values, we can define a function that maps each possible value to a probability. This function is called probability mass function (p.m.f).

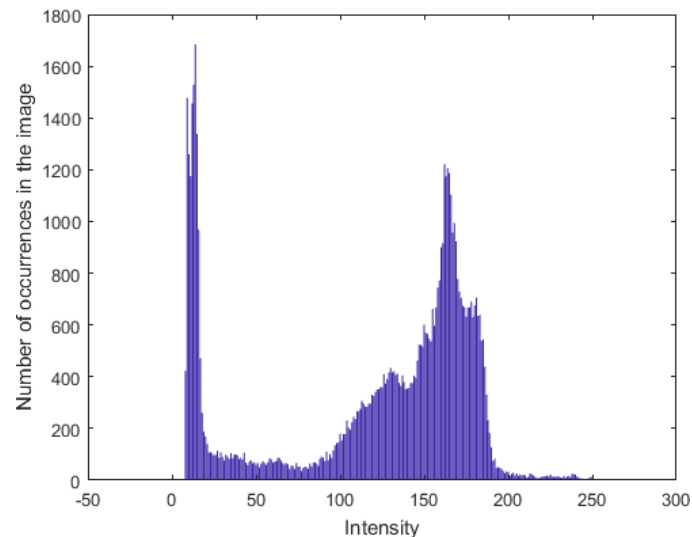
Probability mass function:

$$p_X(x) = P(X = x)$$



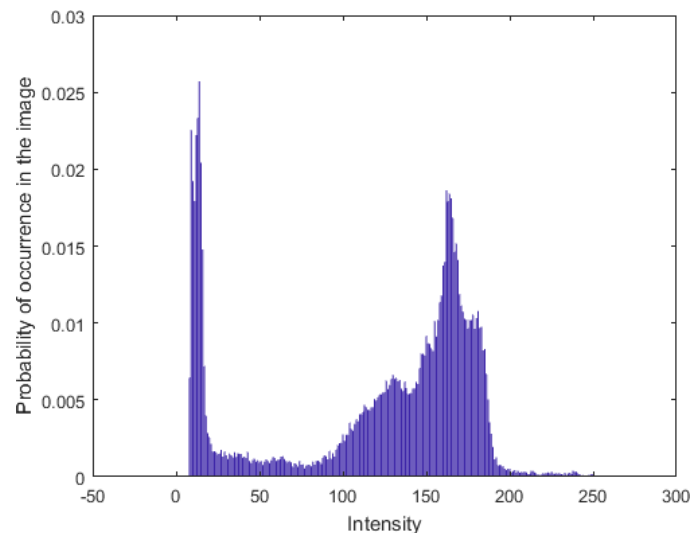
How can we define the probability mass function for the image intensities?

Image histogram – count of the number of occurrences of each intensity value in the image.

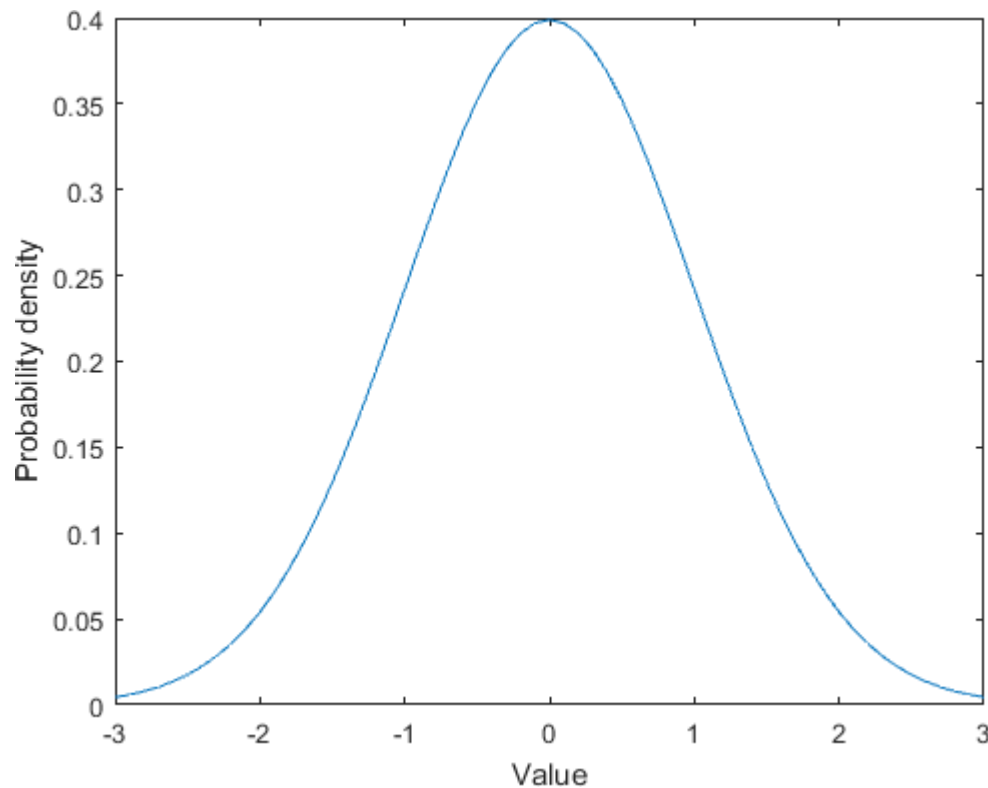


In order to treat the counts of the histogram as probability values, we must normalize the histogram in such a way that all values sum to 1.

This is the probability mass function for the pixel intensity as a random variable.



For continuous random variables (can take infinite number of possible values), we can define the probability density function (p.d.f):





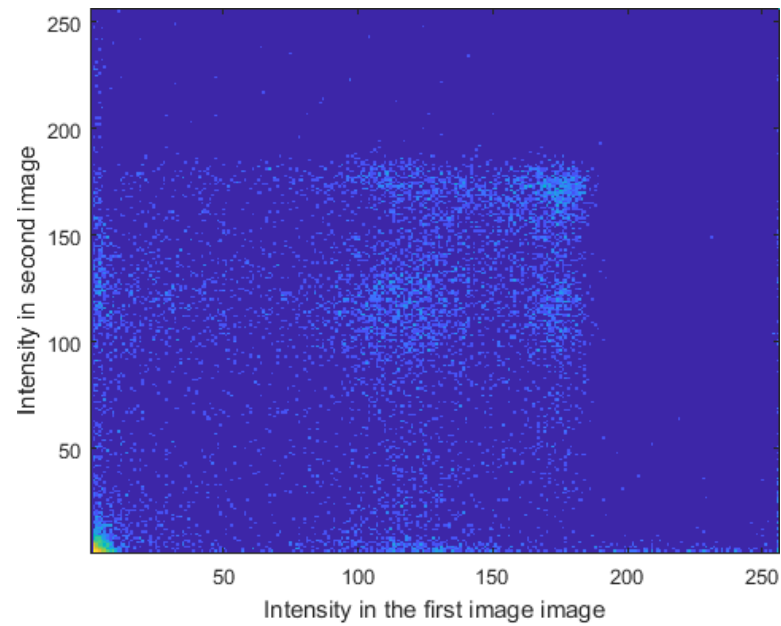
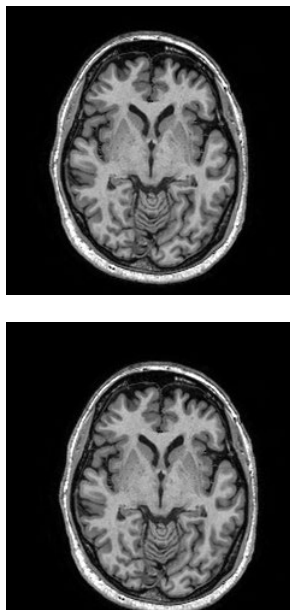
What if we have two random variables? For example, the pixel intensity in two images.

In this case we can define a joint probability mass function:

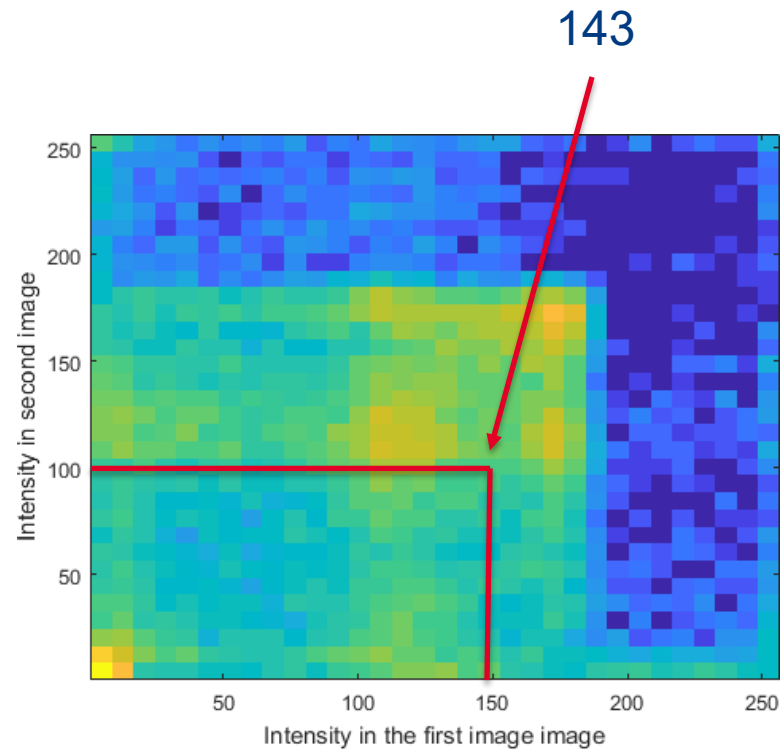
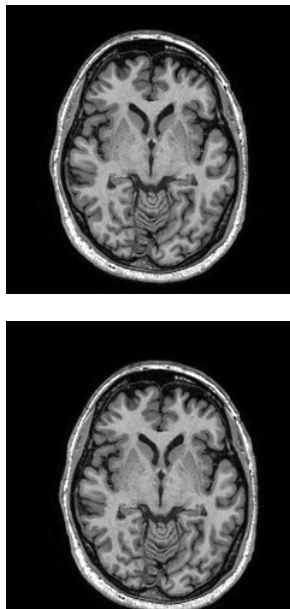
$$p_{X,Y}(x, y) = P(X = x, Y = y)$$



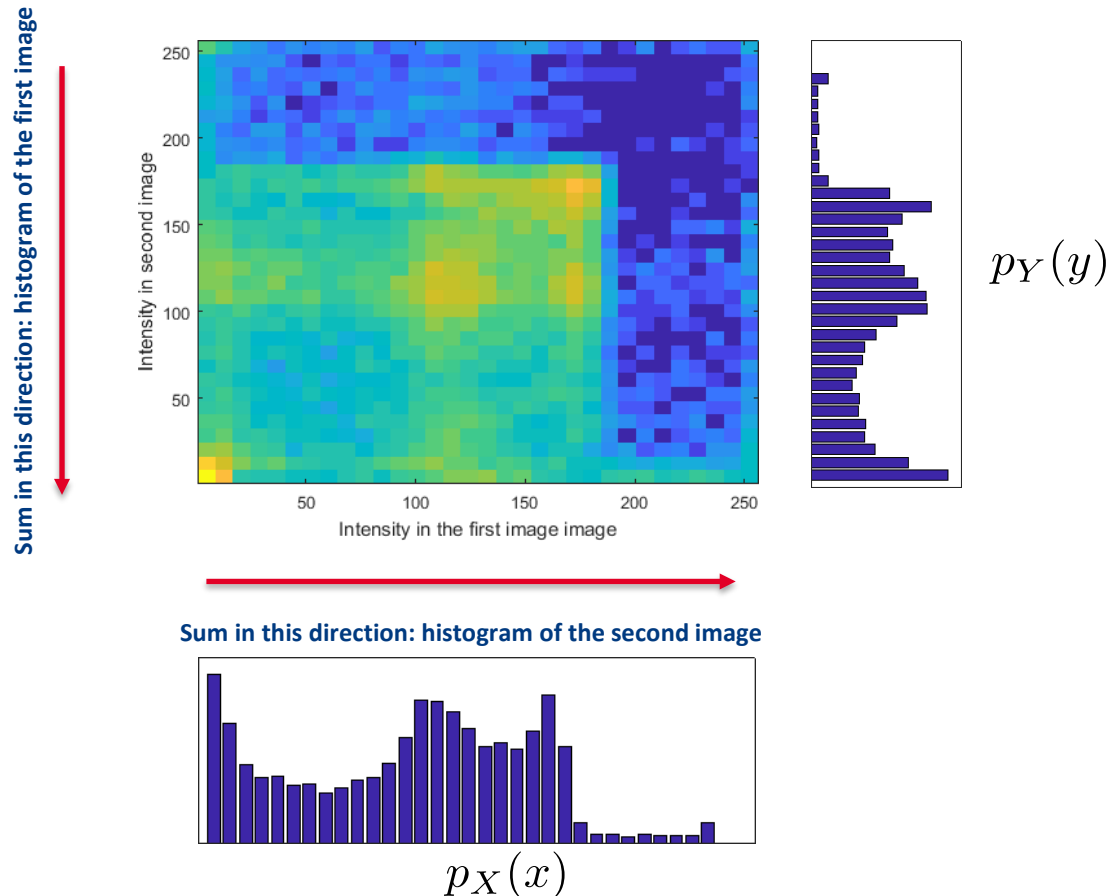
Example: the pixel intensity in two images. We can estimate this joint p.m.f. from the joint histogram of the two images.



Example: the pixel intensity in two images. We can estimate this joint p.m.f. from the joint histogram of the two images.



From the joint p.m.f. we can compute the p.m.f.'s of the individual variables (called marginal p.m.f.'s):



Conditional distributions answer the question: what is the probability of distribution over  $Y$  when we know that  $X$  takes a certain value?

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x, y)}{p_X(x)}$$

$$p_{X,Y}(x, y) = p_{Y|X}(y|x)p_X(x)$$

Example: if we pick a random image location, and the first image has intensity value of 124 at that location, what is the probability distribution for the intensity values in the second image at that location?



The random variables  $X$  and  $Y$  are independent if:

$$p_{X,Y}(x,y) = p_X(x)p_Y(y)$$



Bayes's rule is a very useful formula that we will use later in the computer-aided diagnosis sections of this course:

$$p_{Y|X}(y|x) = \frac{p_{Y|X}(x|y)p_Y(y)}{p_X(x)}$$



# Intensity-based similarity metrics

Continued.

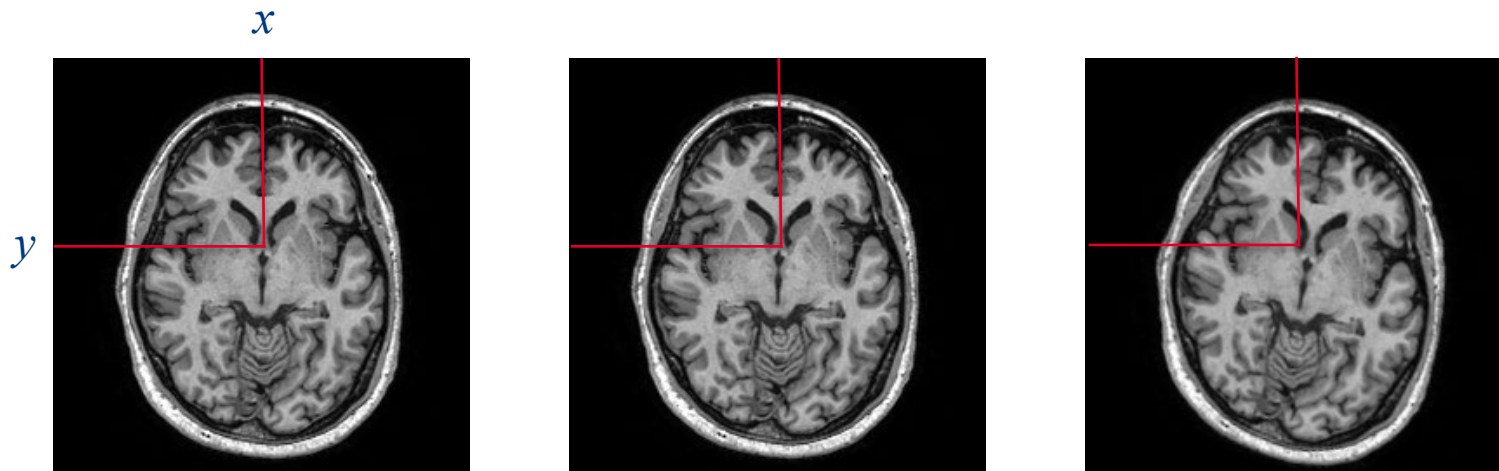




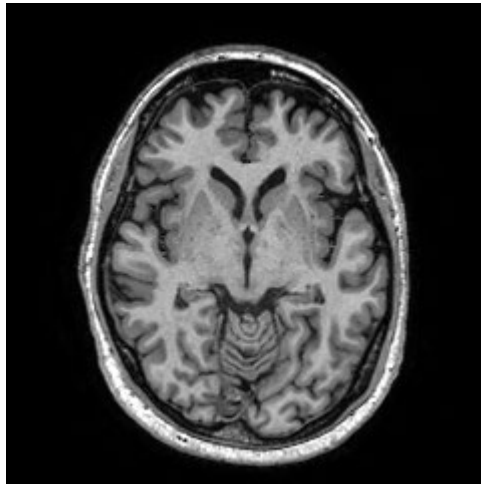
Another measure with even less assumptions: mutual information (MI).

An intuitive interpretation of the MI between two images:

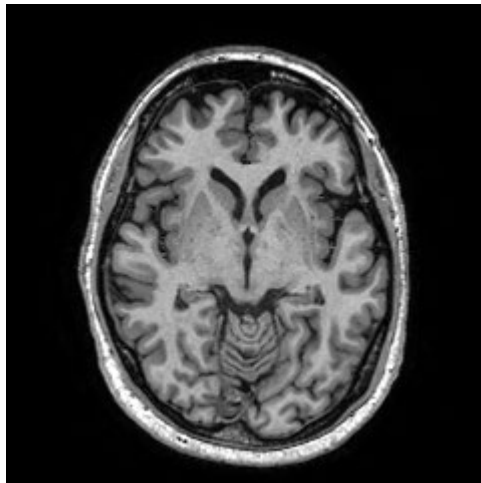
If we know the pixel intensity value at some location in the fixed image, how much information do we have about the pixel intensity at the same location in the other image?



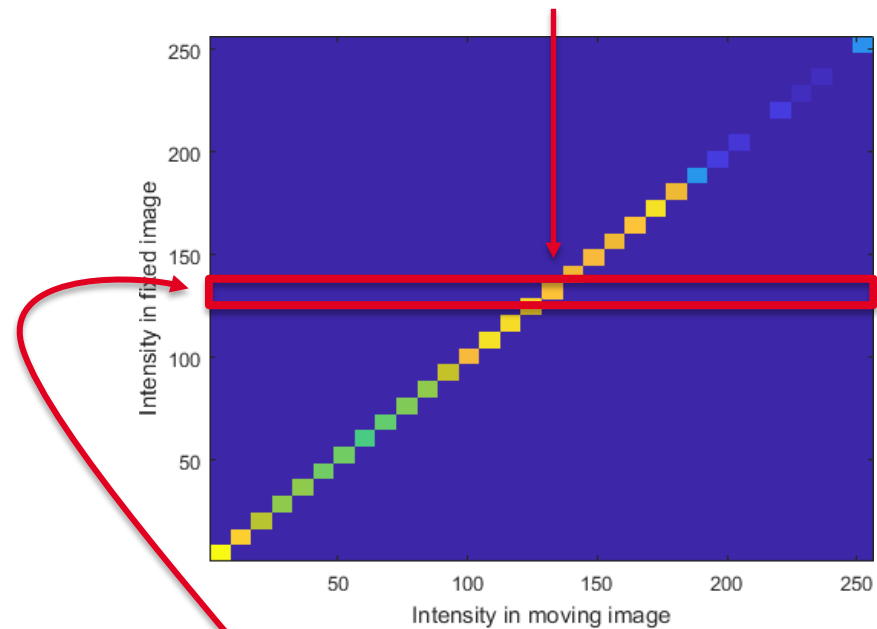
T1



T1



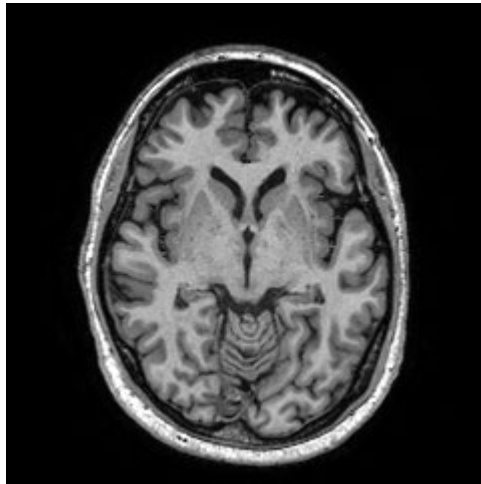
There is only one option for the other value  
(all other values have zero probability).



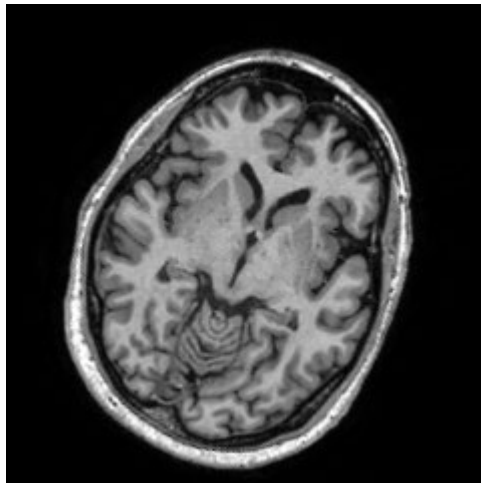
One value is “fixed”.



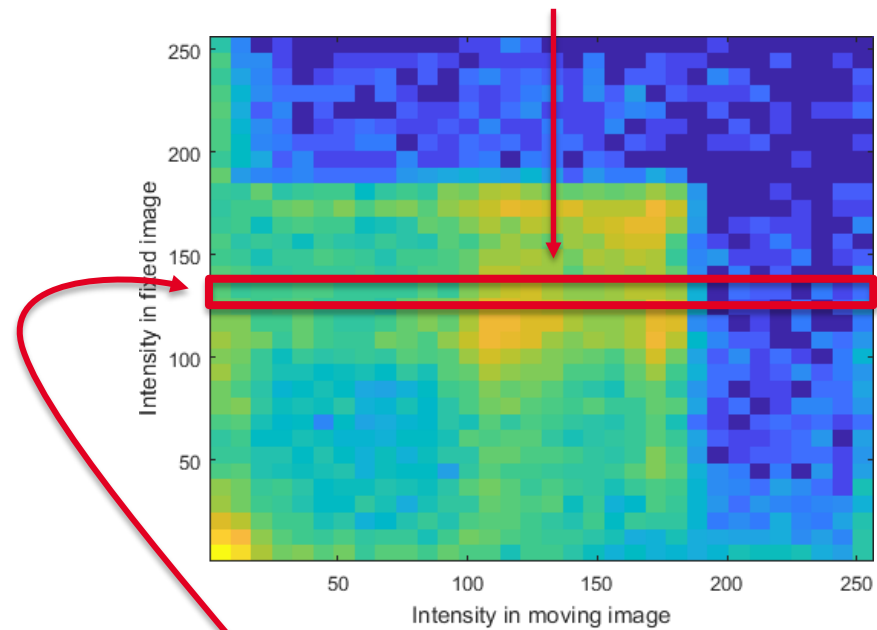
T1



T1



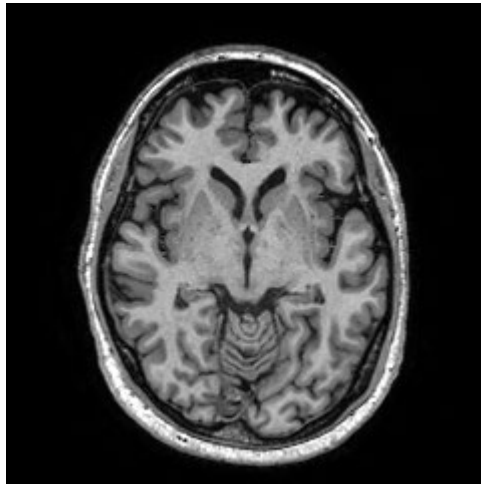
There are many probable values.



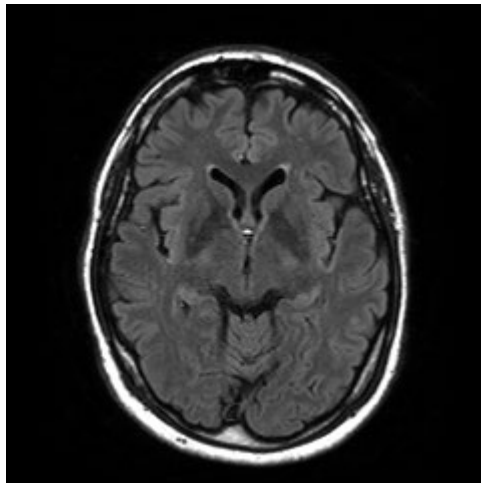
One value is "fixed".



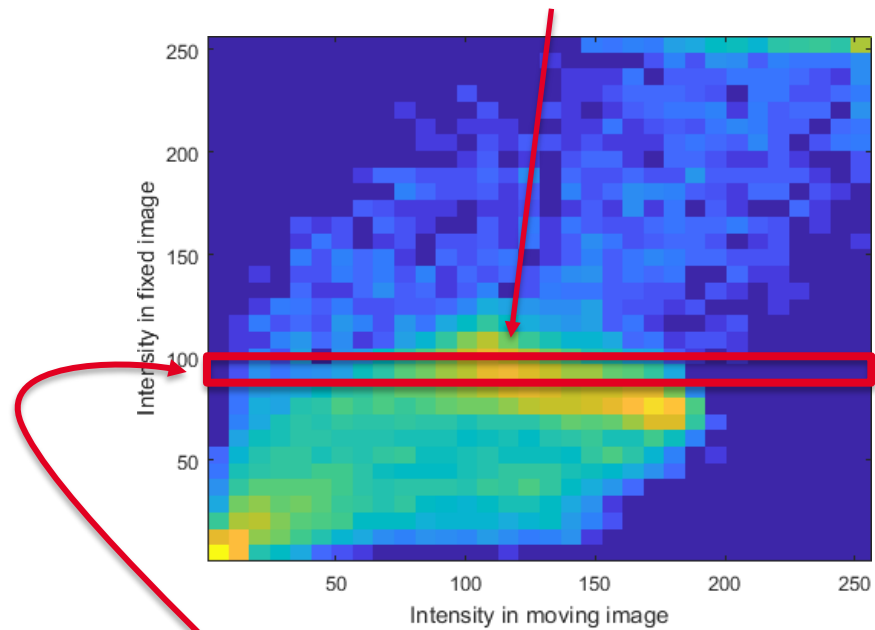
T1



T2



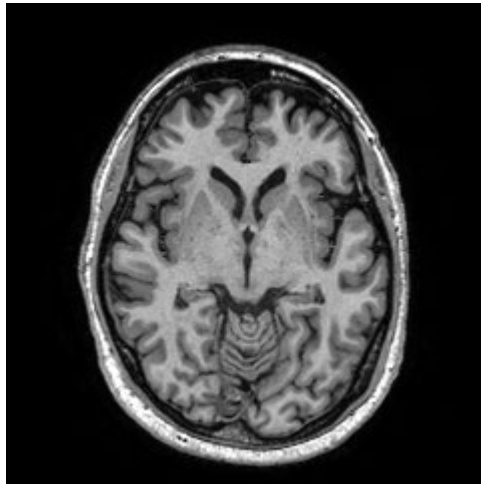
There are a few values with very high probability.



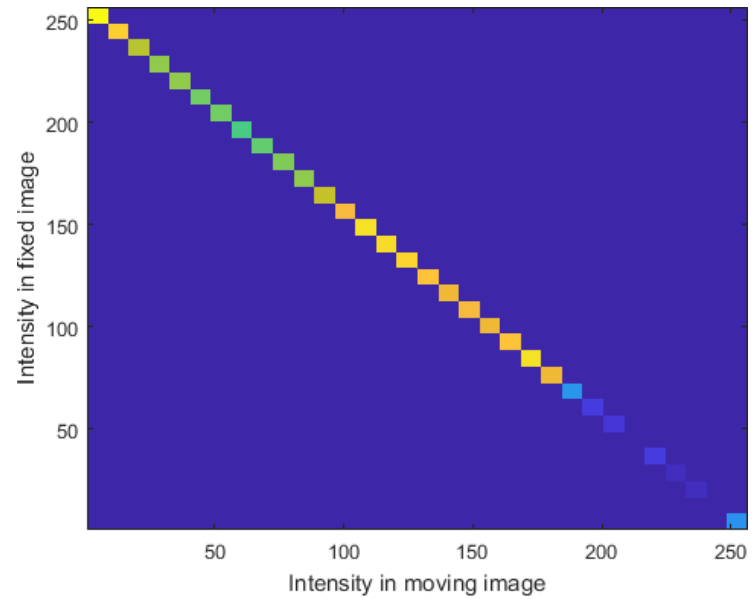
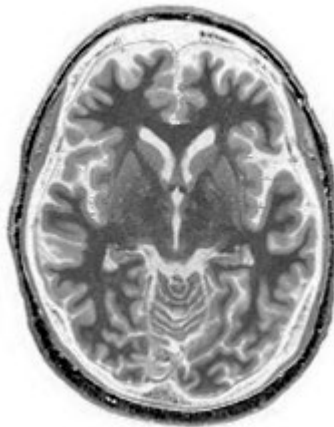
One value is "fixed".



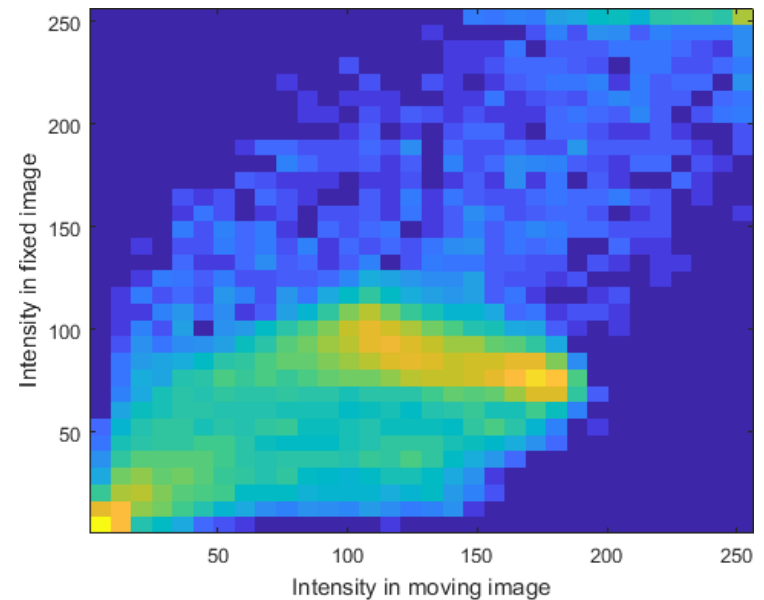
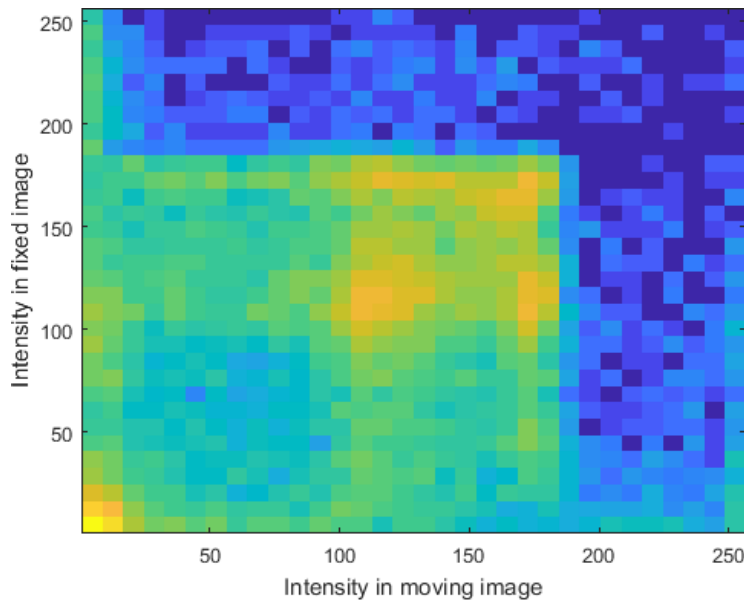
T1



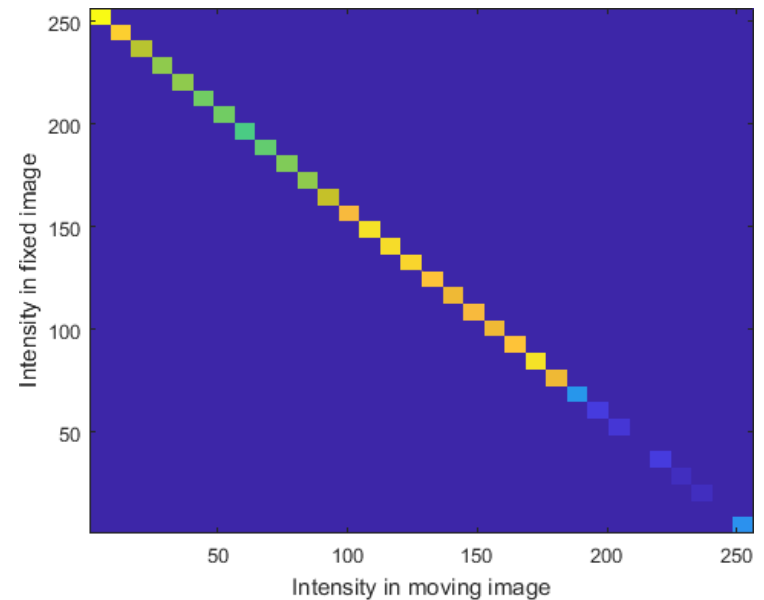
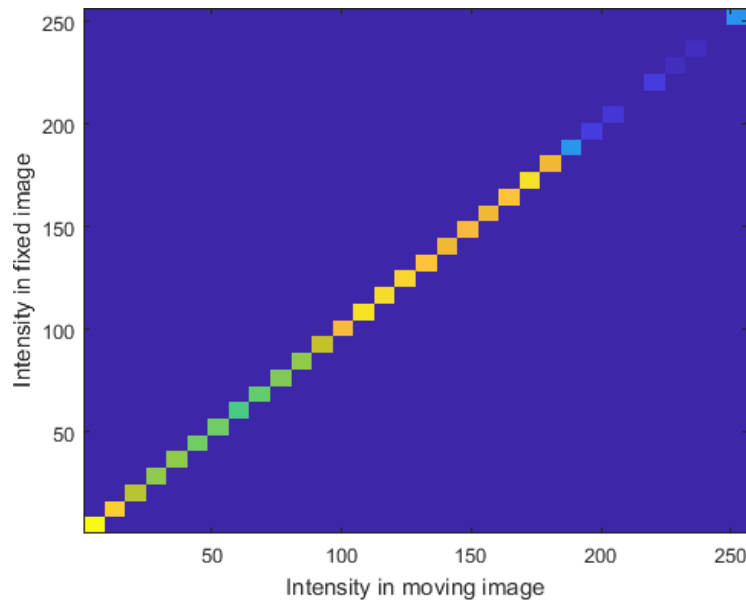
Simulated  
modality



Which pair of images is better aligned according to the joint histogram?  
Which histogram is more “compact”?



Which pair of images is better aligned according to the joint histogram?  
Which histogram is more “compact”?



Given the joint p.m.f. of two images and the two marginal p.m.f.'s, the mutual information between the two images can be computed with the following formula:

$$MI(I, J) = \sum_{i=1}^n \sum_{j=1}^n p_{I,J}(i, j) \log \frac{p_{I,J}(i, j)}{p_I(i)p_J(j)}$$

The unit of MI depends on the particular *log* function: when using the natural logarithm the unit is *nats*, when using base 2 logarithm the unit is *bits*.





MI in essence is a measure of the “compactness” of the joint p.m.f. of two images.

When the two images are well registered the joint p.m.f. is compact.

When the two images are not well aligned the joint p.m.f. is “spread out”.



We have now defined several intensity-based similarity measures.

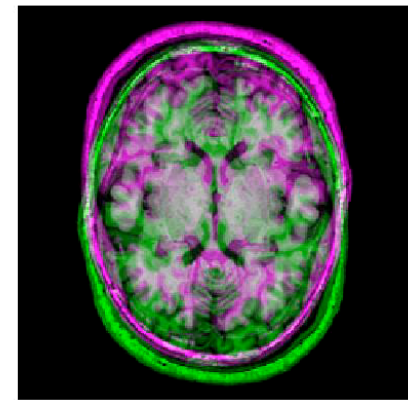
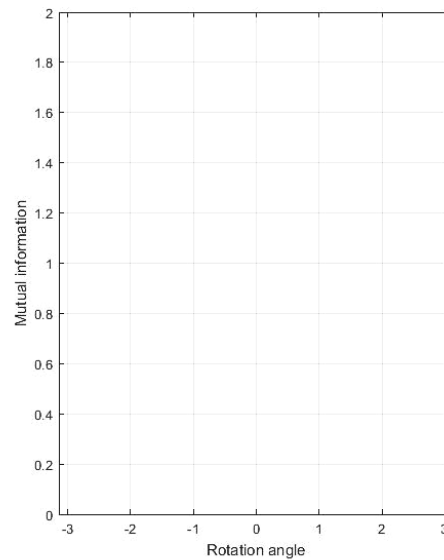
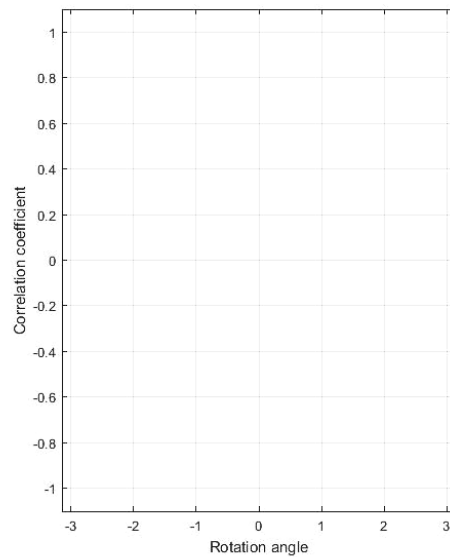
When one of the images is being transformed, the similarity measures are a function of the image transformation.

This is “step 1” in our general approach to registering two images.

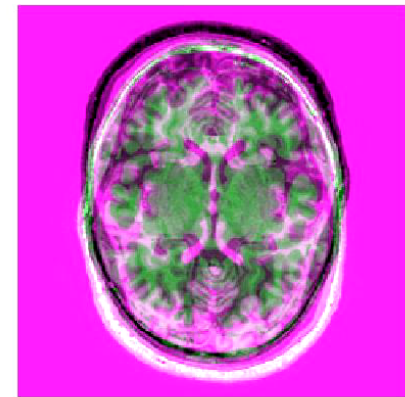
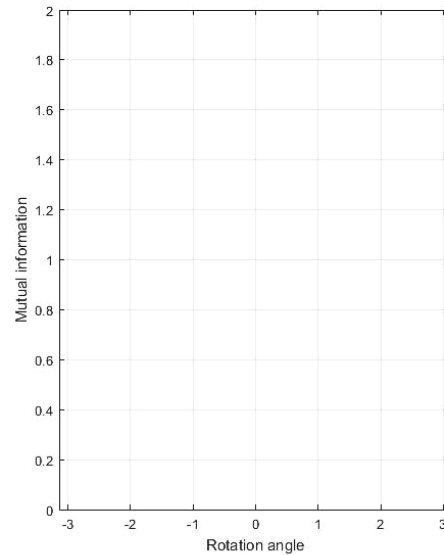
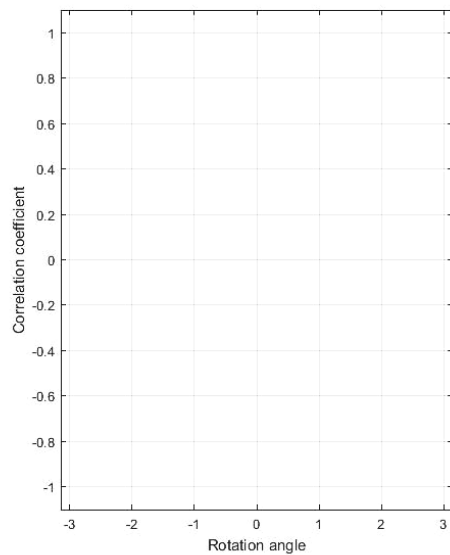
“Step 2” is finding the parameters that find the transformation that maximizes the similarity between two images.



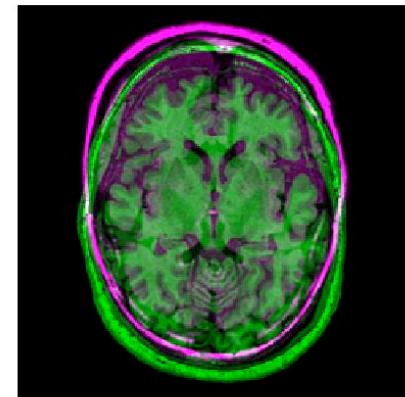
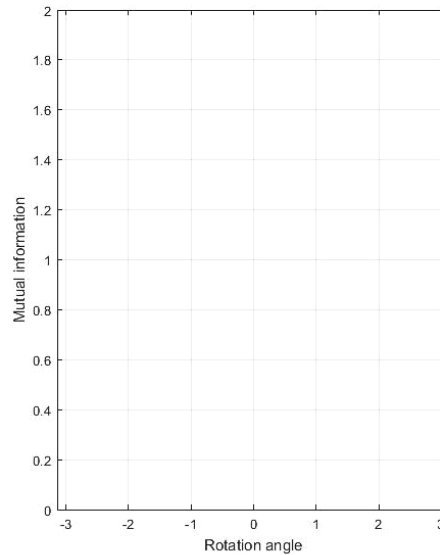
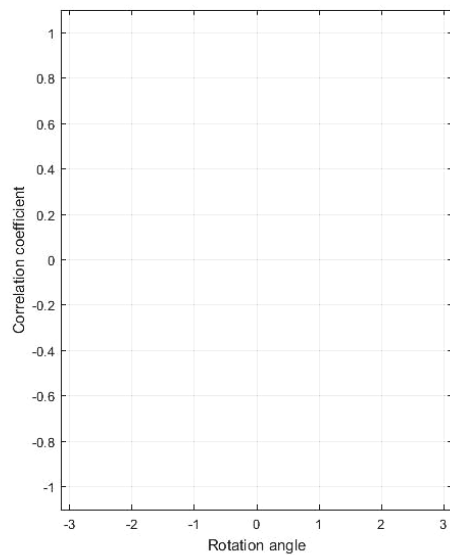
## Similarity as a function of transformation (T1 to T1):



## Similarity as a function of transformation (T1 to sim. modality):



## Similarity as a function of transformation (T1 to T2):

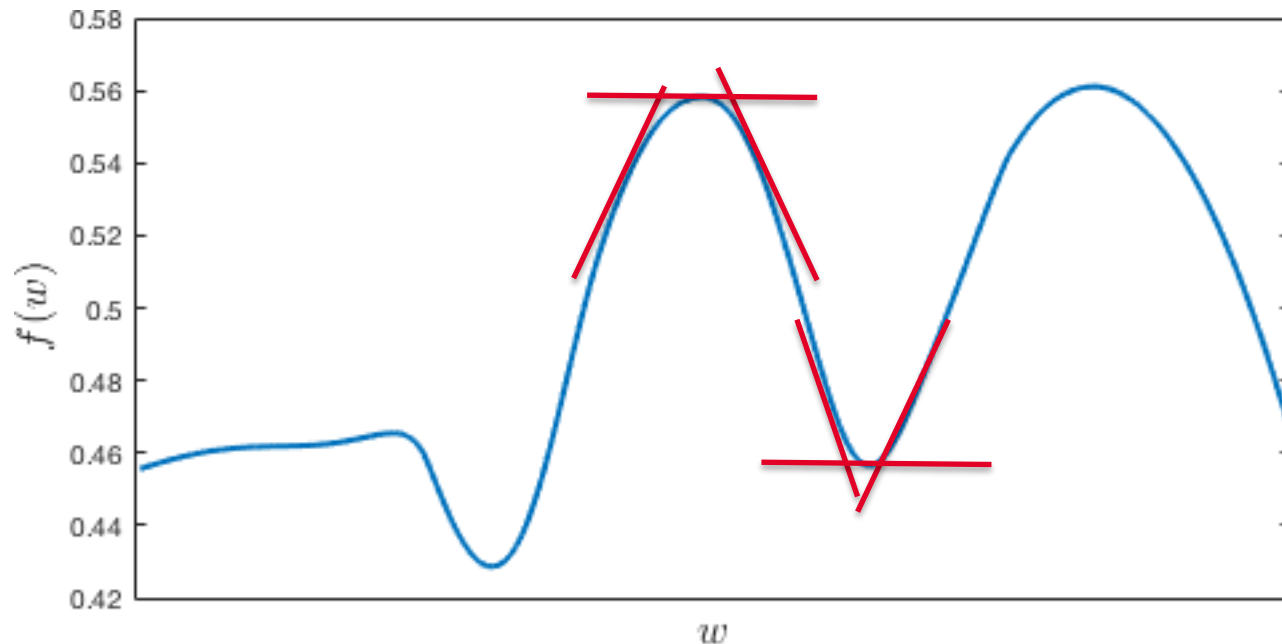


# Optimization

Continued.

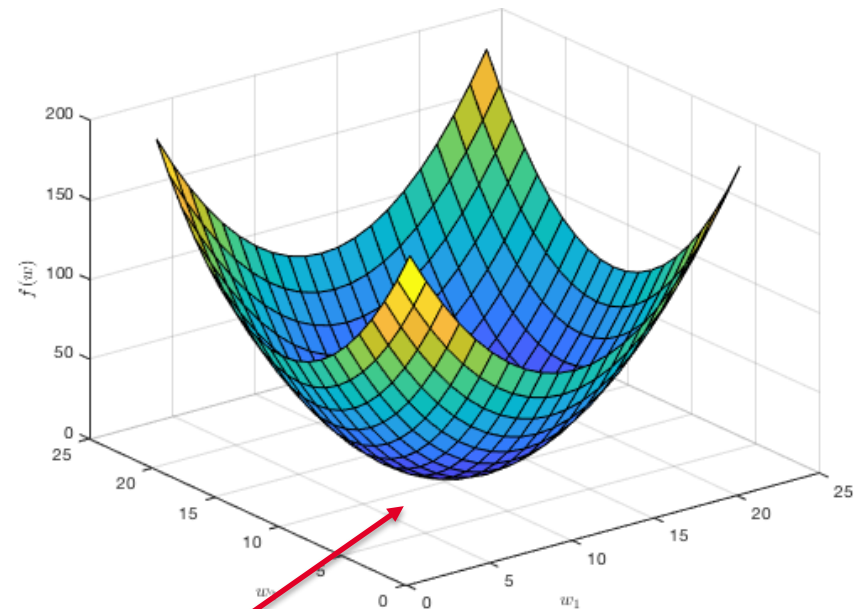
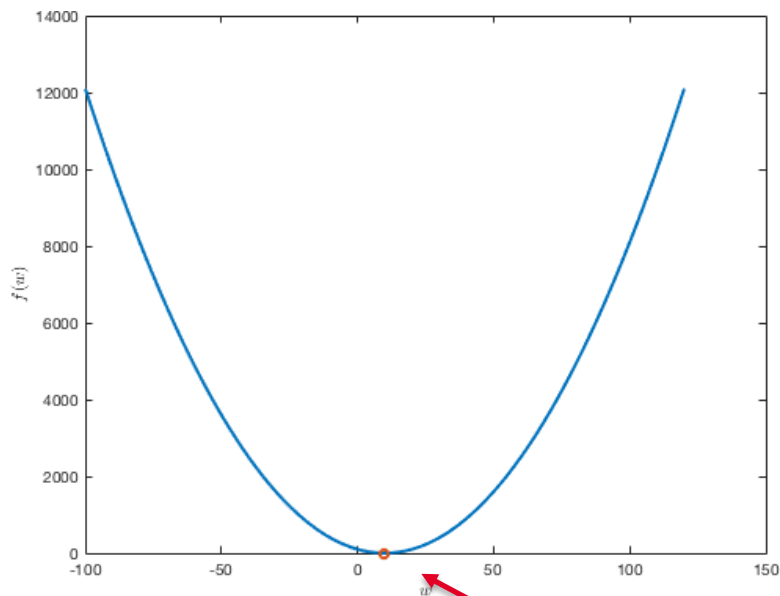


How to find the min. and max. of this function analytically?



Compute the derivative and set it to zero. If the function has more than one variable, set the partial derivatives (or gradient vector) to zero. 

## Convex functions:



Single minimum





For point based affine registration we had:

$$E(\mathbf{T}) = \|\mathbf{T}\mathbf{X}' - \mathbf{X}\|_F^2$$

$$\nabla_{\mathbf{T}} E(\mathbf{T}) = 0$$

$$\mathbf{T} = \mathbf{X}'\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}$$

Find the expression of the gradient and set it to zero. This will result in a system of equations.

The solution of this system is the optimal value of  $\mathbf{T}$ .



However, it might be the case that the system of equations produced by setting the gradient to zero is not solvable.

In that case we have to resort to numerical methods for finding the minimum of the error (or the maximum of the similarity).

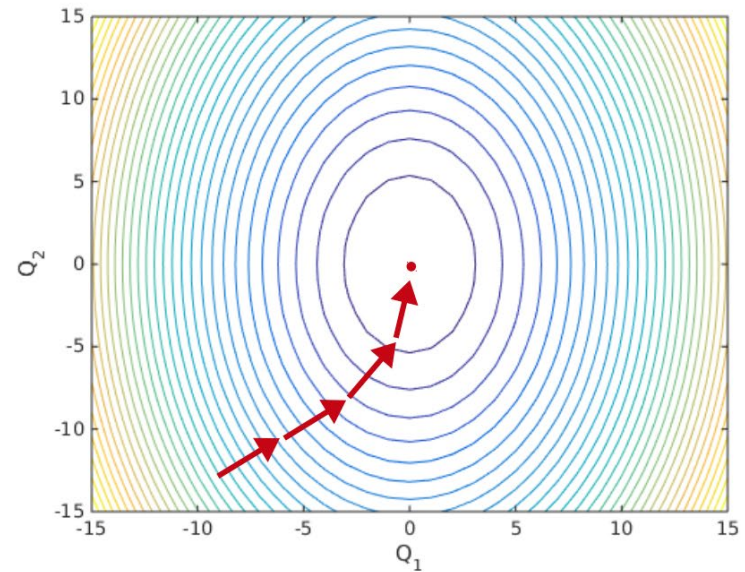
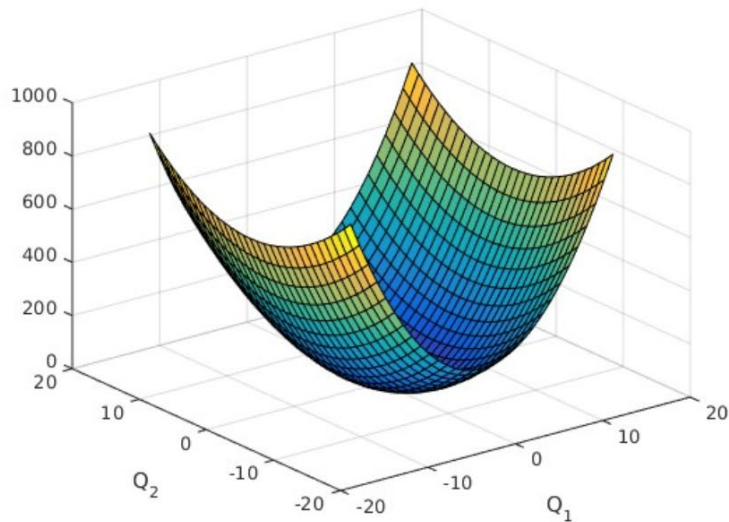
General procedure (for maximization of a similarity function):

1. Start with some initial values for the parameters (in this case the transformation  $\mathbf{T}$ ).
2. Slightly update the parameters in such a way that the similarity will slightly increase.
3. Repeat until the similarity stops increasing.

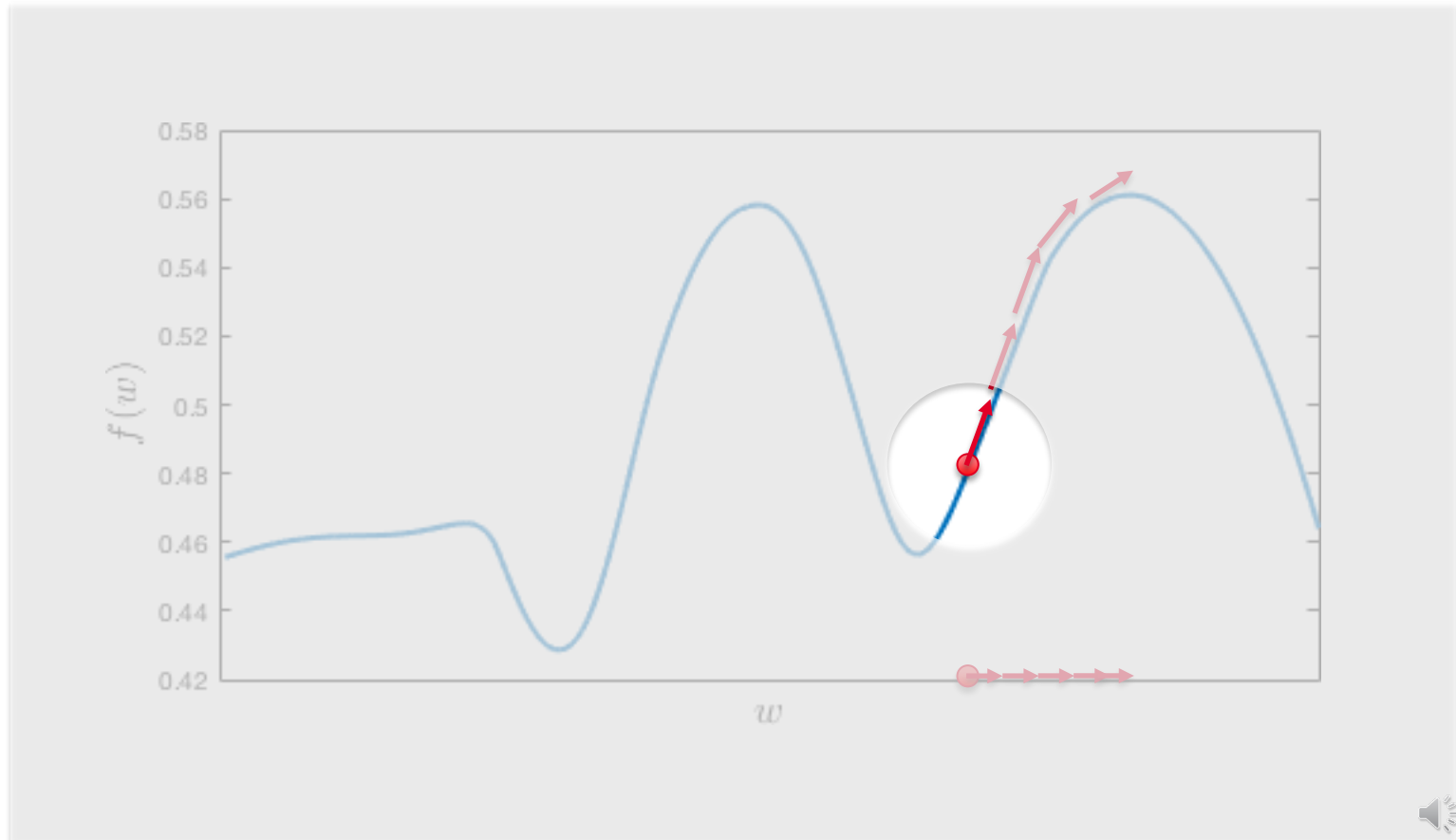




## Minimizing a function with two parameters:



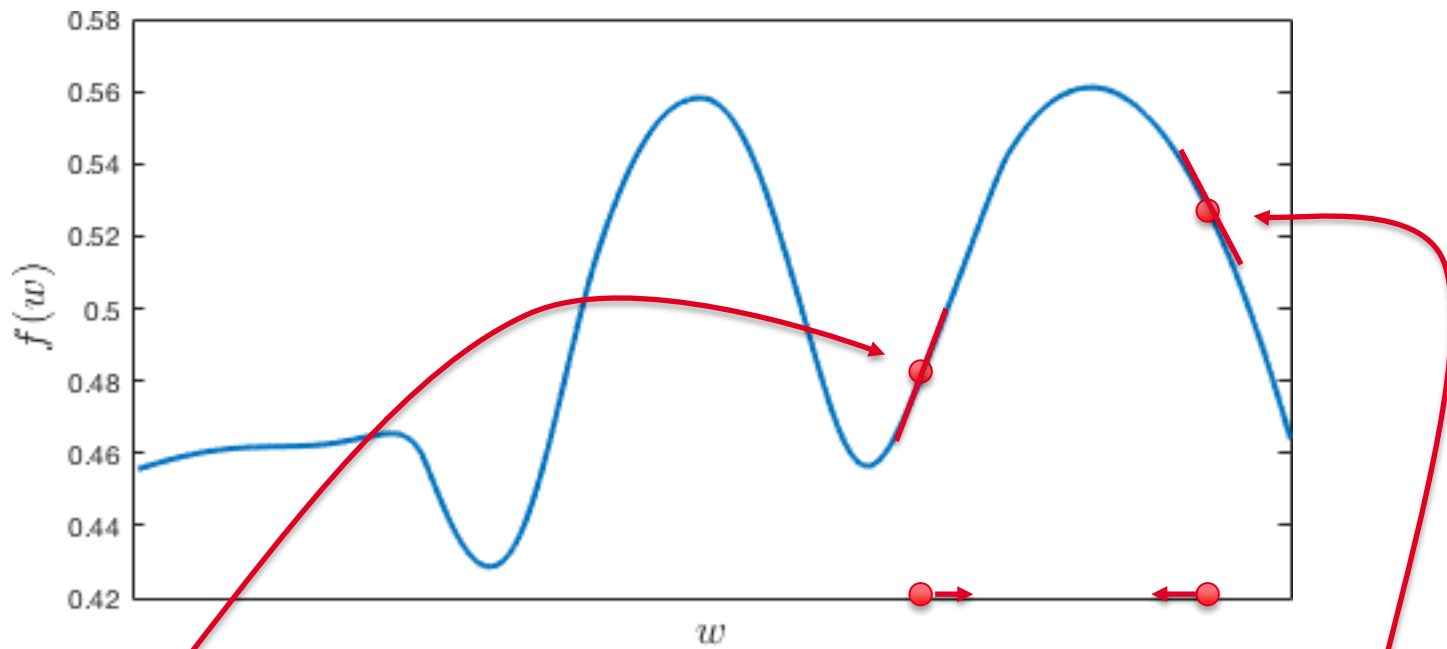
How to find out in which direction to do the parameter update?



How to find out in which direction to do the parameter update?



How to find out in which direction to do the parameter update?



If we are here → positive gradient → increase the parameter (move to the right)

If we are here → negative gradient → decrease the parameter (move to the left)

Gradient ascent algorithm for maximizing a function  $f(\mathbf{w})$ :

1. Choose some initial values of the parameters  $\mathbf{w}$
2. Calculate the value for the gradient of  $f(\mathbf{w})$  for the current parameters
3. Update the parameters in the direction of the gradient:

$$\mathbf{w} \leftarrow \mathbf{w} + \mu \nabla_{\mathbf{w}} f(\mathbf{w})$$

If we want to minimize the function we move in the direction opposite of the gradient (gradient descent):

$$\mathbf{w} \leftarrow \mathbf{w} - \mu \nabla_{\mathbf{w}} f(\mathbf{w})$$

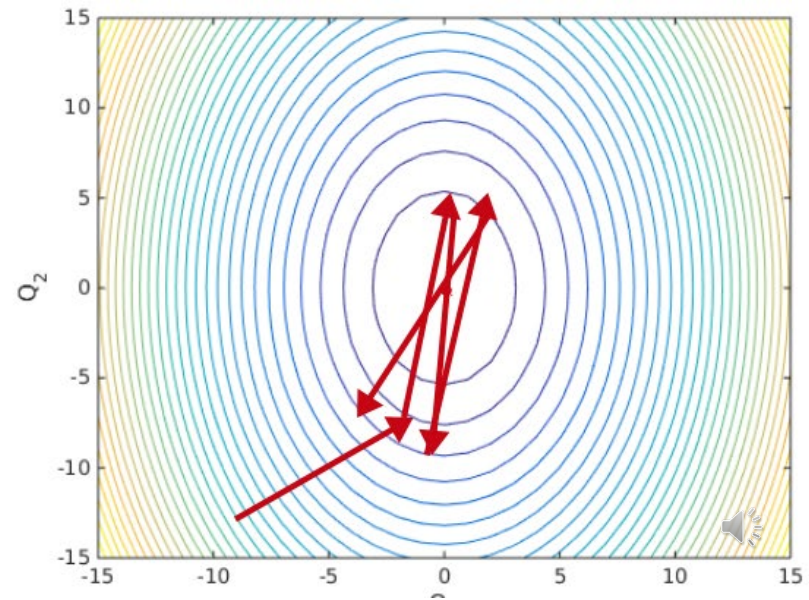
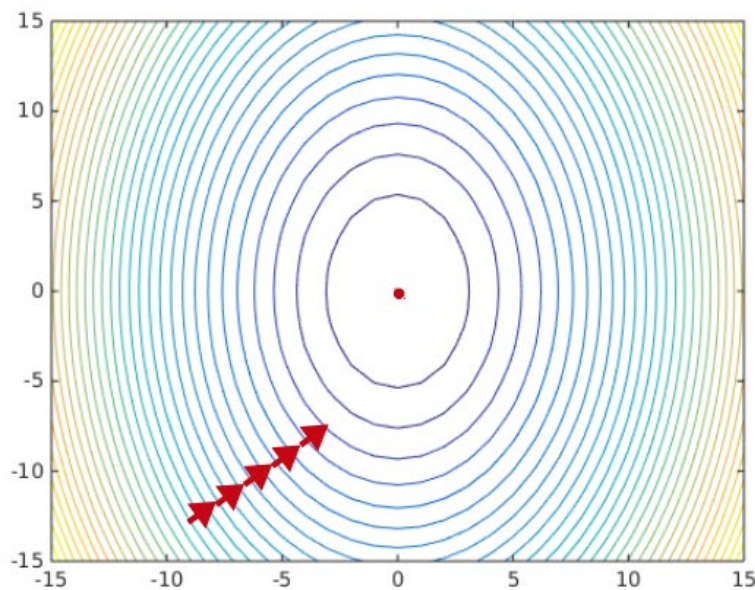




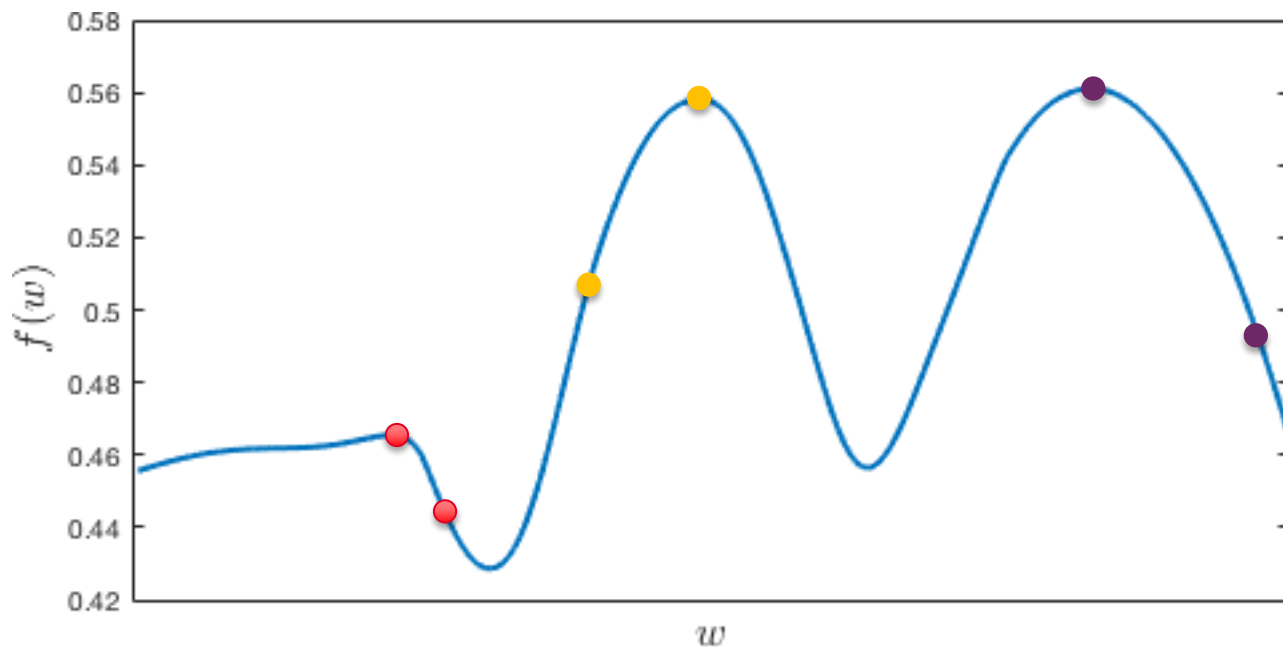
The parameter  $\mu$  is called learning rate. It controls how fast we move towards the maximum (minimum).

If  $\mu$  is too small, the maximum (or minimum) might not be reached in reasonable time.

If  $\mu$  is too large, the maximum (minimum) might be missed.



Initialization is important. Different starting points will result in different found maxima (and not always global).

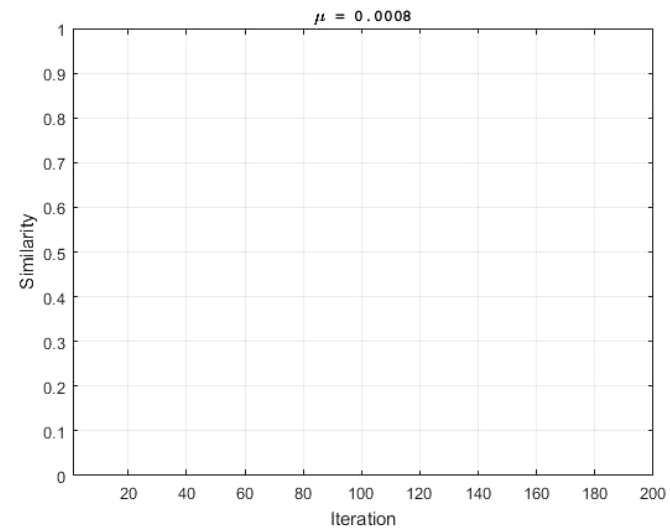
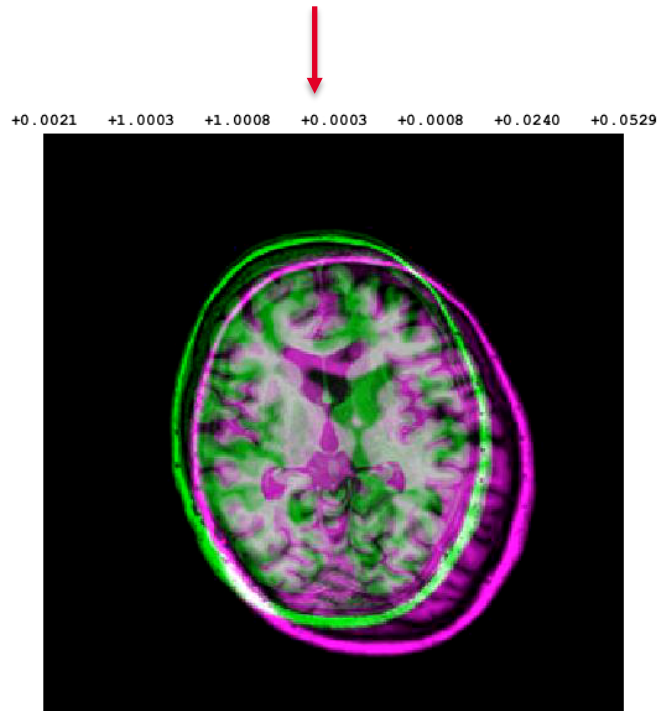


# Intensity- based image registration

examples



parameters of affine transformation

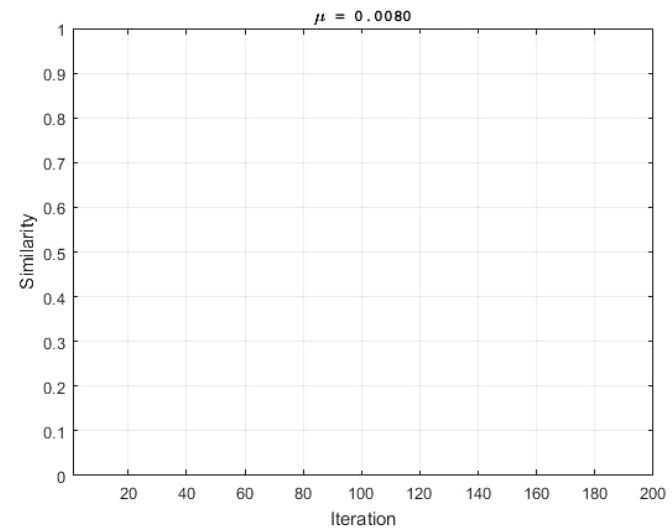
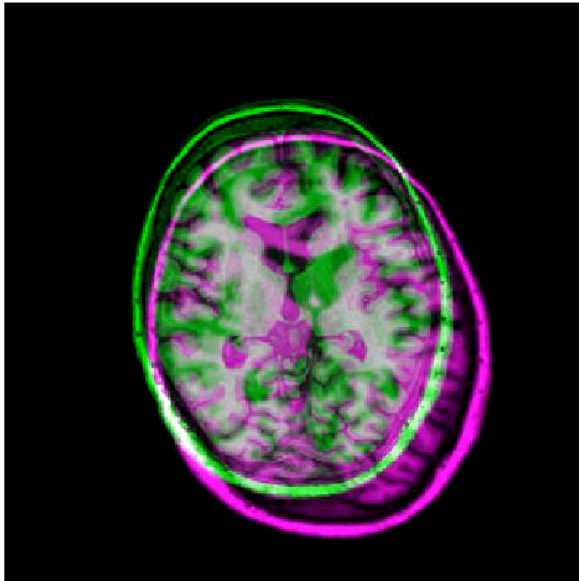


$$\mathbf{w} \leftarrow \mathbf{w} + \mu \nabla_{\mathbf{w}} f(\mathbf{w})$$

gradient of the similarity:  
SSD, CC or MI



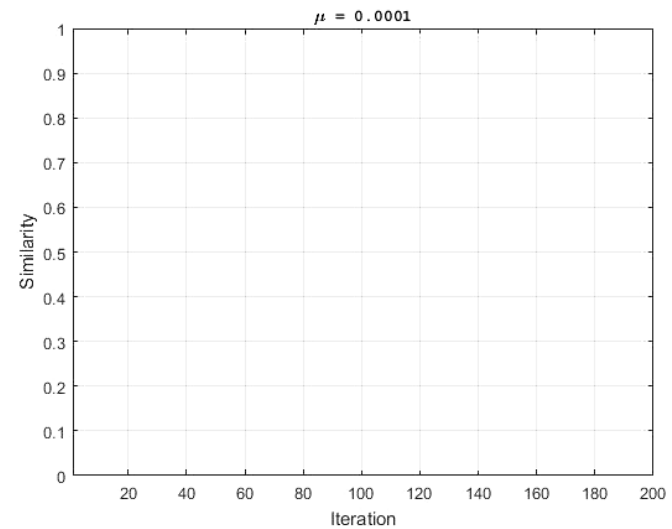
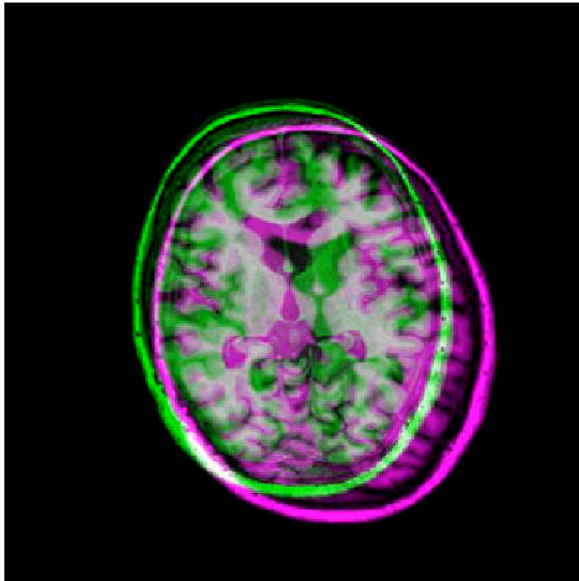
+0.0208 +1.0033 +1.0076 +0.0025 +0.0076 +0.2396 +0.5291



$$\mathbf{w} \leftarrow \mathbf{w} + \mu \nabla_{\mathbf{w}} f(\mathbf{w})$$



+0.0003 +1.0000 +1.0001 +0.0000 +0.0001 +0.0030 +0.0066



$$\mathbf{w} \leftarrow \mathbf{w} + \mu \nabla_{\mathbf{w}} f(\mathbf{w})$$

