

Chapter 7: “Platonic” Holography

Introduction

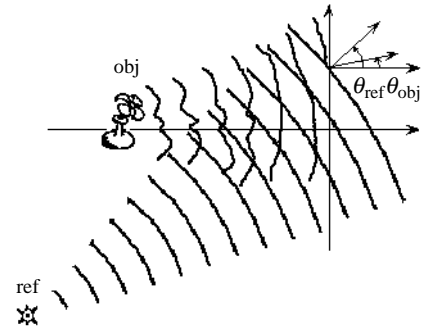
With simple concepts of interference and diffraction, we are ready to “prove” the validity of holography in a fairly simple and interesting way, based only on mathematics. The generality of the proof may come in handy later on, but the lack of practicality of the argument deprives it of much practical value in solving problems. If Gabor had lived in a cave, as Plato proposed to do, his proof might have looked something like this (actually, it does anyway!). But luckily he also spent plenty of time in the laboratory, and showed us how to produce pictures that would convince the doubters that this revolutionary approach to imaging could actually work!

The Object Beam

We represent the optical wave scattered by a generalized diffusely-reflecting object as a wave having an amplitude and phase that are random variables of x and y , and call it the *object beam*¹. The object beam is usually incident roughly perpendicularly to the recording plate. We will assume that the diffuse object reflection preserves the polarization of the beam (for example, that the object is aluminum spray-painted). The object wave has wavelength λ_1 , the recording wavelength, and the corresponding temporal frequency, ν_1 :

$$E_{\text{obj}}(x, y, t) = a_{\text{obj}}(x, y) \sqrt{\frac{2}{\epsilon_0 c}} \sin(2\pi \nu_1 t - \phi_{\text{obj}}(x, y)). \quad (1)$$

The average of the square of the amplitude, a_{obj} , is unity, so that the average *intensity* of the object beam is unity (which is why we included the $\epsilon_0 c$ term in the expression; see Eq. 14 of Ch. 2). Ordinarily, a_{obj} will have an exponential probability distribution function, and the variation of its autocorrelation function with distance will be closely related to the distribution of intensity in the object as measured by the angle it subtends at the plate, which determines the size of the “speckles” or intensity non-uniformities in the object beam. The object beam’s phase is also a random variable, uniformly distributed over $[0, 2\pi]$, so that it has a meaningless average. Note that although a_{obj} and ϕ_{obj} are random functions of x and y , they do not change with time—that is, the exposure system is stable during the exposing time.



The Reference Beam

By contrast, the reference beam is constant in intensity over the plate, but can have any phase variation with x and y . For simplicity, we will assume that it is a plane wave incident at an angle θ_{ref} (perhaps 30°). The reference beam intensity has to be greater than that of the object beam by some factor, K , which we call the “beam ratio.” This typically varies from 5 to 50. We can express this reference wave in the form

$$E_{\text{ref}}(x, y, z, t) = \sqrt{K} \sqrt{\frac{2}{\epsilon_0 c}} \sin\left(2\pi \nu_1 t - \frac{2\pi}{\lambda_1} x \sin \theta_{\text{ref}}\right). \quad (2)$$

The Interference Pattern

Where the object and reference beams overlap, an interference pattern is formed between them. This can be considered as a simple two-beam interference pattern, although now the amplitude and phase of one of the beams is gradually varying with x and y . Thus, continuing from Eq. 7 of Ch. 4, we find the total intensity pattern to be given by,

$$I_{\text{total}}(x, y) = K + a_{\text{obj}}^2(x, y) + 2\sqrt{K} a_{\text{obj}}(x, y) \cos\left(\phi_{\text{obj}}(x, y) - \frac{2\pi}{\lambda_1} x \sin \theta_{\text{ref}}\right). \quad (3)$$

Here, the first two terms are the intensities that would be found if the reference or object beams were turned on separately from each other. The third term is the *holographic* term, the fringe pattern that arises from interference between the two beams. It is the recording of this pattern that provides the necessary information for the reconstruction of an accurate three-dimensional image.

The Holographic Recording Material

The link between the exposure pattern and the reconstructed image is the recording material and its processing. The exposure is a positive real variable, a scalar, as no known material responds to anything but the “heat” of the exposure, the integration of its local intensity or irradiance over the exposure time. Ordinarily, the intensity is a constant over the duration of the exposure, which is gated by a shutter somewhere in front of the laser. The effect of

the exposure is to effect some chemical or physical change in the material, which produces a change in the optical properties of the material (usually after some further steps called “processing”). The properties we are concerned with most are the amplitude and phase transmittances of the material, as discussed in Ch. 6.

The most commonly-used recording material is silver-halide photographic film, a suspension of very small (approx. 35 nanometer diameter) silver bromide (mostly) micro-crystals in gelatin, plus sensitizers and other odds and ends. The absorption of photons creates tiny clusters of silver atoms on the grain surfaces that can catalyze the conversion of the entire microcrystal into a spherical or wormy “grain” of metallic silver during “development”—the grain is first “black” or light absorbing. Further processing steps can significantly improve the results, but let’s stay with this simple approach for now. A subsequent “fixing” step removes all but the developed silver grains, and the film is then washed and dried to produce an apparently-continuous spatial variation of light absorption, described by its transmittance. We will talk here mostly about the *amplitude transmittance*, the ratio of the electric field amplitudes just after and just before the film layer, denoted as t_{amp} ². It is important that the resolution of the material be high enough to allow the film to “follow” very fine-scale variations of exposures. For very low exposures the transmittance is nearly unity, and as the exposure increases the transmittance drops monotonically to less than 0.1.

The response of a recording material (almost any material) can be expressed graphically as a relationship between the amplitude transmittance, t_{amp} , and the exposure, EXP , which is the product of the exposing intensity and the exposure time. That relationship is generally non-linear, and perhaps not even monotonic, as sketched in the margin. However, over some limited range of exposures the transmittance varies nearly linearly with exposure, and be approximated by a straight line, which is the range of exposures where we will try to make holograms. Thus a *linearized* model of a recording material expresses this mathematically as:

$$t_{amp}(x, y) - t_0 = \left. \frac{\partial t_{amp}}{\partial EXP} \right|_{E=E_0} (EXP(x, y) - EXP_0), \quad (4)$$

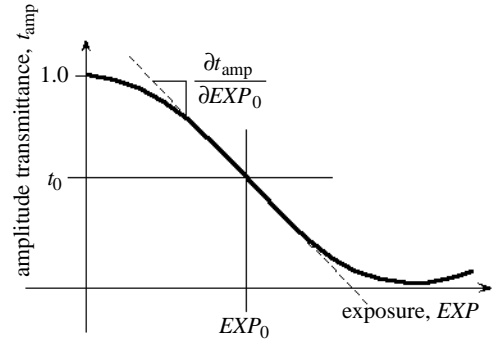
where EXP_0 is the so-called “bias” exposure around which the response is reasonably linear, and t_0 is the transmittance produced by a uniform exposure at that bias level.

The holographic recording material will be exposed to this pattern for some period of time, T_0 , so as to bring the spatially-averaged exposure to the required level, EXP_0 . The exposure at any point is given by $EXP(x, y) = I_{total}(x, y) \cdot T_0$, so the needed exposure time, T_0 , is given by

$$T_0 = \frac{EXP_0}{\langle I_{total}(x, y) \rangle_{spatial\ average}} = \frac{EXP_0}{K + 1}. \quad (5)$$

The amplitude transmittance as a function of intensity then becomes

$$\begin{aligned} t_{amp}(x, y) &= \frac{1}{K + 1} \left(EXP_0 \left. \frac{\partial t_{amp}}{\partial E} \right|_{EXP=EXP_0} \right) I(x, y) + \left(t_0 - \left. \frac{\partial t_{amp}}{\partial E} \right|_{EXP=EXP_0} EXP_0 \right) \\ &= \frac{1}{K + 1} \left(\left. \frac{\partial t_{amp}}{\partial \ln EXP} \right|_{EXP=EXP_0} \right) I(x, y) + \left(t_0 - \left. \frac{\partial t_{amp}}{\partial \ln EXP} \right|_{EXP=EXP_0} \right). \end{aligned} \quad (6)$$



The Holographic Transmittance Pattern

Inserting the expression for the holographic intensity promptly gives the resulting amplitude transmittance pattern. The relevant characteristic of the recording material, the slope of the curve of its amplitude transmittance versus the natural logarithm of its exposure, is usually referred to as the “beta” of the material, β . It is sometimes multiplied by the “modulation transfer function” or MTF of the material at the resolution scale of the hologram (the MTF describes the % response to a sine-wave intensity exposure at the relevant spatial frequency). Thus, substituting Eq. 3 into Eq. 6, we obtain

$$\begin{aligned}
t_{\text{amp}}(x, y) &= \frac{1}{K+1} \beta I(x, y) + (t_0 - \beta) \\
&= \beta \frac{K}{K+1} + (t_0 - \beta) \\
&\quad + \beta \frac{1}{K+1} a_{\text{obj}}^2(x, y) \\
&\quad + 2\beta \frac{\sqrt{K}}{K+1} a_{\text{obj}}(x, y) \cos\left(\phi_{\text{obj}}(x, y) - \frac{2\pi}{\lambda_1} x \sin \theta_{\text{ref}}\right).
\end{aligned} \tag{7}$$

This is the ‘‘hologram!’’ Within its transmittance pattern is imbedded a precise description of the object beam, along with several other terms, awaiting only illumination by a suitable beam to release its information.

The Illuminating Beam

The illumination beam, like the reference beam, may be any uniform-intensity beam (with an arbitrary phase distribution), but we will limit our discussion to a unit-amplitude plane wave inclined at angle θ_{ill} . It has wavelength λ_2 , the reconstruction wavelength.

$$E_{\text{ill}}(x, y, t) = \sin\left(2\pi \nu_2 t - \frac{2\pi}{\lambda_2} x \sin \theta_{\text{ill}}\right). \tag{8}$$

The diffracted output from the hologram is then given by the product of the hologram amplitude transmittance and the illumination amplitude,

$$\begin{aligned}
E_{\text{out}}(x, y, t) &= t_{\text{amp}}(x, y) \cdot E_{\text{ill}}(x, y, t) \\
&= \left(\beta \frac{K}{K+1} + (t_0 - \beta)\right) \sin\left(2\pi \nu_2 t - \frac{2\pi}{\lambda_2} x \sin \theta_{\text{ill}}\right) \\
&\quad + \beta \frac{1}{K+1} a_{\text{obj}}^2(x, y) \sin\left(2\pi \nu_2 t - \frac{2\pi}{\lambda_2} x \sin \theta_{\text{ill}}\right) \\
&\quad + 2\beta \frac{\sqrt{K}}{K+1} a_{\text{obj}}(x, y) \cos\left(\phi_{\text{obj}}(x, y) - \frac{2\pi}{\lambda_1} x \sin \theta_{\text{ref}}\right) \sin\left(2\pi \nu_2 t - \frac{2\pi}{\lambda_2} x \sin \theta_{\text{ill}}\right).
\end{aligned} \tag{9}$$

It is the last of these terms that is of special interest to us, and to explore it we need to apply the same trig identity used previously, $\sin \alpha \cdot \cos \beta = \frac{1}{2} [\sin(\alpha + \beta) + \sin(\alpha - \beta)]$:

$$\begin{aligned}
E_{\text{last}}(x, y, t) &= 2\beta \frac{\sqrt{K}}{K+1} a_{\text{obj}}(x, y) \cos\left(\phi_{\text{obj}}(x, y) - \frac{2\pi}{\lambda_1} x \sin \theta_{\text{ref}}\right) \sin\left(2\pi \nu_2 t - \frac{2\pi}{\lambda_2} x \sin \theta_{\text{ill}}\right) \\
&= \beta \frac{\sqrt{K}}{K+1} a_{\text{obj}}(x, y) \sin\left(2\pi \nu_2 t + \phi_{\text{obj}}(x, y) - \frac{2\pi}{\lambda_1} x \sin \theta_{\text{ref}} - \frac{2\pi}{\lambda_2} x \sin \theta_{\text{ill}}\right) \\
&\quad + \beta \frac{\sqrt{K}}{K+1} a_{\text{obj}}(x, y) \sin\left(2\pi \nu_2 t - \phi_{\text{obj}}(x, y) + \frac{2\pi}{\lambda_1} x \sin \theta_{\text{ref}} - \frac{2\pi}{\lambda_2} x \sin \theta_{\text{ill}}\right).
\end{aligned} \tag{10}$$

We will represent these components as a sum over a variable, m , the ‘‘order number,’’ so that

$$E_{\text{out}}(x, y, t) = \sum_{m=-\infty}^{+\infty} E_m(x, y, t), \tag{11}$$

where

$$\begin{aligned}
E_0(x, y, t) &= \left(\beta \frac{K}{K+1} + (t_0 - \beta) \right) \sin \left(2\pi \nu_2 t - \frac{2\pi}{\lambda_2} x \sin \theta_{\text{ill}} \right) \\
&\quad + \beta \frac{1}{K+1} a_{\text{obj}}^2(x, y) \sin \left(2\pi \nu_2 t - \frac{2\pi}{\lambda_2} x \sin \theta_{\text{ill}} \right), \\
E_{+1}(x, y, t) &= \beta \frac{\sqrt{K}}{K+1} a_{\text{obj}}(x, y) \sin \left(2\pi \nu_2 t - \phi_{\text{obj}}(x, y) + \frac{2\pi}{\lambda_1} x \sin \theta_{\text{ref}} - \frac{2\pi}{\lambda_2} x \sin \theta_{\text{ill}} \right), \\
E_{-1}(x, y, t) &= \beta \frac{\sqrt{K}}{K+1} a_{\text{obj}}(x, y) \sin \left(2\pi \nu_2 t + \phi_{\text{obj}}(x, y) - \frac{2\pi}{\lambda_1} x \sin \theta_{\text{ref}} - \frac{2\pi}{\lambda_2} x \sin \theta_{\text{ill}} \right).
\end{aligned} \tag{12}$$

In general, there will be several higher-order components. It is only our assumption of linearity of the response of the recording material that has limited us to finding only the 0, +1, and -1 terms here. Also, either or both of the first orders may not actually exist, as they may turn out to be *evanescent* upon further analysis.

A Proof of Holography

It is the $m=+1$ diffracted wave that is the potential reconstruction of the object wave. If the angle and wavelength of the illumination beam are made equal to the angle and wavelength of the reference beam, the last two terms in the parentheses cancel out, leaving only the amplitude and phase terms identical to those of the object wave. These are the conditions that we refer to as “perfect reconstruction.” That is,

$$\begin{aligned}
&\text{if } \lambda_2 = \lambda_1 \text{ and } \theta_{\text{ill}} = \theta_{\text{ref}}, \text{ then} \\
E_{+1}(x, y, t) &= (\text{constant}) a_{\text{obj}}(x, y) \sin(2\pi\nu_1 t - \phi_{\text{obj}}(x, y)).
\end{aligned} \tag{13}$$

This represents a general statement of the central property of holography, that it can reproduce an exact replica of the amplitude and phase of the object wave under very general circumstances. The constant term reflects the diffraction efficiency of the hologram, or the brightness of the image it produces. If the object wavefront was produced by a three-dimensional scene, the reconstructed wavefront will be focused by the eyes to produce a three-dimensional perception of that scene. There is no “illusion” involved, the eyes are not being tricked—they are enjoying the same information that the scene itself would have provided, were it still there.

Note that part of the originally-inclined illumination wave has been deflected to travel along the z -axis, in the direction of the object beam’s light. This change of direction is caused by diffraction by an overall grating pattern caused by interference between the object and reference beams, and is sometimes referred to as a “spatial carrier wave” to make an analogy to the radio carrier wave used in AM and FM modulation. Its spatial frequency is determined mainly by the angle of the reference beam; that is, $f_{\text{carrier}} \approx \sin \theta_{\text{ref}} / \lambda_1$. A reference beam angle of 30° thus creates a grating of 790 cy/mm, or a grating spacing d of $1.27 \mu\text{m}$ (using a He-Ne laser). This tiny spacing presents most of the practical challenges of high-quality holography.

There do seem to be some physical paradoxes involved, of course. A purely two-dimensional recording is reconstructing information about a three-dimensional volume, for example! But this is a consequence of Huyghens’ principle (or the ellipticity of the wave equation, if you prefer) that a specification of the amplitude and phase boundary conditions specifies the wave throughout the enclosed volume. And, there is the paradox of reconstructing the amplitude and phase of a quantity from a purely scalar (intensity) recording. This is resolved by noting that we are also reconstructing some other terms that can be regarded as the “extra baggage” required to resolve this paradox.

The Other Reconstructed Components

The most interesting of the “extra baggage” terms is the $m=-1$ component, which is termed the “conjugate” or “twin” image³. Note that under “perfect reconstruction” its terms are

$$E_{-1}(x, y, t) = (\text{constant}) a_{\text{obj}}(x, y) \sin \left(2\pi \nu_1 t - \left(-\phi_{\text{obj}}(x, y) \right) - \frac{4\pi}{\lambda_1} x \sin \theta_{\text{ref}} \right). \tag{14}$$

Which is to say that although the amplitude is the same as for the object beam, the phase has the opposite sign. That is, a diverging object-beam wavefront will produce a converging wavefront in its conjugate, focusing toward a **point to the right** of the hologram. This focus represents a *real* image⁴, focused in space and visible on a white card if it is held in the right place. In the early history of Gabor-style in-line holography, this real image caused considerable corruption of the desired, true, or *virtual* image⁵, the one corresponding the

$m=+1$ term. The introduction of off-axis reference and illumination beams by Leith and Upatnieks caused the output angle of the conjugate wave to be significantly different from that of the desired wave. If the object wave were an on-axis plane wave, $\phi(x,y)=0$, the output angle of that term would be

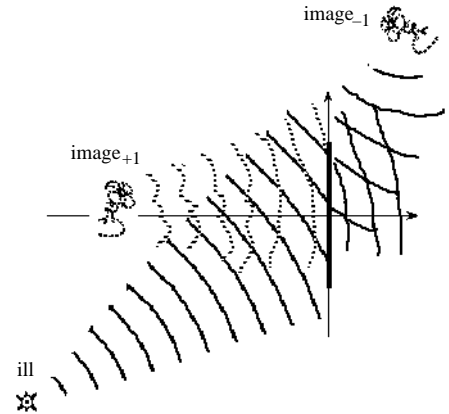
$$\theta_{\text{out},-1} = \sin^{-1}(2 \sin \theta_{\text{ref}}) . \quad (15)$$

note that for reference beam angles of 30° and above, this term is evanescent, and doesn't propagate at all).

The other “extra baggage” is the zero-order component, which has two terms. The first is simply an attenuated version of the illumination beam, headed in the same direction that the illumination was headed before the hologram was placed in the beam. Any energy left in this beam is not available for the desired reconstruction beam, so some effort usually goes into minimizing the zero-order beam to make bright holographic images.

The other zero-order term is more subtle, and deserves a description of its own. We usually call it the “halo” component. This beam is diffracted by the $a_{\text{obj}}^2(x,y)$ term, which is the same transmittance term that would be produced by exposing the hologram to the object alone, without the reference beam. That speckled exposure pattern contains grating patterns caused by interference between all possible pairs of points in the object, and the finest pattern (highest spatial frequency) will be produced by those object points that are the farthest apart. Let's say that these points subtend an angle ω as seen from the hologram plane. Assuming that ω is fairly small, that grating will have a spatial frequency of $\rho = \sin \omega / \lambda$. Including that grating in the hologram means that this modest spatial frequency will diffract the illumination beam over modest angles, roughly equal to ω on either side of the central direction of the beam. Even if the reference beam angle is large enough to allow the illumination beam to clear the desired image beam without overlapping it, the “halo” terms can scatter image-degrading light into that beam. Thus we will have to pay some attention to this component!

The analogies between diffraction by a hologram and diffraction by a simple grating should be becoming clearer to you. Interference of the light from the object with the reference beam creates a grating that is a generalized diffraction grating—that is, it has some variation or modulation of the contrast and location of its fringes (corresponding to amplitude and phase modulation of radio waves). When an illuminating plane wave is scattered by such a grating, it breaks up into the three components we normally see from simple gratings, except that each now has some trace of the object information impressed upon it. The $m=+1$ and $m=-1$ waves correspond to the same orders we observe with diffraction gratings, and most of our analysis will build on these similarities. The third component includes the zero-order and halo terms.



Arbitrary Wavefronts

This analysis can readily be extended to include reference and illumination beams of any wavefront shape—we only require that their amplitudes be reasonably uniform across the area of the hologram (if they are not, then the amplitude of the output wave will be modulated by the product of the two variations, which will generally degrade its image). The phase of the various reconstructed components can then be shown to be given by

$$\phi_{\text{out},m}(x,y) = m(\phi_{\text{obj}}(x,y) - \phi_{\text{ref}}(x,y)) + \phi_{\text{ill}}(x,y) . \quad (16)$$

Thus, whenever the wavefront of the illumination is identical to that of the reference beam, the phase-footprint of the object beam will be reconstructed. If the wavelength of the reconstruction is the same as that of the recording, then the physical properties of the image corresponding to that phase-footprint will be the same as those of the recorded object. This is perhaps the most general formulation of the holographic principle, one that we will use occasionally for fairly high-level proofs; some people have even called it the “Heisenberg’s Equation of Holography!”

$$\phi_{\text{out},m} = m(\phi_{\text{obj}} - \phi_{\text{ref}}) + \phi_{\text{ill}}$$

Diffraction Efficiency

Although we won't worry about just how bright our holograms are (or ought to be) for a while, we can already come to some conclusions about the diffraction efficiency of the Platonic holograms we have just described. Note that the ratio of the intensity of the $m=+1$ output beam to the intensity of the illumination beam is given by the ratio of their averaged-squared amplitudes. We define this ratio to be the *diffraction efficiency*, and note that for large K :

$$DE_{+1} = \eta_{+1} = \left(\beta \frac{\sqrt{K}}{K+1} \right)^2 \left(\left\langle a_{\text{obj}}^2(x,y) \right\rangle_{\text{spatial average}} = 1 \right) \quad (17)$$
$$\Rightarrow \frac{1}{K} \beta^2 \quad \text{for large } K.$$

Thus the fraction of the illumination energy that finds its way into the desired image beam decreases as the beam ratio increases, and depends critically on the slope of the t_{amp} - $\ln EXP$ curve, which we have dubbed the β of the material, which is similar to its "contrast."

Reconstruction Ratio

Another way of thinking about the diffraction efficiency for diffuse objects, and a handy way of gauging it in practice, is to illuminate a processed hologram with the reference beam that originally exposed it (or a replica of it); that is, a beam that has a uniform intensity of K . The diffracted intensity is then divided by the intensity of the original object beam (unity in our case) to yield the ratio of the *luminance* of the image (roughly its brightness) to the luminance of the object, which we call the "reconstruction ratio," denoted by RR . Substituting into the above equation gives

$$RR = \beta^2. \quad (18)$$

All of the absolute and relative beam intensities cancel out and we can aspire, with good reason, to make holographic images that are actually brighter than the objects that created them! It is only a matter of properly chemically processing the material to give a $|\beta| > 1$, and making sure that the holographic setup is tied down tightly enough so that the recorded fringes are as contrasty as they are supposed to be!

Conclusions

A generalized analysis can be very satisfying, and reassuring to our need to know that we haven't just stumbled across some special case or circumstance. But idealized analyses are often useless for solving practical problems! For instance, Eq. 12 tells us nothing about what happens to the $m=+1$ or "true" image if the illumination beam is misaligned a little, or the wavelength isn't quite right, or its radius of curvature is not correct. Those answers are implicit in that equation, of course, but we need a more directly physically-based approach to build up the sense of physical reasonability that will allow us to understand our experimental results, and to predict the likely outcome of proposed new experiments. Thus, we will abandon this domain of modest theoretical luxury, and descend into the dark and greasy pit of slippery approximations and hasty assumptions, with these more precise results safe in our pocket lest we should lose our way.

References:

1. There is some debate as to whether these should be called the *subject* and the *subject beam* instead; we will adopt the more common Leith & Upatnieks convention of *object* and *object beam*.
2. Most applications of photographic film concentrate instead on the *intensity transmittance* of the layer, the ratio of the irradiances just after and just in front of the film, or the "photographic density," which is the negative base-10 logarithm of the intensity transmittance (typically varying between zero and three).
3. conjugate (v. kon'jé gát'; adj., n. kon'jé git, -gát') v. -gated, -gating adj., n. verb transitive
 1. a. to recite or display all or some subsets of the inflected forms of (a verb) in a fixed order: to conjugate the present tense of the verb be.
 - b. to inflect (a verb).
 2. to join together, esp. in marriage.verb intransitive
 3. Biol. to unite; to undergo conjugation.
 4. (of a verb) to be characterized by conjugation.

adjective

5. joined together, esp. in a pair or pairs; coupled.
6. (of words) having a common derivation.
7. Math.
 - a. (of two points, lines, etc.) so related as to be interchangeable in the enunciation of certain properties.
 - b. (of two complex numbers) differing only in the sign of the imaginary part.
8. a. (of an acid and a base) related by the loss or gain of a proton: NH_3 is a base conjugate to NH_4^+ .

noun

9. one of a group of conjugate words.
 10. Math.
 - a. either of two conjugate points, lines, etc.
 - b. either of a pair of complex numbers of the type $a + bi$ and $a - bi$, where a and b are real numbers and i is imaginary.
- [1425-75; late ME (adj.) < LL *conjugatus*, ptp. of *conjugare* to unite (L: to join in marriage) = con- CON - + *jugare* to bind, der. of *jugum* YOKE]
- ⁴. real [1] (rê'él, rêl) adj.
 9. noting an optical image formed by the actual convergence of rays, as the image produced in a camera (opposed to virtual).
 - ⁵. virtual (vûr'ch'Ω él) adj.
 2. a. noting an optical image formed by the apparent convergence of rays geometrically, but not actually, prolonged, as the image formed by a mirror (opposed to real).
 - b. noting a focus of a system forming virtual images.