Surface Reflection I: Random Rough Surfaces, Microfacet Models

Lecture #4: Wednesday, 18 October 2000
Lecturer: Pat Hanrahan
Scribe: Tim Purcell
Reviewer: Adam Phelps

1 Introduction

The world is rich with visually interesting surfaces. Many of the more interesting visual properties of these surfaces tend to be caused by surface roughness. Waves, concrete, and plaster are among some of the surfaces whose appearance properties are related to surface roughness. Given a surface with a perfectly smooth interface, only perfect reflection and transmission are possible. Rough surfaces allow more complex surface interactions to occur, like self shadowing and multiple bounce reflections.

Several physical processes effect the roughness of a surface. Some examples include manufacturing processes such as grinding, etching, polishing and applications of different coatings. Environmental processes and ordinary wear and tear also contribute to surface roughness. Examples include atmospheric corrosion, patination, scratching, and the accumulation of dust.

Along with the physical processes which effect a surface, we need mathematical models to describe how light interacts with a rough surface. The V-groove model considers surfaces to be made up of symmetric grooves with some inclination. Another model considers the surface to be made up of tiny spherical pits - akin to craters on the moon. Surfaces can also be modeled as having cylindrical grooves or gauges in them. Yet another approach is to assume a surface is given by a random distribution of heights due to a random process. The random function generating the heights can have various statistical or frequency properties, like the periodicity of the function. Surfaces can be modeled as having their statistics vary in a Gaussian distribution, a fractal type distribution, or some other distribution.

2 Level of Detail

It is useful to think of surfaces at several different levels of detail. As shown in table 1, many common graphics techniques are really just ways of representing surfaces at different levels of detail.

LOD	Surface Parameterization	Example Representation
Macrostructure	$p(u,v) + h(u,v)n(u,v)$	Displacement Map (also called Offset Map, Relief Map)
Mesostructure	$p(u,v), n'(u,v)$	Bump Map, Normal Map
Microstructure	$p(u,v), n(u,v), D(n')$	Microfacet Distribution

Table 1: Levels of detail of surface representation. u and v represent the surface coordinates, $p(u,v)$ is a point on the surface, $h(u,v)$ is a displacement function, $n(u,v)$ is the surface normal, $n'(u,v)$ is a perturbed normal, and $D(n')$ is a microfacet distribution function.

This classification works at the viewer's level of detail. That is, the solid angle subtended by one photoreceptor intersects the surface in a small patch. The fine structure within that area is beyond the resolving power of the eye. As a result, that fine structure is averaged together and must bs represented in another way.

We have to keep in mind that physical phenomena have their own hierarchies of explanation that are independent of the viewer. At the most detailed level, quantum optics must be used to explain the interaction of light with chemical structures. The next level is the physical optics level, where the wave properties of light are important. And finally there is the level of geometric optics, where light is treated as traveling along rays.

If $h \sim \lambda$ (λ being the wavelength of light), then the interaction of light with the surface involves physical optics and resulting interference patterns formed from reflections at different points on the surface determine the properties of the reflection function. The Beckmann, He, and Stam models are based on physical optics. Some phenomena that need to be modeled using physical optics include reflection from rough metals and irridescence from biological structures (like a butterfly wing). If $|h| \gg \lambda$ then geometric optics will suffice. The Bouguer, Torrance-Sparrow, and Oren-Nayar models all operate in the realm of geometric optics. They produce good simulations of macroscopically rough surfaces like water waves and bricks.

Another consideration in the level of detail view of surfaces is texture. Some examples of texture playing into level of detail are glitter and illumination texture. Figure 1 shows an example of an illumination texture. The overall texture is correlated to surface roughness, and is a function of the angle of inclination of the light source. We might call this a directionally dependent illumination texture. One question for thought is: Can we compute a texture to represent this surface without brute force ray tracing?

Glitter is another example of texture that is caused by surface roughness. Water will often have bright speckles at points where the sun's rays are directly reflected toward the observer, as can be seen in figure 2. The effect of glitter is similar to that of material that has tiny reflectors embedded in it. Rocks, mica, and metallic paints are examples of


Figure 1: Local illumination vs. global illumination. The left image shows a wall illuminated by only considering local lighting effects. The right image shows the same wall, this time incorporating global visibility information.

surfaces that exhibit this property. Calculating glitter points is difficult when using ray tracing, since the probability of a ray hitting the surface at exactly those places where the surface is aligned to reflect the light is very low. However, these glitter points are very bright, and need to be considered to properly recreate the appearance of the surface. The glitter pattern does have some characteristic statistics however, and is another candidate for showing level of detail through texture rather than direct ray tracing.

3 Background Derivations

To help gain an intuition for microfacet reflection functions, we begin by analyzing reflection from a familiar surface - that of wavy water. The analysis of the "macrofacet" reflection scales to surfaces composed of "microfacets." Figures 3 and 4 are sketches of

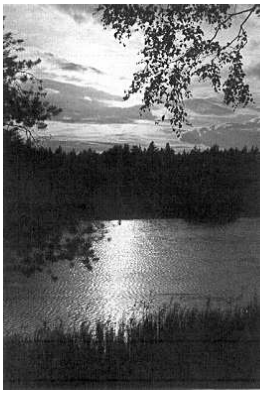

Figure 2: Water often has bright speckles at points where the sun's rays directly reflect toward the observer. This effect is called glitter.

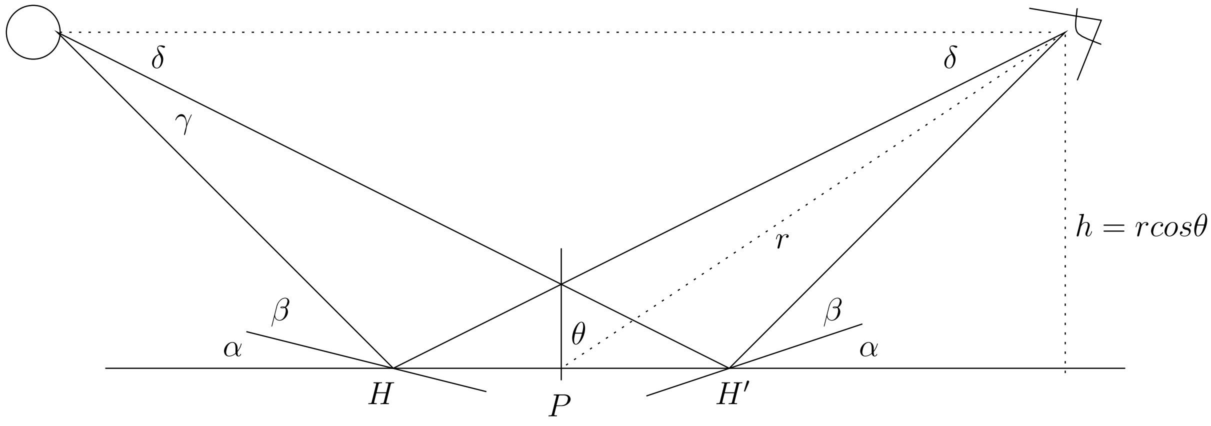

Figure 3: Side view of glitter geometry. We assume the viewer is very far away during observation.

the geometry for looking at waves and will help us determine why glitter is oblong in shape.

We take α to be the maximum inclination of waves. H is the point where waves are maximally inclined towards us and H' is the point where they are maximally inclined away from us. γ is the angle between H and H', or the angle between the minimum and maximum points of reflection from the waves. From the geometry in figure 3, we can derive the following set of equations:

$$\begin{aligned} \alpha + \beta &= \gamma + \delta \\ \beta - \alpha &= \delta \end{aligned}$$

If we substitute the bottom equation and solve for γ, we get

$$\gamma = 2\alpha$$

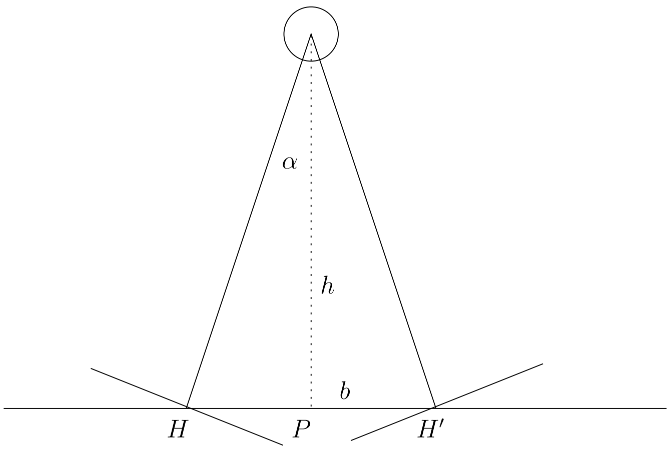

Figure 4: Front view of glitter geometry

We see that the reflection angle is doubled.

We now look at the geometry from the front view, shown in figure 4. We can now derive the following set of equations:

$$\begin{aligned} \tan\alpha &= \frac{b}{h} \\ h &= r\cos\theta \end{aligned}$$

We denote ψ as the angle subtended by two extremes of the waves.

$$\tan\psi = \frac{b}{r} = \tan\alpha\cos\theta$$

The key to this derivation is that $\tan\alpha$ is multiplied by $\cos\theta$. At grazing angles, $\cos\theta$ gets smaller and the angle ψ compresses.

Figure 5 shows another way of thinking about the patch of glitter that may be helpful in understanding why the glitter patch is long and narrow. In this figure, the average surface normal is at the north pole and we assume the surface waves have normals distributed in a cone with half-angle α. $\hat{L}$ represents the position of a point light source. The direction of reflection of the light source depends on the surface normal $\hat{N}$, and is denoted by $R_N(\hat{L})$. Each wave has a different surface normal, and hence reflects light in a different direction. The locus of all these directions is an oblong cone.

Consider now the plane of incidence given by the average surface normal $\hat{N}$ (the north pole) and the $\hat{L}$ direction. Other normal directions in this plane correspond to the near and far points analyzed in figure 4. The reflected directions must also be in this plane and the extrema are at angle γ with respect the ideal specular reflection.

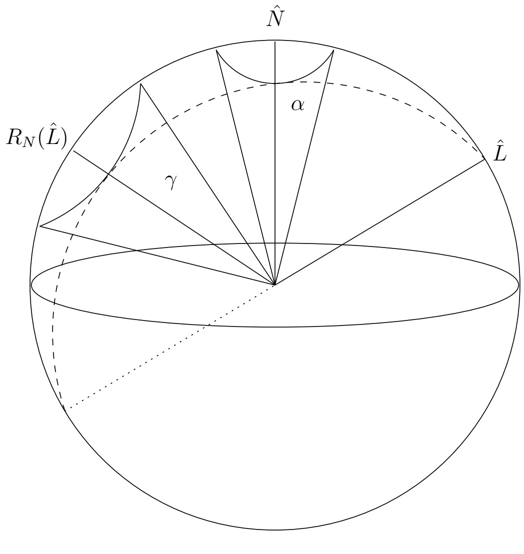

Figure 5: Spherical geometry. $\hat{L}$ is the incoming light direction, and $R_N(\hat{L})$ is the reflected light. The surface is made up of a cone of normals with half angle α.

The oblong nature of the reflected cone may be analyzed by considering planes through the light source and the extreme left and right points in the surface normal distribution. Note that the normal, light and reflection must all lie in a plane, each of these planes intersect the sphere in a great circle. The left and right extrema of the cone of normals define great circles; these great circles must also bound the oblong cone of reflection directions. That is, the oblong reflection must lie between two great circles, both through $\hat{L}$ and each through the left and right extrema. Now consider the case when the angle of incidence changes. At the angle goes towards grazing, the reflection moves towards the antipode of $\hat{L}$. However, the extrema great circles must go through the antipode of $\hat{L}$. As the great cicles approach the antipode they come together. Therefore, as the oblong cone moves towards the antipode, it is compressed. This makes the reflection narrower and narrower. This is like the sun setting and producing a narrow and long glitter patch on the ocean.

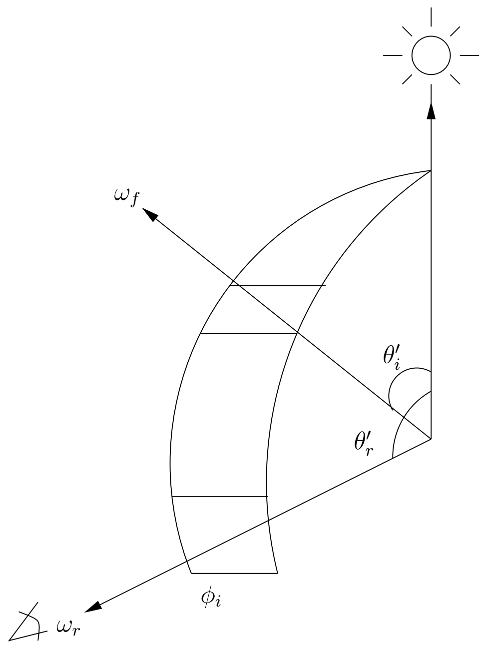

Figure 6: More spherical geometry. We use this to explain the magical factor of 4 found in distribution functions like Torrance-Sparrow.

A related result is the ratio of the solid angle of the reflection cone to the normal cone. We can compute this by considering the ratio of the differential solid angle of an infinitesimal piece of the reflected cone to the normal cone. Using figure 6, and our previous derivations, we know that

$$\begin{aligned} \theta_r' &= 2\theta_i' \\ \phi_r &= \phi_i \end{aligned}$$

Here we deonte ϕ as the angle of rotation about the normal. One might be tempted to think that a differential distribution of reflectors would reflect into equal size distributions of directions. But this is not the case (as we saw previously), so we need to find a correction factor. A simple derivation involving a change of variables will give us the correction factor:

$$\begin{aligned} d\omega_r &= \sin\theta'_r d\theta'_r d\phi_r \\ &= (\sin 2\theta'_i) 2 d\theta'_i d\phi_i \\ &= (2\sin\theta'_i \cos\theta'_i) 2 d\theta'_i d\phi_i \\ &= 4\cos\theta'_i d\omega_f \end{aligned}$$

$$\frac{d\omega_f}{d\omega_r} = \frac{1}{4\cos\theta'_i}$$

This change of variables comes up when we solve the reflection equation

$$\int L_i(\omega_i) f_r(\omega_i \rightarrow \omega_r) \cos\theta_i d\omega_i$$

If the reflection function is a microfacet distribution, we need to change variables from $f_r(\omega_i \rightarrow \omega_r)$ to $f(\omega_f)$, where $\omega_f = (\omega_i + \omega_r)/|\omega_i + \omega_r|$, the halfway vector, or the angle that the microfacet must be oriented at to reflect light from the incident direction to the reflected direction. This change requires changing from an integral over incoming directions $d\omega_i$ to one over microfacet normals $d\omega_f$. The resulting change of variables can be done using the result of the previous derivation.

4 Torrance-Sparrow Model

Bouguer invented the microfacet model with his paper about 'little faces' where he derives it using geometric arguments. In the paper by Torrance and Sparrow, they derive the microfacet model using modern notation and more familiar mathematical techniques.

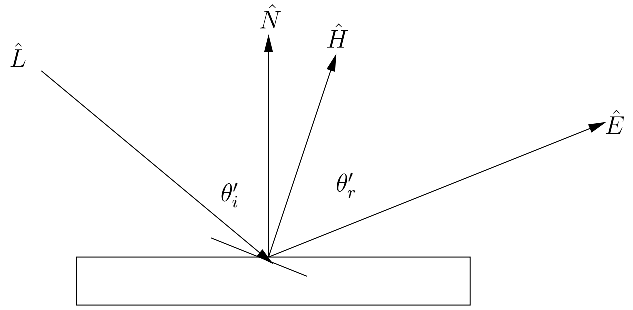

Figure 7: Generic view of a surface. We know $dA' = D(\omega_f) d\omega_f dA$. dA' is the area of a particular microfacet, and ω_f is essentially $\hat{H}$. We define $\cos\theta_i = \hat{L} \cdot \hat{N}$ as measuring angles with respect to the surface, while $\cos\theta'_i = \hat{L} \cdot \hat{H}$ is with respect to the microfacet orientation.

The Torrance-Sparrow model assumes all facets on the surface are mirrors. We start by finding the power falling on the facet

$$d^2\Phi_f \;=\; L_i(\omega_i)\cos\theta_i' d\omega_i dA'$$

Here θ_i' is the angle of incidence *with respect to the microfacet*. dA' is the area of the microfacet oriented in a given direction. This area is given by the microfacet distribution

$$dA' = D(\omega_f) d\omega_f dA$$

So, in terms of the microfacet distribution, the power falling on a given microfacet equals

$$d^2\Phi_f = L_i(\omega_i)\cos\theta_i' d\omega_i D(\omega_f) d\omega_f dA$$

Since the facets are mirrors, the reflected power equals the incident power

$$d^2\Phi_r \;=\; d^2\Phi_f$$

The total power reflected from the surface is

$$d^2\Phi_r \;=\; dL_r(\omega_i \to \omega_r)\cos\theta_r d\omega_r dA$$

Now we just substitute

$$dL_r(\omega_i \to \omega_r)\cos\theta_r d\omega_r dA = L_i(\omega_i)\cos\theta_i' d\omega_i D(\omega_f) d\omega_f dA$$

And we derive the BRDF by remembering it is equal to the reflected radiance divided by the irradiance:

$$\begin{aligned}
f_r(\omega_i \to \omega_r) &= \frac{dL_r(\omega_i \to \omega_r)}{dE_i(\omega_i)} \\
&= \frac{L_i(\omega_i)\cos\theta_i' d\omega_i D(\omega_f) d\omega_f dA}{(\cos\theta_r d\omega_r dA)(L_i(\omega_i)\cos\theta_i d\omega_i)} \\
&= \frac{D(\omega_f)}{\cos\theta_i\cos\theta_r}\cos\theta_i' \frac{d\omega_f}{d\omega_r} \\
&= \frac{D(\omega_f)}{4\cos\theta_i\cos\theta_r}
\end{aligned}$$

Here we used the ratio of the differential facet solid angle to the differential reflected solid angle derived in the last section.

As light moves to grazing angles, $\cos\theta_r$ approaches zero, and the BRDF blows up ($\cos\theta_i$ cancels out when you integrate with incoming irradiance, so it is not a problem). The intuition is that more and more mirror facets are visible at grazing angles. This is the big theoretical problem with microfacet models. On a real surface, many mirrors would be hidden at grazing angles. The BRDF is not correcting for this, and for this reason, many people discount the microfacet model. It can be made to work if the self shadowing correction term divided by $\cos\theta_r$ behaves well. We will discuss this further in the next lecture.

Remember that experiments measuring the reflection of light from different materials often show off-specular highlights. That is the direction of maximum reflection is greater than the mirror direction. For dielectrics, this phenomemon may be explained by appealing to the Fresnel formula, since more light is reflected at grazing angles. For metals, this explanation does not work, since the modified Fresnel equations predict that the reflection is relatively independent of angle of incidence. To account for the off-specular peak for metals, requires a careful consideration of both the $1/\cos\theta_r$ blowup and a geometric attenuation term accounting for shadowing of microfacets. This is the main contribution of the original Torrance and Sparrow paper.

5 Oren-Nayar Model

The Oren-Nayar model differs from the Torrance-Sparrow model by making the facets matte surfaces instead of mirrors. Most of the derivation flows in the same manner. We first find the power incident on the surface:

$$d^2\Phi_f = L_i(\omega_i)\cos\theta_i' d\omega_i D(\omega_f) d\omega_f dA$$

Since the surface is diffuse, the reflected radiance is equal to ρ/π times the incident irradiance. Thus, the reflected power is

$$d^2\Phi_r = \frac{\rho}{\pi} L_i(\omega_i)\cos\theta_i' d\omega_i \cos\theta_r' d\omega_r D(\omega_f) d\omega_f dA$$

Then we find the BRDF the same way we did for the Torrance-Sparrow model:

$$\begin{aligned} f_r(\omega_i \to \omega_r) &= \frac{\frac{\rho}{\pi} L_i(\omega_i)\cos\theta_i' d\omega_i \cos\theta_r' d\omega_r D(\omega_f) d\omega_f dA}{(\cos\theta_r d\omega_r dA)(L_i(\omega_i)\cos\theta_i d\omega_i)} \\ &= \frac{\rho}{\pi}\frac{\cos\theta_i' \cos\theta_r'}{\cos\theta_i \cos\theta_r} D(\omega_f) d\omega_f \end{aligned}$$

It is possible to get the agregate reflection for a single light over all incoming directions by integrating over the microfacet distribution rather than the incoming light. The result is a hairy integral and you have to be careful to think about non-contributing faces.

If the base surface has both specular and diffuse reflection, then it is possible to derive a reflection model for a rough surface by combining the Torrance-Sparrow and Oren-Nayar models. Correctly combining them is difficult, but the combination may be approximated by a linear sum of the two contributions. This is called the TSON model.

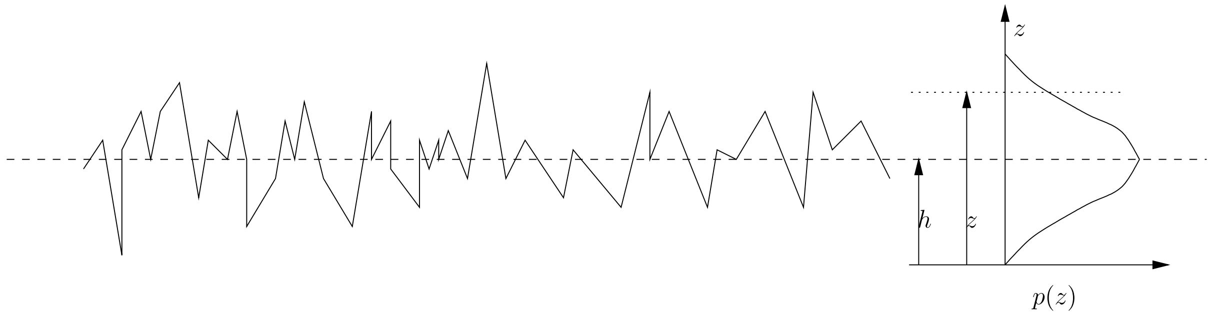

Figure 8: An example of a Gaussian surface.

6 Gaussian Rough Surfaces

Graphicists usually model surface reflection with $c(N \cdot H)^s$ which is a phenomenological model. They usually don't consider self shadowing or mutual illumination, and ignore the various cosines, etc. Gaussian surfaces are a model derived by physicists and may be adapted by graphicists. We say a rough surface obeys Gaussian statistics if the probability distribution of heights is a Gaussian distribution. Figure 8 is an example of what makes a Gaussian surface.

Why Gaussian? Basically the law of large numbers. If lots of small random processes change the height of the surface, then the resulting distribution of heights will tend towards a Gaussian. Think about millions of meteors pounding the moon. The crater sizes and heights might tend towards a gaussian distribution. Of course, this is a hand waving argument. Certain physical processes may change the distribution in a more irregular way. For example, if a surface is rubbed or rolled, then the highest parts of the surface might be removed or flattened. This would lead to a distribution that has a lower probability of high points. Conversely, dirt or dust may accumulate at the low points. This would lead to a distribution with lower probability of low points.

6.1 Definitions

6.1.1 Height Distribution

$P_z(z)$ is height probability.

$$
\begin{aligned}
P_z(z)dz &= \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{z^2}{2\sigma^2}\right)} dz \\
Pr(z < h) &= \int_{-\infty}^{h} P_z(z)dz \\
erf(t) &= \frac{1}{\sqrt{2\pi}} \int_0^t e^{\left(-\frac{s^2}{2}\right)} ds \\
erfc(t) &= 1 - erf(t)
\end{aligned}
$$

6.1.2 Autocorrelation Function

The autocorrelation function gives the distance between which samples become independent. In general, samples close together are correlated, and samples further apart are independent.

$$R(t) = \int_0^\infty z(x)z(x+t)dt$$

The autocorrelation function may be computed directly given the heights on the surface. It is the Fourier transform of the spectral power density, which may also be computed from the heights. Finally, it is often convenient to assume a given form of the autocorrelation function. For example, we could just assume a Gaussian.

$$R(t) = \sigma^2 e^{\left(-\frac{t^2}{T^2}\right)}$$

We can compute the standard deviation of a Gaussian surface from its autocorrelation function. If we apply the ergodic theorem, we can change the integral over the line to an integral over a distribution of heights

$$R(0) = \int z^2 dt = \int z^2 p(z) dz = \sigma^2$$

Of course, there are other possible autocorrelation functions. Example include exponential and power laws (fractal surfaces).

6.1.3 Spectral Power Density

The spectral power density, call it $G(\omega)$, for a height distribution can be computed as

$$\begin{aligned} Z(\omega) &= \int z(t)e^{i\omega t}dt \\ G(\omega) &= |Z(\omega)|^2 \end{aligned}$$

As mentioned in the previous section, the autocorrelation function can be derived from the Fourier transform of the spectral power density

$$R(t) = \int_0^\infty G(\omega)\cos\omega t d\omega$$

This means for distributions where we only know the power spectrum - for example a fractal distribution where the power spectrum has the form $1/\omega^2$ - we can determine the autocorrelation function.

We can also find the spectral power density by taking the Fourier transform of the autocorrelation function

$$G(\omega) = \frac{2}{\pi}\int_0^\infty R(t)\cos\omega t dt$$

6.1.4 Slope Distribution

An obvious question is whether it is possible to derive the slope distribution of a random surface from its height distribution. For Gaussian surfaces this is possible. The derivation is based on a mathematical theorem that states that if we apply a linear operator to a signal that has a Gaussian distibution then the resulting signal also has a Gaussian distribution. Although I won't prove this, let me state that it is plausible. If we form linear combinations of events, given a large enough set of events, the resulting distribution will tend to be Gaussian. Forming an additional linear combinations of events that are already Gaussian, would surely seem to maintain the property of being a Gaussian.

Examples of linear operators are derivatives and integrals, since they can be implemented as filters.

$$
\begin{aligned}
z' &= \frac{\partial z}{\partial x} \\
P(z')dz' &= \frac{1}{\sqrt{2\pi r}} e^{\left(-\frac{z'^2}{2r^2}\right)} dz'
\end{aligned}
$$

Here r is the variance of the slope distribution.

We can relate r to the autocorrelation function of the surface.

$$
\begin{aligned}
R(t) &= \int z(x)z(x+t)dx \\
R' &= \int z(x)\dot{z}(x+t)dx = \int z(x-t)\dot{z}(x)dx \\
R'' &= \frac{d}{dt}\int z(x)\dot{z}(x+t)dx = \frac{d}{dt}\int z(x-t)\dot{z}(x)dx = -\int \dot{z}(x-t)\dot{z}dx \\
&= -\int \dot{z}(x)\dot{z}(x+t)dx
\end{aligned}
$$

Therefore,

$$
\left\langle \dot{z}^2 \right\rangle = r^2 = -R''(0)
$$

If we choose a concrete autocorrelation function, say a Gaussian like that shown above

$$
\begin{aligned}
R(t) &= \sigma^2 e^{\left(-\frac{t^2}{T^2}\right)} \\
R'(t) &= -\frac{2\sigma^2}{T^2} t e^{\left(-\frac{t^2}{T^2}\right)} \\
R''(t) &= -\frac{2\sigma^2}{T^2} e^{\left(-\frac{t^2}{T^2}\right)} + t^2(\ldots) \\
-R''(0) &= \frac{2\sigma^2}{T^2} = r^2
\end{aligned}
$$

This has a simple geometric interpretation. If σ is the height of the surface fluctations and T is the distance over which these fluctuations occur, then r is the mean slope between low and high points on the surface.

6.2 2D Distributions

Now we want to transition from 1D to 2D distributions. We introduce the derivatives

$$z_x = \frac{\partial z}{\partial x}, z_y = \frac{\partial z}{\partial y}$$

6.2.1 Isotropic Slope Distributions

For isotropic surfaces, treat the x and y distributions the same. Furthermore, we assume that the two distributions are uncorrelated. The the 2D distribution is just the product of the two 1D distributions

$$P(z_x)P(z_y)dz_x dz_y = \frac{1}{2\pi r^2} e^{\left(-\frac{z_x^2+z_y^2}{2r^2}\right)} dz_x dz_y$$

We can rewrite this if we reparameterize with polar coordinates

$$z_m = \sqrt{z_x^2 + z_y^2}, \quad \phi = \arctan 2^{-1}(z_x, z_y)$$

Which gives

$$P(z_m)dz_m d\phi = \frac{1}{2\pi r^2} z_m e^{\left(-\frac{z_m^2}{2r^2}\right)} dz_m d\phi$$

Note that distribution of slopes for an isotropic surface is no longer Gaussian. This seems to be a source of inconsistency in many papers.

We can also parameterize the equation in terms of angles

$$\begin{aligned} z_m &= \tan\theta \\ P(\theta,\phi)d\theta d\phi &= \frac{1}{2\pi r^2}\frac{\tan\theta}{\cos^2\theta} e^{\left(-\frac{\tan\theta}{2r^2}\right)} \sin\theta d\theta d\phi \end{aligned}$$

6.2.2 Anisotropic Slope Distribution - Ward Model

General anisotropic functions come from a kernel that looks like a matrix multiply. The equations start to look like differential geometry on a surface. If you find the directions of minimum and maximum curvature on the surface (which are perpendicular), and convert into $\tan\theta$ space

$$\begin{aligned} P(z_x, z_y) &\approx e^{\left(-\begin{bmatrix} z_x & z_y \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} z_x \\ z_y \end{bmatrix} \right)} \\ &= e^{\left(-\tan^2\theta\left[\frac{\cos^2\phi}{rx^2}+\frac{\sin^2\phi}{ry^2}\right]\right)} \end{aligned}$$

We arrive at the formula proposed by Ward for anisotropic rough surfaces.