Sampling Visible GGX Normals with Spherical Caps

Jonathan Dupuy and Anis Benyoub
Intel Corporation

1024 spp, render time: 149.0s
Ours (+7.58% speedup)

1024 spp, render time: 138.5s

4096 spp, render time: 1039.1s
Ours (+3.28% speedup)

Figure 1: We introduce a novel importance-sampling algorithm for GGX microfacet BSDFs. Our algorithm offers the same variance reduction as the state-of-the-art method [Hei18] but with lower computational overhead. This translates into slight but systematic performance gains for the rendering of, e.g., (left) rough conductors and (right) rough dielectrics. Scene credit: Yasutoshi Mori for the PBRT-v4 renderer.

Abstract
Importance sampling the distribution of visible GGX normals requires sampling those of a hemisphere. In this work, we introduce a novel method for sampling such visible normals. Our method builds upon the insight that a hemispherical mirror reflects parallel light rays uniformly within a solid angle shaped as a spherical cap. This spherical cap has the same apex as the hemispherical mirror, and its aperture given by the angle formed by the orientation of that apex and the direction of incident light rays. Based on this insight, we sample GGX visible normals as halfway vectors between a given incident direction and directions drawn from its associated spherical cap. Our resulting implementation is even simpler than that of Heitz and leads to systematic speed-ups in our benchmarks.

CCS Concepts
• Computing methodologies → Reflectance modeling; Ray tracing;

1. Introduction
The GGX microfacet BRDF is the most successful reflectance model in modern computer graphics history: Over the last decade, it has been implemented in many realtime game engines including (but not restricted to) Frostbite [HMD+14], Unreal [MHM+13], and Unity [AMG+18], as well as in offline path-tracers such as those of Walt Disney Animation Studios [MHH+12], Pixar [MHM+13], DreamWorks Animation [HMC+17], and Sony Pictures Imageworks [HMC+17].

In this work, we introduce a novel importance sampling algorithm for the GGX microfacet BRDF. Our algorithm offers the same variance reduction as state-of-the-art methods albeit at faster execution times as shown in Figure 1. Furthermore, it is straightforward to integrate within an existing implementation. It can thus immediately benefit to a large number of physically based renderers including those just mentioned.

Our algorithm effectively offers an alternative way to sample the GGX distribution of visible normals (VNDF) as described by Heitz [Hei18]—this particular method lies foundation for our own algorithm and so we explain it in detail in Section 2. We arrived at our alternative method by linking visible GGX normals to directions enclosed within solid angles shaped as spherical caps. We describe this link and our contributions throughout the remainder of this article as follows:

• In Section 3, we link the GGX VNDF to spherical caps and leverage this link to derive our novel importance sampling algorithm.
• In Section 4, we validate and benchmark our algorithm against that of Heitz both on the CPU and the GPU.
2. Background on Sampling GGX Visible Normals

In this section, we provide the preliminary background suitable for deriving a state-of-the-art GGX VNDF sampler. Note that we refrain from providing more general background on the GGX microfacet BRDF here as this is largely unnecessary for our derivations. We refer the interested reader to Appendix A for such supplementary information.

Definition The GGX VNDF refers to the distribution of normals that lie at the surface formed by the intersection of parallel rays and a truncated ellipsoid [Hei18]. This intersected surface is shaded in green in Figure 2 and varies according to the direction of the parallel rays as well as the shape of the ellipsoid. More formally, the VNDF is parameterized by an incident direction \( \omega_i \) that, by convention, points towards the opposite direction of the incident rays, as well as two scaling parameters \( \alpha_x > 0 \) and \( \alpha_y > 0 \) such that the matrices

\[
M = \begin{bmatrix}
\frac{1}{\alpha_y} & 0 & 0 \\
0 & \frac{1}{\alpha_x} & 0 \\
0 & 0 & 1 \\
\end{bmatrix}, \quad \text{and} \quad M^{-1} = \begin{bmatrix}
\alpha_x & 0 & 0 \\
0 & \alpha_y & 0 \\
0 & 0 & 1 \\
\end{bmatrix},
\]

respectively map a unit hemisphere to the ellipsoid and vice versa. Note that this parameterization implicitly assumes that the ellipsoid is formed by stretching the unit hemisphere centered at the origin and lying on the \( z = 0 \) plane. Next, we describe how to sample this distribution.

Importance Sampling Sampling the GGX VNDF according to the method of Heitz [Hei18] refers to mapping two uniform random numbers \( \theta_1, \theta_2 \in [0, 1) \) to a normal \( \omega_{\text{dir}} \). Intuitively, this mapping selects a ray that intersects the truncated ellipsoid and returns the normal lying at the intersection point. In practise, the GGX VNDF sampler leverages two key components. First, an invariance to linear transformations, which allows to systematically remap the intersection configuration to that of a hemisphere. Second, a routine to sample the visible normals of a hemisphere. The sampling algorithm then consists of the three following steps, which are illustrated in Figure 2 (see also Listing 1 for a GLSL implementation):

1. Warp the entire space using the inverse matrix \( M^{-1} \). This maps the ellipsoid to the unit hemisphere and the incident direction as follows

\[
\omega_i \mapsto M^{-1} \omega_i.
\]

2. Intersect the unit hemisphere based on the new incident direction and record the position of the intersection points. The normal associated to each intersection point is equal to its position. In order to sample such intersection points, Heitz derives an algorithm based on the parameterization of the cross section of the hemisphere. We illustrate this algorithm in Figure 3 and provide a GLSL implementation in Listing 2. Note that this particular algorithm is the one we improve in this work.

3. Transform each intersection point back to the ellipsoid configuration using the matrix \( M \). The normals must be warped by the inverse transpose \( M^{-T} \) and effectively sample the GGX VNDF.

Figure 2: GGX VNDF sampling overview.

Figure 3: Sampling the visible hemisphere using its cross section.
### 3. Sampling GGX Visible Normals with Spherical Caps

In this section, we introduce our alternative method to sample the GGX VNDF. Our method differs from that of Heitz only in the way we sample the visible normals of the unit hemisphere. Our resulting GLSL implementation is shown in Listing 3. As can be seen from a direct comparison with the implementation of Heitz in Listing 2, our code is considerably simpler. We emphasize that this simplification is not due to aggressive optimizations of the original code of Heitz but to a novel insight that leads to an entirely different implementation. We share this insight and detail the implementation it leads to in the remainder of this section.

#### 3.1. Key Insight: The Specular Reflections of a Hemisphere are Distributed as Spherical Caps

We make the key observation that if the hemisphere used for sampling the GGX VNDF acted as a perfect mirror, it would reflect parallel rays uniformly within a spherical cap. We illustrate this property in Figure 4 for two specific direction of incidence—notice how the direction of incidence only affects the height of the cap that shapes the directional distribution of reflected rays. Based on this insight, we devise our sampling algorithm using this spherical cap rather than the cross section of the hemisphere as done by Heitz. Before diving into the details of our algorithm, we first formally prove that our insight is mathematically sound. This is the focus of the next subsection.

#### 3.2. Mathematical Proof

Let us consider that the hemisphere used for sampling the GGX VNDF acts as a perfect mirror, i.e., it deviates incident ray directions \( \omega_i \) towards outgoing ray directions \( \omega_o \) according to the law of specular reflection

\[
\omega_o = 2(\omega_i - \omega_n)\omega_n - \omega_i. \tag{5}
\]

In order to derive the distribution of these outgoing ray directions, we first need to invert Equation (5) to determine the location at which the hemisphere deflects \( \omega_i \) towards \( \omega_o \). This location corresponds to the point on the hemisphere whose normal \( \omega_h \) is aligned with the half-vector

\[
\omega_h = \frac{\omega_i + \omega_n}{\|\omega_i + \omega_n\|}. \tag{6}
\]
The directional distribution of reflected rays at θ = 70° corresponds to specular reflection of parallel rays. The directional distribution of reflected rays at θ = 130° also represents specular reflection of parallel rays.

**Figure 4:** Main insight: a hemispherical mirror reflects parallel light rays towards directions enclosed within a spherical cap.

Thanks to Equation (6), we can relate the density $f_p$ of outgoing ray directions oriented within an infinitesimal solid angle $d\omega_o$ to the probability that the incident rays will intersect the infinitesimal surface $d\omega_h$:

$$f_p(\omega_o, \omega_i) d\omega_o = D_{vis, std}(\omega_h, \omega_i) d\omega_h.$$  

Consequently, we get

$$f_p(\omega_o, \omega_i) = D_{vis, std}(\omega_h, \omega_i) \frac{d\omega_h}{d\omega_o}. \quad (7)$$

Now, to finalize our proof, we show that Equation (7) reduces to a uniform density over a domain bounded by a spherical cap. We first use the fact that the reflection Jacobian satisfies [WMLT07]

$$\frac{d\omega_h}{d\omega_o} = \frac{1}{4(\omega_o \cdot \omega_h)}. \quad (8)$$

Then, plugging Equations (2) and (8) into Equation (7) and using the fact that, by definition, $(\omega_o \cdot \omega_h) = (\omega_i \cdot \omega_h) \geq 0$, we get

$$f_p(\omega_o, \omega_i) = \frac{D_{vis}(\omega_h)}{4\sigma_{std}(\omega_i)}. \quad (9)$$

Finally, using Equations (3) and (4) we obtain the uniform density

$$f_p(\omega_o, \omega_i) = \begin{cases} \frac{1}{2\pi(1+z_i)} & \text{if } z_h > 0 \Rightarrow z_o > -z_i \\ 0 & \text{otherwise.} \end{cases} \quad (10)$$

Such a uniform density is enclosed within a spherical cap oriented upwards and cutoff at the $z = -z_i$ plane, thus concluding our proof. As a final note, we mention that the normalization constant in Equation (10) corresponds to the solid angle of the cap that bounds the distribution.

**3.3. Sampling Algorithm**

A corollary result of our insight is that the half-vectors we retrieve from a spherical cap cutoff at the $z = -z_i$ plane are necessarily distributed according to the normals of the hemisphere that are visible from direction $\omega_i$. This corollary result lies at the core of our sampling algorithm, which is illustrated in Figure 5 and consists of the following steps:

1. Sample a direction $\omega_o$ from a spherical cap with elevation $-z_i$.
2. Compute the half-vector between the sample $\omega_o$ and $\omega_i$.
3. Return this half-vector as a visible normal.

We provide a GLSL implementation of our algorithm in Listing 3, which is meant to be used in conjunction with Listing 1. An important note here is that we purposely omit the normalization of the half-vector at line (14) of Listing 3 (this is equivalent to avoiding step (2.2) in Figure 5). We do this because we know that the direction is normalized later, specifically at line (8) of Listing 1. If one was to use Listing 3 in a stand-alone fashion, then the normalization would be necessary.
4. Results and Validation

In this section, we validate and benchmark our method against that of Heitz through several experiments that we detail below.

CPU Synthetic Benchmark We wrote a C++ program that calls Listing 1 within a loop and measured the median time it takes to run over 100 invocations. We ran our program on an Intel i7-13700K and measured a speed-up of 37.67% in favor of our method. This result is consistent across roughness values and incident directions.

GPU Synthetic Benchmark We wrote a DirectX 12 compute shader that calls Listing 1 64 times per lane and measured the median time taken for the shader to run over 100 invocations. We ran our program on an Intel Arc A770 and an NVIDIA RTX 2080 and measured a speed-up of respectively 52.96% and 39.25% in favor of our method. Again, this result is consistent across roughness values and incident directions.

Rendering Comparisons We validated that our method converges to the same results as the method of Heitz within state-of-the-art path-tracers. We provide an example of such a validation in Figure 1, where we show PBRT-v4 renderings of a scene based on two different GGX microfacet BSDFs. As expected, our sampling algorithm does indeed converge towards the same result as Heitz.

Rendering Profiling We also profiled the execution of a PBRT rendering for scenes of varying geometric complexity. For each scene, we set all materials to either rough conductors or rough dielectric. Under this setup we report the relative time spent for BSDF importance sampling during a PBRT rendering. The scenes are available at the PBRT repository: https://github.com/mmp/pbrt-v4-scenes.

Table 1: Relative time spent for BSDF importance sampling during a PBRT rendering. The scenes are available at the PBRT repository: https://github.com/mmp/pbrt-v4-scenes.

<table>
<thead>
<tr>
<th>scene name</th>
<th>relative time [Hei18]</th>
<th>ours</th>
<th>speed-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>pbrt-book</td>
<td>1.34%</td>
<td>0.54%</td>
<td>×2.48</td>
</tr>
<tr>
<td>lte-orb (conductor)</td>
<td>3.71%</td>
<td>1.49%</td>
<td>×2.49</td>
</tr>
<tr>
<td>lte-orb (dielectric)</td>
<td>1.89%</td>
<td>0.74%</td>
<td>×2.55</td>
</tr>
<tr>
<td>sportscar</td>
<td>0.48%</td>
<td>0.20%</td>
<td>×2.40</td>
</tr>
</tbody>
</table>

5. Conclusion

We introduced a novel importance sampling algorithm for GGX microfacet BSDFs. Our method is easy to implement and offers systematic speed-ups over that of Heitz.

Appendix A: Background on the GGX Microfacet BRDF

The GGX microfacet BRDF refers to an analytic bidirectional reflectance distribution function (BRDF), which results from microfacet theory. We provide here some background and the mathematical expressions suitable for evaluating and importance sampling such a BRDF via the GGX VNDF introduced in Section 2.

Derived from the GGX BRDF The GGX microfacet BRDF is a specialized form of Equation (11), which follows from the two following hypotheses. The first hypothesis is that the microfacets are oriented according to the NDF of a truncated ellipsoid. In the spirit of Section 2, we write this NDF as a function of that of a hemisphere. This leads to the following expression

\[
D(\hat{\omega}_m) = \frac{\det M^T}{\|M^T \hat{\omega}_m\|^2 \|M^T \hat{\omega}_m\|^2} \left( \text{hemisphere NDF see Equation (3)} \right) \text{ Jacobian} \left( \frac{M^T \hat{\omega}_m}{\|M^T \hat{\omega}_m\|} \right) \left( \frac{\|M^T \hat{\omega}_m\|^2}{\text{Jacobian}} \right) \left( \frac{M^T \hat{\omega}_m}{\|M^T \hat{\omega}_m\|} \right) \left( \frac{\|M^T \hat{\omega}_m\|^2}{\text{Jacobian}} \right),
\]

where \(M\) is the matrix that maps the hemisphere to the GGX ellipsoid as defined in Section 2. The second hypothesis is that the microfacets are subject to the Smith shadowing model. We provide an example of such a validation in Figure 1, where we show PBRT-v4 renderings of a scene based on two different GGX microfacet BSDFs. As expected, our sampling algorithm does indeed converge towards the same result as Heitz.

We validated that our method converges systematically lower than 4% for all the scenes we tested. We introduced a novel importance sampling algorithm for GGX microfacet BSDFs. Our method is easy to implement and offers systematic speed-ups over that of Heitz.

Microfacet Theory Intuitively, microfacet theory describes surface reflectance as the result of interactions between incident light and microscopic mirrors (which compose the actual surface) that deviate this incident light. If we only consider single-scattering interactions, then the amount of light a surface reflects towards any given direction becomes proportional to the probability of intersecting a mirror suitably oriented to produce such a reflection. This relation is formally expressed by the Cook-Torrance equation [CT82]

\[
f_r(\omega_i, \omega_o) = \frac{F(\omega_i - \omega_o) G_2(\omega_i, \omega_o, \omega_o) D(\omega_o)}{4 \cos \theta_i \cos \theta_o},
\]

where \(F, G_2, D\) respectively denote the Fresnel term, the microfacet shadowing-and-masking function and the microfacet normal distribution function (NDF) [WMLT07].

Derivation of the GGX BRDF The GGX microfacet BRDF is a specialized form of Equation (11), which follows from the two following hypotheses. The first hypothesis is that the microfacets are oriented according to the NDF of a truncated ellipsoid. In the spirit of Section 2, we write this NDF as a function of that of a hemisphere. This leads to the following expression

\[
D(\hat{\omega}_m) = \frac{\det M^T}{\|M^T \hat{\omega}_m\|^2 \|M^T \hat{\omega}_m\|^2} \left( \text{hemisphere NDF see Equation (3)} \right) \text{ Jacobian} \left( \frac{M^T \hat{\omega}_m}{\|M^T \hat{\omega}_m\|} \right) \left( \frac{\|M^T \hat{\omega}_m\|^2}{\text{Jacobian}} \right) \left( \frac{M^T \hat{\omega}_m}{\|M^T \hat{\omega}_m\|} \right) \left( \frac{\|M^T \hat{\omega}_m\|^2}{\text{Jacobian}} \right),
\]

where \(M\) is the matrix that maps the hemisphere to the GGX ellipsoid as defined in Section 2. The second hypothesis is that the microfacets are subject to the Smith shadowing model. We provide an example of such a validation in Figure 1, where we show PBRT-v4 renderings of a scene based on two different GGX microfacet BSDFs. As expected, our sampling algorithm does indeed converge towards the same result as Heitz.

We validated that our method converges systematically lower than 4% for all the scenes we tested. We introduced a novel importance sampling algorithm for GGX microfacet BSDFs. Our method is easy to implement and offers systematic speed-ups over that of Heitz.

Microfacet Theory Intuitively, microfacet theory describes surface reflectance as the result of interactions between incident light and microscopic mirrors (which compose the actual surface) that deviate this incident light. If we only consider single-scattering interactions, then the amount of light a surface reflects towards any given direction becomes proportional to the probability of intersecting a mirror suitably oriented to produce such a reflection. This relation is formally expressed by the Cook-Torrance equation [CT82]

\[
f_r(\omega_i, \omega_o) = \frac{F(\omega_i - \omega_o) G_2(\omega_i, \omega_o, \omega_o) D(\omega_o)}{4 \cos \theta_i \cos \theta_o},
\]

where \(F, G_2, D\) respectively denote the Fresnel term, the microfacet shadowing-and-masking function and the microfacet normal distribution function (NDF) [WMLT07].

Derivation of the GGX BRDF The GGX microfacet BRDF is a specialized form of Equation (11), which follows from the two following hypotheses. The first hypothesis is that the microfacets are oriented according to the NDF of a truncated ellipsoid. In the spirit of Section 2, we write this NDF as a function of that of a hemisphere. This leads to the following expression

\[
D(\hat{\omega}_m) = \frac{\det M^T}{\|M^T \{\omega}_m\|} \left( \text{hemisphere NDF see Equation (3)} \right) \text{ Jacobian} \left( \frac{M^T \hat{\omega}_m}{\|M^T \hat{\omega}_m\|} \right) \left( \frac{\|M^T \hat{\omega}_m\|^2}{\text{Jacobian}} \right) \left( \frac{M^T \hat{\omega}_m}{\|M^T \hat{\omega}_m\|} \right) \left( \frac{\|M^T \hat{\omega}_m\|^2}{\text{Jacobian}} \right),
\]

where \(M\) is the matrix that maps the hemisphere to the GGX ellipsoid as defined in Section 2. The second hypothesis is that the microfacets are subject to the Smith shadowing model [SMi67]. This leads to the following shadowing term [Hei14, DHI15]

\[
G_2(\hat{\omega}_m, \omega_i, \omega_o) = \frac{G_1 G_0}{G_1 + G_0 - G_1 G_0},
\]

\[
G_1(\hat{\omega}_m, \omega_i, \omega_o) = \frac{\chi^+(\hat{\omega}_m \cdot \omega_i) \chi^+(\hat{\omega}_m \cdot \omega_o)}{\int_{\mathbb{S}^2} \max(0, \hat{\omega}_m \cdot \omega_i, 0) D(\hat{\omega}_m) d\hat{\omega}_m},
\]

where \(G_1\) and \(\chi^+\) respectively denote the Smith monostatic shadowing term and the Heaviside step function. Again, we provide an analytic expression for Equation (15) based on the shadowing term of the hemisphere [AKDW22]

\[
\chi^+(\hat{\omega}_m \cdot \omega_i) = \frac{T(\omega_i \cdot \hat{\omega}_m) \cdot [\omega_i \cdot \omega_i]}{\|\omega_i\|^2},
\]

\[
G_1 = \frac{\chi^+(\hat{\omega}_m \cdot \omega_i) \chi^+(\hat{\omega}_m \cdot \omega_o)}{\sigma_{std}(\omega_o)},
\]

where \(\sigma_{std}\) is the cross-sectional area of the hemisphere defined in Equation (4) and illustrated in Figure 3.
Importance Sampling Using the GGX VNDF

The GGX microfacet BRDF typically occurs in the direct illumination equation

\[ L(\omega_a) = \int_{S^2} L(\omega_b) f_r(\omega_a, \omega_b) \cos \theta_b |d\omega_b|, \]

where \( L \) denotes incident radiance. The GGX VNDF provides a way to solve this equation using the Monte Carlo estimator

\[ L(\omega_a) \approx I(\omega_a) = \frac{1}{N} \sum_{j=0}^{N-1} \frac{L(\omega_{a,j}) f_r(\omega_{a,j}, \omega_b) \cos \theta_{b,j}}{PDF(\omega_{a,j}, \omega_b)}, \]

where \( \omega_{a,j} \) denotes the \( j \)-th sample of the estimator. This sample is produced in the two following steps:

1. Sample the GGX VNDF from direction \( \omega_b \) using Listing 1. This produces a normal \( \omega_{a,j} \) whose PDF is given in Equation (1).
2. Use this normal to reflect the incident direction \( \omega_a \) according to the reflection equation \( \omega_{b,j} = 2(\omega_{a,j} \cdot \omega_a) \omega_a - \omega_a \).

The resulting PDF is given by the VNDF weighted by the Jacobian of the reflection operator as provided in Equation (8):

\[ PDF(\omega_{a,j}, \omega_b) = \frac{G(\omega_{a,j}, \omega_b) D(\omega_{a,j})}{4 \cos \theta_b}. \]

By sampling the GGX VNDF in this way, Equation (19) effectively provides an estimator with state-of-the-art variance [Hd14].

References


© 2023 Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.