Length n Probability Generation

Matthew Seguin

Here I will show how to generate a random set of probability vectors (arbitrary length n). The entries are chosen uniformly from 0 to the left over probability after the previous ones have been chosen then the order is randomized otherwise we will get a pattern of (generally) smaller probabilities near the end.

Generator

```
prob_generator <- function(N, m = 2500){
  p_vecs <- matrix(0, nrow = N, ncol = m)
  sums <- 0
  for (i in 1:(N-1)){
    p_vecs[i,] <- runif(n = m, max = 1 - sums)
    sums <- sums + p_vecs[i,]
  }
  p_vecs[N,] <- 1 - sums
  for (i in 1:m){
    p_vecs[,i] <- p_vecs[sample(1:N, N),i]
  }
  return (p_vecs)
}
```

Quickly note that this actually generates an $n-1$ dimensional vector since once we know $n-1$ values in the vector the final value must make it sum to 1 and hence is fixed. So $n-1$ values are free while one is constrained.

First a little testing that this does indeed produce length n probability vectors.

There is always an issue when it comes to machine precision of floating point numbers so that sums can not be exactly one but we can make a simple tolerance based on machine epsilon.

```
tol <- 5*.Machine$double.eps
tol
```

```
## [1] 1.110223e-15
```

```
for (i in 2:1000){
  m = 1000
  test <- prob_generator(N = i, m = m)
  stopifnot(
    min(test) >= 0,
    max(abs(colSums(test) - 1)) <= tol
  )
}
```

As there are no errors we can see that all of these have non-negative probabilities and that they sum to 1 (within our tolerance based on machine imprecision) and hence define length n probability vectors.

Examining Distribution

The first thing to note is that this generation process for $p = (p_1, ..., p_n)$ is equivalent to that of generating $W_1, ..., W_n \overset{\text{iid}}{\sim} \text{Exponential}(\lambda) = \text{Gamma}(1, \lambda)$ then considering the distribution of $(L_1, ..., L_n) = \frac{(W_1, ..., W_n)}{W_1 + ... + W_n} \sim \text{Dirichlet}(1, ..., 1)$ by the properties of the order statistics for uniform random variables. Therefore we know $p = (p_1, ..., p_n) \sim \text{Dirichlet}(1, ..., 1)$ and hence:

$$f_p(p_1, ..., p_n) = (n-1)!\, I\Big(\sum_{i=1}^{n} p_i = 1\Big) \prod_{i=1}^{n-1} I(p_i \geq 0)$$

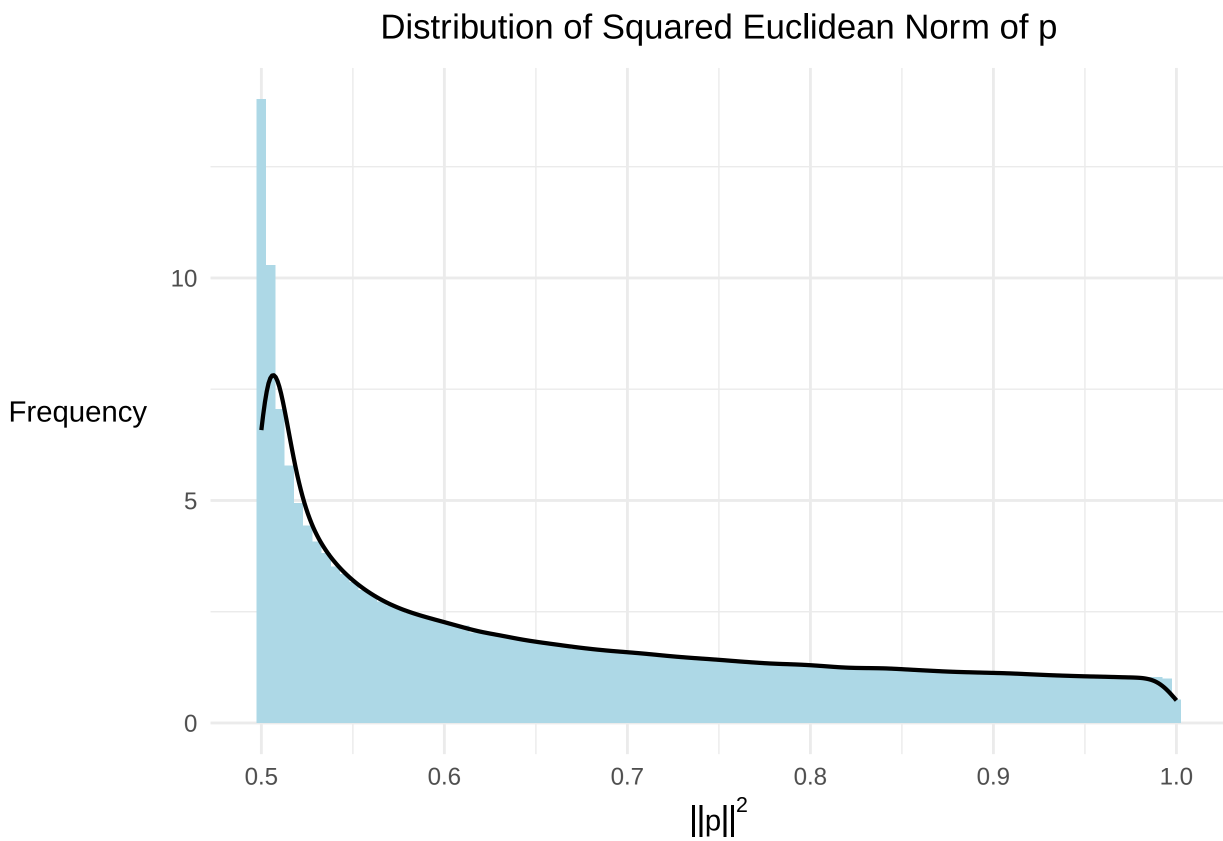
Meaning that p is uniform over the $n-1$ dimensional simplex (a desirable result). Below are some graphs to analyze the distribution of $||p||^2$.

First $n = 2$

```
suppressWarnings(
  suppressMessages({
  library(tidyverse)
  library(latex2exp)
  })
  )

probs <- prob_generator(N = 2, m = 1000000)
sq_norms <- colSums(probs^2)
sq_norms <- data.frame(sq_norm = sq_norms)

sq_norms %>%
  ggplot(aes(x = sq_norm)
         ) +
  geom_histogram(aes(y = after_stat(density)),
                 fill = "lightblue",
                 bins = 100) +
  geom_density(col = "black",
               linewidth = 0.75) +
  labs(x = TeX("$||p||^2$"),
       y = "Frequency  ",
       title = TeX("Distribution of Squared Euclidean Norm of $p$")
       ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        axis.title = element_text(color = "black"),
        axis.title.y = element_text(angle = 0, vjust = 0.5)
        )
```

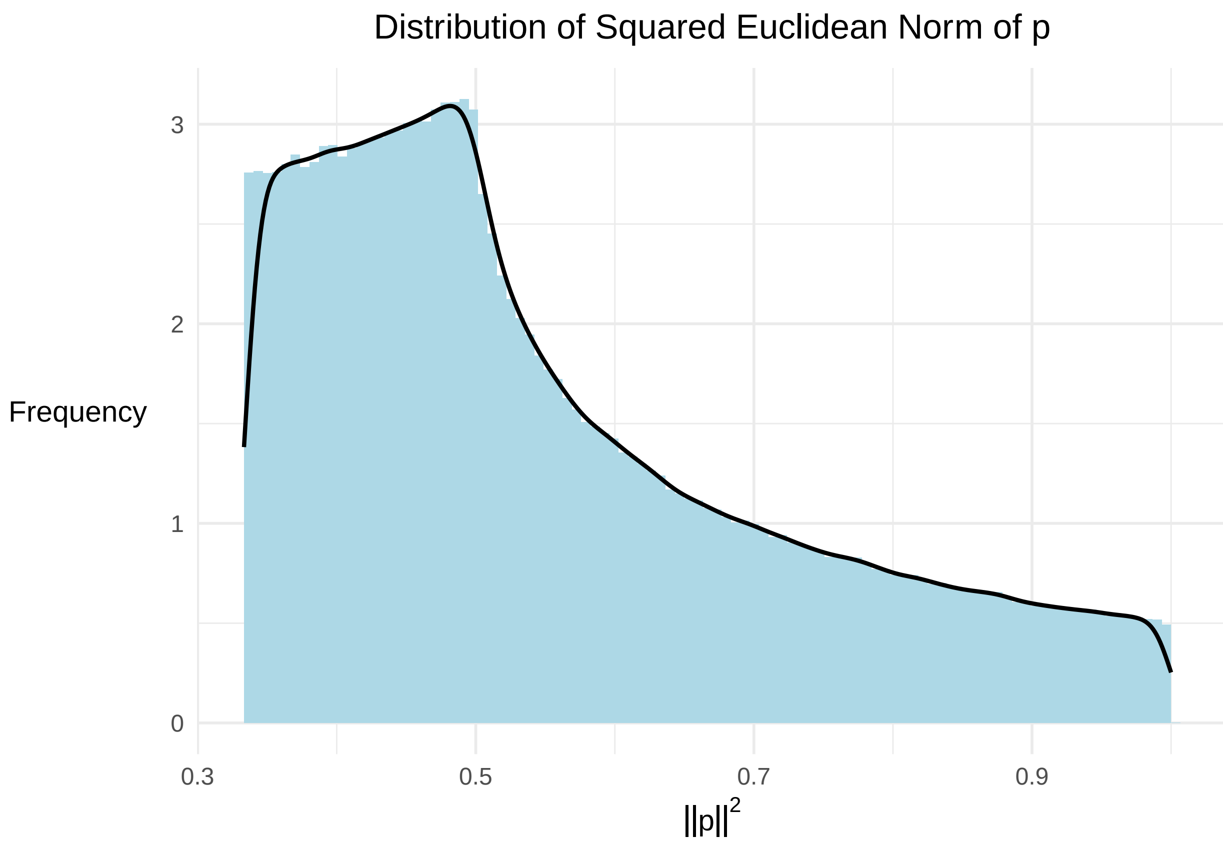


Comment.

Now $n = 3$

```
probs <- prob_generator(N = 3, m = 1000000)
sq_norms <- colSums(probs^2)
sq_norms <- data.frame(sq_norm = sq_norms)

sq_norms %>%
  ggplot(aes(x = sq_norm)
         ) +
  geom_histogram(aes(y = after_stat(density)),
                 fill = "lightblue",
                 bins = 100) +
  geom_density(col = "black",
               linewidth = 0.75) +
  labs(x = TeX("$||p||^2$"),
       y = "Frequency  ",
       title = TeX("Distribution of Squared Euclidean Norm of $p$")
       ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        axis.title = element_text(color = "black"),
        axis.title.y = element_text(angle = 0, vjust = 0.5)
        )
```

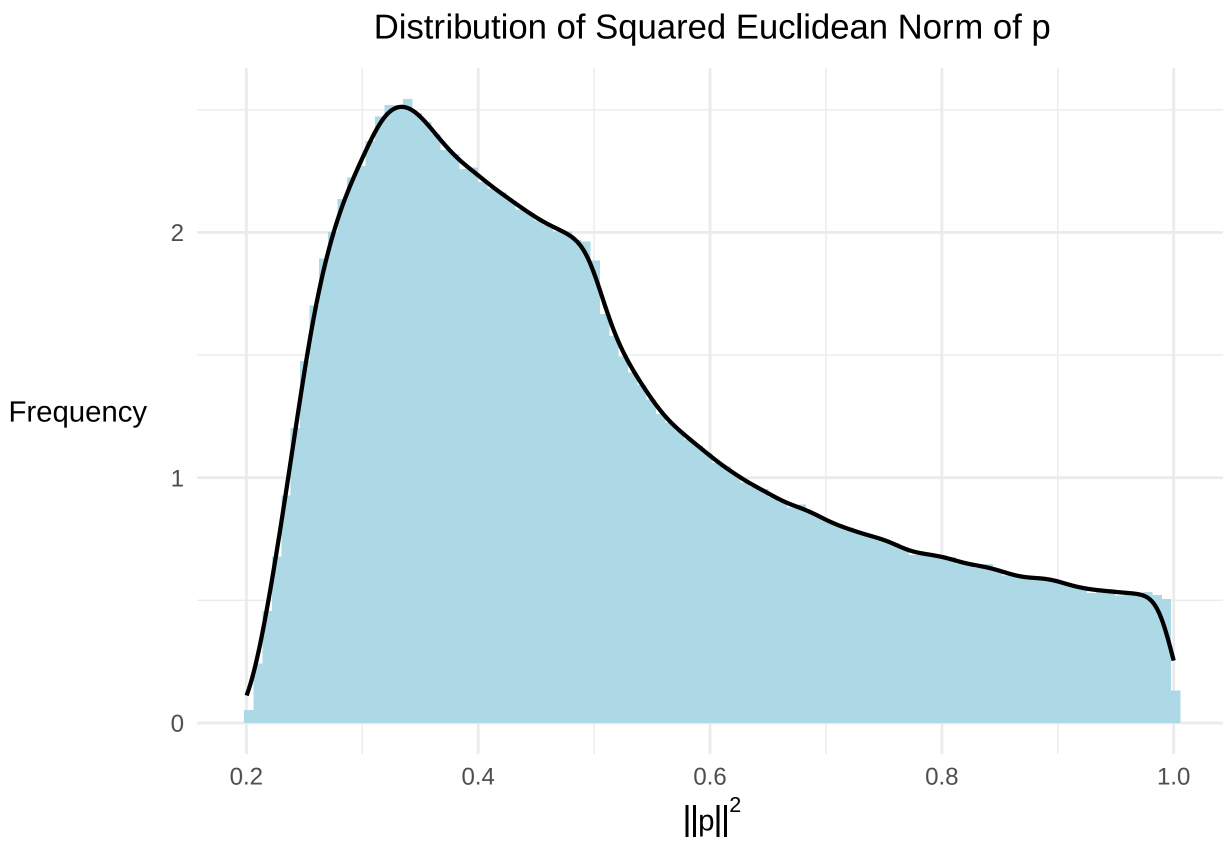


Comment.

Now $n = 5$

```
probs <- prob_generator(N = 5, m = 1000000)
sq_norms <- colSums(probs^2)
sq_norms <- data.frame(sq_norm = sq_norms)

sq_norms %>%
  ggplot(aes(x = sq_norm)
         ) +
  geom_histogram(aes(y = after_stat(density)),
                 fill = "lightblue",
                 bins = 100) +
  geom_density(col = "black",
               linewidth = 0.75) +
  labs(x = TeX("$||p||^2$"),
       y = "Frequency  ",
       title = TeX("Distribution of Squared Euclidean Norm of $p$")
       ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        axis.title = element_text(color = "black"),
        axis.title.y = element_text(angle = 0, vjust = 0.5)
        )
```

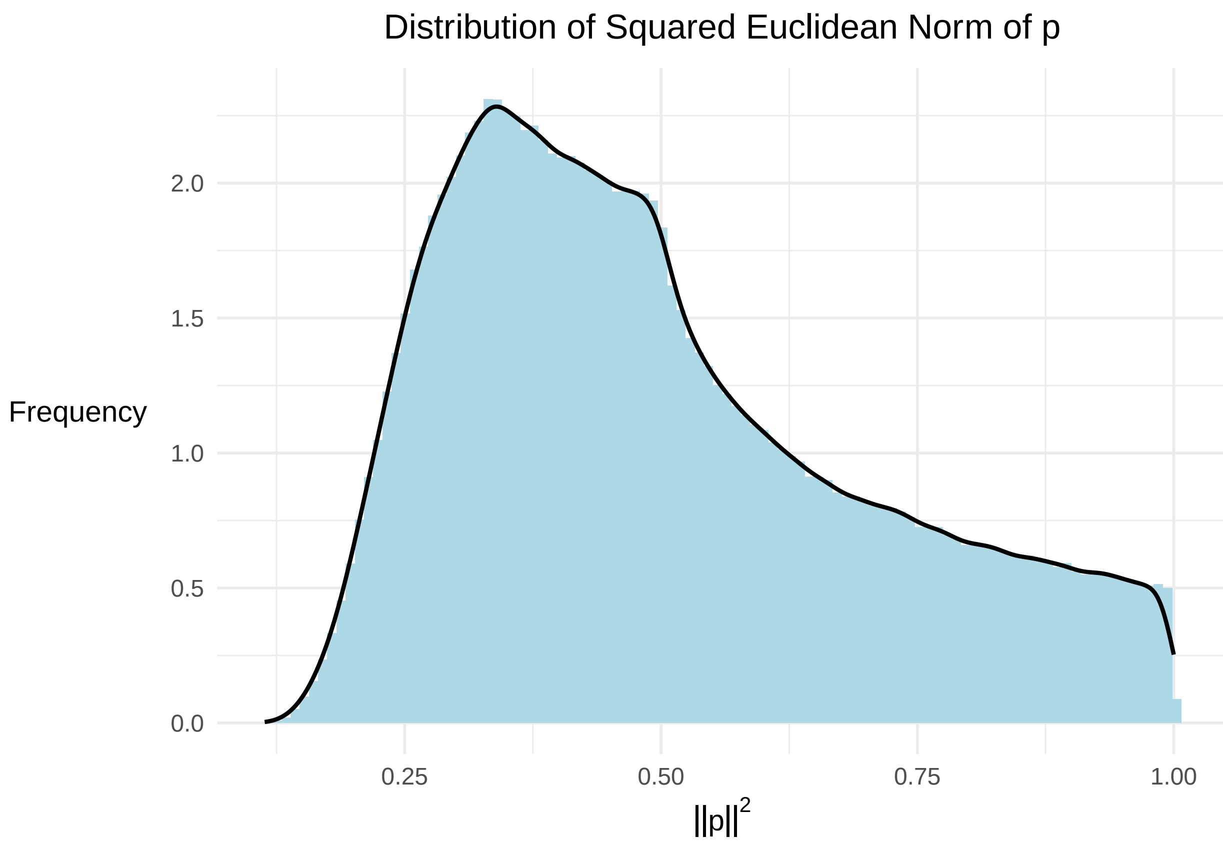


Comment.

Now $n = 10$

```
probs <- prob_generator(N = 10, m = 1000000)
sq_norms <- colSums(probs^2)
sq_norms <- data.frame(sq_norm = sq_norms)

sq_norms %>%
  ggplot(aes(x = sq_norm)
         ) +
  geom_histogram(aes(y = after_stat(density)),
                 fill = "lightblue",
                 bins = 100) +
  geom_density(col = "black",
               linewidth = 0.75) +
  labs(x = TeX("$||p||^2$"),
       y = "Frequency  ",
       title = TeX("Distribution of Squared Euclidean Norm of $p$")
       ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        axis.title = element_text(color = "black"),
        axis.title.y = element_text(angle = 0, vjust = 0.5)
        )
```

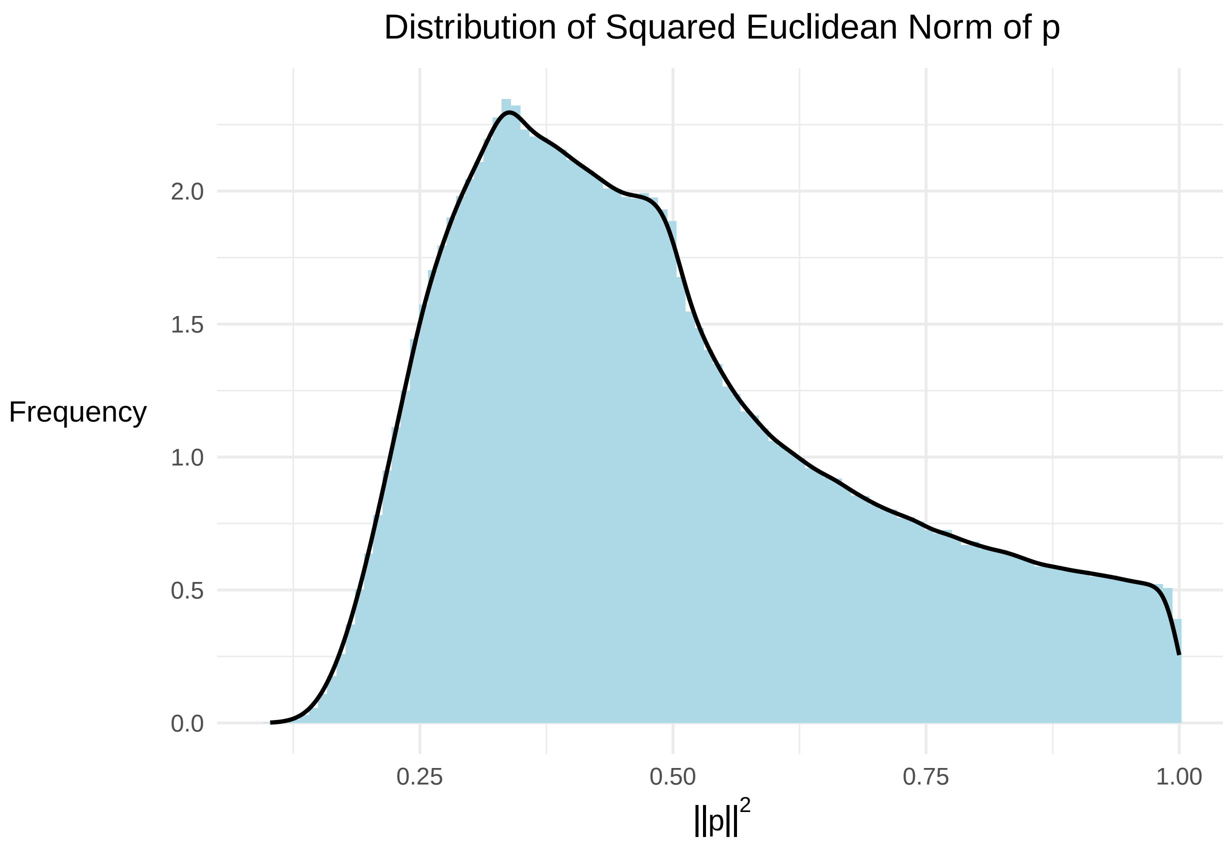


Comment.

Now $n = 50$

```
probs <- prob_generator(N = 50, m = 1000000)
sq_norms <- colSums(probs^2)
sq_norms <- data.frame(sq_norm = sq_norms)

sq_norms %>%
  ggplot(aes(x = sq_norm)
         ) +
  geom_histogram(aes(y = after_stat(density)),
                 fill = "lightblue",
                 bins = 100) +
  geom_density(col = "black",
               linewidth = 0.75) +
  labs(x = TeX("$||p||^2$"),
       y = "Frequency  ",
       title = TeX("Distribution of Squared Euclidean Norm of $p$")
       ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        axis.title = element_text(color = "black"),
        axis.title.y = element_text(angle = 0, vjust = 0.5)
        )
```

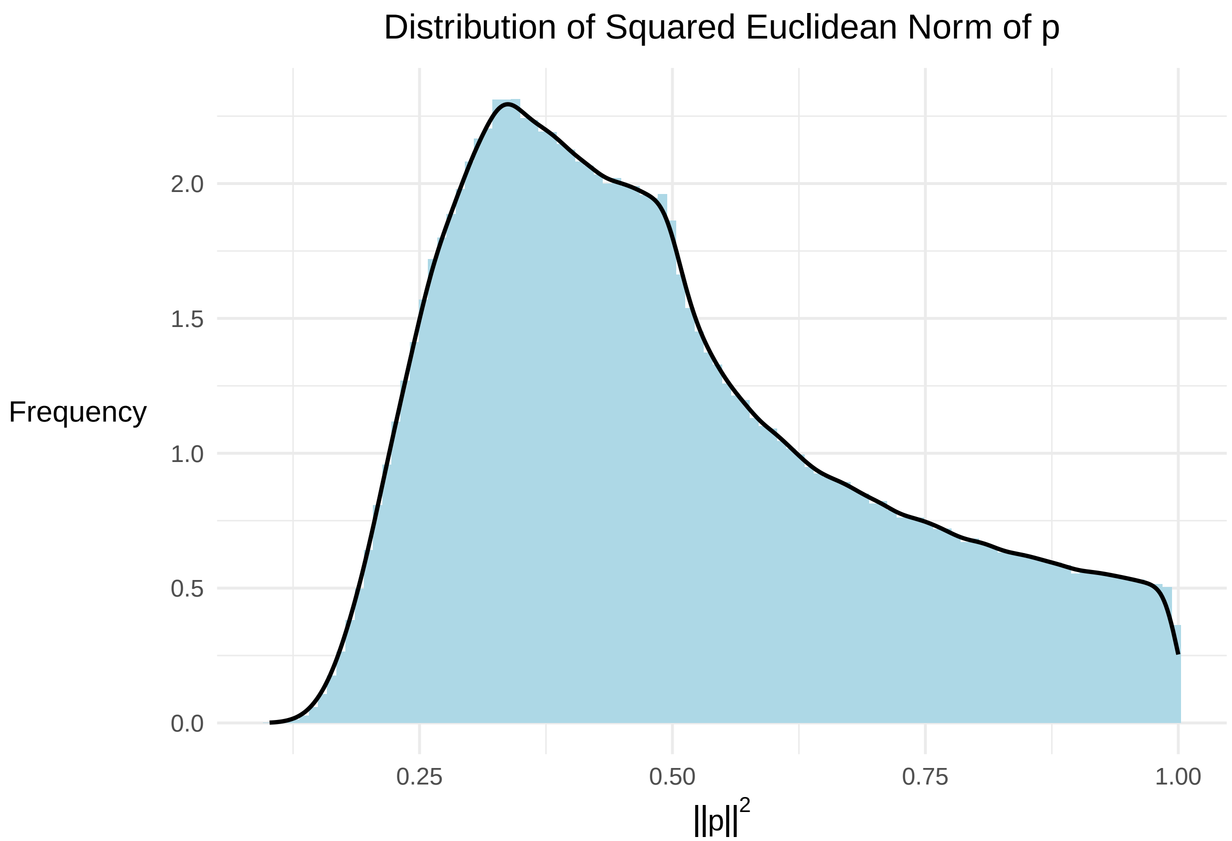


Comment.

Now $n = 1000$

```
probs <- prob_generator(N = 1000, m = 1000000)
sq_norms <- colSums(probs^2)
sq_norms <- data.frame(sq_norm = sq_norms)

sq_norms %>%
  ggplot(aes(x = sq_norm)
         ) +
  geom_histogram(aes(y = after_stat(density)),
                 fill = "lightblue",
                 bins = 100) +
  geom_density(col = "black",
               linewidth = 0.75) +
  labs(x = TeX("$||p||^2$"),
       y = "Frequency  ",
       title = TeX("Distribution of Squared Euclidean Norm of $p$")
       ) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        axis.title = element_text(color = "black"),
        axis.title.y = element_text(angle = 0, vjust = 0.5)
        )
```


Distribution of Squared Euclidean Norm of p
Frequency
2.0
1.5
1.0
0.5
0.0
0.25
0.50
0.75
1.00
||p||²

Comment.