Introduction

Hello everyone, I'm Neel Shah, one of Stan's Google Summer of Code (GSOC) interns. Over the summer, I will add Lambert$W$ transforms to Stan which enable us to model skewed and heavy-tailed data as approximate normals. This post motivates the idea and describes the theory of Lambert$W \times Z$ random variables.

Though the Normal distribution is one of our go to tools for modeling, the real-world often generates observations that are inconsistent with it. The data might appear asymmetric around a central value as opposed to bell-shaped or have extreme values that would be discounted under a normality assumption. When we can't assume normality, we often have to roll up our sleeves and delve into a more complex model. But, by using Lambert$W \times Z$ random variables it is possible for us to model the Skewness and Kurtosis from the data. Then, we continue with our model as if we had a Normal distribution. Later, we can back-transform predictions to account for our Skewness and Kurtosis.

In the first part, we introduce the Lambert$W$ function, also known as the product logarithm. Next, we discuss Skewness and Kurtosis (measures of asymmetry and heavy-tailedness), define the Lambert$W \times Z$ random variables, and share our implementation plans. Finally, we demonstrate how Lambert$W$ transforms can be used for location-hypothesis testing with Cauchy-simulated data.

To simplify matters, we are focusing on the case of skewed and/or heavy-tailed probabilistic systems driven by Gaussian random variables. However, the Lambert$W$ function can also be used to back-transform non-Gaussian latent input. Because Stan allows us to sample from arbitrary distributions, we anticipate that Lambert$W$ transforms would naturally fit into many workflows.

Contents

1 Lambert$W$ as a transform

The magic of the transform is due to the Lambert$W$ function. It is defined as $z = ue^u,\ z, u \in \mathbb{C}$. This function is typically defined for complex numbers, but we are going to restrict our attention to the real line. You'll notice there are no solutions for z in $-\infty$ to approximately -0.35. This trough occurs exactly at

$-1/e$. From $-1/e$ on, there are two real solutions. These two solution domains are split across the $u = -1$ boundary. When $u \geq -1$, z is in the principal branch and when $u \leq 1$, it is in the nonprincipal branch. Georg M Goerg (link to paper) shows that the probability of observations in the nonprincipal branch is really, really low. So, in practice, we stick to the principal branch when transforming from observation space to latent-variable space. Its possible to use the non-principal branch, should we really need it, see paper for more details.

For now all that really matters is that the mapping from z (y-axis) to u (x-axis) unskews the data. You may have noted the extra γ parameter, this controls the curvature and hence the Skewness.

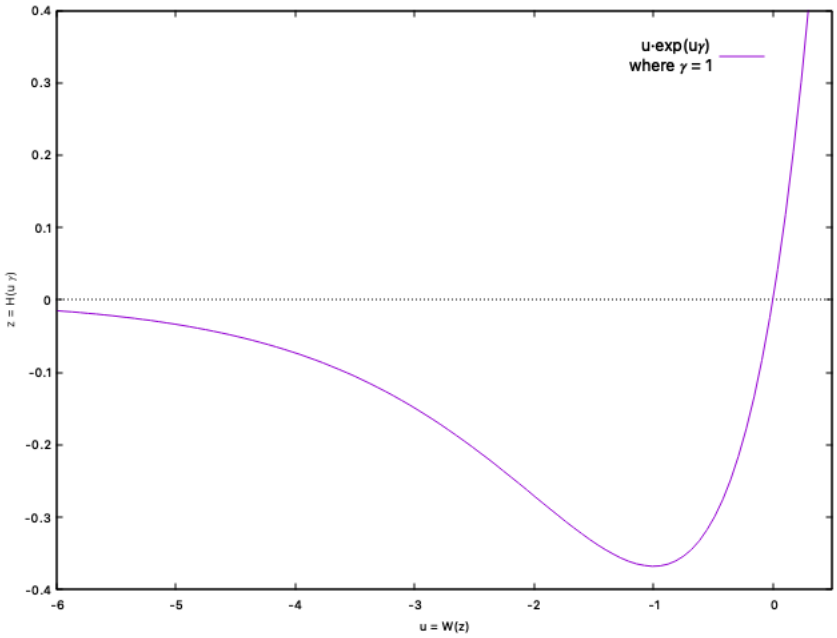

Figure 1: Lambert$W$ function $z = u \exp\{\gamma u\}$ with skewness controller γ

2 Skewness, Kurtosis and Lambert$W \times Z$

For all this concern over asymmetry and heavy-tails, it's worth briefly discussing precisely what these concepts represent. When we speak of heavy-tails, we are usually thinking about Kurtosis in excess of the Normal distribution's value of 3. For a centered-and-scaled random variable X, the Kurtosis is defined as the expected fourth moment $\mathbb{E}[X^4]$, which is influenced by the samples furthest away from the 0. Other measures of heavy-tailedness such as the tail-

index measure how quickly the pdf approaches 0 as we increase $|x|$ relative to a squared-exponential decay. Intuitively, because excess Kurtosis spreads the mass of the distribution over a larger area, it can frustrate inferences assuming some centrality.

Similarly, Skewness measures the degree of asymmetry around the mean. For X as before, it is defined as $\mathbb{E}[X^3]$. Because we're exponentiating a demeaned random variable to an odd-power, we measure the degree to which positive and negative values cancel (cubed and in expectation). Since the Normal distribution can be reflected across the mean, it exhibits 0 Skewness. However, many real-life datasets (e.g. grade inflation, stock volatility) are interesting precisely because they contain high-Skewness; there, we are likelier to find ourselves on one-side of the mean than the other.

A Lambert$W \times Z$ random variable Y can be defined in a variety of similar ways. Each variant has particular advantages and disadvantages, which can be explored by simply making graphs (desmos.com) of the functions for different parameters of (g, h).

Function	Models skew?	Models heavy tails?
$Y = Z \exp\{gZ\},\ g \in \mathbb{R}$	Yes	Yes
$Y = Z \exp\{\frac{h}{2} Z^2\},\ h \geq 0$	No	Yes
$Y = \begin{cases} Z \exp\{\frac{h_r}{2} Z^2\} & Z > 0 \\ Z \exp\{\frac{h_l}{2} Z^2\} & Z < 0 \end{cases},\ h_r, h_l \geq 0$	Yes	Yes

The powerful result of Georg M Goerg (link to paper) is that for any choice of distribution of Z (with continuous support), we have that Lambert$W \times Z$ has a well-defined cdf and pdf. To establish this result, it was necessary to define an inverse transform Y^{-1}, which will be the key to applying this theory. While we focus on $Z \sim \mathcal{N}(\mu, \sigma^2)$, there are other valid choices for Z, such as: Gamma, F, χ^2, Uniform, Student's t, etc. The reason why we focus on Z normal is because we'd like to provide the transforms in Stan within an 8-week period. Also, MLE for joint parameter estimation of transform parameters (g, h) and Normal distribution parameters (μ, σ) is well defined.

When analyzing their data (or observational space) we want them to be able to decide whether they need to model Skewness, Kurtosis or both-simultaneously. Then they would calibrate a prior on the degree of Skewness (g) and either symmetric Kurtosis (h) or asymmetric Kurtosis (h_l, h_r) along with the Normal mean (μ) and standard deviation (σ). Then they would build their model as normal, assuming that their data is sampled from a Lambert$W$ x Normal distribution. Because the MLE is well-defined, so is the log-likelihood and we would provide that to the Stan sampler and construct a posterior on (g, h, mu, sd) simultaneously.

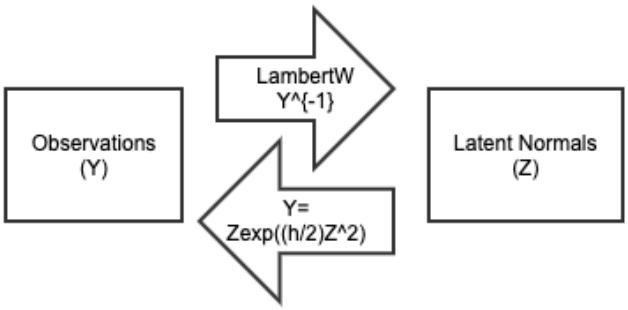

Figure 2: Inferences driven by a normal and precise control over higher moments.

In a Bayesian setting, this can be motivated as a hierarchical model with an unobserved random variable Z. Just like with unobserved parameters, we put a distribution on Z. We can specify such a model, in generative form as follows:

1. $h \sim |\mathcal{N}|(0, 1)$
2. $\mu \sim \mathcal{N}(0, 1)$
3. $\frac{1}{\sigma^2} \sim \gamma(a, b)$
4. $Z \sim \mathcal{N}(0, 1)$
5. $Y = Z \exp\{\frac{h}{2} Z^2\}\sigma + \mu$

3 Cauchy example

The Cauchy distribution is famously heavy-tailed. With an undefined mean and variance, many common statistical techniques such as the Student's t-test do not apply. How would one test a hypothesis about the location of the Cauchy distribution? Drawing on robust statistics, we could first estimate the location and scale using the median and interquartile range (75% - 25% percentiles) respectively, then run a Student's t-test on the centered-and-rescaled samples. While this procedure works in a simulated setting, it is entirely heuristic.

Putting some details in the experiment above: Suppose we observe 100 simulated-samples of Y, a standard Cauchy random variable, and we want to test the hypothesis H0: loc $= 0$ vs. H1: loc $\neq 0$.

```
set.seed(23)
dat <- rcauchy(100)
n <- length(dat)
```

Using R, we come up with an estimate of loc(est.) = -0.13 and scale(est.) = 1.8, which gives a t-stat of -0.72, which is insufficient to reject H0.

```
x0 <- median(dat)
gamma <- IQR(dat)
tstat <- x0/(gamma/sqrt(n))
```

Alternatively, we could try to fit the parameters of the t-distribution using the fitdistr (R, MASS) package. It uses MLE to estimate the t-distr. parameters as: mean=-0.01, standard dev.=0.89, degrees of freedom=1.0.

```
library(MASS)
fitdistr(dat, "t")
```

Both these approaches (correctly) suggest that the location is not far from 0.

While the above analysis indicates that our immediate problem is solved, we don't get any insight from the heuristic and we cannot always fit the t-distr. For example, we can't even say how heavy-tailed our data is. Applying the Lambert$W$ way, we assume our Cauchy simulated data corresponds to a transformed Normal Z using the function $Y = Z \exp\{\frac{h}{2}Z^2\}$ where $h > 0$ is the degree of heavy-tailedness. If we estimated $h = 0$, then $X = Z$ and we would find that our data has Normal tails. For h positive, the function moves Z's tail observations further out, which gives Y heavy-tails. The trick to exploiting this assumption is the Lambert$W$ function W that satisfies the relationship $z = W(z)\exp\{W(z)\}$. Because of this, W crucially defines an inverse Y^{-1}, which can be used to back-transform our observed Y into normal variates Z. If we apply this back-transform on the Cauchy samples (while jointly estimating h, along with Z's mean and standard dev.), then we can run a t-test on the normalized data. The t-stat stat we get is -0.40 and the estimated tail-index h is 1.0 (analogous to the t-dist.'s degrees of freedom). Furthermore, we get a pre-parameterized transform Y^{-1}, which can be applied on the next sample.

```
library(Lambert$W$)

dat_normzld <- Gaussianize(dat)
tstat_normlzd <- mean(dat_normzld)/(sd(dat_normzld)/sqrt(n))
```

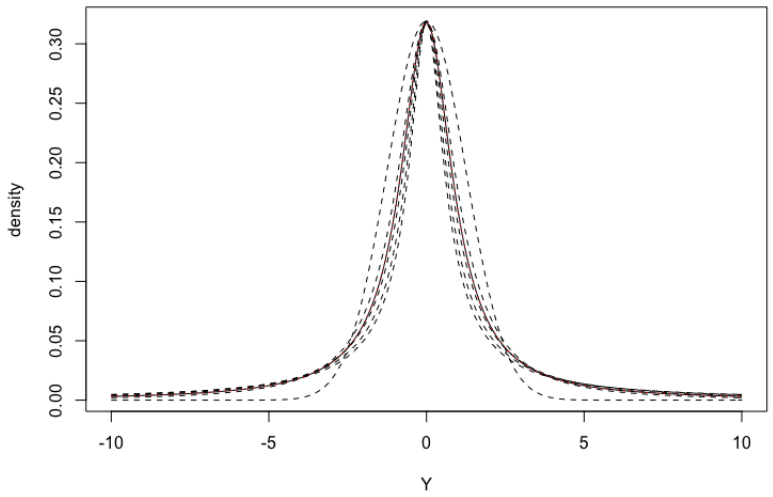

Figure 3: The symmetric Kurtosis variant of LambertW (dashed) can match the t (solid, red) and Cauchy (solid, black) location and tails by varying $h \in [0, 2]$

```
x<-seq(-10,10,0.1)
plot(x, dcauchy(x), type='l', xlab='Y', ylab='density')
for (d in seq(0,2,0.5))
{
    lines(x,
          dLambertW(x,
                    theta=list(delta=d, beta=c(0,1.25)),
                    distname='normal'),
          lty=2)
}
lines(x, dt(x, df=1), lty=3, col='red')
```