# Statistical Modeling and Analysis of SST-Pyr Synapses

Taisuke Yasuda   11/28/2016

1   Research Goal

When we sampled postsynaptic responses from SST-Pyr synapses, the resulting distribution had an oddly long tail relative to the expectation of the distribution. That is, these cells have a very high failure rate, yet exhibit responses of very high amplitude. The research problem of this paper is to explain how postsynaptic response amplitudes that are much higher than the average amplitude are achieved.

Figure 1: A typical distribution with high failure rate and high maximum amplitude
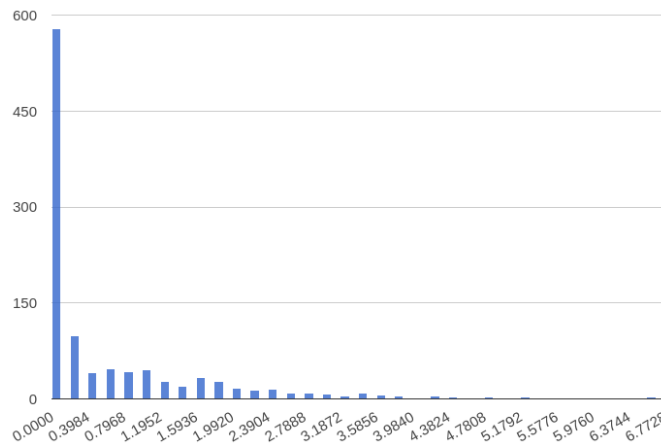
2   Results

2.1   Summary

Our studies show that the observed data is best explained by a distribution of postsynaptic amplitudes over contacts with a high variance. We first find that the simple model obtained by assuming that the release probabilities, which we refer to as p, and response amplitudes, which we refer to as q, of each contact are the same does not sufficiently explain the data obtained from the experiment. We then find that when we take the response amplitudes across contacts to have a uniform distribution with large width parameter, we recover a model that explains the data well, as measured by QQ plots.

2.2   The statistical model

We seek to explain the underlying mechanics of the observed distribution by constructing a statistical model that sufficiently matches the observed distribution. The model we build is the following:

$$\boldsymbol{A} = \sum_{j=1}^{N} \boldsymbol{A}_j, \boldsymbol{A}_j = \begin{cases} N(q_j, \sigma_j^2) & \text{if } \boldsymbol{S}_j = 1 \\ 0 & \text{if } \boldsymbol{S}_j = 0 \end{cases}, \boldsymbol{S}_j = \text{Bern}(p_j)$$

In this model, $\boldsymbol{A}$ is the random variable we observe, $\boldsymbol{A}_j$ are the latent random variables that represent the release amplitudes of each contact, and $\boldsymbol{S}_j$ are the latent random variables that represent the release success of each contact. We assume that the $\boldsymbol{A}_j$ are independent. If the contact $\boldsymbol{A}_j$ releases, which it does with probability p_j, then we observe a response assumed to be normally distributed with mean q_j and variance σ_j^2. Otherwise, we observe a response of 0 from $\boldsymbol{A}_j$. Finally, we observe the sum of the $\boldsymbol{A}_j$, which gives $\boldsymbol{A}$.

<table>
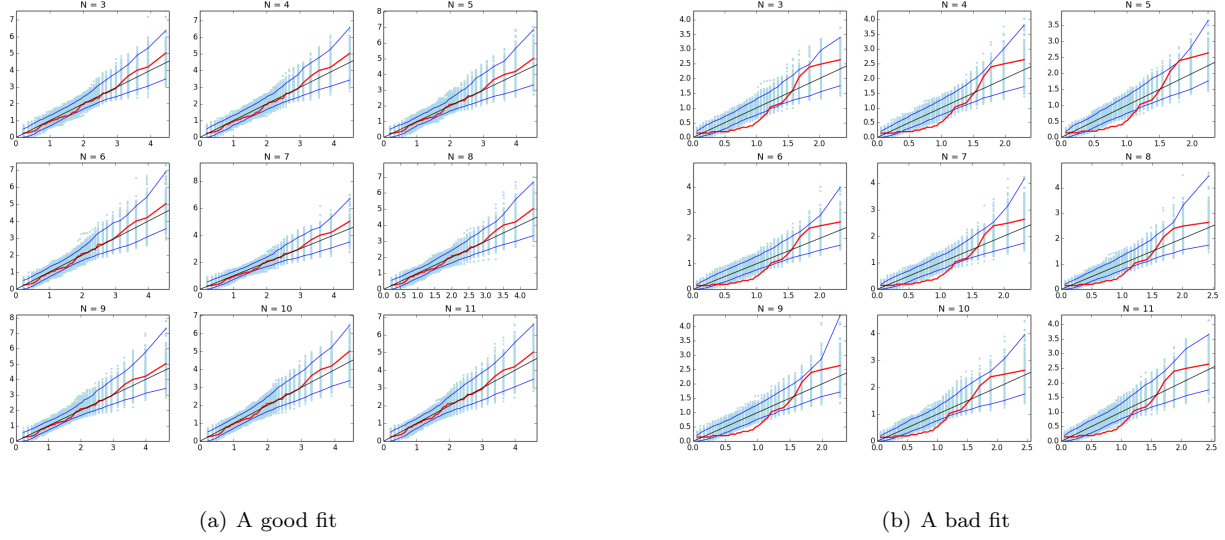<tr><td>(a) A good fit</td><td>(b) A bad fit</td></tr>
</table>

Figure 2: QQ plots for constant parameter model

2.3   Analysis of the constant parameter model

Note that in the above model, we have reduced the distribution to a choice of four parameters, N, the number of contacts, q, the vector of the average response at each contact, p, the vector of the release probability at each contact, and σ^2, the vector of the variance of response at each contact. As an initial analysis of the model, we assume that all the release probabilities, response amplitudes, and variances at each contact are the same. Although we do not expect a model assuming this to sufficiently explain the observed data, we use this analysis to gain a feel for the data.

Given this assumption, we can estimate the parameters p, q, and σ from the statistics of the data to be

$$p = 1 - \sqrt[N]{p_f}, \ q = \frac{\overline{\mu}}{Np}, \ \sigma^2 = \frac{s^2}{Np} - q^2 + q^2 p,$$

where p_f is the failure rate, $\overline{\mu}$ is the sample mean, and s^2 is the sample variance (details in the technical appendix). We do not attempt to estimate N and instead we simulate the model over a range of reasonable values of N, as suggested by data from reconstructions of SST-Pyr synapses.

In figure 2, we simulate the statistical model for 9 values of N, from 3 through 11, and display examples of a cell whose behavior is well-explained by this model and one whose behavior is not explained at all by this model, as indicated by the QQ plots. In fact, in 60 out of the 90 trials, the QQ plots indicated that the observed distribution remained within the 95% confidence intervals at each quantile. However, we also note that in 19 out of the 90 trials, the tail of the distribution was above the identity line in the QQ plot. Thus, we have not yet completely explained the existence of high amplitudes with high probability yet.

2.4   Analysis of distributions on response amplitudes

Equipped with a statistical model with parameters to explore, we attempt to explain the observed data by changing our assumptions on the parameters. We keep the mathematically convenient assumption of constant p and σ and assume distributions other than the point mass distribution on q. We begin with two distributions whose distribution is defined by its width and perturb its width to see the effect on the QQ plots.

2.4.1   Uniform distribution

In order to ensure that our statistical model has the same mean value, we take our uniform distribution with width w to have the pdf

$$f_q(x) = 1_{\left[\frac{\bar{\mu}}{Np} - \frac{1}{2}w, \frac{\bar{\mu}}{Np} + \frac{1}{2}w\right]} \frac{1}{w}$$

where 1_S is the indicator function for the set S. Since we restrict our response amplitudes to be positive, we simulate our distribution for a range of values of w from the interval $\left[0, 2\frac{\bar{\mu}}{Np}\right)$.

However, even at the widest possible interval for the uniform distribution, significant imporvements in the QQ plot were not observed.

2.4.2   Arcsine distribution

We similarly attempted an arcsine distribution, which specifies more weight on the highest and lowest amplitudes in the given width. Similarly, we did not observe significants improvements, though better than the uniform distribution.

2.5   Curve fitting of the cdf

We first show that we can express the cdf of $\boldsymbol{A}$ as a sum of Gaussians cdfs as follows. We observe an amplitude of $\sum_{j \in S} N(q_j, \sigma^2) = N\left(\sum_{j \in S} q_j, |S|\sigma^2\right)$ with probability $\prod_{j \in S} p = p^{|S|}$, where $S \subseteq [N]$ is a set of indices of synaptic contacts and $[N]$ denotes the set $\{1, 2, \ldots, N\}$. Thus, the pdf of the overall distribution is

$$f_{\boldsymbol{A}}(x) = \sum_{S \in \mathcal{P}([N])} p^{|S|} \varphi_{\sum_{j \in S} q_j, |S|\sigma^2}(x).$$

where $\mathcal{P}([N])$ denotes the power set of $[N]$ and $\varphi_{\mu, \sigma^2}(x)$ denotes the pdf of the Gaussian variable with mean μ and variance σ^2. Then, we have the cdf by integrating:

$$F_{\boldsymbol{A}}(x) = \int_0^x f_{\boldsymbol{A}}(t)\, dt = \int_0^x \sum_{S \in \mathcal{P}([N])} p^{|S|} \varphi_{\sum_{j \in S} q_j, |S|\sigma^2}(t)\, dt = \sum_{S \in \mathcal{P}([N])} p^{|S|} \left(\Phi_{\sum_{j \in S} q_j, |S|\sigma^2}(x) - \Phi_{\sum_{j \in S} q_j, |S|\sigma^2}(0) \right)$$

where Φ_{μ, σ^2} denotes the cdf of the Gaussian variable with mean μ and variance σ^2.

Now note that we have estimates of points $(x, F_{\boldsymbol{A}}(x))$ that the cdf satisfies, since we have observed amplitudes x_i and an estimate of $F_{\boldsymbol{A}}(x_i)$ by sorting the observed amplitudes and taking the ith amplitude x_i in the sorted data to have $F_{\boldsymbol{A}}(x_i) = \frac{1}{i+1}$. In fact, this is equivalent to fitting the quantiles of $F_{\boldsymbol{A}}(x)$ to match the data as in the QQ plot analysis.

3   Methods/Technical Appendix

3.1   Estimation of pararmeters for constant parameter model

3.1.1   Release probability

Under any assumption of the parameters, we have that the postsynaptic amplitude is 0 if and only if all N contacts are 0. Then, if we assume that all the p_j are equal, then we can express the probability that the postsynaptic amplitude is 0 as

$$\mathbf{Pr}\,[\boldsymbol{A} = 0] = (1 - p)^N.$$

We have an empirical estimate for $\mathbf{Pr}\,[\boldsymbol{A} = 0]$, the failure rate, which we write as p_f. Then, we can estimate p as

$$p_f = (1 - p)^N \iff p = 1 - \sqrt[N]{p}.$$

3.1.2 Response amplitude

Suppose that p and q are constant across all contacts. Then, by linearity of expectation, we have that

$$
\begin{aligned}
\mathbf{E}[\boldsymbol{A}] &= \sum_{j=1}^{N} \mathbf{E}[\boldsymbol{A}_j] \\
&= \sum_{j=1}^{N} \mathbf{E}[\boldsymbol{A}_j \mid \boldsymbol{S}_j = 1]\mathbf{Pr}\ [\boldsymbol{S}_j = 1] + \mathbf{E}[\boldsymbol{A}_j \mid \boldsymbol{S}_j = 0]\mathbf{Pr}\ [\boldsymbol{S}_j = 0] \\
&= \sum_{j=1}^{N} \mathbf{E}[N(q,\sigma^2)]p + 0 = \sum_{j=1}^{N} qp = Nqp.
\end{aligned}
$$

Then, since we already have an estimate for p and we can find an estimate for $\mathbf{E}[\boldsymbol{A}]$ by taking the average value $\overline{\mu}$ of all our observations, we have that

$$
\overline{\mu} = Nqp \iff q = \frac{\overline{\mu}}{Np}.
$$

3.1.3 Variance

Since we assumed that the $\boldsymbol{A}_j$ are independent, we have, by the linearity of variance for independent random variables, that

$$
\begin{aligned}
\mathbf{Var}[\boldsymbol{A}] &= \sum_{j=1}^{N} \mathbf{Var}[\boldsymbol{A}_j] \\
&= \sum_{j=1}^{N} \mathbf{E}[\boldsymbol{A}_j^2] - \mathbf{E}[\boldsymbol{A}_j]^2 \\
&= \sum_{j=1}^{N} \mathbf{E}[\boldsymbol{A}_j^2 \mid \boldsymbol{S}_j = 1]\mathbf{Pr}\ [\boldsymbol{S}_j = 1] + \mathbf{E}[\boldsymbol{A}_j^2 \mid \boldsymbol{S}_j = 0]\mathbf{Pr}\ [\boldsymbol{S}_j = 0] - \mathbf{E}[\boldsymbol{A}_j]^2 \\
&= \sum_{j=1}^{N} (\sigma^2 + q^2)p + 0 - q^2 p^2 \\
&= N(\sigma^2 p + q^2 p - q^2 p^2).
\end{aligned}
$$

Again, we have an empirical estimate of $\mathbf{Var}[\boldsymbol{A}]$, the sample variance s^2, so we can estimate σ^2 as

$$
s^2 = N(\sigma^2 p + q^2 p - q^2 p^2) \iff \sigma^2 = \frac{s^2}{Np} - q^2 + q^2 p.
$$

3.2 Estimation of parameters for distributed response model

Under this model, the release probabilities are still taken to be constant so we use the same estimation of release probability as the above.

3.2.1 Response amplitude

In order to make a suitable choice of the q_j, we simply require that the mean of the model matches the mean of the observed data. Note that we can enforce this by restricting the average value of the q_j, since

$$
\overline{\mu} = \sum_{j=1}^{N} p_j q_j = p \sum_{j=1}^{N} q_j \iff \frac{1}{N} \sum_{j=1}^{N} q_j = \frac{\overline{\mu}}{Np}.
$$

For both the uniform and arcsine distributions, the mean of the distribution is given by $\frac{1}{2}(a+b)$ where the support of the distribution is $[a,b]$. Furthermore, the distribution is completely determined by the support of the distribution for both the uniform and arcsine distributions. Thus, we determine a and b given this requirement in addition to the assumed value of the width w of the distribution, which is simply $w = b - a$:

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ -1 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \frac{\bar{\mu}}{Np} \\ w \end{pmatrix} \iff \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 1 & -\frac{1}{2} \\ 1 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} \frac{\bar{\mu}}{Np} \\ w \end{pmatrix} = \begin{pmatrix} \frac{\bar{\mu}}{Np} - \frac{1}{2}w \\ \frac{\bar{\mu}}{Np} + \frac{1}{2}w \end{pmatrix}.$$

Thus, we derive the uniform distribution pdf

$$f_q(x) = 1_S \frac{1}{w}$$

and the arcsine distribution pdf

$$f_q(x) = 1_S \frac{1}{\pi \sqrt{(x - (\frac{\bar{\mu}}{Np} - \frac{1}{2}w))((\frac{\bar{\mu}}{Np} + \frac{1}{2}w) - x)}}$$

where 1_S is the indicator function for $\left[\frac{\bar{\mu}}{Np} - \frac{1}{2}w, \frac{\bar{\mu}}{Np} + \frac{1}{2}w\right]$.

3.2.2   Variance

When we do not assume constant q, we modify the previous argument slighty.

$$\mathbf{Var}[A] = \sum_{j=1}^{N} \mathbf{E}[A_j^2 \mid S_j = 1]\mathbf{Pr}\ [S_j = 1] + \mathbf{E}[A_j^2 \mid S_j = 0]\mathbf{Pr}\ [S_j = 0] - \mathbf{E}[A_j]^2$$

$$= \sum_{j=1}^{N} (\sigma^2 + q_j^2)p + 0 - q_j^2 p^2$$

$$= N\sigma^2 p + \sum_{j=1}^{N} q_j^2 (p - p^2).$$

As before, we use the sample variance s^2, to estimate σ^2 as

$$s^2 = N\sigma^2 p + \sum_{j=1}^{N} q_j^2(p - p^2) \iff \sigma^2 = \frac{s^2}{Np} - \frac{1-p}{N}\sum_{j=1}^{N} q_j^2.$$