Mutual Information-based Optimal Design for estimation in 1-D SDEs - Hitting Times Observation Case (LIF MOdels)

Alexandre Iolov, Susanne Ditlevsen, André Longtin
<aiolo040 at uottawa dot ca>, alongtin at uottawa dot ca

March 29, 2015

Abstract

Given a leaky, noisy integrate-and-fire neuronal model - we discuss optimal design-type questions on what is the best external perturbation in order to facilitate parameter estimation using inter-spike intervals data only

Contents

List of Tables

List of Figures

1 Problem Formulation

The basic goal of 'Optimal Design' is to perturb a dynamical system in an 'optimal' way such as to 'best' estimate its structural parameters.

As such the problem is a blend of optimal control and estimation, where the objective of the optimal control is to improve the estimation, for example by minimizing the variance of the estimators.

For illustration sake we return to our favourite LIF model Given a noisy LIF neuronal model:

$$
\begin{aligned}
dX_s = (\underbrace{\alpha(t)}_{\text{control}} + \beta(\mu - X_s)\,\mathrm{d}s + \sigma\,\mathrm{d}W_s,\\
X(0) = .0,\\
X(t_{\mathrm{sp}}) = x_{th} \implies \begin{cases} X(t_{\mathrm{sp}}{}^{+}) & = .0 \\ t_k & = t_{\mathrm{sp}} \\ k & = k+1 \end{cases}
\end{aligned}
\tag{1}
$$

where (a subset of) the parameter set $\theta = \{\mu, \beta, \sigma\}$ is unknown.

Our goal is to choose $\alpha(t)$ as to estimate β 'best' given only that the spike times $\{t_k\}$ are observed

2 Notation

The probability density of the nth interval, conditional on some applied control α:

$$
\begin{array}{rcll}
g_n(\tau)\,\mathrm{d}\tau & := & \mathbb{P}(I_n \in [\tau, \tau + \,\mathrm{d}\tau)\,|\,\alpha(t)) & \text{(probability density)} \\
G_n(t) & := & \mathbb{P}[I_n \leq t\,|\,\alpha(t)] = \int_0^t g_\phi(\tau)\,\mathrm{d}\tau & \text{(cumulative distribution)} \\
\bar{G}_n(t) & := & \mathbb{P}(I_n > t\,|\,\alpha(t)) = 1 - G_\phi(t) & \text{(survivor distribution)}
\end{array}
\tag{2}
$$

We'll drop the n subscript when there is no confusion. There is also the transition distribution for X_t for $t \in [0, I_n)$:

$$
f(x,t) := \mathbb{P}[X_t \in x + \,\mathrm{d}x\,|\,X_0 = 0, X_{s<t} < 1] \quad \text{(transition distribution)}
\tag{3}
$$

$$
\begin{array}{rcl}
\partial_t f(x,t) & = & \underbrace{\dfrac{\beta^2}{2}}_{D} \cdot \partial_x^2 f + \partial_x \left(\underbrace{\Big(\beta(x-\mu) - \alpha(t)\Big)}_{U(x,t)} \cdot f \right) \\
& = & D \cdot \partial_x^2 f + \partial_x \Big(U(x,t) \cdot f \Big) \\
& = & -\partial_x \phi(x,t) \\
& = & \mathcal{L}[f] \\
& & \left\{ \begin{array}{lcl} f(x,0) & = & \delta(x) \\ D\partial_x f + Uf|_{x=x_-} & \equiv & 0 \\ f|_{x=x_{th}} & \equiv & 0. \end{array} \right.
\end{array}
\tag{4}
$$

The probability flux-out at the threshold boundary

$$\phi(x_{th}, s) = D\partial_x f|_{x=x_{th}}$$

is very important as it is related to the spike-time density via

$$g(t) = \phi(x_{th}, t) = D \cdot \partial_x f|_{x=x_{th}}$$

In a Bayesian approach, we have some *a priori* belief over the possible values of θ.

Let us call the prior over the parameters $\theta = \{\mu, \beta, \sigma\}$:

$$\rho(\theta)$$

.

Given a single observation t_{sp}, the posterior of the parameter belief dist'n is

$$p(\theta|t_{\text{sp}};\alpha) = \frac{g(t_{\text{sp}}|\theta;\alpha)\cdot\rho(\theta)}{\int_\Theta g(t_{\text{sp}}|\theta;\alpha)\cdot\rho(\theta)\,\mathrm{d}\theta} \tag{5}$$

Where $g(t_{\text{sp}}|\theta;\alpha)$ is the likelihood of X given in eq. (2).

The idea now, is to choose α such that the mutual information I between the two random variables is maximized. Here the Mutual Information is given by

$$I[\alpha] = \int_\Theta \int_{[0,\infty]} g(t|\theta)\rho(\theta)\cdot\log\left(\frac{g(t|\theta)}{\int_\Theta g(t|\theta)\rho(\theta)\,\mathrm{d}\theta}\right)\mathrm{d}t\,\mathrm{d}\theta. \tag{6}$$

See appendix C for why. Naturally for different controls, $\alpha(\cdot)$, the mutual info, I, will be different since g, the hitting time density depends on the shape of α. (ρ does NOT).

However, there is an added complication b/c we actually will observe many hitting times, and having less informative hitting times happen more often might be better than hitting times which are informative but happen less often.

The rigorous way to deal with this is to consider the mutual information between the parameters and the set of hitting times $\{t_n\}$, however this seems incredibly complicated, so instead we will maximize the 'Mutual Information rate, J, which we define as

$$J[\alpha] = \mathbb{E}[t_{\text{sp}}]^{-1}\cdot I[\alpha] \tag{7}$$

$$= \frac{\int_\Theta \int_{[0,\infty]} g(t|\theta)\rho(\theta)\cdot\log\left(\frac{g(t|\theta)}{\int_\Theta g(t|\theta)\rho(\theta)\,\mathrm{d}\theta}\right)\mathrm{d}t\,\mathrm{d}\theta}{\int_\Theta \int_{[0,\infty]} \tau g(t|\theta)\rho(\theta)\,\mathrm{d}t\,\mathrm{d}\theta} \tag{8}$$

Well, this does NOT look any less complicated..., Note that g (and thus implicitly α) appears 4 times in this expression. Recall that g is related to α via the solution of the Fokker-Planck equation and thus we can also write

$$J[\alpha] = \frac{\int_\Theta \int_{[0,\infty]} \partial_x f(1,t)\rho(\theta)\cdot\log\left(\frac{\partial_x f(1,t)}{\int_\Theta \partial_x f(1,t)\rho(\theta)\,\mathrm{d}\theta}\right)\mathrm{d}t\,\mathrm{d}\theta}{\int_\Theta \int_{[0,\infty]} t\partial_x f(1,t)\rho(\theta)\,\mathrm{d}t\,\mathrm{d}\theta} \tag{9}$$

We want to find the control input $\alpha(t)$, which maximizes J in eq. (9).

3 Gradient Ascent using Maximum Principle in order to maximize eq. (9) or rather the simpler eq. (6)

Let us discuss the optimization problem

$$\alpha(\cdot) = \operatorname*{arg\,max}_{\alpha\sim\text{admissible}} J[\alpha]$$

3.1 The nitty-grity of calculating the (infinite-dimensional) gradient $\nabla_\alpha J$

We would like to maximize eq. (9), but now we realize that doing so is very difficult, b/c we have a ratio of integrals. (The standard theory always works with just one integral).

Let's then drop the denominator integral and focus on the numerator. (i.e. we just go back to eq. (6))

$$I[\alpha] = -\int_\Theta \int_{[0,\infty]} \partial_x f(1,t)\rho(\theta) \cdot \log\left(\frac{\partial_x f(1,t)}{\int_\Theta \partial_x f(1,t)\rho(\theta)\,\mathrm{d}\theta} \right) \mathrm{d}t\,\mathrm{d}\theta \tag{10}$$

To proceed, we apply a Maximum Principle type derivation, in which we first seek the differential of the objective I in eq. (10) wrt. $\alpha(\cdot)$ and proceed from there.

As always, we start by augmenting our objective functional with the dynamics:

$$I = \int_\Theta \int_{[0,\infty]} \partial_x f_\theta(1,t)\rho(\theta) \cdot \log\left(\frac{\partial_x f_\theta(1,t)}{\int_\Theta \partial_x f_\theta(1,t)\rho(\theta)\,\mathrm{d}\theta} \right) \mathrm{d}t\,\mathrm{d}\theta \tag{11}$$

$$- \int_\Theta \int_0^\infty < p_\theta, (\partial_t f_\theta - \mathcal{L}[f_\theta]) > \mathrm{d}s \tag{12}$$

where the inner product, $< f, g >$ is just the space integral $\int f \cdot g \,\mathrm{d}x$ and we write $\mathcal{L}$ for the spatial differential operator in eq. (4).

This is exactly the same problem as we faced in the spike-time optimal control, except there the integrand looked something like $(t - t^*) D \partial_x f$. Thus the equations for the adjoint look exactly the same as there, with the exception of the Terminal Conditions (here assumed 0) and the BCs at the threshold:

In short, the equation for the adjoint function, p, is

$$\begin{aligned} \partial_t p &= -\mathcal{L}^*[p] \\ &= -\Big[D \cdot \partial_x^2 p + U(x,t,\theta) \cdot \partial_x p \Big]. \end{aligned}$$

$$\begin{cases} p_\theta\big|_{x=x_{th}} &= \log\left(\frac{\partial_x f_\theta}{\int_\Theta \partial_x f_\theta \rho(\theta)\,\mathrm{d}\theta)} \right) + 1 - \frac{\partial_x f_\theta}{\int_\Theta \partial_x f_\theta \rho(\theta)\,\mathrm{d}\theta)} \\ \partial_x p_\theta\big|_{x=x_-} &= 0 \\ p_\theta(x,\infty) &= 0 \end{cases} \tag{13}$$

In practice of-course, we will set the terminal conditions for p_θ at some finite value of t, although it is not immediately obvious what that value should be.

The whole goal of this exercise is to calculate the differential of I in eq. (10), wrt. the control $\alpha(t)$, i.e. to calculate $\delta I/\delta\alpha$. After the introduction of the adjoint state, p_θ, that is just:

$$\delta I = \int_\Theta \rho(\theta) \cdot \left(-\int_{x_-}^1 \partial_x p_\theta f_\theta \,\mathrm{d}x + p_\theta f_\theta\Big|_{x_-} \right) \mathrm{d}\theta$$

I.e for a given $\alpha(t)$, we solve for p, f and a few values of θ from the current belief distribution $\rho(\theta)$ and their corresponding probabilities/weights. Compute the above expression and increment α in the direction of increasing δI.

In practice, we usually take a very simple prior, something like three values with equal probability, something like:.

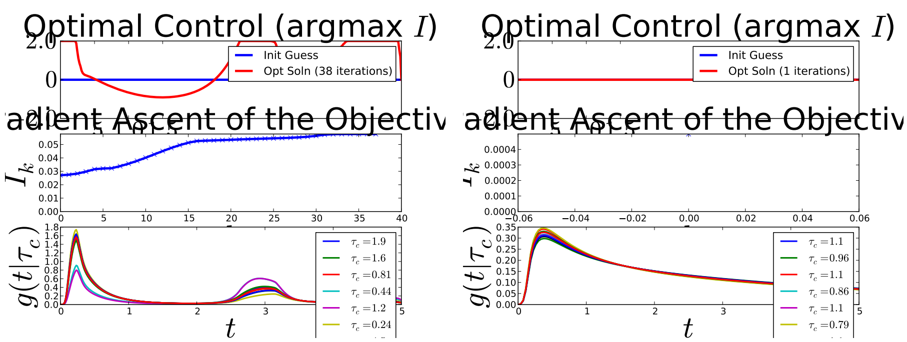
$$\rho(\beta) = \begin{cases} \frac{1}{3} & \text{if } \beta = \in \{.5, 1, 2\} \\ 0 & \text{o/w} \end{cases} \tag{14}$$

3.2 Effect of the Prior

Here we show that the optimal control is sensitive to the *spread* of the prior, for example if we have a tightly clustered vs. loosely spread prior, both centred at roughly the same mean (the log-prior has the same mean).

Very interestingly, we see that while for a wide prior, the optimal control has its characteristic double hump shape, that we have seen already, for a tight prior, that is no longer the case

Thus we see that the shape of the prior *has!* an effect on the optimal control.



(a) Wide Prior

(b) Tight Prior

Figure 1: The effect of the spread (variance) of the prior on the resulting optimal control

Let's look at it another way, we will consider our basic prior as a function of w

$$\rho(\beta) = \begin{cases} \frac{1}{2} & \text{if } \beta = \in \{1 - w, 1/(1 - w)\} \\ 0 & \text{o/w} \end{cases} \tag{15}$$

and sweep for $w = .1 : .1 : .9$ (in matlab notation).

The results are in fig. 2. Looking at fig. 2, we might be optimistic to hypothesize that we should be doing this online and as the uncertainty (roughly speaking w) of the parameter decreases, we should be changing the applied control... This brings us to *adaptive* versions of our scheme which is NOT something we have yet implemented.

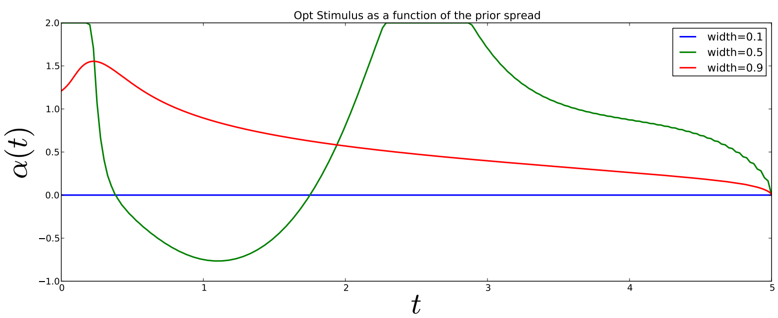


Figure 2: The effect of the width (w, a measure of uncertainty) of the prior on the resulting optimal control

4 Basic Estimation Experiment

We will run the following test:

Assume the true parameters

$$\mu = 0; \tau = 1; \beta = 1.;$$

We will assume we know σ, μ and don't know the time constant, τ so we are trying to maximize the Mutual Information between t_{sp} and τ.

Let's assume a very simple uniform prior on τ, $\tau_i = \{0.25, 1.5, 2.8, 4.0\}$, each with probability 1/4.

Then running the gradient ascent (details of the gradient ascent are omitted) we get the controls, objective and hitting time densities shown in fig. 3. The optimal control seems to be independent of the number of point in the prior (i.e. instead of 4 we could use 32 pts with weight 1/32 and get the same opt. control as in fig. 3).

WARNING: The optimal control search routines is VERY sensitive to the initial conditions. The one shown here is the one that has been empirically through bitter experience been found to give the highest MI, but currently there is no rigorous method to find it starting from arbitrary ICs (or from zero ICs for that mater)

4.0.1 Aside: the nitty-gritty of the estimation procedure

We have posed a fairly-simple estimation objective, in that it amounts to single variable optimization. The negative log-likelihood of an observed hitting-time set $\{t_n\}$ is

$$l(\tau) = -\sum_n \log(g(t_n|\tau)) = -\sum_n \log(-D\partial_x f(t_n|\tau)|_{x_{th}}) \tag{16}$$

The distributions are exemplified in fig. 4, for three different values of $N_s = 1e5$. We see that for the constant stimulations, $\alpha = \alpha_{crit}, \alpha_{max}$ it is very hard to distinguish between different values of τ.

In fig. 4 as well as in fig. 3, we get an indication for why the 'optimal control' is better than the constants. For the constant control the different hitting time densities look like local perturbations of each other, either a little more or a little less, but for the optimal control they are shifted, which means that we see the first indications that the Opt Control, might have some superiority over the 'Crit' Control (for example) as it seems to estimate a τ closer to 1 (the 'true' value). However, on average, the different shapes of $\alpha(t)$ seems to have a very limited impact on the estimates for τ (even though it has a very obvious impact on the shape of the hitting time distribution $g(t)$).

-

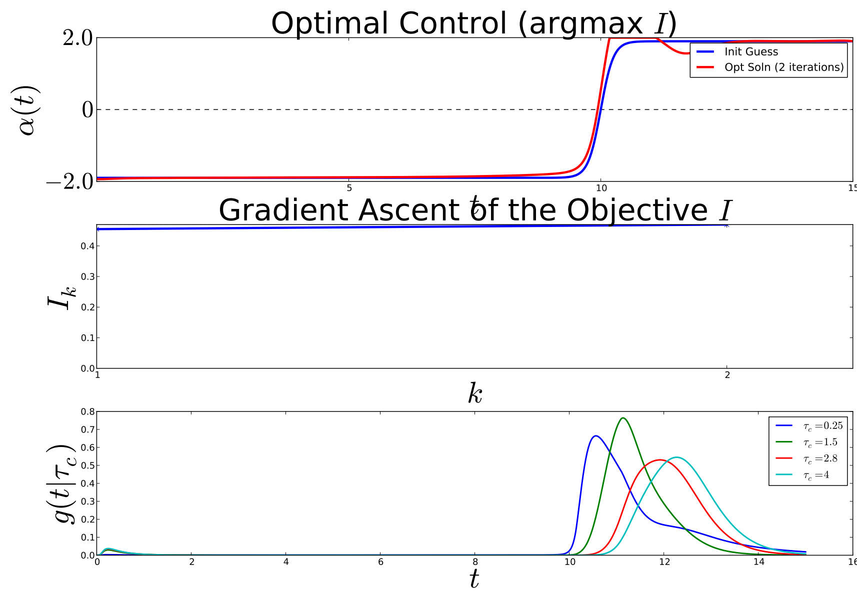
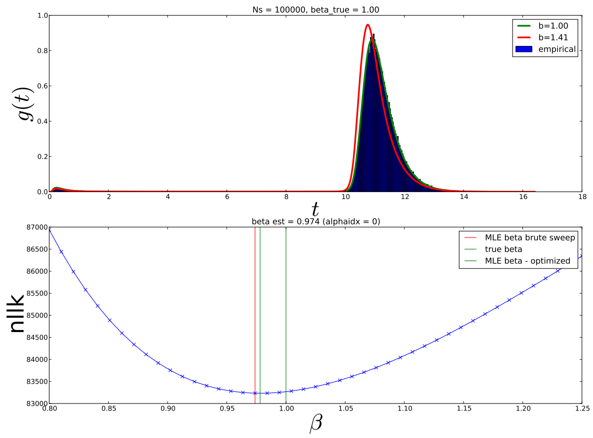
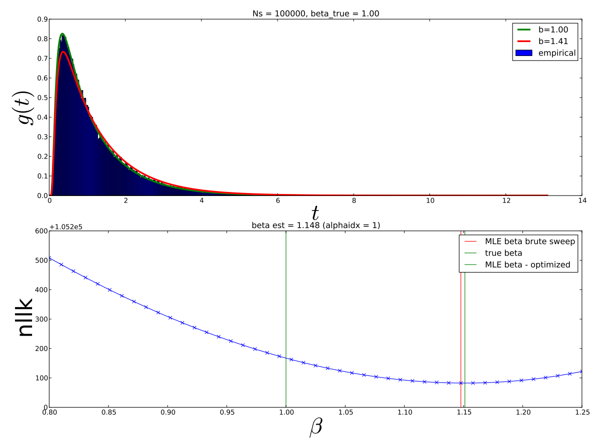


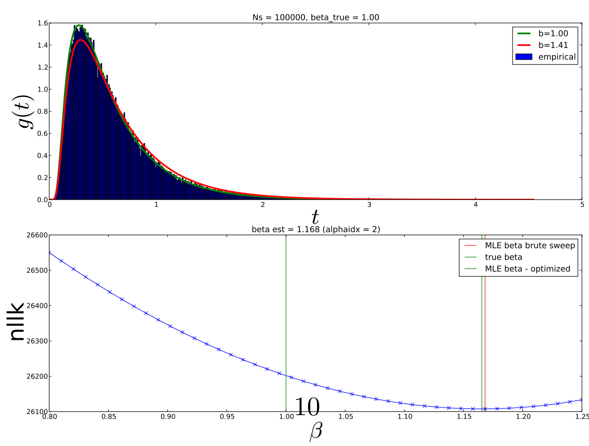
Figure 3: The gradient ascent for the optimization of I in eq. (10). Top panel, the initial and the optimal optimal controls, $\alpha_0(t), \alpha_{opt}(t)$. Middle Panel shows the progress of the Objective Mutual Information (I) (eq. (6)). The bottom plot shows the hitting times $g(t|\tau)$ corresponding to the 3 distinct values of τ in the prior $\rho(\tau)$



(a) opt



(b) crit



(c) max

Figure 4: Example of Empirical vs. Analytical Hitting time distributions, $g(t|\tau;\alpha)$, and the associated log-likelihoods. $N_s = 1e5$ hits

control type	mean(log($\hat{\tau}$))	std($log(\hat{\tau})$)
opt	-0.138	0.39
crit	-0.020	0.44
max	-0.046	0.41

(a) $N_b = 1000, N_s = 1e2$

control type	mean(log($\hat{\tau}$))	std($log(\hat{\tau})$)
opt	0.023	0.03
crit	-0.119	0.10
max	-0.130	0.11

(b) $N_b = 100, N_s = 1e3$

control type	mean(log($\hat{\tau}$))	std($log(\hat{\tau})$)
opt	0.023	0.01
crit	-0.126	0.02
max	-0.136	0.03

(c) $N_b = 10, N_s = 1e4$

control type	mean(log($\hat{\tau}$))	std($log(\hat{\tau})$)
opt	0.023	0.00
crit	-0.126	0.00
max	-0.137	0.00

(d) $N_b = 1, N_s = 1e5$

Table 1: Results for the estimates arising from simulations using various values of α (opt, crit, max). In each sub-table there are N_b parameter estimates for each distinct α, with N_s hitting times used to form an $\tau-$estimate. The 'true' value of τ is $\tau = 1$, i.e. $\log(\tau) = 0$.

4.0.2 Batch Performance of the perturbations over the estimators.

As is we have 3 candidates for perturbing the hitting times:

1. the optimal gradient-ascent-based control α_{opt} (see fig. 3 top panel)
2. the 'critical' constant control α_{crit}, ($\alpha_{crit}(t) = \tau$)
3. the max constant control, $\alpha_{\max}$ $(= 2)$

We now simulate N_b blocks of N_s hitting times each for the 3 alphas and then estimate τ over each set using MaxLikelihood over our computed expression for the density, $g(t|\tau; \alpha(t)))$. Naturally, for each control, we use the same Gaussian random draws per block of N_s Hitting of times). See fig. 4 for a empirical distribution of t_{sp} for $N_s = 1e5$

The estimation results are tabulated in in table 1.

Comments: It certainly looks like there is a marginal advantage to using the Optimal Control, α_{opt} over the simpler, constant controls. In particular both the bias and variance of the estimates seems to be significantly reduced. The strange results for $N_s = 100$ require us to look at the actual distribution of estimates. Results not shown, but the opt-estimates for $N_s = 100$ are bi-modal with a cluster far from the true value. For the higher $N_s = 1e3, 1e4, 1e5$, the opt-estimates are unimodal and tightly centred around the true value.

5 Online MI Optimization

Hre we outline a tentative approach to *online* optimization of the MI, which means

1. Find α_{opt} using the gradient ascent, for the prior ρ
2. Apply α_{opt} and measure several $1, 2 \ldots, N_{s,1}$ hitting times t_k
3. Update the ρ into a posterior conditional on the observed $\{t_k\}$
4. Recalibrate α_{opt} using the new ρ, i.e. go back to 1.

Efficiency considerations aside, we have all the tools to do pts. 1,2,4, it is only the prior update that needs to be discussed.

Of course we start by restating Bayes' formula

$$\rho(\theta|\{t_k\}) = \frac{\rho(\theta) \cdot \prod_k g(t_k|\theta;\alpha)}{\int_\Theta \rho(\theta) \cdot \prod_k g(t_k|\theta;\alpha)\, \mathrm{d}\theta}$$

In practice, exact calculation of $\rho(\theta|\tau_k)$ would not be possible in our context, so an approximation approach needs to be made.

The standard approach is to describe the belief distribution by an ensemble of points (particles). We now describe the basic aspect of how the particle ensemble is constructed, how it is updated and how it is resampled. We use the reference [4], in particular sec 4 therein.

5.1 Quick INtro to Particle Filtering

We have a prior $\rho(\theta)$ and we want to describe it by an ensemble of points θ_i. as

$$\rho(\theta) \sim \sum_i w_i \delta(\theta - \theta_i)$$

This is what we have been doing up to now for the MI optimization routines.

Again recall that a bayesian update, given the kth hitting time is

$$\rho(\theta|t_k) \propto g(t_k|\theta)\rho(\theta)$$

thus the weights are iteratively updated as

$$w_i \to w_i g(t_k|\theta_i)$$

Given the particle ensemble, $\{\theta_i, w_i\}$, we can then approximate the mean/variance of Θ as

$$\mathbb{E}[\Theta] \approx \sum_i w_i \theta_i$$

and

$$\mathbb{V}\mathrm{ar}[\Theta] \approx \sum_i w_i \theta_i \theta_i - (\mathbb{E}[\Theta])^2$$

In fact, in general

$$\mathbb{E}[f(\Theta)] \approx \sum_i w_i f(\theta_i)$$

approximates the expectation of any function, f, of the random variable Θ.

Here's the crux of the update/resample algorithm. Given a new observation, i.e. the latest hitting time t_k, the weights are updated according to

$$w_{i,k} = w_{i,k-1} \cdot g(t_k|\theta_i, \alpha_k),$$

where $g(|\theta_i, \alpha_k)$ is the probability density given the parameter value θ_i and the chosen stimulation that was applied during the kth sample, $\alpha_k(\cdot)$. The weights are then re-normalized so that at all time

$$\sum_i w_i = 1$$

The literature suggests that this procedure will tend to concentrate all the 'mass' on one location and most of the weights will decrease to 0. This concentration is 'bad' since eventually, all the weights go to zero and so effectively the distribution has converged artificially to a point, which might be the most likely point from the initial ensemble, but still be far from the 'true' value of the parameter. This adverse effect can be ameliorated by resampling the *locations* of the particle ensemble, θ_i. This resampling can be done in many ways, but a standard way is described in algorithm. 1 (which is copied from Algo4 in Granade et. al. [4], who in turn closely follow [7], but are very nice and pedagogical).

Algorithm 1 Particle Resampling Algorithm

Given $\{w_i, \theta_i\}_1^N{}_p$ the current particle ensemble (weight, locations)
$\mu \leftarrow \mathbb{E}[\Theta]$
$a = 0.98$ see [4, 7]
$h = \sqrt{1-a^2}$ i.e. $h \approx 0.1990$
$\Xi \leftarrow \mathrm{h}^2 \cdot \mathbb{V}\mathrm{ar}[\Theta]$
for $i = 1 \ldots N_p$ **do**
# *Re-sample each particle individually*
    draw j with probability w_j
the bigger w_i the more likely to choose θ_i
    $m_i \leftarrow \mathrm{a}\ \ \theta_j + (1-a)\mu$
    Resample θ_i' from $N(\mu_i, \Xi)$
    $w_i \leftarrow 1/\mathrm{N}_p$
end for   #$1 \ldots N_p$ *particle resampling*
return w_i, θ_i' the resampled particle ensemble

While, of course, updating happens after every iteration, re-sampling happens only when

$$\frac{1}{\sum_i^{N_p} w_i^2} < \frac{N_p}{2} \implies \text{resample!}$$

In fig. 5 we give an example for the filtering procedure to estimate the rate of a Poissonian process, $1/\tau$, for which the likelihood is just $\frac{1}{\tau}exp(-\frac{t}{\tau})$. From fig. 5 we can conclude that the basic mechanics of the filtering update/resample procedure is working. (We've tried it with other values of τ and it works for those as well.)

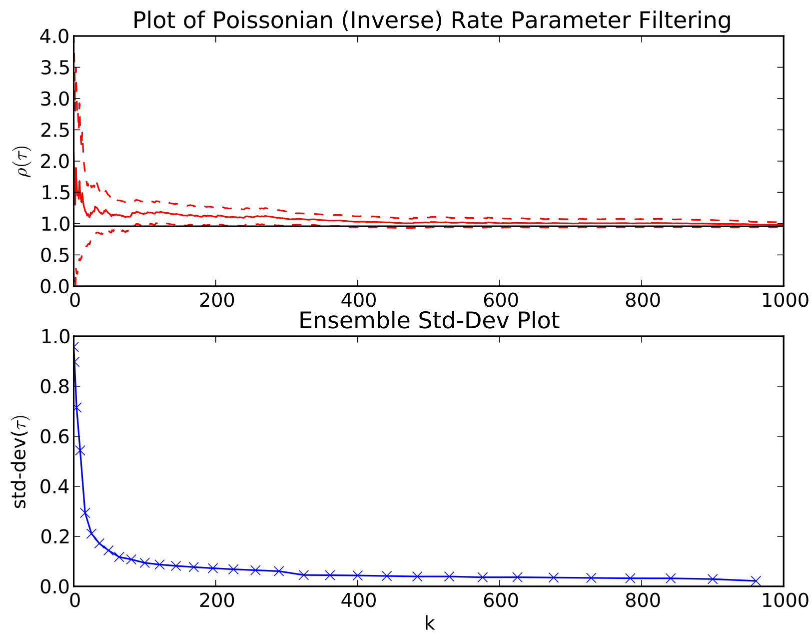


Figure 5: Filtering Example for a test-case. We are trying to estimate the inverse of the Poissonian rate, $1/\tau$, given inter-event intervals t_k. The true value is $\tau = 1$. The Maximum Likelihood estimate for τ in this problem is just the mean of the observed times $\bar{t_k}$, which happens to be 0.9583 in this example, while after the last observation, the ensemble mean $\pm$ std-dev is 0.9811 ± 0.0199

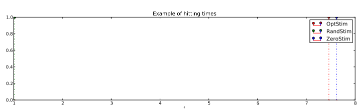
5.2 Example of Particle-Filtering + Online Estimation for hitting time parameter estimation

5.2.1 Single HItting Time Illustration

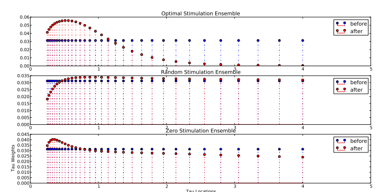
Let's try it. First, in fig. 6, we illustrate one iteration of the update, that is one hitting time given a stimulation from either the MI Optimal Controller or a random controller, which just gives a random constant stimulation per each hitting time.

Let's discuss what happens in fig. 6. Recall that lower values of τ imply higher restoring force and therefore longer hitting times (waiting times). Since in this sample, the observed hitting times were fairly long, especially for the MI-optimal stimulation, weights for smaller τs grow larger, while weights for bigger τs become smaller. However it is immediately clear that the MI-optimal stimulation is more discerning as it has almost entirely discarded (correctly) the possibility that $\tau > 2$, while the other two stimulations have resulted in only mild perturbation in the belief distribution.

WARNING: The new results are much better primarily through a judicious choice of the initial condition of the MI-optimal optimization (the initial guess for the optimal control.)



(a) Example Hitting Times



(b) Resulting Particle Ensemble Updates

Figure 6: Examples of a single iteration of the Online Stimulation-Estimation scheme.

5.2.2 Full Multiple Hitting-Times Experiment(s)

Now let's go through an entire estimation experiment. That is let's stimulate a sequence of hitting times and online update our parameter belief distribution, after every observation and then online-update our MI-optimal stimulation as the belief evolves.

The main result is shown in fig. 7, where we visualize the mean and confidence intervals for the belief distributions for the three protocols (MI-Optimal vs. Random Constant vs Zero), using $N_\tau = 32$ particles and $N_k = 251$ hitting times. In fig. 8, we show the different stimulations that were chosen by the Mutual-Info Maximization Algorithm.

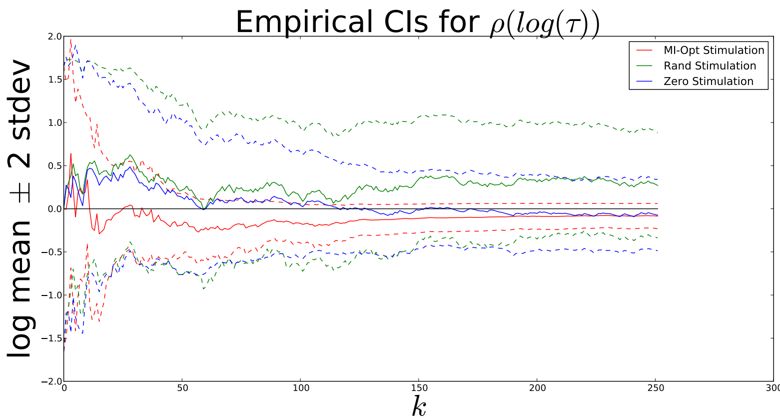


Figure 7: Evolution of the belief distributions given Optimal (red) or Random (green) stimulation. We used $N = 32$ particles to represent the ensemble for both stimulation protocols. There are 250 hitting times used

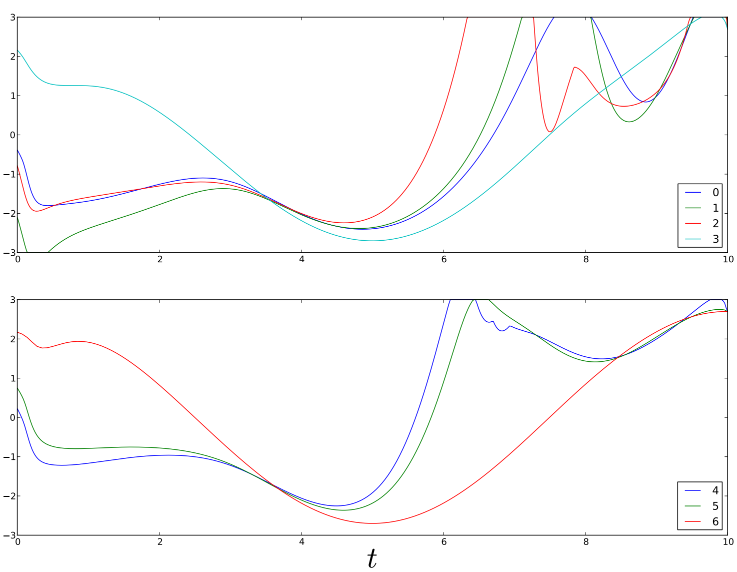


Figure 8: Different MI-OPtimal Stimulations computed as the Experiment evolves and the parameter belief distribution changes

We now simulate 10 independent experiments of approx. 500 hitting times each and visualize them in fig. 9. Here, we see that the MI-optimal stimulation is indeed producing more accurate estimates, faster (earlier in the experiment). This is true in 7 out of the 10 experiments, in 2 it is hard to determine whether there is a 'better' protocol and only once is the MI-based stimulation resulting in worse estimates than one of the alternatives ((f) in fig. 9).

The aggregated version of these results (the averaged evolution distribution from the independent experiments) is shown fig. 10, where the N distributions are averaged out over the different experiments. It is now clear that the MI-Optimal procedure produces more accurate estimates much faster.

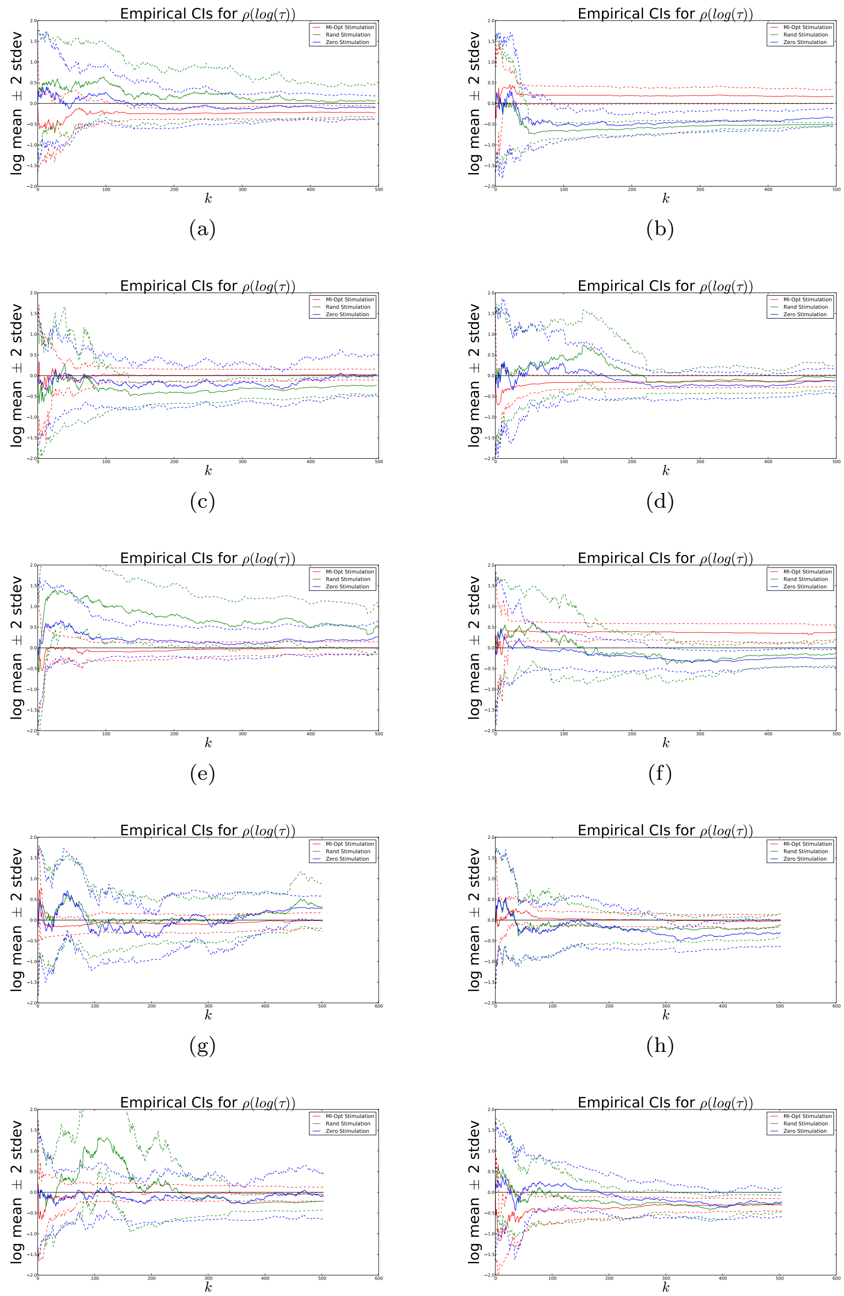


Figure 9: Various examples of the perturbation-estimation protocol with the belief distribution plotted against the hitting time k, with the MI-optimal stimulation (red), the constant random stimulation (green) and the zero, $\alpha \equiv 0 \, \forall t$, stimulation (blue)

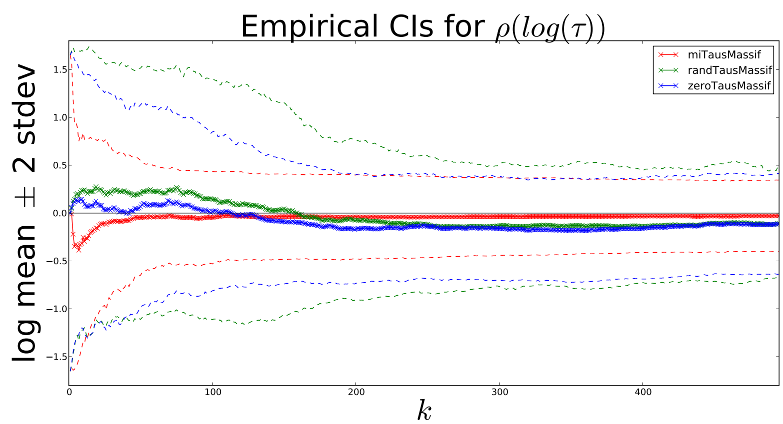


Figure 10: Aggregated belief distribution over $N = 16$ independent belief estimation experiment. The black line indicates the 'true' value of the unknown parameter τ.

A The basic idea of optimal design for SDEs of Lin et al.

Here we sketch the basic idea of Lin et al. [6].

Let us write the dynamics as such

$$dX = \underbrace{f(X, \theta, \alpha)}_{\text{controlled drift}} dt + \beta dW \tag{17}$$

Then given an observed path $\{x_t\}_0^{t_f}$, the log-likelihood, l wrt. the parameter set θ is

$$\begin{aligned} l(\theta|x_t) =& \frac{1}{2}\int_0^{t_f} \frac{f^2(x_t, \theta, \alpha)}{\beta^2}\, \mathrm{d}t \\ &- \int_0^{t_f} \frac{f(x_t, \theta, \alpha)}{\beta^2}\, \mathrm{d}W \end{aligned} \tag{18}$$

The goal then is to choose α in order to facilitate the estimation. The idea in [6] is to to choose α by maximizing the Fisher Information

$$\Phi(\theta, \alpha) = \mathbb{E}\left[\int_0^{t_f} \frac{(\partial_\theta f(x_t, \theta, \alpha))^2}{\beta^2}\, \mathrm{d}t\right] \tag{19}$$

Note that there are two optimizations intertwined. One, to maximize the likelihood l in order to obtain the actual estimate θ, the other - to maximize the Fisher Information evaluated at the (a priori unknown!) estimator θ.

The authors in Lin et al. [6] acknowledge that clearly one cannot form the Fisher Information directly since its evaluation requires the very parameter being sought! To remedy this, they apply a prior of θ. I still need to understand exactly what they do, but as far as I understand, they augment Φ by an outer expectation over the prior for θ, i.e. (I think!) the objective determining the control α becomes

$$\tilde{I}(\theta, \alpha) = \underbrace{\mathbb{E}_\theta \left[\underbrace{\mathbb{E}_X \left[\int_0^{t_f} \frac{(\partial_\theta f(x_t, \theta, \alpha))^2}{\beta^2}\, \mathrm{d}t\right]}_{\text{average over trajectories}} \right]}_{\text{average over prior}} \tag{20}$$

and then they show that the estimator so obtained, i.e. the one which uses the optimal α, is still better than a naive estimator (without any control)

B Aside - an intuition check

This is in reply to Susanne's suggestion that the 'optimal' thing to do is likely to be to stimulate maximally $\alpha = \alpha_{\max}$.

Consider in the simplest case, the linear deterministic ODE:

$$\dot{x} = \alpha - \beta x; \quad x_0 = 0$$

then
$$x(t) = \frac{\alpha}{\beta}(1 - \exp(-\beta t))$$
which, of course, goes to α/β in the long run. Assume $\alpha/\beta > 1$. Then the time $t_{\rm sp}$ to reach $x = x_{th} = 1$ is given by
$$t_{\rm sp}(\tau; \alpha) = -\frac{1}{\beta}\log(1 - \frac{\beta}{\alpha})$$
Thus if we know α and $t_{\rm sp}$ we can determine β. Suppose we could choose α. What would be the value of α that would make λ 'most identifiable'?

Let us equate 'identifiability' with the magnitude of the derivative of $t_{\rm sp}$ wrt. β.
$${t_{\rm sp}}'(\beta) = \frac{\alpha}{\beta(\alpha - \beta)} + \frac{1}{\beta^2}\log(1 - \frac{\beta}{\alpha})$$
Let us check the asymptotics:
$$\lim_{\alpha\uparrow\infty} {t_{\rm sp}}'(\beta) = \frac{1}{\beta}$$
and
$$\lim_{\alpha\downarrow\beta} {t_{\rm sp}}'(\beta) = \infty$$
As I read this, this means that the 'best' thing to do is let $\alpha \approx \beta$.

Now of course, in the noisy case,
$$dX = (\alpha - \frac{X}{\tau})\,\mathrm{d}t + \sigma dW$$
things might not be so simple..., but it does raise the possibility that the best thing to do if you want to identify τ is *not* to excite maximally, $\alpha \rightarrow \infty$, but to excite *critically.*

C Mutual Info calculation

Here we show why eq. (6) for the Mutual Information agrees with the usual definition of the Mutual Information, which for the random variables, X, θ is
$$I(X, \theta) = \int_\Theta \int_X p(x, \theta) \cdot \log\left(\frac{p(x, \theta)}{p(x)p(\theta)}\right) \mathrm{d}x\,\mathrm{d}\theta \tag{21}$$
First of all, the marginal distribution, $p(\theta)$, is just the prior of θ,
$$p(\theta) = \rho(\theta)$$
The joint distribution is
$$p(x, y) = L(x|\theta)\rho(\theta)$$
while the x marginal is
$$p(x) = \int_\Theta L(x|\theta)\rho(\theta)\,\mathrm{d}\theta$$
Plugging the three expressions into the definition in eq. (21) gives:
$$I = \int_\Theta \int_X L(x|\theta)\rho(\theta) \cdot \log\left(\frac{L(x|\theta)\rho(\theta)}{\int_\Theta L(x|\theta)\rho(\theta)\,\mathrm{d}\theta \cdot \rho(\theta)}\right) \mathrm{d}x\,\mathrm{d}\theta. \tag{22}$$
And after canceling $\rho(\theta)$ inside the log, we get eq. (6) .

D Why should we maximize the Mutual Information in the first place

Here we take a step back and discuss some properties of the Mutual Information Functional in order to justify using it as an objective for Optimal Design.

D.1 2013 Math Psych paper: [9]

We start with a tutorial paper from the Journal of Math. Psychology (2013), which begins as follows: *"Imagine an experiment in which each and every stimulus was custom tailored to be maximally informative about the question of interest, so that there were no wasted trials, participants, or redundant data points."*

Their work focuses on both parameter estimation and the larger task of model selection, let's just discuss the parameter estimation bit.

In their terminology, the task of the *experimenter* is to find a *design*, d (this is the stimulation $\alpha(t)$ in our context) that will best facilitate the estimation of the model parameters. After introducing a prior on the parameters $\rho(\theta)$, and denoting the *outcome* of the experiment as t, they state that the design selection can be formalized by optimizing the following expression:

$$d^* = \arg\max_d \int\int u(\theta, t; d) L(t|\theta, d)\rho(\theta)\,\mathrm{d}t\,\mathrm{d}\theta$$

where u is some utility function. For example one could maximize the inverse sum of CVs:

$$u() = \sum_i \frac{\mathbb{E}[\theta_i]}{STD[\theta_i]}$$

where $\mathbb{E}[\theta_i], STD[\theta_i]$ are the posterior mean/std. dev of the estimates, e.g.

$$\mathbb{E}[\theta_i] = \int_{\theta_i} \theta_i p(\theta_i|t, d)\,\mathrm{d}\theta_i = \int_{\theta_i} \theta_i \frac{L(t|\theta, d)\rho(\theta)}{\int_\theta L(t|\theta, d)\rho(\theta)\,\mathrm{d}\theta_i}\,\mathrm{d}\theta_i$$

(An aside, if the parameter estimate is on average zero, or negative I think this will not work at all, but I think they just suggest it as an obvious, but problematic utility function)

They then state that the most common utility function, u, in the literature is the Mutual Information b/w the random variables Θ and T.

Again, quote, *(The Mutual Information) measures the reduction in uncertainty about the values of the parameters that would be provided by the observation of an experimental outcome under design d. In other words, the optimal design is the one that extracts the maximum information about the model's parameters.*

So as far as I can tell, the fact that Mutual Information maximization equates to reduction of parameter uncertainty is here taken dogmatically!

They then go on to talk about *sequential* optimal design, which is the same as above, but then you Bayes-update the prior of θ while the experiment continues and then roll forward.

In our context this could mean that after $N_{s,1}$ spikes, we update the prior of β and redo the calculation of the optimal perturbation $\alpha(t)$, then observe for N_{s_2} spikes, then recompute $\alpha(t)$ and so on. At this point, one can get creative - instead of optimizing $\alpha(t)$ we can just take 1(or 2 or whatever) steps in the gradient descent. In practice, this is probably not very realistic due to computational costs, but in principle very cool.

D.2 Mackay book on Information Theory, Entropy, Mutual Information

The classic on Info Theory is probably, [2], and a popular alternative is [8]. Here we give a quick summary of Chs 2,3 in [8] on Probability, Entropy and INference. HOpefully these will give a more rigorous explanation for why maximizing Mutual Information is a 'good' thing.

Start from first principles, the information content of an outcome, x (of a R.V. X) is measured as:

$$h(x) = -\log_2 p(x)$$

And the entropy of an ensemble is

$$H(X) = -\sum p(x) \log_2 p(x)$$

Mackay postulates that 'entropy' is synonymous with 'uncertainty'

Joint entropy is

$$H(X, Y) = \sum p(x, y) \log_2 p(X, Y)$$

and for independent RVs it is additive (iff).

$$H(X, Y) = H(X) + H(Y) \Leftrightarrow P(X)P(Y) = P(X, Y)$$

There is then the KL divergence and the Gibbs inequality:

$$D_{KL}(P||Q) = \sum p(x) \log \frac{p(x)}{q(x)}$$
$$D_{KL}(P||Q) \geq 0$$

with equality iff $p \equiv q$

Then there is a long section on Shannon Entropy...showing how it is a very sensible measure of the informtion in an experiment / outcome.

IN Ch. 8. Mackay says that the mutual info, $I(X, Y)$ represents the average reduction in uncertainty about x that results from learning the value of y or vice versa!

Since

$$I(\Theta, T) = H(\Theta) - H(\Theta|T)$$

and $H(\Theta)$ in our case is indepenednet of both α and the observed spike times T, we could alternatively maximize $-H(\Theta|T)$, i.e.

$$\alpha* = \arg\max \int_{\Theta, T} L(t|\theta)\rho(\theta) \cdot \log\left(\frac{L(t|\theta)\rho(\theta)}{\int L(t|\theta)\rho(\theta)\,\mathrm{d}\theta}\right)\,\mathrm{d}t\,\mathrm{d}\theta$$

Never mind, that's a tautology. This expression is equivalent to what we had earlier (of course) and does not seem to be any more convenient.

D.3 Further Literature Review

In principle OPtimal Design for Dynamic Systems is a subset of System Identification, which is a very researched topic in the control community. Here's a review paper [3] from the Control Theory Community. They talk about things like Transfer Functions etc. usually things are in Discrete time. There is no notion of SDEs although there is a=often a white noise error term, so in a sense these are discretized SDEs, although without ay of the formalism of Ito Calculus etc.

The papers on Adaptive Optimal Design from the Myung et al group in Ohio is well cited (the main 2010 [1] article has 45 citations)

h, cool, [1] says in its lit. review leading into the paper that

> the desirability and usefulness of (using Mutual Information as the objective functional) was formally justified by Paninski (2005) who proved that under acceptably weak modeling conditions, the adaptive approach with a utility function based on mutual information leads to consistent and efficient parameter estimates.

Ok, so we just need to dig up Paninsky2005 and insert here:) Here it is: [10]. Here is its abridged abstract:

> ... on any given trial, we want to adaptively choose the input in such a way that the mutual information between the (unknown) state of the system and the (stochastic) output is maximal, given any prior information (including data collected on any previous trials). We prove a theorem that quantifies the effectiveness of this strategy... and demonstrate that this method is in a well-defined sense never less efficientand is generically more efficientthan the non-adaptive strategy...

Here's another good quote from

> ...several attempts have been made to devise algorithms to find the optimal stimulus of a neuron, where optimality is defined in terms of firing rate (Tzanakou, Michalak, Harth, 1979; Nelken, Prut, Vaadia, Abeles, 1994; Foldiak, 2001), but we should emphasize that the two concepts of optimality are not related in general and turn out to be typically at odds (maximizing the firing rate of a cell does not maximize and in fact often minimizesthe amount we can expect to learn about the cell; see sections 3 and 4).

Again the punch line in [10] is:

> Our main result (in section 2) states that under acceptably weak conditions on the models $p(t|\alpha, \theta)$ (our notation, not his) the information maximization strategy leads to consistent and efficient estimates of the true underlying model, in a natural sense. In particular, the information maximization strategy is never less efficient, in a well-defined senseand is generically more efficientthan the simpler, non-adaptive, i.i.d. x strategy.

The experimental/computational Lewi, Buttera et Paninsky paper from 2009 is very popular, " Sequential optimal design of neurophysiology experiments", [5]. It has 64 citations, although its main citing authors are Paninsky himself (self-references) + the Myung et Cavagnaro group above :). Another very popular paper from Paninsky (117 cites) is [11], which deals with several things including 'Optimal Stimulus'. Reading Paninsky can be pleasant or unpleasant, the 2009 [5]paper is 73 pages of dense Convex Optimization, which hurts my eyes:) but the 2006 paper is much more fluent, at only 14 pages.

Let's review it and in paritcular let's review the part about 'Optimal Stimulus', this starts in Sec. 3 (p.11)

> If we use the entropy of the posterior distribution on the model parameters (what we call $L(t|\theta)\rho(\theta)$) to quantify this uncertainty, we arrive at the [. . .] mutual information between the response (for us this is t_s) and the model parameters θ given the stimulus and past data.

They then acknowledge that optimizing I is problematic, and even just computing it quickly becomes hard for multi-dimensional params. . .

They then suggest to approximate the posterior of θ with a Gaussian. They self-quote ([10], which seems to be the main theoretical justification paper for using MI).

With that the Mutual Information calculation is reduced to determinig the log-Determinant of the Hessian matrix (the Gaussian Covariance). . .

They then do a very smart observation that for online optimization this Hessian recuses very nicely with itself over observations and so updating it is very cheap (computationally). Thus one can optimize-stimulate-update online, meaning;

1. Optimize the MI for a given param prior
2. stimulate the system with the MI-optimal stimulus
3. update the prior given the new observation
4. Reoptimize the MI using the new prior (the posterior)
5. roll on. . .

OK! Now I understand the rough idea of the 76 page 2009 paper[5].

D.4 Cool (Waterloo) Physics paper

[4] describe something that is in principle very close to what we aim at. They are trying to estimate the 'Hamiltonian' of a quantum experiment (its parameters) and they talk about all the things we talk about - optimal stimulus, belief distn over the params. In Algos, 1-7 of their paper they give a very readable summary of particle filtering! (which in turn sources [7])

References

[1] Daniel R Cavagnaro, Jay I Myung, Mark A Pitt, and Janne V Kujala. Adaptive Design Optimization: A Mutual Information-Based Approach to Model Discrimination in Cognitive Science. 905:887–905, 2010.

[2] Thomas Cover and Joy Thomas. Elements of Information Theory. Wiley-Interscience; 2 edition, 2006 edition, 2006.

[3] Michel Gevers, Xavier Bombois, Roland Hildebrand, and Gabriel Solari. Optimal experiment design for open and closed-loop system identification. 11(3):197–224, 2011.

[4] Christopher E Granade, Christopher Ferrie, Nathan Wiebe, and D G Cory. Robust online Hamiltonian learning. New Journal of Physics, 14(10):103013, October 2012.

[5] Jeremy Lewi, Robert Butera, and Liam Paninski. Sequential optimal design of neurophysiology experiments. Neural computation, 21(3):619–87, March 2009.

[6] Kevin K Lin, Giles Hooker, and Bruce Rogers. Control Theory and Experimental Design in Diffusion Processes. pages 1–23.

[7] J Liu and M West. Combined parameter and state estimation in simulation-based filtering. Sequential Monte Carlo methods in practice, 2001.

[8] David J. C. MacKay. Information Theory, Inference and Learning Algorithms. Cambridge University Press, 2003.

[9] Jay I Myung, Daniel R Cavagnaro, and Mark a Pitt. A Tutorial on Adaptive Design Optimization. Journal of mathematical psychology, 57(3-4):53–67, January 2013.

[10] Liam Paninski. Asymptotic Theory of Information-Theoretic Experimental Design. 1507:1480–1507, 2005.

[11] Liam Paninski, Jonathan Pillow, and Jeremy Lewi. Statistical models for neural encoding, decoding, and optimal stimulus design. pages 1–14, 2006.