Optimal Design for estimation in stochastic LIF models

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

January 14, 2014

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

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:

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

With (a subset of) the parameter set $\theta = \{\mu, \tau_c, \beta\}$ unknown.

Our goal is to choose $\alpha(t)$ as to estimate τ_c. As always we can consider two different contexts:

- X_t is continuously observed.
- Only the spike times $\{t_k\}$ are observed

The first scenario is addressed (for the Morris-Lecar model) in a submitted paper, [**?**] on arXiv, where I got this idea, the second scenario is likely unaddressed in the literature. I say *likely* as it is a 'harder' problem than the first scenario and the first scenario is only now being addressed in [**?**] which claims to be one of the first optimal design papers using stochastic control.

The big idea is that we combine estimation and control into a single task - *control for the benefits of estimation*. This is called *optimal design*.

2 Spike-Only Optimal Design - Exploration

We consider the challenging case of optimal design where we only observe the spikes.

First we need some notation for the probability density of the nth spike, conditional on some applied control α:

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

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

$$F(x,t) := \mathbb{P}\left[X_t < x \,|\, X_0 = 0, X_{s<t} < 1\right] \quad \text{(transition distribution)} \tag{3}$$

which follows a Fokker-Planck PDE:

$$\begin{aligned}\partial_t F(x,t) &= \underbrace{\frac{\beta^2}{2}}_{D}\cdot\partial_x^2 F + \underbrace{\left(\frac{x}{\tau_c}-\alpha(t)-\mu\right)}_{U(x,t)}\cdot\partial_x F \\ &= D\cdot\partial_x^2 F + U(x,t)\cdot\partial_x F \\ &= \mathcal{L}_\theta[F]\end{aligned} \tag{4}$$

$$\begin{cases} F(x,0) &= \text{Heavyside}(x) \\ F(x,t)|_{x=x_-} &\equiv 0 \\ \partial_x F(x,t)|_{x=v_{\text{th}}} &\equiv 0. \end{cases}$$

The spike-time distribution is related to the transition distribution via

$$\bar{G}(t) = F(1,t)$$

and the density follows via

$$g(t) = -\partial_t \bar{G}(t) = -\partial_t F(1,t).$$

In a typical (maximum likelihood) estimation experiment, we will see a lot of spikes and form the likelihood as

$$L(\theta|t_n) = \prod_n g_n(t_n)$$

We will then take logs and proceed as usual:

$$l(\theta|t_n) = \sum_n \log(g_n(t_n)) = \sum_n \log(-\partial_t F(1,t_n))$$

and then maximize l over the parameters θ.

The associated *score* function is

$$S(\theta|t_{\text{sp}}) = \nabla_\theta l(\theta|t_{\text{sp}})$$

The score function is a vector[1].

The typical Maximum Likelihood process is to maximize the likelihood, l or, if one uses a gradient-based approach, to find the roots of the score, S.

The Fisher Information can be thought of as the expected negative Hessian of the Likelihood, where the expectation is taken wrt. the random variable, i.e. the spike time

$$\Phi = -\,\mathbb{E}\left[\,\partial_{\theta_i,\theta_j} l(\theta|t_{\text{sp}})\,\right] = \int_0^\infty \left[\partial_{\theta_i,\theta_j} l(\theta|t)\right]\cdot g(t)\,\mathrm{d}t$$

Up to some regularity conditions, which we will assume hold, the Fisher Info, Φ, is also the second moment of the (log-)likelihood

$$\Phi = \mathbb{E}[\partial_{\theta_i} l()\cdot\partial_{\theta_j} l()]$$

The Fisher Info, Φ, is a matrix.

[1] We write ∇ for the vector differential and ∂ for its scalar components, i.e. $\nabla_\theta = [\partial_{\theta_1},\dots\partial_{\theta_i}],\dots$

Thinking of the Fisher Information as a (expected) Hessian of the likelihood, it is easy to see how a 'large' Φ implies lots of curvature, meaning that the maximum is easily found, as opposed to a 'small' curvature which implies a shallow maximum and many likely candidates for the parameter set θ.

The idea of optimal design is to choose the control, here $\alpha(t)$ such that Φ is maximized. Since Φ is a matrix, one actually maximizes its trace or determinant. The determinant is the commonly used as it is related to the volume of the 'variance ellipsoid'.

Recall that both the trace and determinant are related to the eigenvalues, λ_i, via

$$\det(\Phi) = \prod_i \lambda_i$$

$$tr(\Phi) = \sum_i \lambda_i$$

Naturally, the trace is easiest since it is just

$$tr(\Phi) = \sum_i \mathbb{E}[(\partial_{\theta_i} L(\theta))^2]$$

For pedagogical reasons, let us focus on the case of just one parameter, say the relaxation time, $\theta = \{\tau_c\}$. Thus we don't have to deal with eigenvalues, just the maximization of the negative curvature or equivalently of the score's second moment.

$$\Phi(\alpha) = \mathbb{E}\left[\partial^2_{\tau_c} l(\tau_c|t_{\text{sp}})\right]$$

or

$$\Phi(\alpha) = \mathbb{E}\left[\left(\partial_{\tau_c} l(\tau_c|t_{\text{sp}})\right)^2\right]$$

In fact, it is more convenient to work with the reciprocal

$$\lambda_c = \frac{1}{\tau_c}$$

because eq. (23) is linear in the parameter, λ_c

Now given the relation between the likelihood, l and the transition distribution, $g = -\partial_t F$, we can write the Fisher Info, Φ, explicitly as

$$\Phi[\alpha(\cdot)] = \int_0^\infty \left(\partial_{\lambda_c}[\log(g(s))]\right)^2 \cdot g(s)\,\mathrm{d}s \tag{5a}$$

$$= -\int_0^\infty \left(\partial_{\lambda_c}[\log(-\partial_t F(x_{th}, s))]\right)^2 \cdot \partial_t F(x_{th}, s)\,\mathrm{d}s \tag{5b}$$

We have used the 2nd moment, rather than the second derivative form of Φ. We will see later why using the 2nd moment is more convenient.

We want to find the control input α, which maximizes Φ in eq. (5b). For a *given* $\lambda_c = 1/\tau_c$ that is straight-forward, if numerically challenging. But the real problem is that we are trying to *estimate* τ_c and thus we do not know its value.

The simplest thing to do is to use the estimate for τ_c, optimize Φ, apply the control and roll on. Let us talk about this now, but we MUST bear in bind that this separation ansatz, estimate, then optimize, then estimate again, is not necessarily correct/good/best.

2.1 the nitty-gritty analysis of the Optimal Control, α

Let us discuss the optimization problem - how to maximize Φ given in eq. (5b).

Equation (5b) is a functional, of the transition distribution F, involving first order *sensitivities* of the function F. Let's use F_1 to denote this sensitivity, i.e. $F_1 = \partial_{\lambda_c} F$. Sometimes, to be explicit, we will write the base transition as $F = F_0$ to distinguish it from its sensitivity.

Performing the differentiation wrt. λ_c in eq. (5b) then gives:

$$\begin{aligned}\Phi[\alpha(\cdot)] &= -\int_0^\infty \left(\partial_{\lambda_c}[\log(-\partial_t F(x_{th}, s))]\right)^2 \cdot \partial_t F(x_{th}, s)\,\mathrm{d}s \\ &= -\int_0^\infty \left(\frac{\partial_t F_1(x_{th}, s)}{\partial_t F_0(x_{th}, s)}\right)^2 \cdot \partial_t F_0(x_{th}, s)\,\mathrm{d}s \end{aligned} \tag{6}$$

Where we have assumed that differentiating wrt. t, λ_c commute.

The next step is to obtain a PDE for the sensitivity, F_1. This is done by differentiating the PDE for the transition, F wrt. λ_c. Differentiating wrt. λ in eq. (23) and applying the product rule to terms which contain λ, i.e. the λx term in the drift U, we get

$$\begin{aligned}\partial_t F_1(x,t) &= D \cdot \partial_x^2 F_1 + U \cdot \partial_x F_1 + x \cdot \partial_x F_0 \\ &= \mathcal{L}_1[F_1] \\ &\begin{cases} F_1|_{t=0} &= 0 \\ F_1|_{x=x_-} &= 0 \\ \partial_x F_1|_{x=x_{th}} &= 0. \end{cases}\end{aligned} \tag{7}$$

Note that F_0 appears in the evolution equation for F_1, like a source term.

We can now see why using the 2nd moment form for Φ is more convenient than using the 2nd derivative form - using the 2nd derivative form would require us to take second-order sensitivities, i.e. to take the sensitivity wrt. λ of F_1 in eq. (7) and thus we will have to solve yet another PDE (for F_2).

Equation (6) is the basic objective equation from which we would like to try to obtain the optimal input $\alpha(\cdot)$, given some nominal value for τ_c or equivalently $\lambda = 1/\tau_c$. However into a problem, since the objective is given in terms not only of the states F_0, F_1, but also their time derivatives $\partial_t F_0, \partial_t F_1$

What we must do is similar to the situation in ODEs, when we have a higher-order ODE and we must reduce it to a first-order - Introduce More States!

2.1.1 Introducing more states

We now introduce two new states, the time derivatives of F_0, F_1

$$H_i = \partial_t F_i$$

Just like F_1 satisfies a PDE, related to the PDE of $F = F_0$, so will the H_i's satisfy PDEs related to the PDEs of F_i.

Let us first discuss the time-derivative H_0 of the base distribution, F_0.

First recall the evolution equation for F_0, eq. (23). Then we derive the evolution equation for H as

$$\begin{aligned}\partial_t H =& \partial_t \partial_t F \\ =& \partial_t [D \cdot \partial_x^2 F + U(x,t) \cdot \partial_x F]\end{aligned}$$

Now assume ∂_t, ∂_x commute, and move the ∂_t through the expression on the right.

$$\begin{aligned}=& D \cdot \partial_x^2 \partial_t F + U(x,t) \cdot \partial_x \partial_t F + \partial_t U(x,t) \cdot \partial_x F \\ \implies \partial_t H =& D \cdot \partial_x^2 H + U(x,t) \cdot \partial_x H + \partial_t U(x,t) \cdot \partial_x F\end{aligned}$$

Now we reach another complication. The time derivative of the drift, $\partial_t U(x,t)$ also involves the time derivative of the control $\alpha(t)$, so we must assume that the control has a proper time-derivative *in some sense*, but we have to be very careful if the optimal control turns out to be discontinuous in time, say a bang-bang type controls.

Well, this is the fun part - let's proceed assuming that we can differentiate α. The time derivative of U is:

$$\begin{aligned}\partial_t U(x,t) = \dot{U} =& \partial_t [-\frac{x}{\tau_c} - \alpha(t) - \mu] \\ =& -\frac{d\alpha(t)}{dt} = \dot{\alpha}\end{aligned}$$

for short we will write the time derivative of the control as $\dot{\alpha}$ and $\dot{U}$ as the time-derivative of the velocity field, $\dot{U} = -\dot{\alpha}$

The boundary conditions for $H = \partial_t F$ are fairly straightforward, since the BCs for F itself are time-independent.

The initial conditions however are not so obvious! A simple workaround is to solve, numerically or o/w, the equation for F, eq. (23) and then just take finite-differences to obtain $H_0(x, t = 0)$.

So we can now state the full evolution equation for H:

$$\begin{aligned}\partial_t H &= D \cdot \partial_x^2 H + U(x,t) \cdot \partial_x H + \dot{U}(x,t) \cdot \partial_x F \\ &\begin{cases} H(x,0) & = & \partial_t F(x,t)|_{t=0} \\ H(x,t)|_{x=x_-} & \equiv & 0 \\ \partial_x H(x,t)|_{x=x_{th}} & \equiv & 0. \end{cases}\end{aligned} \tag{8}$$

Similarly, we can obtain the evolution equation for the time-derivative of the sensitivity, $H_1 = \partial_t F_1$. Starting from eq. (7), we will get:

$$\begin{aligned}\partial_t H_1(x,t) &= D \cdot \partial_x^2 H_1 + U \cdot \partial_x H_1 + \dot{U} \cdot \partial_x F_1 + x \cdot \partial_x H_0 \\ &\begin{cases} H_1|_{t=0} & = & \partial_t F_1(x,t)|_{t=0} \\ H_1|_{x=x_-} & = & 0 \\ \partial_x H_1|_{x=x_{th}} & = & 0. \end{cases}\end{aligned} \tag{9}$$

Now we can state the objective, Φ from eq. (6) in terms of H_0, H_1:

$$\Phi[\alpha(\cdot)] = -\int_0^\infty \left(\frac{H_1(x_{th}, s)}{H_0(x_{th}, s)}\right)^2 \cdot H_0(x_{th}, s)\,\mathrm{d}s \tag{10}$$

Note that although the states, F_0, F_1 do not directly appear in the objective for Φ, we still need to keep track of them as they appear in the evolution equations for H_0, H_1. There is one more state we need to introduce and that is the control. Since we now have time-derivatives of the control $\dot{\alpha}$, our control is no longer going to be α itself, but $u = \dot{\alpha}$.

So we have a new state, α which is related to the control as

$$\partial_t \alpha = u(t)$$

This way we only have states (and not their time-derivatives)

$$\Phi[u(\cdot)] = -\int_0^\infty \left(\frac{H_1(x_{th}, s)}{H_0(x_{th}, s)}\right)^2 \cdot H_0(x_{th}, s)\,\mathrm{d}s \tag{11}$$

2.1.2 An Aside: a moment of sanity

So far, we have effectively introduced four states, F_0, F_1, H_0, H_1 (ignoring the control state), each with its own PDE. And all that for only one sensitivity, that wrt. $\lambda = 1/\tau_c$. If we have N sensitivites, going as above will involve, $2N + 2$ states/PDEs. To form a objective differential, $\delta\Phi$ will involve a co-state for each state. Thus to calculate one input signal $\alpha(t)$, for one spike, will require $4N + 4$ PDEs in total. In the case that we are trying to solve for all $N = 3$ parameters in eq. (1) at the same time, we need to solve, 16 PDEs to just to evaluate the differential. Assuming that convergence happens in 5 iterations of the gradient descent, an optimistic estimate, - that means that in order to calculate $\alpha(t)$ we need to solve 100 1-d PDEs. Even if we take an approach that we identify one parameter at a time, that still means $8 \cdot 5$ PDE solves to form $\alpha(t)$ - one needs to wonder whether perhaps that is not 'a little too much'?

2.1.3 Augmenting the objective

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

First we augment our functional with the dynamics:

$$\Phi = \int_0^\infty \frac{\Big[H_1(x_{th}, s)\Big]^2}{H_0(x_{th}, s)}\,\mathrm{d}s \tag{12}$$

$$- \int_0^\infty < p_0, (\partial_t F_0 - \mathcal{L}[F_0]) > \,\mathrm{d}s \tag{13}$$

$$- \int_0^\infty < p_1, (\partial_t F_1 - \mathcal{L}[F_1]) > \,\mathrm{d}s \tag{14}$$

$$- \int_0^\infty < q_0, (\partial_t H_0 - \mathcal{L}[H_0]) > \,\mathrm{d}s \tag{15}$$

$$- \int_0^\infty < q_1, (\partial_t F_1 - \mathcal{L}[F_1]) > \,\mathrm{d}s \tag{16}$$

$$- \int_0^\infty < z, \dot{\alpha} - u > \,\mathrm{d}s \tag{17}$$

where the inner product, $< f, g >$ is just the space integral $\int f \cdot g \,\mathrm{d}x$

We use the generic $\mathcal{L}$ even though, of course, the spatial differential operator is different for each F_i, H_i as given explicitly in eqs. (7) to (9) and (23).

Before we proceed with taking the differential of Φ, we need to take the adjoints, so that we only have terms involving F, H and not their spatial or time derivatives. That is we need to integrate-by-parts in order to move all the partials from F, H to p, q

Let us give the example of how that is done for the first pair, p_0, F_0.

$$\begin{aligned}
- \int_0^\infty < p_0, (\partial_t F_0 - \mathcal{L}[F_0]) > \,\mathrm{d}s &= - \int_0^\infty < p_0, (\partial_t F_0 - D \cdot \partial_x^2 F_0 - U \cdot \partial_x F_0 > \,\mathrm{d}s \\
&= \int_0^\infty < \partial_t p_0, F_0 > \,\mathrm{d}s + < \partial_t p_0, F_0 > \Big|_{t=0}^{T} \\
&\quad + \int_0^\infty < \partial_x^2 [D p_0] - \partial_x [U p_0], F_0 > \,\mathrm{d}s \\
&\quad + \int_0^\infty \Big(U p_0 \cdot F_0 - \partial_x [D p_0] \cdot F_0 + D p_0 \cdot \partial_x F_0 \Big|_{x=x_-}^{x_{th}} \,\mathrm{d}s
\end{aligned}$$

The whole point is to have only terms involving F in there. So terms that have $\partial_x F$, for example $Dp \cdot \partial_x F$ in the BCs, will have to be removed, by posing appropriate BCs on p. However, the case of p_0, F_0 is the simplest as it has no coupling with the other states F_1, H_i. On the other hand, all the others have coupling. That is, one of the other states comes up in their evolution equations (the $\mathcal{L}$). So we should show how the integration-by-parts works for all three cases.

Next is, p_1, F_1, the only difference here is that the spatial operator on

F_1 has an additional term of the form $x \cdot \partial_x F_0$

$$
\begin{aligned}
-\int_0^\infty < p_1, (\partial_t F_1 - \mathcal{L}[F_1]) > \, \mathrm{d}s &= -\int_0^\infty < p_0, (\partial_t F_1 - D \cdot \partial_x^2 F_1 - U \cdot \partial_x F_1 - x \cdot \partial_x F_0 > \, \mathrm{d}s \\
&= \int_0^\infty < \partial_t p_1, F_1 > \, \mathrm{d}s + < \partial_t p_1, F_1 > \Big|_{t=0}^{T} \\
&\quad + \int_0^\infty < \partial_x^2 [D p_1] - \partial_x [U p_1], F_1 > - < \partial_x [x p_1], F_0 > \, \mathrm{d}s \\
&\quad + \int_0^\infty \Big(U p_1 \cdot F_1 - \partial_x [D p_1] \cdot F_1 + D p_1 \cdot \partial_x F_1 + x p_1 F_0 \Big|_{x=x_-}^{x_{th}} \, \mathrm{d}s
\end{aligned}
$$

Similarly we can do for q_i, H_i

The whole goal of this exercise is to calculate the differential of Φ in eq. (11), wrt. the control u, i.e. to calculate $\delta\Phi/\delta u$. First, the differential of Φ in terms of all the other states F_i, H_i, α is:

$$
\begin{aligned}
\delta\Phi = &-\int_0^\infty 2\frac{H_1}{H_0} \cdot \delta H_1 - \frac{H_1^2}{H_0^2} \cdot \delta H_0 \, \mathrm{d}s \\
&+ \dots
\end{aligned}
$$

where the dots signify all the terms that will come from the adjoint terms. We will choose p, q to satisfy PDEs exactly so that the coefficients in front of $\delta F_i, H_i$ are zero. Whatever remains will be the differential of Φ wrt. the control u. Implementing a gradient ascent strategy we can climb to the top of Φ...

3 Big Questions

It is clear that even for fixed parameters, calculating the optimal control $\alpha^*(\cdot)$ via gradient ascent is very difficult. It is highly non-trivial and very error-prone to calculate the differential $\delta\Phi$. And once the analysis is done the numerics will still involve numerically calculating up to 100 parabolic PDEs.

It is difficult to guess beforehand what the added value of all this, meaning how much better estimates will be obtained with the optimal control α^* compared to the pure-observation case $\alpha = 0$.

Moreover there is the very important detail that the parameters are NOT known - they are being estimated - as such it is unclear whether using the optimal control obtained from 'nominal' parameters is at all useful in estimating the 'real' parameters.

The combination of high analytical / computational difficulty in obtaining the optimal control and the fact that it is not clear whether all this effort will be of any value, is stopping me from exploring the topic further...

4 The Bayesian Approach

Let's try something else. Suppose we consider the parameters, as probability densities. Say

$$\rho(\lambda) = \mathbb{P}[\lambda_c = \lambda] \tag{18}$$

Given an observation I_n, then the Bayes-rule update prior $\to$ posterior looks like

$$\rho(\lambda|I_n) = \frac{g(I_n|\lambda)\cdot\rho(\lambda)}{\int g(I_n|\lambda)\cdot\rho(\lambda)\,\mathrm{d}\lambda} \tag{19}$$

This way, having observed an interval I_n, we can recalculate the posterior-distribution, $\rho(\lambda|I_n)$.

A natural goal is to seek a control α which will minimize the variance of the posterior ρ.

$$\mathbb{V}\mathrm{ar}[\rho; i_n] = \int \lambda^2 \rho(\lambda|i_n)\,\mathrm{d}\lambda - \left(\int \lambda\rho(\lambda|i_n)\,\mathrm{d}\lambda\right)^2$$

That is we need to choose the control α before observing i_n and so we could, ostensibly, choose α as the optimal control α^* st.

$$\alpha^* = \arg\min \mathbb{E}[\mathbb{V}\mathrm{ar}[\rho; i_n]]$$

where the expectation is taken with respect to the marginal density of g i.e.

$$\mathbb{E}[\mathbb{V}\mathrm{ar}[\rho; i_n]] = \int \mathbb{V}\mathrm{ar}[\rho; i_n] g(i)\,\mathrm{d}i$$

And in turn the marginal, $g(i)$ is given with respect to the prior

$$g(i) = \int g(I_n|\lambda)\cdot\rho(\lambda)\,\mathrm{d}\lambda$$

So that the optimal control is to minimize the expected variance, which in all gory details reads as

$$J[\alpha(\cdot)] = \int \left[\int \lambda^2 \frac{g(i|\lambda)\cdot\rho(\lambda)}{\int g(i|\lambda)\cdot\rho(\lambda)\,\mathrm{d}\lambda}\,\mathrm{d}\lambda - \left(\int \lambda \frac{g(i|\lambda)\cdot\rho(\lambda)}{\int g(i|\lambda)\cdot\rho(\lambda)\,\mathrm{d}\lambda}\,\mathrm{d}\lambda\right)^2\right]\cdot\left[\int g(i|\lambda)\cdot\rho(\lambda)\,\mathrm{d}\lambda\right]\mathrm{d}i \tag{20}$$

where the prior ρ is known before the nth interval and $g(i|\lambda)$ is found by solving a PDE...

well, great!!!

Imagine trying to compute the differential δJ with respect to α in eq. (20). There are integrals within integrals within integrals... Sigh...

Note that minimizing the variance is not the only reasonable objective, one could try to minimize the entropy, H, of the posterior:

$$H[\rho; i_n] = \int \rho(\lambda)\log\rho(\lambda)\,\mathrm{d}\lambda$$

but this has the same conceptual difficulties as minimizing the (expected) variance. that is we still need to evaluate a lot of PDEs, and a lot of

integrals..., except that J will look like:

$$J[\alpha(\cdot)] = \int \Bigg[\underbrace{\int \frac{g(i|\lambda)\cdot\rho(\lambda)}{\int g(i|\lambda)\cdot\rho(\lambda)\,\mathrm{d}\lambda} \cdot \log\left(\frac{g(i|\lambda)\cdot\rho(\lambda)}{\int g(i|\lambda)\cdot\rho(\lambda)\,\mathrm{d}\lambda}\right)\mathrm{d}\lambda}_{H(\rho,i)} \Bigg] \cdot \Bigg[\underbrace{\int g(i|\lambda)\cdot\rho(\lambda)\,\mathrm{d}\lambda}_{g(i)} \Bigg]\,\mathrm{d}i \tag{21}$$

Well, this can be somewhat simplified. First we can take $g(i)$ out of the inner $\mathrm{d}\lambda$ integral and cxl it out

$$J[\alpha(\cdot)] = \int \Bigg[\underbrace{\int g(i|\lambda)\cdot\rho(\lambda)\cdot\log\left(\frac{g(i|\lambda)\cdot\rho(\lambda)}{\int g(i|\lambda)\cdot\rho(\lambda)\,\mathrm{d}\lambda}\right)\mathrm{d}\lambda}_{H(\rho,i)} \Bigg] \cdot \frac{\cancel{g(i)}}{\cancel{g(i)}}\,\mathrm{d}i$$

leaving

$$J[\alpha(\cdot)] = \int_I \int_\Lambda \left[g(i|\lambda)\cdot\rho(\lambda)\cdot\log\left(\frac{g(i|\lambda)\cdot\rho(\lambda)}{\int g(i|\lambda)\cdot\rho(\lambda)\,\mathrm{d}\lambda}\right)\right] \cdot\, \mathrm{d}\lambda\,\mathrm{d}i$$

But this is still a very complicated expression - especially because of the integral inside the log (the normalizing constant).

A possible avenue is to realize that after each observation i_n, ρ will not change much. So that

$$\rho(\lambda|i_n) \approx \rho(\lambda)$$

or equivalently that

$$\frac{\rho(\lambda|i_n)}{\rho(\lambda)} \approx 1$$

but I don't see how to use that. The most obvious way is to expand the log:

$$\log(\rho(i|\lambda) = \log(\rho(\lambda)) + \log(\frac{\rho(i|\lambda)}{\rho(\lambda)}) \approx \log(\rho(\lambda)) + \frac{\rho(i|\lambda)}{\rho(\lambda)} - 1)$$

and the objective looks like

$$J[\alpha(\cdot)] = \int_I \int_\Lambda g(i|\lambda)\cdot\rho(\lambda)\cdot\left[\log(\rho(\lambda)) + \frac{g(i|\lambda)}{\int g(i|\lambda)\cdot\rho(\lambda)\,\mathrm{d}\lambda} - 1\right]\cdot\,\mathrm{d}\lambda\,\mathrm{d}i$$

I don't know... is that more useful? As usual in Bayesian matters, the big problem is in that normalizing constant, $\int g(i|\lambda)\cdot\rho(\lambda)\,\mathrm{d}\lambda$. In any case, if we try to take the differential of the expression above, we can (?), but it will have exactly the same issue as before, where we have four states F_0, F_1, H_0, H_1, and we are calculating 4 PDEs... But maybe we can agree to do only one iteration of the algorithm. That is increment α only once and then apply this. Still, integrating over λ, means that we have to calculate F_0, F_1, H_0, H_1 for many λ's.

Ufff...!!!

This just doesn't 'feel' right!

One last, and speculative approach is to consider the K-L entropy from ρ to $\rho(\lambda|i)$.

$$KL(\rho, i) = \int_\Lambda \rho(\lambda) \cdot \log[\frac{\rho(\lambda|i)}{\rho(\lambda)}] \, \mathrm{d}\lambda$$

Let's say we want to maximize KL, that is we want to find $\rho(\lambda|i)$ that is maximally away (wrt. KL quasi-metric). While the variance and entropy objectives were fairly intuitive, this objective is much more speculative. But let's see...

Since $\rho(\lambda)$ does not depend on i, this is the same as minimizing

$$J = \int_I \Big[\int_\Lambda \rho(\lambda) \log \rho(\lambda|i) \, \mathrm{d}\lambda \int_\Lambda g(i|\lambda)\rho(\lambda) \, \mathrm{d}\lambda \Big] \, \mathrm{d}i$$

And of course, $\rho(\lambda|i)$ is related to g st.

$$J = \int_I \left[\int_\Lambda \rho(\lambda) \log \Big(\frac{g(i|\lambda)\rho(\lambda)}{\int_\Lambda g(i|\lambda)\rho(\lambda) \, \mathrm{d}\lambda} \Big) \, \mathrm{d}\lambda \cdot \int_\Lambda g(i|\lambda)\rho(\lambda) \, \mathrm{d}\lambda \right] \mathrm{d}i$$

5 Working with the transition density to reduce the number of states

When working with the transition distribution F, using $g = \partial_t F$ we were forced to also work with $\partial_t F = H$, i.e. we had to introduce another state and also another state for each sensitivity. However if we work with the density, this might not be necessary, lets see how:

$$f(x,t) := \partial_x F \quad \text{(transition density)} \tag{22}$$

which follows a Fokker-Planck PDE similar to eq. (23)

$$\begin{aligned}
\partial_t f(x,t) &= \underbrace{\frac{\beta^2}{2}}_{D} \cdot \partial_x^2 F + \partial_x \Bigg(\underbrace{\Big(\lambda_c x - \alpha(t) - \mu\Big)}_{U(x,t)} \cdot \partial_x f \Bigg) \\
&= D \cdot \partial_x^2 f + \partial_x \Big(U(x,t) \cdot f \Big) \\
&= -\partial_x \phi(x,t) \\
&\begin{cases} f(x,0) &= \delta(x) \\ D\partial_x f + U f|_{x=x_-} &\equiv 0 \\ f|_{x=x_{th}} &\equiv 0. \end{cases}
\end{aligned} \tag{23}$$

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

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

This means that the Fisher Info objective from eq. (5a) will look like:

$$\begin{aligned}
\Phi[\alpha(\cdot)] &= \int_0^\infty \Big(\partial_{\lambda_c} [\log(g(s))] \Big)^2 \cdot g(s)\, \mathrm{d}s \\
&= \int_0^\infty \Big(\partial_{\lambda_c} [\log(\phi(x_{th}, s))] \Big)^2 \cdot \phi(x_{th}, s)\, \mathrm{d}s \\
&= \int_0^\infty \left(\frac{D \cdot \partial_x f_1(x_{th}, s)}{D \cdot \partial_x f_0(x_{th}, s)} \right)^2 \cdot D \cdot \partial_x f_0(x_{th}, s)\, \mathrm{d}s
\end{aligned} \tag{24}$$

where as before $f_1 = \partial_{\lambda_c} f$ is the parameter sensitivity.

Well, the objective no longer has any ∂_t terms, but now there are a lot of ∂_x terms... of course δf being small, does not mean that $\partial_x \delta f$ is small... take for example $\sin(nx)/n$ which goes to zero as n increases, but whose derivative is always of order 1.

Let me try to explain in short. If the objective looked like so:

$$\Phi = \int \frac{1}{f}(x_{th}, s)$$

then its differential would be

$$\delta\Phi = -\int \frac{\delta f}{f^2(x_{th}, s)}$$

we would then introduce an adjoint state, p and use its boundary conditions to knock out the term $\frac{\delta f}{f^2(x_{th},s)}$. In the simplest case this would be by setting $p(x_{th},s) = 1/f^2$. But it depends on the dynamics/BCs of f.

However with the non-linearity involving the partial $\partial_x f$ in the integrand

$$\Phi = \int \frac{1}{\partial_x f(x_{th},s)}$$

then $\delta\Phi$ looks like:

$$\delta\Phi = -\int \frac{\partial_x \delta f}{(\partial_x f)^2 + \partial_x f \cdot \partial_x \delta f}$$

and it is hard to isolate $\delta_x f$ from this.

The naive approach is to take a new state $h = \partial_x f$ and use it as a state, but then we are back to exactly the same issue we had before (of too many states)

6 Intuition check

Consider in the simplest case, the linear deterministic ODE:

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

then

$$x(t) = \alpha\tau(1 - \exp(-t/\tau))$$

which, of course, goes to $\alpha\tau$ in the long run. Assume $\alpha\tau > 1$. Then the time $t_{\rm sp}$ to reach $x = x_{th} = 1$ is given by

$$t_{\rm sp}(\tau;\alpha) = -\tau \log(1 - \frac{1}{\alpha\tau})$$

Thus if we know α and $t_{\rm sp}$ we can determine τ. Suppose we could choose α. What would be the value of α that would make λ 'most identifiable'?

Let us equate 'identifiability' with the derivative of $t_{\rm sp}$ wrt. τ (or rather with its magnitude).

$$t_{\rm sp}{}'(\tau) = \frac{\tau\alpha}{1-\alpha\tau} - \log(1 - \frac{1}{\alpha\tau})$$

Let us check the asymptotics:

$$\lim_{\alpha\to\infty} t_{\rm sp}{}'(\tau) = -1$$

and

$$\lim_{\alpha\leftarrow 1/\tau} t_{\rm sp}{}'(\tau) = -\infty$$

As I read this, this means that the 'best' thing to do is let $\alpha \approx 1/\tau$ or more generally, $\alpha \approx x_{th}/\tau$

Now of course, in the noisy case,

$$dX = (\alpha - \frac{X}{\tau})\,{\rm d}t + \sigma dW$$

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, $\alpha\to\infty$, but to excite *critically.*

7 Using particles

What if we use particles $\{\lambda_k\}$ to represent the prior/posterior?

Then the entropy of the posterior for a given interval i_n,

$$H[\rho; i_n] = -\int \frac{g_\alpha(i_n|\lambda)\cdot\rho(\lambda)}{\int g_\alpha(i_n|\lambda)\cdot\rho(\lambda)\,\mathrm{d}\lambda}\cdot\log\left(\frac{g_\alpha(i_n|\lambda)\cdot\rho(\lambda)}{\int g_\alpha(i_n|\lambda)\cdot\rho(\lambda)\,\mathrm{d}\lambda}\right)\mathrm{d}\lambda$$

in terms of the particles λ_k becomes

$$H[\{\lambda_k\}; i_n] \approx -\frac{1}{K}\sum_k \frac{g_\alpha(i_n|\lambda_k)}{\sum_k g_\alpha(i_n|\lambda_k)}\cdot\log\left(\frac{g_\alpha(i_n|\lambda_k)\rho(\lambda_k)}{\sum_k g_\alpha(i_n|\lambda_k)}\right)$$

ah great! When we take a particle approach, we don't have an access to the prior $\rho(\lambda)$ analytically. Why is that a problem? Because we need to evaluate $\rho(\lambda_k)$ inside the log.

However!

If instead of minimizing the entropy of the posterior, we maximize the relative entropy of the posterior *relative* to the prior,

$$KL[\rho; i_n] = \int \frac{g_\alpha(i_n|\lambda)\cdot\rho(\lambda)}{\int g_\alpha(i_n|\lambda)\cdot\rho(\lambda)\,\mathrm{d}\lambda}\cdot\log\left(\frac{\left(\frac{g_\alpha(i_n|\lambda)\cdot\rho(\lambda)}{\int g_\alpha(i_n|\lambda)\cdot\rho(\lambda)\,\mathrm{d}\lambda}\right)}{\rho(\lambda)}\right)\mathrm{d}\lambda$$

Then the two ρ's inside the log will cxl!

$$KL[\rho; i_n] = \int \frac{g_\alpha(i_n|\lambda)\cdot\rho(\lambda)}{\int g_\alpha(i_n|\lambda)\cdot\rho(\lambda)\,\mathrm{d}\lambda}\cdot\log\left(\frac{g_\alpha(i_n|\lambda)}{\int g_\alpha(i_n|\lambda)\cdot\rho(\lambda)\,\mathrm{d}\lambda}\right)\mathrm{d}\lambda$$

With that we *can* write KL completely in terms of our particle ensemble $\{\lambda_k\}$.

$$KL[\{\lambda_k\}; i_n] \approx \frac{1}{K}\sum_k \frac{g_\alpha(i_n|\lambda_k)}{\frac{1}{K}\sum_k g_\alpha(i_n|\lambda_k)}\cdot\log\left(\frac{g_\alpha(i_n|\lambda_k)}{\frac{1}{K}\sum_k g_\alpha(i_n|\lambda_k)}\right)$$

We want to maximize KL, so our objective becomes

$$J = \int_{s\in[0,\infty]} KL[\{\lambda_k\}; s]\cdot g_\alpha(s)\,\mathrm{d}s$$

where the marginal hitting density is given by

$$g_\alpha(s) = \frac{1}{K}\sum_k g_\alpha(s|\lambda_k)$$

The unconditional density appears both on numerator and denominator in the integrand so it can be cancelled and J simplifies to

$$J[\alpha(\cdot)] = \int_{s\in[0,\infty]} \frac{1}{K}\sum_k g_\alpha(s|\lambda_k)\cdot\log\left(\frac{g_\alpha(s|\lambda_k)}{\frac{1}{K}\sum_k g_\alpha(s|\lambda_k)}\right)\mathrm{d}s \qquad (25)$$

Our goal is to maximize eq. (25) over the allowed controls $\alpha(\cdot)$.

Note that if the prior, ρ, is uninformative, i.e. $\rho(\lambda) = \text{const}$ then *minimizing* H is the same as *maximizing* KL; in particular for $\rho = 1$, $KL = -H$.

Why should we be maximizing the relative entropy, KL? The idea is that we equate a 'maximally informative experiment' with 'moving the posterior as far away from the prior as possible'. *'The justification of such a mathematical construct is solely and precisely that it is expected to work.'*

8 Parametrizing the controls

So far we have taken a Maximum Principle approach to create an equation for the optimal control. That has proved very difficult, in particular deriving and implementing the adjoint equations is highly complex.

Instead let us parametrize the controls via a set of basis functions:

$$\alpha(t) = \sum_i a_i \phi_i(t)$$

where the basis functions ϕ_i can be anything, for example Chebyshev or Legendre polynomials.

The question now becomes how to enforce $\alpha(t) \in [\alpha_{\min}, \alpha_{\max}]$. What we want to do is minimize $J[\{a_i\}]$ first and we'll think about how to update the particles, $\{\lambda_k\}$, given a spike-time observation at a later point.

So the objective is now framed in terms of the basis coefficients $\{a_k\}$:

$$J[a_i] = \frac{1}{K} \sum_{s_n} \left[\sum_{\lambda_k} g_\alpha(s_n|\lambda_k) \cdot \log \left(\frac{g_\alpha(s_n|\lambda_k)}{\frac{1}{K}\sum_k g_\alpha(s_n|\lambda_k)} \right) \right] \Delta_n$$

at appropriate time-nodes $s_n \in [0, T]$ where T is taken big enough so that nothing interesting happens for $s > T$.

As always, we will add a quadratic penalty for the control:

$$J[a_i] = \sum_{s_n \in [0,T)} \left[\frac{1}{K} \sum_{\lambda_k} g_\alpha(s_n|\lambda_k) \cdot \log \left(\frac{g_\alpha(s_n|\lambda_k)}{\frac{1}{K}\sum_k g_\alpha(s_n|\lambda_k)} \right) \right] - \epsilon \left(\sum_i a_i \phi_i(s_n) \right)^2 \Delta_n \tag{26}$$

And we are trying to maximize J over the coefficients $\{a_i\}$.

A big problem is that α is supposed to be bounded, but any polynomial is inherently unbounded as $T \to \infty$. So hopefully T is small enough and the basis is expressive enough for that not to be a problem.

Now, the million dollar question is can we form the gradient of J wrt. a_i

$$\nabla_{\{a_i\}} J = ?$$

All the a_i's are the 'same'. So it suffices to calculate the sensitivity for one of them, we will write this as

$$\partial_i J = \frac{\partial J}{\partial a_i}$$

Basically, a_i is a sensitivity in a PDE. We will need to calculate things of the form: $\partial_{a_i} F$, which satisfy sensitivity PDEs related to the main

PDE for F. So we are going to have to calculate PDEs for each pair of i, k, i.e. $I \times K$ PDEs, to evaluate the gradient of J.

Ufff!

We could try to calculate the adjoint sensitivity, that is calculate the gradient all at once for all a_i not for each separately, which will be complex but may work.... However our experience with adjoint sensitivities has not been stelar.

9 Basic Test

Ok, let's try it.

We will run the following test:

Given the true parameters

$$\beta = 1.; \mu = 0; \lambda = .5;$$

We will assume we know β, μ and don't know λ.

Let's assume a uniform prior on λ, $\lambda_k = .1 \ldots 2.$, with $N_p = 8$ number of particles. Sample solutions for the hitting-time density, $g_(s)$, for the two basic inputs $\alpha = 0$ and $\alpha = .5$ and for $\lambda_k = .1, .5, 1., 2.$ are shown in figs. 1 and 2.

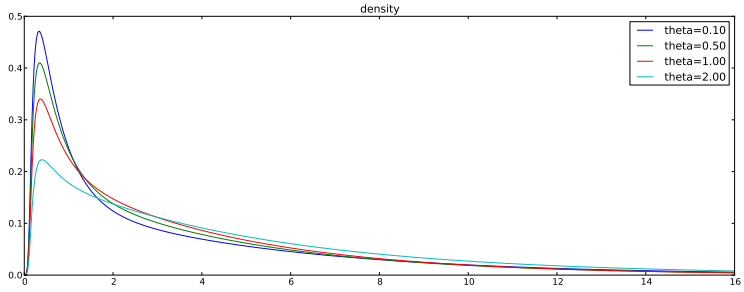

Figure 1: $g_{\alpha_0}(s|\lambda = .1, 2)$, i.e. the input $\alpha(t) = 0$

Now we are going to attempt to optimize J in eq. (26) over the set of piecewise constant $\alpha(t)$, breaking α over the intervals $[0, T/N_i \ldots T]$, where N_i is the number of interval (first we'll try 5);

We use the simplest function-only optimizer - the Nelder-Mead routine. We also compare against three simple signals that are constant $\alpha = 0, .5, 2$. The objective values are in table 1 and plots of the controls $\alpha(t)$ are in fig. 3. In table 1, we see that the optimal input signal is not just better than the max value b/c it has lower energy (time-integral of α^2), but it also has a higher value of the K-L divergence (slightly higher - $0.0125 > 0.0118$).

I don't know what to make of these results. Indeed, we came up with an improvement of the objective and a non-trivial solution. But the improvement is very marginal and the optimal control is not really

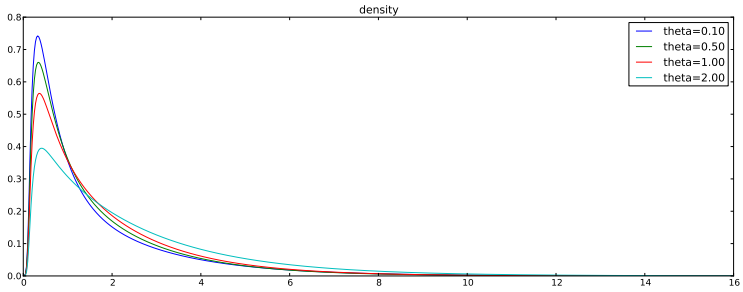

Figure 2: $g_{\alpha_c}(s|\lambda = .1, 2$, i.e. the input $\alpha(t) = .5$

α	KL	$\int \alpha^2 \, \mathrm{d}t$	J
$\alpha = 0$	0.0079	0.0000	0.0079
$\alpha = 0.5$	0.0112	0.0003	0.0109
$\alpha = 2$	0.0118	0.0048	0.0070
$\alpha = \alpha^*(t)$	0.0125	0.0005	0.0120

Table 1: The objective values corresponding to some basic controls. J refers to the total objective $= KL + \int \alpha^2$ while KL refers to the relative entropy only (so omitting the quadratic energy penalty). The value of $\epsilon = .0001$, so that energy cost should really have minimum impact here

that exciting. However, if I try a bang-bang control, $\alpha = \pm 2$ alternating on intervals, I get $KL = 0.0118$, which is still worse than the optimal. If I try a bang-bang solution with $N_i = 16$, this improves (slightly) to $KL = 0.0121$, but still below the optimal. For a sinusoidal $\alpha = \sin(t)$, I get $KL = 0.0092$, which is worse than the optimal. We can do (slightly) better if we increase the frequency of the sinusoidal, $KL = 0.0096$.

Anyway. I suppose the thing to do now is to actually run the system under $\alpha = 0$ and $\alpha = \alpha^*$ and see if estimates are better with α^*...

Btw, the results hold if we double the number of particles to $N_p = 16$. (A very similar optimal α^* is calculated to the one in fig. 3).

Another btw, if we are working with a non-uniform prior (particles are 'bunched up') then the results are even less encouraging...

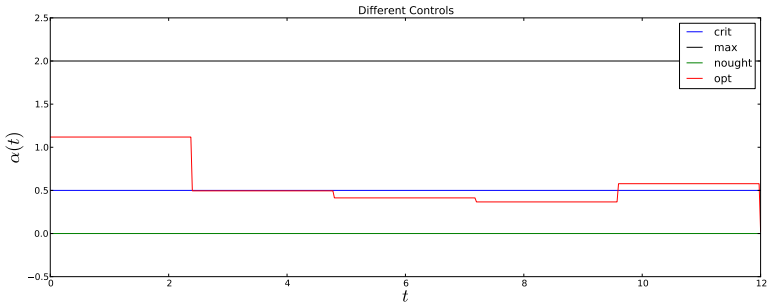

Figure 3: The three constant $\alpha(\cdot)$ used for comparison and the optimal $\alpha^*(\cdot)$ calculated by maximizing eq. (26) over $N_i = 5$ intervals

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

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

Let us write the dynamics as such

$$dX = \underbrace{f(X,\theta,\alpha)}_{\text{controlled drift}}\ dt + \beta dW \tag{27}$$

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

$$\begin{aligned} l(\theta|x_t) =& \frac{1}{2}\int_0^{t_f} \frac{f^2(x_t,\theta,\alpha)}{\beta^2}\,\mathrm{d}t \\ & - \int_0^{t_f} \frac{f(x_t,\theta,\alpha)}{\beta^2}\,\mathrm{d}W \end{aligned} \tag{28}$$

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

$$\Phi(\theta,\alpha) = \mathbb{E}\left[\int_0^{t_f} \frac{\left(d_\theta f(x_t,\theta,\alpha)\right)^2}{\beta^2}\,\mathrm{d}t\right] \tag{29}$$

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. [**?**] 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

$$\tilde{I}(\theta,\alpha) = \underbrace{\mathbb{E}_\theta\left[\underbrace{\mathbb{E}_X\left[\int_0^{t_f} \frac{\left(d_\theta f(x_t,\theta,\alpha)\right)^2}{\beta^2}\,\mathrm{d}t\right]}_{\text{average over trajectories}}\right]}_{\text{average over prior}} \tag{30}$$

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)

References

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