# 1 Introduction

In this document we provide detailed proofs for the equations presented in the main text and discuss in more detail some of the underlying principles.

The derivation of hard performance bounds in the presence of noisy feedback has been extensively addressed by control theorists over the last century, with many results going back to the early work of Wiener, Bode and Kalman. In a biochemical reaction network, however, where the noise is associated with the spontaneous births and deaths of individual molecules, many of the more conventional methodologies[1] come to a halt, and need to be appropriately refined and extended. Our approach is to address such complications by quantifying the way these spontaneous fluctuations restrict the ability to transmit information reliably, i.e. they lead to communication channels with finite capacity in the sense of Shannon, which can be explicitly calculated. We show that such restrictions in information transmission fundamentally limit the ability to suppress fluctuations in molecular concentrations, thus leading to hard bounds for noise suppression that hold for *arbitrary* feedback policies.

More precisely, it follows from the law of total variance in probability theory, that for a random process $x_t$ and a signal $u_0^t$, the variance of $x_t$ is lower bounded by the estimation error $E\{[x_t - \hat{x}_t]^2\}$, where the conditional expectation $\hat{x} = E\{x_t | u_0^t\}$ is also the minimum mean squared error estimator of $x_t$ given $u_0^t$. In our setting $x_t$ is the molecule numbers of the controlled species $x_1$, and $u_0^t$ is the control signal that determines the rate of formation of $x_1$, with any information about $x_1$ conveyed to $u_0^t$ via an intermediate signaling species $x_2$ made in probabilistic birth events.

A central part of the paper is to treat signaling pathways as Poisson communication channels, and quantify restrictions in reliable information transmission by means of information theoretic notions: the 'mutual information' and its smallest upper bound known as the 'Shannon capacity'. A restricted mutual information between $x_1$ and $x_2$, i.e. a bounded capacity of the 'Poisson channel' from $x_1$ to $x_2$, is shown to fundamentally restrict the ability to reduce the estimation error and hence the $x_1$ variance.

---

[1]Such as approaches that rely on diffusion approximations for the noisy feedback signals.

Such constraints in the channel capacity arise from the fact that the intensity of the $x_2$ births is not allowed to vary infinitely, due to restrictions in its maximum value or moments. In particular, if $x_1$ directly affects the rate of formation of $x_2$, then a small $x_1$ variance can restrict the capacity of the channel from $x_1$ to $x_2$; i.e. a smaller variance makes it harder to transmit the information that is fundamentally needed to reduce it. This tradeoff leads to the quartic root hard limits discussed in the main text.

In order to be able to quantify these limits we derive an explicit expression for the capacity of the Poisson channel $x_2 \xrightarrow{\;f\;} x_2 + 1$ , where the mean and variance of the intensity $f$ are constrained (the derivation is also extended to channels with serial cascades, bursts and parallel reactions). The limits are derived in section 4 and are shown to hold for an arbitrary feedback control law by interpreting the latter as being part of a channel with feedback (i.e. a channel where its output can be used in the 'encoding' of the transmitted information). The analysis also makes use of Pinsker's nonanticipatory epsilon entropy, an information theoretic notion that takes into account the fact that information is transmitted in real time.

The document is organized as follows. To improve readability of the manuscript we start with a self contained discussion and leave more detailed derivations for the appendix. Section 2 gives an introduction to some basic concepts in the theory of information transmission that are used within the paper. In section 3 we discuss how signalling pathways can be treated as Poisson communication channels and derive expressions for the corresponding Shannon capacity under constraints that are relevant for the development that follows. In section 4 we derive hard bounds for the variance of the controlled species in terms of the Shannon capacity of the feedback channel. These are used to derive the hard limits that appear in the main text. In section 5 we show that analogous more restrictive bounds hold when feedback information is conveyed by means of multiple species, either in the form of a serial cascade or as the superposition of many parallel systems. Finally, section 6 discusses noise tradeoffs that arise in specific feedback mechanisms. The appendix gives an overview of Pinsker's nonanticipatory epsilon entropy, and also includes detailed proofs of the results on which the discussion in the previous sections is based.

## 2   Preliminaries

Notation. For a random variable $X$ we denote its mean by $E\{X\}$ or $\langle X \rangle$, and its variance by $Var\{X\}$ or $\sigma_X^2$. Random process $\{x_t : t_1 \leq t \leq t_2\}$ is denoted by $x_{t_1}^{t_2}$, where subscript $t$ denotes time i.e. $x_t = x(t)$. For convenience in the presentation, natural numbers will be used as subscripts to denote species number, i.e. $x_1(t)$ denotes the number of molecules of species $x_1$ at time $t$ (the time argument will often be omitted for simplicity).

This section gives an overview on some basic notions in the theory of information transmission that are used throughout the manuscript. Information theory is associated with the efficient and reliable transmission of information through channels that

can lead to its corruption. A key concept in this context is that of channel capacity, which quantifies the maximum rate with which information can be transmitted reliably through a channel. One of Shannon's fundamental contributions has been to relate this to the notion of mutual information, which we define in more detail below.

For two discrete random variables $X, Y$ with probability mass function $P_X, P_Y$ respectively and joint probability function $P_{XY}$, the *mutual information* $I(X;Y)$ is defined as

$$I(X;Y) := \sum_{x,y} P_{XY}(x,y) \log \left[ \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \right]$$
$$= E \left\{ \log \left[ \frac{P_{XY}}{P_X P_Y} \right] \right\}$$

a definition that can also be extended to continuous random variables and random processes[2].

Consider now a random process $\xi_0^T$, which is a message that needs to be transmitted through a noisy channel. For our purposes $\xi$ is associated with the fluctuations in $x_1$ and will be defined in detail in section 4. The message $\xi$ is first 'encoded' into signal $f_0^T$ (meaning that $f_0^T$ is an arbitrary function of the time series $\xi_0^T$) that is considered as the input to the channel and random process $y_0^T$ denotes the output from the channel, which is a corrupted version of $f_0^T$. The channel is allowed to have feedback if $f_0^T$ is a function of $y_0^T$ in addition to the message $\xi_0^T$ (see figure 1). The *capacity* of the channel is defined as

$$\sup \frac{1}{T} I(\xi_0^T; y_0^T)$$

where the supremum is taken over all distributions of $\xi_0^T$ and all allowable input signals[3] $f_0^T$.

Throughout the manuscript we concentrate on Poisson channels. In such channels the output is a Poisson process with intensity at any time $t$ the input signal $f_t$. If, for example, information about a species $x_1$ is communicated via an intermediate species $x_2$ as in the main text, then $y$ denotes the signalling species $x_2$ and its birth rate $f$ is a function of $x_1$.

Finally, a useful property of the mutual information that is used in the derivations in section 4 is the *data processing inequality*. Random variables $Z, Y, X$ are said to form a Markov chain if $Z$ conditioned on $Y$ is independent of $X$. This is, for example, the case if $Z = g(Y)$, i.e. $Z$ follows from a 'processing' of $Y$. The data processing inequality states that if $Z, Y, X$ form a Markov chain then then $I(X;Z) \leq I(X;Y)$, i.e. any processing cannot increase the mutual information.

---

[2]For random variables $X, Y$ with probability measure $\mu_X, \mu_Y$ and joint measure $\mu_{XY}$ the mutual information is given by

$$I(X;Y) = E \left\{ \log \left[ \frac{d\mu_{XY}}{d(\mu_X \times \mu_Y)} \right] \right\}$$

where $d\mu_{XY}/d(\mu_X \times \mu_Y)$ denotes the Radon-Nikodym derivative of $\mu_{XY}$ with respect to the product measure $\mu_X \times \mu_Y$.

[3]Shannon [1] gave a physical interpretation to the capacity as the highest rate with which information can be transmitted, such that this can be recovered with an arbitrarily small probability of error from the channel output (an interpretation that also holds for Poisson channels [2]).
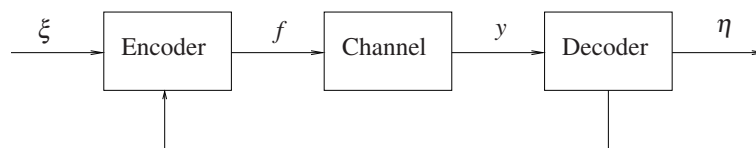
Figure 1: Channel with feedback. Random process $\xi_0^T$ is the message to be transmitted, $f_0^T$ is the input to the channel which is an encoded version of $\xi_0^T$, $y_0^T$ is the channel output, and $\eta_0^T$ is the estimate of the transmitted message $\xi_0^T$.

In order to be able to derive the hard variance bounds presented in the main text, analytical expression are required for the channel capacity of Poisson channels. These are given in section 3 (and extended in section 5) under various kinds of constraints on the birth rate.

The variance limits are derived in section 4, where an appropriate transformation of the problem allows to interpret feedback control laws as being part of a channel with feedback. The limits then follow from a fundamental tradeoff we show between the variance of a species and the ability to communicate information about that species reliably through a Poisson channel.

## 3 The information capacity of molecular channels

Central to the development of the paper is the treatment of signalling pathways as Poisson communication channels. The birth rate $f$ of the second species $x_2$ depends on the state of the first species $x_1$, and possibly of $x_2$ itself. The question is how much information about $x_1$ can be inferred from $x_2$ under various different constraints on the intensity $f$.

$$x_2 \xrightarrow{f} x_2 + 1 \qquad (1)$$

This channel has been studied in the optical communications literature, where it is known as the direct detection photon channel. Some of the known results are of use to us here, but we also need to determine the channel capacity under novel forms of constraints.

First, if $f$ is simply constrained to lie between the values 0 and some maximum value $f_{\max}$ then it is known [3] that the capacity is[4]

$$C = f_{\max}/e,$$

If, in addition, the mean of the intensity $f$ is constrained to be constant, it has been shown in [4] that the capacity is given by

$$C = K\langle f \rangle$$

---

[4]Constant $e$ in the expression for the capacity denotes the base of the natural logarithm.

where

$$K = \log(f_{\max}/\langle f \rangle).$$

In each of these cases, it is known that the capacity is realized by switching $f$ between 0 and $f_{\max}$ at an infinitely fast rate.

Since our aim is to establish limits in the $x_1$ variance, we are particularly interested in the situation where $x_1$ directly modulates the channel and restrictions on statistical properties (such as mean and variance) of $x_1$ thus impose restrictions on the statistical properties of the intensity $f$. We show (Theorem 1 in appendix B) that when the first two moments of the intensity are fixed, the channel capacity is given by

$$C = \langle f \rangle \log \left( 1 + \frac{\sigma_f^2}{\langle f \rangle^2} \right) \leq \frac{\sigma_f^2}{\langle f \rangle}.$$

The idea, roughly speaking, behind the derivation is to apply Jensen's inequality on the Liptser-Shiryaev formula [5] for the mutual information, and then maximize over the distributions of the intensity $f$ to get an upper bound for the capacity, which can then be shown to be achievable.

Note that all the capacities given above allow for feedback, that is they are not increased by allowing the intensity $f$ to depend on $x_2$.

## 4 The bounds

The main ideas behind the bounds we derive can be summarized as follows:

1. There is a bound on the variance in $x_1$ whenever there is a bound on the error with which $x_1$ can be estimated.

2. To achieve a sufficiently small estimation error, a minimal capacity is required for the channel through which information is conveyed about $x_1$.

3. To achieve a certain capacity, a minimal variance in the reaction rates of intermediate species is required.

4. If these rates depend on $x_1$, then a high variance of $x_1$ is needed to increase the channel capacity.

Steps 1-4 thus mean that reducing the variance reduces the channel capacity which in turn makes it harder to further reduce the variance. This eventually causes a hard bound beyond which no improvements can be made.

As indicated in the main paper, we consider now the following Stochastic Differential Equation for $x_1$

$$dx_1 = (u - x_1/\tau_1)dt + \sqrt{2\langle x_1 \rangle/\tau_1}\, dw \tag{2}$$

where $u$ is to be chosen to minimize fluctuations in $x_1$ about its stationary mean $\langle x_1 \rangle$, but with the constraint that any information about $x_1$ is only to be relayed to $u$ via a

communication channel with feedback capacity $C$ (in our case via another species $x_2$, i.e. $x_2 \xrightarrow{\ f\ } x_2 + 1$ with $f$ being a function of $x_1$). It is important to note here that we do not specify or linearize the feedback scheme, or make any diffusion approximations for the signaling system, where the discrete nature of individual events is key. This is, we will allow an *arbitrary* nonlinear feedback mechanism from $x_2$ to $u$ (the first "control demon" of the paper)[5].

We consider a decomposition of $x_1$ as $x_1 = \xi + \phi$, where $\xi$ satisfies (2) when the control signal $u$ is set to zero (i.e. $d\xi = -b\xi dt + \beta dw$ with $b = 1/\tau_1$ and $\beta^2 = 2\langle x_1 \rangle / \tau_1$) and $\phi$ satisfies

$$d\phi = (u - \phi/\tau_1)dt$$

i.e. it is a function of only the signal[6] $u$.

Using now the law of total variance (the decomposition in (3)) we derive a lower bound for the variance of $x_1$, which we then write in terms of $\xi_t$ and its conditional mean, i.e.

$$Var\{x_1(t)\} = E\left\{ \left[ x_1(t) - E\left\{ x_1(t)|u_0^t \right\} \right]^2 \right\} + E\left\{ \left[ E\left\{ x_1(t)|u_0^t \right\} - E\{x_1(t)\} \right]^2 \right\} \quad (3)$$

$$\geq E\left\{ \left[ x_1(t) - E\left\{ x_1(t)|u_0^t \right\} \right]^2 \right\} \quad (4)$$

$$= E\left\{ \left[ \xi_t + \phi_t - E\left\{ \xi_t + \phi_t|u_0^t \right\} \right]^2 \right\}$$

$$= E\left\{ \left[ \xi_t - E\left\{ \xi_t|u_0^t \right\} \right]^2 \right\} \quad (5)$$

Note that $E\left\{ x_1(t)|u_0^t \right\}$ is the minimum mean squared error (MMSE) estimate of $x_1(t)$ given the feedback signal $u_0^t$, and the expression in (4) is the corresponding estimation error. Therefore inequality (4) captures the fact that the variance of $x_1(t)$ is restricted when there are lower bounds in the error with which it can be estimated[7].

Let $\eta_t := E\left\{ \xi_t|u_0^t \right\}$. A bound for the estimation error in (4) (and hence the $x_1$ variance) can be obtained in terms of the channel capacity by taking into account the fact that the latter restricts the mutual information rate $\frac{1}{t}I(\xi_0^t; \eta_0^t)$. This constraint on the mutual information follows from the data processing inequality, i.e. (we denote for

---

[5]At any time $t$, control signal $u_t$ is an arbitrary function of $\{x_2(\tau) : 0 \leq x_2(\tau) \leq t\}$, the history of $x_2$ up to time $t$.

[6]The importance of process $\xi$ is that it quantifies the deviation of $x_1$ about its conditional mean, i.e. $x_1(t) - E\left\{ x_1(t)|u_0^t \right\} = \xi_t - E\left\{ \xi_t|u_0^t \right\}$. This will be exploited later in our analysis when we make use of the law of total variance.

[7]The limits derived in this section can still be deduced in an analogous way if in the conditional expectation in (3) we condition on both $u_0^t$ and the channel output $y_0^t$. It should be noted in this case that if a gaussian approximation was used for the Poisson channel, then what is known as separation principle in optimal control theory holds. That is, the first term in (3) does not depend on $u_0^t$, and hence $u_0^t$ can be chosen to make the second term equal to zero. Such a separation principle does not hold in the case of a Poisson channel, thus making the problem of minimizing the $x_1$ variance non trivial. Using, however, information theoretic concepts hard variance limits can be obtained, as we show in the remainder of the section.

convenience $y_t := x_2(t)$)

$$\frac{1}{t}I(\xi_0^t;\eta_0^t) \leq \frac{1}{t}I(\xi_0^t;u_0^t) \tag{6}$$

$$\leq \frac{1}{t}I(\xi_0^t;y_0^t) \tag{7}$$

$$\leq C \tag{8}$$

where $C$ is the feedback capacity of the Poisson channel[8]. We have used here in (6) the fact that $\xi_0^t$ conditioned on $u_0^t$ is independent of $\eta_0^t$. Similarly in (7) $\xi_0^t$ conditioned on $y_0^t$ is independent of $u_0^t$, since any information about $x_1$ is conveyed to signal $u$ by means of $x_2$.

To sum up, we have from (5) and (8) respectively

$$Var\{x_1(t)\} \geq E\left\{(\xi_t - \eta_t)^2\right\} \quad\text{and}\quad \frac{1}{t}I(\xi_0^t;\eta_0^t) \leq C$$

Hence, for a constant variance of $x_1$ up to time $T$, minimizing over $\eta_0^T$ the distortion $d(\xi_0^T,\eta_0^T) := \sup_{0 \leq t \leq T} E\left\{(\xi_t - \eta_t)^2\right\}$, under the mutual information constraint $\frac{1}{T}I(\xi_0^T;\eta_0^T) \leq C$ gives a hard lower bound for the variance of $x_1$; that is,

$$Var\{x_1\} \geq \inf_{\eta_0^T} d(\xi_0^T,\eta_0^T) \tag{9}$$

$$\text{such that } \frac{1}{T}I(\xi_0^T;\eta_0^T) \leq C \text{ and}$$

$$\eta_0^t,\xi_0^t,\xi_t^T \quad 0 \leq t \leq T \text{ form a Markov chain}[9] \tag{10}$$

where (10) is a causality constraint that captures the natural ordering of the signals; i.e. estimator $\eta_0^t$ carries no information about future $\xi$ values other than information that can be inferred from the past of $\xi$ .

Minimizing a distortion function under mutual information constraints lies within the context of a branch of information theory known as rate distortion theory. However, in order to additionally take into account causality constraint (10) we need to make use of Pinsker's nonanticipatory epsilon entropy [6], [7] (see appendix A).

The nonanticipatory epsilon entropy of $\xi$ solves the 'inverse' problem of minimizing the mutual information $I(\xi_0^T;\eta_0^T)$ given an upper bound on the distortion $d(\xi_0^T,\eta_0^T)$ and causality constraint (10). This can be used to obtain a lower bound for the distortion in (9), as discussed in appendix A. In particular, for a first order process

$$d\xi = -b\xi dt + \beta dw \tag{11}$$

the distortion is lower bounded by

$$d(\xi_0^T,\eta_0^T) \geq \frac{\beta^2}{2(C+b)} \tag{12}$$

---

[8]The feedback control law is interpreted here as feedback around the channel from $\xi$ to $y$, which is why we need the feedback capacity of the channel.

[9]Signal $\eta_0^t$ conditioned on $\xi_0^t$, the past of $\xi$, is independent of its future $\xi_t^T$.

If $\xi$ is a Wiener process $d\xi = \beta dw$ then (12) reduces to

$$d(\xi_0^T, \eta_0^T) \geq \frac{\beta^2}{2C} \tag{13}$$

Noting the way $\xi$ is analogously defined in this section (i.e. it satisfies (11) with $b = 1/\tau_1$, $\beta^2 = 2\langle x_1\rangle/\tau_1$)) we therefore have from (9), (12) that for a constant mean and variance of $x_1$

$$\frac{\sigma_{x_1}^2}{\langle x_1\rangle} \geq \frac{1}{C\tau_1 + 1} \tag{14}$$

This is a hard limit for the $x_1$ variance in terms of the capacity, which captures the fact that a large channel capacity is a fundamental requirement when we need to reduce variances.

The capacity expressions given in section 3 can therefore now be used to derive limits in the variance of $x_1$ under various settings. If we allow arbitrary encoding schemes[10] with the birth rate $f$ of the second species $x_2$ constrained to have a mean $\langle f\rangle$ and maximum value $f_{\max}$, it follows from [4] that the capacity of the channel from $x_1$ to $x_2$ is given by[11]

$$C = K\langle f\rangle \tag{15}$$

where $K = \log\frac{f_{\max}}{\langle f\rangle}$. Substituting in (14), we obtain

$$\frac{\sigma_{x_1}^2}{\langle x_1\rangle^2} \geq \frac{1}{N_1(KN_2 + 1)} \tag{16}$$

where $N_1 = \langle x_1\rangle$ and $N_2 = \langle f\rangle\tau_1$. Therefore arbitrary encoding leads to a bound for the $x_1$ coefficient of variation that decreases with the square root of $N_2$. As we show below if the encoding functions for $f$ are linear then the bound becomes more restrictive decaying with the quartic root of $N_2$.

We consider now the case where $f$ is proportional to $x_1$, i.e. (1) is replaced by

$$x_2 \xrightarrow{\alpha x_1} x_2 + 1 \tag{17}$$

This leads to explicit bounds for the variance of the controlled species $x_1$, since the latter directly restricts the channel capacity. More precisely, for a given mean and variance of $x_1$ the first two moments of the intensity are also restricted. Hence as discussed

---

[10]That is, at any time $t$, the birth rate $f_t$ of $x_2$ is an arbitrary function of $\{x_1(\tau) : 0 \leq \tau \leq t\}$ with the mean and maximum value of $f_t$ constrained. Such a setting is not relevant when the $x_2$ intensity depends proportionally on $x_1$, i.e. $f_t = \alpha x_1(t)$. It should be noted that when arbitrary encoding is instead allowed, the use of a stochastic differential equation for $x_1$ also avoids extreme scenarios whereby an 'encoding demon' monitors both individual $x_1$ and $x_2$ events and induces $x_2$ events immediately after corresponding $x_1$ ones.

[11]Note that the capacity of the Poisson channel does not increase with feedback, so this is also the capacity of the feedback channel from $\xi$ to $y$.

in section 3 and proved in Theorem 1 in appendix B, the capacity is given by[12]

$$C = \alpha\langle x_1\rangle \log\left(1 + \frac{\sigma_{x_1}^2}{\langle x_1\rangle^2}\right) \leq \alpha\frac{\sigma_{x_1}^2}{\langle x_1\rangle} = \frac{N_2}{\tau_1}\frac{\sigma_{x_1}^2}{\langle x_1\rangle^2} \tag{18}$$

This means that a large $x_1$ variance is needed to increase the channel capacity. There is therefore a fundamental limit on how much the variance can be decreased due to an increase in the capacity (as this was quantified in (14)). The limit follows by substituting (18) in (14) and then solving the corresponding quadratic inequality, thus giving

$$\frac{\sigma_{x_1}^2}{\langle x_1\rangle^2} \geq \frac{1}{\langle x_1\rangle} \times \frac{2}{1 + \sqrt{1 + 4N_2/N_1}} \tag{19}$$

Analogous arguments can be used for the case where the birth rate of $x_1$ is given by $ux_1$. The diffusion approximation of $x_1$ is now $dx_1 = u\langle x_1\rangle dt + \sqrt{2\langle x_1\rangle/\tau_1}dw$ and it follows from (9), (13), (18) that for linear encoding functions for $f$ as in (17) we have

$$\frac{\sigma_{x_1}^2}{\langle x_1\rangle^2} \geq \frac{1}{\sqrt{N_1 N_2}} \tag{20}$$

Similarly for the case of arbitrary encoding, we use the variance limit in terms of the capacity that follows from (9) and (13), i.e. $\frac{\sigma_{x_1}^2}{\langle x_1\rangle} \geq \frac{1}{\tau_1 C}$, and substitute in the latter the capacity expression (15). This therefore leads to the the same lower bound as in (16) but without the +1.

## 4.1  Hill functions for nonlinear encoding

It should be noted that the variance bounds in (19) and (20), where a quartic root dependence of the $\sigma_{x_1}$ limit on $N_2$ is featured, are based on an upper bound on the capacity of the form in (18). Analogous bounds for the variance can therefore be obtained in cases $x_1$ is affecting nonlinearly the birth rate of $x_2$, i.e.

$$x_2 \xrightarrow{f(x_1)} x_2 + 1 \tag{21}$$

if upper bounds on the capacity of the same form can be obtained. In particular, for a Hill function

$$f(x_1) = v\frac{K_1 + x_1^h}{K_2 + x_1^h}$$

the capacity is bounded by

$$C \leq \frac{\gamma N_{2\max}}{\tau_1}\frac{\sigma_{x_1}^2}{\langle x_1\rangle^2} \tag{22}$$

---

[12]Note that for the model setting to be well defined the birth rate $f_t$ needs to be truncated so that it is always positive i.e. $f_t = \alpha x_1^+(t)$, where $x_1^+$ denotes the truncated process $x_1^+(t) = x_1(t)$ if $x_1(t) \geq 0$ and 0 otherwise. This does not affect the upper bound on the capacity given below since $\langle x_1^+\rangle \geq \langle x_1\rangle$ and $\sigma_{x_1^+}^2 \leq \sigma_{x_1}^2$.

where $N_{2\max} := \nu\tau_1$ is the maximum average number of molecules that are made per $x_1$ lifetime, i.e. the average number when the Hill function is saturated; $\gamma$ is a constant that depends on $K_1/K_2$ and the Hill coefficient $h$ and can be determined by solving a corresponding optimization problem[13,14]. Therefore comparing (22) with the capacity bound in (18), the same variance limits as in (19) and (20) hold but with $N_2$ now replaced by $\gamma N_{2\max}$. Parameter $\gamma$ is shown in figure 2 for different values of $K_1/K_2$ and $h$.
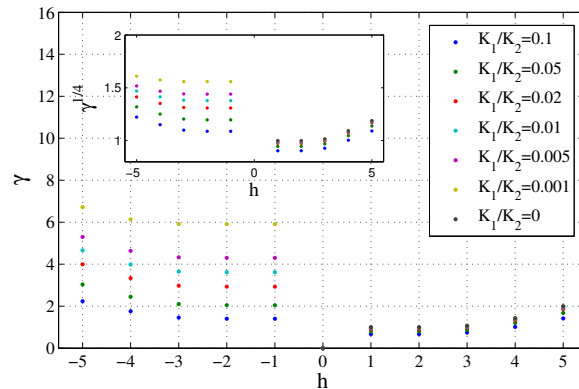


Figure 2: Parameter $\gamma$ in the capacity bound $C \leq \frac{\gamma N_{2\max}}{\tau_1} \frac{\sigma_x^2}{\langle x \rangle^2}$ for the Poisson channel in (21) with the $x_2$ birth rate being a Hill function $f(x_1) = \nu \frac{K_1 + x_1^h}{K_2 + x_1^h}$. (Inset) The quartic root of the same parameter $\gamma$.

For example, for a positive Hill function with $K_1 = 0$ and $0 \leq h \leq 4$ we have $\gamma < 2$, hence $N_2$ is replaced by $2N_{2\max}$ in the limits (19), (20). Similarly for a negative Hill function with $K_1/K_2 = 0.01$ and $-4 \leq h \leq 0$ we have that $\gamma < 4$, so $N_2$ is now replaced by $4N_{2\max}$.

## 4.2 Extrinsic noise

The methodology followed in section 4 also holds when the stochastic differential equation for $x_1$ in (2) is of higher order, with the noise source being correlated. This

---

[13] Parameter $\gamma$ can be obtained by solving the optimization problem $\max \frac{\hat{I}}{\nu\sigma_{x_1}^2/\langle x_1 \rangle^2}$, where $\hat{I}$ is the upper bound on the mutual information rate that follows from (42) in appendix B, and the maximization is over all $x_1$ distributions. It follows from Hoeffding's theorem on the extrema of expectations [8] that the optimum distribution is that of a discrete random variable with at most four points. This is therefore a tractable optimization problem that involves only a finite number parameters and can hence be solved numerically. Note that analogous ideas have been used in the exact calculation of the capacity when $f(x_1)$ is linear in $x_1$ (Theorem 1 in appendix B).

[14] It should be noted that $\gamma$ depends on the ratio $K_1/K_2$, rather than the individual values of $K_1$ and $K_2$, since the Hill function can be written as $f(x_1) = \nu(K_1/K_2 + \tilde{x}_1^h)/(1 + \tilde{x}_1^h)$, where $\tilde{x}_1 = x_1/K_2^{1/h}$, and note also that $\sigma_{\tilde{x}_1}^2/\langle \tilde{x}_1 \rangle^2 = \sigma_{x_1}^2/\langle x_1 \rangle^2$.

occurs when extrinsic noise is additionally included.

Consider, for example, the case where the birth rate of $x_1$ is given by $ux_3$, where $x_3$ is a species with a constant birth rate and exponential decay, i.e. $x_3 \xrightarrow{k} x_3 + 1$ $x_3 \xrightarrow{x_3/\tau_3} x_3 - 1$ , with the $x_3$ fluctuations acting as extrinsic noise for $x_1$. A corresponding diffusion approximation is given by

$$dx_3 = (k - x_3/\tau_3)dt + \sqrt{2\langle x_3\rangle/\tau_3}dw_3$$
$$dx_1 = \left(\langle x_3\rangle u + \frac{\langle x_1\rangle}{\tau_1\langle x_3\rangle}(x_3 - \langle x_3\rangle) - x_1/\tau_1\right)dt + \sqrt{2\langle x_1\rangle/\tau_1}dw_1 \qquad (23)$$

As in (2) $u$ is an arbitrary feedback control law that is chosen to minimize the fluctuations in $x_1$ about its stationary mean $\langle x_1\rangle$, but with any information relayed to $u$ via a second species[15] $x_2$.

Equation (23) is equivalent to a process

$$dx_1 = (\tilde{u} - x_1/\tau_1)dt + d\tilde{w} \qquad (24)$$

where $\tilde{u} = \langle x_3\rangle u$, and $\tilde{w}$ is a correlated noise process that satisfies $d\tilde{w} = \frac{\langle x_1\rangle}{\tau_1\langle x_3\rangle}(x_3 - \langle x_3\rangle)dt + \sqrt{2\langle x_1\rangle/\tau_1}dw_1$.

Following the procedure described in this section, we use the decomposition $x_1 = \tilde{\xi} + \tilde{\phi}$, where $\tilde{\xi}$ satisfies $d\tilde{\xi} = -x_1/\tau_1 dt + d\tilde{w}$. Therefore a lower bound for the $x_1$ variance in terms of the capacity is given by the distortion rate function of $\tilde{\xi}$ as defined in (9), with $\tilde{\xi}$ in the place of $\xi$. This computation is possible if classical rate distortion theory is used, with causality constraint (10) relaxed[16,17]. Combining as before this lower bound on the $x_1$ variance in terms of the capacity, together with the upper bound (18) on the capacity in terms of the $x_1$ variance, we get a system of equations for the $x_1$ variance limit that can be solved numerically.

In order to illustrate the effect of extrinsic noise we show in figure 3 the $x_1$ variance limit in the regime where the intrinsic noise is much smaller than the extrinsic noise[18] (i.e. $N_1 \to \infty$ for given $\langle x_3\rangle, \tau_1, \tau_3$). The limit in $\sigma_{x_1}^2/\langle x_1\rangle^2$ is normalized by the corresponding value when the input $u$ is constant, i.e. no feedback is present. $N_2$

---

[15]That is, $x_2 \xrightarrow{f(x_1)} x_2 + 1$ with $u(t)$ being an arbitrary function of the $x_2$ history $\{x_2(\tau) : 0 \leq \tau \leq t)\}$.

[16]The distortion rate function $D$ of $\tilde{\xi}$ for a given capacity constraint $C$, that is $D = \inf_{\eta_0^T} d(\tilde{\xi}^T, \eta_0^T)$ s.t. $\frac{1}{T}I(\tilde{\xi}_0^T; \eta_0^T) \leq C$, can be derived at steady state by what is known as a water-filling argument (e.g. [9]), i.e.

$$D = \frac{1}{2\pi}\int_{-\infty}^{\infty}\min(S(\omega), S(\omega_c))d\omega, \qquad C = \frac{1}{4\pi}\int_{-\infty}^{\infty}\log\max(S(\omega)/S(\omega_c), 1)d\omega \qquad (25)$$

where $S(\omega)$ is the power spectrum of $\tilde{\xi}$, and $\omega_c$ is a constant that depends on $C$.

[17]Note that this relaxation decreases the bounds in (19) and (20) by only a factor of at most $\pi/2$.

[18]The power spectrum of $\tilde{\xi}$ in this regime is $S(\omega) = \frac{2\tau_1 N_1^2/N_3}{(\tau_1^2\omega^2+1)(\tau_3^2\omega^2+1)}$, and from the first equation in (25) the normalized variance limit $(\sigma_{x_1}^2 \text{ limit})/(\sigma_{x_1}^2 \text{ open-loop})$ can be expressed in terms of $\omega_c$ and $\tau_3/\tau_1$ ($\sigma_{x_1}^2$ open-loop denotes the $x_1$ variance for constant $u$ and is given by $\frac{\sigma_{x_1}^2}{\langle x_1\rangle^2} = \frac{1}{\langle x_3\rangle}\frac{\tau_3}{\tau_1+\tau_3}$). Similarly from the two equations in (25), and the capacity bound in (18) $N_2/N_3$ can also be expressed in terms of the $\omega_c$ and $\tau_3/\tau_1$; hence figure 3 can be generated by varying these parameters.

is normalized by $N_3 = (\langle x_3 \rangle / \tau_3) \tau_1$, the average number of $x_3$ molecules made per $x_1$ lifetime. Note also that $N_2/N_3$ is equal to the ratio of the average rates of formation of $x_2$ and $x_3$ respectively.
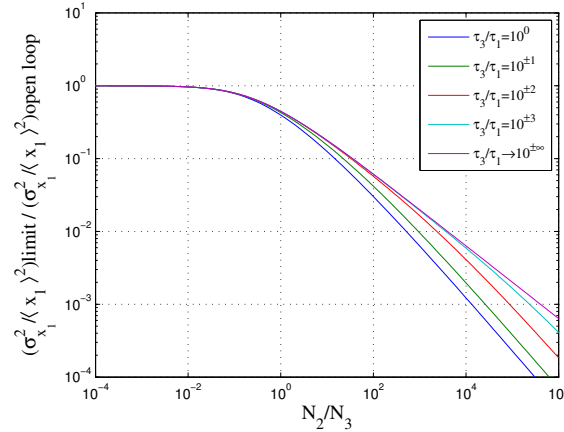


Figure 3: Lower limit on $\sigma_{x_1}^2 / \langle x_1 \rangle^2$ scaled by the value of $\sigma_{x_1}^2 / \langle x_1 \rangle^2$ when the control input $u$ is constant, i.e. no feedback is present. $N_2/N_3$ is equal to the ratio of the average rates of formation of $x_2$ and $x_3$ respectively.

For either $\tau_1 \gg \tau_3$ or $\tau_3 \gg \tau_1$ the quartic root bound in $\sigma_{x_1}$ still applies over a large range of relevant parameters, in the former case because the $x_3$ noise appears in the same way as intrinsic noise in the $x_1$ equation and in the latter case because $x_1$ simply follows the $x_3$ fluctuations, and the problem is then mathematically equivalent to suppressing the $x_3$ fluctuations. Note that, for large $\tau_3$, $N_3$ would be expected to be relatively small[19] - so in absolute terms the noise could then be regarded as being easy to suppress. For $\tau_1 \approx \tau_3$, though, it is easier to suppress the noise in the sense that the limit in $\sigma_{x_1}$ decreases with a higher rate with increasing $N_2$, however this is always smaller than the square root of $N_2$ ($\sigma_{x_1}/\langle x_1 \rangle > \text{const} \times N_2^{-3/8}$ for large $N_2$).

It should be noted that the $x_1$ variance limits in (19) and (20) also increase when the $x_1$ molecules are made in bursts, i.e. $x_1 \xrightarrow{u} x_1 + b_1$ where constant $b_1 \geq 1$ denotes the $x_1$ burst size. The stochastic differential equation (2) for $x_1$ becomes in this case

$$dx_1 = (b_1 u - x_1/\tau_1)dt + \sqrt{2r\langle x_1 \rangle / \tau_1}dw \,, \quad \text{where} \quad r = (1+b_1)/2 \geq 1 \quad (26)$$

Following the procedure in this section the $x_1$ variance bound in terms of the capacity (14) is now $\dfrac{\sigma_{x_1}^2}{\langle x_1 \rangle^2} \geq \dfrac{r}{\langle x_1 \rangle} \dfrac{1}{C\tau_1+1}$, and combining this with (18) we get the same bound as

---

[19]Parameter $N_3$ is proportional to $(1 + \tau_3/\tau_1)^{-1}$ for constant $(\sigma_{x_1}^2 / \langle x_1 \rangle^2)_{\text{open loop}}$ ( $(\sigma_{x_1}^2 / \langle x_1 \rangle^2)_{\text{open loop}} = 1/(N_3(1 + \tau_3/\tau_1))$ ).
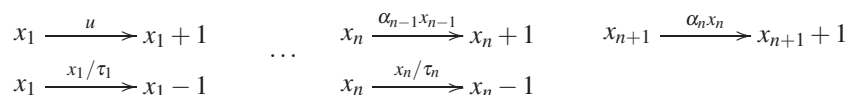
in (19), but with $\frac{1}{\langle x_1 \rangle}, \frac{1}{N_1}$ in the right hand side scaled by $r$, i.e.

$$\frac{\sigma_{x_1}^2}{\langle x_1 \rangle^2} \geq \frac{r}{\langle x_1 \rangle} \times \frac{2}{1 + \sqrt{1 + 4rN_2/N_1}}$$

# 5   Serial and parallel cascades

We investigate here the case where information about $x_1$ is conveyed by means of a cascade of multiple molecular species, rather than a single species $x_2$. We first consider a serial cascade where information is lost at each step, and then consider parallel signaling pathways where $x_1$ communicates with the control system through several parallel paths.

Each species is proportionally affecting the birth rate of the next species in the cascade and each one is also decaying exponentially, i.e.

$$x_1 \xrightarrow{\ u\ } x_1 + 1 \qquad \qquad x_n \xrightarrow{\alpha_{n-1}x_{n-1}} x_n + 1 \qquad x_{n+1} \xrightarrow{\alpha_n x_n} x_{n+1} + 1$$
$$x_1 \xrightarrow{x_1/\tau_1} x_1 - 1 \qquad \cdots \qquad x_n \xrightarrow{x_n/\tau_n} x_n - 1$$

We denote $N_i$ the average number of births of species $x_i$ in time $\tau_1$, i.e. $N_1 = \langle u \rangle \tau_1$, $N_i = \alpha_{i-1} \langle x_{i-1} \rangle \tau_1$ for $i = 2, \ldots, n+1$ (note that $N_{i+1} = \alpha_i \tau_i N_i$ using the equilibrium relations $\alpha_{i-1} \langle x_{i-1} \rangle = \langle x_i \rangle / \tau_i$, $\langle u \rangle = \langle x_1 \rangle / \tau_1$) ).

We first note that linearity of the transition rates in the cascade means that each molecule of a species $x_i$ decays exponentially and causes births of species $x_{i+1}$ with fixed rates $1/\tau_i$ and $\alpha_i$ respectively. Therefore the birth of a particular $x_2$ molecule can be inferred at the channel output only from its $x_{n+1}$ descendants. Now the probability $q$ that an $x_2$ molecule will lead to the birth of at least one $x_{n+1}$ molecule is fixed for given $N_i$, $i \geq 1$ (see Lemma 1). One the other hand, letting $\tau_i \to 0$ for $i = 2, \ldots, n$ will enable to recover the time of an $x_2$ birth from a corresponding $x_{n+1}$ birth with arbitrarily small probability of error, since the $x_{n+1}$ birth will occur in an arbitrarily small time after the $x_2$ birth (see Remark 4 in appendix C). We therefore calculate below the capacity of the channel with input $x_1$ and output $x_{n+1}$ in the limit $\tau_i \to 0$ for $i = 2, \ldots, n$, which is the regime where the capacity is maximized for given $N_i, i \geq 1$.

The fact that in this limiting regime the birth of an $x_{n+1}$ descendant occurs in an arbitrarily small time after the corresponding $x_2$ birth, also implies that the cascade will tend to a Poisson process with bursts

$$x_{n+1} \xrightarrow{\alpha_1 x_1} x_{n+1} + b$$
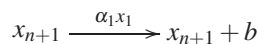
where $P(b > 0) = q = N_{\text{eff}}/N_2$ and

$$\frac{1}{N_{\text{eff}}} = \sum_{k=2}^{n+1} \frac{1}{N_k} \tag{27}$$
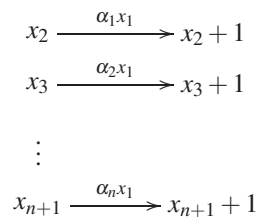
(see Lemma 1, Remark 4). As shown in Theorem 2 in appendix B, the capacity of this channel for a given mean and variance of $x_1$ is the same as the capacity of the

corresponding Poisson channel in Theorem 1 but scaled by $q$, i.e.

$$C = q\alpha_1 \langle x_1 \rangle \log\left(1 + \frac{\sigma_{x_1}^2}{\langle x_1 \rangle^2}\right) = q\frac{N_2}{\tau_1}\log\left(1 + \frac{\sigma_{x_1}^2}{\langle x_1 \rangle^2}\right)$$

Therefore the bound (18) for the capacity and also the lower bounds (19), (20) for the variance of $x_1$ remain the same but with $N_2$ replaced by $N_{\text{eff}}$.

Consider now the case of parallel signaling cascades where $x_1$ affects directly the birth rate of more than one signalling molecular species, with each of the $n$ birth processes having an intensity $\alpha_i x_1$ for $i = 1, \ldots, n$ respectively.

$$x_2 \xrightarrow{\alpha_1 x_1} x_2 + 1$$

$$x_3 \xrightarrow{\alpha_2 x_1} x_3 + 1$$

$$\vdots$$

$$x_{n+1} \xrightarrow{\alpha_n x_1} x_{n+1} + 1$$

This parallel channel is equivalent to a single birth process

$$y \xrightarrow{\lambda} y + 1$$
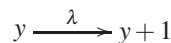
with intensity $\lambda = \sum_{i=1}^{n} \alpha_i x_1$ (a $y$ birth represents a birth in one of the species at the channel output), followed by a process that determines the species each $y$ birth corresponds to. This is an $x_k$ birth with probability $p_k$ proportional to the birth rate of the species, i.e., $p_k = \alpha_{k-1}/\sum_{i=1}^{n} \alpha_i$ depends only on the $\alpha_i$'s and is independent of $x_1$. Since this second process provides no further information about $x_1$, the capacity of the parallel channel is equal to that of the channel with input $x_1$ and output $y$. For a given mean and variance of $x_1$, this is equal to (Theorem 1)

$$C = \frac{N_{2\text{tot}}}{\tau_1}\log\left(1 + \frac{\sigma_{x_1}^2}{\langle x_1 \rangle^2}\right) \leq \frac{N_{2\text{tot}}}{\tau_1}\frac{\sigma_{x_1}^2}{\langle x_1 \rangle^2} \tag{28}$$

where $N_{2\text{tot}} = \sum_{i=1}^{n} \alpha_i \langle x_1 \rangle \tau_1$ is the total number of birth events during the interval $\tau_1$ affected directly by $x_1$. Comparing with the capacity bound in (18) when $x_1$ affects directly only a single birth reaction, we see that the corresponding variance bounds (19), (20) in section 4 remain the same but with $N_2$ replaced by $N_{2\text{tot}}$.

Finally it should be noted that bursts do not increase the capacity, i.e. if in the signalling molecule $x_2$ births occur with bursts

$$x_2 \xrightarrow{\alpha_1 x_1} x_2 + b$$
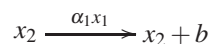
with the increment $b$ when a birth occurs being a random variable independent of $x_1$, the capacity of the Poisson channel with output $x_2$ and input $x_1$ does not increase relative to the case where the increments are fixed (see Theorem 2). The capacity also remains unchanged if $b > 0$. This can be important, as many biological processes involve large bursts of transcription, signaling, or transport.

# 6   Tradeoffs

In this section we analyze the noise in molecular numbers when specific feedback mechanisms are present, and discuss tradeoffs in the suppression of fluctuations.

An overview of the chemical master equation and the corresponding moment equations is given first. These are then used to analyze the direct inhibition mechanism discussed in the main text.

Consider a biochemical system with $n$ molecular species and $m$ elementary reactions

$$x \xrightarrow{W_i(x)} x + r_i , \qquad i = 1, 2, \ldots, m$$
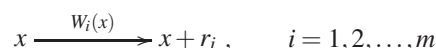
where $x = [x_1, x_2, \ldots, x_n]^T$ is the vector of molecule number with $x_j$ denoting the number of molecules of species $j$, $W_i(x)$ is the rate of reaction $i$, and the $j$'th element of $r_i \in \mathbb{Z}^n$ (this is denoted as $r_i^j$) gives the number of molecules by which species $j$ changes due to reaction $i$. The master equation for the system is

$$\frac{dP(k,t)}{dt} = \sum_i W_i(k - r_i)P(k - r_i, t) - W_i(k)P(k,t)$$

where $P(k,t)$ is the probability $x$ takes value $k$ at time t. This can be used to derive equations for the moments of $x$ which can be explicitly calculated when the transition rates $W_i(x)$ are linear functions of $x$.

More precisely, defining the fluxes

$$J_j^+ = \sum_{i: r_i^j > 0} r_i^j W_i(x), \quad J_j^- = \sum_{i: r_i^j < 0} |r_i^j| W_i(x), \quad J_j^{\text{tot}} = J_j^+ - J_j^-,$$

$$J^{\text{tot}} = [J_1^{\text{tot}}, \ldots, J_m^{\text{tot}}]^T \quad \text{and} \quad B_{js} = \sum_i r_i^j r_i^s W_i(x)$$

it can easily be shown [10] that

$$\frac{\partial \langle x \rangle}{\partial t} = \langle J^{\text{tot}} \rangle,$$

$$\frac{\partial \langle C \rangle}{\partial t} = \langle (J^{\text{tot}} - \langle J^{\text{tot}} \rangle) \times (x - \langle x \rangle)^T \rangle + \langle (x - \langle x \rangle) \times (J^{\text{tot}} - \langle J^{\text{tot}} \rangle)^T \rangle + \langle B \rangle$$

where $C$ is the covariance matrix, i.e. $C = \langle xx^T \rangle - \langle x \rangle \langle x^T \rangle$. When transition rates $W_i(x)$ are nonlinear functions of $x$ the equation for each moment will depend on higher moments and hence exact analytical solutions are not generally possible. A first order approximation that renders the problem analytically tractable is to linearize $J^{\text{tot}}$ about $\langle x \rangle$, i.e. $J^{\text{tot}}(x) \approx J^{\text{tot}}(\langle x \rangle) + [\partial J^{\text{tot}}(\langle x \rangle)/\partial x](x - \langle x \rangle)$ where $\partial J^{\text{tot}}(\langle x \rangle)/\partial x$ is the Jacobian of $J^{\text{tot}}$ evaluated at $\langle x \rangle$. This leads at steady state to a Lyapunov equation for the covariance matrix (a version of the Fluctuation Dissipation Theorem in physics), which can be interpreted by means of biological observables by introducing appropriate scaling in the equation parameters; i.e. , for $\eta_{ij} = C_{ij}/(\langle x_i \rangle \langle x_j \rangle)$ we have [11]
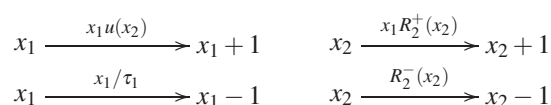
$$M\eta + \eta M^T = D, \tag{29}$$

where

$$D_{ij} = \frac{\langle B_{ij} \rangle}{\langle x_i \rangle \langle x_j \rangle}, \quad M_{ij} = \frac{\langle J_i^- \rangle}{\langle x_i \rangle} H_{ij}, \quad H_{ij} = \frac{\partial \ln(J_i^- / J_i^+)}{\partial \ln x_j} \tag{30}$$

with the partial derivative in $H_{ij}$ evaluated at $x = \langle x \rangle$. We also define $\tau_i = \frac{\langle x_i \rangle}{\langle J_i^- \rangle}$ as the average lifetime of species $x_i$.

We consider now the feedback system discussed in the main text, where two species $x_1$ and $x_2$ are in a negative feedback interconnection by means of a direct inhibition mechanism.

$$x_1 \xrightarrow{\;x_1 u(x_2)\;} x_1 + 1 \qquad x_2 \xrightarrow{\;x_1 R_2^+(x_2)\;} x_2 + 1$$
$$x_1 \xrightarrow{\;x_1 / \tau_1\;} x_1 - 1 \qquad x_2 \xrightarrow{\;R_2^-(x_2)\;} x_2 - 1$$

It should be noted that $u(x_2)$ at time $t$ is a static function of $x_2(t)$ rather than an arbitrary function of the history of $x_2$ up to time $t$, as in the bounds derived in the previous sections. Applying the approximate fluctuation dissipation theorem (29) we get the following expression for the variance of $x_1$

$$\frac{\sigma_{x_1}^2}{\langle x_1 \rangle^2} = \underbrace{\frac{1}{\langle x_1 \rangle} \left( \frac{H_{22}}{H_{12}} + \frac{\tau_2}{\tau_1} \frac{1}{H_{22}} \right)}_{\text{noise from } x_1} + \underbrace{\frac{1}{\langle x_2 \rangle} \frac{H_{12}}{H_{22}} \frac{\tau_2}{\tau_1}}_{\text{noise from } x_2} \tag{31}$$

Note that there is a tradeoff in the suppression of the two noise terms by varying $H_{11}$ and $H_{22}$. The $x_1$ variance is minimized by letting $H_{22} \to \infty$ and setting $H_{12}/H_{22} = \sqrt{N_2/N_1}$ where $N_1 = \langle x_1 \rangle$, $N_2 = \langle x_2 \rangle \tau_1 / \tau_2$, thus giving

$$\frac{\sigma_{x_1}^2}{\langle x_1 \rangle^2} \geq \frac{2}{\sqrt{N_1 N_2}}$$

# APPENDIX

In the appendix we provide detailed derivations of results discussed in sections 3-5. Appendix A gives on overview of Pinker's nonanticipatory epsilon entropy. In appendix B we provide a detailed proof for the capacity of a Poisson channel under mean and variance constraints. This is also derived for channels with bursts, and further extended in appendix C to cascades of multiple birth death processes.

## A    Causality constraints: the nonanticipatory epsilon entropy

A major plank to our development is Pinsker's notion of nonanticipatory epsilon entropy [6], [7]. This is a rate distortion function which takes into account a causality constraint.

Rate distortion theory becomes relevant when transmitted signals are not exactly recovered, but rather have a certain recovery error, called distortion. This leads to the notion of a rate distortion function (or Kolmogorov's epsilon entropy [12]). For a random process $\xi_0^T$ and an allowable distortion $\varepsilon$, the rate distortion function $R(\varepsilon; \xi_0^T)$ is defined as

$$R(\varepsilon; \xi_0^T) := \inf_{\eta_0^T} I(\xi_0^T; \eta_0^T) \quad \text{such that} \quad d(\xi_0^T, \eta_0^T) \leq \varepsilon \tag{32}$$

where random process $\eta_0^T$ is the estimate of the transmitted message $\xi_0^T$ and $d(\xi_0^T, \eta_0^T)$ is the distortion measure[20], i.e., $R(\varepsilon; \xi_0^T)$ is the minimum mutual information that needs to be transmitted in order to reproduce $\xi_0^T$ within a distortion $\varepsilon$. [21]

It should be noted that causality is often not a constraint when information theory is traditionally used to study communication systems, as is the case with the notion of a rate distortion function defined in (32). By this we mean that at any time $t$ the signal estimate $\eta_0^t$ in (32), contains also information about $\xi_t^T$, the future values of the transmitted message, rather than just its past values $\xi_0^t$. In practice, this refers to the fact that we wait to observe the whole of the signal $\xi_0^T$ before we encode it and transmit it through the channel, thus introducing a delay of at least $T$, before the estimate $\eta$ becomes available. Such delays might sometimes not be an issue in communication systems. They cannot, however, be tolerated when any inference about the transmitted message is needed in real time within a feedback control loop.

---

[20]In our analysis we will be using a distortion measure of the form $d(\xi_0^T, \eta_0^T) = \sup_{0 \leq t \leq T} E\{(\xi_t - \eta_t)^2\}$.

[21]Similarly, the distortion rate function $D(r; \xi_0^T)$ of signal $\xi_0^T$ is defined as

$$D(r; \xi_0^T) := \inf_{\eta_0^T} d(\xi_0^T, \eta_0^T) \quad \text{such that} \quad I(\xi_0^T; \eta_0^T) \leq r$$

i.e. the minimum distortion achievable for a given mutual information constraint between the transmitted signal $\xi_0^T$ and its estimate. It can be shown that the distortion rate function is the inverse of the rate distortion function, i.e. $R(D(r; \xi_0^T); \xi_0^T) = r$.

Causality becomes here a very real physical constraint that cannot be violated. At any time $t$ the message estimate $\eta_0^t$ carries no information about the future of the transmitted message $\xi$, other than information that can be inferred from its past $\xi_0^t$. This time ordering of the signals can be imposed by requiring $\eta_0^t, \xi_0^t, \xi_t^T$ to form a Markov chain, i.e. for all $t \in [0, T]$, signal $\eta_0^t$ conditioned on $\xi_0^t$, the past of $\xi$, is independent of its future $\xi_t^T$.

The nonanticipatory epsilon entropy $H_\varepsilon$ of a stochastic process $\xi_0^T$ is defined as the smallest mutual information between it and any other stochastic process $\eta_0^T$ under the constraints that the distortion between $\xi$ and $\eta$ is sufficiently small and that $\eta$ can only depend causally on $\xi$. To be specific

$$H_\varepsilon(\xi_0^T) := \inf I(\xi_0^T; \eta_0^T)$$

where the infimum is taken over stochastic processes $\eta_0^T$ s.t.

$$d(\xi_0^T, \eta_0^T) \le \varepsilon \tag{33}$$

$$\eta_0^t, \xi_0^t, \xi_t^T \quad 0 \le t \le T \text{ form a Markov chain} \tag{34}$$

with distortion $d(\xi_0^T, \eta_0^T)$ defined as $\quad d(\xi_0^T, \eta_0^T) := \sup_{0 \le t \le T} E\left\{(\xi_t - \eta_t)^2\right\}$

i.e., it corresponds to a rate distortion function with causality constraint (34).

This has been calculated in [7] for the case of a Markov Gaussian source. In the special case of a Wiener process $\xi$ satisfying

$$d\xi_t = \beta dw_t$$

where $\{w_t : t \ge 0\}$ is a standard Brownian motion we have[22]

$$\lim_{T \to \infty} \frac{1}{T} H_\varepsilon(\xi_0^T) = \frac{\beta^2}{2\varepsilon} \tag{35}$$

A distortion rate function with causality constraints can similarly be defined as

$$D_r(\xi_0^T) := \inf_{\eta_0^T} d(\xi_0^T, \eta_0^T) \quad \text{such that } \frac{1}{T} I(\xi_0^T; \eta_0^T) \le r \tag{36}$$

$$\text{(34) holds}$$

Making use of the fact that the distortion rate function is the inverse of the rate distortion function, solving for $\varepsilon$ in (35) we get a lower bound on the distortion for an upper bound $C$ on the mutual information rate[23], i.e.[24]

$$d(\xi_0^T, \eta_0^T) \ge \frac{1}{2C} \beta^2 \tag{37}$$

---

[22]The equality below is also a lower bound for any finite time i.e. $\frac{1}{T} H_\varepsilon(\xi_0^T) \ge \frac{\beta^2}{2\varepsilon}$ if $E\left\{(\xi_t)^2\right\} \ge \varepsilon$ at $t = 0$.

[23]This can be seen here as follows. Since $D_r(\xi_0^T)$ is the infimum distortion under constraints (36), (34), then for all $\delta > 0$ there exists $\eta_0^T$ such that (34) holds, $\frac{1}{T} I(\xi_0^T; \eta_0^T) \le r$ and $d(\xi_0^T, \eta_0^T) \le D_r(\xi_0^T) + \delta$. Hence $\frac{1}{T} H_{D_r(\xi_0^T)+\delta}(\xi_0^T) \le \frac{1}{T} I(\xi_0^T; \eta_0^T) \le r$.

[24]Note that without the causality constraint (34), the lower bound for the distortion in (37) would be reduced by a factor of $4/\pi^2$ (see [12], [13]).

Similar arguments for the first order source

$$d\xi_t = -b\xi_t + \beta dw_t, \tag{38}$$

yield the lower bound

$$d(\xi_0^T, \eta_0^T) \geq \frac{1}{2(C+b)}\beta^2. \tag{39}$$

# B    On the Capacity of the Poisson channel

Consider a Poisson process $n_t$, $t \in [0,T]$ with its intensity $\lambda_t$ being a function of the input signal to be transmitted $\theta_0^t$, and also the past values $n_0^t$ of the Poisson process, i.e. $n_t \xrightarrow{\lambda_t} n_t + 1$ with $\lambda_t = \lambda(t, \theta_0^t, n_0^t)$. Then the capacity of the Poisson channel with output $n_t$ is defined as

$$C := \sup_{\theta,\lambda} \frac{1}{T} I(\theta_0^T; n_0^T) \tag{40}$$

This is finite if appropriate constraints are imposed on the moments of the intensity $\lambda_t$.

**Theorem 1.** *The capacity of a Poisson channel with intensity $\lambda_t$ s.t. $E\{\lambda_t\} = m$, $Var\{\lambda_t\} \leq \sigma^2$ is given by*

$$C = m\log\left(1 + \frac{\sigma^2}{m^2}\right)$$

**Remark 1.** *The capacity with a maximum value constraint $0 \leq \lambda_t \leq c$ instead of a variance constraint was calculated in [4], [3]. This is equal to $C = c/e$ if the mean is unconstrained and $m\log(c/m)$ if the mean is also constrained as in Theorem 1.*

**Remark 2.** *Noting the inequality $\log(1+x) \leq x$, a simple upper bound for the capacity is given by $\sigma^2/m$.*

*Proof.* From the Liptser/Shiryaev formula [5] we have

$$I(\theta_0^T; n_0^T) = E\left\{\int_0^T \left[\phi(\lambda_t) - \phi(\hat{\lambda}_t)\right] dt\right\} \tag{41}$$

where $\phi(x) := x\log(x)$, $\hat{\lambda}_t = E\{\lambda_t/n_0^t\}$. Since $\phi$ is a convex function, Jensen's inequality gives

$$I(\theta_0^T; N_0^T) \leq \int_0^T \left[E\{\phi(\lambda_t)\} - \phi(E\{\hat{\lambda}_t\})\right] dt \tag{42}$$

Since $E\{\hat{\lambda}_t\} = m$ an upper bound is obtained by maximizing $\int_0^T E\{\phi(\lambda_t)\}dt$ under the two moment constraints specified in the Theorem. According to Hoeffding's theorem on the extrema of expectations [8] the maximum is achieved with a cumulative distribution that is a step function with 3 steps, i.e. the optimum distribution of $\lambda_t$ is that

of a discrete random variable that is nonzero at at most three points. Therefore an upper bound for $E\{\phi(\lambda_t)\}$ is obtained from the maximization problem below, where $p_i$ denotes the probability that $\lambda_t$ takes value $k_i$

$$\sup_{p_i,k_i} \sum_{i=1}^{M} p_i\phi(k_i)$$

$$\text{s.t.} \sum_i p_i k_i = m \tag{43}$$

$$\sum_i p_i k_i^2 \leq \sigma^2 + m^2 =: Q^2 \tag{44}$$

$$\sum_i p_i = 1, \quad p_i \geq 0, k_i \geq 0 \quad \forall i \tag{45}$$

The Lagrangian for the problem is

$$L = \sum \phi(k_i)p_i + \mu(\sum_i p_i k_i^2 - Q^2) + \nu(\sum_i k_i p_i - m) + \xi(\sum_i p_i - 1) - \sum_i \rho_i k_i - \sum_i \psi_i p_i$$

The Karush-Kuhn-Tucker conditions give necessary conditions for optimality, and these take the form

$$\frac{\partial L}{\partial p_i} = 0, \frac{\partial L}{\partial k_i} = 0, \ \mu \leq 0, \rho_i \leq 0, \psi_i \leq 0 \ \forall i,$$

$$\mu(\sum_i p_i k_i^2 - Q^2) = 0, \rho_i k_i = 0, \psi_i p_i = 0 \ \forall i$$

in addition to the primal feasibility constraints (43-45). Since $\phi(k_i)$ is not differentiable at $k_i = 0$ we consider the maximization problem above with $\phi(x)$ defined as $\phi(x) = (\alpha + x)\log(\alpha + x)$, for some $\alpha > 0$, and then take the limit $\alpha \to 0$.

$$\frac{\partial L}{\partial p_i} = 0 \ \text{ gives } \ \phi(k_i) + \mu k_i^2 + \nu k_i + \xi - \psi_i = 0. \tag{46}$$

For $p_i > 0$ we have $\psi_i = 0$ and hence (46) reduces to

$$(\alpha + k_i)\log(\alpha + k_i) + \mu k_i^2 + \nu k_i + \xi = 0 \tag{47}$$

$$\frac{\partial L}{\partial k_i} = 0 \ \text{ gives } \ p_i(\log(\alpha + k_i) + 1) + 2\mu p_i k_i + \nu p_i - \rho_i = 0 \tag{48}$$

For $k_i > 0$ then $\rho_i = 0$ hence for $p_i > 0$ we get

$$\log(\alpha + k_i) + 2\mu k_i + \nu + 1 = 0 \tag{49}$$

Also combining (47), (49) we get

$$\mu k_i^2 + k_i - \xi - \alpha \log(\alpha + k_i) = 0 \tag{50}$$

The system of equations (49), (50) has at most one solution when solved for $k_i$. To see this note that if $k_1$ and $k_2$ are two distinct solutions then, substituting these into (49),

(50) one can deduce that $\log x = 2\frac{x-1}{x+1}$ for $x = \frac{k_1+\alpha}{k_2+\alpha}$, which has a unique solution $x = 1$, i.e. $k_1 = k_2$. Hence $k_i$ is constant for all $i$ such that $k_i > 0$ and $p_i > 0$. We denote this value as $k_u$ and let $p_u := \sum_{i:k_i>0} p_i$. Note that a $k_i > 0$ with $p_i > 0$ will always exist unless $m = 0$, in which case the maximum value of the cost function is zero. In order to satisfy the primal constraints on the mean and variance, $p_i$ needs to be positive also for $i$ s.t. $k_i = 0$ (unless $Q^2 = m^2$, i.e. the variance is zero, in which case the solution is trivial). Substituting $k_i = 0$ in (47) we get $\xi + \alpha\log(\alpha) = 0$, and hence for $k_i > 0$ we get from equation (50) that as $\alpha \to 0$ then $\mu \to -1/k_u < 0$ i.e. the variance constraint is active. Therefore from the mean and variance constraints we get

$$p_u k_u = m, \quad p_u k_u^2 = \sigma^2 + m^2$$

So $k_u = (\sigma^2 + m^2)/m$. Now letting $\alpha \to 0$ the upper bound on the capacity in (42) is

$$E\{\phi(\lambda_t)\} - \phi(E\{\hat{\lambda}_t\}) = k_u p_u log(k_u) - m log m$$
$$= m log\left(1 + \frac{\sigma^2}{m^2}\right)$$

It is shown in [4], [3] that this bound is achievable arbitrarily closely without feedback, by means of a random telegraph signal with sufficiently fast transitions between 0 and $k_u$. $\qquad\square$

Consider now a Poisson channel as the one considered in Theorem 1, but when a birth occurs in point process $n_t$ the increment size is given by a discrete, non-negative, integer valued random variable $b$, i.e.

$$n_t \xrightarrow{\lambda_t} n_t + b \qquad (51)$$

with $b$ having a probability mass function $p_k = P(b = k)$. The capacity of the Poisson channel with output $n_t$ and input $\lambda_t$ is defined as in (40), but with $n_t$ being now the point process (51).

**Theorem 2.** *The capacity of the Poisson channel (51) with intensity $\lambda_t$ s.t. $E\{\lambda_t\} = m$, $Var\{\lambda_t\} \le \sigma^2$ is given by*

$$C = qm\log\left(1 + \frac{\sigma^2}{m^2}\right) \qquad (52)$$

*where $q = \sum_{k>0} p_k$ is the probability the increment $b$ is positive.*

*Proof.* The point process in (51) is equivalent to the superposition of point processes with fixed increments, equal to the various v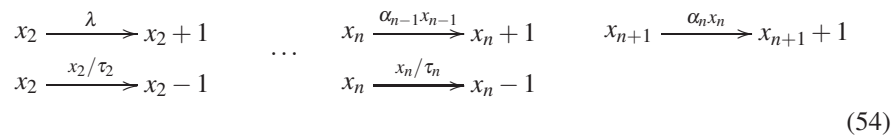alues $k$ random variable $b$ can take, i.e., $n_t = \sum_k n_k(t)$ where $n_k(t)$ is a birth processes of the form $n_k(t) \xrightarrow{p_k\lambda_t} n_k(t) + k$ . Since having no births, or having a birth with the increment $b = 0$, have the same effect on $n_t$ in (51), the point process is equivalent to one with strictly positive increments $\bar{b}$ and with the birth rate $\lambda_t$ scaled by $q$, i.e.

$$n_t \xrightarrow{q\lambda_t} n_t + \bar{b} \qquad (53)$$

where for $k > 0$, $P(\bar{b} = k) = P(b = k | b > 0) = p_k / q$. As $\bar{b}$ is independent of the input $\lambda_t$ the capacity of the channel from $\lambda_t$ to $n_t$ in (53) is equal to that of a Poisson channel with the same rate $q\lambda_t$ but unit increments, which is given from Theorem 1 by (52). $\square$

## C Cascades

Consider now the case where instead of a single Poisson process we have a cascade of birth/death processes as shown below

$$
x_2 \xrightarrow{\lambda} x_2 + 1 \qquad \cdots \qquad x_n \xrightarrow{\alpha_{n-1}x_{n-1}} x_n + 1 \qquad x_{n+1} \xrightarrow{\alpha_n x_n} x_{n+1} + 1
$$
$$
x_2 \xrightarrow{x_2/\tau_2} x_2 - 1 \qquad\qquad x_n \xrightarrow{x_n/\tau_n} x_n - 1
$$

$$(54)$$

This is equivalent to each molecule of species $x_i$, $i = 2, \ldots, n$ decaying exponentially and giving birth to molecules of species $x_{i+1}$ with fixed rates $1/\tau_i$ and $\alpha_i$ respectively. Note that a molecule $x_2$ will cause the birth of a molecule $x_{n+1}$ only with a certain probability which is less than 1, since it is likely that intermediate molecules decay before a birth occurs in $x_{n+1}$. This probability is calculated below.

**Lemma 1.** *Consider the cascade in (54). The probability $P_{n+1}$ that a particular molecule of species $x_2$ will cause the birth of at least one molecule of species $x_{n+1}$ is given by*

$$
\frac{1}{P_{n+1}} = N_2 \sum_{k=2}^{n+1} \frac{1}{N_k}
$$

*where for $k \geq 2$, $\quad N_{k+1} = N_2 \prod_{i=1}^{k} \alpha_k \tau_k$ .*

**Remark 3.** *Note that $P_{n+1}$ for a particular molecule of species $x_2$ does not change in the presence of other $x_2$ molecules since each molecule $x_i$ decays and causes births of $x_{i+1}$ molecules with fixed rates that do not depend on other molecules.*

*Proof.* Let $P_{n-k+3}$, for $k = 2, \ldots, n$ denote the probability that a particular molecule $x_k$ will cause the birth of at least one molecule $x_{n+1}$. We evaluate $P_n$ by deriving a recursion in $k$ for $P_{n-k+3}$. Note that

$$
P_{n+1} = \sum_{r=1}^{\infty} P_{3,r} [1 - Q_{3,r}] \quad \text{where}
$$
$$
P_{3,r} = P(x_2 \text{ causes the birth of } r \ x_3 \text{ before it decays})
$$
$$
Q_{3,r} = P(\text{none of the } r \ x_3 \text{ leads to a birth of } x_{n+1})
$$

We have

$$
P_{3,r} = \int_0^{\infty} \underbrace{\frac{1}{\tau_2} e^{-t/\tau_2}}_{\text{pdf of } x_2 \text{ decay time } t} \times \underbrace{e^{-\alpha_2 t} \frac{(\alpha_2 t)^r}{r!}}_{P(r \ x_3 \text{ in time } t)} dt \tag{55}
$$

$$
= \frac{(\alpha_2 \tau_2)^r}{(1 + \alpha_2 \tau_2)^{1+r}} \tag{56}
$$

and $Q_{3,r} = (1 - P_n)^r$. So

$$P_{n+1} = \sum_{r=1}^{\infty} \frac{(\alpha_2 \tau_2)^r}{(1 + \alpha_2 \tau_2)^{1+r}} \left[1 - (1 - P_n)^r\right]$$

$$= \frac{\alpha_2 \tau_2 P_n}{1 + \alpha_2 \tau_2 P_n}$$

Also in general for $k = 2, \dots, n$

$$P_{n-k+3} = \frac{\alpha_k \tau_k P_{n-k+2}}{1 + \alpha_k \tau_k P_{n-k+2}} \qquad \text{or} \qquad \frac{1}{P_{n-k+3}} = 1 + \frac{1}{\alpha_k \tau_k P_{n-k+2}} \qquad (57)$$

It can easily be shown using the arguments above that $1/P_3 = 1 + 1/(\alpha_n \tau_n)$ hence $P_{n+1}$ follows from the recursion (57). $\qquad \square$

**Remark 4.** *Note that if a birth in $x_{n+1}$ occurs due to an $x_2$ molecule formed at time $t_1$, the $x_{n+1}$ birth takes place in an infinitesimally small time after $t_1$, in the limit where $\tau_i \to 0$ for $i = 2, \dots, n$. This is because the integrand in (55) tends to a $\delta$-function centered arbitrarily close to 0. Therefore the number of molecules in the last species $x_{n+1}$ in the cascade will tend to satisfy a birth process $x_{n+1} \xrightarrow{\lambda} x_{n+1} + b$ with $P(b > 0)$, the probability of a positive increment given that a birth occurs, given by $P_{n+1}$.*

As discussed in section 5, it follows from Remarks 3, 4 that for given $N_i/N_2, i = 2, \dots, n+1$, the capacity of the channel from $\lambda$ to $x_{n+1}$ is maximized by letting $\tau_i \to 0$ for $i = 2, \dots, n$. This is because the probability $P_{n+1}$ of detecting an $x_2$ birth at the channel output remains unchanged, whereas, letting $\tau_i \to 0$ enables to recover the time of an $x_2$ birth with an arbitrarily small probability of error. Therefore the capacity of the Poisson channel with bursts discussed in Remark 4 gives an upper bound for the capacity of the cascade. For a given mean and variance of $\lambda$, this can be calculated from Theorem 2 and Lemma 1, and is given by (52) with $q = P_{n+1}$.

# References

[1] C. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[2] A. D. Wyner. Capacity and error exponent for the direct detection photon channel-Part I, II. *IEEE Transactions on Information Theory*, 34:1449–1471, 1988.

[3] YU. M. Kabanov. The capacity of a channel of the Poisson type. *Theory of Probability and its Applications*, 23(1):143–147, 1978.

[4] M. H. A. Davis. Capacity and Cutoff Rate for Poisson-Type Channels. *IEEE Transactions on Information Theory*, 26(6):710–715, 1980.

[5] R. S. Liptser and A. N. Shiryaev. *Statistics of Random Processes II: Applications*. Berlin, Germany: Springer-Verlang, 2nd edition, 2001. Original Russian edition published by Nauka, Moscow, 1974.

[6] A. K. Gorbunov and M. S. Pinsker. Nonanticipatory and prognostic epsilon entropies and message generation rates. *Problems of Information Transmission*, 9(3):184–191, 1973. A translation from Problemy Peredachi Informatsii.

[7] A. K. Gorbunov and M. S. Pinsker. Prognostic epsilon entropy of a gaussian message and a gaussian source. *Problems of Information Transmission*, 10(2):93–99, 1974. A translation from Problemy Peredachi Informatsii.

[8] W. Hoeffding. The extrema of the expected value of a function of independent random variables. *The Annals of Mathematical Statistics*, 26(2):268–275, 1955.

[9] S. Ihara. *Information theory for continuous systems*. World Scientiifc Publishing, Singapore, 1993.

[10] M. Lax. Fluctuations from nonequilibrium steady state. *Reviews of Modern Physics*, 32(1):25–64, 1960.

[11] J. Paulsson. Summing up the noise in gene networks. *Nature*, 427:415–418, 2004.

[12] A. Kolmogorov. On the Shannon theory of information transmission in the case of continuous signals. *IRE Transactions on Information Theory*, 2(4):102–108, 1956.

[13] T. Berger. Information Rates of Wiener Processes. *IEEE Transactions on Information Theory*, 16(2):134–139, 1970.