Published in final edited form as:
*Phys Rev E Stat Nonlin Soft Matter Phys*. 2008 July ; 78(1 Pt 1): 011910.

# Information capacity of genetic regulatory elements

**Gašper Tkačik**[1,2], **Curtis G. Callan Jr.**[1,2,3], and **William Bialek**[1,2,3]

[1] Joseph Henry Laboratories of Physics, Princeton University, Princeton, New Jersey 08544, USA

[2] Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey 08544, USA

[3] Princeton Center for Theoretical Physics, Princeton University, Princeton, New Jersey 08544, USA

## Abstract

Changes in a cell's external or internal conditions are usually reflected in the concentrations of the relevant transcription factors. These proteins in turn modulate the expression levels of the genes under their control and sometimes need to perform nontrivial computations that integrate several inputs and affect multiple genes. At the same time, the activities of the regulated genes would fluctuate even if the inputs were held fixed, as a consequence of the intrinsic noise in the system, and such noise must fundamentally limit the reliability of any genetic computation. Here we use information theory to formalize the notion of information transmission in simple genetic regulatory elements in the presence of physically realistic noise sources. The dependence of this "channel capacity" on noise parameters, cooperativity and cost of making signaling molecules is explored systematically. We find that, in the range of parameters probed by recent *in vivo* measurements, capacities higher than one bit should be achievable. It is of course generally accepted that gene regulatory elements must, in order to function properly, have a capacity of at least one bit. The central point of our analysis is the demonstration that simple physical models of noisy gene transcription, with realistic parameters, can indeed achieve this capacity: it was not self-evident that this should be so. We also demonstrate that capacities significantly greater than one bit are possible, so that transcriptional regulation need not be limited to simple "on-off" components. The question whether real systems actually exploit this richer possibility is beyond the scope of this investigation.

## I. INTRODUCTION

Networks of interacting genes coordinate complex cellular processes, such as responding to stress, adapting the metabolism to a varying diet, maintaining the circadian cycle, or producing an intricate spatial arrangement of differentiated cells during development [1–4]. The success of such regulatory modules is at least partially characterized by their ability to produce reliable responses to repeated stimuli or changes in the environment over a wide dynamic range, and to perform the genetic computations reproducibly, on either a day-by-day or generation time scale. In doing so the regulatory elements are confronted by noise arising from physical processes that implement such genetic computations, and this noise ultimately traces its origins back to the fact that the state variables of the system are concentrations of chemicals, and "computations" are really reactions between individual molecules, usually present at low copy numbers [5,6].

It is useful to picture the regulatory module as a device that, given some input, computes an output, which in our case will be a set of expression levels of regulated genes. Sometimes the inputs to the module are easily identified, such as when they are the actual chemicals that a system detects and responds to, for example chemoattractant molecules, hormones, or transcription factors (TFs). There are cases, however, when it is beneficial to think about the inputs on a more abstract level: in embryonic development we talk of "positional information" and think of the regulatory module as trying to produce a different gene expression footprint at each spatial location [7]; alternatively, circadian clocks generate distinguishable gene expression profiles corresponding to various phases of the day [1]. Regardless of whether we view the input as a physical concentration of some transcription factor or perhaps a position within the embryo, and whether the computation is complicated or as simple as an inversion produced by a repressor, we want to quantify its reliability in the presence of noise, and ask what the biological system can do to maximize this reliability.

If we make many observations of a genetic regulatory element in its natural conditions we are collecting samples drawn from a distribution $p(\mathcal{I},\mathcal{O})$, where $\mathcal{I}$ describes the state of the input and $\mathcal{O}$ the state of the output. Saying that the system is able to produce a reliable response $\mathcal{O}$ across the spectrum of naturally occurring input conditions, $p(\mathcal{I})$, amounts to saying that the dependency—either linear or strongly nonlinear—between the input and output is high, i.e., far from random. Shannon has shown how to associate a unique measure, the mutual information $I$, with the notion of dependency between two quantities drawn from a joint distribution [8–10]:

$$I(\mathcal{I};\mathcal{O})=\int\int d\mathcal{I}\, d\mathcal{O}\, p(\mathcal{I},\mathcal{O})\log_2\frac{p(\mathcal{I},\mathcal{O})}{p(\mathcal{I})p(\mathcal{O})}. \tag{1}$$

The resulting quantity is a measure in bits and is essentially the logarithm of the number of states in the input that produce distinguishable outputs given the noise. A device that has one bit of capacity can be thought of as an “on-off” switch, two bits correspond to four distinguishable regulatory settings, and so on. Although the input is usually a continuous quantity, such as nutrient concentration or phase of the day, the noise present in the regulatory element corrupts the computation and does not allow the arbitrary resolution of a real-valued input to propagate to the output; instead, the mutual information tells us how precisely different inputs are distinguishable to the organism.

Experimental or theoretical characterization of the joint distribution $p(\mathcal{I},\mathcal{O})$ for a regulatory module can be very difficult if the inputs and outputs exist in a high-dimensional space. We can proceed, nevertheless, by remembering that the building blocks of complex modules are much simpler, and finally must reduce to the point where a single gene is controlled by transcription factors that bind to its promoter region and tune the level of its expression. While taking a simple element out of its network will not be illuminating about how the network as a whole behaves in general—especially if there are feedback loops—there may be cases where the information flow is “bottlenecked” through a single gene, and its reliability will therefore limit that of the network. In addition, the analysis of a simple regulatory element will provide directions for taking on more complicated systems; see Ref. [11] for a recent related analysis.

Before proceeding, we wish to briefly review the current state of understanding of gene regulation through the lens of information theory. In particular, we would like to give several examples of theoretical models that have been used to successfully make connection with experimental measurements in various regulatory systems, and stress that the information capacities of such proposed mechanisms can be very different. Asked to name a particular “typical” number for the capacity of a simple genetic regulatory element, one is thus in difficulty, as we hope to illustrate below.

It would seem that requiring that the transcriptional apparatus provide at least two distinguishable levels of expression in a regulated gene (and thus one bit of capacity) is a safe bet for a lower bound on capacity, and this would be consistent with the notion of a gene as a switchable element. However, even this “obvious” view can be challenged: recent work, for example, has shown that adaptive behavior for a population of organisms could emerge even in the case of stochastic switching, where the output of the regulatory process is a (biased) random “guess” about the input [12,13]. This can provide information beneficial to the organism, even if the information capacity in the technical sense is significantly less than one bit.

A large body of work considers deterministic switch like behavior, reflecting the bias toward the study of genetic systems that are essential to the proper functioning of the organism and for which the requirement of reproducibility of responses argues for accurate, or even deterministic, control. Boolean networks, for example, provided interesting insights into the dynamics of the cell cycle [14], but, interestingly, they had to be extended to higher-than-binary logical functions (with some genes requiring three or four distinct levels of expression) in a model constructed to explain how the four gap genes respond to primary maternal morphogens in the fruit fly *Drosophila* wild type and several mutant phenotypes [15].

Assuming the noise is small enough, one can progress beyond two (or several) discrete states toward true continuous control (which has, formally at least, infinite information capacity). This level of description is typical of models grounded in mass-action kinetics, such as models of the cell cycle in fission yeast or of circadian clocks [16,17]. However, the behavior of the organism is ultimately often explained by the (qualitative) features of the phase portrait, such as the bifurcation points, stable states and limit cycles, without invoking the benefits of capacity available through continuous precision. In these models that do not include noise *ab initio*, it is important to check whether the qualitative behavior of the phase portrait is robust with respect to the addition of small amounts of noise; if so, then the model is robust, the continuous description was only a mathematical convenience, and the true capacity will be finite; if the model is qualitatively sensitive to small noise, however, the biological relevance of the model is questionable.

In this paper we prepare the framework in which the questions of information capacity and its heuristic interpretation as the number of distinguishable states of gene expression can be precisely formulated. Given what is known about noise in transcriptional regulation, we first want to check a simple, but important idea, namely that at least binary control *is* achievable; we then proceed to the computation of higher capacities with biologically realistic noise models. While answering the broader question about the diversity of genetic regulatory mechanisms is not our aim here—and the field also lacks such a correspondingly broad experimental survey of gene expression noise—we believe that our results nevertheless offer a nontrivial insight into the capacities and limitations of simple genetic regulatory elements.

## II. MAXIMIZING INFORMATION TRANSMISSION

Our aim is to understand the reliability of a simple genetic regulatory element, that is, of a single activator or repressor transcription factor controlling the expression level of its downstream gene. We will identify the concentration $c$ of the transcription factor as the only input, $\mathcal{I} \equiv \{c\}$, and the expression level of the downstream gene $g$ as the only relevant output, $\mathcal{O} \equiv \{g\}$. The regulatory element itself will be parametrized by the input-output kernel $p(g|c)$, i.e., the distribution [as opposed to a “deterministic” function $g = g(c)$ in the case of a noiseless system] of possible outputs given that the input is fixed to some particular level $c$. For each such kernel, we will then compute the maximum amount of information $I(c;g)$ that can be transmitted through it, and examine how this capacity depends on the properties of the kernel.

For the sake of discussion let us split the transcriptional regulatory processes into the “system” under study, i.e., the direct control of expression level $g$ by the concentration of transcription factor, $c$, and all the remaining regulatory processes, grouped collectively into an internal cellular “environment,” which could possibly include the downstream readout of $g$ and the control over the production of $c$. The input-output kernel of a simple regulatory element, $p(g|$

$c$), is determined by the properties of our system, namely, by the biophysics of transcription factor–DNA interaction, and transcription and translation for gene $g$. In contrast, the distribution of input transcription factor concentrations, $p_{\mathrm{TF}}(c)$, that the cell uses during its typical lifetime, is not a property of the $c \to g$ regulatory process directly, but of other processes constituting the internal environment; considered as an input into regulation of $g$, the cell's transcription factor expression footprint $p_{\mathrm{TF}}(c)$ can be viewed as its representation of the external world and internal state to which the cell responds by modulating its expression level of $g$. Together, the input-output kernel and the distribution of inputs define the joint distribution $p(c,g) = p(g|c)p_{\mathrm{TF}}(c)$, and consequently the mutual information of Eq. (1) between the input and the output, $I(c;g)$.

Maximizing the information between the inputs and outputs, which corresponds to our notions of reliability in representation and computation, will therefore imply a specific matching between the given input-output kernel and the distribution of inputs, $p_{\mathrm{TF}}(c)$. Because we consider our system, the genetic regulatory element $p(g|c)$, as given, we are going to ask about the context in which this element can be optimally used, i.e., we are going to find the optimal $p_{\mathrm{TF}}(c)$. Parenthetically, we note that $p(g|c)$and $p_{\mathrm{TF}}(c)$ are both internal to the cell and are thus potentially targets of either adaptation or evolutionary processes; as we argue below, the choice to treat $p(g|c)$as given and optimize $p_{\mathrm{TF}}(c)$ is a convenience that allows easier connection with the current state of experiments rather than a necessity. Then, if one believes that a specific regulatory element has been tuned for maximal information transmission, the optimal solution for the inputs, $p^*_{\mathrm{TF}}(c)$, and the resulting optimal distribution of output expression levels, $p^*_{\exp}(g)=\int dc\; p(g|c)p^*_{\mathrm{TF}}(c)$, become experimentally verifiable predictions. If, on the other hand, the system is not really maximizing information transmission, then the capacity achievable with a given kernel and its optimal input distribution, $I[\;p(g|c),\, p^*_{\mathrm{TF}}(c)]$, can still be regarded as a (hopefully revealing) upper bound on the true information transmission of the system.

One could also consider maximizing the information by holding the distribution of inputs fixed and adjusting the input-output kernel, and this is indeed the setup in which information transmission arguments have been applied before to explain adaptation in neural systems that encode visual stimuli [18–21]. In that context the organism does not control the ensemble of images that it perceives over its lifetime, but can tune the properties of its input-output kernel to the image statistics. If the matching between the input-output kernel and the input distribution is mutual, however, either measuring the input-output kernel and optimizing the input distribution, or, alternatively, measuring the input distribution and optimizing the kernel, should yield the information capacity of the system. In the case of gene regulation, we have good experimental access to the input-output relation $p(g|c)$, but not to the "natural" distribution of input concentrations, $p_{\mathrm{TF}}(c)$ with the exception of Refs. [22,23]); we therefore choose to infer $p(g|c)$ from experimental measurements directly and find a corresponding optimal $p^*_{\mathrm{TF}}(c)$.

During the past decades the measurements of regulatory elements have focused on recovering the mean response of a gene under the control of a transcription factor that had its activity modulated by experimentally adjustable levels of inducer or inhibitor molecules [24]. Typically, a sigmoidal response is observed with a single regulator, as in Fig. 1, and more complicated regulatory "surfaces" are possible when there are two or more simultaneous inputs to the system [25,26]. In our notation, these experiments measure the conditional average over the distribution of outputs, $\bar{g}(c) = \int dg\; gp(g|c)$. Developments in flow cytometry and single-cell microscopy enabled the experimenters to start tracking in time and across the population of cells the expression levels of fluorescent reporter genes and thus open a window into the behavior of fluctuations. Consequently, work exploring the noise in gene expression, or $\sigma^2_g(c)=\int dg(g-\overline{g})^2\,p(g|c)$, has begun to accumulate, on both the experimental and biophysical modeling sides [27–30]. The efforts to further characterize and understand this noise were renewed by theoretical work by Swain and co-workers [31] that has shown how to separate intrinsic and extrinsic components of the noise, i.e., the noise due to the stochasticity of the observed regulatory process in a single cell, and the noise contribution that arises because typical experiments make many single-cell measurements and the internal chemical environments of these cells differ across the population.

## A. Small-noise approximation

We start by showing how the optimal distributions can be computed analytically if the input-output kernel is Gaussian and the noise is small, and proceed by presenting the exact numerical solution later. Let us assume then that the first and second moments of the conditional distribution are given, and write the input-output kernel as a set of Gaussian distributions $\mathcal{G}(g;\bar{g}(c), \sigma_g(c))$, or explicitly,

$$p(g|c)=\frac{1}{\sqrt{2\pi\sigma_g^2(c)}}\exp\left(-\frac{[g-\bar{g}(c)]^2}{2\sigma_g^2(c)}\right), \tag{2}$$

where both the mean response $\bar{g}(c)$ and the noise $\sigma_g(c)$ depend on the input, as illustrated in Fig. 1.

We rewrite the mutual information between the input and the output of Eq. (1) in the following way:

$$I(c;g)=\int dc\, p_{\mathrm{TF}}(c)\int dg\, p(g|c)\log_2 p(g|c) - \int dc\, p_{\mathrm{TF}}(c)\int dg\, p(g|c)\log_2 p_{\exp}(g). \tag{3}$$

The first term can be evaluated exactly for Gaussian distributions, $p(g|c) = \mathcal{G}(g;\bar{g}(c), \sigma_g(c))$. The integral over $g$ is just the calculation of the (negative of the) entropy of the Gaussian, and the first term therefore evaluates to $-\langle S[\mathcal{G}(g;\bar{g},\sigma_g]\rangle_{p_{\mathrm{TF}}(c)} = -\frac{1}{2}\langle\log_2 2\pi e\sigma_g^2(c)\rangle_{p_{\mathrm{TF}}(c)}$.

In the second term of Eq. (3), the integral over $g$ can be viewed as calculating $\langle\log_2 p_{\exp}(g)\rangle$ under the distribution $p(g|c)$. For an arbitrary continuous function $f(g)$ we can expand the integrals with the Gaussian measure around the mean:

$$\langle f(g)\rangle_{\mathcal{G}(g;\bar{g},\sigma_g)}=\int dg\,\mathcal{G}(g)f(\bar{g})+\int dg\,\mathcal{G}(g)\frac{\partial f}{\partial g}\Big|_{\bar{g}}(g-\bar{g})+\frac{1}{2}\int dg\,\mathcal{G}(g)\frac{\partial f^2}{\partial g^2}\Big|_{\bar{g}}(g-\bar{g})^2+\cdots. \tag{4}$$

The first term of the expansion simply evaluates to $f(\bar{g})$. The series expansion would end at the first term if we were to take the small-noise limit, $\lim_{\sigma_g\to 0}\mathcal{G}(g;\bar{g}, \sigma_g) = \delta(g - \bar{g})$. The second term of the expansion is zero because of symmetry, and the third term evaluates to $\frac{1}{2}\sigma_g^2 f''(\bar{g})$. We apply the expansion of Eq. (4) and compute the second term in the expression for the mutual information, Eq. (3), with $f(g) = \log_2 p_{\exp}(g)$. Taking only the zeroth order of the expansion, we get

$$I(c;g)=-\int dc\, p_{\mathrm{TF}}(c)[\log_2\sqrt{2\pi e}\sigma_g(c)+\log_2 p_{\exp}(\bar{g})]. \tag{5}$$

We can rewrite the probability distributions in terms of $\bar{g}$, using $p_{\mathrm{TF}}(c)dc = \hat{p}_{\mathrm{exp}}(\bar{g})d\bar{g}$ (the caret denotes the distribution of *mean* expression levels), and assume that, in the small-noise approximation, $p_{\mathrm{exp}}(\bar{g}) = \bar{p}_{\mathrm{exp}}(\bar{g})$. To maximize the information transmission we form the following Lagrangian and introduce the multiplier Λ that keeps the resulting distribution normalized:

$$\mathcal{L}[\widehat{p}_{\mathrm{exp}}(\overline{g})]= -\int d\overline{g}\,\widehat{p}_{\mathrm{exp}}(\overline{g})\log_2[\sqrt{2\pi e}\sigma_g(\overline{g})\widehat{p}_{\mathrm{exp}}(\overline{g})] - \Lambda\int d\overline{g}\,\widehat{p}_{\mathrm{exp}}(\overline{g}). \tag{6}$$

The optimal solution is obtained by taking a variational derivative with respect to $\hat{p}_{\mathrm{exp}}(\bar{g})$, $\frac{\delta\mathcal{L}[\widehat{p}_{\mathrm{exp}}(\overline{g})]}{\delta\widehat{p}_{\mathrm{exp}}(\overline{g})}=0$. The solution is

$$\widehat{p}^{*}_{\mathrm{exp}}(\overline{g})=\frac{1}{Z}\frac{1}{\sigma_g(\overline{g})}. \tag{7}$$

By inserting the optimal solution, Eq. (7), into the expression for mutual information, Eq. (3), we get an explicit result for the capacity:

$$I_{\mathrm{opt}}(c;g)=\log_2\left(\frac{Z}{\sqrt{2\pi e}}\right), \tag{8}$$

where $Z$ is the normalization of the optimal solution in Eq. (7):

$$Z=\int_0^1\frac{d\overline{g}}{\sigma_g(\overline{g})}. \tag{9}$$

The optimization with respect to the distribution of inputs, $p_{\mathrm{TF}}(c)$, has led us to the result for the optimal distribution of *mean outputs*, Eq. (7). We had to assume that the input-output kernel is Gaussian and that the noise is small, and we refer to this result as the small-noise approximation (SNA) for channel capacity. Note that in this approximation only the knowledge of the noise in the output as a function of mean output, $\sigma_g(\bar{g})$, matters for capacity computation, and the direct dependence on the input $c$ is irrelevant. This is important because the behavior of intrinsic noise as a function of the mean output is an experimentally accessible quantity [28]. Note also that for big enough noise the normalization constant $Z$ will be small compared to $\sqrt{2\pi e}$, and the small-noise capacity approximation of Eq. (8) will break down by predicting negative information values.

## B. Large-noise approximation

Simple regulatory elements usually have a monotonic, saturating input-output relation, as shown in Fig. 1, and (at least) a shot noise component whose variance scales with the mean. If the noise strength is increased, the information transmission must drop and, even with the optimally tuned input distribution, eventually yield only a bit or less of capacity. Intuitively, the best such a noisy system can do is to utilize only the lowest and highest achievable input concentrations, and ignore the continuous range in between. Thus, the mean responses will be as different as possible, and the noise at low expression will also be low because it scales with the mean. More formally, if only $\{c_{\mathrm{min}}, c_{\mathrm{max}}\}$ are used as inputs, then the result is either $p(g|c_{\mathrm{min}})$ or $p(g|c_{\mathrm{max}})$; the optimization of channel capacity reduces to finding $p_{\mathrm{TF}}(c_{\mathrm{min}})$, with $p_{\mathrm{TF}}(c_{\mathrm{max}}) = 1 - p_{\mathrm{TF}}(c_{\mathrm{min}})$. This problem can be solved by realizing that each of the two possible input concentrations produces its respective Gaussian output distributions, and by maximizing information by varying $p_{\mathrm{TF}}(c_{\min})$. Simplifying even further, we can threshold the outputs and allow $g$ to take on only two values instead of a continuous range; then each of the two possible inputs "min" and "max" maps into two possible outputs, on and off, and confusion in the channel arises because the min input might be misunderstood as on output, and vice versa, with probabilities given by the output distribution overlaps, as shown schematically in Fig. 2.

In the latter case we can use the analytic formula for the capacity of the binary asymmetric channel. If $\eta$ is the probability of detecting an off output if max input was sent, and $\xi$ is the probability of receiving an off output if min input was sent, and $H(\cdot)$ is a binary entropy function,

$$H(p) = -p\log_2 p - (1-p)\log_2(1-p), \tag{10}$$

then the capacity of this asymmetric channel is [32]

$$I(c;g) = \frac{-\eta H(\xi) + \xi H(\eta)}{\eta - \xi} + \log_2\left(1 + 2^{[H(\xi) - H(\eta)]/(\eta - \xi)}\right). \tag{11}$$

Because this approximation reduces the continuous distribution of outputs to only two choices, on or off, it can underestimate the true channel capacity and is therefore a lower bound.

### C. Exact solution

The information between the input and output in Eq. (3) can be maximized numerically for any input-output kernel $p(g|c)$ if the variables $c$ and $g$ are discretized, making the solution space that needs to be searched, $p_{\mathrm{TF}}(c_i)$, finite. One possibility is to use a gradient descent-based method and make sure that the solution procedure always stays within the domain boundaries $\Sigma_i p_{\mathrm{TF}}(c_i) = 1$, $p_{\mathrm{TF}}(c_j) \geq 0$ for every $j$. Alternatively, a procedure known as the Blahut-Arimoto algorithm has been derived specifically for the purpose of finding optimal channel capacities [33]. Both methods yield consistent solutions, but we prefer to use the second one because of faster convergence and convenient inclusion of constraints on the cost of coding (see Appendix A for details).

One should be careful in interpreting the results of such naive optimization and worry about the artifacts introduced by discretization of input and output domains. After discretization, the formal optimal solution is no longer required to be smooth and could, in fact, be composed of a collection of Dirac $\delta$ function spikes. On the other hand, the real, physical concentration $c$ cannot be tuned with arbitrary precision in the cell; it is a result of noisy gene expression, and even if this noise source were removed, the *local* concentration at the binding site would still be subject to fluctuations caused by randomness in diffusive flux [34,35]. The Blahut-Arimoto algorithm is completely agnostic as to which (physical) concentrations belong to which bins after concentration has been discretized, and so it could assign wildly different probabilities to concentration bins that differ in concentration by less than $\sigma_c$ (i.e., the scale of local concentration fluctuations), making such a naive solution physically unrealizable. In Appendix A we describe how to properly use the Blahut-Arimoto algorithm despite the difficulties induced by discretization.

## III. A MODEL OF SIGNALS AND NOISE

If enough data were available, one could directly sample $p(g|c)$ and proceed by calculating the optimal solutions as described previously. Here we start, in contrast, by assuming the Gaussian model of Eq. (2), in which the mean, $\bar{g}(c)$, and the output variance, $\sigma_g^2(c)$, are functions of the transcription factor concentration $c$. Our goal for this section is to build an effective microscopic model of transcriptional regulation and gene expression, and therefore define both functions with a small number of biologically interpretable parameters. In the subsequent discussion we plan to vary those and thus systematically observe the changes in information capacity.

In the simplest picture, the interaction of the TF with the promoter site consists of binding with a (second-order) rate constant $k_+$ and unbinding at a rate $k_-$. In a somewhat more complicated case where $h$ TF molecules cooperatively activate the promoter, the analysis still remains simple as long as the favorable interaction energy between the TFs is sufficient to make only the fully occupied (and thus activated) and the empty (and thus inactivated) states of the promoter likely; this effective two-state system is once more describable with a single rate for switching off the promoter, $k_-$, and the corresponding activation rate has to be proportional to $c^h$ (see Ref. [35], in particular Appendix B). Generally, therefore, the equilibrium occupancy of the site will be

$$n=\frac{c^h}{c^h+K_d^h}, \tag{12}$$

where the Hill coefficient $h$ captures the effects of cooperative binding, and $K_d$ is the equilibrium constant of binding. The mean expression level $g$ is then

$$g(c)=g_0\bar{g}=g_0\begin{cases} n, & \text{activator,} \\ 1-n, & \text{repressor,} \end{cases} \tag{13}$$

where $\bar{g}$ has been normalized to vary between 0 and 1, and $g_0$ is the maximum expression level. In what follows we will assume the activator case, where $\bar{g} = n$, and present the result for the repressor at the end.

The fluctuations in occupancy have a (binomial) variance $\sigma_n^2=n(1-n)$ and a correlation time $\tau_c = 1/(k_+c^h+k_-)$ [35]. If the expression level of the target gene is effectively determined by the average of the promoter site occupancy over some window of time $\tau_{\text{int}}$, then the contribution to variance in the expression level due to the on-off promoter switching will be

$$\left(\frac{\sigma_g}{g_0}\right)^2=\sigma_n^2\frac{\tau_c}{\tau_{\text{int}}}=\frac{n(1-n)}{(k_+c^h+k_-)\tau_{\text{int}}}=\frac{n(1-n)^2}{k_-\tau_{\text{int}}}, \tag{14}$$

where in the last step we use the fact that $k_+c_{\,h}(1 - n) = k_-n$.

At low TF concentrations the arrival times of single transcription factor molecules to the binding site are random events. Recent measurements [22] as well as simulations [36] seem to be consistent with the hypothesis that this variability in diffusive flux contributes an additional noise term [34,35,37,46], similar to the Berg-Purcell limit to chemoattractant detection in chemotaxis. The noise in expression level due to fluctuations in the binding site occupancy, or the total *input* noise, is therefore a sum of this diffusive component [see Eq. (11) of Ref. [35]] and the switching component of Eq. (14):

$$\left(\frac{\sigma_g}{g_0}\right)^2_{\text{input}}=\frac{n(1-n)^2}{k_-\tau_{\text{int}}}+\frac{h^2(1-n)^2n^2}{\pi Dac\tau_{\text{int}}}, \tag{15}$$

where $D$ is the diffusion constant for the TF and $a$ is the receptor site size ($a \sim 3$ nm for a typical binding site on the DNA); see Ref. [38] for the case where the transport of TF molecules to the binding site involves one-dimensional (1D) diffusion along the DNA contour in addition to the 3D diffusion in the cytoplasm.

To compute the information capacity in the small-noise limit using the simple model developed so far, we need the constant $Z$ from Eq. (9), which is defined as an integral over expression levels. As both input noise terms are proportional to $(1-\bar{g})^2$, the integral must take the form

$$Z\propto\int_0^1\frac{d\bar{g}}{(1-\bar{g})F(\bar{g})}, \tag{16}$$

where $F(\bar{g})$ is a function that approaches a constant as $\bar{g}\rightarrow 1$. Strangely, we see that this integral diverges near full induction ($\bar{g}=1$), which means that the information capacity also diverges.

Naively, we expect that modulations in transcription factor concentration are *not* especially effective at transmitting regulatory information once the relevant binding sites are close to complete occupancy. More quantitatively, the sensitivity of the site occupancy to changes in TF concentration, $\partial n/\partial c$, vanishes as $n\rightarrow 1$, and hence small changes in TF concentration will have vanishingly small effects. Our intuition breaks down, however, because in thinking only about the mean occupancy we forget that even very small changes in occupancy could be effective if the noise level is sufficiently small. As we approach complete saturation, the variance in occupancy decreases, and the correlation time of fluctuations becomes shorter and shorter; together these effects cause the standard deviation as seen through an averaging time $\tau_{\text{int}}$ to decrease faster than $\partial n/\partial c$, and this mismatch is the origin of the divergence in information capacity. Of course, the information capacity of a physical system cannot really be infinite; there must be an extra source of noise (or reduced sensitivity) that becomes limiting as $n\rightarrow 1$.

The noise in Eq. (15) captures only the *input* noise, i.e., the noise in the protein level caused by the fluctuations in the occupancy of the binding site. In contrast, the *output* noise arises even when the occupancy of the binding site is fixed (for example, at full induction), and originates in the stochasticity in transcription and translation. The simplest model postulates that when the activator binding site is occupied with fractional occupancy $n$, mRNA molecules are synthesized in a Poisson process at a rate $R_e$ that generates $R_e\tau_e n$ mRNA molecules on average during the lifetime of a single mRNA molecule, $\tau_e$. Every message is a template for the production of proteins, which is another Poisson process with rate $R_g$. If the integration time is larger than the lifetime of single mRNA molecules, $\tau_{\text{int}}\gg\tau_e$, the mean number of proteins produced is $g=R_g\tau_{\text{int}}R_e\tau_e n=g_0 n$, and the variance associated with both Poisson processes is [35]

$$\left(\frac{\sigma_g}{g_0}\right)^2_{\text{output}}=\frac{1+R_g\tau_e}{g_0}n, \tag{17}$$

where $b = R_g\tau_e$ is the burst size, or the number of proteins synthesized per mRNA.

We can finally put the results together by adding the input noise Eq. (15) and the output noise Eq. (17), and expressing both in terms of the normalized expression level $\bar{g}(c)$. We consider both activators and repressors, following Eq. (13):

$$\left(\frac{\sigma_g}{g_0}\right)^2_{\text{act}} = \alpha\overline{g} + +\beta(1-\overline{g})^{2+1/h}\overline{g}^{2-1/h} + \gamma\overline{g}(1-\overline{g})^2, \tag{18}$$

$$\left(\frac{\sigma_g}{g_0}\right)^2_{\text{rep}} = \alpha\overline{g} + +\beta(1-\overline{g})^{2-1/h}\overline{g}^{2+1/h} + \gamma\overline{g}^2(1-\overline{g}), \tag{19}$$

with the relevant parameters $\{\alpha, \beta, \gamma, h\}$ explained in Table I. Note that both repressor and activator cases differ only in the shape of the input noise contributions (especially for low cooperativity $h$). Note further that the output noise increases monotonically with mean expression $\bar{g}$, while the input noise peaks at intermediate levels of expression.

In addition to the noise sources that we have discussed here, there can be other sizable contributions to the total noise in the output gene expression level, $\sigma_g(c)$, such as the global variations in protein concentrations and transmitted noise from the fluctuations in the production of the regulating TF, among others; see, e.g., Ref. [39] for the analysis of noise propagation and transmitted noise, the work by Elowitz and colleagues [28] in the synthetic *lac* system where global noise seems to be important, or the general discussion about extrinsic noise in [31]. For any particular genetic regulatory element for which there is experimental data on noise and for which one wants to compute the information capacity, the noise model has to account for all those terms that contribute significantly to the total noise in the output, $\sigma_g(c)$, across the whole range of input concentrations $c$. In this theoretical paper we focus on the output ($\alpha$) and diffusional input ($\beta$) noise contributions, in order to make the noise parameter space easy to analyze and visualize. Furthermore, these two noise sources are sufficient to explain the experimental data of Refs. [22,40], which we will use to illustrate our theoretical results. It has been easy to extend the noise models and capacity calculations to include global noise (which scales as $\sigma_g \propto \bar{g}$), or a nonzero amount of input switching noise (parametrized by $\gamma$) for the systems in which these contributions are significant, although these additions do not change the capacity results in a qualitatively interesting way [41].

# IV. RESULTS

## A. Capacity of simple regulatory elements

Having at our disposal both a simple model of signals and noise and a numerical way of finding the optimal solutions given an arbitrary input-output kernel, we are now ready to examine the channel capacity as a function of the noise parameters from Table I. Our first result, shown in Fig. 3, concerns the simplest case of an activator with no cooperativity, $h = 1$; for this case, the noise in Eq. (18) simplifies to

$$\left(\frac{\sigma_g}{g_0}\right)^2 = \alpha\overline{g} + \beta(1-\overline{g})^3\overline{g}. \tag{20}$$

Here we have assumed that there are two relevant sources of noise, i.e., the output noise (which we parametrize by $\alpha$ and plot on the horizontal axis) and the input diffusion noise (parametrized by $\beta$, vertical axis). Each point of the noise plane in Fig. 3(a) therefore represents a system characterized by a Gaussian noise model, Eq. (2), with variance given by Eq. (20) above.

As expected, the capacity increases most rapidly when the origin of the noise plane is approached approximately along its diagonal, whereas along each of the edges one of the two noise sources effectively disappears, leaving the system dominated by either output or input noise alone. We pick two illustrative examples, the blue and the red systems of Figs. 3(b) and 3(c), that have realistic noise parameters. The blue system has, apart for the decreased cooperativity ($h$ = 1 instead of 5), the characteristics of the Bicoid-Hunchback regulatory element in *Drosophila melanogaster* [23,35]; the red system is dominated by output noise with characteristics measured recently for about 40 yeast genes [40]. We would like to emphasize that both the small-noise approximation and the exact solution predict that these realistic systems are capable of transmitting more than 1 bit of regulatory information and that they, indeed, could transmit up to about 2 bits. In addition, we are also reminded that while the distributions [for example, the optimal output distribution in Fig. 3(b)] can look bimodal and this has often been taken as an indication that there are two relevant states of the output, such distributions really can have capacities above 1 bit; similarly, distributions without prominent features, such as the monotonically decreasing optimal output distribution of Fig. 3(c), should also not be expected necessarily to have low capacities.

A closer look at the overall agreement between the small-noise approximation [dashed lines in Fig. 3(a)] and the exact solution (thick lines) shows that the small-noise approximation underestimates the true capacity, consistent with our remark that for large noise the approximation will incorrectly produce negative results; at the 2-bit information contour the approximation is about ~15% off but improves as the capacity is increased.

In the high-noise regime we are making yet another approximation, the validity of which we now need to examine. In our discussion about the models of signals and noise we assumed that we can talk about the fractional occupancy of the binding site and *continuous* concentrations of mRNA, transcription factors and protein, instead of counting these species in discrete units, and that noise can effectively be treated as Gaussian. Both of these assumptions are the cornerstones of the Langevin approximation for calculating the noise variance [42]. If the parameters $\alpha$ and $\beta$ actually arise due to the underlying microscopic mechanisms described in Sec. III on signals and noise, we expect that at least for some large-noise regions of the noise plane the discreteness in the number of mRNA molecules will become important and the Langevin approximation will fail. In such cases (a much more time-consuming) exact calculation of the input-output relations using the master equation is possible for some noise models (see Appendix B); we show that in the region where $\log \alpha > -2$ the channel capacities calculated with Gaussian kernels can be overestimated by ~10% or more; there the Langevin calculation gives the correct second moment, but misses the true shape of the distribution. Although both examples with realistic noise parameters, Figs. 3(b) and 3(c), lie safely in the region where the Langevin approximation is valid, care should be used whenever both output noise $\alpha$ and burst size are large, and $\alpha$ is consequently dominated by a small number of transcripts.

Is there any difference between activators and repressors in their capacity to convey information about the input? We concluded Sec. III on the noise models with separate expressions for activator noise, Eq. (18), and repressor noise, Eq. (19); focusing now on the repressor case, we recompute the information in the same manner as we did for the activator in Fig. 3(a), and display the difference between the capacities of the repressor and activator with the same noise parameters in Fig. 4. As expected, the biggest difference occurs above the main diagonal, where the input noise dominates over the output noise. In this region the capacity of the repressor can be bigger by as much as third than that of the corresponding activator. Note that, as $h \to \infty$, the activator and repressor noise expressions become indistinguishable and the difference in capacity vanishes for the noise models with output and input diffusion noise contributions, Eqs. (18) and (19).

The behavior of the regulatory element is conveniently visualized in Fig. 5 by plotting a cut through the noise plane along its main diagonal. Moving along this cut scales the total noise variance of the system up or down by a multiplicative factor, and allows us to observe the overall agreement between the exact solution and small- and large-noise approximations. In addition we point out the following interesting features of Fig. 5 that will be examined more closely in subsequent sections.

First, the parameter region in the noncooperative case, in which the capacity falls below 1 bit and the large-noise approximation is applicable, is small and shrinks further at higher cooperativity. This suggests that a biological implementation of a reliable binary channel could be relatively straightforward, assuming our noise models are appropriate. Moreover, there exist distributions not specifically optimized for the input-output kernel, such as the input distribution uniform in $\ln(c/K_d)$ that we pick as an illustrative example in Fig. 5 (thick black line); we find that this simple choice can achieve considerable information transmission, and are therefore motivated to raise a more general question about the sensitivity of channel capacity with respect to perturbations in the optimal solution $p^*_{\mathrm{TF}}(c)$. We reconsider this idea more systematically in the next section.

Second, it can be seen from Fig. 5 that at small noise the cooperativity has a minor effect on the channel capacity. This is perhaps unexpected, as the shape of the mean response $\bar{g}(c)$ strongly depends on $h$. We recall, however, that mutual information $I(c;g)$ is invariant to any invertible reparametrization of either $g$ or $c$. In particular, changing the cooperativity or the value of the equilibrium binding constant $K_d$ in theory only results in an invertible change in the input variable $c$, and therefore the change in the steepness or midpoint of the mean response must not have any effect on $I(c;g)$. This argument does break down in the high-noise regime, where the cooperative system achieves capacities above 1 bit while the noncooperative system fails to do so. Reparametrization invariance would work only if the input concentration could extend over the whole positive interval, from zero to infinity. The substantial difference between capacities of cooperative and non-cooperative systems in Fig. 5 at *low capacity* stems from the fact that in reality the cell (and our computation) is limited to a finite range of concentrations $c \in [c_{\mathrm{min}}, c_{\mathrm{max}}]$, instead of the whole positive half axis $c \in [0,\infty)$. We explore the issue of limited input dynamic range further in the next section.

Finally, we draw attention to the simple linear scaling of the channel capacity with the logarithm of the total noise strength in the small-noise approximation, as explained in the caption of Fig. 5. In general, increasing the number of input and output molecules by a factor of 4 will decrease the relative input and output noise by a factor of $\sqrt{4}=2$, and therefore, in the small-noise approximation, increase the capacity by $\log_2 2=1$ bit. If one assumes that the cell can make transcription factor and output protein molecules at no cost, then scaling of the noise variance along the horizontal axis of Fig. 5 is inversely proportional to the total number of signaling molecules used by the regulatory element, and its capacity can grow without bound as more and more signaling molecules are used. If, however, there are metabolic or time costs to making more molecules, our optimization needs to be modified appropriately, and we present the relevant computation in Sec. IV C on the costs of coding.

### B. Cooperativity, dynamic range, and the tuning of solutions

In the analysis presented so far we have not paid any particular attention to the question of whether the optimal input distributions are biologically realizable or not. We will proceed to relax some of the idealizations made until now and analyze the corresponding changes in the information capacity.

We start by considering the impact on channel capacity of changing the allowed dynamic range to which the input concentration is restricted. Figure 6(a) displays the capacity as a function of the dynamic range, output noise and cooperativity. The main feature of the plot is the difference between the low- and high-cooperativity cases at each noise level; regardless of cooperativity the total information at infinite dynamic range would saturate at approximately the same value (which depends on the output noise magnitude). However, highly cooperative systems manage to reach a high fraction (80% or more) of their saturated information capacity even at reasonable dynamic ranges of 25- to 100-fold (meaning that the input concentration varies in the range $[\frac{1}{5}K_d, 5K_d]$ or $[\frac{1}{10}K_d, 10K_d]$, respectively), whereas low-cooperativity systems require a much bigger dynamic range for the same effect. The decrease in capacity with decreasing dynamic range is a direct consequence of the nonlinear relationship between the concentration and occupancy, Eq. (12), and for low-cooperativity systems means being unable to fully shut down or fully induce the promoter. In theory, Eq. (18) predicts that $\sigma_g(\bar{g}(c) \to 0)=0$, making the state in which the gene is off very informative about the input. If, however, the gene cannot be fully repressed either because there is always some residual input, $c_{\text{min}}$, or because there is leaky expression even when the input is exactly zero, then at any biologically reasonable input dynamic range some capacity will be lost.

Next, we briefly discuss how precisely tuned the resulting optimal distributions have to be to take full advantage of the regulatory element's capacity. For each point in the noise plane of Fig. 3(a) the optimal input distribution $p^*_{\text{TF}}(c)$ is perturbed many times to create an ensemble of suboptimal inputs $p^i_{\text{TF}}(c)$ (see Appendix C). For each $p^i_{\text{TF}}(c)$, we compute, first, its distance away from the optimal solution by means of Jensen-Shannon divergence, $d_i = D_{\text{JS}}(p^i_{\text{TF}}, p^*_{\text{TF}})$ [43]; next, we use the $p^i_{\text{TF}}(c)$ to compute the suboptimal information transmission $I_i$. The divergence $d_i$ is a measure of similarity between two distributions and ranges between 0 (distributions are the same) and 1 (distributions are very different); $1/d_i(p^i_{\text{TF}}, p^*_{\text{TF}})$ approximately corresponds to the number of samples one would have to draw to say with confidence that they were selected from either $p^i_{\text{TF}}(c)$ or $p^*_{\text{TF}}(c)$. A scatter plot of many such pairs $(d_i, I_i)$ obtained with various perturbations $p^i_{\text{TF}}(c)$ for each system of the noise plane characterizes the sensitivity of the optimal solution for that system; the main feature of such a plot, Fig. 9 below, is the linear (negative) slope that describes the fraction of channel capacity lost for a unit of Jensen-Shannon distance away from the optimal solution. Figure 6(b) displays these fractions as a function of the optimal capacity, and each system from the noise plane shown in Fig. 3 is represented by a black dot. We note that systems with higher capacities require more finely tuned solutions and suffer a larger fractional (and thus absolute) loss if the optimal input distribution is perturbed. Importantly, if the linear slopes are taken seriously and are used to extrapolate toward distributions that are very different from optimal, $D_{\text{JS}} \to 1$, we observe that for most of the noise plane the leftover capacity still remains about 1 bit, indicating that biological regulatory elements capable of transmitting an on-off decision perhaps are not difficult to construct. On the other hand, transmitting significantly more than one bit requires some degree of tuning that matches the distribution of inputs to the characteristics of the regulatory element.

### C. Costs of higher capacity

Real regulatory elements must balance the pressure to convey information reliably with the cost of maintaining the cell's internal state, represented by the expression levels of transcription factors. The fidelity of the representation is increased (and the fractional fluctuation in their number is decreased) by having more molecules "encode" a given state. On the other hand, making or degrading more transcription factors puts a metabolic burden on the cell, and frequent transitions between various regulatory states could involve large time lags as, for example, the regulation machinery attempts to keep up with a changed environmental condition, by accumulating or degrading the corresponding TF molecules. In addition, the output genes themselves that get switched on or off by transcription factors and therefore read out the internal state must not be too noisy, otherwise the advantage of maintaining precise transcription factor levels is lost. For recent work on costs of regulation, see Refs. [13,44,45].

Suppose that there is a cost to the cell for each molecule of output gene that it needs to produce, and that this incremental cost per molecule is independent of the number of molecules already present. Then, on the output side, the cost must be proportional to $\langle g\rangle=\int dg\, g p_{\mathrm{exp}}(g)$. We remember that in optimal distribution calculations $g$ is expressed as relative to the maximal expression, such that its mean is between zero and one. To get an absolute cost in terms of the number of molecules, this normalized $\bar{g}$ therefore needs to be multiplied by the inverse of the output noise strength, $\alpha^{-1}$, as the latter scales with $g_0$ (see Table I). The contribution of the output cost is thus proportional to $\alpha^{-1}\bar{g}$.

On the input side, the situation is similar: the cost must be proportional to $K_d\langle\tilde{c}\rangle = K_d\int d\tilde{c}\tilde{c}p_{\mathrm{TF}}(\tilde{c})$, where our optimal solutions are expressed, as usual, in dimensionless concentration units $\tilde{c}=c/K_d$. In either of the two input noise models (i.e., diffusion or switching input noise), with diffusion constant held fixed, $K_d\propto\beta^{-1}$ or $K_d\propto\gamma^{-1}$. See Appendix D for notes on the effects of nonspecific binding of transcription factors to the DNA.

Collecting all our thoughts on the costs of coding, we can write down the "cost functional" as the sum of input and output cost contributions:

$$\langle C[p_{\mathrm{TF}}(c)]\rangle=\frac{v_1}{\beta}\int dc\, p_{\mathrm{TF}}(c)c+\frac{v_2}{\alpha}\int dc\, p_{\mathrm{TF}}(c)\int dg\, p(g|c)g, \tag{21}$$

where $v_1$ and $v_2$ are proportional to the unknown costs per molecule of input or output, respectively, and $\alpha$ and $\beta$ are noise parameters of Table I. This ansatz captures the intuition that, while decreasing noise strengths will increase information transmission, it will also increase the cost. Instead of maximizing the information without regard to the cost, the new problem to extremize is

$$\mathcal{L}[p_{\mathrm{TF}}(c)]=I[p_{\mathrm{TF}}(c)]-\Phi\langle C[p_{\mathrm{TF}}(c)]\rangle-\Lambda\int dc\, p_{\mathrm{TF}}(c), \tag{22}$$

and the Lagrange multiplier $\Phi$ has to be chosen so that the cost of the resulting optimal solution $\langle C[p^*_{\mathrm{TF}}(c)]\rangle$ equals some predefined cost $C_0$ that the cell is prepared to pay.

We now wish to recreate the noise plane of Fig. 3, while constraining the total cost of each solution to $C_0$. To be concrete and pick the value for the cost and proportionality constants in Eq. (21), we use the estimates from *Drosophila* noise measurements and analysis in Refs. [22,35], which assign to the system denoted by a blue dot in Fig. 3(a) the values of ~800 bicoid molecules of input at $K_d$, and a maximal induction of $g_0$ ~ 4000 hunchback molecules if the burst size $b$ is 10. Figure 7(a) is the noise plane for an activator with no cooperativity, as in Fig. 3, but with the cost limited to an average total of $C_0$ ~ 7000 molecules of input and output per nucleus. There is now one optimal solution denoted by a green dot (with a dominant input noise contribution); if one tries to choose a system with lower input or output noise, the cost constraint forces the input distribution $p_{\mathrm{TF}}(c)$ and the output distribution $p_{\exp}(g)$ to have very low probabilities at high induction, consequently limiting the capacity.

Clearly, a different system will be optimal if another total allowed cost $C_0$ is selected. The dark green line on the noise plane in Fig. 7(a) corresponds to the flow of the optimal solution for an activator with no cooperativity if the allowed cost is increased, and the corresponding cost-capacity curve is shown in Fig. 7(b). The light green line is the trajectory of the optimal solution in the noise plane of the activator system with cooperativity $h$=3, and the dark and light red trajectories are shown for the repressor with $h$=1 and 3, respectively. We note first that the behavior of the cost function is quite different for the activator (where low input implies low output and therefore low cost; and conversely high input means high output and also high cost) and the repressor (where input and output are mutually exclusively high or low and the cost is intermediate in both cases). Second, in Fig. 7(b) we observe that the optimal capacity as a function of cost is similar for the activators and repressors, in contrast to the comparison of Fig. 4, where repressors provided higher capacities. Third, we note in the same figure that increasing the cooperativity at fixed noise strength $\beta$ brings a substantial increase, of almost 1 bit over the whole cost range, in the channel capacity, in agreement with our previous observations about the interaction between capacity and the dynamic range. The last and perhaps the most significant conclusion is that, even with input distributions matched to maximize the transmission at a fixed cost, the capacity still only scales roughly linearly with the logarithm of the number of available signaling molecules, and this fact must ultimately be limiting in a single regulatory element.

Finally, we comment on the balance between input and output noise. The noise, which limits information transmission, ultimately depends on the number of molecules used for signaling, and our formulation of the cost constraints allows, in effect, for some fixed total number of molecules to be distributed between input and output. When this assignment of molecules (and hence metabolic cost) to the input and output is chosen to maximize information transmission, we see for the example in Fig. 7(a) that the dominant noise source is at the input. This is interesting not least because, until recently, most analyses of noise in gene expression have focused on output noise sources, the randomness in transcription and translation. Our results suggest that emerging evidence for the importance of input noise sources [22,34–36,45] may reflect selection for biological mechanisms that maximize information transmission at fixed cost.

## V. DISCUSSION

We have tried to analyze a simple regulatory element as an information processing device. One of our major results is that one cannot discuss an element in isolation from the statistics of the input that it is exposed to. Yet in cells the inputs are often transcription factor concentrations that encode the state of various genetic switches, from those responsible for cellular identity to those that control the rates of metabolism and cell division, and the cell exerts control over these concentrations. While it *could* use different distributions to represent various regulatory settings, we argue that the cell *should* use the one distribution that allows it to make the most of its genetic circuitry—the distribution that maximizes the dependency, or mutual information, between inputs and outputs. Mutual information can then be seen both as a measure of how well the cell is doing by using *its* encoding scheme, and the best it could have done using the *optimal* scheme, which we can compute; comparison between the optimal and measured distributions gives us a sense of how close the organism is to the achievable bound [23].

Moreover, mutual information has absolute units, i.e., bits, that have a clear interpretation in terms of the number of discrete distinguishable states that the regulatory element can resolve. This last fact helps clarify the ongoing debates about what is the proper noise measure for genetic circuits, and in what context a certain noise is either "big" or "small" (as it is really a function of the inputs). Information does not replace the standard noise-over-the-mean measure —noise calculations or measurements are still necessary to compute the element's capacity— but does give it a functional interpretation.

We emphasize that formulating a correct model of signals and noise is a separate problem from computing the information capacity and the optimal input distribution once the transfer function and noise are known. Even though microscopic noise models can, in combination with experiment, be used to *infer* (and are thus a probe for) the underlying molecular processes that give rise to the observed noise [35], for the capacity calculation itself this microscopic level of detail is not required. When making a connection between theory and experiment for a concrete genetic regulatory element, one only needs a good phenomenological model, i.e., a good fit for the measurements of $\bar{g}(c)$ and $\sigma_g(c)$. This is perhaps most clearly demonstrated in Ref. [23], where the information theoretic framework developed here is applied to the high-precision measurements of the noise in the bicoid-hunchback system of the fruit fly. For a theoretical discussion like the one presented here, it is, however, convenient to parametrize the noise with a small number of tunable "knobs" that regulate the strengths of various kinds of noise sources, and observe the scaling of channel capacity as the noise parameters are varied.

In this paper we have considered a class of simple parametrizations of signals and noise that can be used to fit measurements and capture all of the measured noise in several model systems, such as the bicoid-hunchback system of the fruit fly, a number of yeast genes, and the *lac* repressor in *Escherichia coli* (see Refs. [23,41] for the last). We find that the capacities of these realistic elements are generally larger than 1 bit, and can be as high as 2 bits. By simple inspection of optimal output distributions in Fig. 3(b) or Fig. 3(c) it is difficult to say anything about the capacity: the distribution might look bimodal yet carry more than one bit, or might even be a monotonic function without any obvious structure, indicating that the information is encoded in the graded response of the element. When the noise is sufficiently high, on the other hand, the optimal strategy is that of achieving one bit of capacity and only utilizing maximum and minimum available levels of transcription factors for signaling. The set of distributions that achieve capacities close to the optimal one is large, suggesting that perhaps 1-bit switches are not difficult to implement biologically, while in contrast we find that transmission of much more than one bit requires some tuning of the system.

Finally, we discussed how additional biophysical constraints can modify the channel capacity. By assuming a linear cost model for signaling molecules and a limited input dynamic range, the capacity and cost couple in an interesting way and the maximization principle allows new questions to be asked. For example, increasing the cooperativity reduces the cost, as we have shown; on the other hand, it increases the sensitivity to fluctuations in the input, because the input noise strength $\beta$ is proportional to the square of Hill's coefficient, $h^2$. In a given system we could therefore predict the optimal effective cooperativity, if we knew the real cost per molecule. Further work is needed to tease out the consequences of cost (if any) from experimental data.

The principle of information maximization clearly is not the only possible lens through which regulatory networks are to be viewed. One can think of examples where the system either does not reach a steady state or there are constraints on the *dynamics*, something that our analysis has ignored by only looking at steady state behavior; for instance, consider genetic regulatory circuits that execute limit cycle oscillations, to which our analysis is not applicable (but for which we also do not have any experimental data on noise); or the chemotactic network of

*Escherichia coli* that has to perfectly adapt in order for the bacterium to be able to climb the attractant gradients. Alternatively, suppose that a system needs to convey only a single bit, but it has to be done reliably in a fluctuating environment, perhaps by being robust to the changes in outside temperature. In this case it seems that both concepts, that of maximal information transmission and the robustness to fluctuations in certain auxiliary variables which also influence the noise, could be included into the same framework, but the issue needs further work. More generally, however, these and similar examples assume that one has identified in advance the biologically relevant features of the system, e.g., perfect adaptation or robustness, and that there exists a problem-specific error measure which the regulatory network is trying to minimize. Such a measure could then either replace or complement the assumption-free information theoretic approach presented here.

We emphasize that the kind of analysis carried out here is not restricted to a single regulatory element. As was pointed out in the introduction, the inputs $\mathcal{I}$ and the outputs $\mathcal{O}$ of the regulatory module can be multidimensional, and the module could implement complex internal logic with multiple feedback loops. It seems that especially in such cases, when our intuition about the noise—now a function of multiple variables—starts breaking down, the information formalism could prove to be helpful. Although the solution space that needs to be searched in the optimization problem grows exponentially with the inputs, there are biologically relevant situations that nevertheless appear tractable: for example, when there are multiple readouts of the same input, or combinatorial regulation of a single output by a pair of inputs; in addition, knowing that the capacities of a single input-output chain are on the order of a few bits also means that only a small number of distinct input levels for each input need to be considered. Some cases of interest therefore appear immediately amenable to biophysical modeling approaches and the computation of channel capacities, as presented in this paper.

We have focused here on the theoretical exploration of information capacity in simple models. It is natural to ask how our results relate to experiment. Perhaps the strongest connection would be if biological systems really were selected by evolution to optimize information flow in the sense we have discussed. If this optimization principle applies to real regulatory elements, then, for example, given measurements on the input-output relation and noise in the system we can make parameter free predictions for the distribution of expression levels that cells will use. Initial efforts in this direction, using the bicoid-hunchback element in the *Drosophila* embryo as an example, are described in Ref. [23]. It is worth noting that a parallel discussion of optimization principles for information transmission has a long history in the context of neural coding, where we can think of the distribution on inputs as given by the sensory environment and optimization is used to predict the form of the input-output relation [18–21]. Although there are many open questions, it would be attractive if a single principle could unify our understanding of information flow across such a wide range of biological systems.

## Acknowledgments

We thank T. Gregor, J. B. Kinney, D. W. Tank, and E. F. Wieschaus for helpful discussions. This work was supported in part by NIH Grants No. P50 GM071508 and No. R01 GM077599, NSF Grant No. PHY-0650617, the Swartz Foundation, the Burroughs Wellcome Fund Program in Biological Dynamics (G.T.), and U.S. Department of Energy Grant No. DE-FG02-91ER40671 (C.C.).

## References

1. Leloup JC, Goldbeter A. Proc Natl Acad Sci USA 2003;100:7051. [PubMed: 12775757]
2. Lawrence, PA. The Making of a Fly: The Genetics of Animal Design. Blackwell; Oxford: 1992.
3. Levine M, Davidson EH. Proc Natl Acad Sci USA 2005;102:4936. [PubMed: 15788537]
4. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO. Mol Biol Cell 2000;11:4241. [PubMed: 11102521]

5. McAdams HH, Arkin A. Proc Natl Acad Sci USA 1997;94:814. [PubMed: 9023339]

6. Kepler TB, Elston TC. Biophys J 2001;81:3116. [PubMed: 11720979]

7. Wolpert L. J Theor Biol 1969;25:1. [PubMed: 4390734]

8. Shannon CE. Bell Syst Tech J 1948;27:379. **27**, 623, (1938); reprinted in C. E. Shannon and W Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, 1949).

9. Shannon CE. Proc IRE 1949;37:10.

10. Cover, TM.; Thomas, JA. Elements of Information Theory. John Wiley; New York: 1991.

11. Ziv E, Nemenman I, Wiggins C. PLoS ONE 2007;2:e1077. [PubMed: 17957259]

12. Kussell E, Leibler S. Science 2005;309:2075. [PubMed: 16123265]

13. Taylor S, Tishby N, Bialek W. e-print arXiv:0712.4382.

14. Li F, Long T, Lu Y, Ouyang Q, Tang C. Proc Natl Acad Sci USA 2004;101:4781. [PubMed: 15037758]

15. Sanchez L, Thieffry D. J Theor Biol 2001;211:115. [PubMed: 11419955]

16. Tyson JJ, Chen K, Novak B. Nat Rev Mol Cell Biol 2001;2:908. [PubMed: 11733770]

17. Goldbeter A. Nature (London) 2002;420:238. [PubMed: 12432409]

18. Barlow, HB. Sensory Communication. Rosenblith, W., editor. MIT Press; Cambridge, MA: 1961. p. 217-234.

19. Laughlin, SB.; Naturforsch, Z. C. Vol. 36c. 1981. p. 910

20. Atick JJ, Redlich AN. Neural Comput 1990;2:308.

21. Brenner N, Bialek W, de Ruyter van Steveninck R. Neuron 2000;26:695. [PubMed: 10896164]

22. Gregor T, Tank DW, Wieschaus EF, Bialek W. Cell 2007;130:153. [PubMed: 17632062]

23. Tkačik G, Callan CG Jr, Bialek W. Proc Natl Acad Sci USA. to be published.

24. Watson, JD.; Hopkins, NH.; Roberts, JW.; Steitz, JA.; Weiner, AM. Molecular Biology of the Gene. 4. Benjamin/Cummings Publishing; Menlo Park, CA: 1988. p. 465-502.

25. Setty Y, Mayo AE, Surette MG, Alon U. Proc Natl Acad Sci USA 2003;100:7702. [PubMed: 12805558]

26. Kuhlman T, Zhang Z, Saier MH Jr, Hwa T. Proc Natl Acad Sci USA 2007;104:6043. [PubMed: 17376875]

27. Thattai M, van Oudenaarden A. Proc Natl Acad Sci USA 2001;98:8614. [PubMed: 11438714]

28. Elowitz MB, Levine AJ, Siggia ED, Swain PD. Science 2002;297:1183. [PubMed: 12183631]

29. Ozbudak E, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A. Nat Genet 2002;31:69. [PubMed: 11967532]

30. Blake WJ, Kaern M, Cantor CR, Collins JJ. Nature (London) 2003;422:633. [PubMed: 12687005]

31. Swain PS, Elowitz MB, Siggia ED. Proc Natl Acad Sci USA 2002;99:12795. [PubMed: 12237400]

32. Silverman R. IRE Trans Inf Theory 1955;1:19.

33. Blahut RE. IEEE Trans Inf Theory 1972;18:460.

34. Bialek W, Setayeshgar S. Proc Natl Acad Sci USA 2005;102:10040. [PubMed: 16006514]

35. Tkačik G, Gregor T, Bialek W. PLoS ONE. to be published.

36. van Zon JS, Morelli MJ, Tanase–Nicola S, ten Wolde PR. Biophys J 2006;91:4350. [PubMed: 17012327]

37. Metzler R. Phys Rev Lett 2001;87:068103. [PubMed: 11497866]

38. Tkačik G, Bialek W. e-print arXiv:0712.1852.

39. Pedraza JM, van Oudenaarden A. Science 2005;307:1965. [PubMed: 15790857]

40. Bar–Even A, Paulsson J, Maheshri N, Carmi M, O'Shea E, Pilpel Y, Barkai N. Nat Genet 2006;38:636. [PubMed: 16715097]

41. Tkačik, G. PhD dissertation. Princeton University; 2007.

42. van Kampen, NG. Stochastic Processes in Physics and Chemistry. 3. North-Holland; Amsterdam: 2007.

43. Lin J. IEEE Trans Inf Theory 1991;37:145.

44. Dekel, E.; Alon, U. Nature (London). Vol. 436. 2005. p. 588

45. Tanase-Nicola S, ten Wolde PR. PLoS Comput Biol. to be published.

46. Bialek W, Setayeshgar S. Phys Rev Lett 2008;100:258101. [PubMed: 18643705]

## APPENDIX A: FINDING OPTIMAL CHANNEL CAPACITIES

If we treat the kernel on a discrete ($c$,$g$) grid we can easily choose such $p_{\text{TF}}(c)$ as to maximize the mutual information $I(c;g)$ between the expression level and the concentration. The problem can be stated in terms of the following variational principle:

$$\mathcal{L}[\,p_{\text{TF}}(c)]=\sum_{c,g}p(g|c)p_{\text{TF}}(c)\ln\frac{p(g|c)}{p_{\exp}(g)}-\Lambda\sum_{c}p_{\text{TF}}(c), \tag{A1}$$

where the multiplier Λ enforces the normalization of $p_{\text{TF}}(c)$, and $p_{\exp}(g)$ itself is a function of the unknown distribution [since $p_{\exp}(g) = \Sigma_c p(g|c)p_{\text{TF}}(c)$]. The solution $p^{*}_{\text{TF}}(c)$ of this problem achieves the capacity $I(c;g)$ of the channel. For brevity, in the rest of this appendix, we explicitly suppress subscripts on probability distributions, and distinguish them by using the argument name [i.e., $p(c)$ is $p_{\text{TF}}(c)$ and $p(g)$ is $p_{\exp}(g)$].

The original idea behind the Blahut-Arimoto approach [33] was to understand that the maximization of Eq. (A1) using variational objects $p(c_i)$ is equivalent to the following maximization:

$$\max_{p(c)}\mathcal{L}[\,p(c)]-\max_{p(c)}\max_{p(c|g)}\mathcal{L}'[\,p(c),\,p(c|g)], \tag{A2}$$

where

$$\mathcal{L}'[\,p(c),\,p(c|g)]=\sum_{g,c}p(c)p(g|c)\ln\frac{p(c|g)}{p(c)}-\Lambda\sum_{c}p(c). \tag{A3}$$

In words, finding the extremum in the variational object $p(c)$ is equivalent to a double maximization of a modified Lagrangian, where both $p(c)$ and $p(g|c)$ are treated as independent variational objects. The extremum of the modified Lagrangian is achieved exactly when the consistency condition $p(c|g)=\frac{p(g|c)p(c)}{\sum_c p(g|c)p(c)}$ holds. This allows us to make an iterative algorithm that we detail below, where Eq. (A3) is solved for the optimal $p(c)$ and evaluated at some “known” $p(c|g)$, which is in turn updated with the newly obtained estimate of $p(c)$.

Before describing the algorithm let us also suppose that each input signal $c$ carries some metabolic or time cost to the cell. Then we can introduce a cost vector $v(c)$ that assigns a cost to each codeword $c$, and require of the solution the following:

$$\sum_{c}p(c)v(c)\le C_0, \tag{A4}$$

where $C_0$ is the maximum allowed expense. The constraint can be introduced into the functional, Eq. (A1) or Eq. (A3), through an appropriate Lagrange multiplier; the same approach can be taken to introduce the cost of coding for the output words, $\Sigma_g\Sigma_c p(g|c)p(c)v(g)$, because it reduces to an additional “effective” cost for the input, $\tilde{v}(c) = \Sigma_g p(g|c)v(g)$.

As was pointed out in the main text, after discretization we have no guarantees that the optimal distribution $p(c_i)$ is going to be smooth. One way to address this problem is to enforce the smoothness on the scale set by the precision at which the input concentration can be controlled by the cell, $\sigma_c(\bar{c})$, by penalizing big derivatives in the Lagrangian of Eq. (A3). An alternative way is to find the spiky solution (with-out imposing any direct penalty term), but interpret it not as a real, "physical," concentration distribution, but rather as the distribution of concentrations that the cell attempts to generate, $c^*$. In this case, however, the limited resolution of the input $\sigma_c(\bar{c})$ must be referred to the output as an additional effective noise in gene expression, $\sigma_g^2 = \sigma_c^2 (\frac{\partial \bar{g}}{\partial c})^2$. The optimal solution $p(c^*)$ is therefore the distribution of the levels that the cell would use if it had infinitely precise control over choosing various $c^*$ (i.e., if the input noise were absent), but the physical concentrations are obtained by convolving this optimal result $p(c^*)$ with a Gaussian of width $\sigma_c(c^*)$. Although we chose to use the second approach to compute the results of this paper, we will, for completeness, describe next how to include the smoothness constraint into the functional explicitly.

If the smoothness of the input distribution $p(c)$ is explicitly constrained in the optimization problem, then it will be controllable through an additional Lagrange multiplier, and both ways of computing the capacity—that of referring the limited input resolution $\sigma_c(\bar{c})$ to the noise in the output, and that of including it as a smoothness constraint on the input distribution—will be possible within a single framework. We proceed by analogy to field theories in which the kinetic energy terms of the form $\int|\nabla f(x)|^2 dx$ constrain the gradient magnitude, and form the following functional:

$$\mathcal{L}[\,p(c)] = \mathrm{I}(c;g) - \lambda_0 \sum_c p(c) \tag{A5}$$

$$-\Phi_1 \sum_c p(c) v_1(c) - \Phi_2 \sum_g p(g) v_2(g) \tag{A6}$$

$$-\Theta \sum_c \left( \frac{\Delta p}{\Delta c} \sigma(c) \right)^2 . \tag{A7}$$

Equation (A5) maximizes the capacity with respect to variational objects $p(c)$ while keeping the distribution normalized; Eq. (A6) imposes cost $v_1(c)$ on input symbols and cost $v_2(g)$ on output symbols; finally, Eq. (A7) limits the derivative of the resulting solution. The difference operator Δ is defined for an arbitrary function $f(c)$:

$$\Delta f(c) = f(c_{i+1}) - f(c_i). \tag{A8}$$

$\sigma(c)$ assigns a different weight to various intervals on the input axis $c$. If the input cannot be precisely controlled, but has an uncertainty of $\sigma(c)$ at mean input level $c$, we require that the optimal probability distribution must not change much as the input fluctuates on the scale $\sigma(c)$. In other words, we require for each input concentration that

$$\delta p=\frac{\Delta p}{\Delta c}\sigma(c) \ll 1; \tag{A9}$$

the term in Eq. (A7) constrained by Lagrange multiplier Θ can be seen as the sum of squares of such variations over all possible values of the input.

By differentiating the functional Eq. (A3) that includes the relevant constraints, with respect to $p(c_i)$ we get the following equation:

$$0=\sum_g p(g|c_i)\ln p(c_i|g) - \ln p(c_i) - \lambda - \Phi_1 v_1(c_i) - \Phi_2 \sum_g p(g|c_i) v_2(g) \tag{A10}$$

$$+\Theta\left( [\, p(c_{i+1}) - p(c_i)] \frac{\sigma^2(c_i)}{(c_{i+1}-c_i)^2} - [\, p(c_i) - p(c_{i-1})] \frac{\sigma^2(c_{i-1})}{(c_i - c_{i-1})^2} \right). \tag{A11}$$

Let us denote by $F(c, p(c))=\Delta \frac{\Delta p}{(\Delta c)^2}\sigma^2$ the term in large parentheses. The solution for $p(c)$ is therefore given by

$$p(c)=\frac{1}{Z}\exp\left( \sum_g p(g|c)\ln p(c|g) - \Phi_1 v_1(c) - \Phi_2 \sum_g p(g|c) v_2(g) + \Theta\, F(c, p(c)) \right). \tag{A12}$$

We can now continue to use the Blahut-Arimoto trick of pretending that $p(c|g)$ is an independent variational object, and that $p(c)$ has to be solved with $p(c|g)$ held fixed; however, even in that case, Eq. (A12) is an implicit equation for $p(c)$ which needs to be solved by numerical means. The complete iterative prescription is therefore as follows:

$$p^n(g)=\sum_c p(g|c)p^n(c), \tag{A13}$$

$$p^n(c|g)=\frac{p(g|c)p^n(c)}{p^n(g)}, \tag{A14}$$

$$p^{n+1}(c)=\frac{1}{Z}\exp\left( \sum_g p(g|c)\ln p^n(c|g) - \Phi_1 v_1(c) - \Phi_2 \sum_g p(g|c) v_2(g) + \Theta F(c, p^{n+1}(c)) \right). \tag{A15}$$

Again, Eq. (A15) has to be solved on its own by numerical means as the variational objects for iteration ($n$+1) appear both on the left- and right-hand sides. The input and output costs of coding are neglected if one sets $\Phi_1 = \Phi_2 = 0$; likewise, the smoothness constraint is ignored for Θ=0, in which case Eq. (A15) is the same as in the original Blahut-Arimoto derivation and it gives the value of $p^{n+1}(c)$ explicitly.

For the capacities computed in this paper we have calculated the effective output noise that includes the intrinsic output noise as well as the input noise that has been referred to the output (see Sec. III); we can therefore set Θ=0. This approach treats all sources of noise on the same footing and allows us to directly compare the magnitudes of noise sources at the input and the output. We also note that it makes sense to compute and compare the optimal distribution of outputs rather than inputs: the input-output kernels are degenerate and there are various input distributions (differing either in the regions that give saturated or zero response, or by having variations on a scale below $\sigma_c$) that will yield essentially the same distribution of outputs.

## APPENDIX B: VALIDITY OF LANGEVIN APPROXIMATIONS

The Langevin approximation assumes that the fluctuations of a quantity around its mean are Gaussian and proceeds to calculate their variance [42]. For the calculation of exact channel capacity we must calculate the full input-output relation, $p(g|c)$. Even if the Langevin approach ends up giving the correct variance as a function of the input, $\sigma_g(c)$, the shape of the distribution might be far from Gaussian. We expect such a failure when the number of mRNA molecules is very small: the distribution of expression levels might be then multipeaked, with peaks corresponding to $b$,$2b$,$3b$,… proteins, where $b$ is the burst size (number of proteins produced per transcript lifetime).

In the model used in Eq. (18), the parameter $\alpha=(1+b)/g_0$ determines the output noise; $g_0 = b\bar{m}$, where $\bar{m}$ is the average number of transcripts produced during the integration time (i.e., the longest averaging time scale in the problem, for example the protein lifetime or cell doubling time). If $b\gg 1$, then the output noise is effectively determined only by the number of transcripts, $\alpha \approx 1/\bar{m}$. We should therefore be particularly concerned what happens as $\bar{m}$ gets small.

Our plan is therefore to solve for $p(g|c)$ exactly by finding the stationary solution of the master equation in the case where the noise consists of the output and switching input contributions. In this approach, we explicitly treat the fact that the number of transcribed messages, designated by $m$, is discrete. We start by calculating $p_i(m|c,t)$. The state of the promoter is described by index $i$, which can be 0 or 1, depending on whether the promoter is bound by the transcription factor or not, respectively. Normalization requires that for each value of $c$

$$\sum_{i=0,1}\sum_{m} p_i(m|c,t)=1. \tag{B1}$$

The time evolution of the system is described by the following set of equations for an activator:

$$\frac{\partial p_0(m|c,t)}{\partial t}=R_e[\,p_0(m-1|c,t)-p_0(m|c,t)]-\frac{1}{\tau}[\,mp_0(m|c,t)-(m+1)p_0(m+1|c,t)]-k_-p_0(m|c,t)+k_+cp_1(m|c,t), \tag{B2}$$

$$\frac{\partial p_1(m|c,t)}{\partial t}=-\frac{1}{\tau}[\,mp_1(m|c,t)-(m+1)p_1(m+1|c,t)]+k_-p_0(m|c,t)-k_+cp_1(m|c,t), \tag{B3}$$

where $\tau$ is the integration time, $k_-$ is the rate for switching into the inactive state (off rate of the activator), $k_+$ is the second-order on rate, and $R_e$ is the rate of mRNA synthesis. These constants combine to give $\bar{m} = R_e\tau$ and the input switching noise strength $\gamma=(k_-\tau)^{-1}$; see Table I. This set of equations is supplemented by appropriate boundary conditions for $m$=0. To find a steady state distribution $p(m|c,t \to \infty) = p(m|c)$, we set the left-hand side to zero and rewrite the set of equations (with high enough cutoff value of $m_{\mathrm{max}}$) in matrix form:

$$\mathbf{M}(c)\mathbf{p}(c)=\mathbf{b}, \tag{B4}$$

where $\mathbf{p}=(p_0(0|c), p_1(0|c), p_0(1|c), p_1(1|c),\ldots)$ and $\mathbf{b}=(0,0,\ldots,0,1)$. The matrix $\mathbf{M}$ [of dimension $2(m_{\mathrm{max}}+1)+1$ rows and $2(m_{\mathrm{max}}+1)$ columns] contains, in its last row, only 1's, which enforces normalization. The resulting system is a nonsingular band-diagonal system that can be easily inverted. The input-output relation for the number of messages is given by taking $p(m|c) = p_0(m|c) + p_1(m|c)$.

Having found the distribution for the number of transcripts, we then convolve it with another Poisson process $p(g|\langle g\rangle = b\, m)$, i.e., $p(g|c) = \Sigma_m p(m|c) p(g|\langle g\rangle = b\, m)$. Finally, the result is rediscretized such that the mean expression $\bar{g}$ runs from 0 to 1.

Note that the Langevin approximation depends only on the combination of the burst size $b$ and the mean number of transcripts $\bar{m}$ through $\alpha$; in contrast, the master equation solution depends on both $b$ and $\bar{m}$ independently. The generalization of this calculation to repressors or Hill-coefficient-type cooperativity is straightforward.

Figure 8(c) shows that the Langevin approximation yields correct second moments of the output distribution; however, Gaussian distributions themselves are, for large burst sizes and small numbers of messages, inconsistent with the exact solutions, as can be seen in Fig. 8(a). In the opposite limit, where the number of messages is increased and burst size kept small [see Fig. 8(b)], normal distributions are an excellent approximation. Despite these difficulties, the information capacity calculated with either Gaussian or master input-output relations differs by at most 12% over a large range of burst sizes $b$ and values for $\alpha$, illustrated by Fig. 8(d).

## APPENDIX C: FINE TUNING OF OPTIMAL DISTRIBUTIONS

To examine the sensitivity to the perturbations in the optimal input distributions for Fig. 6 we need to generate an ensemble of perturbations. We pick an *ad hoc* prescription, whereby the optimal solution is taken, and we add to it the five lowest harmonic modes on the input domain, each with a weight that is uniformly distributed on some range. The range determines whether the perturbation is small or not. The resulting distribution is clipped to be positive and renormalized. This choice was made to induce low-frequency perturbations (high-frequency perturbations get averaged out because the kernel is smooth). Then, for an ensemble of 100 such perturbations $p_i(c)$, $i=1, \ldots, 100$, and for every system of the noise plane in Fig. 3(a), the divergence of the perturbed input distribution from the true solution $d_i = D_{\mathrm{JS}}(p_i(c), p^*_{\mathrm{TF}}(c))$ is computed, as well as the information transmission $I_i = I[p(g|c), p_i(c)]$. Figure 9 plots the $(d_i, I_i)$ scatter plots for 3×3 representative systems with varying amounts of output ($\alpha$) and input ($\beta$) noise, taken from Fig. 3(a) uniformly along the horizontal and vertical axes.

Figure 9 shows that, as we move toward systems with higher capacity (lower left corner), perturbations to the optimal solution that are at the same distance from the optimum as in the low-capacity systems (upper right corner), will cause greater relative loss (and therefore an even greater absolute loss) in capacity. As expected, higher-capacity systems must be better tuned, but even for the highest-capacity system considered, a perturbation of around $d_{\mathrm{JS}} \approx 0.2$ will cause only an average 15% loss in capacity. We also note that for systems with high capacity the linear relationship between the the divergence $d_i$ and capacity $I_i$ provides a better fit than for systems with small capacity.

## APPENDIX D: NONSPECIFIC BINDING

One needs to make a careful distinction between the total concentration of the input transcription factors, $c_t$, and the free concentration $c_f$, diffusing in solution in the nucleus. We imagine the true binding site embedded in a pool of nonspecific binding sites—perhaps all other short fragments of DNA—and there being an ongoing competition between one functional site (with strong affinity) and large number of weaker nonspecific sites. If these nonspecific sites are present at concentration $\rho$ in the cell, and have affinities drawn from some distribution $p(K)$, the relationship between the free and the total concentration of the input is

$$c_t = c_f + \rho \int dK\, p(K) \frac{c_f}{c_f + K}. \tag{D1}$$

Importantly, the concentration that enters all information capacity calculations is the *free* concentration $c_f$, because it directly determines both the promoter occupancy in Eq. (12) as well as the diffusive noise; on the other hand, the cell can influence the free concentration only by producing more or less of the transcription factor, i.e., by varying (and paying for) the *total* concentration. If the free concentration is well below the strength of the nonspecific binding $\langle K \rangle$, Eq. (D1) can be approximated by $c_t \approx c_f(1+\rho/\langle K \rangle)$, with the total and free concentrations being proportional to each other. Because the cost functional Eq. (21) is only determined to within a factor anyway, the presence of nonspecific sites will effectively just rescale the cost per free molecule of transcription factor. A separate calculation is needed to show that the presence of nonspecific binding does not appreciably increase the noise in gene expression (to be presented elsewhere).

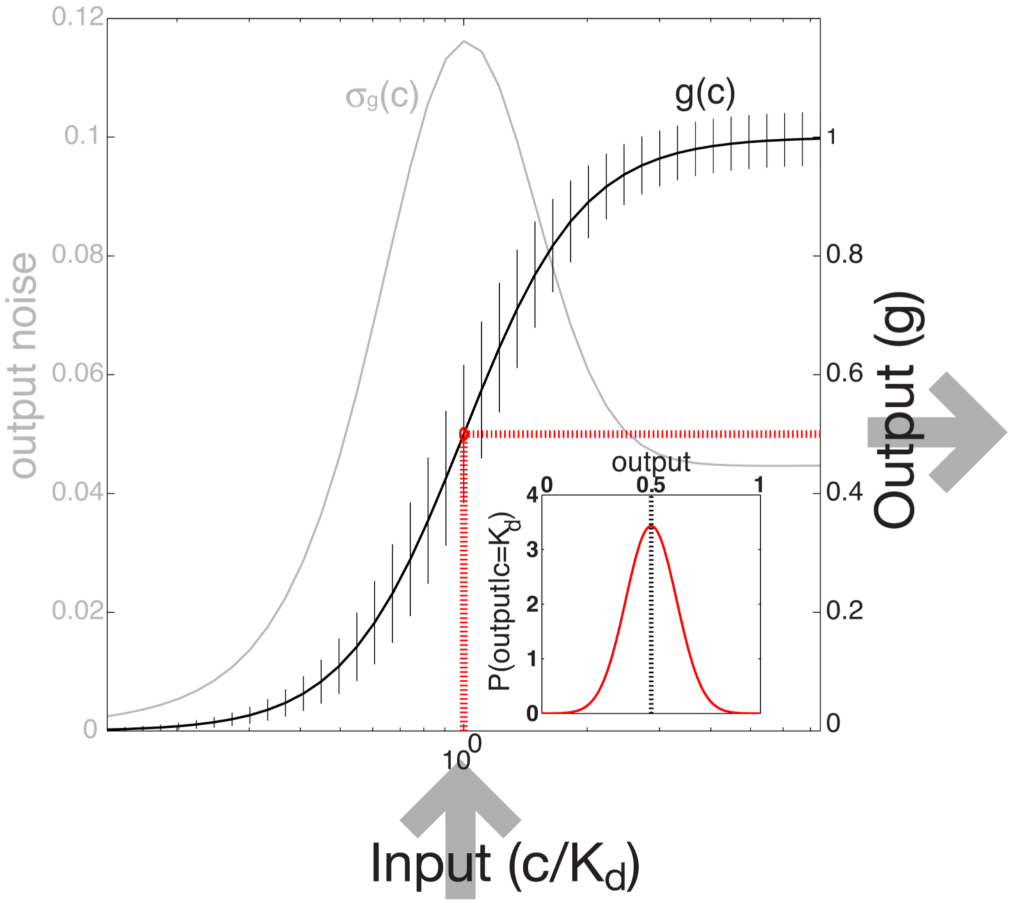
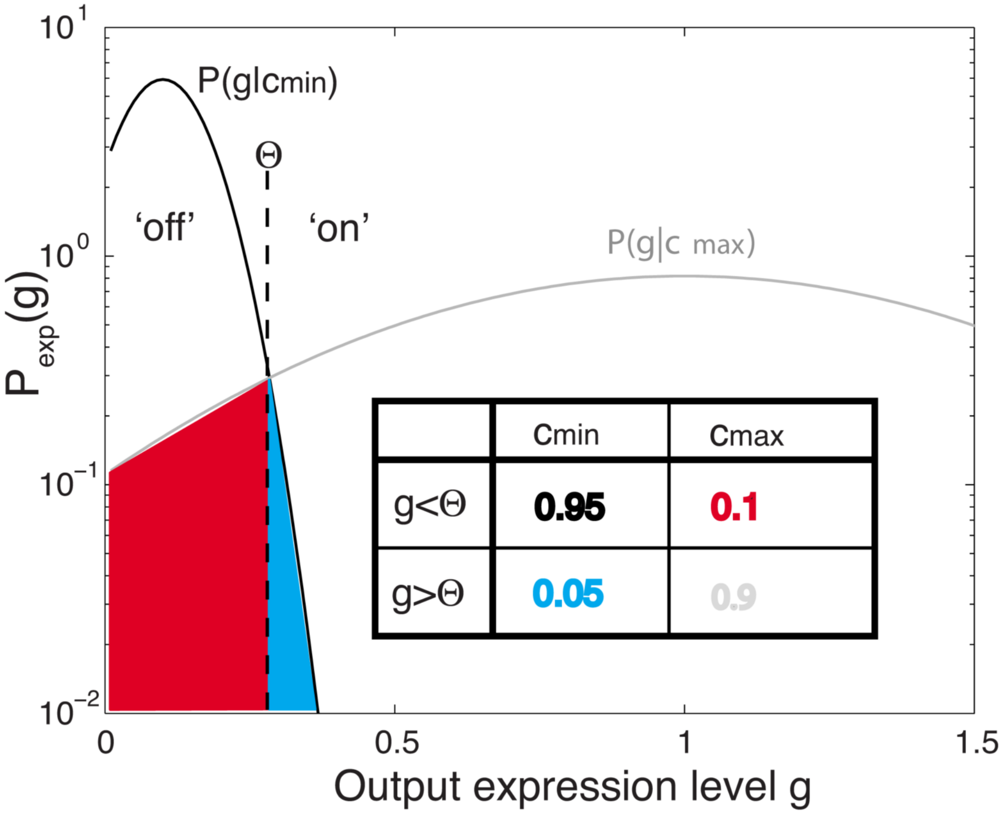


**FIG. 1.**

(Color online) A schematic diagram of a simple regulatory element. Each input is mapped to a mean output according to the input-output relation (thick sigmoidal black line). Because the system is noisy, the output fluctuates about the mean. This noise is plotted in gray as a function of the input and shown in addition as error bars on the mean input-output relation. Inset shows the probability distribution of outputs at half saturation, $p(g|c = K_d)$ (red dotted lines); in this simple example we assume that the distribution is Gaussian and therefore fully characterized by its mean and variance.



|  | Cmin | Cmax |
|---|---|---|
| g<Θ | 0.95 | 0.1 |
| g>Θ | 0.05 | 0.9 |

**FIG. 2.**

(Color online) An illustration of the large-noise approximation. We consider distributions of the output at minimal ($c_{\text{min}}$) and full ($c_{\text{max}}$) induction as trying to convey a single binary decision, and construct the corresponding encoding table (inset) by discretizing the output using the threshold Θ. The capacity of such an asymmetric binary channel is degraded from the theoretical maximum of 1 bit, because the distributions overlap (blue and red). For unclipped Gaussians the optimal threshold Θ is at the intersection of two alternative probability distributions, but in general one searches for the optimal Θ that maximizes information in Eq. (11).

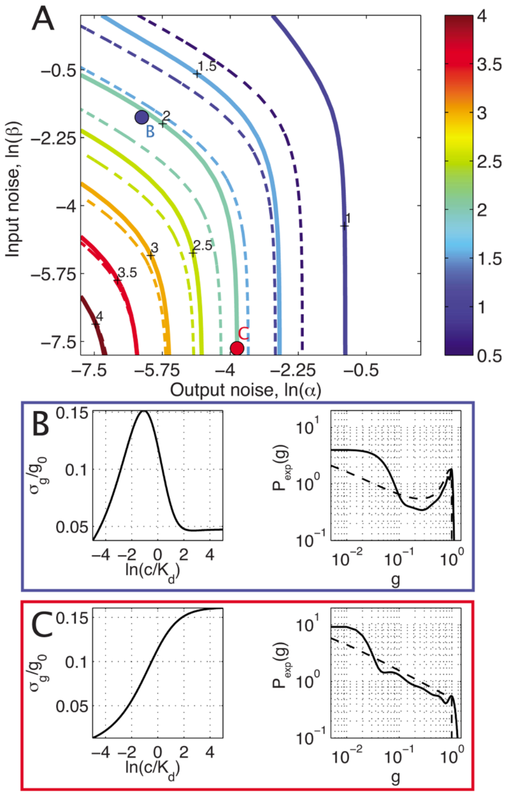


**FIG. 3.**

(Color online) Information capacity (color code, in bits) as a function of input and output noise using the activator input-output relation with Gaussian noise given by Eq. (20) and no cooperativity ($h = 1$). (a) shows the exact capacity calculation (thick line) and the small-noise approximation (dashed line). (b) displays the details of the blue point in (a): the noise in the output is shown as a function of the input, with a peak that is characteristic of a dominant input noise contribution; also shown is the exact solution (thick black line) and the small-noise approximation (dashed black line) to the optimal distribution of output expression levels. (c) similarly displays details of the system denoted by a red dot in (a); here the output noise is dominant and both approximate and exact solutions for the optimal distribution of outputs show a trend monotonically decreasing with the mean output.

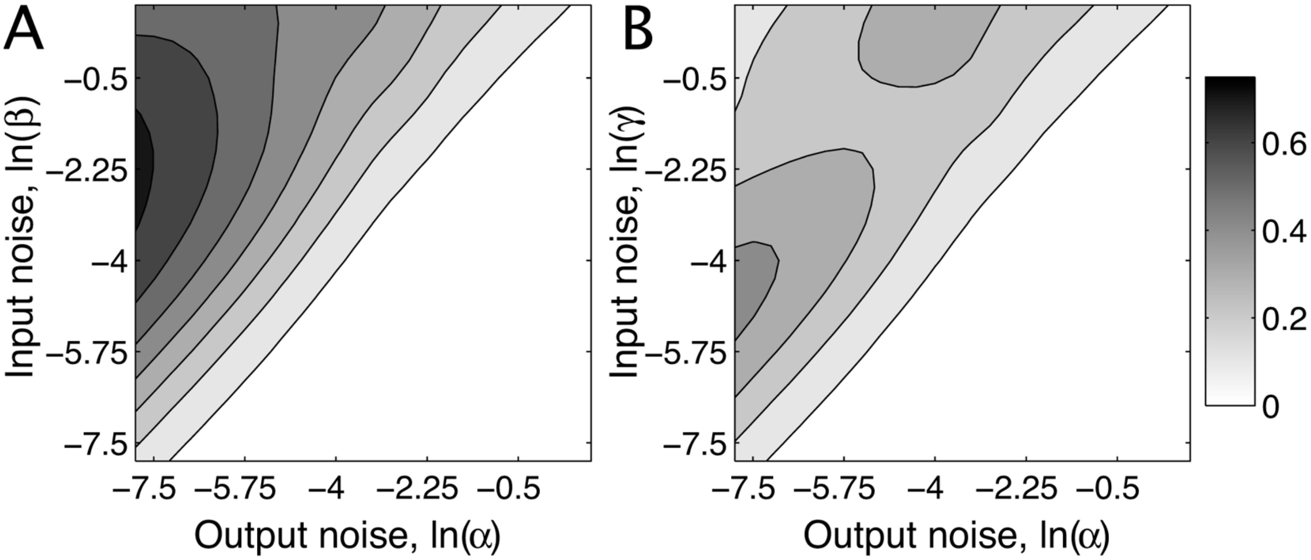


**FIG. 4.**

Difference in the information capacity between the re-pressors and activators (color code in bits). (a) shows $I_{\text{rep}}(h = 1) - I_{\text{act}}(h = 1)$, with the noise model that includes output ($\alpha$) and input diffusion noise ($\beta$) contributions [see Fig. 3 for absolute values of $I_{\text{act}}(h = 1)$]. (b) shows $I_{\text{rep}} - I_{\text{act}}$ for the noise model that includes output noise ($\alpha$)and input switching noise ($\gamma$) contributions; this plot is independent of cooperativity $h$.

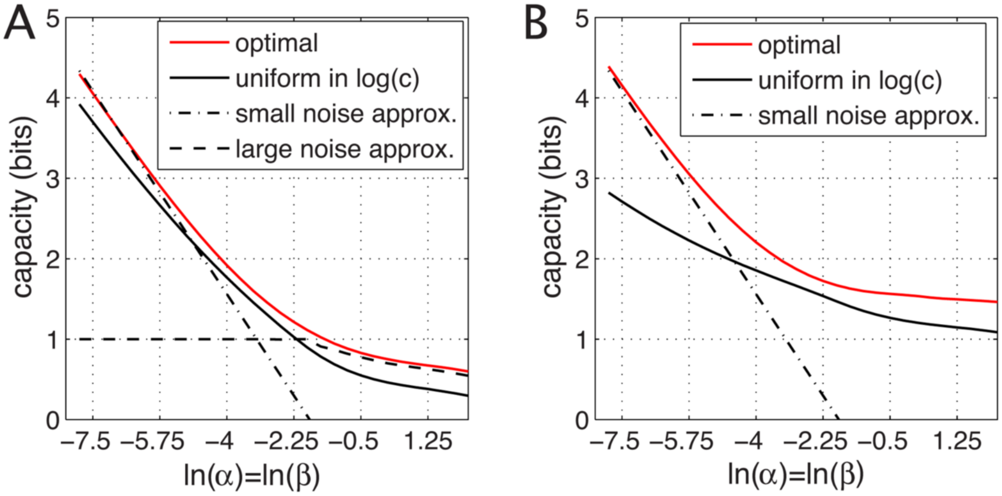


**FIG. 5.**

(Color online) Comparison of exact channel capacities and various approximate solutions. For both panels [(a) no cooperativity, $h = 1$; (b) strong cooperativity, $h = 3$] we take a cross section through the noise plane in Fig. 3 along the main diagonal, where the values for noise strength parameters $\alpha$ and $\beta$ are equal. The exact optimal solution is shown in red. By moving along the diagonal of the noise plane (and along the horizontal axis in the plots above) one changes both input and output noise by the same multiplicative factor $s$, and since, in the small-noise approximation, $I_{\mathrm{SNA}} \propto \log_2 Z$, $Z = \int \sigma_g(\bar{g})^{-1} d\bar{g}$, that factor results in an additive change in capacity by $\log_2 s$. We can use the large-noise-approximation lower bound on capacity for the case $h = 1$, in the parameter region where capacities fall below 1 bit.

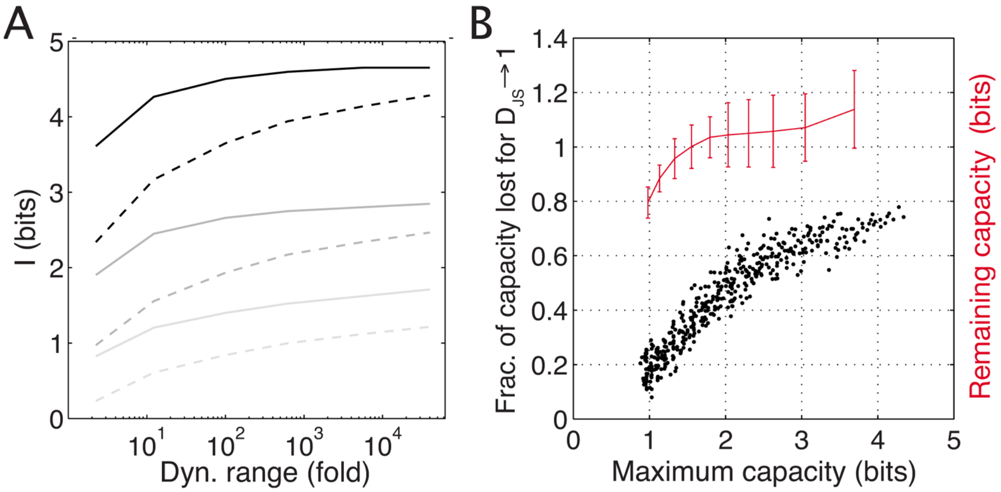


**FIG. 6.**

(Color online) Effects of imposing realistic constraints on the space of allowed input distributions. (a) shows the change in capacity if the dynamic range of the input around $K_d$ is changed ("25-fold range" means $c \in [K_d/5, 5K_d]$). The regulatory element is a repressor with either no cooperativity (dashed line) or high cooperativity $h$=3 (thick line). We plot three high-low cooperativity pairs for different choices of the output noise magnitude (high noise in light gray, $\ln \alpha \approx -2.5$; medium noise in dark gray, $\ln \alpha \approx -5$; low noise in black, $\ln \alpha \approx -7.5$). (b) shows the sensitivity of channel capacity to perturbations in the optimal input distribution. For various systems from Fig. 3 we construct suboptimal input distributions, as described in the text, compute the fraction of capacity lost relative to the unperturbed optimal solution, and plot this fraction against the optimal capacity of that system (black dots); extrapolated absolute capacity left when the input tends to be very different from optimal, i.e., $D_{\mathrm{JS}} \to 1$, is plotted in red.

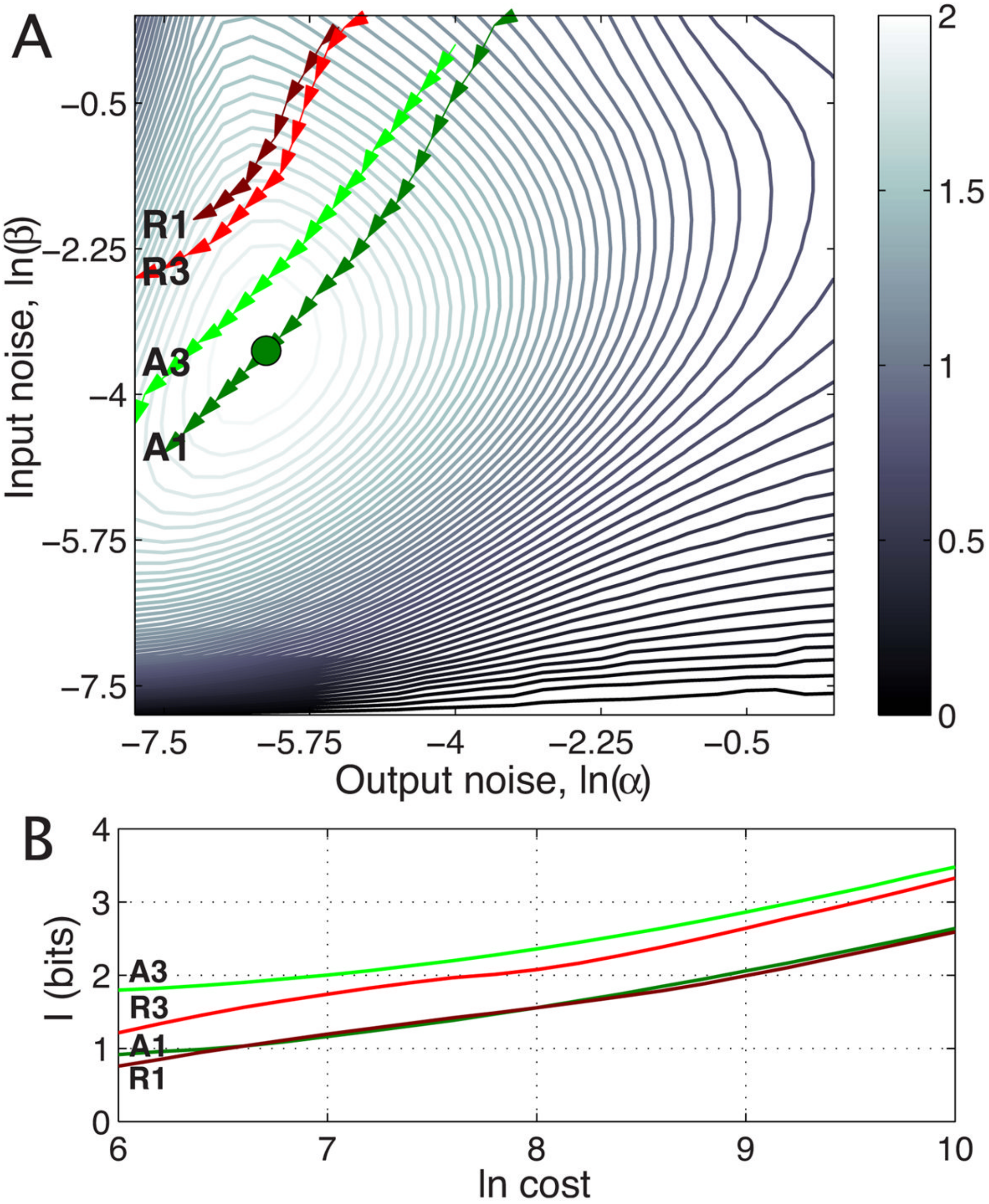


**FIG. 7.**

(Color online) Effects of metabolic or time costs on the achievable capacity of simple regulatory elements. Contours in (a) show the noise plane for noncooperative activator from Fig. 3, with the imposed constraint that the average total (input+output) cost is fixed to some $C_0$; as the cost is increased, the optimal solution (green dot) moves along the arrows on a dark green line (A1); the contours change correspondingly, not shown. Light green line (A3) shows activator with cooperativity $h$=3, dark (R1) and light red lines (R3) show repressors without and with cooperativity ($h$=3). (b) shows the achievable capacity as a function of cost for each line in (a).

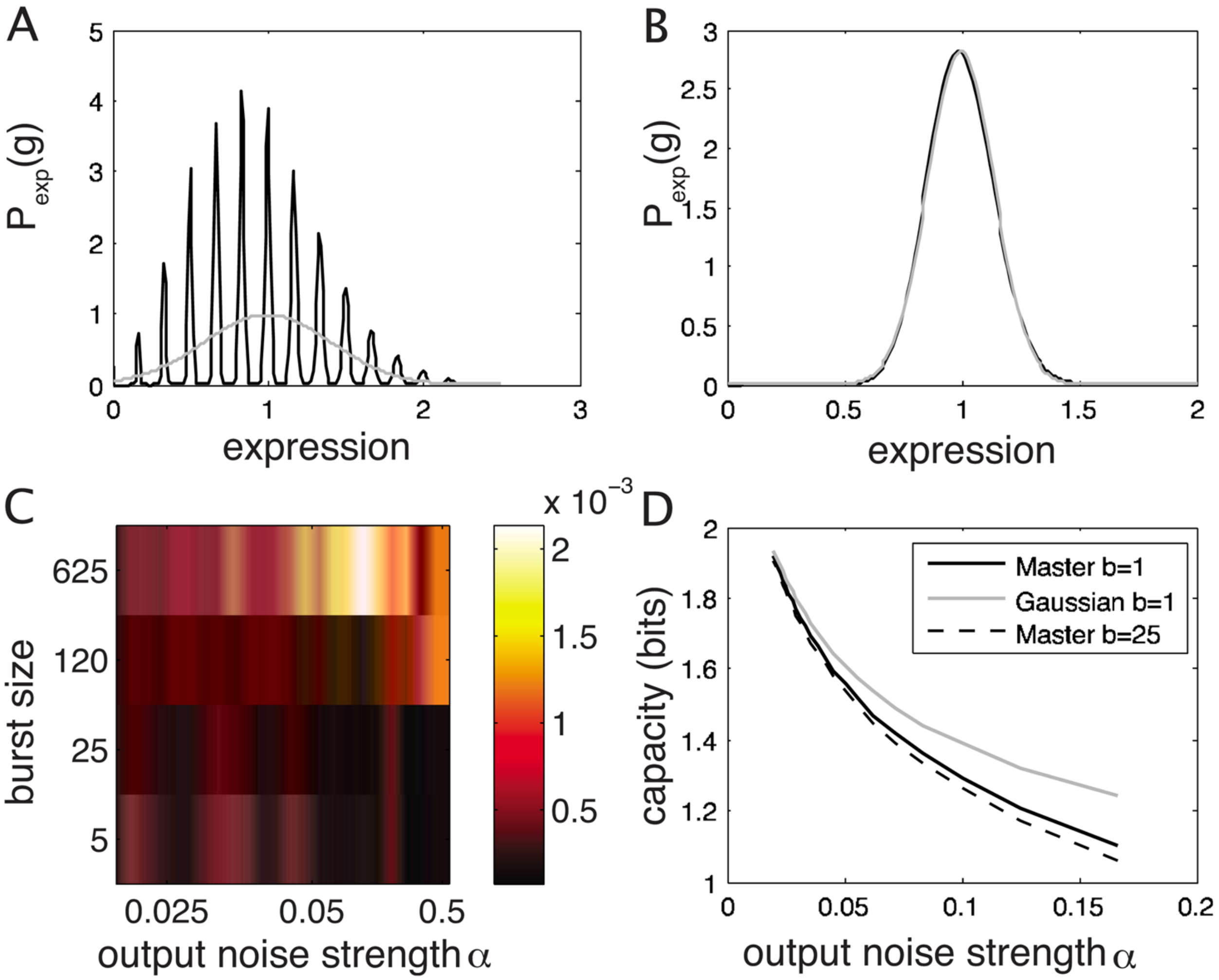


**FIG. 8.**

(Color online) Exact solutions (black) for input-output relations, $p(g|c)$, compared to their Gaussian approximations (gray). (a) shows the distribution of outputs at maximal induction, $p(g|c_{\max})$, for a system with a large burst size $b=5^4$ and a large output noise $\alpha = 1/6$ (i.e., the average number of messages is 6, as is evident from the number of peaks, each of which corresponds to a burst of translation at different number of messages). (b) shows the same distribution for smaller output noise, $b=5^2$ and $\alpha=1/50$; here the Gaussian approximation performs well. Both cases are computed with switching noise parameter $\gamma=1/50$ and cooperativity of $h=2$. (c) shows in color code the error made in computing the standard deviation of the output given $c$; the error measure we use is the maximum difference between the exact and Gaussian results over the full range of concentrations $\max_c \text{abs}\{[\sigma_g(c)/g_0]_{\text{master}} - [\sigma_g(c)/g_0]_{\text{Gaussian}}\}$. As expected the error decreases with decreasing output noise. (d) shows that the capacity is overestimated by using an approximate kernel, but the error again decreases with decreasing noise as Langevin becomes an increasingly good approximation to the true distribution. In the worst case the approximation is about 12% off. The Gaussian computation depends only on $\alpha$ and not separately on burst size, so we plot only one curve for $b=1$.

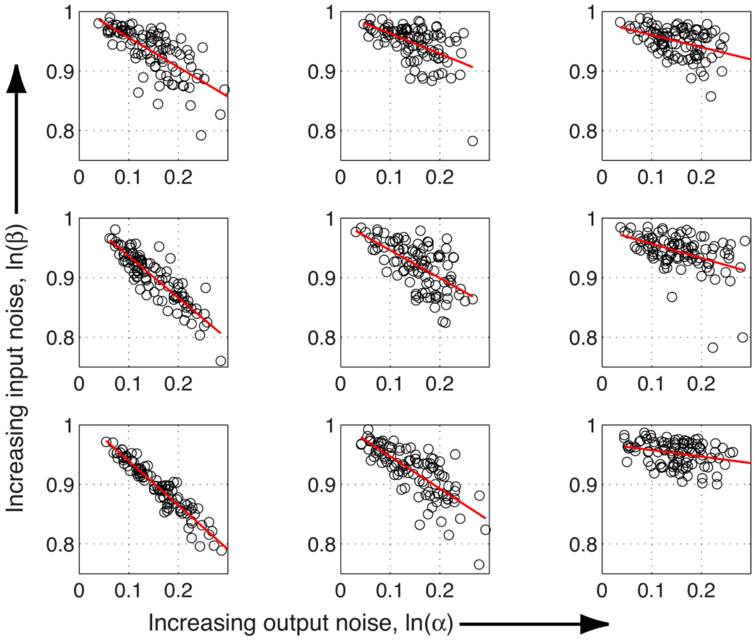


**FIG. 9.**

(Color online) Robustness of the optimal solutions to perturbations in the input distribution. Activator systems with no cooperativity are plotted; their parameters are taken from a uniformly spaced 3×3 grid of points in the noise plane of Fig. 3(a), such that the output noise increases along the horizontal edge of the figure and the input noise along the vertical edge. Each subplot shows a scatter plot of 100 perturbations from the ideal solution; the Jensen-Shannon distance from the optimal solution, $d_i$, is plotted on the horizontal axis and the channel capacity (normalized to maximum when there is no perturbation), $Ii_{/I\max}$, on the vertical axis. Red lines are best linear fits.

## TABLE I

Gaussian noise model parameters. Note that if burst size $b \gg 1$ then the output noise is determined by the average number of mRNA molecules, $\alpha \sim (\langle \mathrm{mRNA} \rangle)^{-1}$. Note also that if the on rate is diffusion limited, i.e., $k_+ = 4\pi Da$, then both input noise magnitudes $\beta$ and $\gamma$ are proportional to each other and decrease with increasing $k_-$, or alternatively, with increasing $K_d = k_-/k_+$. In steady state the system averages the fluctuations on the time scale of $\tau_{\mathrm{int}}$, so that all noise strengths $\alpha$, $\beta$, and $\gamma$ are inversely proportional to the integration time.

| Parameter | Value | Description |
|---|---|---|
| $\alpha$ | $(1+b)/g_0$ | Output noise strength |
| $\beta$ | $h^2/\pi DaK_d\tau_{\mathrm{int}}$ | Diffusion input noise strength |
| $\gamma$ | $(k_-\tau_{\mathrm{int}})^{-1}$ | Switching input noise strength |
| $h$ |  | Cooperativity (Hill coefficient) |