Fast inference of ill-posed problems within a convex space

J Fernandez-de-Cossio-Diaz and R Mulet

1 Department of Systems Biology, Center of Molecular Immunology, La Habana, Cuba
2 Group of Complex Systems and Statistical Physics, Department of Theoretical Physics, Physics Faculty, University of Havana, Cuba

E-mail: cossio@cim.sld.cu and mulet@fisica.uh.cu

Received 16 February 2016, revised 21 June 2016
Accepted for publication 28 June 2016
Published 21 July 2016

Abstract. In multiple scientific and technological applications we face the problem of having low dimensional data to be justified by a linear model defined in a high dimensional parameter space. The difference in dimensionality makes the problem ill-defined: the model is consistent with the data for many values of its parameters. The objective is to find the probability distribution of parameter values consistent with the data, a problem that can be cast as the exploration of a high dimensional convex polytope. In this work we introduce a novel algorithm to solve this problem efficiently. It provides results that are statistically indistinguishable from currently used numerical techniques while its running time scales linearly with the system size. We show that the algorithm performs robustly in many abstract and practical applications. As working examples we simulate the effects of restricting reaction fluxes on the space of feasible phenotypes of a genome scale Escherichia coli metabolic network and infer the traffic flow between origin and destination nodes in a real communication network.

Keywords: disordered systems (theory), message-passing algorithms, metabolic networks, peer-to-peer networks
1. Introduction

Although we live in the time of high-throughput data, or perhaps because of it, today many problems in science and technology have to deal with experiments that provide low dimensional data and a huge parameter space to be explored in search for a comprehension of this data. For example, in metabolic network analysis [1] a stoichiometric matrix of $N$ reactions between $M$ metabolites is given, and one is interested in estimating the fluxes through these reactions. In network tomography [2] data flows on $M$ peer-to-peer router links are known and the interest is to estimate the net flow of data between $N$ pairs of origin to destination nodes. In both cases $N$ is larger than $M$. Similar problems arise in other fields like compressed sensing [3], image super-resolution [4], positron emission tomography [5], the freezing transition of hard spheres [6] and density reconstruction from gravitational lensing in astrophysics [7]. All these examples may be recognized as linear ill-posed inverse problems where the dimension of the data available is smaller than the number of parameters to be explored looking for a consistent model. The problem is usually cast as the set of linear equations: $\mathbf{y} = \mathbf{S}\mathbf{x}$ with $\mathbf{x} \in [\mathbf{a}, \mathbf{b}]$, where the dimension of $\mathbf{y}$ is lower than the dimension of $\mathbf{x}$. Given $\mathbf{y}$ and $\mathbf{S}$ the task is to make inferences about $\mathbf{x}$.

In specific contexts it is known that the vector $\mathbf{x}$ is sparse, where the sparseness is defined by a properly chosen (sometimes arbitrarily) distance. In this case the problem
translates into an optimization problem: finding the sparsest vector in the set of possible solutions \[8, 9\]. Alternatively, the linear set of equations may be complemented with a linear objective function to be optimized \[10, 11\]. However, it seems difficult to justify similar strategies in other fields. A more general approach is to introduce a maximum entropy-like principle to be satisfied by the data, usually with the addition of some sort of side information that, depending on its pertinence or quality, may be good or bad for the inference. In this case the solution will depend, not only on the technique used to infer \(\tilde{x}\), i.e. entropy maximization (EM), maximum likelihood (ML), Markov random fields, hierarchical clustering, vector auto-regressive models (VAR) etc \[12–15\], but also on the prior proposed \[16–20\].

Within these approaches the ill-defined linear problem is transformed into a well defined one, with one solution relatively easy to find. The price to pay is the introduction of ad hoc assumptions that may be unjustified or may introduce unwanted biases hard to control. Without these assumptions it is mandatory to infer all the solutions of the system \(\tilde{y} = S\tilde{x}\), more specifically the probability of occurrence of each solution. This amounts to finding the points inside the convex polytope defined by a linear set of equations, or equivalently the volume of the high dimensional convex body \[21, 22\].

The exact determination of this volume is known to be \#P-hard and therefore exact algebraic methods are computationally infeasible even in low dimensional problems \[21\]. On the other hand stochastic approaches, like the hit and run Markov chain Monte Carlo (MCMC) \[23–25\] are widely used but still too expensive for very high dimensional problems. A different stochastic algorithm, based on message passing techniques borrowed from the information theory community, was recently introduced in \[26\] to study the volume of the phenotypic space of metabolic networks, and latter extended in \[27, 28\]. Unfortunately, the performance and efficiency of message passing techniques in problems with binary variables \[29–31\] is limited here because of the continuous and bounded support of the quantities involved.

To overcome this limitation we introduce a proper parametrization of the messages. The idea has been borrowed from the study of message passing equations with continuous variables \[3, 29\], but its application in problems with bounded support has been elusive for years. The reason is that in this case one must consider, together with the moments of the distributions, their corresponding bounds. We present in this work the solution to this problem as a set of fixed point equations and its corresponding implementation as a message passing algorithm over a finite number of parameters.

We test the algorithm performance on different ensembles of artificial graphs and show that it has very good convergence properties, provides results that are statistically equivalent to those given by hit and run Markov chain Monte Carlo (MCMC), while it is much faster—it scales linearly with system size. To further prove the capabilities of our algorithm we explore the effects of reaction knock-down experiments in a genome scale Escherichia coli metabolic network, unveiling the importance of redundancy in the network. Finally, we use our algorithm to infer the traffic flow between origin and destination nodes in the Abilene Network of the USA and show that our predictions compare remarkably well with the experimental available data.
2. Methods

We want to study the space of solutions $\mathbf{x}$ subject to the constraints:

$$\mathbf{y} = \mathbf{S}\mathbf{x}, \quad a \leq \mathbf{x} \leq b,$$

where $\mathbf{S}$ is an $M \times N$ matrix. We can set $\mathbf{y} = 0$ without loss of generality (it is enough to translate the origin of coordinates by a particular solution). If the matrix $\mathbf{S}$ is invertible, equation (1) can be solved for $\mathbf{x}$. However, in many applications of interest $M < N$ and the problem is ill-posed, because there may be many vectors $\mathbf{x}$ consistent with (1). In this case (1) defines a convex polytope in a $(N - M)$-dimensional manifold embedded in the $N$-dimensional space of the variables (we assume that $\mathbf{S}$ is full rank). In the absence of further information, there is no general criteria to favor one solution over another. This means that a uniform probability distribution over the solutions is the best way to describe our knowledge about the system:

$$P(\mathbf{x}) \propto \prod_{a=1}^{M} \delta\left(\sum_{i \in a} S_{ai}x_i\right)$$

for $a \leq \mathbf{x} \leq b$, and $P(\mathbf{x}) = 0$ otherwise, where $\delta(x)$ is Dirac’s delta function and the proportionality factor is a normalization constant. It is helpful to picture the system (1), or equivalently (2), as a factor graph [32]. This is a bipartite graph connecting variable node $i$ to equation node $a$ whenever variable $i$ participates in equation $a$, see figure 1. The Bethe approximation is the simplest approach to the problem and assumes that the factor graph is a tree. In this case:

$$P(\mathbf{x}) = \left[\prod_{a=1}^{M} P_{a}(\{x_{j}\}_{j \in a})\right] \left[\prod_{i=1}^{N} P_{i}(x_{i})^{1-d_{i}}\right]$$

where $i$ is a variable (node) index, $a$ is an equation (factor node) index, $d_{i}$ is the number of equations in which the variable $x_{i}$ participates (i.e. the degree of variable node $i$ in the factor graph) and

$$P_{a}(\{x_{j}\}_{j \in a}) = \int P(\mathbf{x}) d\{x_{j}\}_{j \notin a}$$

represent marginal probability distributions (see theorem 14.2 in [29] for a derivation). Of course one may wonder whether this approximation is useful in a real-world situation. It turns out that, in fact, it has worked successfully in a number of applications, for instance, constraint satisfaction problems [29], error correcting codes [33], perceptron learning [34] and metabolic networks [26]. The intuition is that if typical loop lengths are large enough, a tree is a good approximation to the statistical correlations in the network.
Fast inference of ill-posed problems within a convex space

Within this approximation the problem can be reduced to the solution of an iterative set of equations, usually called belief propagation algorithm, for two kinds of messages, \( m_{a \rightarrow i}(x_i) \) from the equation (factor node) \( a \) to the variable (node) \( x_i \) and \( n_{j \rightarrow a}(x_j) \) from the variable \( x_j \) to the factor node \( a \) [29, 32],

\[
m_{a \rightarrow i}(x_i) \propto \int_{\{x_i\} \in \mathcal{A}_i} \delta \left( \sum_{i \in a} S_{ai} x_i \right) \prod_{j \in a \setminus i} n_{j \rightarrow a}(x_j) \tag{6}
\]

\[
n_{i \rightarrow a}(x_i) \propto \prod_{c \in \lambda \setminus a} m_{c \rightarrow i}(x_i), \tag{7}
\]

which in turn are related to the marginals (4), (5):

\[
P_{a}(\{x_j\}_{j \in a}) \propto \delta \left( \sum_{j \in a} S_{aj} x_j \right) \prod_{i \in a} n_{i \rightarrow a}(x_i),
\]

\[
P_a(x_i) \propto \prod_{a \in i} m_{a \rightarrow i}(x_i). \tag{8}
\]

In practical terms an iteration like equations (6) and (7) is particularly cumbersome in the current problem of interest because \( x_i \) is defined on a continuous and bounded support [26]. In fact, although the number of message passing equations scales linearly with the system size, the continuous support imposes convolution and multiplication operations, whose costs scale as \( D^{d_a} \), where \( D \) is the discretization chosen for the distribution and \( d_a \) is the number of variables connected to a factor node.

An efficient solution is to parameterize the messages and to write the update rules (6) and (7) in terms of these parameters, significantly reducing the dimensionality of the stored arrays of values and the computations per iteration. In a convex body, the

\[
doi:10.1088/1742-5468/2016/07/073207
\]
Brunn-Minkowski theorem [35] implies that the marginal distributions are unimodal. In general they will be non-symmetric in the vicinity of the mode. A simple distribution capturing these properties is a generalized Beta distribution, which motivates the following ansatz:

\[
m_{a\to i}(x_i) \propto (x_i - A_{a\to i})^{\alpha_{a\to i}-1}(B_{a\to i} - x_i)^{\beta_{a\to i}-1}, \quad A_{a\to i} \leq x_i \leq B_{a\to i} \\
n_{i\to a}(x_i) \propto (x_i - A_{i\to a})^{\alpha_{i\to a}-1}(B_{i\to a} - x_i)^{\beta_{i\to a}-1}, \quad A_{i\to a} \leq x_i \leq B_{i\to a}.
\]

This is a proper parametrization of the messages, with the constrains \(\alpha_{a\to i}, \beta_{a\to i}, \beta_{i\to a} \geq 1\), \(a_i \leq A_{a\to i} \leq B_{a\to i} \leq b_i\), and \(a_i \leq A_{i\to a} \leq B_{i\to a} \leq b_i\). An illustrative example of the appropriateness of Beta distributions to reproduce the shape of marginal distributions in convex problems of the type (1) is given in the supplementary materials (see the appendix). The implementation of belief propagation that follows from this parameterization will be called BP/\(\beta\) in the rest of this paper. We now rewrite equations (6) and (7) in a manner consistent with (9). Noting that (6) is a convolution, it follows that \(\mu_{a\to i}\) and \(\sigma_{a\to i}\), the mean and variance of \(m_{a\to i}\), are given by:

\[
\mu_{a\to i} = \sum_{j \in a^+} S_{ai} S_{aj} \mu_{j\to a}, \quad \sigma_{a\to i}^2 = \sum_{j \in a^+} S_{ai}^2 S_{aj}^2 \sigma_{j\to a}^2
\]

where \(\mu_{i\to a}\) and \(\sigma_{i\to a}\) are the mean and variance of \(n_{i\to a}\). Similarly, the bounds \(A_{a\to i}, B_{a\to i}\) are the minimum and maximum values of the sum \(-\sum_{j \in a^+}(S_{aj}/S_{ai})x_j\), that is:

\[
A_{a\to i} = -\sum_{j \in a^+} S_{aj} A_{j\to a} - \sum_{j \in a^-} S_{aj} B_{j\to a} \\
B_{a\to i} = -\sum_{j \in a^+} S_{aj} B_{j\to a} - \sum_{j \in a^-} S_{aj} A_{j\to a}
\]

where \(j \in a^+\) denotes the set of j’s for which \(S_{aj}/S_{ai} > 0\), \(j \in a^-\) those for which \(S_{aj}/S_{ai} < 0\), and \(A_{i\to a}, B_{i\to a}\) are the bounds of the messages \(n_{i\to a}\). We obtain \(\alpha_{a\to i}, \beta_{a\to i}\) by inverting the equations giving the mean and variance of a generalized Beta distribution:

\[
\mu = \frac{A\beta + B\alpha}{\alpha + \beta}, \quad \sigma^2 = \frac{\alpha \beta (B - A)^2}{(\alpha + \beta)^2(1 + \alpha + \beta)},
\]

where the subindices are omitted for brevity. Finally, in the cases where \(A_{i\to a} < a_i\) or \(B_{i\to a} > b_i\), we compute the mean and variance of the Beta distribution truncated to \([\max(A_{a\to i}, a_i), \min(B_{a\to i}, b_i)]\), and from these recompute the values \(\alpha_{a\to i}, \beta_{a\to i}\), and reset \(A_{a\to i} := \max(A_{a\to i}, a_i), B_{a\to i} := \min(B_{a\to i}, b_i)\). This way we guarantee that message \(m_{a\to i}\) always has its support contained in the interval \([a_i, b_i]\). This way we complete the updating of messages from factor to variable node.

To update messages from variable to equation node, we multiply the messages \(m_{a\to i}\) as in equation (7), and then by fast single variable numerical integrations obtain the mean and variance of the resulting distributions:
Fast inference of ill-posed problems within a convex space

\[ \mu_i \rightarrow a = \frac{\int_{A_i \rightarrow a} x \prod_{c \in \land a} m_{c \rightarrow i}(x)dx}{\int_{B_i \rightarrow a} \prod_{c \in \land a} m_{c \rightarrow i}(x)dx}, \]

\[ \sigma_i^2 \rightarrow a = \frac{\int_{A_i \rightarrow a} x^2 \prod_{c \in \land a} m_{c \rightarrow i}(x)dx}{\int_{B_i \rightarrow a} \prod_{c \in \land a} m_{c \rightarrow i}(x)dx} - \mu_i^2 \rightarrow a. \]  

(13)

\( A_i \rightarrow a \) and \( B_i \rightarrow a \) are the bounds of the interval where the product (7) is non-zero. By the parameterization of the messages this implies that:

\[ A_i \rightarrow a = \max_{c \in \land a} A_c \rightarrow i, \quad B_i \rightarrow a = \min_{c \in \land a} B_c \rightarrow i. \]  

(14)

It can be shown that \( A_i \rightarrow a \leq B_i \rightarrow a \) if the system is feasible, because for any solution \( \bar{x} \) of (1), the inequality \( \max_a A_a \rightarrow i \leq x_i \leq \min_a B_a \rightarrow i \) will hold at all iterations if it is true at the first iteration, as can be easily shown.

On the other hand, it is possible to define an entropy in terms of the marginal probability distributions. This amounts to the logarithm of the volume of the convex solution space. Still assuming that the factor graph is a tree, one obtains the following expression:

\[ H = \sum_a H_a - \sum_i (d_i - 1)H_i, \]  

(15)

where \( H_a \) is the joint entropy of the variables participating in equation \( a \), and \( H_i \) is the entropy of variable \( i \),

\[ H_a = \ln \int \delta \left( \sum_{i \in a} S_i x_i \right) \prod_{i \in a} n_{i \rightarrow a}(x_i)dx_a - \sum_{(a)} \int P(x_i) \ln n_{i \rightarrow a}(x_i)dx_i \]  

(16)

\[ H_i = \int P_i(x_i) \ln P_i(x_i)dx_i. \]  

(17)

The entropy is calculated once the belief propagation iteration has converged. In our implementation we used the Fourier transform of the Beta distribution (a confluent hypergeometric function of the first kind) to evaluate the convolution in the first term of equation (16).

3. Results

3.1. Statistical analysis

We start this section comparing the output of our algorithm with the results of exact techniques in small instances of the problem. This is shown in figure 2(a) where we present a comparison (black rhombuses) between the volumes estimated using our implementation of \( \beta \) with the results obtained with lexicographic reverse search (lrs) [36], an exact scheme to compute the volume of low dimensional polytopes. In the data
shown, we worked with random sparse adjacency matrices $S$ of dimensions $N = 12$ and $M = 3$, with an average of $(k) = 5$ variables participating in each equation. All the non-null elements of matrices were extracted from the set $\{1, -1\}$ with equal probability but guaranteeing, to avoid inconsistencies, that each equation contains both elements of the set. All the variables were constrained to the interval $[0, 1]$.

To have a reference about the quality of the algorithm we present in the same figure results obtained with *hit and run* Monte Carlo (abbreviated to MC from now on) simulations. Since it is known that the mixing time of MC (the sample size to guarantee statistical independence) scales quadratically with the system size [25], we used a sampling time quadratic in $N$ for all our simulations involving MC. In figure 2(a) two sampling times were used: $t_{MC,1} = 10 \times N^2 = 1440$ (white circles) and $t_{MC,2} = 50 \times N^2 = 7200$ (crosses). As can be clearly observed the dispersion of results from MC simulations decreases as $t_{MC}$ increases, becoming similar to the dispersion of BP$\beta$. A more quantitative assessment of the dispersion may be obtained by using, as a figure of merit, the average relative error between distinct methods and lrs (i.e. $\epsilon = \frac{1}{Q} \sum_{i=1}^{Q} \frac{|V_i^l - V_i^s|}{V_i^s}$, where $Q$ is the number of random networks studied and $V_{x,y}^i$ is the volume computed with the method on the $x$ or $y$ axes).

![Figure 2](https://example.com/figure2.png)

**Figure 2.** Comparison of algorithms for the computation of the volume of the solution space of (1). Panel (a) compares the exact volume computations given by lrs with our algorithm (BP-$\beta$) and Monte Carlo with two sampling times. Each point is a random realization of (1) with 12 variables and 3 linear equations involving 5 variables each. A comparison of Monte Carlo and BP-$\beta$ for a larger system (100 variables, 25 linear equations, 5 variables per equation average) is shown in panel (b). Three sampling times are shown for Monte Carlo: $t_{MC,1} = 10^4$, $t_{MC,2} = 10^5$ and $t_{MC,3} = 5 \times 10^5$. As the sample time increases, the results agree more with those given by BP$\beta$, indicating that BP$\beta$ is more accurate than Monte Carlo with these sample times. In both panels $\epsilon$ is the mean relative error of the method on the $y$-axis with respect to method on the $x$-axis. (i.e. $\epsilon = \frac{1}{Q} \sum_{i=1}^{Q} \frac{|V_i^l - V_i^s|}{V_i^s}$, where $Q$ is the number of random networks studied and $V_{x,y}^i$ is the volume computed with the method on the $x$ or $y$ axes).
In figure 2(b) we checked the differences between the volumes computed using BP and MC simulations for larger networks, where exact computations using lrs are not feasible. We consider random systems [37] of size $N = 100$, with $M = 25$ linear equations and $\langle k \rangle = 5$ average variables per equation. We show results for three different MC times, $t_{MC,1} = N^2 = 10^4$, $t_{MC,2} = 10 \times N^2 = 10^5$ and $t_{MC,3} = 50 \times N^2 = 5 \times 10^5$. As can be seen in the figure, the larger the MC time, the lower the dispersion of the data, indicating that BP and MC are converging to the same results. We measured the relative error between BP and MC (i.e. $\epsilon = \frac{1}{Q} \sum_{i=1}^{Q} \frac{|V_{MC,i} - V_{BP,i}|}{V_{MC,i}}$). The figure also suggests that if $t_{MC}$ is small, MC underestimates the volumes. Similar results were obtained for other ensembles of adjacency matrices reflecting small world and scale free networks topologies (supplementary information).

Finally, in figure 3 we present the probability of convergence of the algorithm as a function of the number of variables per equation $\langle k \rangle$. In the figure we present data from two ensembles of random matrices, one with $N = 128$ (variables), $M = 64$ (equations), and the other with $N = 256$, $M = 64$. In each case, for each value of $\langle k \rangle$ (mean number of variables per equation) we generated 20 random instances and counted the fraction of systems that converged before 1000 iterations. Although this result clearly indicates that our algorithm fails when the matrices involved are too dense, we show below that it may still be useful in many practical situations. Moreover, we see that a larger ratio $N/M$ expands the range of values of $\langle k \rangle$ for which the algorithm converges. Increasing the size of the network with a fixed ratio $N/M$ and fixed $k$ does not affect the convergence probability significantly (supplementary materials).

3.2. Metabolic networks

A metabolic network is an engine that converts metabolites into other metabolites through a series of intra-cellular intermediate steps. The fundamental equation characterizing all functional states of a reconstructed biochemical reaction network is a mass

Figure 3. Probability of convergence (P) as a function of the average number of variables per equation ($\langle k \rangle$) in random networks of 64 equations, with 128 and 256 variables respectively. If a network takes more than $10^3$ iterations, we classify it as non-convergence. The fitted curves are intended as visual guides only.
Fast inference of ill-posed problems within a convex space

conservation law that imposes simple linear constraints between incoming and outgoing fluxes on every metabolite:

\[
\frac{d\bar{\rho}}{dt} = S\bar{x},
\]

where \(\bar{\rho}\) is the vector of the concentrations of the metabolites in the network and \(\bar{x}\) is the vector of the reaction fluxes. In this application the matrix \(S\) (called the stoichiometric matrix) contains the stoichiometric coefficients of each metabolite in each reaction. As long as just steady-state cellular properties are concerned one can assume that a variation in the concentration of metabolites in a cell can be ignored and considered constant. Therefore in case of fixed external conditions one can assume a (quasi) stationarity of the metabolite concentrations and consequently the lhs of (18) can be set to zero. Moreover, the reaction fluxes are usually restricted by lower and upper bounds defined by biochemical or thermodynamic considerations. Under these very general hypotheses the problem of describing the set of metabolic fluxes compatible with flux-balance constraints (18) is described mathematically by a system of equations identical to (1).

To validate the applicability of our algorithm in this scenario, we computed the marginal flux distributions of the red blood cell metabolic network, using the stoichiometric matrix presented in [38]. The results compare well with the standard Monte Carlo hit and run sampling (see figure A4 in supplementary materials).

We start studying the core of the \(E.\ coli\) metabolism to understand the effects of knock-down of single reactions in the volume of the solution space of the network [39]. It consists initially of 95 reactions and 72 metabolites [40]. The intuition is that the solution space defined by the set of linear equation (18) and the bounds of the reactions characterizes the set of phenotypic states of the metabolic network.

To reduce the small-loop burden of the network, we removed the smallest molecules. We also eliminated reactions that were inactive for any metabolic state in the medium defined by the exchange fluxes in the network [40] (see supplementary materials for details). Mathematically this means that there are variables which are fixed to zero by the constrains imposed which can be safely eliminated. Following [39] we reduce to half the bound of one of the reactions while keeping the other bounds fixed, and compute the volume of the modified network. The initial bounds through each reaction are taken from the literature [40], to approach as closely as possible a set of metabolic states that actually occurs in physiological conditions. The experiment is repeated for every reaction independently. See [39] for further discussion on the biological relevance of estimating the solution space volume in this context.

In figure 4 we present a map of the internal reactions of \(E.\ coli\) core metabolic network, where reactions are colored according to the impact on the solution space of the network after a 50% reduction of their bounds, as explained above. We can observe a highly heterogeneous impact of the distinct reactions. Reactions leading to secretion of lactate or alcohol are very significant in this experiment. Intuitively, this means that blocking the main secretion outlets will severely limit the space of metabolic states available. Also significant are the first reactions of glycolysis and the pentose-phosphate pathways, which are direct destinations of glucose in the core network. The production of glutamine is also important. In the core network glutamine can only be secreted,
but this is only because the core network misses most amino-acid metabolism. Another observation is that reactions with lower impacts are those that can be bypassed by alternative pathways. For example, inspection of figure 4 reveals that there is only one path to convert glutamate to glutamine, and that is through the reaction GLNS (glutamine synthetase), which has a high impact on the volume. However, there are two alternative routes to the reverse conversion of glutamine to glutamate, through GLUN (glutaminase) and through GLUSy (glutamate synthase), which may explain their smaller impact.

The role of redundancy is more dramatic when we compare these results with a similar experiment on the E. coli iJO1366 genome-scale metabolic network [41]. This large-scale reconstruction has 2583 reactions and 1805 metabolites. We submitted this network to the same reductions as E. coli core (see supplementary materials), leaving a working model with 1417 reactions and 950 metabolites. We performed knock-down of single reactions as described in the previous paragraph for each of the 70 reactions

![Figure 4. Internal reactions of E. coli core metabolism. In this drawing the nodes represent metabolites and the arrows connecting them are reactions. Each reaction is colored according to the entropy decrease of the space of steady metabolic flux distributions induced by a 50% flux reduction on that reaction.](image-url)
that are also present in the core network. The convergence of the algorithm after a knock-down starting from the wild type solution took nearly a minute in a personal computer with an Intel i5-processor. Most of the time of the calculation, approximately 10 minutes per knock-down, was spent calculating the volume.

Figure 5 shows a comparison of the ranking of the 70 core reactions according to the impact of the knock-down on the core network and on the genome-scale iJO1366 network. Interestingly, many reactions have different effects on the genome-scale network than they did in the core network, such as the group highlighted in red in figure 5. For the most part this is explained by the higher redundancy of iJO1366. The larger reconstruction contains alternative pathways for most reactions of the core network, reducing the global impact of knock-downs affecting these reactions. On the other hand, some reactions are seen to be very important in both networks. These are the reactions which are not bypassed by alternative pathways in the iJO1366 reconstruction. Two examples (highlighted in blue in the figure) are the production of alcohol (ALCD2x) and glutamine (GLNS). In both networks there is a single reaction capable of producing alcohol, thus lack of redundancy explains its importance in both reconstructions. Glutamine, on the other hand, is a central hub of amino acid metabolism in iJO1366. But in the core network it is secreted as a surrogate of the amino acid components of biomass. Again the lack of redundant pathways to reach this important metabolite in iJO1366 explains the simultaneous importance of GLNS in both reconstructions.

3.3. Network tomography

As a second application example of our algorithm we present a problem from the field of network tomography. A communication network is a collection of nodes representing computer terminals, routers, or subnetworks. Two nodes are linked if there is a direct connection between them that does not involve other nodes. Messages are transmitted
Fast inference of ill-posed problems within a convex space

by sending packets of bits from a source node to a destination node along a path consisting of one or more links and which generally passes through other nodes. In many real world applications it is not feasible to keep a centralized record of the origins and destinations of all the packets that have traversed the network, either because of bandwidth restraints or hardware limitations. The count of packets traversing each link is a more readily available datum. In this section we show that the BP\(\beta\) algorithm can be used to estimate the traffic flow between all source-destination pairs, given measurements of the traffic flow in each link.

We consider fixed routing networks, where the path between all source-destination pairs is known and remains the same for all packets traveling between these two nodes. For simplicity of notation source-destination pairs will be identified with a single letter index. Thus, let \(x_a(t)\) be the traffic intensity (packets per unit time) on the \(a\)th source-destination pair at time \(t\). The total traffic through the \(i\)th link, which we denote as \(y_i(t)\), is given by the sum of the source-destination flows where the path traverses the \(i\)th link. If we define a routing matrix \(S\) with components \(S_{ai} = 1\) if the \(i\)th link participates in the path associated to the \(a\)th source-destination pair, and \(S_{ai} = 0\) otherwise, it follows that

\[
\bar{y}(t) = S\bar{x}(t). \tag{19}
\]

Then, given measurements of link traffic \(\bar{y}(t)\), the problem is to estimate \(\bar{x}(t)\), subject to the non-negativity constraint \(\bar{x}(t) \geq 0\). Equation (19) is of the same form as (1). To make an analogy with metabolic networks, the metabolites are mapped in this context to the links of the communication network, and the reactions become source-destination pairs.

As a working example we use real world data from a portion of the internet, a sub-network known as Abilene [15]. Measurements of origin-destination flows were taken continuously over a seven-day period, starting on December 22, 2003. The relevant portion of the Abilene network, as it was at that time, may be conceptualized as a graph consisting of 11 nodes and 14 links (see figure 6). The original traffic flow counts were aggregated to five-minute intervals, a standard approach used to avoid issues of time synchronization across the network. We further averaged to 1 h intervals to mitigate the effects of noise [42]. From the origin-destination traffic data, we generated link flows using equation (19). Then from the link flow data, we try to estimate the original origin-destination traffic. These estimates are calculated as averages over the marginal distributions returned by BP-\(\beta\) using as upper bounds the maximum flow observed throughout the experiment.

Figure 7(a) shows a scatter plot of the aggregated traffic over all origin-destination pairs and for all times. It is clear from the figure that there is a good correlation between real and inferred flows, with an aggregated root mean square error of 0.07. A peculiar cluster of points in the lower-left corner shows small but systematic overestimation. Closer inspection reveals that these points originate from a single origin-destination pair in the network: Denver-Indianapolis. The path associated with this pair consists of the single link Denver-Indianapolis, which participates in over 15 other origin-destination paths in the network. Thus the traffic measurements on this link give very poor information about the flow on the Denver-Indianapolis path. Under these circumstances inference about the path Denver-Indianapolis is hard, and our algorithm...
is doing the best it can with the information given. Any improvement on this should come from additional information about the network.

To have a more intuitive picture of the quality of our algorithm we show in figure 7(b) the estimates and the experimental traffic as functions of time for four different origin-destination links. The relative error of our estimations, defined as the difference between real and estimated values divided by the real value, has an average of 0.3 over all origin-destination pairs and all times. More sophisticated algorithms, which make
additional assumptions about time correlations or incorporate prior probabilities, score average relative errors between 0.1 and 0.4 [15, 43, 44]. Our simple method highlights the value of inferences drawn directly from the structure of the problem without incorporating *a priori* assumptions.

4. Conclusions

In this work we presented an efficient and robust algorithm to estimate the volume of a convex polytope in high dimensions. The algorithm allows the direct computation of the marginal distributions of variables defined as the solution of linear ill-posed problems in a convex space. Our implementation provides results that are statistically indistinguishable from other numerical techniques, with much less computational effort, it scales linearly with the system size, and has a finite operation cost per iteration. The algorithm was tested in various applications, starting with random systems of equations to show its robustness. Then, taking profit of its efficiency, we studied an *E. coli* genome scale metabolic network for multiple knock-downs, unveiling the importance of non-redundant reactions in the functional flexibility of the network. We compare with the effects of knock-downs in the core metabolic network of *E. coli*, obtaining significant differences. This highlights the abundant redundancy in the genome-scale model that is not present in the core model. On the other hand, knock-down of some reactions have similar effects in both networks, for example alcohol production, since its important role as a secreted metabolite is captured in both models. Differences in other reactions, such as ACALD, which is essential in the core network for the production of ethanol, can be explained by the presence of alternative pathways to the same products in the genome-scale model. As a final working example we studied a problem from the field of Network Tomography, the estimation of the origin-destination traffic in a communication network, from the traffic flow in the individual links. Again, our results compare statistically well with other numerical estimates that make assumptions about correlations or incorporate prior probabilities about the traffic flow. As a by-product we also showed that, for specific origin-destination pairs, in particular the Denver-Indianapolis pair of the Abilene network, the correct estimation of the corresponding origin-destination flow cannot be done without introducing additional external information.

Appendix. Supplementary material

A.1. Convergence time & comparison to discretized BP

We generated random networks of different sizes and ran our algorithm on each, measuring the time taken until it converged. We also ran hit-and-run Monte Carlo, using the standard prescription of $N$ steps between points so as to ensure independence [25]. The results are shown in figure A1. We also compared our algorithm to a discretization of the standard belief propagation equations, as is done in [26]. The advantage of our approach is the significant gain in speed, and as figure A2 shows, the Beta ansatz is able
Fast inference of ill-posed problems within a convex space

Figure A1. Convergence time versus number of variables in a random network, BP$\beta$ compared to Monte-Carlo hit and run.

Figure A2. Single-variable beliefs computed according to discretized belief propagation (as in [26]) and BP-$\beta$ in a small random network of 6 variables and 3 equations. Only three variable beliefs are shown; the other three are very similar.

to reproduce the shapes of distributions in the convex problems that we consider. Finally, figure A3 shows that increasing the network size maintaining a constant ratio $N/M$ and fixed $k$ has a small impact on the probability of convergence over a wide range. The greatest impact seems to come from a variation in $k$, as shown in the main text, figure 3.

A.2. Red blood cell

We used BP$\beta$ algorithm and Monte Carlo hit and run sampling to obtain the marginal flux distributions for each of the reactions in the red blood cell metabolism taking the same stoichiometric matrix presented in [38]. The network contains 46 reactions and 34 metabolites. All reactions are irreversible, with upper bounds set to realistic physiological values given in [38]. We compared BP$\beta$ with a set of 10 000 feasible solutions generated by Monte Carlo sampling. As seen in figure A4, the predictions of both methods compare rather well.

A.2.1. Simplifications of E. coli metabolic network

To reduce the small-cycle burden of the network, we eliminated the smallest molecules from the list of metabolites. By this criteria we manually eliminated the following metabolites: $O_2$, $H_2O$, $NH_4$, Pi, H, $CO_2$. We also deleted reactions that were constrained to zero-flux by the nutrients available,
as defined by the exchange fluxes in the network (reaction bounds were taken from [40]). After these two steps, we deleted metabolites that did not participate in any reaction, as well as empty reactions.

We also removed the reaction of biomass production, since it was not our interest to bias the solution space with this objective function. In order to model the net production of metabolites in the network, which in physiological conditions are diverted to various activities in the cell, such as biomass production, protein synthesis, or even

Figure A3. Probability of convergence over random ensembles of random networks, with a fixed ratio of variables to equation of $N/M = 2$, varying the average variables per equation ($k$). The probability and error bars are estimated from 40 repetitions per point.

Figure A4. Flux distributions of the reactions in the red blood cell metabolic network, computed with Monte Carlo hit and run (gray, filled) and with our BP-β algorithm (black contours).
degradation, we included drain reactions for every metabolite. Finally, pairs of irreversible reactions that were mirrors of each other were merged into a single reversible reaction.

The resulting reduced model of \textit{E. coli} core metabolism has 70 reactions and 56 metabolites.

### A.3. Scale free and small-world random networks

Figure 2(b) in the main text compares the accuracy of our algorithm with that of a Monte Carlo \textit{hit and run} method on a set of random networks generated according to the Erdős & Rényi model \cite{37}. We also did this comparison on two additional models of random networks: the small-world model and the scale-free model \cite{45}. As can be seen in figures A5(a) and (b), the results are qualitatively the same as in figure 2(b), validating the applicability of the algorithm on real networks.

In the small world model the random construction begins with a network embedded in space where each node is linked to some of its closest neighbors. Then random links are inserted between random pairs of nodes independently of the distance between them, bringing closer together nodes that are far apart in the original spatial arrangement of the network (hence the name ‘small world’). In the scale-free model nodes are linked randomly with a preference for attachment to nodes that already have the largest degrees. The small-world and the scale-free models of random networks were created to try to explain properties observed in real large-scale networks that weren’t present in the original random network model of Erdős & Rényi \cite{45}. To generate small-world random systems of equations, we started with a factor graph consisting of a cycle of alternating equations and variables. Then we added links...
Fast inference of ill-posed problems within a convex space

connecting random pairs of equations and variables. To generate scale-free random systems of equations, we started with a network where each equation is connected to two variables with opposite signs. This ensures that the network is consistent. Then we added random variables, which have a probability of linking an equation proportional to that equation’s degree. Each added variable has a fixed degree of 3.

A.4. Supplementary programs

The C++ source code of the BPβ algorithm used in the simulations in this paper are available from the authors upon request.

References

[1] Schellenberger J and Palsson B Ø 2009 Use of randomized sampling for analysis of metabolic networks J. Biol. Chem. 284 5457–61
[2] Vardi Y 1996 Network tomography: estimating source-destination traffic intensities from link data J. Am. Stat. Assoc. 91 365–77
[3] Krzakala F, Mézard M, Sausset F, Sun Y and Zdeborová L 2012 Statistical physics-based reconstruction in compressed sensing Phys. Rev. X 2 021005
[4] Shepp L A and Krouskal J B 1978 Computerized tomography: the new medical x-ray technology Am. Math. Mon. 85 420
[5] Vardi Y, Shepp L A and Kaufman L 1985 A statistical model for positron emission tomography J. Am. Stat. Assoc. 80 8–20
[6] Kapfer S and Krauth W 2013 Sampling from a polytope and hard disk monte carlo arXiv:1301.4901
[7] Lubini M and Coles J 2012 A sampling strategy for high dimensional spaces applied to free-form gravitational lensing Mon. Not. R. Astron. Soc. 425 3077
[8] Donoho D L and Elad M 2003 Optimally sparse representation in general (non-orthogonal) dictionaries via L1 minimization Proc. Natl Acad. Sci. 100 2197
[9] Candès E, Romberg J and Tao T 2006 Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information IEEE Trans. Inf. Theory 52 48
[10] Palsson B Ø 2006 Systems Biology. Properties of Reconstructed Networks (Cambridge: Cambridge University Press)
[11] Papadimitriou Ch H 1998 Combinatorial Optimization: Algorithms and Complexity (New York: Dover)
[12] Werhli A V 2012 A comparison of algorithms for inference and learning in probabilistic graphical models IEEE Trans. Pattern Anal. Mach. Intell. 27 1
[13] Kolaczyk E D 2009 Statistical Analysis of Network Data (New York: Springer)
[14] Medina A, Taft N, Salamatian K, Bhattacharyya S and Diot C 2002 Traffic matrix estimation: existing techniques and new directions ACM SIGCOMM (Pittsburg, PA, USA)
[15] Olsen C, Bontempi G, Emmert-Streib F, Quackenbush J and Haibe-Kains B 2014 Frontiers Genet. 5 177
[16] Scharnagl B, Vrught J A, Vereecken H and Herbst M 2011 Hydrol. Earth Syst. Sci. 15 3043
[17] Verbyla K L, Bowman P J, Hayes B J and Goddard M E 2010 Sensitivity of genomic selection to using different prior distributions BMC Proc. 4 S5
[18] Lalush D S 1992 Simulation evaluation of gibbs prior distributions for use in maximum a posteriori spect reconstructions IEEE Trans. Med. Imaging 11 267
[19] Dyer M E and Frieze A M 1988 On the complexity of computing the volume of a polyhedron SIAM J. Comput. 17 967
[20] Büeler B, Enge A and Fukuda K 2000 Exact volume computation for polytopes: a practical study Polytopes—Combinatorics and Computation (Basel: Birkhäuser) pp 131–54
[21] Smith R L 1984 Hit-and-run algorithms for generating multivariate distributions Oper. Res. 32 1296
[22] Turcin V 1971 On the computation of multidimensional integrals by the monte-carlo method Theory Probab. Appl. 16 720

doi:10.1088/1742-5468/2016/07/073207
Fast inference of ill-posed problems within a convex space

[25] De Martino D, Mori M and Parisi V 2014 Uniform sampling of steady states in metabolic networks: heterogeneous scales and rounding arXiv:1312.5228

[26] Braunstein A, Mulet R and Pagnani A 2008 Estimating the size of the solution space of metabolic networks BMC Bioinf. 9 240

[27] Massucci F A, Font-Clos F, De Martino A and Castillo I P 2013 A novel methodology to estimate metabolic flux distributions in constraint-based models Metabolites 3 838–52

[28] Font-Clos F, Massucci F A and Castillo I P 2012 A weighted belief-propagation algorithm for estimating volume-related properties of random polytopes J. Stat. Mech. P11003

[29] Mézard M and Montanari A 2009 Information, Physics and, Computation (Oxford: Oxford University Press)

[30] Mézard M and Zecchina R 2002 Random k-satisfiability problem: from an analytic solution to an efficient algorithm Phys. Rev. E 66 056126

[31] Braunstein A, Mulet R, Pagnani A and Zecchina R 2003 Polynomial iterative algorithms for coloring and analyzing random graphs Phys. Rev. E 68 036702

[32] Kschischang F R and Loeliger H-A 2001 Factor graphs and the sum-product algorithm IEEE Trans. Inf. Theory 47 498

[33] Richardson T and Urbanke R 2001 The capacity of low-density parity check codes under message passing decoding IEEE Trans. Inf. Theory 47 599

[34] Baldassi C, Braunstein A, Brunel N and Zecchina R 2007 Efficient supervised learning in networks with binary synapses Proc. Natl Acad. Sci. 104 11079

[35] Guggenheimer H W 1977 Applicable Geometry (Malabar, FL: Krieger)

[36] Avis D and Fukuda K 1990 A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra Technical Report B-237 Tokyo Institute of Technology

[37] Erdős P and Rényi A 1960 On the evolution of random graphs Publ. Math. Inst. Hung. Acad. Sci. 5 17–61

[38] Wiback S J, Famili I, Greenberg H J and Palsson B Ø 2004 Monte Carlo sampling can be used to determine the size and shape of the steady-state flux space J. Theor. Biol. 228 437–47

[39] Price N D, Schellenberger J and Palsson B Ø 2004 Uniform sampling of steady-state flux spaces: Means to design experiments and to interpret enzymopathies Biophys. J. 87 2172–86

[40] Orth J D, Fleming R M T and Palsson B Ø 2010 Reconstruction and use of microbial metabolic networks: the core Escherichia coli metabolic model as an educational guide EcoSal Plus 2010 5538–48

[41] Orth J D, Conrad T M, Na J, Lerman J A, Nam H, Feist A M and Palsson B Ø 2011 A comprehensive genome-scale reconstruction of Escherichia coli metabolism Mol. Syst. Biol. 7 535

[42] Zhang Y, Roughan M, Duffield N and Greenberg A 2003 Fast accurate computation of large-scale IP traffic matrices from link loads ACM SIGMETRICS Performance Evaluation Review vol 31 (New York: ACM) pp 206–17

[43] Zhang Y, Roughan M, Lund C and Donoho D 2003 An information-theoretic approach to traffic matrix estimation Proc. of the 2003 Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communications (New York: ACM) pp 301–12

[44] Lakhina A, Papagiannaki K, Crovella M, Diot C, Kolaczyk E D and Taft N 2004 Structural analysis of network traffic flows ACM SIGMETRICS Performance Evaluation Review 32 61

[45] Réka A and Barabási A-L 2002 Statistical mechanics of complex networks Rev. Mod. Phys. 74 47

doi:10.1088/1742-5468/2016/07/073207