On Constraint Problems with Incomplete or Erroneous Data

Neil Yorke-Smith and Carmen Gervet

IC–Parc, Imperial College, London, SW7 2AZ, U.K.
{nys,cg6}@icparc.ic.ac.uk

Abstract. Real-world constraint problems abound with uncertainty. Problems with incomplete or erroneous data are often simplified at present to tractable deterministic models, or modified using error correction methods, with the aim of seeking a solution. However, this can lead us to solve the wrong problem because of the approximations made, an outcome of little help to the user who expects the right problem to be tackled and correct information returned. The certainty closure framework aims at fulfilling these expectations of correct, reliable reasoning in the presence of uncertain data. In this short paper we give an intuition and brief overview of the framework. We define the certainty closure to an uncertain constraint problem and show how it can be derived by transformation to an equivalent certain problem. We outline an application of the framework to a real-world network traffic analysis problem.

1 Motivation

Real-world Large Scale Combinatorial Optimisation problems (LSCOs) have inherent data uncertainties. The uncertainty can be due to the dynamic and unpredictable nature of the commercial world, but also due to the information available to those modelling the problem. In this paper we are concerned with the latter form of uncertainty, which can arise when the data is not fully known or is even erroneous. This work is motivated by practical issues we faced when working on two real-world applications, the first in energy trading [4] and the second in network traffic analysis [5].

In both applications the data information is incomplete or erroneous. In the energy trading problem, the demand and cost profiles have evolved due to market privatisation; the only data available is the profiles from previous years. Thus the existing simulation or stochastic data models would not help address the actual problem after market deregulation. In the network traffic analysis problem, we are forced to use partial data due to the overwhelming amount of information. Further, the data can be erroneous due to unrecorded packet loss or time gaps between readings of router tables.

In such cases, to grasp the actual complexity of the LSCO and bring useful feedback to the user, it is crucial to deal with real data despite its incompleteness or intrinsic errors. The core question is how to ensure that the correct problem is formulated and that correct information is derived.

We propose a formal framework, based on the CSP formalism, that can be applied to the heterogeneous constraint classes and computation domains found in real-world LSCOs. The contribution of the certainty closure framework is three-fold:

- **Data as a primary concept.** We complete the CSP-based description of a LSCO by bringing data as a primary concept into the CSP formalism. We model explicitly what is known about the uncertain data (for instance, by an interval of values) in terms of an *Uncertain Constraint Satisfaction Problem*. The absence of approximation and data correction allows us to separate the data issues from those related to the constraint model.
- **Closure of a UCSP**. We introduce the *certainty closure*, a correct set of solutions to the UCSP that excludes no solution possible given the present knowledge of the data. We do not attempt to compute an unwarranted single 'solution'. Rather, the certainty closure can provide useful insight into the uncertain problem, for instance by identifying actual reasons for unsatisfiability.
- **Correctness and tractability.** The paradigm we follow is model accuracy: correctness of the information derived without approximations or assumptions. We propose two resolution forms to derive the certainty closure: enumeration and transformation. The latter, central to our contribution, aims at reasoning about the UCSP by transforming it into an equivalent CSP.

Our purpose in essence consists of introducing to constraint programming concepts and properties that enable us to reason about these classes of problems. In order to ensure the practical value of the framework, we address specifically how the closure of an uncertain constraint problem can be derived in a tractable way.

2 Uncertain CSP: Intuition

We give an intuition of the certainty closure framework by applying it to an example of the networking problem. Consider the fragment of an IP network shown in Fig. 1. Four nodes, corresponding to routers and designated A to D, are shown, together with the bidirectional traffic flow on each link. Each router makes decisions on how to direct traffic it receives, based on a routing algorithm and local flow information. The problem is to determine guaranteed bounds for the traffic flow between end-points from measurements made at routers [5].

The network can be modelled as a CSP. The variables are the volume of traffic, $V_{ij} \in \mathbb{R}^+$, entering at node i and leaving at node j. The constraints state that the volume of traffic through each link in each direction is the sum of the traffic entering the link in that direction. There is also an upper bound on the flows that use only a single link, such as V_{ab} and V_{ad}.

Due to the volume of information required to represent the traffic data exactly, and to the time gap between reading router tables, it is not possible to sample the true traffic in an entire network at one instant. Instead, we measure the aggregated flow volume on a link over a given time interval, and derive the volume of traffic on the link as the difference between end and start measurements. The result is that the data information obtained is inevitably incomplete and erroneous. On the link D→C, for example, the flow might measure as 70 at D and as 80 at C, whereas the true value, equal at both nodes, is presumably somewhere between. Classically, we might choose one value in the possible range to work with: the median, for instance.

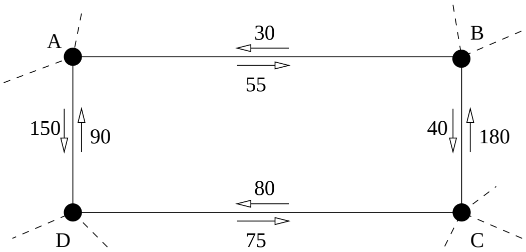

Fig. 1. Traffic flow in a network fragment

Secondly, when there are two paths of equal cost (from the perspective of the routing algorithm), the traffic is split equally between them in 90% of cases. This is true for flows from A to C, for example. To simplify the model, the current approach assumes that the traffic is split equally in all such cases. Supposing these modelling decisions, consider the traffic flow problem. There are eight traffic constraints:

$$\begin{array}{ll}
\mathtt{A}\rightarrow\mathtt{D} & V_{ad} + 0.5V_{ac} + 0.5V_{bd} = 150 \\
\mathtt{D}\rightarrow\mathtt{C} & V_{dc} + 0.5V_{db} + 0.5V_{ac} = 75 \\
\mathtt{C}\rightarrow\mathtt{B} & V_{cb} + 0.5V_{ca} + 0.5V_{db} = 180 \\
\mathtt{B}\rightarrow\mathtt{A} & V_{ba} + 0.5V_{bd} + 0.5V_{ca} = 30
\end{array}$$

(the other four similarly in the clockwise direction).

Unfortunately, the CSP is unsatisfiable. The current approach to the problem uses a data correction procedure (minimising deviation on the link volumes) in order to reach a satisfiable model. Another common interpretation would be that the problem is simply over-constrained. But could the lack of solution be due to the assumptions and approximations made? Suppose we do not attempt error correction and, further, model the actual dispatching of traffic, known to be anywhere between 30–70% between two paths. The more accurate CSP that results is again unsatisfiable.

At this point the erroneous approximation to the traffic flow data becomes clear. Let us remove all approximations and represent the uncertain flow measurements explicitly. Modelling the problem as an uncertain CSP, we have:

$$\begin{array}{ll}
\mathtt{A}\rightarrow\mathtt{D} & V_{ad} + [0.3, 0.7]V_{ac} + [0.3, 0.7]V_{bd} = [135, 160] \\
\mathtt{D}\rightarrow\mathtt{C} & V_{dc} + [0.3, 0.7]V_{db} + [0.3, 0.7]V_{ac} = [70, 80] \\
\mathtt{C}\rightarrow\mathtt{B} & V_{cb} + [0.3, 0.7]V_{ca} + [0.3, 0.7]V_{db} = [180, 190] \\
\mathtt{B}\rightarrow\mathtt{A} & V_{ba} + [0.3, 0.7]V_{bd} + [0.3, 0.7]V_{ca} = [25, 40]
\end{array}$$

We cannot hope to give a single value assignment and declare it to be 'the' true solution; however we can give an interval for each variable and assert that any possible solution lies within. Hence the best information we can produce based on the new traffic flow model is: $V_{ac} \in [0, 150], V_{ad} \in [30, 64], V_{db} \in [32, 200], V_{dc} \in [0, 40], V_{ca} \in [0, 133], V_{cb} \in [17, 64], V_{bd} \in [0, 133]$, and $V_{ba} \in [0, 20]$ (and the four single-link flows in the clockwise direction). This is the certainty closure. The system tells us that there is a solution to the problem that corresponds to at least one possible realisation of the data. If more information about the data becomes available to us, then it will be possible to refine the closure.

This illustrative introduction to uncertain constraint problems highlights the main aspects of uncertain constraint models and benefits of the certainty closure framework. It shows in particular that model accuracy and reliable quantitative results can be more valuable than seeking a solution at the cost of correctness. Further, representing incomplete or erroneous data adequately can reveal the true reason for model unsatisfiability.

In the sequel we outline how the closure can be obtained in a tractable manner.

3 Finding the Closure: Overview

Informally, an *uncertain CSP* is a simple extension to a classical CSP with an explicit description of the data. The *certainty closure* of an uncertain CSP P is the union of solutions to P such that every solution that is feasible under one or more realisations of P is contained in the closure. A *realisation* of the data is the selection of one value for each data coefficient from its *uncertainty set* of possible values; each realisation gives rise from P to a classical, certain CSP. While the examples here use intervals for the uncertainty set, since this is the form of the data in the networking problem, in general the data may be represented in any way — a set of values, an interval, an ellipsoid, or otherwise — as fits the computation domain.

We have identified two resolution forms to derive the certainty closure: enumeration and transformation. The first is to consider every data realisation: each gives rise to a certain CSP, which we solve, and the closure is then formed from all the solutions to the satisfiable CSPs. Unfortunately this approach can be exponential in the number of CSPs generated and is not suited to handling continuous data. The second form is to transform the uncertain CSP into a single, equivalent certain CSP. The objective is to ensure both correctness and tractability by seeking a CSP which can be solved efficiently using existing techniques, and whose complete solution set coincides with the certainty closure of the UCSP. For space reasons we will only outline the transformation method.

Solving an Equivalent CSP This resolution form consists of transforming the uncertain constraint problem to an equivalent certain problem, solvable by existing resolution techniques. The transformation aims at deriving a standard CSP model that has (or contains, in the worst case) the same set of solutions as the UCSP. The theoretical issues related to the approach are: (i) the identification of the constraint classes that allow us to find a transformation operator, and (ii) the definition of the transformation operator and its properties such that the two problems are solution-equivalent.

Without describing in details the properties of the operator, we briefly outline the transformation operation we have implemented for the networking problem. The UCSP corresponding to the problem has linear network flow constraints with positive real intervals as coefficients; it is an instance of a *positive orthant interval linear system*[1]. We illustrate the transformation on linear constraints over two variables X_1 and X_2 with

[1] In 2D, for instance, the positive orthant is the upper-right quadrant.

domains in $\mathbb{R}^+$. Suppose uncertain constraints of the form $\mathbf{a_1}X_1 + \mathbf{a_2}X_2 \le \mathbf{a_3}$, where $\mathbf{a_i} = [\underline{\mathbf{a_i}}, \overline{\mathbf{a_i}}]$ are real, closed intervals. Then the operator transforms each constraint separately in the following way:

$$\mathbf{a_1}X_1 + \mathbf{a_2}X_2 \le \mathbf{a_3} \rightarrow \begin{cases} \underline{\mathbf{a_1}}X_1 + \underline{\mathbf{a_2}}X_2 \le \overline{\mathbf{a_3}} & \text{if } \underline{\mathbf{a_2}} \ge 0 \\ \underline{\mathbf{a_1}}X_1 + \overline{\mathbf{a_2}}X_2 \le \underline{\mathbf{a_3}} & \text{if } 0 \in \mathbf{a_2} \\ \overline{\mathbf{a_1}}X_1 + \overline{\mathbf{a_2}}X_2 \le \underline{\mathbf{a_3}} & \text{if } \overline{\mathbf{a_2}} < 0 \end{cases}$$

We can prove that the transformation is generic to any system of linear constraints, and that it correctly and tightly yields the certainty closure. The transformed constraint problem can be solved by using linear programming. The upper and lower bounds on the possible values for each variable X_i are found by solving two linear programs, with objective $\max X_i$ and $\min X_i$ respectively. We thus obtain the projection of the certainty closure onto the domain of each variable. This gives guaranteed bounds for each end-to-end traffic flow.

4 Discussion

We have introduced the certainty closure as a generic framework to reason about constraint problems with data uncertainty, focusing on incomplete and erroneous data. The framework is complementary to stochastic data models (e.g. [6]). Our current research looks at finding uncertain constraint classes and their corresponding tractable transformation operators.

Existing approaches to data uncertainty in CP come from a different perspective. Most authors consider uncertainty due to a dynamic and unpredictable environment and thus seek robust solutions that hold under the most possible realisations of the data. We are not aware of any work in CP aimed at building correct solution sets in the presence of erroneous data, the closest being work on interval CSPs in the presence of universally quantified variables [2]. Our work in concept is more closely related to that in operational research [1] and control theory [3].

Acknowledgements. This work was partially supported by the EPSRC under grant GR/N64373/01. The authors thank the anonymous reviewers for their suggestions.

References

1. Ben-Tal, A. and Nemirovski, A. Robust convex optimization. *Mathematics of Operations Research*, **23** (1998).
2. Benhamou, F. and Goualard, F. Universally quantified interval constraints. In: *CP-2000*.
3. Elishakoff, I. Convex modeling — a generalization of interval analysis for nonprobabilistic treatment of uncertainty. In: *Proc. of the Intl. Workshop on Applications of Interval Computations (APIC'95)*, El Paso, TX, 76–79 (1995).
4. Gervet, C., Caseau, Y. and Montaut, D. On refining ill-defined constraint problems: A case study in iterative prototyping. In: *Proc. of PACLP'99*, London, 255–275 (1999).
5. Gervet, C. and Rodošek, R. RiskWise-2 problem definition. IC–Parc Internal Report (2000).
6. Walsh, T. Stochastic constraint programming. In: *Proc. of AAAI'01 Fall Symposium on Using Uncertainty within Computation*, Cape Cod, MA, 129–135 (2001).