# Marginal Consistency: Unifying Constraint Propagation on Commutative Semirings

Tomáš Werner

Center for Machine Perception
Dept. of Cybernetics, Faculty of Electrical Engineering, Czech Technical University
Karlovo náměstí 13, 12135 Praha, Czech Republic

**Abstract.** We generalise the linear programming relaxation approach to Weighted CSP by Schlesinger and the max-sum diffusion algorithm by Koval and Kovalevsky twice: from Weighted CSP to Semiring CSP, and from binary networks to networks of arbitrary arity. This generalisation reveals a deep property of constraint networks on commutative semirings: by locally changing constraint values, any network can be transformed into an equivalent form in which all corresponding marginals of each constraint pair coincide. We call this state marginal consistency. It corresponds to a local minimum of an upper bound on the Semiring CSP. We further show that a hierarchy of gradually tighter bounds is obtained by adding neutral constraints with higher arity. We argue that marginal consistency is a fundamental concept to unify local consistency techniques in constraint networks on commutative semirings.

## 1 Introduction

Given[1] a set of variables, the *Weighted Constraint Satisfaction Problem* (WCSP) [1] is defined as maximising a sum of given functions of subsets of the variables. One of the approaches to WCSP is the linear programming (LP) relaxation of its integer LP formulation, first proposed by Schlesinger [2]. The LP dual of this relaxation can be interpreted as minimising an upper bound on the WCSP objective function by equivalent transformations, which are changes of constraint weights that preserve the objective function. Schlesinger's co-workers Kovalevsky and Koval [3] formulated a simple network algorithm, the *max-sum diffusion*, to minimise the upper bound. The works [2, 3] have been recently surveyed in [4].

The LP relaxation approach to WCSP has been independently proposed by others [5–9] and has been shown very successful on a number of sparse large-scale binary WCSP instances originating from practical computer vision problems [10].

In our previous works, we proposed two generalisations of the LP relaxation approach [2] and the max-sum diffusion. First, observing that the max-sum diffusion resembles algorithms to enforce arc consistency in CSP, in [11, 12] we generalised it to Semirings CSPs [13]. We remark that enforcing *soft arc consistency* as proposed in [14, 9] is related to but different from the max-sum diffusion (though it yields the same upper bound); it is closer to the *augmenting DAG*

[1] This research has been supported by the European Commission grant 215078 (DIPLECS) and the Czech government grant MSM6840770038.

*algorithm* by Koval and Schlesinger [15, 4, 16]. Second, in [17] we generalised the max-sum diffusion, originally formulated for binary problems [3], to problems of arbitrary arity. In [17] we also showed that a hierarchy of gradually tighter LP relaxations of WCSP can be obtained simply by adding *zero constraints* of higher arity. Suitable zero constraints can be added dynamically, which in combinatorial optimisation terms can be understood as a cutting plane algorithm.

Here we apply the two generalisations simultaneously. This reveals a deep property of constraint networks on commutative semirings which was not fully apparent before: *by locally changing constraint values, any constraint network can be transformed to an equivalent form in which all corresponding marginals of each pair of constraints coincide.* We call this state *marginal consistency.* It corresponds to a local minimum of an upper bound on the true objective function. Marginal consistency can be enforced for any semiring under mild assumptions, including the max-min, max-sum, and sum-product semirings.

An important feature of the proposed framework is its simplicity: it hinges on only two concepts, *local equivalent transformations* and *marginal consistency*, which combined together yield *enforcing marginal consistency.* This simplicity suggests that marginal consistency is a fundamental concept to unify local consistencies in constraint networks on commutative semirings.

As our second contribution, we show that by adding *neutral constraints* of higher arity to the network, a hierarchy of gradually tighter levels of consistency is obtained, corresponding to gradually tighter upper bounds on the problem.

For the crisp CSP, marginal consistency corresponds to *pairwise consistency.* Several other crisp local consistencies (generalised arc consistency, $k$-consistency) are obtained by straightforward modifications of marginal consistency.

## 2 Notation and basic definitions

In the sequel, $2^V$ resp. $\binom{V}{k}$ denotes the set of all resp. of $k$-element subsets of set $V$. $\mathbb{R}$ denotes the reals, $\mathbb{R}_+$ the non-negative reals, and $\mathbb{R}_{++}$ the positive reals. Equality of functions is denoted by $\equiv$. Strict (non-strict) inclusion is $\subset$ ($\subseteq$).

The notation we are going to introduce in this section is not typical in the constraints literature but is common in machine learning and statistical mechanics (e.g. [18]). Unfortunately, it is often used without a rigorous definition. To prevent confusion, we will develop this notation in more detail than is usual.

Let $V$ be a finite, totally ordered set of **variables**. To emphasise the variable ordering, when defining a subset of $V$ by enumerating its elements we will use round brackets ( ) instead of curly brackets { }. Each variable $v \in V$ is assigned a finite set $X_v$, the **domain** of the variable. An element of $X_v$ is a **state** of variable $v$ and is denoted by $x_v$. The **joint domain** of variables $A \subseteq V$ is the Cartesian product $X_A = \bigtimes_{v\in A} X_v$, where the order of the factors in the product is given by the order on $V$. A tuple $x_A \in X_A$ is a **joint state**[2] of variables $A$.

*Example 1.* Let $V = (1, 2, 3, 4)$ and $X_1 = X_2 = X_3 = X_4 = \{\mathsf{a}, \mathsf{b}\}$. A joint state $x_{134} = (x_1, x_3, x_4) \in X_{134}$ of variables $A = (1, 3, 4) \subseteq V$ is e.g. $(\mathsf{a}, \mathsf{a}, \mathsf{b})$. □

[2] In other works, a joint state is called a (partial) *instantiation* or simply a *tuple*.

For $B \subseteq A$, whenever symbols $x_A$ and $x_B$ appear in the scope of a single logical expression they do not denote independent joint states but $x_B$ denotes the *restriction* of $x_A$ to variables $B$. This 'implicit restriction' results in simpler expressions than if the restriction were denoted explicitly, e.g. by $x_A|_B$.

*Example 2.* Let $A = (1,2,3,4)$ and $B = (1,2)$. The set $\{ x_A \mid x_{A\setminus B} \in X_{A\setminus B} \}$ denotes the set $\{ (x_1, x_2, x_3, x_4) \mid x_3 \in X_3,\ x_4 \in X_4 \}$ for some fixed $x_1 \in X_1$ and $x_2 \in X_2$. The sentence "let $y_A$ be such that $y_B = x_B$" means "let quadruple $(y_1, y_2, y_3, y_4)$ be such that $y_1 = x_1$ and $y_2 = x_2$". $\square$

Let $S$ be a set of **weights**. A **constraint** with **scope** $A \subseteq V$ is a function $f_A\colon X_A \to S$. The **arity** of the constraint is the size of its scope, $|A|$.

Let $E \subseteq 2^V$ be a set of subsets of $V$, i.e., a hypergraph. Each hyperedge $A \in E$ is assigned a constraint $f_A\colon X_A \to S$. Note, this notation assumes that no two constraint have the same scope. All the constraints together can be understood as a single mapping $f\colon T(E) \to S$ with $T(E) = \{ (A, x_A) \mid A \in E,\ x_A \in X_A \}$, where the image of $(A, x_A) \in T(E)$ under $f$ is denoted by $f_A(x_A)$.

A **constraint network** is a tuple $(V, \{X_v\}, E, f)$. However, since the variable set $V$ and the domains $X_v$ will be the same in the whole paper, we will refer to a constraint network only as 'network $f$ with structure $E$' or, when also $E$ is clear from context, as 'network $f$'. The **arity of the network** is $\max_{A\in E} |A|$.

Let $\oplus$ and $\odot$ be binary operations closed on $S$. Our aim is to calculate or approximate the expression

$$\bigoplus_{x_V \in X_V} \bigodot_{A\in E} f_A(x_A) \tag{1}$$

For this expression to be uniquely defined, both operations are assumed to be associative and commutative, thus $(S, \oplus)$ and $(S, \odot)$ are *commutative semigroups*. For later purposes, we further assume that $\odot$ distributes over $\oplus$, thus $(S, \oplus, \odot)$ is a *commutative semiring*. Then, (1) is known as the *Semiring CSP* [13, 1].

*Example 3.* Let $V = (1,2,3,4)$ and $E = \{(2,3,4), (1,2), (3,4), (3)\}$. Expression (1) reads $\bigoplus_{x_1,x_2,x_3,x_4} [f_{234}(x_2,x_3,x_4) \odot f_{12}(x_1,x_2) \odot f_{34}(x_3,x_4) \odot f_3(x_3)]$ where e.g. $f_{234}$ is a ternary function of variables $2,3,4$. $\square$

We denote $a^{\oplus n} = a \oplus \cdots \oplus a$ ($n$-times). The solution of equation $a^{\oplus n} = b$ (if it exists) is denoted $a = b^{\oplus 1/n}$. Similarly we define $a^{\odot n}$ and $a^{\odot 1/n}$.

A commutative semiring may have a zero element $\mathbf{0}$ satisfying $a \oplus \mathbf{0} = a$ and $a \odot \mathbf{0} = \mathbf{0}$ and/or a unit element $\mathbf{1}$ satisfying $a \odot \mathbf{1} = a$.

## 3 Equivalent networks

Different networks $f$ can give rise to the same objective function of (1), i.e., changing the weights of certain constraints in certain ways preserves the objective function $\bigodot_{A\in E} f_A(x_A)$. This is captured by the concept of equivalent networks.

**Definition 1.** *Constraint networks $f, f'$: $T(E) \to S$ are* **equivalent** *iff they have the same variables $V$, domains $X_v$ and structure $E$, and*

$$\forall x_V \in X_V: \quad \bigodot_{A \in E} f_A(x_A) = \bigodot_{A \in E} f'_A(x_A)$$

*A change of $f$ to an equivalent network is an* **equivalent transformation**.

This equivalence factorises the set of all constraint networks to *equivalence classes*, each class giving rise to one objective function.

Some equivalent transformations are *local* in the sense sense that they affect only a small part of the network. More precisely, any such transformation changes some of the weights of a single pair of constraints, $f_A$ and $f_B$.

**Definition 2.** *A* **pencil** *is a triplet $(A, B, x_B)$ such that $A \in E$, $B \in E$, $B \subset A$, and $x_B \in X_B$.*

**Definition 3.** *An equivalent transformation is* **local** *if there exists a pencil $(A, B, x_B)$ such that the transformation changes only the weights $f_C(x_C)$ for $(C, x_C) \in \{(B, x_B)\} \cup \{\, (A, x_A) \mid x_{A \setminus B} \in X_{A \setminus B} \,\}$ and it changes them in such a way that the expression $f_A(x_A) \odot f_B(x_B)$ is preserved for all $x_{A \setminus B} \in X_{A \setminus B}$.*

*Example 4.* Let $A = (1,2,3)$, $B = (1,3)$, $X_v = \{\mathsf{a},\mathsf{b},\mathsf{c}\}$ for $v \in V$, $x_B = (x_1, x_3) = (\mathsf{a},\mathsf{b})$. A local equivalent transformation on pencil $(A, B, x_B)$ is any change of the weights $f_{123}(\mathsf{a},\mathsf{a},\mathsf{b})$, $f_{123}(\mathsf{a},\mathsf{b},\mathsf{b})$, $f_{123}(\mathsf{a},\mathsf{c},\mathsf{b})$, $f_{13}(\mathsf{a},\mathsf{b})$ that preserves the expressions $f_{123}(\mathsf{a},\mathsf{a},\mathsf{b}) \odot f_{13}(\mathsf{a},\mathsf{b})$, $f_{123}(\mathsf{a},\mathsf{b},\mathsf{b}) \odot f_{13}(\mathsf{a},\mathsf{b})$, $f_{123}(\mathsf{a},\mathsf{c},\mathsf{b}) \odot f_{13}(\mathsf{a},\mathsf{b})$. □

*Remark 1.* Definition 3 is not the most general form of local equivalent transformations, e.g., it allows no transformation between $f_A$ and $f_B$ if $B \not\subset A$. The most general form is as follows: for any quadruple $(A, B, C, x_C)$ with $A \in E$, $B \in E$, $C \subseteq A \cap B$ (where not necessarily $C \in E$), change the weights $f_D(x_D)$ for $(D, x_D) \in \{\, (A, x_A) \mid x_{A \setminus C} \in X_{A \setminus C} \,\} \cup \{\, (B, x_B) \mid x_{B \setminus C} \in X_{B \setminus C} \,\}$ such that the expression $f_A(x_A) \odot f_B(x_B)$ is preserved for all $x_{(A \cup B) \setminus C}$. Definition 3 is obtained from this for $B \subset A$ and $C = A \cap B = B$. □

Composing local equivalent transformations yields an equivalent transformation. The converse is not true in general: there may be two equivalent networks that are not connected by a sequence of local equivalent transformations.

*Example 5.* Let $(S, \odot) = (\{0,1\}, \min)$. Let $f$: $T(E) \to \{0,1\}$ represent an unsatisfiable crisp CSP. Then $f$ is equivalent to the zero network $f \equiv 0$ but this in general is not achieved by a sequence of local equivalent transformations. Moreover, testing if $f$ is equivalent to the zero network $f \equiv 0$ is NP-complete. □

### 3.1 Equivalent transformations for a group

An important special case is when $(S, \odot)$ is a group, i.e., it has division, denoted by $\oslash$. Then, all possible local equivalent transformations on a pencil $(A, B, x_B)$ can be parameterised by a single weight $\varphi_{A,B}(x_B)$ as follows:

$$f'_A(x_A) = f_A(x_A) \odot \varphi_{A,B}(x_B), \qquad \forall x_{A \setminus B} \in X_{A \setminus B} \tag{2a}$$

$$f'_B(x_B) = f_B(x_B) \oslash \varphi_{A,B}(x_B) \tag{2b}$$

because $[f_A(x_A) \odot \varphi_{A,B}(x_B)] \odot [f_B(x_B) \oslash \varphi_{A,B}(x_B)] = f_A(x_A) \odot f_B(x_B)$.

Let a weight $\varphi_{A,B}(x_B)$ be assigned to every pencil $(A, B, x_B)$. All these weights together can be understood as a single mapping $\varphi$ from the set of all pencils to $S$. Let $f^\varphi$ denote the network obtained by composing local equivalent transformations (2) on all the pencils (in any order). It is given by

$$f^\varphi_A(x_A) = f_A(x_A) \odot \bigodot_{B|B\in E, B\subset A} \varphi_{A,B}(x_B) \oslash \bigodot_{B|B\in E, B\supset A} \varphi_{B,A}(x_A) \tag{3}$$

For binary networks, it was proved that any equivalent transformation can be obtained as a composition of local equivalent transformations [19, 4]. We conjecture, currently without proof, that this extends to networks of any arity.

## 4 Marginal consistency

**Definition 4.** *Given a function $f_A\colon X_A \to S$ and a set $B \subseteq A$, we define the function $f_A|_B\colon X_B \to S$ by*

$$f_A|_B(x_B) = \bigoplus_{x_{A\setminus B}} f_A(x_A) \tag{4}$$

*We call $f_A|_B(x_B) \in S$ the* **marginal**[3] *of $f_A$ associated with $(B, x_B) \in T(E)$.*

*Example 6.* Let $A = (1, 2, 3, 4)$ and $B = (1, 3)$. The marginal of $f_A$ associated with $(B, x_B)$ is given by $f_{1234}|_{13}(x_1, x_3) = \bigoplus_{x_2,x_4} f_{1234}(x_1, x_2, x_3, x_4)$. □

**Definition 5.** *A constraint network $f\colon T(E) \to S$ is* **marginal consistent** *iff $f_A|_{A\cap B}(x_{A\cap B}) = f_B|_{A\cap B}(x_{A\cap B})$ for all $A \in E$, $B \in E$, and $x_{A\cap B} \in X_{A\cap B}$.*

*Example 7.* A network $f$ with structure $E = \{(1), (1, 2), (2, 3)\}$ is marginal consistent iff $f_1 \equiv f_{12}|_1$, $f_1|_\emptyset \equiv f_{23}|_\emptyset$, and $f_{12}|_2 \equiv f_{23}|_2$. Note that $f_1|_1 \equiv f_1$. □

**Definition 6.** *A pencil $(A, B, x_B)$ is* **marginal consistent** *iff $f_A|_B(x_B) = f_B(x_B)$.*

One would expect that making all the pencils marginal consistent makes the network marginal consistent. However, this holds only under an additional condition on hypergraph $E$, e.g., if $E$ is closed to intersection ($A, B \in E$ implies $A \cap B \in E$). As we will show in §8, this does not cause much loss of generality.

*Example 8.* The network in Example 7 has pencils $\{\, ((1, 2), (1), x_1) \mid x_1 \in X_1 \,\}$. Making these pencils marginal consistent covers only the equality $f_1 \equiv f_{12}|_1$.

Closing $E$ by intersection yields $E = \{\emptyset, (1), (2), (1, 2), (2, 3)\}$. Now, marginal consistency of all the pencils in the network implies all the three equalities. □

*Remark 2.* The requirement that $E$ is closed to intersection could be avoided by defining marginal consistency not for a pencil $(A, B, x_B)$ but for a triplet $(A, B, x_{A\cap B})$ as follows: a triplet $(A, B, x_{A\cap B})$ with $A, B \in E$ is marginal consistent iff $f_A|_{A\cap B}(x_{A\cap B}) = f_B|_{A\cap B}(x_{A\cap B})$. However, Definition 6 is more convenient because there are fewer pairs $(A, B)$ with $B \subset A$ in $E$ than all the pairs.

[3] In the constraints literature, *marginals* are more often called *projections*.

In fact, marginal consistency of a part of a network could be defined even in a more general way: a quadruple $(A, B, C, x_C)$ with $A \in E$, $B \in E$, $C \subseteq A \cap B$ is marginal consistent iff $f_A|_C(x_C) = f_B|_C(x_C)$. Compare to Remark 1. $\square$

## 5 Enforcing marginal consistency

### 5.1 Enforcing marginal consistency of a pencil

Enforcing marginal consistency of a pencil is a local equivalent transformation on the pencil that makes it marginal consistent. Denoting for brevity

$$f_A(x_A) = a_i, \qquad i = x_{A\setminus B} \in X_{A\setminus B} \tag{5a}$$
$$f_B(x_B) = b \tag{5b}$$

enforcing marginal consistency of pencil $(A, B, x_B)$ means, by Definitions 3 and 6, that we are given weights $a_i, b$ and look for weights $a'_i, b'$ satisfying

$$a'_i \odot b' = a_i \odot b \qquad \forall\, i \tag{6a}$$
$$\bigoplus_i a'_i = b' \tag{6b}$$

For the enforcing to be possible, system (6) must be solvable for any $a_i, b$. We in addition require (although this may be arguable) that the solution be *unique*.

While solving (6) in a concrete semiring is typically easy, solving it in the abstract semiring is difficult. That would mean, characterise the semirings in which (6) is uniquely solvable and give an algorithm to find the solution. We have not accomplished this. Yet we observed that in almost all semirings in which (6) is uniquely solvable, the solution can be written in a closed form as

$$b' = \Big(b \odot \bigoplus_i a_i\Big)^{\odot 1/2} \tag{7a}$$
$$a'_i = a_i \odot b \oslash b' \tag{7b}$$

where the operations $a^{\odot 1/2}$ and $\oslash$ are defined below.

In (7b), we assume that the semigroup $(S, \odot)$ is *uniquely 2-divisible*, i.e., the equation $a \odot a = b$ has a unique solution for every $b$, denoted by $a = b^{\odot 1/2}$. Then (7a) is verified by summing (6a) over $i$ and substituting $\bigoplus_i a'_i$ from (6b) to (6a). This shows that for (6) to have a unique solution, it is necessary that the semigroup $(S, \odot)$ is uniquely 2-divisible.

Formula (7b) is obvious if $(S, \odot)$ is a group. If $(S, \odot)$ is only a semigroup, we do not have division. However, it is a classical result that in some semigroups, a weaker form of division can be introduced. A semigroup $(S, \odot)$ is an *inverse semigroup* (e.g. [20]) iff for every $a$ there is a unique $b$ such that

$$a \odot b \odot a = a, \quad b \odot a \odot b = b$$

We call $a$ and $b$ the *inverse* of each other and define division by $c \oslash a = c \odot b$. It has been shown [21, §5] that so defined division is useful for local computations in a constraint network with a tractable structure (join trees or hypertrees). With this division, (7b) uniquely solves (6a) in many semirings.

*Example 9.* Examples of semirings in which system (6) is uniquely solvable are $(R_{++}, +, \times)$, $(\mathbb{R}_+, +, \times)$, $(\mathbb{R}, \max, +)$, $(\mathbb{R} \cup \{-\infty\}, \max, +)$, any distributive lattice, and the semiring defined in [13, §2.4.5]. For all of them, $(S, \odot)$ is a uniquely 2-divisible inverse semigroup. In semiring $(\mathbb{R}_+, \min, +)$ system (6) is uniquely solvable despite that the semigroup $(\mathbb{R}_+, +)$ is not inverse. □

*Example 10.* Examples of semirings in which system (6) is not uniquely solvable are $(\mathbb{N}, \max, +)$, $([k, 0], \max, +^k)$ where $k < 0$ and $+^k$ is the truncated addition defined by $a +^k b = \max\{a + b, k\}$, and $(\mathbb{R}, +, \times)$. □

Switching back from our temporary notation by plugging (5) into (7) yields the formula to enforce marginal consistency of pencil $(A, B, x_B)$:

$$f'_B(x_B) = \left[f_B(x_B) \odot f_A|_B(x_B)\right]^{\odot 1/2} \tag{8a}$$

$$f'_A(x_A) = f_A(x_A) \odot f_B(x_B) \oslash f'_B(x_B), \qquad \forall x_{A \setminus B} \tag{8b}$$

Formula (8) changes values of the constraints, which has the drawback that if $f_A$ was originally represented intensionally (as a black box function), we need to switch to extensional representation (storing all its weights explicitly in memory). This can be avoided by using the parameterised form (3) of equivalent transformations, in which we need to store only the variables $\varphi_{A,B}(x_B)$ but the constraints themselves are not changed. This may mean an important saving of memory when $|X_A| \gg |X_B|$. By (8) and (2), variable $\varphi_{A,B}(x_B)$ is updated as

$$\varphi_{A,B}(x_B) \leftarrow \varphi_{A,B}(x_B) \odot \left[f^\varphi_B(x_B) \oslash f^\varphi_A|_B(x_B)\right]^{\odot 1/2} \tag{9}$$

### 5.2 Enforcing marginal consistency of a network

Having solved enforcing marginal consistency of a single pencil $(A, B, x_B)$, let us now turn to enforcing it for the whole network.

**Observation 1** *Let the semiring $(S, \oplus, \odot)$ be such that system (6) is uniquely solvable. Enforcing marginal consistency repetitively for different pencils converges to a state when the whole network is marginal consistent. The pencils can be visited in any order such that each has a non-zero probability to be visited.*

Currently, we have neither a prove of this observation (convergence has not been proved even for semiring $(\mathbb{R}, \max, +)$, see §7.2) nor a counter-example.

Using the parameterised form (9), the algorithm can thus look like this:

**loop**
  **for** $(A, B, x_B)$ such that $(A, B) \in J$, $x_B \in X_B$ **do**
    $\varphi_{A,B}(x_B) \leftarrow \varphi_{A,B}(x_B) \odot \left[f^\varphi_B(x_B) \oslash f^\varphi_A|_B(x_B)\right]^{\odot 1/2}$
  **end for**
**end loop**

We will refer to the algorithm as the $(\oplus, \odot)$ **marginal consistency algorithm** and to any of its fixed points as a $(\oplus, \odot)$ **marginal consistency closure**. Depending on the semiring, convergence may be achieved in a finite or an infinite number of iterations and the closure may be unique or non-unique.

## 6 Upper bound

A relation $\leq$ can be defined on semigroup $(S, \oplus)$ by

$$a \leq b \quad \Longleftrightarrow \quad (a = b) \text{ or } (\exists c \in S\colon\; a \oplus c = b) \tag{10}$$

Note, the condition $a = b$ in (10) is redundant if the semiring has a zero element. The relation $\leq$ is reflexive and transitive, hence a preorder. It is known as *Green's preorder* [20, §II.1]. It follows [21, §2] that the semiring operations are monotonic with respect to $\leq$, i.e., $a \leq b$ implies $a \oplus c \leq b \oplus c$ and $a \odot c \leq b \odot c$. Often the semigroup $(S, \oplus)$ is such that the relation $\leq$ is antisymmetric, hence a partial or a total order. Then both $(S, \oplus)$ and $(S, \odot)$ become *ordered semigroups*.

In this section, we assume that the semiring $(S, \oplus, \odot)$ satisfies two additional properties we did not need before. First, we assume that the equation $a^{\oplus n} = b$ has a solution, $a = b^{\oplus 1/n}$, for every $b$, i.e., the semigroup $(S, \oplus)$ is divisible. Second, we assume the following inequality holds for any $a_1, \dots, a_n \in S$:

$$\bigodot_{i=1}^{n} a_i \leq \Big[ \Big( \bigoplus_{i=1}^{n} a_i \Big)^{\oplus 1/n} \Big]^{\odot n} \tag{11}$$

In semiring $(\mathbb{R}_+, +, \times)$, (11) is the well-known arithmetic-geometric average inequality $\big(\prod_i a_i\big)^{1/n} \leq \frac{1}{n}\big(\sum_i a_i\big)$. We do not know whether (11) can be derived from more elementary properties of the semiring, hence we assume it as an independent axiom. Note, for all the semirings in Example 9, the preorder $\leq$ is a partial or total order and the semiring satisfies the above two properties.

**Theorem 1.** *For any $f\colon T(E) \to S$, we have*

$$\bigoplus_{x_V} \bigodot_{A \in E} f_A(x_A) \leq \bigodot_{A \in E} \bigoplus_{x_A} f_A(x_A) \leq \Big[ \Big[ \bigoplus_{A \in E} \bigoplus_{x_A} f_A(x_A) \Big]^{\oplus 1/|E|} \Big]^{\odot |E|} \tag{12}$$

*If $f$ is marginal consistent then the middle and right-hand expressions equal.*

*Proof.* Using distributivity, multiply out the factors in the middle expression. This yields the terms present in the left-hand expression and some additional terms. The first inequality in (12) follows from (10), where $c$ are the extra terms.

The second inequality in (12) follows from (11).

At marginal consistency, the weights $\bigoplus_{x_A} f_A(x_A) = f_A|_\emptyset$ for $A \in E$ are all equal. The equality claim trivially follows. □

Theorem 1 gives two upper bounds on problem (1). We will refer to them as **the first** resp. **the second** $(\oplus, \odot)$ **upper bound**. At marginal consistency, the two bounds coincide, therefore we will speak only about a single bound.

The bounds can be improved by equivalent transformations. Ideally, we would like to find the least upper bound in the whole equivalence class. If this is not possible, we at least want to decrease it by local equivalent transformations. This is in a simple way related to marginal consistency.

**Theorem 2.** *Enforcing marginal consistency of any pencil does not increase the value of expression* $\bigoplus_{A\in E}\bigoplus_{x_A} f_A(x_A)$.

*Proof.* Let us use notation (5) and denote $a=\bigoplus_i a_i$. It suffices to prove that enforcing marginal consistency does not increase the expression $a\oplus b$. By (7a) and (6b), after enforcing marginal consistency we have $a'=b'=(a\odot b)^{\odot 1/2}$. Thus $a'\oplus b'=[(a\odot b)^{\odot 1/2}]^{\oplus 2}$. By (11), we have $a'\oplus b'\leq a\oplus b$. □

The theorem only says that enforcing marginal consistency of a pencil does not make the upper bound worse, it does not ensure it strictly improves. However, a stronger statement relating the upper bound and marginal consistency holds: *if the network is marginal consistent then the upper bound cannot be improved by any local equivalent transformation.*

### 6.1 The case of idempotent semiring addition and total order

If $\leq$ is a total order and $\oplus$ is idempotent, $\oplus$ necessarily is the maximum with respect to $\leq$ and problem (1) is known as the *Valued CSP* [22, 13]. In this case, the networks for which the upper bound is tight can be characterised in a simple way, by satisfiability of a CSP. We will formulate this only for the first bound in (12), the second bound is analogical. We omit the proof which is rather simple.

**Definition 7.** *A joint state* $(A,x_A)\in T(E)$ *is* **active** *iff* $f_A(x_A)=\max_{y_A} f_A(y_A)$. *The CSP* $\bar f\colon T(E)\to\{0,1\}$ *is defined such that* $\bar f_A(x_A)=1$ *iff* $(A,x_A)$ *is active.*

**Theorem 3.** *Let the operation* $\oplus$ *be idempotent and the preorder* $\leq$ *be a total order. The first inequality in (12) holds with equality iff* $\bar f$ *is satisfiable.*

For semiring $(\{0,1\},\max,\min)$, Theorem 3 holds trivially because $\bar f=f$. However, the theorem is important for semiring $([0,1],\max,\min)$ and mainly for $(\mathbb{R},\max,+)$, where $\bar f$ is crucial for reasoning about optimality of the upper bound. Marginal consistency of $\bar f$ relates to marginal consistency of $f$ as follows.

**Theorem 4.** *For* $B\subset A$, *if* $f_A|_B(x_B)=f_B(x_B)$ *then* $\bar f_A|_B(x_B)=\bar f_B(x_B)$.

## 7 Enforcing marginal consistency for important semirings

In this section, we discuss the properties of the marginal consistency algorithm for important semirings in detail. The selection of the semirings is generic in the sense that it covers the case of $\oplus$ and $\odot$ both idempotent, of idempotent $\oplus$ and non-idempotent $\odot$, and of $\oplus$ and $\odot$ both non-idempotent. For each semiring, we discuss the following questions:

– Does the marginal consistency algorithm converge in a finite or infinite number of iterations?

– Is the marginal consistency closure unique?
– Does the marginal consistency algorithm finds the global minimum of the upper bound (12) in the class of equivalent networks?
– Is the upper bound evaluated at marginal consistency useful?

### 7.1 Distributive lattice

Here $(S, \oplus, \odot)$ is any distributive lattice $(S, \wedge, \vee)$, i.e., $\leq$ is a partial order and $\wedge$ ($\vee$) is the infimum (supremum) given by $\leq$. We have $a \oslash b = a \wedge b$ and $a^{\odot 1/n} = a$.

The marginal consistency algorithm converges in a finite number of iterations and the closure is unique [13, 23]. Since not every equivalent transformation can be obtained by composing local equivalent transformations, the algorithm does not find the global minimum of the upper bound in the equivalence class. This is expected because minimising the upper bound is NP-hard.

Special cases are semirings $([0, 1], \max, \min)$ (the *Fuzzy CSP*, i.e., the order $\leq$ is total) and $(\{0, 1\}, \max, \min)$ (the crisp CSP). Note what the upper bounds mean for $(\{0, 1\}, \max, \min)$: the first bound equals 1 if all the constraints are non-empty and 0 if at least one constraint is empty; the second bound equals 1 if at least one constraint is non-empty and 0 if all the constraints are empty.

### 7.2 The max-sum semiring

Here $(S, \oplus, \odot) = (\mathbb{R}, \max, +)$. We have $a \oslash b = a - b$ and $a^{\odot 1/n} = a/n$.

The marginal consistency algorithm is the *max-sum diffusion*, proposed for binary networks in [3] and generalised to networks of higher arity in [17]. It converges in an infinite number of iterations. The behaviour of the algorithm is highly non-trivial, e.g., its convergence was only conjectured and never proved. The convergence rate is rather slow, but still practical even for large instances.

Finding the least upper bound is tractable. As $(\mathbb{R}, +)$ is a group, parameterisation (3) covers the whole equivalence class. Then, minimising the first resp. second upper bound leads to the convex nonsmooth minimisation problem

$$\min_{\varphi} \sum_{A \in E} \max_{x_A} f_A^{\varphi}(x_A) = |E| \min_{\varphi} \max_{A \in E} \max_{x_A} f_A^{\varphi}(x_A) \tag{13}$$

which can be written as a linear program [2, 4, 17]. The LP dual of (13) reads

$$\max \left\{ \mu \cdot f \;\middle|\; \mu \geq 0,\;\; \forall A\colon \sum_{x_A} \mu_A(x_A) = 1,\;\; \forall B \subset A\colon \sum_{x_{A \setminus B}} \mu_A(x_A) = \mu_B(x_B) \right\}$$

where $\mu \cdot f$ denotes the scalar product of $\mu$ and $f$ understood as vectors. This is a linear programming relaxation of problem (1). Note that in the dual we meet marginal consistency again, but this time with respect to operation $+$.

In general, the marginal consistency closure is not unique and is not a solution of problem (13) [24, 8], see [4, Figures 5b,c,d] for examples. This is because the marginal consistency algorithm can get stuck in a point when the upper bound can be improved by changing no variable $\varphi_{A,B}(x_B)$ separately but only by changing several of them simultaneously. In other words, the upper bound can be improved by a sequence of local equivalent transformations but the marginal consistency algorithm does not find this sequence. This can be seen as a manifestation of the fact that coordinate descent in general does not find the global minimum of a convex nonsmooth function [25]. Despite that, the marginal consistency closures are useful upper bounds on (1), which are often tight for highly non-trivial instances [4, 17, 8, 10].

For semiring $(\mathbb{R} \cup \{-\infty\}, \max, +)$, we have $\mathbf{0} = -\infty$, $a \odot \mathbf{0} = a \oslash \mathbf{0} = \mathbf{0}$, and $\mathbf{0}^{\odot 1/2} = \mathbf{0}$. Finding the least upper bound becomes intractable because of the absorbing element $-\infty$, for the reason as in Example 5. Two stages can be discerned in the marginal consistency algorithm: after a finite number of iterations the set of joint states $(A, x_A)$ with $f_A(x_A) = -\infty$ stabilises, then the algorithm proceeds on the finite joint states as in semiring $(\mathbb{R}, \max, +)$.

If functions $f_A$ are supermodular for all $A \in E$, then at the marginal consistency closure the upper bound is tight, i.e., (12) holds with equality [17].

### 7.3 The sum-product semiring

Here $(S, \oplus, \odot) = (\mathbb{R}_{++}, +, \times)$. We have $a \oslash b = a/b$ and $a^{\odot 1/n} = a^{1/n}$.

Semiring $(\mathbb{R}_{++}, +, \times)$ is isomorphic (via logarithm) with semiring $(\mathbb{R}, \circledast, +)$ where $a \circledast b = \log(\mathrm{e}^a + \mathrm{e}^b)$. Since the function $\circledast$ is convex and smooth, minimising the first resp. second upper bound is a convex smooth minimisation task,

$$\min_{\varphi} \sum_{A \in E} \bigcircledast_{x_A} f_A^{\varphi}(x_A) = |E| \Big[ -\log |E| + \min_{\varphi} \bigcircledast_{A \in E} \bigcircledast_{x_A} f_A^{\varphi}(x_A) \Big] \tag{14}$$

which is a geometric programming task [26, §4.5]. The marginal consistency algorithm can be understood as a coordinate descent method to minimise the upper bound and it converges in an infinite number of iterations to a unique closure, which is the least the upper bound in the equivalence class [11, §6].

For semiring $(\mathbb{R}_+, +, \times)$, finding the least upper bound becomes intractable because of the absorbing element 0. The behaviour is similar as for semiring $(\mathbb{R} \cup \{-\infty\}, \max, +)$, discussed in §7.2.

In statistical mechanics, (1) is known as the *partition function*, very hard to approximate. Thus one can expect that the upper bound will be not very useful. Indeed, it is typically very loose, and it is tight only for uninterestingly simple instances. Though this may suggest that the $(+, \times)$ marginal consistency algorithm is useless, it has two interesting applications which we will now describe.

*Finding the optimal* $(\max, +)$ *marginal consistency closure.* Let the operation $\circledast_\tau$ be defined by $a \circledast_\tau b = [(\tau a) \circledast (\tau b)]/\tau$, for any $\tau > 0$. Then $(\mathbb{R}, \circledast_\tau, +)$ is a semiring. Furthermore, $\circledast_\tau$ for increasing $\tau$ approaches the ordinary maximum, i.e., $\lim_{\tau \to \infty} a \circledast_\tau b = \max\{a, b\}$.

Recall from §7.2 that the $(\max, +)$ marginal consistency closure is not unique and in general does not minimise the upper bound in the equivalence class. However, the sequence of the $(\circledast_\tau, \times)$ marginal consistency closures of $f$ for increasing $\tau$ converges to the *optimal* $(\max, +)$ marginal consistency closure.

This can be understood as an interior point algorithm to solve the LP (13). We remark that for various reasons, this algorithm to compute the optimal $(\max, +)$ closure does not seem very practical. See [11, §7.4] for more details.

*A necessary condition for satisfiability of a CSP.* Suppose that $f\colon T(E) \to \{0,1\}$. Such $f$ represents a crisp CSP and (1) equals the *number of solutions* of this CSP (this is known as #CSP problem). If the $(+, \times)$ upper bound is strictly less than 1 then inevitably this CSP is not satisfiable. This gives us a necessary condition for satisfiability of a CSP. Interestingly, this condition is *strictly stronger* than the condition obtained from $(\max, \times)$ marginal consistency closure.

To state this precisely, let us compare the $(+, \times)$ marginal consistency closure and the $(\max, \times)$ marginal consistency closure of a network $f\colon T(E) \to \{0,1\}$. If we run the $(\max, \times)$ marginal consistency algorithm on $f$, all the weights will stay in $\{0,1\}$ during the algorithm. This corresponds to enforcing a crisp local consistency. However, if we run the $(+, \times)$ marginal consistency algorithm, the weights will not stay in $\{0,1\}$, they will end up in $\mathbb{R}_+$. We claim that if the $(\max, \times)$ marginal consistency closure of $f$ is zero then also the $(+, \times)$ consistency closure of $f$ is zero (here, by 'zero' we mean the zero network $f \equiv 0$). This is obvious because the absorbing element 0 behaves the same way both in $(\max, \times)$ and $(+, \times)$ marginal consistency algorithms. However, there exist networks (e.g. [4, Figures 5b,c,d]) for which the $(\max, \times)$ closure is non-zero but the $(+, \times)$ closure is zero. See [11, §7.5] for more details.

## 8 Adding neutral constraints

In this section, we assume that operation $\oplus$ is idempotent.

If $f_A(x_A) = \mathbf{1}$ for all $x_A \in X_A$ (in short, $f_A \equiv \mathbf{1}$) where $\mathbf{1}$ is the identity element of the semiring, we call $f_A$ the *neutral constraint* because $\mathbf{1}$ is the neutral element for the constraint combination operator $\odot$, i.e., $a \odot \mathbf{1} = a$. Suppose that for some $A \notin E$, we add the neutral constraint $f_A \equiv \mathbf{1}$ to the network. This changes neither the objective function of (1) nor the upper bounds (12). However, this allows for new, previously impossible (local) equivalent transformations. Therefore, the new upper bound may be possible to decrease even if the old upper bound was not.

Recalling §4 and Example 8, one application is to add neutral constraint $f_{A\cap B} \equiv \mathbf{1}$ for every $A, B \in E$, to close $E$ by intersection. Then, marginal consistency of all the pencils will imply marginal consistency of the network.

More interestingly, we can add neutral constraints of arity higher than the arity of the network. This may seem impractical as we would have to store all its values in memory. However, this is avoided by using the parameterised form (9) of the update: adding a constraint $f_A \equiv \mathbf{1}$ with $A \notin E$, we newly need to store only the variables $\varphi_{A,B}(x_B)$ for all $B \in E \cap 2^A$. The only restriction is that calculating the marginals $f^\varphi_A|_B(x_B)$ in (9) has to be tractable. Since $f_A \equiv \mathbf{1}$ and there is no $B \in E$ with $B \supset A$, equivalent transformation (3) reads simply

$$f^\varphi_A(x_A) = \bigodot_{B \mid B \in E, B \subset A} \varphi_{A,B}(x_B) \tag{15}$$

Enforcing marginal consistency of pencil $(A, B, x_B)$ is tractable if problem (15) is tractable. Note, (15) is an instance of problem (1) with structure $E \cap 2^A$.

*Example 11.* We showed in [17] that for many non-trivial binary problems with a sparse structure on semiring $(\mathbb{R}, \max, +)$, after adding certain neutral constraints of arity 4 the upper bounds was tight much more often than before. □

Adding different sets of neutral constraints provides a hierarchy of gradually tighter upper bounds. The hierarchy can be made even finer because we need not impose marginal consistency for all the pencils in the network – if only a subset of pencils is visited in the marginal consistency algorithm, it will also converge. The top element of the hierarchy is obtained by adding the constraint $f_V \equiv \mathbf{1}$ and imposing $f_V|_A \equiv f_A$ for all $A \in E$, which yields the exact solution of problem (1). The hierarchy is partially ordered by inclusion on the sets of marginal consistent pencils.

We can add a set of neutral constraints before running the marginal consistency algorithm. This has a disadvantage that we may add many neutral constraints that did not help improve the bound. Alternatively, neutral constraints can be added one by one during the algorithm, such that a constraint is added only if it is sure to enable a strict improvement of the bound.

### 8.1 Relation to existing local consistencies for the crisp CSP

For semiring $(\{0, 1\}, \max, \min)$, problem (1) is the crisp CSP. Imposing marginal consistency of different sets of pencils, possibly after adding necessary neutral constraints, yields several existing crisp local consistencies [27] as follows:

- Marginal consistency of a network, as given by Definition 5, corresponds to *pairwise consistency* [28], [27, §5.4]. Importantly, this shows that marginal consistency can be seen as the semiring generalisation of pairwise consistency.
- If we enforce $f_A|_v \equiv f_v$ for all $A \in E$ and $v \in A$, we obtain *(generalised) arc consistency*. For that, we need that $\binom{V}{1} \subseteq E$, which is achieved by adding necessary unary neutral constraints.
- Marginal consistency of a pencil $(A, B, x_B)$ has the following meaning: joint state $x_B$ satisfies constraint $f_B$ iff it can be extended to a joint state $x_A$ satisfying constraint $f_A$. It follows that if $\binom{V}{k-1} \cup \binom{V}{k} \subseteq E$ and the network is marginal consistent, then the network is *$k$-consistent*. If $E$ is smaller than required it is extended by adding neutral constraints. Similarly, we obtain *strong $k$-consistency* and *$(i, j)$-consistency*.

### 8.2 Idempotent *versus* non-idempotent semiring multiplication

As adding neutral constraints is concerned, there is a qualitative difference between semirings $([0, 1], \max, \min)$ and $(\mathbb{R}, \max, +)$. Suppose a constraint $f_A \equiv \mathbf{1}$ with $A \notin E$ has been added to a network and then marginal consistency enforced. This changes some weights of both the original network and the constraint $f_A$, which is now a part of the network. The difference is in the fact that:

– In semiring $([0,1], \max, \min)$, we can now *remove* $f_A$ from the network because this leaves the network equivalent with the original network (we leave this claim unproved, referring to well-known properties of the Fuzzy CSP). Thus, $f_A$ can be in fact added only temporarily, which results in changing (in fact, decreasing) some weights of the original network but does not increase the number of constraints in the network.
– In semiring $(\mathbb{R}, \max, +)$, we cannot remove $f_A$ because this would yield an non-equivalent network. Thus, adding a neutral constraint and enforcing marginal consistency *does* increase the number of constraints in the network.

The difference can be explained still in other words. Adding a neutral constraint makes new equivalent networks reachable by local equivalent transformations. Then:

– In semiring $([0,1], \max, \min)$, some equivalent transformations cannot be composed from local ones. Adding a neutral constraint makes some new equivalent transformations possible to compose from local ones. Thus, adding a neutral constraint, enforcing marginal consistency and removing the constraint can be seen as a non-local equivalent transformation on the original network.
– In semiring $(\mathbb{R}, \max, +)$, the whole equivalence class is reachable by local equivalent transformations and thus the only way how to make new equivalent transformations possible is to add a new constraint.

## References


1. Meseguer, P., Rossi, F., Schiex, T.: Soft constraints. In: Handbook of Constraint Programming. Elsevier (2006)
2. Shlezinger, M.I.: Syntactic analysis of two-dimensional visual signals in noisy conditions. Cybernetics and Systems Analysis **12**(4) (1976) 612–628 Translated from Russian original: M.I.Schlesinger, Sintaksicheskiy analiz dvumernykh zritelnikh signalov v usloviyakh pomekh, Kibernetika, No. 4, pp. 113-130, 1976.
3. Kovalevsky, V.A., Koval, V.K.: A diffusion algorithm for decreasing energy of max-sum labeling problem. Glushkov Institute of Cybernetics, Kiev, USSR. Unpublished (approx. 1975)
4. Werner, T.: A linear programming approach to max-sum problem: A review. IEEE Trans. Pattern Analysis and Machine Intelligence **29**(7) (2007) 1165–1179
5. Koster, A., van Hoesel, C.P.M., Kolen, A.W.J.: The partial constraint satisfaction problem: Facets and lifting theorems. Operations Research Letters **23**(3–5) (1998) 89–97
6. Chekuri, C., Khanna, S., Naor, J., Zosin, L.: Approximation algorithms for the metric labeling problem via a new linear programming formulation. In: Symposium on Discrete Algorithms. (2001) 109–118
7. Wainwright, M., Jaakkola, T., Willsky, A.: MAP estimation via agreement on (hyper)trees: message passing and linear programming approaches. IEEE Trans. Information Theory **51**(11) (2005) 3697–3717
8. Kolmogorov, V.: Convergent tree-reweighted message passing for energy minimization. IEEE Trans. Pattern Analysis and Machine Intelligence **28**(10) (2006) 1568–1583

9. Cooper, M., de Givry, S., Schiex, T.: Optimal soft arc consistency. In: Proc. IJCAI, Hyderabad, India. (2007)
10. Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwal, A., Tappen, M., Rother, C.: A comparative study of energy minimization methods for Markov random fields. In: Eur. Conf. Computer Vision (ECCV). (2006) II: 16–29
11. Werner, T., Shekhovtsov, A.: Unified framework for semiring-based arc consistency and relaxation labeling. In: 12th Computer Vision Winter Workshop, St. Lambrecht, Austria, Graz University of Technology (2007) 27–34 ISSN 3902465603.
12. Werner, T.: What is decreased by the max-sum arc consistency algorithm? In: Intl. Conf. on Machine Learning, Oregon, USA. (2007)
13. Bistarelli, S., Montanari, U., Rossi, F., Schiex, T., Verfaillie, G., Fargier, H.: Semiring-based CSPs and valued CSPs: Frameworks, properties,and comparison. Constraints **4**(3) (1999) 199–240
14. Schiex, T.: Arc consistency for soft constraints. In: Principles and Practice of Constraint Programming. (2000) 411–424
15. Koval, V.K., Schlesinger, M.I.: Two-dimensional programming in image analysis problems. USSR Academy of Science, Automatics and Telemechanics **8** (1976) 149–168 In Russian.
16. Werner, T.: A linear programming approach to max-sum problem: A review. Technical Report CTU–CMP–2005–25, Center for Machine Perception, Czech Technical University (2005)
17. Werner, T.: High-arity interactions, polyhedral relaxations, and cutting plane algorithm for soft constraint optimisation (MAP-MRF). In: Computer Vision and Pattern Recognition (CVPR) Conference, Anchorage, USA. (2008)
18. Yedidia, J.S., Freeman, W.T., Weiss, Y.: Constructing free-energy approximations and generalized belief propagation algorithms. IEEE Trans. Information Theory **51**(7) (2005) 2282–2312
19. Schlesinger, M.I.: Lectures on labeling problems, Kiev, Prague, Dresden (1996-2006) Unpublished.
20. Grillet, P.: Semigroups: an Introduction to the Structure Theory. Marcel Dekker, New York (1995)
21. Kohlas, J., Wilson, N.: Exact and approximate local computation in semiring induced valuation algebras. Technical Report 06-06, Department of Informatics, University of Fribourg (2006)
22. Schiex, T., Fargier, H., Verfaillie, G.: Valued constraint satisfaction problems: Hard and easy problems. In: Intl. Joint Conf. on Artificial Intelligence (IJCAI), Montreal (1995)
23. Bistarelli, S.: Semirings for Soft Constraint Solving and Programming. Lecture Notes in Computer Science. Springer Verlag (2004)
24. Schlesinger, M.I.: False minima of the algorithm for minimizing energy of max-sum labeling problem. Glushkov Institute of Cybernetics, Kiev, USSR. Unpublished (1976)
25. Bertsekas, D.P.: Nonlinear Programming. Athena Scientific, Belmont, MA (1999)
26. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press (2004)
27. Bessiere, C.: Constraint propagation. In: Handbook of Constraint Programming. Elsevier (2006) Also technical report LIRMM 06020.
28. Janssen, P., Jegou, P., Nouguier, B., Vilarem, M.: A filtering process for general constraint satisfaction problems: Achieving pairwise consistency using an associated binary representation. In: Proc. IEEE Workshop on Tools for Artificial Intelligence. (1989) 420–427