Simplification and extension of the SPREAD Constraint

Pierre Schaus[1], Yves Deville[1], Pierre Dupont[1]
Jean-Charles Régin[2]

[1] Université catholique de Louvain, Belgium
{pschaus, yde, pdupont}@info.ucl.ac.be
[2] ILOG, France
regin@ilog.fr

Abstract. Many assignment problems require the solution to be balanced. Such a problem is the Balanced Academic Curriculum Problem (BACP) [1]. Standard deviation is a common way to measure the balance of a set of values. A recent constraint presented by Pesant and Régin [2] enforces the mean μ and the standard deviation σ of a set of variables. Our work extends [2] by showing a more simple propagator from σ and μ to X and by introducing new propagators: from σ together with X to μ and from X together with μ to σ.

1 Introduction

In assignation problems, it is often desirable to have a fair or balanced solution. One example of such a problem is BACP. The goal is to assign periods to courses such that the academic load of each period is balanced, i.e., as similar as possible [1]. A perfectly balanced solution is generally not possible. A standard approach is to include the balance property in the objective function. Alternatively the constraint *SPREAD* introduced by Pesant and Régin [2] could be used to reduce the search tree while simplifying the model. Constraining the variance of assignments to fall below an upper bound is a proper way to enforce the balance property.

Given a set of variables X and two variables μ and σ, $SPREAD(X, \mu, \sigma)$ states that the collection of values taken by the variables of X exhibits an arithmetic mean μ and a standard deviation σ. While the *SPREAD* constraint in [2] also involves the median, this will not be considered here.

The *SPREAD* constraint can be seen as a special kind of soft constraint opening new perspectives in CSP modeling. As a perfect balanced solution is mostly not possible, the perfect balance constraint can be soften with *SPREAD* allowing a positive maximum standard deviation and an interval for the mean. *SPREAD* could also be used to combine a set of soft constraints. The usual way to combine a set of soft constraints is to minimize the sum of the violation cost of each of them. The drawback with this approach is that some constraints could be much more violated that the others. A clever way could be to use *SPREAD* to enforce the violation costs to be balanced among the soft constraints.

To appear in the Third International Workshop on Constraint Propagation And Implementation (CPAI'06), Nantes, France, September 25, 2006.

Section 2 mainly reviews the material from [2] used in this paper and introduces some statistical background and definitions relative to constraint programming. The problem of the variance minimization over the set of variables X is the starting point of our propagation algorithms. This problem is solved in [2] and explained in Section 2.

The propagator described in [2] filters from standard deviation σ to the set of variables X with quadratic time complexity with respect to the number of variables. It is possible to achieve better pruning by taking also the mean μ into consideration. Although the propagator from [2] can be easily extended to take also μ into account together with σ, this is not explicitly described in [2]. Section 3 presents a simpler filtering algorithm from μ and σ to X with the same time complexity.

We show in Section 4 that the problem of the variance minimization is a convex one. This implies that it admits a global minimum. This result allows us to design a propagator not present in [2]: from X and σ to μ. The filtering algorithm presented in Section 5 also performs in quadratic time with respect to the number of variables.

The filtering of the upper bound of the standard deviation requires a solution to the problem of the variance maximization. Section 6 shows that this problem is NP-hard and presents an algorithm to find an upper bound on the variance running in quadratic time with respect to the number of variables.

2 Background

We start this section with some statistical background and definitions relative to constraint programming. Next we present the problem of the variance minimization over the set of variables X. This problem is solved in [2] and is the starting point of our propagation algorithms.

We assume the reader familiar with common statistical notions such as *mean*, *standard deviation* and *variance* (these notions are defined in Section 2 of [2]). Note simply that a convenient way to compute the variance of a set of values $\{v_1, v_2, ..., v_n\}$ is the following: $\sigma^2 = \left(\frac{1}{n}\sum_{i=1}^{n} v_i^2\right) - \mu^2$.

We use the following notations for the variables and domains considered in this paper:

- A finite-domain (discrete) variable x takes a value in $D(x)$, a finite set called its domain. We denote the smallest (resp. largest) value x may take as $x^{\min}$ (resp. $x^{\max}$).
- A bounded-domain (continuous) variable y takes a value in $I_D(y) = [y^{\min}, y^{\max}]$, an interval on $\mathbb{R}$ called its domain as well.
- Given a finite-domain variable x, $I_D(x)$ denotes its domain relaxed to the continuous interval $[x^{\min}, x^{\max}]$. By extension for a union of domains $\mathcal{D} = \bigcup_{i=1}^{n} D(x_i)$, $I_{\mathcal{D}}$ represents the interval $[\min_{i=1}^{n} x_i^{\min}, \max_{i=1}^{n} x_i^{\max}]$.

The remaining of the section reviews the problem of the variance minimization solved in [2]. We detail successively the key points to find a solution:

1. A property of an optimal solution.
2. An optimal solution can be found by iterating once over a set of contiguous intervals.
3. The construction of this set of contiguous interval is based on the bounds of the domains.
4. For each interval, the optimal solution property can be checked in constant time.

Let first define formally the problem we want to solve. The variance minimization is an optimization problem under a sum constraint:

Definition 1 (Minimization of the variance on X). *Let $X = \{x_1, x_2, ..., x_n\}$ be a set of finite-domain (discrete) variables. For some fixed number q we denote by $\Pi_1(X, q)$ the problem:* $\min \sum_{i=1}^{n} (x_i - q/n)^2$ *such that* $\sum_{i=1}^{n} x_i = q$, $x_i \in I_D(x_i)$, $1 \leq i \leq n$ *and we denote by opt($\Pi_1(X, q)$), or simply opt(Π_1), the optimal value to this problem.*

In the above definition, $opt(\Pi_1)$ corresponds to n times the minimal variance and q to n times μ.

The following definition and lemma characterize an optimal solution to $\Pi_1(X, q)$. This property is a particular assignment of a variable x to a value of its relaxed domain to the continuous interval $I_D(x)$.

Definition 2. *An assignment $A : x \rightarrow I_D(x)$ is said to be a v-centered assignment when:*

$$A(x) = \begin{cases} x^{\max} \; if \; x^{\max} \leq v \\ x^{\min} \; if \; x^{\min} \geq v \\ v \quad otherwise \end{cases}$$

Lemma 1. *[2]. Any optimal solution to $\Pi_1(X, q)$ is a v-centered assignment.*

Lemma 1 gives a necessary condition for an assignment to be optimal for $\Pi_1(X, q)$ but the v value can be anywhere in $I_{\mathcal{D}}$. [2] introduces a splitting of $I_{\mathcal{D}}$ into contiguous intervals based on the bounds of the domains of variables. The v value of the v-centered assignment characterizing an optimal solution can be found by iterating once over this set of contiguous intervals. Any such interval is either subsumed by a domain or has an empty intersection with it but partial overlap cannot occur.

Definition 3. *Let $B(X)$ be the sorted sequence of bounds of the relaxed domains of the variables of X, in non-decreasing order and with duplicates removed. Define $\mathcal{I}(X)$ as the set of intervals defined by a pair of two consecutive elements of $B(X)$. The k^{th} interval of $\mathcal{I}(X)$ is denoted by I_k. For an interval $I = I_k$ we define the operator $prev(I) = I_{k-1}, (k > 1)$ and $succ(I) = I_{k+1}$.*

Example 1 (Building $\mathcal{I}(X)$). Let $X = \{x_1, x_2, x_3\}$ with $I_D(x_1) = [1, 3]$, $I_D(x_2) = [2, 6]$ and $I_D(x_3) = [3, 9]$ then $\mathcal{I}(X) = \{I_1, I_2, I_3, I_4\}$ with $I_1 = [1, 2]$, $I_2 = [2, 3]$, $I_3 = [3, 6]$, $I_4 = [6, 9]$. We have $prev(I_3) = I_2$ and $succ(I_3) = I_4$.

There are at most $2.n-1$ intervals in $\mathcal{I}(X)$. Let assume that the value v of the optimal solution to $\Pi_1(X,q)$ lies in the interval $I \in \mathcal{I}(X)$. We denote by $R(I) = \{x | x^{\min} \geq \max(I)\}$ the variables lying to the right of I and by $L(I) = \{x | x^{\max} \leq \min(I)\}$ the variables lying to the left of I. By Lemma 1, all variables $x \in L(I)$ take their value $x^{\max}$ and all variables in $R(I)$ take their value $x^{\min}$. It remains to assign the variables subsuming I. We denote these variables by $M(I) = \{x | I \subseteq I_D(x)\}$ and the cardinality of this set by $m = |M(I)|$. By Lemma 1, the variables of $M(I)$ must take a common value v. The sum constraint (see Definition 1) of $\Pi_1(X,q)$ can be rewritten as

$$\sum_{x \in R(I)} x^{\min} + \sum_{x \in L(I)} x^{\max} + \sum_{x \in M(I)} v = q. \tag{1}$$

Let denote the sum of extrema by

$$ES(I) = \sum_{x \in R(I)} x^{\min} + \sum_{x \in L(I)} x^{\max}.$$

The sum constraint in Equation (1) implies that v must be equal to a specific value $v^* = (q - ES(I))/m$. This results in a valid assignment only if $v^* \in I$. This condition is satisfied if

$$q \in V(I) = [ES(I) + \min(I).m, ES(I) + \max(I).m].$$

We previously said that an optimal solution the problem $\Pi_1(X,q)$ (see Definition 1) could be found by iterating once over a set of contiguous intervals by checking for each interval a property in constant time. The set of intervals is naturally $\mathcal{I}(X)$ introduced in Definition 3 and for each $I \in \mathcal{I}(X)$, the test is: does q belong to $V(I)$? If it is true that $q \in V(I)$, the value v of the v-centered assignment defined in Definition 2 characterizing an optimal solution $\Pi_1(X,q)$ lies in the interval I and has a value of $(q - ES(I))/m$.

We denote the overall minimal (resp. maximal) sum by $\underline{S}(X) = \sum_{x \in X} x^{\min}$ (resp. $\overline{S}(X) = \sum_{x \in X} x^{\max}$). We are sure that for every value $q \in [\underline{S}(X), \overline{S}(X)]$ there is one $I \in \mathcal{I}(X)$ such that $q \in V(I)$. Indeed, we have $\min(V(I_1)) = \underline{S}(X)$, $\max(V(I_{|\mathcal{I}(X)|})) = \overline{S}(X)$ and for two consecutive intervals I_k, I_{k+1} from $\mathcal{I}(X)$, we have $\min(V(I_{k+1})) = \max(V(I_k))$, thus leaving no gap.

Theorem 1. *[2] Given a value q such that $q \in [\underline{S}(X), \overline{S}(X)]$ and $I^q \in \mathcal{I}(X)$ such that $q \in V(I^q)$, the following assignment gives the optimal value to $\Pi_1(X,q)$:*

$$A(x) = \begin{cases} x^{\max} & if\ x \in L(I^q) \\ x^{\min} & if\ x \in R(I^q) \\ v = \frac{q - ES(I^q)}{m} & if\ x \in M(I^q) \end{cases}$$

Example 2 (Solving $\Pi_1(X,q)$). Variables and domains are from Example 1. We obtain the following values:

i	I_i	$R(I_i)$	$L(I_i)$	$M(I_i)$	$ES(I_i)$	$V(I_i)$
1	$[1, 2]$	x_2, x_3	ϕ	x_1	5	$[6, 7]$
2	$[2, 3]$	x_3	ϕ	x_1, x_2	3	$[7, 9]$
3	$[3, 6]$	ϕ	x_1	x_2, x_3	3	$[9, 15]$
4	$[6, 9]$	ϕ	x_1, x_2	x_3	9	$[15, 18]$

For $q = 10$ we have $q \in V(I_3)$ thus $I^{10} = I_3$. $A(x_1) = 3$, $A(x_2) = A(x_3) = 3.5$. For $q = 9$, we have $q \in V(I_2)$ and $q \in V(I_3)$. Whichever interval we choose between I_2 and I_3, we find the same optimal assignment $A(x_1) = 3$, $A(x_2) = 3$ and $A(x_3) = 3$.

3 Propagation from μ and σ to X

In this section, we propose to reformulate, simplify and extend (by considering explicitly the mean together with the standard deviation) the propagator given in [2]. The procedure to filter the domain of a variable $x \in X$ is the following:

- Shift the domain of x by a positive real quantity d.
- For some maximal shift $d = d^{\max}$, the minimum standard deviation reaches the upper bound $\sigma^{\max}$ of the domain of σ.
- The computed value $d^{\max}$ allows us to filter $D(x)$ since a shift larger that $d^{\max}$ would render the constraint inconsistent.

To clarify the presentation, we first assume that σ is an interval $[\sigma^{\min}, \sigma^{\max}]$ and μ is a given value. We consider afterward the general case where μ is an interval.

We recall and introduce some notations to explain more precisely the filtering procedure. We denote $q = n\mu$, $\pi_1^{\max} = n(\sigma^{\max})^2$ and $I^q \in \mathcal{I}(X)$ is such that $q \in V(I^q)$. In the following we use the shift operation $I + d$ by a positive quantity d on an interval $[I^{\min}, I^{\max}]$ to denote interval $[I^{\min} + d, I^{\max} + d]$. This operation also applies on domains of variables: we simply denote by $x' = x + d$ the variable x with a shifted domain $D(x) + d$.

These notations allow us to explain more precisely the filtering of the domain of one variable $x \in X$. First, the constraint fails if the minimum standard deviation is larger than the upper bound of σ. In this case, there exists no consistent assignment. This happens if $opt(\Pi_1) > \pi_1^{\max}$. When the constraint is consistent ($opt(\Pi_1) \leq \pi_1^{\max}$) we can consider the filtering of each variable $x \in X$. In particular for a variable $x \in R(I^q)$ (resp. $\in L(I^q)$), we compute its maximal consistent value (resp. minimal consistent value) by computing the maximal shift $d^{\max}$. For a variable $x \in M(I^q)$ we compute both. Each value larger (resp. smaller) than the maximal (resp. minimal) consistent value must be filtered. As the problem is symmetrical we only consider the computation of the maximal consistent value for $x \in R(I^q) \cup M(I^q)$.

For a variable $x \in R(I^q) \cup M(I^q)$ we show that shifting its domain $(D(x) + d)$ by $d \in \Re^+$ increases $opt(\Pi_1)$ ($n\times$ the minimum variance) quadratically with d. The bound $\pi_1^{\max}$ is reached for some d denoted by $d^{\max}$. The propagation on X

considers each variable $x \in X$ in turn, computes its maximum shift $d^{\max}$ and prunes $D(x)$ as follows: $D(x) \leftarrow D(x) \cap [x^{\min}, x^{\min} + d^{\max}]$. All the domains can be updated once after consideration of all variables in X. Alternatively, each pruned domain can directly be used for the propagation on the other variables.

Searching $d^{\max}$ for $x \in R(I^q)$ X' denotes X after the shift $x' = x + d$. Let $\Pi_1(X', q)$, $ES'(I^q)$ and $V'(I^q)$ be the corresponding quantities for X'. We have $ES'(I^q) = ES(I^q) + d$ and $V'(I^q) = V(I^q) + d$.

Let assume that $d \leq d_1 = q - \min(V(I^q))$ such that v' remains in I^q. In this case, the value of q leading to the value of $opt(\Pi_1)$ does not change (i.e. $q' = q$). Only the v value will change in the optimal assignment: $v' = v - d/m$. We have $opt(\Pi_1(X', q)) = \left(\sum_{x_i \in L(I^q)} (x_i^{\max})^2\right) + \left(\sum_{x_i \in R(I^q)} (x_i^{\min})^2\right) + d^2 + 2dx^{\min} + \left(\sum_{x_i \in M(I^q)} (v - \frac{d}{m})^2\right) - \frac{q^2}{n} = opt(\Pi_1(X, q)) + d^2 + 2dx^{\min} + m\left(\frac{d^2}{m^2} - 2\frac{d}{m}v\right)$. The value $d^{\max}$ is the positive solution of a second degree equation $ad^2 + 2bd + c$, where $a = (1 + \frac{1}{m})$, $b = x^{\min} - v$ and $c = opt(\Pi_1(X, q)) - \pi_1^{\max}$.

Until now, we made the assumption that $d \leq d_1$. If $d^{\max} > d_1$ this value is not valid since v does not lie within I^q anymore. In this case x is shifted by d_1 (i.e. $x^{\min}$ is increased by d) and the interval $I'^q = prev(I^q)$ is considered. The resulting Algorithm 1 searching for $d^{\max}$ runs in $O(n)$ since there are at most $|\mathcal{I}(\mathcal{X})| < n$ recursive calls and that the body runs in $O(1)$.

Algorithm: FindDMax(x, I^q)
Data: $x \in R(I^q)$; $I^q \in \mathcal{I}$; $q \in V(I^q)$;
Result: $d^{\max}$ s.t. $opt(\Pi_1(X', q)) = \pi_1^{\max}$ with $x' = x + d^{\max}$

```
d_1 = q − min(V(I^q))
d^max = (−b + √(b² − ac)) / a
if d^max < d_1 then
    return d^max
else
    if I^q = I_1 then
        return d_1
    else
        return d_1 + FindDMax(x + d_1, prev(I^q))
    end
end
```

Algorithm 1: FindDMax

Searching $d^{\max}$ with $x \in M(I^q)$ can be reduced to searching for $d^{\max}$ with a new variable x' with $x'^{\min} = v$. When x is increased ($x' = x + d$), the optimal assignment does not change if $d \leq v - x^{\min}$ (i.e. the values of $A(x)$ remain the same) . For $d = v - x^{\min}$ two new intervals are created replacing the old I^q: $I_j = [\min(I^q), v]$ and $I_k = [v, \max(I^q)]$ with $q = \max(V'(I_j)) = \min(V'(I_k))$. The optimal assignment is the same but a new problem $\Pi_1(X', q)$ is created with $q \in V'(I_j)$ and $x' \in R(I_j)$. This case reduced to searching for $d^{\max}$ with $x' \in R(I_j)$ is exposed above. The final $d^{\max}$ relative to the variable x is given by:

$$d^{\max} = v - x^{\min} + FindDMax(x', I_j) \;\; where \;\; x' = x + v - x^{\min} \qquad (2)$$

Example 3 (Filtering of one domain). Variables and domains are from Example 1. This example shows the filtering of the domain of variable x_2 for $q = 10$ and $\pi_1^{\max} = 8$. We are in the case of searching $d^{\max}$ with $x \in M(I^q)$. We have $I^{10} = [3, 6]$ because $10 \in V([3, 6]) = [9, 15]$ and the value v of the v-centered assignment is 3.5 (see Example 2). From Equation (2) we have $d^{\max} = 3.5 - 2 + FindDMax(x_2', [3, 3.5])$ where $x_2' = x_2 + 1.5$. We now analyze the successive calls to Algorithm 1.

1. $FindDMax(x_2 + 1.5, [3, 3.5])$. We have $ES([3, 3.5]) = 6.5$, $V([3, 3.5]) = [9.5, 10]$, $d_1 = 0.5$, $a = 2$, $b = 0$ and $c = (3 - 10/3)^2 + 2 * (3.5 - 10/3)^2 - 8 \approx -7.83$. We can compute $d^{\max} \approx 1.98$. Since $d^{\max} > d_1$ we have the recursive call $FindDMax(x_2 + 1.5 + 0.5, [1, 3])$.
2. $FindDMax(x_2 + 1.5 + 0.5, [1, 3])$. We have $ES([1, 3]) = 7$, $V([1, 3]) = [8, 10]$, $d_1 = 2$, $a = 2$, $b = 1$ and $c = 2 * (3 - 10/3)^2 + (4 - 10/3)^2 - 8 \approx -7.33$. We can compute $d^{\max} \approx 1.48$. Since $d^{\max} < d_1$ we can return 1.48.

In conclusion, the $d^{\max}$ value of x_2 is $1.5 + 0.5 + 1.48 = 3.48$ and the variable x_2 can be filtered $D(x_2) \leftarrow D(x_2) \cap [2, 5.48]$.

An example of the application of Algorithm 1 is given in Figure 1. The complexity analysis of Algorithm 1 shows that $d^{\max}$ is computed in $O(n)$ making the propagation on whole X running in $O(n^2)$.

Extension to $\boldsymbol{\mu = [\mu^{\min}, \mu^{\max}]}$ The generalization $\mu = [\mu^{\min}, \mu^{\max}]$ is equivalent to $q \in [q^{\min} = n\mu^{\min}, q^{\max} = n\mu^{\max}]$. This extension does not affect our propagator but only requires an additional step before the call to $FindDMax$ for each variable: the computation of a suitable $q \in [q^{\min}, q^{\max}]$. The computation of $d^{\max}$ in the algorithm depends on the value of q. To express this explicitly we denote $d^{\max}$ as a function of q: $d^{\max}(q)$. Since it can be shown to be concave and derivable, one can search a q^0 such that $d^{\max}(q)$ is maximum: $\frac{\partial d^{\max}}{\partial q} \mid_{q=q^0} = 0$. It can be shown that q^0 is the only valid solution of a second degree equation . As $q \in [q^{\min}, q^{\max}]$, if $q^0 > q^{\max}$ (resp. $< q^{\min}$) then $FindDmax$ is called with $q = q^{\max}$ (resp. $q = q^{\min}$). If $q^0 \in [q^{\min}, q^{\max}]$, $FindDmax$ is called with $q = q^0$.

4 Study of $\Pi_1(X, q)$

We show in this section that the problem of the variance minimization with given mean is convex. This result allows us to design a propagator from X and

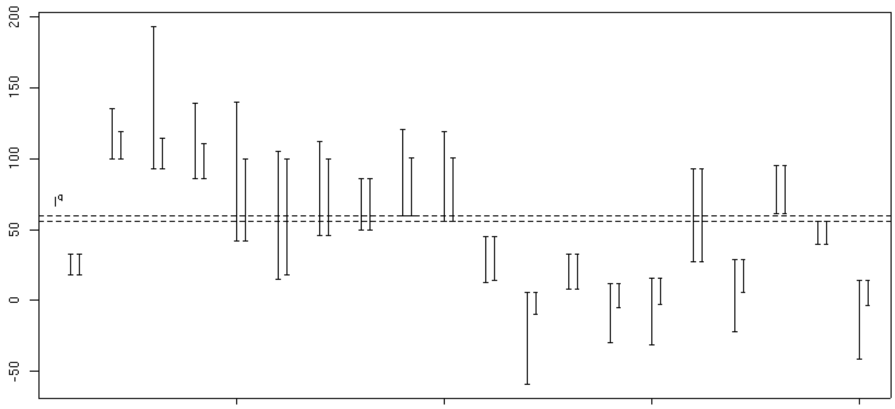


Fig. 1. The propagation on a typical run. The I^q interval lies between the two horizontal lines. The posted constraint is $SPREAD(X, 50, [0, 28])$. There are 20 variables and the domains after the propagations are represented on the right of each original domain.

σ to μ in Section 5. Indeed, the values for the mean leading to a minimum standard deviation larger than the upper bound $\sigma^{\max}$ must be filtered. Thanks to the convexity property, all inconsistent values for the mean will be filtered by computing only two values for the mean such that the upper bound $\sigma^{\max}$ is reached. All the values for the mean not between these two computed values are inconsistent.

More precisely, in this section we characterize completely the function of q $opt(\Pi_1(X, q))$ which is the minimization of the variance for a fixed mean (see Definition 1). We demonstrate that $opt(\Pi_1(X, q))$ is continuous, derivable, convex and accepts one global minimum on $[\underline{S}(X), \overline{S}(X)]$. Figure 2 shows a typical set X of variables with their domains and the corresponding functions $opt(\Pi_1(X, q))$. You can see on the figure that $opt(\Pi_1(X, q))$ is continuous, convex with one global minimum.

Theorem 2 (Characteristics of $opt(\Pi_1(X, q))$). *Assuming a domain for q in the interval $[\underline{S}(X), \overline{S}(X)]$ and a given set of variables X the optimal value to $\Pi_1(X, q)$ denoted by $opt(\Pi_1(X, q))$ is continuous, differentiable and convex having a global minimum for some $q \in [\underline{S}(X), \overline{S}(X)]$.*

Proof. It is sufficient to show that $\partial opt(\Pi_1(X, q))/\partial q$ is a continuous (1) increasing (2) function with $\partial opt(\Pi_1(X, q))/\partial q \mid_{q=\underline{S}(X)} \leq 0$ and $\partial opt(\Pi_1(X, q))/\partial q \mid_{q=\overline{S}(X)} \geq 0$ (3).

The function $opt(\Pi_1(X, q))$ is piecewise defined on $[\underline{S}(X), \overline{S}(X)]$: for $q \in V(I_k)$, $opt(\Pi_1(X, q)) = C + \sum_{x_i \in M(I_k)} ((q - ES(I_k))/m)^2 - q^2/n$ where $C = \sum_{x_i \in L(I_k)} (x_i^{\max})^2 + \sum_{x_i \in R(I_k)} (x_i^{\min})^2$. The derivative is also piecewise defined:

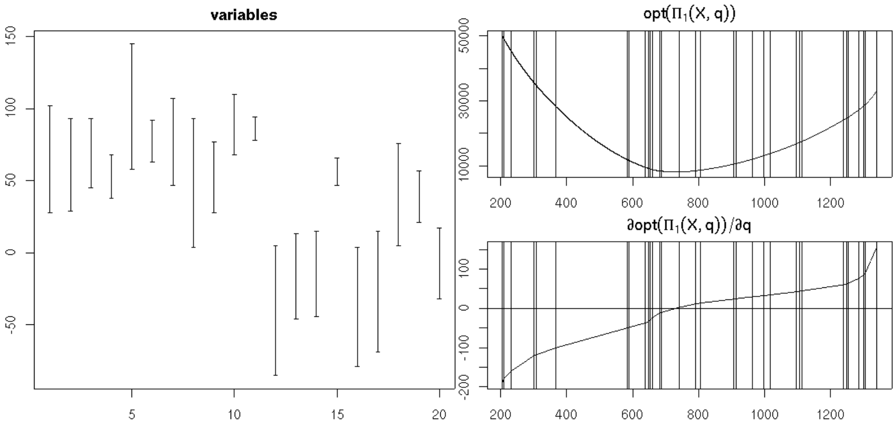


Fig. 2. On the left a typical set X is represented with the domain of each variables. On the right top and bottom $opt(\mathit{\Pi}_1(X,q))$ and $\partial opt(\mathit{\Pi}_1(X,q))/\partial q$ are respectively represented for $q \in [\underline{S}(X), \overline{S}(X)]$. The vertical lines represent $V(I_k), 1 \leq k \leq |\mathcal{I}(X)|$.

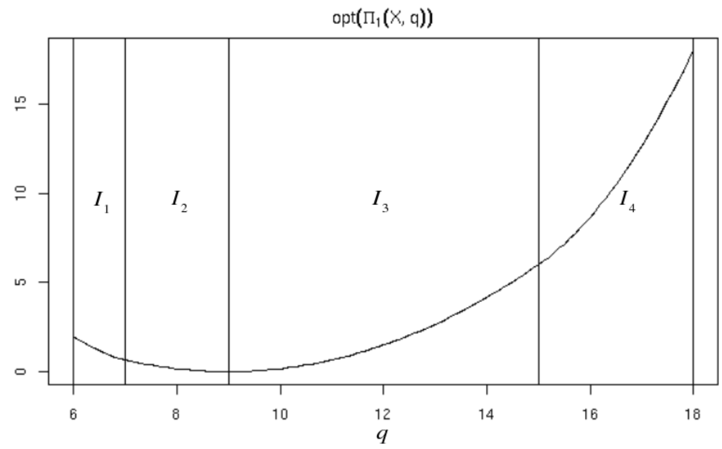
for $q \in V(I_k)$, $\partial opt(\mathit{\Pi}_1(X,q))/\partial q = 2(q - ES(I_k))/m - 2q/n$. The proofs for (1),(2) and (3) are:

1. The derivative is continuous because for $q = \max(V(I_k)) = \min(V(I_{k+1}))$, the values obtained on interval $V(I_k)$ and $V(I_{k+1})$ are the same: $2.\frac{q-ES(I_k)}{|M(I_k)|} - 2\frac{q}{n} = 2\frac{q-ES(I_{K+1})}{|M(I_{K+1})|} - 2\frac{q}{n}$. Indeed, by denoting $\delta^m = |M(I_{k+1})| - |M(I_k)|$ we have $\frac{q-ES(I_k)}{|M(I_k)|} = \frac{ES(I_k)+|M(I_k)|\max(I_k)-ES(I_k)}{|M(I_k)|} = \max(I_k)$ and $\frac{q-ES(I_{k+1})}{|M(I_{k+1})|} = \frac{ES(I_k)+|M(I_k)|\max(I_k)-ES(I_k)+\delta^m\max(I_k)}{|M(I_k)|+\delta^m} = \max(I_k)$.
2. Since $\partial^2 opt(\mathit{\Pi}_1(X,q))/\partial q^2 = 2(\frac{1}{|M(I_k)|} - \frac{1}{n}) \geq 0$, $\partial opt(\mathit{\Pi}_1(X,q))/\partial q$ is non decreasing on $V(I_k)$. Because $\partial opt(\mathit{\Pi}_1(X,q))/\partial q$ is continuous (1) and non decreasing on each interval the function is globally convex on $[\underline{S}(X), \overline{S}(X)]$.
3. Note that $ES(I_1) = \underline{S}(X) - m\min(I_1)$ and $ES(I_{|\mathcal{I}(X)|}) = \overline{S}(X) - m\max(I_{|\mathcal{I}(X)|})$. $\partial opt(\mathit{\Pi}_1(X,q))/\partial q\ |_{q=\underline{S}(X)} = \frac{\underline{S}(X)-\underline{S}(X)+m\min(I_1)}{m} - \frac{\underline{S}(X)}{n} = \min(I_1) - \frac{\underline{S}(X)}{n} \leq 0$. $\partial opt(\mathit{\Pi}_1(X,q))/\partial q\ |_{q=\overline{S}(X)} = \frac{\overline{S}(X)-\overline{S}(X)+m\max(I_{|\mathcal{I}(X)|})}{m} - \frac{\overline{S}(X)}{n} = \max(I_{|\mathcal{I}(X)|}) - \frac{\overline{S}(X)}{n} \geq 0$. □

Example 4 (Study of $opt(\mathit{\Pi}_1(X,q))$). Variables and domains are from Example 1. We study the function $opt(\mathit{\Pi}_1(X,q))$. With help of the table from Example 2 we add one column which is the definition of $opt(\mathit{\Pi}_1(X,q))$ for $q \in V(I_i)$.

i	I_i	$R(I_i)$	$L(I_i)$	$M(I_i)$	$ES(I_i)$	$V(I_i)$	C	$opt(\Pi_1(X,q))$ for $q \in V(I_i)$
1	$[1,2]$	x_2, x_3	ϕ	x_1	5	$[6,7]$	13	$13 + 1 * (\frac{q-5}{1})^2 - \frac{q^2}{3}$
2	$[2,3]$	x_3	ϕ	x_1, x_2	3	$[7,9]$	9	$9 + 2 * (\frac{q-3}{2})^2 - \frac{q^2}{3}$
3	$[3,6]$	ϕ	x_1	x_2, x_3	3	$[9,15]$	9	$9 + 2 * (\frac{q-3}{2})^2 - \frac{q^2}{3}$
4	$[6,9]$	ϕ	x_1, x_2	x_3	9	$[15,18]$	45	$45 + 1 * (\frac{q-9}{1})^2 - \frac{q^2}{3}$

Each function $opt(\Pi_1(X,q))$ for $q \in V(I_i)$ is plotted on the following graphics. Clearly the minimum is reached for $q = 9$ in this example.



5 Propagation from X and σ to μ

As already explained at the beginning of Section 4, the convexity property of the problem of variance minimization (see Theorem 2) with given mean allows us to design an efficient propagator from X and σ to μ. All the values for the mean leading to a minimum standard deviation larger than the upper bound $\sigma^{\max}$ can be filtered. Thanks to the convexity property, all inconsistent values for the mean will be filtered by computing only two values for the mean such that the upper bound $\sigma^{\max}$ is reached. All the values for the mean not between these two computed values are inconsistent.

We now explain more precisely the narrowing of μ with help of Figure 3. The function $\underline{\sigma}(\mu)$ depicted on figure 3 is the function $\sqrt{opt(\Pi_1(X,q))/n}$ with $\mu = q/n$. Naturally this function has the same properties than the function $opt(\Pi_1(X,q))$. The constraint $\sigma \leq \sigma^{\max}$ is represented by a vertical line crossing $\underline{\sigma}(\mu)$ in two points. The projection of these two points on the mean axis gives the two values μ_1, μ_2 for the mean such that the minimum standard deviation is equal to the upper bound of σ. All mean values outside the interval $[\mu_1, \mu_2]$ are inconsistent and can be filtered.

It is possible that the maximum standard deviation is so large that it does not constraint the mean. In this case $\mu_1 = \underline{S}/n$ and $\mu_2 = \overline{S}/n$ and we have simply a propagation from X to μ.

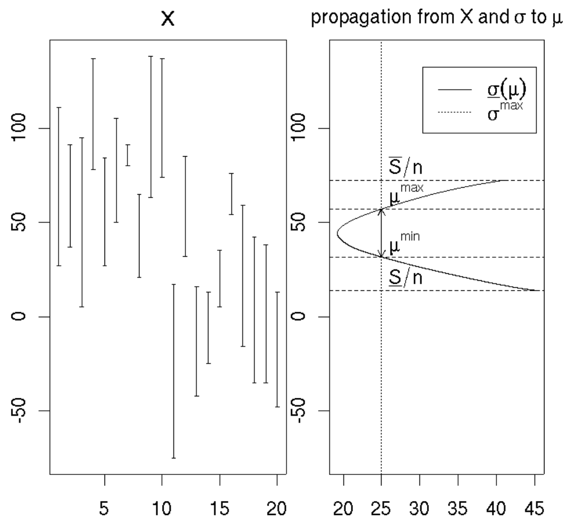


Fig. 3. Propagation from X and σ to μ.

In the remaining of this section we explain how the two values μ_1, μ_2 are found and finally we give the resulting filtering algorithm for μ

As already said μ_1, μ_2 are the projection on the mean axis of the two cross points of $\sqrt{opt(\Pi_1(X,q))/n}$ with $\sigma^{\max}$ (see Figure 3). These two cross points are obtained by considering each interval $V(I_k)$ in turn. It is possible to find two values $n.\mu_1 = q_1 \leq q_2 = n.\mu_2$ for q such that $opt(\Pi_1(X,q_1)) = opt(\Pi_1(X,q_2)) = \pi_1^{\max} = n(\sigma^{\max})^2$ and $\forall q \in [q_1,q_2], opt(\Pi_1(X,q)) \leq \pi_1^{\max}$. The two values q_1, q_2 are found as follows. For every value of q: $opt(\Pi_1(q)) = C + \sum_{x_i \in M(I_k)} \left(\frac{q-ES(I_k)}{m}\right)^2 - \frac{q^2}{n}$ where $C = \sum_{x_i \in L(I_k)} (x_i^{\max})^2 + \sum_{x_i \in R(I_k)} (x_i^{\min})^2$. Then, q_1 and q_2 are the solutions of the second degree equation $aq^2 + 2bq + c$ where $a = (1/m - 1/n)$, $b = -ES(I_k)/m$ and $c = C + (1/m).ES(I_k)^2 - \pi_1^{\max}$. If $q_1 = (-b - \sqrt{b^2 - ac})/a \in V(I_k)$ then $\mu_1 = q_1/n$ is a lower bound of the permitted interval for μ. If $q_2 = (-b + \sqrt{b^2 - ac})/a \in V(I_k)$ then $\mu_2 = q_2/n$ is the upper bound of the permitted interval for μ. Else there is no bounds in $V(I_k)$. The resulting Algorithm 2 narrows the interval μ if possible.

Example 5 (Filtering of μ).

Variables and domains are from Example 1. We search the permitted values for μ under the constraint $\pi_1^{\max} = 8$. Clearly, if we look at the figure of Example 4, we can deduce that $q^{\min} = 6$ but the upper bound $q^{\max}$ must be computed. All we know by looking at the figure is that $q^{\max} \in V(I_4)$ because the curve $opt(\Pi_1(X,q))$ intersects $\pi_1^{\max} = 8$ in this interval. We can take the expression of $opt(\Pi_1(X,q))$ on the interval $V(I_4)$ (see Example 4) and compute the value $q^{\max}$ such that $opt(\Pi_1(X,q^{\max})) = \pi_1^{\max} = 8$. We have the equation $45 + 1 *$

Algorithm: MeanPruning

Result: narrowing of μ
set $\mu^{min} \geq \underline{S}/n$
set $\mu^{max} \leq \overline{S}/n$
for $1 \leq k \leq |\mathcal{I}(X)|$ **do**
    $q_1 = (-b - \sqrt{b^2 - ac})/a$
    if $q_1 \in V(I_k)$ **then**
        set $\mu^{min} \geq q_1/n$
        break
    end
end
for $|\mathcal{I}(X)| \geq k \geq 1$ **do**
    $q_2 = (-b + \sqrt{b^2 - ac})/a$
    if $q_2 \in V(I_k)$ **then**
        set $\mu^{max} \leq q_2/n$
        break
    end
end

Algorithm 2: MeanPruning

$(\frac{q^{\max}-9}{1})^2 - \frac{(q^{\max})^2}{3} = 8$ and we find $q^{\max} \approx 15.79$. A bound consistent interval for the mean is thus $[6/3, 15.79/3] = [2, 3.74]$.

6 Narrowing of σ

The propagation from X and μ to $\sigma^{\min}$ is detailed in [2]. We propose to study the propagation from X and μ to $\sigma^{\max}$.

The decreasing of the upper bound of σ requires to compute the maximal variance on X with a given mean. This can be shown to be a convex maximization problem (NP-hard in general [3]). Even the relaxed problem without the sum constraint remains a convex maximization problem but it is easier to design an upper bound on it because of a known characterization of the optimal solution with respect to the extrema of the domains. We describe in this section a quadratic running time algorithm (with respect to the number of variables) to find an upper bound on the variance.

The maximization problem we want to solve is:

Definition 4 (Maximization of the variance on X). *Let $X = \{x_1, x_2, ..., x_n\}$ be a set of finite-domain (discrete) variables. We denote by $\Pi_2(X)$ the problem: $\max \sum_{i=1}^{n}(x_i - \sum_{j=1}^{n} x_j/n)^2$. We denote by $opt(\Pi_2(X))$ the optimal value for the problem.*

Since $opt(\Pi_2(X)) = \sum_i x_i^2 - \left(\sum_i x_i\right)^2/n$, an upper bound $\overline{opt}(\Pi_2(X))$ can be computed using the bound values $x_i^{\max}$ (resp. $x_i^{\min}$) in the first (resp. second) sum. This upper bound can be used to narrow the interval σ by setting $n.\sigma^2 \leq \overline{opt}(\Pi_2(X))$.

Example 6 (Upper bound). We consider the same variables and domains as in Example 1. We have $X = \{x_1, x_2, x_3\}$ with $I_D(x_1) = [1, 3]$, $I_D(x_2) = [2, 6]$ and $I_D(x_3) = [3, 9]$. $\overline{opt}(\Pi_2(X)) = (3^2 + 6^2 + 9^2) - (1 + 2 + 3)^2/3 = 114$.

The following lemma gives a property on an optimal assignment for the variance maximization problem. We will use this property to improve the upper bound in $O(n^2)$.

Lemma 2 (Optimal solution to $\Pi_2(X)$). *Any optimal solution to $\Pi_2(X)$ must be an assignment on the extrema of the domains i.e. on $x^{\max}$ or $x^{\min}$.*

Proof (Proof of Lemma 2). It is sufficient to show that starting from an arbitrary assignment and choosing an arbitrary variable $x_i > \sum_j x_j/n$, assigning a greater value to x_i i.e. $x_i \leftarrow x_i + d$ will increase the variance on X. The previous variance was $\sigma^2 = \frac{1}{n}\sum_j x_j^2 - \frac{1}{n^2}(\sum_j x_j)^2$ the variance with the modified variable is $\sigma'^2 = \frac{1}{n}\sum_j x_j^2 + \frac{1}{n}(d^2 + 2dx_i) - \frac{1}{n^2}(\sum_j x_j)^2 - \frac{1}{n2}(d^2 + 2\sum_j(d.x_j))$. The result is $\sigma'^2 = \sigma^2 + \frac{1}{n}(d^2 + 2dx_i) - \frac{1}{n^2}(d^2 + 2d\sum_j(x_j)) > \sigma^2 + \frac{1}{n}(d^2 + 2dx_i) - \frac{1}{n^2}(d^2 + 2dnx_i) = \sigma^2 + \frac{1}{n}d^2 - \frac{1}{n^2}d^2$ with $\sigma'^2 > \sigma^2$. The same result holds by symmetry for a variable $x_i < \sum_j x_j/n$ if it is decreased $x_i \leftarrow x_i - d$. □

As already explained, an upper bound for $\overline{opt}(\Pi_2(X))$ can be computed using the values and $x_i^{\max}$ (resp. $x_i^{\min}$) in the first (resp. second) sum of $\sum_i x_i^2 - \left(\sum_i x_i\right)^2/n$. With Lemma 2, it is possible to improve this bound. In each case where the lower-bound using an extrema is larger than the upper-bound using the other extrema, the optimal assignment corresponds to the first extrema. If for one variable, the extrema assignment can be found, then we can use this extrema value in the first and in the second sum to decrease the upper bound. If all the extrema assignment could be found the upper bound would be optimal (equal to the maximum variance). There are 2^n possible extrema assignments on X. We suggest an $O(n^2)$ algorithm to deduce as much extrema assignments as possible.

We now detail the method to deduce the correct extrema assignment of some variables. We denote $\underline{\mu} = \underline{S}(X)/n$ and $\overline{\mu} = \overline{S}(X)/n$. For some variables the optimal assignment can be deduced immediately. Indeed if $x^{\min} > \overline{\mu}$, an optimal solution to $\Pi_2(X)$ is such that $x = x^{\max}$. The case $x^{\max} < \underline{\mu}$ is symmetrical. There are additional cases where extrema assignment can be deduced. Note that if x would be assigned to $x^{\min}$, the upper bound for μ would become $\overline{\mu}^* = \overline{\mu} - \frac{x^{\max} - x^{\min}}{n}$.

In the example on the left of Figure 4, an optimal solution would assign $x = x^{\max}$ because the lower bound on the distance of $x^{\max}$ to μ is greater than the upper bound on the distance of $x^{\min}$ to $\overline{\mu}^*$. More generally, in each case where the lower-bound using an extrema is larger than the upper-bound using the other extrema, the optimal assignment corresponds to the first extrema.

Assigning a variable x to $x^{\min}$ will decrease $\overline{\mu}$ and assigning a variable x to $x^{\max}$ will increase $\overline{\mu}$ resulting possibly in a larger set of variables for which an optimal assignment can be deduced. All such extrema can be found in $O(n^2)$.

Example 7 (Deducing extrema assignment). We consider the same variables and domains as in Example 1. We have $X = \{x_1, x_2, x_3\}$ with $I_D(x_1) = [1, 3]$, $I_D(x_2) = [2, 6]$ and $I_D(x_3) = [3, 9]$. We have $\underline{\mu} = 2$ and $\overline{\mu} = 6$.

- x_3: If we assign x_3 to 9 then we have $\underline{\mu} = 4, \overline{\mu} = 6$ and the smallest distance from 9 to μ is $9 - 6 = 3$. If we assign x_3 to 3 then we have $\underline{\mu} = 2, \overline{\mu} = 4$ and the largest distance from 3 to μ is 1. We are sure that the correct extrema assignment for x_3 is 9 because whichever the assignment on other variables is, the distance to μ (and thus the variance also) will always be greater with x_3 assigned to 9. The new values for the bounds on the mean are now $\underline{\mu} = 4, \overline{\mu} = 6$.
- x_2: A similar argument as for x_3 leads to the conclusion that the extrema assignment on x_2 is 2.
- x_1: Since the distance to the mean is always larger with x_1 assigned to 1 because $x_1^{\max} = 3 < \underline{\mu} = 4$ we are sure that it is the correct extrema assignment.

Example 8 (Upper bound with extrema assignments). The extrema assignment computed in Example 7 can be used to compute $\overline{opt}(\Pi_2(X))$. In this example, all the extrema assignments could be deduced. Consequently we have $\overline{opt}(\Pi_2(X)) = opt(\Pi_2(X)) = (1-4)^2 + (2-4)^2 + (9-4)^2 = 38$.

For the example on the right of Figure 4 with 50 variables, the algorithm find the optimal solution i.e. $\overline{opt}(\Pi_2(X)) = opt(\Pi_2(X))$. The deduced extrema are indicated with a $\oplus$. The worst case for propagating on σ would correspond to all variables with an identical domain.

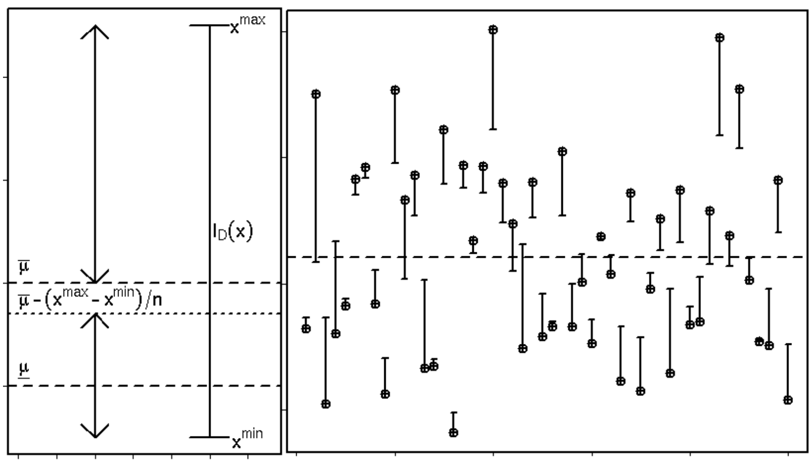


Fig. 4. Left figure: $x = x^{\max}$ because the lower bound on the distance from $x^{\max}$ to μ is smaller than the upper bound on the distance from $x^{\min}$ to μ. Right figure: $\overline{opt}(\Pi_2(X)) = opt(\Pi_2(X))$. The deduced extrema are indicated with a $\oplus$

7 Conclusion

In this paper we have considered a constraint dealing with statistics: the Spread constraint. This constraint and some filtering algorithms associated with it have been proposed by [2]. First, we have shown that simpler filtering algorithms with the same efficiency can be designed. Then, we have studied the main problem on which the constraint is based, that is the minimization of the standard deviation, and we have proved that this problem has a unique optimal value. From this result, we have proposed for the first time an algorithm reducing the values of the mean from the variables and the standard deviation. At last, we have shown that the computation of the maximal value of the standard deviation is NP-hard and we have given an algorithm to compute an upper bound of that value.

References

1. *Problem 30 of CSPLIB (www.csplib.org).*
2. Jean-Charles Regin Gilles Pesant. Spread: A balancing constraint based on statistics. *Lecture Notes in Computer Science*, 3709:460–474, 2005.
3. Lieven Vandenberghe Stephen Boyd. *Convex Optimization.* Cambridge University Press, 2004.