

# Esame di Programmazione su Architetture Parallele

## Metodo del simplesso per la risoluzione della programmazione lineare

Belliato Riccardo (mat. 142652)

Simone Tomada

2022-09-12

### Abstract

In questa relazione si propone una implementazione del metodo del simplesso a due fasi in CUDA per la risoluzione dei problemi di programmazione lineare in forma canonica.

Dopo una breve descrizione dell'algoritmo, seguirà la discussione su alcune scelte implementative.

Infine verranno valutate performance e scalabilità della soluzione proposta confrontando i tempi di esecuzione dell'algoritmo su istanze a dimensione crescente generate casualmente.

## Contents

<b>Introduzione al metodo del simplesso</b>	<b>1</b>
Problemi di ottimizzazione lineare . . . . .	1
Forma canonica e forma standard . . . . .	2
Metodo del simplesso a due fasi . . . . .	3
<b>Scelte implementative</b>	<b>5</b>
Gestione della memoria . . . . .	5
Schema di parallelizzazione . . . . .	7
<b>Risultati sperimentali</b>	<b>9</b>
Misurazione dei tempi . . . . .	9

## Introduzione al metodo del simplesso

### Problemi di ottimizzazione lineare

Nell'ambito della Ricerca Operativa (Operations Research) uno dei principali argomenti è la cosiddetta **ottimizzazione lineare**, ossia lo studio di una classe di problemi del tipo:

$$\begin{aligned} & \min / \max c^T x \\ & \text{subject to} \\ & Ax \preceq b \end{aligned}$$

con  $x \in \mathbb{R}^n$ ,  $c \in \mathbb{R}^n$ ,  $A \in \mathbb{R}^{m \times n}$ ,  $b \in \mathbb{R}^{1 \times m}$ .

In altre parole, si vogliono trovare dei valori per le componenti del vettore  $x$  (di dimensione  $n$ ) tali da massimizzare (o minimizzare) il valore di una *funzione lineare* variabili (detta **funzione obiettivo**, mentre il vettore  $c$  è chiamato **vettore dei costi**), dati una serie di  $m$  vincoli espressi nel sistema di disequazioni lineari  $Ax \preceq b$  (dove  $A$  è detta **matrice dei vincoli** e  $b$  **vettore dei termini noti**).

Questa classe di problemi permette di modellare un gran numero di situazioni reali in molteplici ambiti (ottimizzazione dei costi, creazione di orari, *etc.*), oltre che alcuni problemi NP-hard come il *Knapsack* o il *Vertex Cover*.

Dal punto di vista dell'algebra lineare, un problema in  $n$  variabili non è altro che uno spazio  $\mathbb{R}^n$ , la funzione obiettivo è una retta nello spazio, mentre i vincoli definiscono un *poliedro* nello spazio.

Utilizzando i teoremi e le tecniche dell'algebra lineare è stato possibile creare degli algoritmi per risolvere i problemi di ottimizzazione, come il **simplex**, i quali si prestano molto bene ad essere parallelizzati (in quanto operano su matrici).

### Problemi risolvibili, non risolvibili, illimitati

Dato un problema di ottimizzazione, questo può essere di tre tipi:

- ammissibile (*feasible*): ossia esistono uno o più vettori che moltiplicati per il vettore dei costi assegnano alla funzione obiettivo il valore massimo (o minimo possibile) e tutti i vincoli sono rispettati,
- non ammissibile (*infeasible*): se non esistono soluzioni
- illimitati (*unbounded*): se per una o più componenti della soluzione è possibile aumentarne (o ridurne) il valore all'infinito senza mai violare i vincoli
- degeneri (*degenerate*): sono problemi ammissibili in cui più soluzioni di base corrispondono allo stesso vertice del poliedro. In linea teorica possono far ciclare l'algoritmo (ossia l'algoritmo torna a controllare basi già viste in precedenza), ma all'atto pratico questo può essere evitato introducendo delle euristiche. Per semplicità la nostra implementazione una volta riconosciuto un problema come degenero lo segnala all'utente e termina senza cercare di risolverlo.

### Forma canonica e forma standard

I problemi di massimizzazione possono essere convertiti in problemi di minimizzazione (e viceversa), così come è possibile manipolare le singole disequazioni. Queste operazioni servono a riportare i problemi in una forma che permetta di utilizzarli da parte dei solver. In particolare vengono utilizzate la forma *canonica* e la forma *standard*.

Un problema (di massimizzazione) è in forma canonica se è nella forma

$$\begin{aligned} &\max c^T x \\ &\text{subject to} \\ &Ax \leq b \\ &x \geq 0 \end{aligned}$$

Considerando che qualsiasi disequazione nella forma  $\alpha x \leq y$  può essere convertita in una equazione equivalente  $\alpha x + \delta = b$  definiamo la forma standard di un problema in forma canonica

$$\begin{aligned} &\max(c|0)^T x \\ &\text{subject to} \\ &(A|I)x = b \\ &x \geq 0 \end{aligned}$$

con  $(A|I) \in \mathbb{R}^{m \times (n+m)}$ ,  $(c|0) \in \mathbb{R}^{n+m}$  e  $I$  la matrice di identità di dimensione  $m$ .

In altre parole aggiungiamo una nuova variabile al problema (detta variabile *slack*) per ogni disequazione.

La forma standard è quella che viene utilizzata dagli algoritmi di soluzione.

## Metodo del simplesso a due fasi

Il metodo del simplesso è un'algoritmo per risolvere i problemi di ottimizzazione lineari.

Si basa sul teorema secondo il quale le soluzioni (dette *soluzioni di base*) di un qualsiasi problema in forma standard sono i *vertici* del poliedro costruito sui vincoli.

Ogni vertice è individuato da una o più *soluzioni di base ammissibili*, cioè tutti i valori delle variabili in base sono maggiori o uguali a 0, mentre il valore delle altre variabili è 0. Se in base sono presenti uno o più valori uguali a 0, si dice che la base è *degenere*.

Esistono diverse implementazioni del metodo del simplesso, per questo progetto verrà implementato il cosiddetto *simplesso a due fasi con il metodo del tableau*

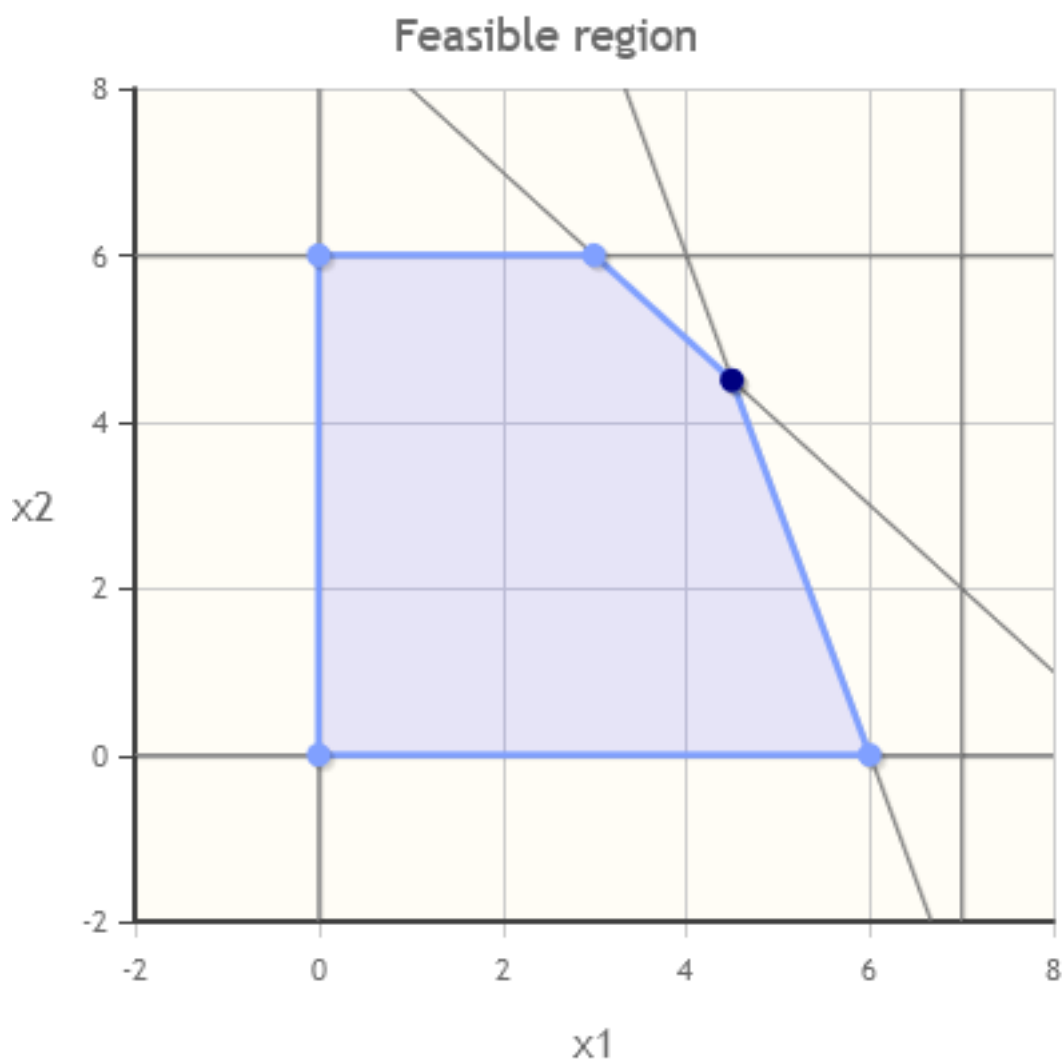


Figure 1: Esempio di problema con due variabili sul piano cartesiano. La regione evidenziata è il poliedro costruito sui vincoli. Il vertice evidenziato in blu scuro è la soluzione.

### Cos'è il tableau

Il tableau di un simplesso è la struttura dati su cui viene eseguito l'algoritmo del simplesso.

Si tratta di una matrice composta nel seguente modo:

Vettore dei costi	$x_1$	$x_2$	$s_1$	$s_2$	$s_3$	RHS	Valore funzione obiettivo
	-3	-2	0	0	0	0	
	2	1	1	0	0	10	
	1	1	0	1	0	8	
	1	0	0	0	1	4	
Matrice dei vincoli						Termini noti	

Nella prima riga della matrice ( $R_0$ ) è presente il vettore dei costi. Inizialmente il vettore dei costi contiene i coefficienti della funzione obiettivo con segno invertito. La base viene memorizzata in un vettore a parte.

### Algoritmo di soluzione

---

#### Algorithm 1: Simplex

---

**Data:**  $T \in \mathbb{R}^{(m+1) \times (n+1)}$ ,  $\mathcal{B} \in \mathbb{N}^m$

**Output:**  $x$  come soluzione,  $\mathcal{B}$  base della soluzione e  $T[0][n]$  valore ottimo della funzione obiettivo

OPPURE problema UNBOUNDED

```

1  $A \leftarrow T[1..m][0..n-1]$  ;      /* La notazione  $[a..b]$  indica una porzione di array dall'indice  $a$ 
   all'indice  $b$  */
2  $b \leftarrow T[1..m][n]$ ;
3  $c \leftarrow T[0][0..n-1]$ ;
4 while  $\exists i : c[i] < 0$  do
5   Pick  $h \in \{0, \dots, n-1\} : c[h] = \min_{i=0, \dots, n-1} c[i]$ ;
6   if  $\forall i \in \{0, \dots, m-1\} \Rightarrow A[0..m-1][h] \leq 0$  then
7     return problem UNBOUNDED;
8   end
9   Pick  $k \in \{0, \dots, m-1\} : c[k] = \min_{i=0, \dots, m-1} \left\{ \frac{b[i]}{A[i][h]} \wedge A[i][h] > 0 \right\}$ ;
10   $p \leftarrow A[k][h]$ ;
11   $\mathcal{B}[k] \leftarrow h$ ;
12   $R_p \leftarrow \text{copy}(T[k][0..n])$ ;
13   $C_p \leftarrow \text{copy}(T[0..m][h+1])$ ;
14  for  $i \leftarrow 0$  to  $m$  do
15    for  $j \leftarrow 0$  to  $n$  do
16      if  $i = k$  then
17         $T[i][j] = \frac{1}{p} T[i][j]$ ;
18      end
19      else
20         $T[i][j] = T[i][j] - \frac{C_p[i]}{p} R_p[j]$ ;
21      end
22    end
23  end
24 end
25  $x[0..n-1] \leftarrow 0$ ;
26 for  $i \leftarrow 0$  to  $m-1$  do
27    $x[\mathcal{B}[i]] \leftarrow b[i]$ ;
28 end
29 return  $(x, \mathcal{B}, T[0][n])$  ;      /*  $T[0][n]$  è il valore della funzione obiettivo */

```

---

**Complessità** Considerando che l'algoritmo non fa altro che esplorare le soluzioni di base ammissibili, l'algoritmo esplora al massimo  $O\binom{n+m}{m}$  (con  $n$  numero di variabili e  $m$  il numero di vincoli).

Nei casi peggiori il simplesso può raggiungere complessità esponenziali. Tuttavia nel corso degli anni si è visto EMPIRICAMENTE come questa complessità venga raggiunta in pochissimi casi specifici, anzi l'algoritmo si è rivelato molto efficiente su problemi “reali”, dove in media raggiunge una complessità nell'ordine di  $O(mn)$ , anche a confronto con altri algoritmi di soluzione (polinomiali per costruzione) come il metodo dei punti interni.

Da queste considerazioni si può concludere anche come la complessità del simplesso dipenda molto dall'istanza del problema fornito in input.

## Fase 1

In questa fase si costruisce il tableau e si verifica se il problema di partenza è ammissibile o meno. Per fare ciò si risolve il cosiddetto **problema ausiliario**

$$\begin{aligned} &\max (0|-1)^T x \\ &\text{subject to} \\ &(A|I|I)x = b \\ &x \geq 0 \end{aligned}$$

In pratica aggiungo al problema originario  $m$  nuove variabili (dette **variabili artificiali**) e cerco di azzerarne la somma negata. Se ciò avviene il problema di partenza è ammissibile e la soluzione di base ottenuta può essere utilizzata come base di partenza per la fase 2 dell'algoritmo, altrimenti il problema originale non è ammissibile. Se nella base finale una delle variabili artificiali è rimasta in base il problema è degenerare. Per lo pseudocodice si veda l'algoritmo 2.

## Fase 2

Terminata la prima fase si eliminano le variabili artificiali dal tableau e si sostituisce la funzione obiettivo con quella del problema originale, in seguito si procede a risolvere il problema originale. Per lo pseudocodice si veda l'algoritmo 3.

## Simplesso a due fasi

Si veda l'algoritmo 4.

# Scelte implementative

## Gestione della memoria

- I dati del problema originale sono memorizzati in memoria host, mentre il tableau viene costruito e memorizzato nella memoria globale della scheda video.
- I dati del problema originale vengono memorizzati utilizzando memoria *page-locked*, questo per permettere i trasferimenti di memoria in parallelo con i CUDA Stream (e quindi ottimizzare la fase di costruzione del tableau in global memory).
- Il vettore della base è memorizzato come memoria *mapped page-locked*, in modo da potervi accedere sia lato host che lato GPU in qualsiasi momento senza dover gestire i trasferimenti.
- Siccome l'algoritmo opera spesso su singole righe o colonne è fondamentale memorizzare i dati in memoria globale in modo da minimizzare il numero di accessi strided o disallineati. Per questo motivi si è deciso il seguente schema di memorizzazione (visibile in figura 2):

---

**Algorithm 2:** Simplex-Phase1

---

**Data:**  $A \in \mathbb{R}^{m \times n}$ ,  $b \in \mathbb{R}^m$  in forma standard (senza var. slack)

**Output:**  $T \in \mathbb{R}^{(n+m+1) \times (m+1)}$  tableau per la fase 2 e  $\mathcal{B} \in \mathbb{N}^m$  base ammissibile OPPURE problema INFEASIBLE o DEGENERARE

```
/* riempimento tableau */
1  $T \in \mathbb{R}^{(m+1) \times (n+2m+1)}$ ;
2  $T[0][0..n+m-1] \leftarrow 0$ ;
3  $T[0][n+m..n+2m-1] \leftarrow 1$ ;
4  $T[0][n+2m] \leftarrow 0$ ;
5  $T[1..m][n+2m] \leftarrow b$ ;
6  $T[1..m][0..n-1] \leftarrow A$ ;
7  $T[1..m][n..n+m-1] \leftarrow I$ ; /* variabili slack */
8  $T[1..m][n+m..n+2m-1] \leftarrow I$ ; /* variabili artificiali */
/* Nego eventuali disequazioni con termine noto negativo (in modo da rendere la base
iniziale ammissibile) */
9 for  $i \leftarrow 1$  to  $m$  do
10   if  $T[i][n+2m] < 0$  then
11     for  $j \leftarrow 0$  to  $n+2m$  do
12        $T[i][j] = -T[i][j]$ ;
13     end
14   end
15 end
16  $\mathcal{B} \leftarrow [n+m, \dots, n+2m-1]$ ; /* base iniziale con le variabili artificiali */
/* esprimo la funzione obiettivo solo in funzione delle variabili non di base */
17 for  $i \leftarrow 1$  to  $m$  do
18   for  $j \leftarrow 0$  to  $n+2m$  do
19      $T[0][j] \leftarrow T[0][j] - T[0][\mathcal{B}[i]] * T[i][j]$ 
20   end
21 end
22  $(x, \mathcal{B}, f) \leftarrow \text{Simplex}(T, \mathcal{B})$ ;
23 if  $f < 0$  then
24   return problem INFEASIBLE;
25 end
26 if  $\exists i : n+m \leq \mathcal{B}[i] < n+2m$  then
27   return problem DEGENERARE;
28 end
/* Elimino le variabili artificiali */
29  $T[0..m][n+m] \leftarrow \text{copy}(T[0..m][n+2m])$ ;
30 drop( $T[0..m][n+m+1..n+2m]$ );
31 return  $(T, \mathcal{B})$ 
```

---

---

**Algorithm 3: Simplex-Phase2**

---

**Data:**  $T \in \mathbb{R}^{(m+1) \times (n+m+1)}, \mathcal{B} \in \mathbb{N}^m, c \in \mathbb{R}^n$ **Output:** soluzione al problema iniziale OPPURE problema UNBOUNDED

```
1  $T[0][0..n-1] \leftarrow -c;$ 
2  $T[0][n..n+m] \leftarrow 0;$ 
   /* esprimo la funzione obiettivo solo in funzione delle variabili non di base */
3 for  $i \leftarrow 1$  to  $m$  do
4   for  $j \leftarrow 0$  to  $n+m$  do
5      $T[0][j] \leftarrow T[0][j] - T[0][\mathcal{B}[i]] * T[i][j]$ 
6   end
7 end
8 return  $\text{Simplex}(T, \mathcal{B})$ 
```

---

---

**Algorithm 4: 2Phases-Simplex**

---

**Data:**  $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, c \in \mathbb{R}^n$ **Output:** soluzione al problema OPPURE problema INFEASIBLE o UNBOUNDED o DEGENERERE

```
1  $(T, \mathcal{B}) \leftarrow \text{Simplex} - \text{Phase1}(A, b);$ 
2 return  $\text{Simplex} - \text{Phase2}(T, \mathcal{B}, c)$ 
```

---

- la matrice del tableau è memorizzata linearizzata per **colonne**. In altre parole l’implementazione utilizza il tableau trasposto, ossia le colonne del tableau sono i vincoli del problema, mentre le righe sono le variabili. L’algoritmo, infatti, tende di più ad accedere ai vettori delle singole variabili che a quelli dei singoli vincoli. Per copiare singole colonne della matrice in vettori in global memory si è visto che l’approccio migliore (dal punto di vista del tempo necessario ad eseguire l’operazione) è quello di utilizzare comunque gli accessi strided, invece di altri pattern di accesso più efficienti dal punto di vista della bandwidth utilizzata (Es. copiare un tile della matrice in shared memory e poi copiare la colonna desiderata dalla shared memory in global memory);
- il vettore dei termini noti non viene memorizzato in fondo al tableau, ma nella prima riga della matrice. In questo modo nel passaggio dalla fase 1 alla fase 2 non è necessario spostare la colonna nella matrice, ma è sufficiente troncare il numero di righe senza dover intervenire sull’allocazione della memoria;
- il vettore dei costi non viene memorizzato nella matrice del tableau, ma in un vettore a parte e il valore della funzione obiettivo è nel primo elemento di questo vettore (analogamente a quanto fatto con il vettore dei termini noti al punto precedente).

## Schema di parallelizzazione

Per introdurre parallelismo si è deciso di parallelizzare le singole operazioni svolte sui vettori e sulle matrici, mantenendo intatta la struttura dell’algoritmo.

Così facendo non si prevede una riduzione della complessità teorica dell’algoritmo, ma ottenere un miglioramento empirico sul tempo necessario a risolvere un’istanza rispetto a un’algoritmo puramente seriale.

### Creazione tableau (Algoritmo 2, righe 1-16)

Il tableau viene creato a partire dai dati del problema (matrice dei vincoli già linearizzata per colonne, vettore dei termini noti e dei coefficienti della funzione obiettivo). Per prima cosa si istanziano in global memory:

- il vettore che memorizza la matrice dei vincoli e il vettore dei termini noti con *cudaMallocPitch* di dimensione  $m \times (n + 2m + 1)$ . La larghezza reale della matrice (in byte) in memoria viene salvata nel campo *pitch* di *tabular\_t*. Per comodità si salvano in due campi distinti di *tabular\_t* i puntatori al

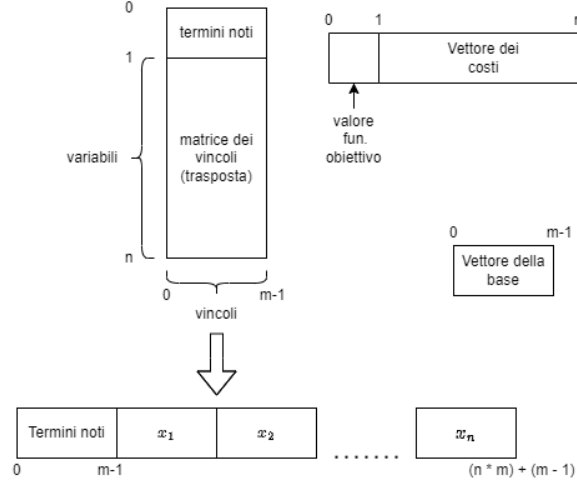


Figure 2: Schema di memorizzazione in memoria globale

vettore dei termini noti (lo stesso della matrice) e quello alla matrice dei vincoli (la seconda riga della matrice);

- il vettore dei costi di dimensione  $n + 2m + 1$  con *cudaMalloc*
- il vettore della base di dimensione  $m$  utilizzando *cudaHostAlloc* con il flag *cudaHostAllocMapped*

Successivamente si procede a riempire questi vettori. Per fare ciò si utilizzano sei *cudaStream* differenti:

1. il primo stream imposta a 0 i primi  $n + m + 1$  valori del vettore dei costi a 0 con *cudaMemset*,
2. il secondo stream imposta a 1 i restanti  $m$  valori della funzione dei costi (utilizzando un kernel apposito),
3. il terzo stream copia la matrice dei vincoli dalla memoria host alla matrice in global memory con *cudaMemcpy2D*,
4. il quarto stream costruisce in coda alla matrice dei vincoli le due matrici di identità, una per le variabili slack e una per le variabili artificiali, utilizzando un kernel apposito,
5. il quinto stream copia il vettore degli indicatori dalla memoria host alla global memory con *cudaMemcpy*,
6. il sesto stream inizializza il vettore della base con numeri progressivi da  $n + m$  a  $n + 2m - 1$  con un kernel apposito.

Una volta terminata questa prima fase si procede a scandire il vettore dei termini noti per verificare che tutti i valori siano maggiori o uguali a 0 (in modo che la base iniziale sia ammissibile). Se un elemento è negativo viene lanciato un nuovo kernel (tramite parallelismo dinamico) che inverte i segni a tutti gli elementi della colonna corrispondente. Questo equivale a moltiplicare l'equazione della colonna per -1.

### Aggiornamento funzione obiettivo (Algoritmo 2, righe 17-21 e Algoritmo 3, righe 3-7)

Lo scopo di questa operazione è quello di azzerare tutti gli elementi del vettore dei costi in corrispondenza delle variabili in base (una variabile è in base se nel suo vettore è presente un solo elemento a 1 e il resto è a 0).

Per fare ciò si effettuano una serie di eliminazioni di Gauss sottraendo al vettore dei costi tutte le righe della matrice opportunamente moltiplicate per il valore della  $i$ -esima variabile di base nel vettore dei costi.

Per fare questa operazione sono stati provati due approcci:

-



### Verifica di degenerazione (Algoritmo 2, riga 26)

Per verificare se il problema è degenerare si scandisce il vettore della base e si verifica se l'elemento in questione sia compreso tra  $n + m$  e  $n + 2m$  (non compreso). Se sì, si incrementa di 1 (con *atomicAdd*) un contatore in global memory e una volta terminata la scansione si controlla se il valore di questo contatore sia maggiore di 0.

### Sostituzione vettore dei costi (Algoritmo 3, righe 1-2)

Ricerca dell'elemento minimo nel vettore dei costi e in quello degli indicatori (Algoritmo 1, righe 4-5 e 9)

Verifica di unbounding (Algoritmo 1, riga 6)

Aggiornamento tableau (Algoritmo 1, righe 12-24)

Considerazioni sulle operazioni in virgola mobile

Estrazione della soluzione (Algoritmo 1, righe 25-29)

## Risultati sperimentali

Al fine di analizzare le prestazioni del solver è stata predisposta una funzione di benchmark con dati generati da un generatore casuale di istanze.

Il benchmark consiste nella risoluzione di una serie di problemi con dimensione crescente a partire da 256 variabili per 256 vincoli a 8192 variabili per 8192 vincoli. Il passo tra un benchmark e l'altro è esponenziale sulle potenze di due.

Per i benchmark il generatore utilizza come seed la dimensione del problema, per cui si andranno a generare ogni volta gli stessi problemi. In questo modo è possibile rieseguire il benchmark su dispositivi diversi e poter confrontare le prestazioni.

### Misurazione dei tempi

Per misurare i tempi sono stati utilizzati gli eventi di CUDA così come visto a lezione e sulla guida ufficiale di NVIDIA. Nonostante il metodo *cudaEventElapsedTime* misuri i tempi in ms, per evitare numeri eccessivamente piccoli (soprattutto sulle istanze più piccole) i tempi dei benchmark sono espressi in  $\mu s$  (ottenuti moltiplicando il valore della misurazione per 1000).

Come già discusso nel paragrafo 29 sulla complessità del simplesso, non ha molto senso misurare il tempo globale dell'algoritmo, per cui si è deciso di misurare i tempi delle singole operazioni che lo compongono, in particolare il tempo necessario a svolgere una singola iterazione del ciclo while da riga 4 a riga 24 dell'algoritmo 1, che d'ora in poi chiameremo **ciclo di solve**.