フローサイトメトリーデータ解析ための方向制限付きアース ムーバー距離の効率的な計算法

奥 牧人1,a)

概要:本稿では、フローサイトメトリーデータ解析のための方向制限付きアースムーバー距離 (一般に $EMD-L_1$ と呼ばれている) の効率的な計算法を紹介する.この距離は一方の確率分布をもう一方の確率分布へ変形するために必要な最小コストを表す.ただし,確率分布はいずれも 2 次元とし,確率質量の移動が特定の方向に制限されているものとする.方向制限付きアースムーバー距離は小さな問題を除いて一般のアースムーバー距離よりも速く計算出来る.また,分布変形に関する流れを可視化することで各細胞集団がどう変化したかを推測するのに役立つ可能性もある.

キーワード:フローサイトメトリー,アースムーバー距離,Wasserstein 距離,2次元分布

An efficient method for calculating earth mover's distance with direction restriction for flow cytometry data analysis

MAKITO OKU^{1,a)}

Abstract: In this paper, I introduce an efficient method for calculating earth mover's distance with direction restriction (generally called $EMD-L_1$) for flow cytometry data analysis. The distance represents the minimum cost required for transforming one probability distribution to the other, in which both probability distributions are two-dimensional, and movement of probability mass is restricted to certain directions. The earth mover's distance with direction restriction can be calculated faster than general earth mover's distance except for small problems. Moreover, visualization of the flow associated with the distribution transformation may be useful to guess how each cell population has changed.

Keywords: flow cytometry, earth mover's distance, Wasserstein distance, two-dimensional distribution

1. はじめに

フローサイトメトリー [1] とは、細胞などの粒子を流体中に分散させ、それを細い管の中に流し、1つ1つの粒子を光学的に測定する手法である。粒子にレーザーを当てて前方散乱光 (forward scatter, FSC) や側方散乱光 (side scatter, SSC) を測定するのに加え、予め蛍光標識しておいた数種類の蛋白質の量をそれぞれ対応する蛍光波長のパルス面積(以降では強度と呼ぶ)により測定することが出来る。FSC

の強度は粒子の相対的な大きさを、SSCの強度は粒子の構造的な複雑性をそれぞれ表す [1]. 得られた測定データを解析することで、どのようなタイプの細胞がどれだけ沢山いるのかを知ることが出来る。蛍光強度はそのままでは解釈しづらいため、通常は何らかのスケール変換を行う [2].

FSC 強度などの各変数の値の範囲を指定して特定の細胞集団のデータのみを抽出することをゲーティングと呼ぶ. 一度の抽出で使用する変数は通常は1つか2つであり,使用する変数を変えつつ抽出を複数回繰り返すことで,最終的に目的の細胞集団のデータを得ることが出来る. ゲーティングはフローサイトメトリーのデータ解析において一般的に用いられる.

¹ 富山大学 和漢医薬学総合研究所 Institute of Natural Medicine, University of Toyama, Toyama 930-0194, Japan

a) oku@inm.u-toyama.ac.jp

一方,複数の細胞集団が混在したデータ同士を直接比較する方法もある.本稿では,文献 [3] でフローサイトメトリーデータへ適用されたアースムーバー距離 (earth mover's distance, EMD) [4] に着目する. EMD とは何かについては次の節で説明する.

EMD は計算に時間がかかる。そこで本稿では、一般に EMD- L_1 [5] と呼ばれている方向制限付きの EMD を考え、それをフローサイトメトリーデータ解析において効率的に 計算する方法を紹介する。

2. アースムーバー距離

EMD は 2 つの分布の間の距離であり、それらが異なるほど大きな値をとる。本稿では 2 つ分布はいずれも総和が1 の確率分布と仮定する。その場合、EMD は一方の分布をもう一方の分布に変形するために必要なコストの最小値である (図 1)。そのコストは移動する量と移動距離の積 (仕事)の総和を表す。EMD 自体が距離であるため、コスト計算に使用する距離のことを区別のため地面距離 (ground distance) と呼ぶことがある。アースムーバーとはブルドーザーなどの土工機械のことであり、分布を変形する様子が土の運搬と似ているためそう呼ばれているようである。比較対象の分布は 2 次元以上であっても構わない。

比較対象の 2 つの分布 P, Q がいずれも確率分布の場合の EMD の定義を述べる。空間全体を有限個の点で表し、点i における P, Q の確率質量をそれぞれ p_i , q_i と表す。これらは非負であり $\sum_i p_i = \sum_i q_i = 1$ を満たす。また、点i からj へ移動する量を f_{ij} で表し、その距離を d_{ij} で表す。このとき P と Q の EMD は次の式で表される [3]:

$$EMD(P,Q) = \min_{\{f_{ij}\}} \sum_{i,j} f_{ij} d_{ij}$$

subject to $f_{ij} \ge 0$, $\sum_{j} f_{ij} = p_i$, $\sum_{j} f_{ji} = q_i$. (1)

これは中継輸送を認める次の式と等価である [6]:

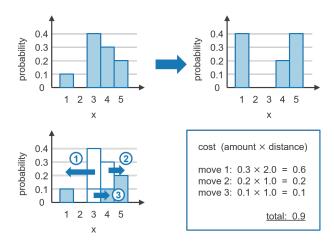


図 1 分布変形の例

Fig. 1 An example of distribution transformation

$$EMD(P,Q) = \min_{\{f_{ij}\} \setminus \{f_{ii}\}} \sum_{i,j,i \neq j} f_{ij} d_{ij}$$

subject to $f_{ij} \ge 0$, $\sum_{j,j \neq i} (f_{ij} - f_{ji}) = p_i - q_i$. (2)

ただし、式(2)を使う場合は冗長性を避けるためにいずれか1つの等式制約を予め取り除いておく必要がある.

 $P \ge Q$ が 1 次元の場合は、EMD はそれぞれの累積分布 関数の間の面積に等しいため、簡単に計算出来る。一方、P $\ge Q$ が 2 次元以上の場合は、一般には線形計画法などを用 いて式 (1) または (2) の最適化問題を解く必要がある [3].

EMD はしばしば1次の Wasserstein 距離と同じものして 扱われる [7]. しかし, EMD を扱った文献では式 (1) のような離散分布に基づく定義を用いるのに対し, Wasserstein 距離を扱った文献では連続確率分布に基づく定義を用いることが多いようである. Wasserstein 距離は 2017 年に敵対的生成ネットワークへ応用された [7] ことをきっかけに近年注目されている.

3. 方向制限付きアースムーバー距離

本稿では、一般に EMD- L_1 [5] と呼ばれている、分布変形において確率質量の移動方向が制限されている場合の EMD を考え、EMD-DR (EMD with direction restriction) と呼ぶことにする。特に、4 方向制限の場合の EMD-DR は、地面距離としてユークリッド距離の代わりにマンハッタン距離 (L_1 距離)を用いた場合の EMD と本質的に等価である。これが EMD- L_1 という呼称の由来と考えられる。しかし、8 方向制限など他の場合には EMD-DR は L_1 距離と関係が無いため、本稿では呼び方を変えている。

EMD-DR の計算法について説明する。まず,2次元平面に埋め込まれた有向グラフG=(V,E)を用意する(図 2)。ここでVは頂点集合,Eは枝集合を表す。そして,比較を行う2つのデータ点集合それぞれについて,Gの各頂点における確率密度をカーネル密度推定,ヒストグラム,k 近傍法などを用いて推定し,それらの総和で割って確率質量と見なせるようにしておく。頂点iにおける2つの分布の確率質量を p_i , q_i とする。

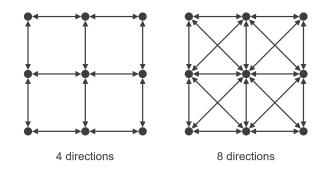


図 2 グラフ G の例. 黒丸と矢印は頂点と枝をそれぞれ表す.

Fig. 2 Examples of graph G. Black circles and arrows show vertices and edges, respectively.

IPSJ SIG Technical Report

次に,グラフGが頂点iからjへの枝を持つ場合にのみ,その枝に沿った移動量 f_{ij} と距離 d_{ij} を考える.それらを式 (2) に当てはめたものを線形計画法により解く.詳細な式は付録 A.1 に示す.未知変数 f_{ij} の個数は枝の数 |E| と等しいが,方向制限の無い場合はその数が |V|(|V|-1) であるのに対し,4 方向制限の場合は 4|V| 未満,8 方向制限の場合は 8|V| 未満と少なくなっている.このことが EMD-DR が EMD よりも高速に計算出来る主な理由と考えられている [5].

最後に,得られた最小値に補正係数 $(4 \, \mathrm{ End} \, \mathrm { End}$

得られた変数値 f_{ij} を用いて分布変形に関する流れを簡単に計算することも出来る. 具体的には、各頂点でそれと接続する枝における流れベクトルを平均することで、その頂点における流れベクトルが得られる. ただし、そのようにして得られたベクトル場はあくまで変形の過程で各頂点を通過した確率質量の総量および平均方向を表すものであり、そのベクトル場が一定期間持続しそれに従って確率質量が動くという意味ではないことに注意する.

4. EMDとEMD-DRの比較

本節では、地面距離としてユークリッド距離を用いた場合の EMD と、EMD-DR との比較について述べる。 図 2 に示す 4 方向と 8 方向の場合をそれぞれ EMD-4DR、EMD-8DR と呼ぶことにする.

比較条件について述べる. グラフ G の頂点は $n \times n$ の正 方格子状に配置し,その正方格子全体の一辺の長さは 1 とした. 比較対象の確率分布 P と Q は中点変位法 [8] によりランダムに生成した. 中点変位法を用いた理由は,頂点数が大きくなっても非局所的な輸送が必要な例を生成するためである. EMD の計算には python の pyemd パッケージ [9,10] を用いた. 他にも汎用の線形計画法のソルバーを複数試したが,pyemd より遅かったため,それらの結果は省略する. EMD-DR の計算には python の cvxopt パッケージを用いた.

図 3 左上と右上に EMD と EMD-DR の値の散布図を示す (100 回試行、n=10). EMD-4DR と EMD-8DR はいずれも EMD と高い相関を示し、特に EMD-8DR は EMD とほぼ同じ値であった.頂点数を変えた場合も同様の結果が得られた.図 3 下に EMD と EMD-DR の計算時間を示す.EMD-4DR と EMD-8DR の計算はいずれも n=15以上で EMD の計算よりも速かった.また,n=40 では EMD-4DR が EMD の平均 500 倍以上,EMD-8DR が EMD の平均 200 倍以上高速に計算出来た.

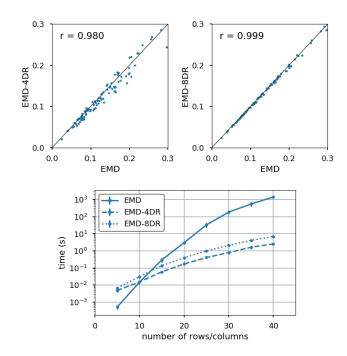


図 3 値 (左上と右上) と計算時間 (下) に関する EMD と EMD-DR の比較. 下図のエラーバーは 10 回試行の標準偏差を表す.

Fig. 3 Comparison between EMD and EMD-DR with respect to their values (top left and top right) and computational time (bottom). Error bars in the bottom panel show the standard deviation obtained from 10 trials.

適用例

5.1 使用データ

以降では EMD-DR を FlowRepository [11] で公開されている登録番号 FR-FCM-ZZCB のデータに適用した例について述べる. このデータを選んだ理由は, FlowRepositoryで "flow cytometry" と検索した際に一番上に表示されたためである. 同データは文献 [12] の Figure 4 に対応する.

このデータは、ヒトの末梢血単核細胞 (血液を遠心分離機にかけ、リンパ球などがいる層を取り出したもの) にサイトカイン産生を誘導する薬剤を添加し、0,1,5 時間後に測定したものである。産生されたサイトカインが細胞外へ分泌されるのを妨げるモネンシンという薬剤を添加した場合としない場合のデータがある。測定項目にはサイトカインの一種である IFN- γ (interferon-gamma) の mRNA と蛋白質の量が含まれる。

データの内訳は、サンプル毎の FCS (flow cytometry standard) 形式のファイルと、xml 形式の設定ファイルが1つである。FCS ファイルはテキスト部とバイナリ部が合わさったものだが、python の fcsparser パッケージを用いて読み込んだ。また、xml ファイルは BD FACSDiva というソフトウェア用のものだが、ウェブブラウザで開いて内容を読み、必要に応じて python を用いて情報を抽出した。

5.2 蛍光補正

蛍光補正とは、2種類以上の蛍光色素を用いた場合に生じる蛍光漏れ込みを補正することである。蛍光漏れ込みとは、ある蛍光色素の発する蛍光が別の蛍光色素用の検出器に捉えらえてしまうことである。今回使用したデータは蛍光補正前なのか補正後なのかが分からなかったが、FCSファイルのテキスト部に書かれていた蛍光漏れ込みの割合はどの蛍光色素の組に関してもほぼ0であったため、そのまま扱った。データが補正前だと仮定して補正を行った場合も結果はほとんど変わらなかった。

5.3 ゲーティング

ゲーティングは設定ファイルに書かれていた通りに行った. 凹多角形を用いたゲートは複数の凸多角形に分割して処理した.

まず最初に、FSC-AとSSC-Aに関してリンパ球のデータを抽出した(図4左). Aは area の頭文字であり、パルス面積を意味する. この処理はリンパ球以外の粒子 (細胞破片、単球、顆粒球など)のデータを取り除くために行うものである. 細胞破片はリンパ球より小さく、単球はリンパ球より大きいので、主にFSC強度によりリンパ球と区別出来る. 顆粒球はリンパ球より構造的複雑性が高いので主にSSC強度によりリンパ球と区別出来る.

次に、FSC-AとFSC-Hに関して一粒子データを抽出した (図 4 右). H は height の頭文字であり、パルスの高さを意味する. この処理は、パルスの高さに対して面積が極端に大きい場合は、一度に 2 つ以上の粒子が流路を通過してしまったと考えられるため、そのようなデータを取り除くために行うものである.

5.4 スケール変換

蛍光強度で表される IFN- γ の mRNA と蛋白質の量はスケール変換を行った。その詳細について設定ファイルには書かれていなかったが、おそらく文献 [12] の該当図は BD FACSDiva に実装されている biexponential scaling [2] を使ったものと推察される。しかし、biexponential scaling

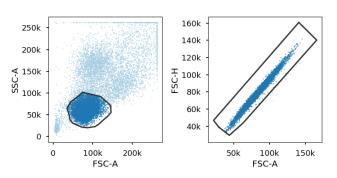


図 4 リンパ球 (左) と一粒子 (右) のゲーティング

Fig. 4 Gating of lymphocytes (left) and singlets (right)

はパラメータ数が多く,それらの値が設定ファイルに書かれていなかったため,代わりにフローサイトメトリーデータのスケール変換でしばしば用いられる $\arcsinh(ax)$ を用いた.パラメータの値は文献 [12] の該当図となるべく近くなるよう調節し,a=0.01 とした.描画範囲は変換前の値で -150 から $2^{18}-1$ とした.この $2^{18}-1=262$ 143 という値は今回扱った FCS 形式のデータの最大値である.

図5にスケール変換後の散布図を示す. 文献 [12] の該当図と完全には一致しないものの, よく似た分布が得られた.

5.5 EMD-DR の計算結果

表1に図5のサンプル間のEMD-DRを示す.移動は8方向とした.グラフの頂点は50×50の正方格子状に配置し,確率分布の推定には等方ガウス関数によるカーネル密度推定を用いた.モネンシン無しよりもモネンシン有りの方が,異なる時間のサンプル間のEMD-DRが大きかった.また,同じ時間のモネンシン無しとモネンシン有りのサンプル間のEMD-DRは薬剤添加5時間後が最も大きかった.

5.6 流れの可視化

図6に,薬剤添加5時間後のモネンシン無しと有りのサンプル間の分布変形に関する流れを示す.全体的に上向きの移動が多いことが分かる.この結果は,モネンシンがIFN-γの細胞外分泌を抑制したため細胞内の蛋白質量が増えたことを示唆する.

6. おわりに

本稿では、フローサイトメトリーデータを対象とした EMD-DR の効率的な計算法を紹介し、その実データへの 適用例について述べた.この計算法は高速である (図 3 下) だけでなく、分布変形に関する流れ (図 6) も簡単に計算出 来るという利点があった。EMD-DR を使えば、図 5 のように性質の異なる複数の細胞集団が混在したデータ同士を 追加のゲーティング無しに簡単に比較することが出来る.加えて、分布変形に関する流れの可視化により、各細胞集団がどう変化したかを推測するのに役立つ可能性もある.

表 **1** 図 5 のサンプル間の EMD-DR

Table 1 EMD-DR between samples shown in Fig. 5

sample 1	sample 2	EMD-DR
0 h without monensin	1 h without monensin	1.05
1 h without monensin	5 h without monensin	0.86
0 h without monensin	5 h without monensin	1.48
0 h with monensin	1 h with monensin	1.21
1 h with monensin	5 h with monensin	1.93
0 h with monensin	5 h with monensin	2.42
0 h without monensin	0 h with monensin	0.27
1 h without monensin	1 h with monensin	0.35
5 h without monensin	5 h with monensin	1.31

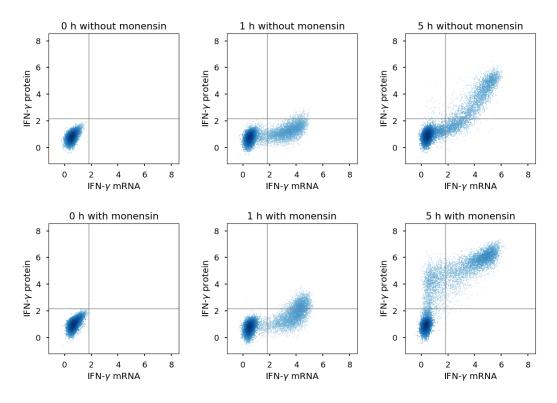


図 5 文献 [12] の Figure 4 に対応する散布図. ここではスケール変換に $\arcsinh(ax)$ を用いた.

Fig. 5 Scatter plots corresponding to Figure 4 in reference [12]. Here, arsinh(ax) was used for scaling.

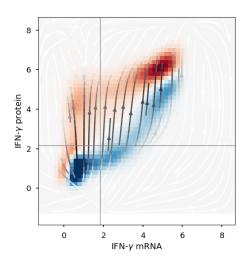


図 6 薬剤添加 5 時間後のモネンシン無しと有りのサンプル間の分 布変形に関する流れ

Fig. 6 Flow associated with the distribution transformation between the sample at 5 h without monensin and the sample at 5 h with monensin

さらに、もしも特定の疾患に関連した何らかの特徴的な分布形状の変化を捉えることが出来れば、それを新しいバイオマーカーとして利用出来るかもしれない.

EMD-DR はフローサイトメトリーデータだけでなくマスサイトメトリーや 1 細胞 RNA-seq によって計測されたデータにも適用出来る可能性がある。また,t-distributed

Stochastic Neighbor Embedding (t-SNE) [13] によって 2 次元に埋め込まれたデータにも適用出来るかもしれない. 詳細な検討は今後の課題とする.

EMD-DR は規則格子だけでなくデータの密度に応じて 頂点を配置したグラフなどにも使える可能性がある.ま た,3次元以上の分布にも適用出来るかもしれない.詳細 な検討は今後の課題とする.

EMD-DR をさらに高速化するには、仕事 (移動量と移動距離の積) の総和の代わりにそれらの 2 乗和を最小化するなどの近似が考えられる。移動距離だけでなく移動量も 2 乗するので、Wasserstein 距離で p=2 とした場合とは異なる。この場合の流れを可視化すると磁力線のようになる。詳細な検討は今後の課題とする。

謝辞 本研究は JSPS 科研費 JP15H05707 の助成を受けたものである.

参考文献

- Picot, J., Guerin, C. L., Le Van Kim, C. and Boulanger, C. M.: Flow cytometry: Retrospective, fundamentals and recent instrumentation, *Cytotechnology*, Vol. 64, No. 2, pp. 109–130 (online), DOI: 10.1007/s10616-011-9415-0 (2012).
- [2] Parks, D. R., Roederer, M. and Moore, W. A.: A new "Logicle" display method avoids deceptive effects of logarithmic scaling for low signals and compensated data, Cytometry A, Vol. 69, No. 6, pp. 541–551 (online), DOI:

IPSJ SIG Technical Report

10.1002/cyto.a.20258 (2006).

- [3] Orlova, D. Y., Zimmerman, N., Meehan, S., Meehan, C., Waters, J., Ghosn, E. E. B., Filatenkov, A., Kolyagin, G. A., Gernez, Y., Tsuda, S., Moore, W., Moss, R. B., Herzenberg, L. A. and Walther, G.: Earth mover's distance (EMD): A true metric for comparing biomarker expression levels in cell populations, *PLoS ONE*, Vol. 11, No. 3, p. e0151859 (online), DOI: 10.1371/journal.pone.0151859 (2016).
- [4] Rubner, Y., Tomasi, C. and Guibas, L. J.: The earth mover's distance as a metric for image retrieval, *Int. J. Comput. Vis.*, Vol. 40, No. 2, pp. 99–121 (online), DOI: 10.1023/A:1026543900054 (2000).
- [5] Ling, H. and Okada, K.: An efficient earth mover's distance algorithm for robust histogram comparison, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 29, No. 5, pp. 840–853 (online), DOI: 10.1109/TPAMI.2007.1058 (2007).
- [6] Essid, M. and Solomon, J.: Quadratically regularized optimal transport on graphs, SIAM J. Sci. Comput., Vol. 40, No. 4, pp. A1961–A1986 (online), DOI: 10.1137/17M1132665 (2018).
- [7] Arjovsky, M., Chintala, S. and Bottou, L.: Wasserstein generative adversarial networks, *ICML*, pp. 214– 223 (2017).
- [8] Fournier, A., Fussell, D. and Carpenter, L.: Computer rendering of stochastic models, Comm. ACM, Vol. 25, No. 6, pp. 371–384 (online), DOI: 10.1145/358523.358553 (1982).
- [9] Pele, O. and Werman, M.: A linear time histogram metric for improved SIFT matching, Lect. Notes Comput. Sci., Vol. 5304, pp. 495–508 (online), DOI: 10.1007/978-3-540-88690-7_37 (2008).
- [10] Pele, O. and Werman, M.: Fast and robust earth mover's distances, *Proc. ICCV'09*, pp. 460–467 (online), DOI: 10.1109/ICCV.2009.5459199 (2009).
- [11] Spidlen, J., Breuer, K., Rosenberg, C., Kotecha, N. and Brinkman, R. R.: FlowRepository: A resource of annotated flow cytometry datasets associated with peerreviewed publications, *Cytometry A*, Vol. 81, No. 9, pp. 727–731 (online), DOI: 10.1002/cyto.a.22106 (2012).
- [12] Van Hoof, D., Lomas, W., Hanley, M. B. and Park, E.: Simultaneous flow cytometric analysis of IFN- γ and CD4 mRNA and protein expression kinetics in human peripheral blood mononuclear cells during activation, Cytometry A, Vol. 85, No. 10, pp. 894–900 (online), DOI: 10.1002/cyto.a.22521 (2014).
- [13] van der Maaten, L. and Hinton, G.: Visualizing data using t-SNE, J. Machine Learn. Res., Vol. 9, pp. 2579–2605 (online), available from (http://www.jmlr.org/papers/v9/vandermaaten08a.html) (2008).

付 録

A.1 線形計画問題の詳細な式

2次元平面に埋め込まれた有向グラフG=(V,E)と、その各項点iにおける2つの分布の確率質量 p_i, q_i が与えられたとき、EMD-DRの値を線形計画法で求める方法を説明する.まず、頂点と枝にそれぞれ番号を振っておく、次に、行列 $A \in \mathbf{R}^{(|V|-1)\times |E|}$ 、ベクトル $b \in \mathbf{R}^{|V|-1}$ 、 $c \in \mathbf{R}^{|E|}$ を用意し、それらの各要素を次のように定める [6]:

$$A_{ik} = \begin{cases} 1 & \text{if } \operatorname{tail}(k) = i, \\ -1 & \text{if } \operatorname{head}(k) = i, \\ 0 & \text{otherwise,} \end{cases}$$
 (A.1)

$$b_i = p_i - q_i, \tag{A.2}$$

$$c_k = d_k. (A.3)$$

ここで tail(k) と head(k) は枝 k の始点と終点の頂点番号をそれぞれ表す。また、 d_k は枝 k の長さ、すなわちその両端の頂点間の地面距離を表す。冗長性を避けるため、一番最後の頂点に対応する A と b の要素は考えない。行列 A は各列の非零要素数が高々 2 のスパース行列である。

続いて,変数ベクトル $x \in \mathbf{R}^{|E|}$ を用意する.ここでx の各要素は式 (2) の f_{ij} に対応する.例えば,頂点i からj への枝が存在し,その枝番号がk だったとき, $x_k = f_{ij}$ である.分布変形に必要な最小コストおよびそれと対応する変数ベクトルx は以下の最適化問題を解くことによって求める:

$$\min_{x} c^{T} x \quad \text{subject to} \quad A x = b, \ x \geq 0. \tag{A.4} \label{eq:A.4}$$

なお,グラフGの2つの頂点を結ぶ最短経路は1通りとは限らないため,xは一意には定まらない.そのため,流れを可視化する際はソルバーによって結果が変わり得ることに注意しておく必要がある.解を一意にするため正則化項を加えるなどの方法も提案されている [6].