Multi-Feasibility Variable Selection*

!Optimizer

Sharif Optimization and Applications Laboratory
Department of Mathematical Sciences
Sharif University of Technology

February 2021

1 Motivation and Background

In inverse optimization models, one wants to learn the parameters of a family of optimization problems in a way that some desirable points would become optimal for their associated instances [YZ99]. In this contest, we consider the following class of parameterized feasibility *linear programs* (LP) extensively studied in the field of *constraint-based reconstruction and analysis* (COBRA)

$$\begin{array}{ll} \text{find} & v \\ \text{subject to} & S^I v = 0, \\ & l^I \preceq v \preceq u^I, \\ & I = \{i_1, i_2, \ldots, i_k\} \subseteq \{1, 2, \ldots, n\}. \end{array} \tag{1}$$

where $S = [S_1, S_2, \ldots, S_n]$ is an $m \times n$ stoichiometric matrix whose columns S_i characterize the stoichiometry of the biochemical reactions happening inside a cell,

$$l = \begin{bmatrix} l_1 \\ l_2 \\ \vdots \\ l_n \end{bmatrix} \quad \text{and} \quad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

Problem proposed for the *Optimizer Competition* 2021 on February 19, 2021. Please refer to this problem as:
[SOAL, *Multi-feasibility variable selection*, Optimizer Competition (2021), Sharif Optimization and Applications Laboratory, Department of Mathematical Sciences, Sharif University of Technology, February 2021, optimizer.math.sharif.edu].

are lower and upper bound vectors of dimension n as determined by the growth media capacity and thermodynamic constraints, while the subvectors

$$l^I = \begin{bmatrix} l_{i_1} \\ l_{i_2} \\ \vdots \\ l_{i_k} \end{bmatrix} \quad \text{and} \quad u^I = \begin{bmatrix} u_{i_1} \\ u_{i_2} \\ \vdots \\ u_{i_k} \end{bmatrix}$$

as well as the submatrix $S^I = [S_{i_1}, S_{i_2}, \ldots, S_{i_k}]$ contain the corresponding entries and columns, respectively.

A metabolic network comprises the collection of all chemical pathways of the metabolism of an organism, *e.g.*, a bacterium. In systems biology, it is observed that if we reconstruct a metabolic network by determining the set of indices I representing its reactions, whether the associated organism is viable or not under circumstances specified by l and u can, to some extent, be predicted by whether LP (1) is feasible or not [OTP10]. Moreover, numerous experiments have demonstrated the consistency of cell viability predictions derived by the feasibility of LP (1) with wet lab measurements [LNP12].

Turning this argument around, one may try to reconstruct a genome-scale metabolic network, given a list of viable and nonviable scenarios, by exploiting the fact that all the solutions to different instances of LP (1) have the same sparsity pattern because they belong to different strains of the same species [HFP06, KM09]. Therefore, one may select the candidate I by utilizing methods based on joint group sparsity. Other possible applications of this developed framework include, but are not limited to, signal processing [HB14], astrophysics [STA16], photoplethysmography [FB18, FB19], and inverse scattering problem [SKY18, SDX20].

2 Submission Logistics and Requirements

Let (L, U) and $(\tilde{L}, \tilde{U})$ denote two pairs of matrices such that $L \preceq U$ and $\tilde{L} \preceq \tilde{U}$ are satisfied, where the inequalities are understood entry-wise. Assume that a sparse matrix S, a constant λ, a positive integer K, and the matrices L and U are provided. Furthermore, the private matrices $\tilde{L}$ and $\tilde{U}$ are also predetermined but not available to you.

This year, the competition consists of five different rounds, each one with one question. Each round involves three increasingly large datasets in which the given instances of these matrices have tens, hundreds, and thousands of rows and columns. Each participating team can make as many submissions as they want for each dataset until the time for that round is up. At each time before the deadline, only the highest score for each dataset will be displayed for each team. Upon receiving new submissions achieving better results, the online leaderboard, visible to all participants, will be updated accordingly in a real-time way.

All the required input files in both `.mat` and `.asdf` formats will be made available via the official website according to the announced timeline and can be imported in several popular programming languages including Matlab, Python and Julia. Whenever submitting your best current solution, you should only upload the returned matrix V (vector v in the first two rounds), in either `.mat` or `.asdf` format, to the announced portal for the online judge system. Details on the timeline of data releases and submission deadlines are to be announced in the upcoming weeks.

For now, you can download the example data and output files from the following links to familiarize yourself with the formats. In this example submission, the vector v is optimal for the first two rounds and the matrix V is feasible for the remaining three ones. Note that this toy example does not necessarily represent the actual data in any manner, including the sparsity pattern, typical values for entries, etc.

example.mat   example.asdf   submission.mat   submission.asdf

3 Problem Formulation

Warm-up round: Steady-state flux distributions

The s_{ij} entry of S represents the molar rate of either consumption (if $s_{ij} \leq 0$) or production (if $s_{ij} \geq 0$) of the metabolite i in the reaction j per unit of dry cell weight. If all the metabolites are in mass balance at specific concentrations, *i.e.*, $Sv = 0$, we say that the metabolic network is in the steady-state condition.

Find the vector v of the rates of reactions subject to the constraints

$$Sv = 0, \qquad l^1 \preceq v \preceq u^1,$$

as predicted by *flux balance analysis* (FBA) [LGP06], *i.e.*,

$$\begin{array}{ll} \text{find} & v \\ \text{subject to} & Sv = 0, \\ & l^1 \preceq v \preceq u^1. \end{array}$$

Hint. This problem is an LP. You can learn more about LP here.

Scoring. If your solution is feasible, you get the full mark.

Novice optimizer round: Convex relaxation of cardinality optimization problems

Find the most sparse flux vector satisfying the constraints

$$Sv = 0, \qquad l^1 \preceq v \preceq u^1,$$

by minimizing $\|v\|_0$, *i.e.*,

$$\begin{array}{ll} \text{minimize} & \|v\|_0 \\ \text{subject to} & Sv = 0, \\ & l^1 \preceq v \preceq u^1. \end{array}$$

Hint. Read more about sparse recovery in the compressed sensing framework to learn about the relevant theoretical literature [CRT05, DON06]. Moreover, the biological intuition behind the theory is to minimize the total enzyme load imposed on the organism [MZ16].

Scoring. If your solution is feasible, then your final score is the number of zero entries of v.

Expert optimizer round: Exact multi-feasibility variable selection

Find the unknown matrix V with jointly sparse columns which satisfies the following constraints

$$SV = 0, \qquad L \preceq V \preceq U.$$

Joint sparsity for an arbitrary set of sparse vectors means that all members of the set share a common sparse support set, *i.e.*,

$$\begin{array}{ll} \text{minimize} & \|V\|_{2,0} \\ \text{subject to} & SV = 0, \\ & L \preceq V \preceq U, \end{array} \tag{2}$$

where the mixed norm is defined as follows

$$\|X\|_{p,q} = \|(\|x_1\|_p, \|x_2\|_p, \ldots, \|x_m\|_p)\|_q.$$

To the end of this paper, we assume that $x_1^T, x_2^T, \ldots, x_m^T$ are the rows of X in this definition, but some authors use another convention of considering the columns instead of rows. Apart from the difference in notation, the two definitions become clearly equivalent to one another if applied to the transpose matrix.

Hint. This non-convex problem can be formulated as an MILP [BNCM16]. However, consider minimizing the $l_{2,1}$ norm $\|V\|_{2,1}$ as a proxy cost function to derive a convex relaxation of (2) [SPP18]. This is a generalization of one suggested approach to *multiple-measurement vector* (MMV) problem [CRKK05].

Scoring. If your solution is feasible, then your final score is the number of zero rows of V.

Expert++ optimizer round: Multi-feasibility variable selection in the presence of error

Find the unknown matrix V with jointly sparse columns when the matrix SV is constrained to have jointly sparse rows and $L \preceq V \preceq U$, *i.e.*,

$$\begin{array}{ll} \text{minimize} & \left(\|V\|_{2,0}, \|(SV)^T\|_{2,0}\right) \\ \text{subject to} & L \preceq V \preceq U. \end{array}$$

Hint. This is a multicriterion optimization problem as stated [BV04]. One common technique to find Pareto optimal points is scalarization that is to minimize the scalar objective function $\|V\|_{2,1} + \lambda\|(SV)^T\|_{2,1}$ [KBW+17].

Scoring. If your solution is feasible, then your final score is the number of zero rows of V plus λ times the number of zero columns of SV.

The !Optimizer round: Multi-feasibility/infeasibility variable selection

Solve the previous task with the additional constraint that at most K columns of SV may have nonzero entries, *i.e.*,

$$\begin{array}{ll} \text{minimize} & \|V\|_{2,0} \\ \text{subject to} & \|(SV)^T\|_{2,0} \leq K, \\ & L \preceq V \preceq U. \end{array} \tag{3}$$

Suppose that $\tilde{L}$ and $\tilde{U}$ have t columns denoted by $\tilde{l}_1, \tilde{l}_2, \ldots, \tilde{l}_t$ and $\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_t$, respectively. Consider the following feasibility problems for $1 \leq k \leq t$

$$\begin{array}{ll} \text{find} & v \\ \text{subject to} & S^I v = 0, \\ & \tilde{l}_k^I \preceq v \preceq \tilde{u}_k^I, \end{array} \tag{4}$$

where I is defined as follows

$$I = \{i \mid \max_j |V_{ij}| > 0\}.$$

According to the biological model, these feasibility problems should be infeasible and we will validate each model by the percentage of infeasible instances. Note that, solving (3) helps to get a better score since we know *a priori* that the smaller the set of indices I, the higher the probability of infeasibility for each LP of the form (4).

Hint. Use your answer to the previous problem and find the minimum possible λ for which this additional condition holds.

Scoring. If your solution is feasible, then your final score is the percentage of infeasible instances. You cannot calculate your score locally and must submit the matrix V in order to get your score on the online leaderboard which is visible to everyone.

References

[BNCM16] Sébastien Bourguignon, Jordan Ninin, Hervé Carfantan, and Marcel Mongeau. Exact Sparse Approximation Problems via Mixed-Integer Programming: Formulations and Computational Performance. *IEEE Transactions on Signal Processing*, 64(6):1405–1419, March 2016.

[BV04] Stephen P Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[CRKK05] S. F. Cotter, B. D. Rao, Kjersti Engan, and K. Kreutz-Delgado. Sparse solutions to linear inverse problems with multiple measurement vectors. *IEEE Transactions on Signal Processing*, 53(7):2477–2488, July 2005.

[CRT05] E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59:1207–1223, 2005.

[DON06] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[FB18] S. H. Fouladi and I. Balasingham. Recovery of linearly mixed sparse sources from multiple measurement vectors using L1-minimization. In 26*th European Signal Processing Conference (EUSIPCO)*, pages 563–567, 2018.

[FB19] S. H. Fouladi and I. Balasingham. On improving recovery performance in multiple measurement vector having dependency. *IEEE Access*, 7:3287–3297, 2019.

[HB14] R. W. Heckel and H. Bölcskei. *Sparse Signal Processing: Subspace Clustering and System Identification*. Series in communication theory. Hartung-Gorre, 2014.

[HFP06] Markus J Herrgård, Stephen S Fong, and Bernhard Ø Palsson. Identification of genome-scale metabolic network models using experimentally measured flux profiles. *PLOS Computational Biology*, 2(7):1–11, 07 2006.

[KBW+17] A. Kazemipour, B. Babadi, M. Wu, K. Podgorski, and S. Druckmann. Multiplicative updates for optimization problems with dynamics. In *2017 51st Asilomar Conference on Signals, Systems, and Computers*, pages 2025–2029, 2017.

[KM09] Vinay Satish Kumar and Costas D. Maranas. Growmatch: An automated method for reconciling in silico/in vivo growth predictions. *PLOS Computational Biology*, 5(3):1–13, 03 2009.

[LGP06] Jong Min Lee, Erwin P. Gianchandani, and Jason A. Papin. Flux balance analysis in the era of metabolomics. *Briefings in Bioinformatics*, 7(2):140–150, 04 2006.

[LNP12] Nathan E Lewis, Harish Nagarajan, and Bernhard Ø Palsson. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nature Reviews Microbiology*, 10(4):291–305, 2012.

[MZ16] Costas D Maranas and Ali R Zomorrodi. *Optimization methods in metabolic networks*. John Wiley & Sons, 2016.

[OTP10] Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. What is flux balance analysis? *Nature Biotechnology*, 28(3):245–248, 2010.

[SDX20] S. Sun, D. Dai, and Z. Xu. Application of a GMMV-based multi-frequency inversion method to 3-d half-space inverse scattering problems. In *XXXIIIrd General Assembly and Scientific Symposium of the International Union of Radio Science*, pages 1–4, 2020.

[SKY18] S. Sun, B. J. Kooij, and A. G. Yarovoy. Extension of the GMMV-based linear method to quantitative inverse scattering. *IEEE Antennas and Wireless Propagation Letters*, 17(1):94–97, 2018.

[SPP18] C. Steffens, M. Pesavento, and M. E. Pfetsch. A compact formulation for the $\ell_{2,1}$ mixed-norm minimization problem. *IEEE Transactions on Signal Processing*, 66(6):1483–1497, 2018.

[STA16] Jean-Luc Starck. Sparsity and inverse problems in astrophysics. *Journal of Physics: Conference Series*, 699:012010, mar 2016.

[YZ99] C. Yang and J. Zhang. Two general methods for inverse optimization problems. *Applied Mathematics Letters*, 12(2):69 – 72, 1999.

!Optimizer

Sharif Optimization and Applications Laboratory,
Department of Mathematical Sciences,
Sharif University of Technology,
Tehran, Iran.
optimizer.math.sharif.edu