Discrete Optimization and Machine Learning
RIKEN Center for Advanced Intelligence Project
Tokyo, Japan, 29-31 July, 2019

Organizers: Antoine Deza, Sebastian Pokutta, Takanori Maehara

# ABSTRACT BOOKLET

## Contents

# Monday, July 29

## Session A1: 10h00-12h00 *Chair: Amitabh Basu*

A11. **A convex integer programming approach for optimal sparse PCA**
Santanu Dey (Georgia Tech)

Principal component analysis (PCA) is one of the most widely used dimensionality reduction tools in scientific data analysis. The PCA direction, given by the leading eigenvector of a covariance matrix, is a linear combination of all features with nonzero loadingsthis impedes interpretability. Sparse principal component analysis (SPCA) is a framework that enhances interpretability by incorporating an additional sparsity requirement in the feature weights (factor loadings) while finding a direction that explains the maximal variation in the data. However, unlike PCA, the optimization problem associated with the SPCA problem is NP-hard. While many heuristic algorithms based on variants of the power method are used to obtain good solutions, they do not provide certificates of optimality on the solution-quality via associated dual bounds. Dual bounds are available via standard semidefinite programming (SDP) based relaxations, which may not be tight and the SDPs are difficult to scale using off-the-shelf solvers. In this paper, we present a convex integer programming (IP) framework to solve the SPCA problem to near-optimality, with an emphasis on deriving associated dual bounds. We present worst-case results on the quality of the dual bound provided by the convex IP. We empirically observe that the dual bounds are significantly better than worst-case performance, and are superior to the SDP bounds on some real-life instances. Moreover, solving the convex IP model using commercial IP solvers appears to scale much better that solving the SDP-relaxation using commercial solvers. To the best of our knowledge, we obtain the best dual bounds for real and artificial instances for SPCA problems involving covariance matrices of size up to $2000 \times 2000$. This is co-authored by Rahul Mazumder and Guanyi Wang.

A12. **Subspace communication driven search for high dimensional optimization**
Logan Mathesen (Arizona State University)

Techniques for global optimization often suffer from the curse of dimensionality. In an attempt to face this challenge, high dimensional search techniques try to identify and leverage upon the effective, lower, dimensionality of the problem-either in the original or in a transformed space. As a result, algorithms can search for and exploit a projection or create a suitable random embedding. Our proposed approach avoids modeling of high dimensional spaces, and the assumption of low effective dimensionality. In fact, we argue that effectively high dimensional functions can be recursively optimized over sets of complementary lower dimensional subspaces that constitute a partition of the original space. In this light, we propose the novel Subspace COmmunication for OPtimization (SCOOP) algorithm, which enables intelligent information sharing among subspaces such that each subspace can guide the other towards improved locations. In particular, we present three methods, which differ for the level of trust in the shared information. The experiments show that the accuracy of

SCOOP rivals the state-of-the-art global optimization techniques, while being several orders of magnitude faster and having better scalability against the problem dimensionality.

A13. **A limiting analysis of regularization of SDP and its implication to infeasible interior-point algorithms**
Takashi Tsuchiya (Graduate Research Institute for Policy Studies)

Though the interior-point algorithm has been established a practical and standard way to solve SDPs, its convergence analysis assumes primal-dual strong feasibility as a standard assumption. Singular SDPs which do not satisfy this assumption often arises in real world applications. A typical and conventional approach to recover primal-dual standard feasibility is to add on the both sides the identity matrices multiplied by small positive numbers $t_P$ and $t_D$, say. In this talk, we conduct a limiting analysis of the perturbed system when $t_P$ and $t_D$ goes to zero, and apply it to an analysis of infeasible interior-point algorithms in the absence of strong feasibility. This is a joint work with Bruno F. Lourenço, Masakazu Muramatsu, and Takayuki Okuno.

A14. **Stochastic proximal methods for non-smooth non-convex constrained sparse optimization**
Michael Metel (RIKEN AIP)

This talk will focus on first-order stochastic gradient methods for solving non-smooth non-convex optimization problems with applications in machine learning. An overview of convergence rates will be presented with new results for optimizing a smooth non-convex loss function with a non-smooth non-convex regularizer and convex constraints.

## Session A2: 13h00-14h30 *Chair: Santanu Dey*

A21. **Locally accelerated conditional gradients**
Alejandro Carderera (Georgia Tech)

Conditional gradient methods form a class of projection-free first-order algorithms for solving smooth convex optimization problems. Apart from eschewing projections, these methods are attractive because of their simplicity, numerical performance, and the sparsity of the solutions outputted. However, they do not achieve optimal convergence rates for smooth convex and strongly convex functions. We present the Locally Accelerated Conditional Gradients algorithm that relaxes the projection-freeness requirement to only require projection onto (typically low-dimensional) simplices and mixes accelerated steps with conditional gradient steps to achieve local acceleration. We derive asymptotically optimal convergence rates for this algorithm. Our experimental results demonstrate the practicality of our approach; in particular, the speedup is achieved both in wall-clock time and per-iteration progress compared to standard conditional gradient methods and a Catalyst-accelerated Away-Step Frank-Wolfe algorithm.

A22. **No-regret algorithms for online $k$-submodular maximization**
Tasuku Soma (University of Tokyo)

We present a polynomial time algorithm for online maximization of $k$-submodular maximization. For online (non-monotone) $k$-submodular maximization, our algorithm achieves a tight approximate factor in the approximate regret. For online monotone $k$-submodular maximization, our approximate-regret matches to the best-known approximation ratio, which is tight asymptotically as $k$ tends to infinity. Our approach is based on the Blackwell approachability theorem and online linear optimization.

A23. **Parallel depth-first search and the applications to data mining**
Kazuki Yoshizoe (RIKEN AIP)

Large-scale parallelization of graph search is a challenging problem in general. Even for the Depth-First Search (DFS), which is one of the simplest graph search algorithms, parallel speedup rarely exceeds 100-fold when solving real-world problems. The main challenge is the difficulty of equally distributing the workloads to multiple processors with small communication overhead. We show that by combining recent work stealing techniques with old knowledge in parallel algorithms, the estimated parallel speedup of our parallel DFS reached 110K-fold or higher using up to 140K CPU cores on the supercomputer K, when solving two data mining problems that is Frequent Itemset Mining (also known as the "beer diaper problem") and Genetic Disease Analysis (Genome-Wide Association Studies).

## Session A3: 15h00-16h30 *Chair: Takanori Maehara*

A31. **alsification of cyber physical systems: a stochastic optimization perspective**
Giulia Pedrielli (Arizona State University)

Technological advancement, including Artificial Intelligence, Scientific Machine Learning, has led to the development of new classes of Systems that require humongous effort to be effectively controlled. In the area of control of Cyberphysical systems, stochastic optimization has started to attract increasing attention. Companies that are increasingly marketing CPSs (e.g., automotive, energy, medical) have impellent need of new methods that can give guarantees about the quality and safety of the devices. In fact, safety has become a matter of probabilistic guarantees, and the way to provide those in a short time is a critical research challenge. In this talk, we refer at large to Stochastic Optimization methods, where the focus is on families of algorithms that deliberately inject randomness in the search process (whether or not the original dynamics is stochastic). In this context, we develop methods to be applied in cases where, (1) there is no homogeneous dynamics of the systems, (2) we can construct clever approximations of the system behavior to facilitate the process. We show the improved performance of the proposed approaches with respect to state of the art Bayesian Optimization solvers. We also offer an analysis of the verification results for several applications ranging from automotive to aerospace.

A32. **Learning-algorithms from Bayesian principles**
Emtiyaz Khan (RIKEN AIP)

In machine learning, new learning algorithms are designed by borrowing ideas from optimization and statistics followed by an extensive empirical efforts to make them practical. However, there is a lack of underlying principles to guide this process. I will present a stochastic learning algorithm derived from Bayesian principle. Using this algorithm, we can obtain a range of existing algorithms: from classical methods such as least-squares, Newton's method, and Kalman filter to deep-learning algorithms such as RMSprop and Adam. Surprisingly, using the same principles, new algorithms can be naturally obtained even for the challenging learning tasks such as online learning, continual learning, and reinforcement learning. This talk will summarize recent works and outline future directions on how this principle can be used to make algorithms that mimic the learning behaviour of living beings.

A33. **Matrix co-completion for multi-label classification with missing features and labels**
Miao Xu (RIKEN AIP)

We consider a challenging multi-label classification problem where both feature matrix $X$ and label matrix $Y$ have missing entries. An existing method concatenated $X$ and $\sigma(Y)$ together as $[X, \sigma(Y)]$ where $\sigma(.)$ is a sigmoid function. Then under the assumption of low-rankness, a matrix completion (MC) method is applied to fill the missing entries. However, there is no theoretical guarantee for the recovery result of this method. In this talk, I will present an upper bound on the recovery error of the method. In deriving the bound, we found that adding another trace norm constraint on recovering $X$ will lead to a guaranteed recovery of the whole matrix. Such phenomenon coincides with Elastic Net where both $L_1$ norm and $L_2$ are used for regularization. The practical usefulness of the proposed method is demonstrated through experiments on both synthetic and benchmark data.

# Tuesday, July 30

## Session B1: 10h00-12h00 *Chair: Masashi Sugiyama*

B11. **Statistical decision theory perspectives on learning and stochastic optimization**
Amitabh Basu (Johns Hopkins University)

We look at stochastic optimization problems through the lens of statistical decision theory. In particular, we address admissibility, in the statistical decision theory sense, of the natural sample average estimator for a stochastic optimization problem (which is also known as the empirical risk minimization (ERM) rule in learning literature). It is well known that for general stochastic optimization problems, the sample average estimator may not be admissible. This is known as Stein's phenomenon in the statistics literature. We show in this paper that for optimizing stochastic linear or convex quadratic functions over compact sets, the sample average estimator is admissible.

B12. **On solving mixed integer non linear programs with separable non convexities**
Claudia D'Ambrosio (CNRS & Ecole Polytechnique)

In this talk, we focus on mixed integer non linear programming (MINLP) problems with non convexities that can be formulated as sums of univariate functions. D'Ambrosio et al. 2009 and D'Ambrosio et al. 2012 proposed a method called Sequential Convex MINLP (SC-MINLP), an iterative algorithm based on lower and upper bounds obtained by solving a convex MINLP and a non convex non linear program, respectively. The method aims at finding a global solution of the tackled MINLP and exploits the fact that the convex or concave parts of univariate functions can be identified numerically. The weaknesses of the original version of the SC-MINLP method are mainly two: on the one hand, solving several (one per iteration) convex MINLPs is time-consuming; on the other hand, at each iteration, the convex MINLP is modified to improve the lower bound and no information about the previous convex MINLP and its optimal solution is exploited. These two weaknesses are addressed in two recent works: in the first, a strengthening of the convex MINLP relaxation is proposed based on perspective reformulation. In the second, a disjunctive programming approach was explored to better approximate the concave parts of each univariate function. Extensive computational experiments show a significant speedup of the original SC-MINLP method.

B13. **Blended matching pursuit**
Cyrille Combettes (Georgia Tech)

Matching pursuit algorithms are an important class of algorithms in signal processing and machine learning. We present a blended matching pursuit algorithm, combining coordinate descent-like steps with stronger gradient descent steps, for minimizing a smooth convex function over a linear space spanned by a set of atoms. We derive sublinear to linear convergence rates according to the smoothness and sharpness orders of the function and demonstrate computational superiority of our approach. In particular, we derive linear rates for a wide class of non-strongly convex functions, and we demonstrate in experiments that our algorithm enjoys very fast rates of convergence and wall-clock speed while maintaining a sparsity of iterates very comparable to that of the (much slower) orthogonal matching pursuit.

B14. **Simulation based mixed integer programming**
Alexander Martin (Universität Erlangen-Nürnberg)

MIP techniques are often successfully applied for the solution of problems containing nonlinear, physical side constraints by remodeling them for instance via piecewise linear approximations and relaxations. We report on this success using the example of gas networks, but simultaneously ask the question whether we might be able to directly include the physics via an iterative call of existing simulation algorithms. We show first promising results in this direction and prove that this way the physics may even be incorporated more accurately than MIP techniques themselves might ever be able to do.

## Session B2: 13h00-14h30 *Chair: Edwin Romeijn*

B21. **Strong mixed-integer programming formulations for trained neural networks**
Joey Huchette (Massachusetts Institute of Technology)

We present strong mixed-integer programming (MIP) formulations for high-dimensional piecewise linear functions that correspond to trained neural networks. These formulations can be used for a number of important tasks, such as verifying that an image classification network is robust to adversarial inputs, or solving decision problems where the objective function is a machine learning model. We present a generic framework, which may be of independent interest, that provides a way to construct sharp or ideal formulations for the maximum of d affine functions over arbitrary polyhedral input domains. We apply this result to derive MIP formulations for a number of the most popular nonlinear operations (e.g. ReLU and max pooling) that are strictly stronger than other approaches from the literature. We corroborate this computationally, showing that our formulations are able to offer substantial improvements in solve time on verification tasks for image classification networks.

B22. **Airline schedule planning under infrastructure constraint and with customer choice evaluation**
Sébastien Deschamps (Ecole Nationale des Ponts et Chaussées)

The schedule planning problem aims at choosing the set of flight legs operated by an airline so as to maximize its revenue. The difficulty of this optimization problem is that it links two sub-problems, both challenging on their own. On one side, the planning must be operable with the assets of the airline. An ever growing constraint is the scarcity of slots in hubs in a context where air traffic grows faster than airport infrastructures. And on the other side, evaluating the revenue requires to model demand and customer choices on each origin-destination operated by the company, spill and recover, and decision of the airline revenue management to affect capacity between different itineraries sharing the same leg. Due to connections, the number of itineraries potentially operated is very large. Leveraging a logit model for customer choice, we propose a linear program for this second subproblem that models spill and recover and airline choice. And we integrate it in a mixed integer linear program for the complete schedule planing problem. This problem being challenging for present day MILP solvers, we propose a Benders decomposition approach to solve it.

B23. **Adding variables - speed up by including new binary variables**
Robert Hildebrand (Virginia Tech)

Although complexity theoretically speaking, fewer variables are better, we often see in practice that adding variables can improve a model and even solve faster. With the speed of branch and bound algorithms, adding binary variables can substantially improve solve times for many problems. We discuss applications of converting both integer and continuous variables to binary variables (exactly and approximately) and the theory behind these choices related to both cutting planes and neural networks.

## Session B3: 15h00-16h30 *Chair: Alexander Martin*

B31. **Causal-retro-causal Systems induce a dynamics on a manifold**
Hans-Georg Zimmermann (Fraunhofer Nürnberg)

Fraunhofer and Siemens work together on commodity price forecast models up to one year ahead (This effort is ongoing since 15 years). We have develop a neural network model that covers the causal parts of the economy together with a retro-causal part to describe goal driven human behavior. It shows that the combined causal-retro-causal model generates implicitly an underlying manifold on which such systems have to stay. The combined identification of manifold and dynamics with past data is a sophisticated task - even more demanding is the question to keep the dynamics on the manifold in the future without referring to target data. In the last years Siemens has done an extensive comparison study relative to other advisory companies showing that our system performed best.

B32. **Adaptive algorithm for finding connected dominating sets in uncertain graphs**
Takuro Fukunaga (Chuo University)

The problem of finding a minimum-weight connected dominating set (CDS) of a given undirected graph has been studied actively, motivated by operations of wireless ad hoc networks. In this talk, we discuss a new stochastic variant of the problem, where each node in the graph has a hidden random state, which represents whether the node is active or inactive, and we seek a CDS of the graph that consists of the active nodes. We consider adaptive algorithms for this problem, which repeat choosing nodes and observing the states of the nodes around the chosen nodes until a CDS is found. Our algorithms have a theoretical performance guarantee that the sum of the weights of the nodes chosen by the algorithm is at most $O(\alpha \log(1/\delta))$ times that of any adaptive algorithm in expectation, where $\alpha$ is an approximation factor for the node-weighted polymatroid Steiner tree problem and $\delta$ is the minimum probability of possible scenarios on the node states.

B33. **Computing full conformal prediction set with approximate homotopy**
Eugène Ndiaye (RIKEN AIP)

If you are predicting the label $y$ of a new object with $\hat{y}$, how confident are you that $y = \hat{y}$? Conformal prediction methods provide an elegant framework for answering such question by building a $100(1-\alpha)\%$ confidence region without assumptions on the distribution of the data. It is based on a refitting procedure that parses all the possibilities for $y$ to select the most likely ones. Although providing strong coverage guarantees, conformal set is impractical to compute exactly for many regression problems. We propose efficient algorithms to compute conformal prediction set using approximated solution of (convex) regularized empirical risk minimization. Our approaches rely on an homotopy continuation techniques for tracking the solution path with respect to sequential changes of the observations. We provide a detailed analysis quantifying its complexity.

# Wednesday, July 31

## Session C1: 10h00-12h00 *Chair: Claudia D'Ambrosio*

C11. **Sampling from** log **supermodular distribution**
Shuji Kijima (Kyushu University)

Log supermodular distribution is deeply related to topics in combinatorics such as submodular function minimization, Tutte polynomial and stable matching. In this talk, we remark some background of log supermodular distribution, and show that an approximate sampling is #BIS-hard even for a special class of log $M$ convex distribution.

C12. **Fair dimensionality reduction and iterative rounding for SDPs**
Uthaipon Tantipongpipat (Georgia Tech)

Dimensionality reduction is a classical technique widely used for data analysis. One foundational instantiation is Principal Component Analysis (PCA), which minimizes the average reconstruction error. In this paper, we introduce the "multi-criteria dimensionality reduction" problem where we are given multiple objectives that need to be optimized simultaneously. As an application, our model captures several fairness criteria for dimensionality reduction such as the Fair-PCA problem introduced by Samadi, et. al. 2018 and the Nash Social Welfare (NSW) problem. In the Fair-PCA problem, the input data is divided into $k$ groups, and the goal is to find a single d-dimensional representation for all groups for which the maximum reconstruction error of any one group is minimized. In NSW the goal is to maximize the product of the individual variances of the groups achieved by the common low-dimensional space. Our main result is an exact polynomial-time algorithm for the two-criteria dimensionality reduction problem when the two criteria are increasing concave functions. As an application of this result, we obtain a polynomial time algorithm for Fair-PCA for $k = 2$ groups, resolving an open problem of Samadi, et. al. 2018, and a polynomial time algorithm for NSW objective for $k = 2$ groups. We also give approximation algorithms for $k > 2$. Our technical contribution in the above results is to prove new low-rank properties of extreme point solutions to semi-definite programs. We conclude with experiments indicating the effectiveness of algorithms based on extreme point solutions of semi-definite programs on several real-world datasets.

C13. **Differentiable ranks using optimal transport: The Sinkhorn CDF and quantile operator**
Marco Cuturi (Google Brain)

We propose a framework to sort values that is algorithmically differentiable. We leverage the fact that sorting can be seen as a particular instance of the optimal transport (OT) problem on $\mathbb{R}$, from input values to a predefined array of sorted values (e.g. $1, 2, \ldots, n$ if the input array has $n$ elements). Building upon this link, we propose generalized ranks, CDFs and quantile operators by varying the size and weights of the target pre-sorted array. We

recover differentiable algorithms by adding to the OT problem an entropic regularization, and approximate it using a few Sinkhorn iterations. We call these operators $S$-ranks, $S$-CDFs and $S$-quantiles, and use them in various learning settings: we benchmark them against the recently proposed neuralsort [Grover et al. 2019], propose applications to quantile regression and introduce differentiable formulations of the top-$k$ accuracy that deliver state-of-the art performance.

C14. **Random projection for conic programming**
Pierre-Louis Poirion (RIKEN AIP)

We present a new random projection method that allows the reduction of the number of inequalities of a Linear Program (LP). More precisely, we randomly aggregate the constraints of a LP into a new one with fewer constraints, while approximately preserving the optimal value of the LP. We will also see how to extend this idea to conic programming.

## Session C2: 14h00-15h30 *Chair: Albert Heuberger*

C21. **Multi-armed bandit problem in piece-wise stationary environment**
Chi-Guhn Lee (University of Toronto)

We propose new variants of Thompson sampling for an extension of the multi-armed bandit (MAB) problem, called the piece-wise stationary MAB, in which the reward distributions for the arms are piece-wise stationary and will shift at some unknown time steps. Several algorithms have been proposed, but they show poor performance or require large memory and computation. Different from existing works, we consider distinctive stationary periods as states, which are only partially observable, and estimate the distribution of the states, called the belief state. The integration of Thompson sampling and belief update leads to two variants of Thompson sampling: Thompson sampling with Belief Update - Finite (TS-BU-Fin) and Thompson Sampling with Belief Update - Infinite (TS-BU-Inf). Extensive empirical studies show that TS-BU-Fin outperforms the state-of-the-art algorithms, and TS-BU-Inf compares similarly with Global-STS-BA and Global-CTS. It is worthwhile noting that both of the proposed algorithms have significant advantages in memory and computation required over Global-STS-BA and Global-CTS.

C22. **Policy transfer via greedy state recoding**
Christopher Mutschler (Universität Erlangen-Nürnberg)

End-to-end model-free reinforcement learning that uses raw sensory input achieves remarkable results. However, if the environment representation, i.e., the encoding of the observations, changes after the policy has been deployed we can no longer use this policy as we cannot correctly process the (new) observations. Approaches that break up the end-to-end pipeline, e.g. by encoding the environment with variational auto-encoders and training a policy on that representation, are also no option as we cannot guarantee that both latent

representations capture the generative factors in the same way. We propose to transfer the original policy by translating the environment representation. We sample observations from both the environment and a forward model that we may have acquired from executing the policy under the old representation. This generates pairs of states (new representation from the environment, (biased) old representation from the forward model) that allow to bootstrap a neural network model for state translation as the states are correlated through the forward model.

C23. **Stochastic monotone submodular maximization with queries**
Takanori Maehara (RIKEN AIP)

Stochastic optimization with queries is an optimization problem that has stochastically uncertain parameters, and we can reduce the uncertainty by conducting queries. The purpose of the problem is to find a query strategy such that after conducting the queries we can obtain a good solution. In this study, we consider a monotone submodular objective function for the first time, and propose a technique to find query strategies that have bounded degree of adaptivity and provable approximation factors. Using our technique, we derive query strategies for matching, k-exchange system, and knapsack constrained problems.

## Session C3: 16h00-17h30 *Chair: Joey Huchette*

C31. **Tensor network representation in machine learning**
Qibin Zhao (RIKEN AIP)

Tensor network is a promising technology for model compression and fast computation. In this talk, we will present our study on tensor network model and algorithms. Then, various applications to model representation and data representation are also presented. Finally, we will discuss other potential problems and applications of tensor network.

C32. **Machine learning approaches in brain correlates of dementia elucidation - tensor machine lLearning and beyond**
Tomasz Rutkowski (University of Tokyo / RIKEN AIP)

A proliferation of dementia cases in aging societies creates a remarkable economic as well as medical problems in many communities around the world. We present an attempt and exploratory results of a brain signal (EEG) classification to establish digital biomarkers for dementia stages' elucidation. We discuss a comparison of various machine learning approaches for event-related potential (ERPs) and brain micro-state pattern classification of healthy young and cognitively frail elderly subjects. We review developed by our team machine learning methods using tensor- and fully end-to-end deep learning-based models. The presented results are steps forward to develop AI-based approaches for a subsequent application for subjective and mildcognitive impairment (SCI and MCI) diagnostics.

C33. **Towards robust deep learning**
Bo Han (RIKEN AIP)

It is challenging to train deep neural networks robustly with noisy labels, as the capacity of deep neural networks is so high that they can totally overfit on these noisy labels. In this talk, I will introduce three orthogonal techniques in robust deep learning with noisy labels, namely data perspective estimating the noise transition matrix; training perspective training on selected samples; and regularization perspective conducting scaled stochastic gradient ascent. First, as an approximation of real-world corruption, noisy labels are corrupted from ground-truth labels by an unknown noise transition matrix. Thus, the accuracy of classifiers can be improved by estimating this matrix. We present a human-assisted approach called Masking. Masking conveys human cognition of invalid class transitions, and naturally speculates the structure of the noise transition matrix. Given the structure information, we only learn the noise transition probability to reduce the estimation burden. Second, motivated by the memorization effects of deep networks, which shows networks fit clean instances first and then noisy ones, we present a new paradigm called Co-teaching even combating with extremely noisy labels. We train two networks simultaneously. First, in each mini-batch data, each network filters noisy instances based on the memorization effects. Then, it teaches the remaining instances to its peer network for updating the parameters. To tackle the consensus issue in Co-teaching, we propose a robust learning paradigm called Co-teaching+, which bridges the Update by Disagreement strategy with the original Co-teaching. Third, deep networks inevitably memorize some noisy labels, which will degrade their generalization. We propose a meta algorithm called Pumpout to overcome the problem of memorizing noisy labels. By using scaled stochastic gradient ascent, Pumpout actively squeezes out the negative effects of noisy labels from the training model, instead of passively forgetting these effects. We leverage Pumpout to robustify two representative methods: MentorNet and Backward Correction.

# Speaker index