

Apprentissage multi-tâches

Sylvain Arlot (sylvain.arlot@ens.fr)

En apprentissage statistique, on dispose souvent de peu de données pour pouvoir réellement “apprendre” de manière satisfaisante. L’apprentissage multi-tâches propose une manière de résoudre cette difficulté, à nombre de données constant. L’idée est de considérer simultanément plusieurs problèmes d’apprentissage, que l’on sait similaires, et de profiter de cette similarité pour les résoudre mieux ensemble que si on les avait considérés séparément. Par exemple, pour la classification d’images, les questions “y a-t-il un chat sur cette photo?” et “y a-t-il un chien sur cette photo?” sont suffisamment similaires pour qu’il soit intéressant de les aborder de front, en commençant par apprendre à repérer la présence d’un chat ou d’un chien, avant de résoudre chaque problème spécifique. Le but de ce mémoire est de comprendre l’approche multi-tâches proposée par [1] pour le problème de régression avec des estimateurs à noyaux, et en particulier préciser de quel type de “similarité” entre tâches cette approche permet de profiter pour l’apprentissage.

References

- [1] T. Evgeniou, C. Micchelli and M. Pontil. *Learning multiple tasks with kernel methods*. Journal of Machine Learning Research, 6:615–637, 2005. <http://www.jmlr.org/papers/volume6/evgeniou05a/evgeniou05a.pdf>

Convex Relaxations for Permutation Problems

Alexandre d’Aspremont (aspremon@ens.fr)

Seriation seeks to reconstruct a linear order between variables using unsorted, pairwise similarity information. It has direct applications in archeology and DNA sequencing for example. We write seriation as an optimization problem by proving the equivalence between the seriation and combinatorial 2-SUM problems on similarity matrices (2-SUM is a quadratic minimization problem over permutations). The seriation problem can be solved exactly by a spectral algorithm in the noiseless case and we derive several convex relaxations for 2-SUM to improve the robustness of seriation solutions in noisy settings. These convex formulations also allow us to impose structural constraints on the solution, hence solve semi-supervised seriation problems. We study new approximation bounds for some of these relaxations and present numerical experiments on archeological data, Markov chains and DNA assembly from shotgun gene sequencing data.

DNA presents unique challenges for seriation because human sequences contain large repeated sequences which make local reconstruction problems degenerate. It is not clear if and how these degeneracies can be overcome algorithmically in a reliable way. The goal of this work will be to test new algorithms on realistic DNA data sets from the assemblathon.

References

- [1] J.E. Atkins and E.G. Boman and B. Hendrickson and others. A spectral algorithm for seriation and the consecutive ones problem, SIAM J. Comput. 1998
- [2] F. Fogel, R. Jenatton, F. Bach, and A. d’Aspremont. Convex relaxations for permutation problems. NIPS 2013, arXiv:1306.4805

Apprentissage statistique a grande echelle

Francis Bach (francis.bach@ens.fr)

De nombreux problèmes d'apprentissage statistique à grande échelle sont formulés comme l'optimisation d'une fonction convexe dont on n'observe que des gradients bruités: cette fonction est typiquement l'erreur de généralisation, et seulement l'erreur sur une observation est disponible à chaque itération. Les algorithmes utilisés en pratique donnent lieu à des garanties de convergence dont l'étude [1] est le but du mémoire.

References

- [1] F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. Technical report, HAL 00831977, 2013. To appear in Advances in Neural Information Processing Systems (NIPS).

Échantillonnage parfait pour les processus spaciaux

Anne Bouillard (anne.bouillard@ens.fr)

L'algorithme de simulation parfaite développé par Propp et Wilson en 1996 [1] permet d'échantillonner de manière exacte une variable aléatoire selon la distribution stationnaire d'une chaîne de Markov ergodique finie. La complexité de l'algorithme est proportionnelle au temps de couplage de la chaîne, ainsi qu'à son nombre d'états rendant son application directe inefficace pour des espaces d'états très grands. Cette limite est contournée par l'introduction de chaînes bornantes [2]. Le but de ce mémoire est de comprendre ces concepts et son application à l'échantillonnage de processus spatiaux répulsifs [3, 4].

References

- [1] Olle Häggström. Finite Markov Chains and Algorithmic Applications *London Mathematical Society Student Texts*, Cambridge university press, 2002.
- [2] Mark L. Huber. Perfect sampling using bounding chains. *The Annals of Applied Probability*, vol. 14 no. 2 (August, 2004), pp. 734–753.
- [3] Spatial Birth-Death-Swap Chains. Mark L. Huber. *Bernoulli*, vol. 18, no. 3 (2012) pp. 1031–1041.
- [4] Mark L. Huber, Elise Vilella, Daniel Rozenfeld and Jason Xu. Bounds on the artificial phase transition for perfect simulation of hard core Gibbs processes. *Involve*, vol. 5, no. 3 (2012), pp. 247–255.

Hybrid reactive systems: semantics and causality

Timothy Bourke (PARKAS, Timothy.Bourke@ens.fr)

Marc Pouzet (PARKAS, Marc.Pouzet@ens.fr)

Benoit Caillaud (Inria Rennes, Benoit.Caillaud@inria.fr)

This project is about the semantics and compilation of languages for programming critical embedded controls software and modeling their physical environment. Important examples can be found, for example, in avionics. The flight control systems of Airbus, Embraer, and other companies are programmed in dedicated languages with strong mathematical foundations and their environments are modeled using differential equations and simulated numerically.

From a mathematical point of view, the objects manipulated are difference equations (that define discrete sequences), automata, and Ordinary Differential Equations (ODEs). The interaction between discrete and continuous elements is the source of serious bugs in the most used tools (such as SIMULINK/STATEFLOW,¹; of which there are 1 000 000 licenses worldwide).

Causality loops are an important source of problems. They arise naturally when modeling the interdependence between reactive components and their environments and also from feedback dynamics within the environment itself. Compilers must untangle such loops to produce a sequence of instructions for a computer to execute—whether in simulation or in reality.

We proposed [2] that programming languages for such systems be defined using *non-standard analysis* [4, 3], with the idea that the progression of time can be modeled as a sequence of steps of infinitesimal length: $\partial \approx 0$, and $\forall x \in \mathbb{R}. 0 < \partial < x$. In this context, the ODE $\dot{x} = f(x)$ is equal to

$$\begin{aligned} *x(n) &= *x(n-1) + \partial \cdot f(*x(n-1)) \\ x(0) &= 0 \end{aligned}$$

A program is causal when there are no instantaneous cycles between its defining equations. This property can be checked by a dedicated type system and it can be shown that the signals of well-typed programs are piecewise constant [1].

The objective of this subject is to study [2] and its application to causality analysis in [1], both of which are implemented in the language ZÉLUS.² This work would be followed by an internship with Benoit Caillaud at Inria Rennes during which the aim is to develop an alternative semantics based on the *superdense time* of Maler, Manna, and Pnueli [5, 6], and to use it to address the issues of causality raised in [1]—these are important scientific and practical questions. This project would suit a student interested in the semantics and implementation of programming languages and who has a strong taste for mathematics.

¹<http://www.mathworks.com/products/simulink>

²<http://zelus.di.ens.fr>

References

- [1] A. Benveniste, T. Bourke, B. Caillaud, B. Pagano, and M. Pouzet. A type-based analysis of causality loops in hybrid modelers. In M. Fränzle and J. Lygeros, editors, *Proc. 17th Int. Conf. on Hybrid Systems: Computation and Control (HSCC 2014)*, Berlin, Germany, Apr. 2014. ACM Press. to appear.
- [2] A. Benveniste, T. Bourke, B. Caillaud, and M. Pouzet. Non-Standard Semantics of Hybrid Systems Modelers. *Journal of Computer and System Sciences (JCSS)*, 78(3):877–910, May 2012.
- [3] H. J. Keisler. *Elementary Calculus: An Infinitesimal Approach*. Prindle, Weber and Schmidt, 2 edition, 1986. available online: <http://www.math.wisc.edu/~keisler/calc.html>.
- [4] T. Lindstrom. An invitation to non standard analysis. In N. Cutland, editor, *Nonstandard analysis and its applications*. Cambridge Univ. Press, 1988.
- [5] O. Maler, Z. Manna, and A. Pnueli. Prom timed to hybrid systems. In J. de Bakker, C. Huizing, W. de Roever, and G. Rozenberg, editors, *Real-Time: Theory in Practice*, volume 600 of *Lecture Notes in Computer Science*, pages 447–484. Springer Berlin / Heidelberg, 1992.
- [6] Z. Manna and A. Pnueli. Verifying hybrid systems. In R. L. Grossman, A. Nerode, A. P. Ravn, and H. Rischel, editors, *Hybrid Systems*, volume 736 of *Lecture Notes in Computer Science*, pages 4–35. Springer-Verlag, 1993.

Représentations et propriétés des “string graphs”

Éric Colin de Verdière (eric.colin.de.verdiere@ens.fr)

Soit C un ensemble fini de courbes dans le plan. Soit $S(C)$ le *string graph* associé à C : c’est le graphe ayant C pour sommets et tel que deux éléments c_1 et c_2 de C sont adjacents si et seulement si les courbes c_1 et c_2 s’intersectent. Les *string graphs* sont intéressants et assez généraux: de nombreuses classes de graphes étudiées pour leurs propriétés algorithmiques ou combinatoires, comme les graphes d’intersections, sont des *string graphs*.

Diverses questions naturelles ont été résolues. Par exemple, quel est le nombre minimal d’intersections d’une famille de courbes C tel que $S(C)$ est un graphe G donné? Quelles sont les propriétés structurelles des *string graphs*? Nous verrons aussi que l’étude des *string graphs* permet de démontrer des résultats subtils sur les différentes notions de *crossing number* d’un graphe (dans sa version la plus basique, le *crossing number* d’un graphe G est le nombre minimal de croisements d’un dessin de G dans le plan).

Matoušek [1] a très récemment écrit un article (presque auto-suffisant) sur les *string graphs*; le but de ce travail est d’abord d’étudier cet article, et ensuite, en fonction du temps restant et des goûts, de creuser dans une ou deux directions suggérées dans ce texte. Cela sera aussi l’occasion d’apprendre et d’utiliser quelques outils généraux de théorie géométrique et topologique des graphes et de dualité en programmation linéaire.

References

- [1] Jiří. Matoušek, String graphs and separators. arXiv:1311.5048.

Détection de communautés et matrices aléatoires

Marc Lelarge (marc.lelarge@ens.fr)

Le but de ce mémoire est d'étudier des réseaux ayant une structure: des groupes de nœuds sont fortement connectés. Le problème de retrouver ces communautés à partir du graphe intervient par exemple dans l'étude de graphes sociaux où le réseaux est alors le graphe des interactions. Récemment, des méthodes basées sur le spectre de la matrice d'adjacence ont été proposées [1]. D'un point de vue mathématique, la performance de ces méthodes est reliée à un phénomène de transition de phase pour le spectre de matrices aléatoires symétriques [2].

References

- [1] Raj Rao Nadakuditi, M. E. J. Newman. Graph spectra and the detectability of community structure in networks, 2012 arXiv:1205.1813
- [2] Florent Benaych-Georges, Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices, Adv. Math. 227 (2011), no. 1, 494–521.

Proof techniques for graphs excluding induced subgraphs

Zhentao Li (zhentao.li@ens.fr)

While the idea of developing structural decomposition theorems for graphs excluding some induced subgraphs is not entirely new, recent work in this area has allowed breakthrough for long standing open questions such as the Strong Perfect Graph Theorem [1]. For this project, we aim to understand the new tools introduced by such work and will study one of these new structure theorems in detail: the structure theorem for claw-free graphs [7]. As the proof is very long, the aim is to read the survey article [7] and then one of the articles containing details [2, 3, 4, 5, 6].

The goal of this project is to render explicit the proof techniques developed in those paper. I.e., suppose we know that graphs excluding some set of induced subgraphs has one structure amongst some list of structures, what proof techniques would be needed to show this (if we are allowed a finite but arbitrarily large number of case analysis)? What techniques tend to cut down on the number of cases the most?

References

- [1] Maria Chudnovsky, Neil Robertson, Paul Seymour, and Robin Thomas. The strong perfect graph theorem. *Annals of Mathematics*, pages 51–229, 2006.
- [2] Maria Chudnovsky and Paul Seymour. Claw-free graphs. I. orientable prismatic graphs. *Journal of Combinatorial Theory, Series B*, 97(6):867–903, 2007.
- [3] Maria Chudnovsky and Paul Seymour. Claw-free graphs. II. non-orientable prismatic graphs. *Journal of Combinatorial Theory, Series B*, 98(2):249–290, 2008.
- [4] Maria Chudnovsky and Paul Seymour. Claw-free graphs. III. circular interval graphs. *Journal of Combinatorial Theory, Series B*, 98(4):812–834, 2008.
- [5] Maria Chudnovsky and Paul Seymour. Claw-free graphs. IV. decomposition theorem. *Journal of Combinatorial Theory, Series B*, 98(5):839–938, 2008.
- [6] Maria Chudnovsky and Paul Seymour. Claw-free graphs. V. global structure. *Journal of Combinatorial Theory, Series B*, 98(6):1373–1410, 2008.
- [7] Maria Chudnovsky and Paul D Seymour. The structure of claw-free graphs. *Surveys in combinatorics*, 327:153–171, 2005.

The Learning Parity with Noise over Rings Problem

Vadim Lyubashevsky (vadim.lyubash@gmail.com)

In this internship, we will study the Learning Parity with Noise over Rings (Ring-LPN) problem - a problem stemming from the field of cryptography. The problem is defined as follows: there is some polynomial ring $R = \mathbb{Z}_2[x]/(f(x))$ and a secret polynomial s . We are given many samples of the form $(a, as + e)$ where a is randomly-chosen from R and e is chosen from R such that its weight (i.e. the number of 1's in it) is small. Every sample has a different a and e , while the s is always kept the same. The goal is to find s . The algorithmic hardness of this problem may depend on the ring R , and in particular the structure of the polynomial $f(x)$. For example, if $f(x)$ splits modulo 2, then the problem may be easier than when $f(x)$ is irreducible (and thus R is a field). We will look into what other properties of $f(x)$ affect the hardness of the Ring-LPN problem.

References

- [1] Lapin: An Efficient Authentication Protocol Based on Ring-LPN : <http://www.di.ens.fr/~lyubash/papers/lapin.pdf>

Représentation Parcimonieuse de Signaux et d'Images

Stéphane Mallat (stephane.mallat@ens.fr)

Beaucoup de problèmes de traitement du signal ou de l'image nécessitent de construire une représentation parcimonieuse, ce qui veut dire représenter un signal de grande dimension avec un nombre réduit de paramètres. Les applications concernent la compression de signaux ou d'images, le débruitage, la restauration d'images dégradées ainsi que des problèmes de classification. Le but de ce projet est d'étudier les techniques de représentation parcimonieuses et une application particulière qui pourra être choisie par l'élève. Le travail se fera à partir d'un article scientifique qui dépendra de l'application choisie et inclura des simulations numériques pour tester les algorithmes étudiés. Différents stages sont possibles dans des laboratoires travaillant sur ces problèmes, aussi bien en France qu'aux US. Ce projet nécessite de suivre le cours de traitement du signal.

Analyse d’algorithmes distribués d’accès au canal pour les réseaux sans fil avec une dynamique spatiale

Florian Simatos (florian.simatos@inria.fr)

Un problème fondamental dans le cadre des réseaux sans fil est de concevoir des algorithmes d’accès au canal distribués, simples et efficaces. Récemment, ce problème a reçu une attention considérable et plusieurs algorithmes ont été proposés [2, 3]. Néanmoins, ces algorithmes ont principalement été étudiés dans un cadre statique et le but de ce mémoire est d’étudier le comportement de ces algorithmes en présence d’une dynamique spatiale des utilisateurs. Le travail de mémoire pourra par exemple s’appuyer sur la lecture d’articles récents [1, 4] traitant de problèmes proches, ainsi que sur des résultats de simulation.

References

- [1] N. Bouman, S. C. Borst, and J. S. H. van Leeuwaarden. Stability of spatial wireless systems with random admissible-set scheduling. In *Proceedings of the 5th International ICST Conference on Performance Evaluation Methodologies and Tools, VALUETOOLS ’11*, pages 57–65, ICST, Brussels, Belgium, Belgium, 2011. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).
- [2] Libin Jiang, Devavrat Shah, Jinwoo Shin, and Jean Walrand. Distributed random access algorithm: scheduling and congestion control. *IEEE Trans. Inform. Theory*, 56(12):6182–6207, 2010.
- [3] Devavrat Shah, Jinwoo Shin, and Prasad Tetali. Medium access using queues. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science—FOCS 2011*, pages 698–707. IEEE Computer Soc., Los Alamitos, CA, 2011.
- [4] P. Van De Ven, S. Borst, and Lei Ying. Spatial inefficiency of maxweight scheduling. In *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2011 International Symposium on*, pages 62–69, 2011.