Learning with Submodular Functions: A Convex Optimization Perspective

Francis Bach
INRIA - Ecole Normale Supérieure, Paris, France
francis.bach@ens.fr

the essence of knowledge
Boston — Delft

Foundations and Trends® in Machine Learning

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

F. Bach. *Learning with Submodular Functions: A Convex Optimization Perspective.* Foundations and Trends® in Machine Learning, vol. 6, no. 2-3, pp. 145–373, 2013.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-60198-757-0

Foundations and Trends® in Machine Learning
Volume 6, Issue 2-3, 2013

Editorial Scope

Topics

Foundations and Trends® in Machine Learning publishes survey and tutorial articles on the theory, algorithms and applications of machine learning, including the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive, and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians

Foundations and Trends® in Machine Learning, 2013, Volume 6, 4 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.

Foundations and Trends® in Machine Learning
Vol. 6, No. 2-3 (2013) 145–373

DOI: 10.1561/2200000039

Learning with Submodular Functions: A Convex Optimization Perspective

Francis Bach
INRIA - Ecole Normale Supérieure, Paris, France
francis.bach@ens.fr

Contents

Abstract

Submodular functions are relevant to machine learning for at least two reasons: (1) some problems may be expressed directly as the optimization of submodular functions and (2) the Lovász extension of submodular functions provides a useful set of regularization functions for supervised and unsupervised learning. In this monograph, we present the theory of submodular functions from a convex analysis perspective, presenting tight links between certain polyhedra, combinatorial optimization and convex optimization problems. In particular, we show how submodular function minimization is equivalent to solving a wide variety of convex optimization problems. This allows the derivation of new efficient algorithms for approximate and exact submodular function minimization with theoretical guarantees and good practical performance. By listing many examples of submodular functions, we review various applications to machine learning, such as clustering, experimental design, sensor placement, graphical model structure learning or subset selection, as well as a family of structured sparsity-inducing norms that can be derived and used from submodular functions.

F. Bach. *Learning with Submodular Functions: A Convex Optimization Perspective*. Foundations and Trends® in Machine Learning, vol. 6, no. 2-3, pp. 145–373, 2013.
DOI: 10.1561/2200000039.

1

Introduction

Many combinatorial optimization problems may be cast as the minimization of a *set-function*, that is a function defined on the set of subsets of a given base set V. Equivalently, they may be defined as functions on the vertices of the hyper-cube, i.e, $\{0,1\}^p$ where p is the cardinality of the base set V—they are then often referred to as pseudo-boolean functions [27]. Among these set-functions, submodular functions play an important role, similar to convex functions on vector spaces, as many functions that occur in practical problems turn out to be submodular functions or slight modifications thereof, with applications in many areas areas of computer science and applied mathematics, such as machine learning [125, 154, 117, 124], computer vision [31, 97], operations research [99, 179], electrical networks [159] or economics [200]. Since submodular functions may be minimized exactly, and maximized approximately with some guarantees, in polynomial time, they readily lead to efficient algorithms for all the numerous problems they apply to. They also appear in several areas of theoretical computer science, such as matroid theory [186].

However, the interest for submodular functions is not limited to discrete optimization problems. Indeed, the rich structure of submodular functions and their link with convex analysis through the Lovász extension [134] and the various associated polytopes makes them particularly adapted to problems beyond combinatorial optimization, namely as regularizers in signal processing and machine learning problems [38, 7]. Indeed, many continuous optimization problems exhibit an underlying discrete structure (e.g., based on chains, trees or more general graphs), and submodular functions provide an efficient and versatile tool to capture such combinatorial structures.

In this monograph, the theory of submodular functions is presented in a self-contained way, with all results proved from first principles of convex analysis common in machine learning, rather than relying on combinatorial optimization and traditional theoretical computer science concepts such as matroids or flows (see, e.g., [72] for a reference book on such approaches). Moreover, the algorithms that we present are based on traditional convex optimization algorithms such as the simplex method for linear programming, active set method for quadratic programming, ellipsoid method, cutting planes, and conditional gradient. These will be presented in details, in particular in the context of submodular function minimization and its various continuous extensions. A good knowledge of convex analysis is assumed (see, e.g., [30, 28]) and a short review of important concepts is presented in Appendix A—for more details, see, e.g., [96, 30, 28, 182].

Monograph outline. The monograph is organized in several chapters, which are summarized below (in the table of contents, sections that can be skipped in a first reading are marked with a star*):

– **Definitions**: In Chapter 2, we give the different definitions of submodular functions and of the associated polyhedra, in particular, the base polyhedron and the submodular polyhedron. They are crucial in submodular analysis as many algorithms and models may be expressed naturally using these polyhedra.

– **Lovász extension**: In Chapter 3, we define the Lovász extension as an extension from a function defined on $\{0,1\}^p$ to a function defined on $[0,1]^p$ (and then $\mathbb{R}^p$), and give its main properties. In particular we present key results in submodular analysis: the Lovász extension is convex if and only if the set-function is submodular; moreover, minimizing the submodular set-function F is equivalent to minimizing the Lovász extension on $[0,1]^p$. This implies notably that submodular function minimization may be solved in polynomial time. Finally, the link between the Lovász extension and the submodular polyhedra through the so-called "greedy algorithm" is established: the Lovász extension is the support function of the base polyhedron and may be computed in closed form.

– **Polyhedra**: Associated polyhedra are further studied in Chapter 4, where support functions and the associated maximizers of linear functions are computed. We also detail the facial structure of such polyhedra, which will be useful when related to the sparsity-inducing properties of the Lovász extension in Chapter 5.

– **Convex relaxation of submodular penalties**: While submodular functions may be used directly (for minimization of maximization of set-functions), we show in Chapter 5 how they may be used to penalize supports or level sets of vectors. The resulting mixed combinatorial/continuous optimization problems may be naturally relaxed into convex optimization problems using the Lovász extension.

– **Examples**: In Chapter 6, we present classical examples of submodular functions, together with several applications in machine learning, in particular, cuts, set covers, network flows, entropies, spectral functions and matroids.

– **Non-smooth convex optimization**: In Chapter 7, we review classical iterative algorithms adapted to the minimization of non-smooth polyhedral functions, such as subgradient, ellipsoid, simplicial, cutting-planes, active-set, and conditional gradient methods. A particular attention is put on providing when applicable primal/dual interpretations of these algorithms.

– **Separable optimization - Analysis**: In Chapter 8, we consider *separable* optimization problems regularized by the Lovász extension $w \mapsto f(w)$, i.e., problems of the form $\min_{w\in\mathbb{R}^p} \sum_{k\in V} \psi_k(w_k) + f(w)$, and show how this is equivalent to a sequence of submodular function minimization problems. This is a key theoretical link between combinatorial and convex optimization problems related to submodular functions, that will be used in later chapters.

– **Separable optimization - Algorithms**: In Chapter 9, we present two sets of algorithms for separable optimization problems. The first algorithm is an exact algorithm which relies on the availability of an efficient submodular function minimization algorithm, while the second set of algorithms are based on existing iterative algorithms for convex optimization, some of which come with online and offline theoretical guarantees. We consider active-set methods ("min-norm-point" algorithm) and conditional gradient methods.

– **Submodular function minimization**: In Chapter 10, we present various approaches to submodular function minimization. We present briefly the combinatorial algorithms for exact submodular function minimization, and focus in more depth on the use of specific convex optimization problems, which can be solved iteratively to obtain approximate or exact solutions for submodular function minimization, with sometimes theoretical guarantees and approximate optimality certificates. We consider the subgradient method, the ellipsoid method, the simplex algorithm and analytic center cutting planes. We also show how the separable optimization problems from Chapters 8 and 9 may be used for submodular function minimization. These methods are then empirically compared in Chapter 12.

– **Submodular optimization problems**: In Chapter 11, we present other combinatorial optimization problems which can be partially solved using submodular analysis, such as submodular function maximization and the optimization of differences of submodular functions, and relate these to non-convex optimization problems on the submodular polyhedra. While these problems typically cannot be solved in polynomial time, many algorithms come with approximation guarantees based on submodularity.

– **Experiments**: In Chapter 12, we provide illustrations of the optimization algorithms described earlier, for submodular function minimization, as well as for convex optimization problems (separable or not). The Matlab code for all these experiments may be found at `http://www.di.ens.fr/~fbach/submodular/`.

In Appendix A, we review relevant notions from convex analysis (such as Fenchel duality, dual norms, gauge functions, and polar sets), while in Appendix B, we present in details operations that preserve submodularity.

Several books and monograph articles already exist on the same topic and the material presented in this monograph rely on those [72, 159]. However, in order to present the material in the simplest way, ideas from related research papers have also been used, and a stronger emphasis is put on convex analysis and optimization.

Notations. We consider the set $V = \{1,\dots,p\}$, and its power set 2^V, composed of the 2^p subsets of V. Given a vector $s \in \mathbb{R}^p$, s also denotes the modular set-function defined as $s(A) = \sum_{k\in A} s_k$. Moreover, $A \subseteq B$ means that A is a subset of B, potentially equal to B. We denote by $|A|$ the cardinality of the set A, and, for $A \subseteq V = \{1,\dots,p\}$, $1_A \in \mathbb{R}^p$ denotes the indicator vector of the set A. If $w \in \mathbb{R}^p$, and $\alpha \in \mathbb{R}$, then $\{w \geqslant \alpha\}$ (resp. $\{w > \alpha\}$) denotes the subset of $V = \{1,\dots,p\}$ defined as $\{k \in V,\ w_k \geqslant \alpha\}$ (resp. $\{k \in V,\ w_k > \alpha\}$), which we refer to as the weak (resp. strong) α-sup-level sets of w. Similarly if $v \in \mathbb{R}^p$, we denote $\{w \geqslant v\} = \{k \in V,\ w_k \geqslant v_k\}$.

For $q \in [1,+\infty]$, we denote by $\|w\|_q$ the ℓ_q-norm of w, defined as $\|w\|_q = \big(\sum_{k\in V} |w_k|^q\big)^{1/q}$ for $q \in [1,\infty)$ and $\|w\|_\infty = \max_{k\in V} |w_k|$. Finally, we denote by $\mathbb{R}_+$ the set of non-negative real numbers, by $\mathbb{R}^*$ the set of non-zero real numbers, and by $\mathbb{R}_+^*$ the set of strictly positive real numbers.

References

[1] S. Ahmed and A. Atamtürk. Maximizing a class of submodular utility functions. *Mathematical Programming: Series A and B*, 128(1-2):149–169, 2011.

[2] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network flows: theory, algorithms, and applications*. Prentice hall, 1993.

[3] T. Ando. Concavity of certain maps on positive definite matrices and applications to Hadamard products. *Linear Algebra and its Applications*, 26:203–241, 1979.

[4] M. Babenko, J. Derryberry, A. Goldberg, R. Tarjan, and Y. Zhou. Experimental evaluation of parametric max-flow algorithms. In *Proceedings of the International Conference on Experimental algorithms (WEA)*, 2007.

[5] F. Bach. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

[6] F. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[7] F. Bach. Structured sparsity-inducing norms through submodular functions. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[8] F. Bach. Shaping level sets with submodular functions. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[9] F. Bach. Convex relaxations of structured matrix factorizations. Technical Report 00861118, HAL, 2013.

[10] F. Bach. Duality between subgradient and conditional gradient methods. Technical Report 00757696-v3, HAL, 2013.

[11] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2011.

[12] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Structured sparsity through convex optimization. *Statistical Science*, 27(4):450–468, 2012.

[13] M. F. Balcan and N. J. A. Harvey. Learning submodular functions. In *Proceedings of the Symposium on Theory of Computing (STOC)*, 2011.

[14] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[15] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56(4):1982–2001, 2010.

[16] A. Barbero and S. Sra. Fast Fewton-type methods for total variation regularization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2011.

[17] R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical Inference under Order Restrictions: The Theory and Application of Isotonic Regression*. John Wiley, 1972.

[18] H. H. Bauschke, P. L. Combettes, and D. R. Luke. Finding best approximation pairs relative to two closed convex sets in Hilbert spaces. *Journal of Approximation Theory*, 127(2):178–192, 2004.

[19] A. Beck and M. Teboulle. A conditional gradient method with linear rate of convergence for solving convex linear systems. *Mathematical Methods of Operations Research*, 59(2):235–247, 2004.

[20] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[21] S. Becker, J. Bobin, and E. Candès. NESTA: A fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):1–39, 2011.

[22] D. Bertsekas. *Nonlinear programming*. Athena Scientific, 1995.

[23] D. P. Bertsekas and H. Yu. A unifying polyhedral approximation framework for convex optimization. *SIAM Journal on Optimization*, 21(1):333–360, 2011.

[24] D. Bertsimas and J. N. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1997.

[25] M. J. Best and N. Chakravarti. Active set algorithms for isotonic regression: a unifying framework. *Mathematical Programming*, 47(1):425–439, 1990.

[26] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal. *Numerical Optimization Theoretical and Practical Aspects*. Springer, 2003.

[27] E. Boros and P.L. Hammer. Pseudo-Boolean optimization. *Discrete Applied Mathematics*, 123(1-3):155–225, 2002.

[28] J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer, 2006.

[29] M. Bouhtou, S. Gaubert, and G. Sagnol. Submodularity and randomized rounding techniques for optimal experimental design. *Electronic Notes in Discrete Mathematics*, 36:679–686, 2010.

[30] S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[31] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 23(11):1222–1239, 2001.

[32] P. Brucker. An $O(n)$ algorithm for quadratic knapsack problems. *Operations Research Letters*, 3(3):163–166, 1984.

[33] N. Buchbinder, M. Feldman, J. Naor, and R. Schwartz. A tight linear time (1/2)-approximation for unconstrained submodular maximization. In *Proceedings of the Symposium on Foundations of Computer Science (FOCS)*, 2012.

[34] G. Calinescu, C. Chekuri, M. Pál, and J. Vondrák. Maximizing a monotone submodular function subject to a matroid constraint. *SIAM Journal on Computing*, 40(6):1740–1766, 2011.

[35] J. F. Cardoso. Dependence, correlation and gaussianity in independent component analysis. *Journal of Machine Learning Research*, 4:1177–1203, 2003.

[36] V. Cevher, M. F. Duarte, C. Hegde, and R. G. Baraniuk. Sparse signal recovery using Markov random fields. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[37] A. Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical imaging and vision*, 20(1):89–97, 2004.

[38] A. Chambolle and J. Darbon. On total variation minimization and surface evolution using parametric maximum flows. *International Journal of Computer Vision*, 84(3):288–307, 2009.

[39] G. Charpiat. Exhaustive family of energies minimizable exactly by a graph cut. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[40] A. Chechetka and C. Guestrin. Efficient principled learning of thin junction trees. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[41] C. Chekuri and A. Ene. Approximation algorithms for submodular multiway partition. In *Proceedings of the Symposium on Foundations of Computer Science (FOCS)*, 2011.

[42] C. Chekuri, J. Vondrák, and R. Zenklusen. Submodular function maximization via the multilinear relaxation and contention resolution schemes. In *Proceedings of the Symposium on Theory of Computing (STOC)*, 2011.

[43] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[44] B. V. Cherkassky and A. V. Goldberg. On implementing the push-relabel method for the maximum flow problem. *Algorithmica*, 19(4):390–410, 1997.

[45] S. Chopra. On the spanning tree polyhedron. *Operations Research Letters*, 8(1):25–29, 1989.

[46] G. Choquet. Theory of capacities. *Annales de l'Institut Fourier*, 5:131–295, 1954.

[47] F. R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

[48] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer, 2010.

[49] T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms.* MIT Press, 1989.

[50] G. Cornuejols, M. L. Fisher, and G. L. Nemhauser. Location of bank accounts to optimize float: An analytic study of exact and approximate algorithms. *Management Science*, 23(8):789–810, 1977.

[51] G. Cornuejols, M. L. Fisher, and G. L. Nemhauser. On the uncapacitated location problem. *Annals of Discrete Mathematics*, 1:163–177, 1977.

[52] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.

[53] W. H. Cunningham. Testing membership in matroid polyhedra. *Journal of Combinatorial Theory, Series B*, 36(2):161–188, 1984.

[54] W. H. Cunningham. Minimum cuts, modular functions, and matroid polyhedra. *Networks*, 15(2):205–215, 1985.

[55] J. Darbon. Global optimization for first order Markov random fields with submodular priors. In *Combinatorial Image Analysis*, pages 229–237. Springer, 2008.

[56] A. Das and D. Kempe. Algorithms for subset selection in linear regression. In *Proceedings of the Symposium on Theory of Computing (STOC)*, 2008.

[57] A. Das and D. Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2011.

[58] B. A. Davey and H. A. Priestley. *Introduction to Lattices and Order*. Cambridge University Press, 2002.

[59] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.

[60] M. Dudik, Z. Harchaoui, and J. Malick. Lifted coordinate descent for learning with trace-norm regularization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

[61] J. C. Dunn. Convergence rates for conditional gradient sequences generated by implicit step length rules. *SIAM Journal on Control and Optimization*, 18:473–487, 1980.

[62] J. C. Dunn and S. Harshbarger. Conditional gradient algorithms with open loop step size rules. *Journal of Mathematical Analysis and Applications*, 62(2):432–444, 1978.

[63] J. Edmonds. Submodular functions, matroids, and certain polyhedra. In *Combinatorial optimization - Eureka, you shrink!*, pages 11–26. Springer, 2003.

[64] V. V. Fedorov. *Theory of optimal experiments*. Academic Press, 1972.

[65] U. Feige. A threshold of $\ln(n)$ for approximating set cover. *Journal of the ACM (JACM)*, 45(4):634–652, 1998.

[66] U. Feige. On maximizing welfare when utility functions are subadditive. In *Proceedings of the Symposium on Theory of Computing (STOC)*, 2006.

[67] U. Feige, V. S. Mirrokni, and J. Vondrák. Maximizing non-monotone submodular functions. *SIAM Journal on Computing*, 40(4):1133–1153, 2011.

[68] M. Feldman, J. Naor, and R. Schwartz. A unified continuous greedy algorithm for submodular maximization. In *Proceedings of the Symposium on Foundations of Computer Science (FOCS)*, 2011.

[69] S. Foldes and P. L. Hammer. Submodularity, supermodularity, and higher-order monotonicities of pseudo-Boolean functions. *Mathematics of Operations Research*, 30(2):453–461, 2005.

[70] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[71] J. Friedman, T. Hastie, and R. Tibshirani. A note on the group Lasso and a sparse group Lasso. Technical Report 1001.0736, ArXiv, 2010.

[72] S. Fujishige. *Submodular Functions and Optimization*. Elsevier, 2005.

[73] S. Fujishige and S. Isotani. A submodular function minimization algorithm based on the minimum-norm base. *Pacific Journal of Optimization*, 7:3–17, 2011.

[74] G. Gallo, M.D. Grigoriadis, and R.E. Tarjan. A fast parametric maximum flow algorithm and applications. *SIAM Journal on Computing*, 18(1):30–55, 1989.

[75] A. Gelman. *Bayesian Data Analysis*. CRC Press, 2004.

[76] E. Girlich and N. N. Pisaruk. The simplex method for submodular function minimization. Technical Report 97-42, University of Magdeburg, 1997.

[77] M. X. Goemans, N. J. A. Harvey, S. Iwata, and V. Mirrokni. Approximating submodular functions everywhere. In *Proceedings of the Symposium on Discrete Algorithms (SODA)*, 2009.

[78] J.-L. Goffin and J.-P. Vial. On the computation of weighted analytic centers and dual ellipsoids with the projective algorithm. *Mathematical Programming*, 60(1-3):81–92, 1993.

[79] A. V. Goldberg and R. E. Tarjan. A new approach to the maximum-flow problem. *Journal of the ACM (JACM)*, 35(4):921–940, 1988.

[80] B. Goldengorin, G. Sierksma, G.A. Tijssen, and M. Tso. The data-correcting algorithm for the minimization of supermodular functions. *Management Science*, 41(11):1539–1551, 1999.

[81] D. Golovin and A. Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research*, 42(1):427–486, 2011.

[82] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996.

[83] H. Groenevelt. Two algorithms for maximizing a separable concave function over a polymatroid feasible region. *European Journal of Operational Research*, 54(2):227–236, 1991.

[84] M. Grötschel, L. Lovász, and A. Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1(2):169–197, 1981.

[85] S. J. Grotzinger and C. Witzgall. Projections onto order simplexes. *Applied mathematics and Optimization*, 12(1):247–270, 1984.

[86] B. Grünbaum. *Convex polytopes*, volume 221. Springer Verlag, 2003.

[87] J. Guelat and P. Marcotte. Some comments on Wolfe's "away step". *Mathematical Programming*, 35(1):110–119, 1986.

[88] A. Guillory and J. Bilmes. Online submodular set cover, ranking, and repeated active learning. *Advance in Neural Information Processing Systems (NIPS)*, 2011.

[89] Z. Harchaoui, A. Juditsky, and A. Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. Technical Report 1302.2325, arXiv, 2013.

[90] Z. Harchaoui and C. Lévy-Leduc. Catching change-points with Lasso. *Advances in Neural Information Processing Systems (NIPS)*, 20, 2008.

[91] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, 2001.

[92] J. Haupt and R. Nowak. Signal reconstruction from noisy random projections. *IEEE Transactions on Information Theory*, 52(9):4036–4048, 2006.

[93] E. Hazan and S. Kale. Online submodular minimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[94] M. Hebiri and S. van de Geer. The Smooth-Lasso and other $\ell_1 + \ell_2$-penalized methods. *Electronic Journal of Statistics*, 5:1184–1226, 2011.

[95] D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

[96] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms: Part 1: Fundamentals*. Springer, 1996.

[97] D. S. Hochbaum. An efficient algorithm for image segmentation, Markov random fields and related problems. *Journal of the ACM*, 48(4):686–701, 2001.

[98] D. S. Hochbaum. Multi-label Markov random fields as an efficient and effective tool for image segmentation, total variations and regularization. *Numerical Mathematics: Theory, Methods and Applications*, 6(1):169–198, 2013.

[99] D. S. Hochbaum and S. P. Hong. About strongly polynomial time algorithms for quadratic optimization over submodular constraints. *Mathematical Programming*, 69(1):269–309, 1995.

[100] T. Hocking, A. Joulin, F. Bach, and J.-P. Vert. Clusterpath: an algorithm for clustering using convex fusion penalties. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2011.

[101] H. Hoefling. A path algorithm for the fused Lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4):984–1006, 2010.

[102] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, 1990.

[103] R. Horst and N. V. Thoai. DC programming: overview. *Journal of Optimization Theory and Applications*, 103(1):1–43, 1999.

[104] J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

[105] D. R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37, 2004.

[106] H. Ishikawa. Exact optimization for Markov random fields with convex priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 25(10):1333–1336, 2003.

[107] S. Iwata, L. Fleischer, and S. Fujishige. A combinatorial strongly polynomial algorithm for minimizing submodular functions. *Journal of the ACM*, 48(4):761–777, 2001.

[108] L. Jacob, G. Obozinski, and J.-P. Vert. Group Lasso with overlaps and graph Lasso. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

[109] M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2013.

[110] S. Jegelka, F. Bach, and S. Sra. Reflection methods for user-friendly submodular optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[111] S. Jegelka, H. Lin, and J. A. Bilmes. Fast approximate submodular minimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[112] R. Jenatton, J-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12:2777–2824, 2011.

[113] R. Jenatton, A. Gramfort, V. Michel, G. Obozinski, E. Eger, F. Bach, and B. Thirion. Multiscale mining of fMRI data with hierarchical structured sparsity. *SIAM Journal on Imaging Sciences*, 5(3):835–856, 2012.

[114] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. *Journal Machine Learning Research*, 12:2297–2334, 2011.

[115] R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009.

[116] K. Kavukcuoglu, M. A. Ranzato, R. Fergus, and Y. Le-Cun. Learning invariant features through topographic filter maps. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[117] Y. Kawahara, K. Nagano, K. Tsuda, and J. A. Bilmes. Submodularity cuts and applications. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[118] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Conference on Knoledge Discovery and Data Mining (KDD)*, 2003.

[119] S. Kim and E. Xing. Tree-guided group Lasso for multi-task regression with structured sparsity. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

[120] V. Klee and G. J. Minty. How good is the simplex algorithm? In O. Shisha, editor, *Inequalities*, volume 3, pages 159–175. Academic Press, 1972.

[121] V. Kolmogorov. Minimizing a sum of submodular functions. *Discrete Applied Mathematics*, 160(15):2246–2258, 2012.

[122] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 26(2):147–159, 2004.

[123] N. Komodakis, N. Paragios, and G. Tziritas. Mrf energy minimization and beyond via dual decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(3):531–552, 2011.

[124] A. Krause and V. Cevher. Submodular dictionary selection for sparse representation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

[125] A. Krause and C. Guestrin. Near-optimal nonmyopic value of information in graphical models. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2005.

[126] A. Krause and C. Guestrin. Submodularity and its applications in optimized information gathering. *ACM Transactions on Intelligent Systems and Technology*, 2(4), 2011.

[127] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[128] K. S. S. Kumar and F. Bach. Maximizing submodular functions using probabilistic graphical models. Technical Report 00860575, HAL, 2013.

[129] S. L. Lauritzen. *Graphical Models*. Oxford University Press, July 1996.

[130] A. Lefèvre, F. Bach, and C. Févotte. Itakura-Saito nonnegative matrix factorization with group sparsity. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.

[131] H. Lin and J. Bilmes. A class of submodular functions for document summarization. In *Proceedings of the North American chapter of the Association for Computational Linguistics/Human Language Technology Conference (NAACL/HLT)*, 2011.

[132] H. Lin and J. A. Bilmes. Optimal selection of limited vocabulary speech corpora. In *Proceedings of INTERSPEECH*, 2011.

[133] F. Lindsten, H. Ohlsson, and L. Ljung. Clustering using sum-of-norms regularization: With application to particle filter output computation. In *Workshop on Statistical Signal Processing Workshop*, 2011.

[134] L. Lovász. Submodular functions and convexity. *Mathematical programming: The state of the art, Bonn*, pages 235–257, 1982.

[135] M. E. Lübbecke and J. Desrosiers. Selected topics in column generation. *Operations Research*, 53(6):1007–1023, 2005.

[136] R. Luss, S. Rosset, and M. Shahar. Decomposing isotonic regression for efficiently solving large problems. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[137] R. Luss, S. Rosset, and M. Shahar. Efficient regularized isotonic regression with application to gene–gene interaction search. *Annals of Applied Statistics*, 6(1):253–283, 2012.

[138] N. Maculan and G. Galdino de Paula. A linear-time median-finding algorithm for projecting a vector on the simplex of $\mathbb{R}^n$. *Operations Research Letters*, 8(4):219–222, 1989.

[139] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(1):19–60, 2010.

[140] J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Convex and network flow optimization for structured sparsity. *Journal of Machine Learning Research*, 12:2681–2720, 2011.

[141] J. Mairal and B. Yu. Complexity analysis of the Lasso regularization path. In *Proceedings of the International Conference on Machine Learning (ICML*, 2012.

[142] C. L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.

[143] J. L. Marichal. An axiomatic approach of the discrete Choquet integral as a tool to aggregate interacting criteria. *IEEE Transactions on Fuzzy Systems*, 8(6):800–807, 2000.

[144] H. Markowitz. Portfolio selection. *Journal of Finance*, 7(1):77–91, 1952.

[145] S. T. McCormick. Submodular function minimization. *Discrete Optimization*, 12:321–391, 2005.

[146] N. Megiddo. Optimal flows in networks with multiple sources and sinks. *Mathematical Programming*, 7(1):97–107, 1974.

[147] M. Minoux. Accelerated greedy algorithms for maximizing submodular set functions. *Optimization Techniques*, pages 234–243, 1978.

[148] J. J. Moreau. Fonctions convexes duales et points proximaux dans un espace Hilbertien. *Comptes-Rendus de l'Académie des Sciences, Série A Mathématiques*, 255:2897–2899, 1962.

[149] J. R. Munkres. *Elements of algebraic topology*, volume 2. Addison-Wesley Reading, MA, 1984.

[150] K. Murota. *Discrete convex analysis*, volume 10. Society for Industrial Mathematics, 1987.

[151] H. Nagamochi and T. Ibaraki. A note on minimizing submodular functions. *Information Processing Letters*, 67(5):239–244, 1998.

[152] K. Nagano. A strongly polynomial algorithm for line search in submodular polyhedra. *Discrete Optimization*, 4(3-4):349–359, 2007.

[153] K. Nagano, Y. Kawahara, and K. Aihara. Size-constrained submodular minimization through minimum norm base. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2011.

[154] M. Narasimhan and J. Bilmes. PAC-learning bounded tree-width graphical models. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2004.

[155] M. Narasimhan and J. Bilmes. A submodular-supermodular procedure with applications to discriminative structure learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

[156] M. Narasimhan and J. Bilmes. Local search for balanced submodular clusterings. In *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI)*, 2007.

[157] M. Narasimhan, N. Jojic, and J. Bilmes. Q-clustering. *Advances in Neural Information Processing Systems (NIPS)*, 2006.

[158] H. Narayanan. A rounding technique for the polymatroid membership problem. *Linear algebra and its applications*, 221:41–57, 1995.

[159] H. Narayanan. *Submodular Functions and Electrical Networks*. North-Holland, 2009. Second edition.

[160] A. Nedić and A. Ozdaglar. Approximate primal solutions and rate analysis for dual subgradient methods. *SIAM Journal on Optimization*, 19(4), February 2009.

[161] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[162] S. Negahban and M. J. Wainwright. Joint support recovery under high-dimensional scaling: Benefits and perils of ℓ_1-ℓ_∞-regularization. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[163] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions–I. *Mathematical Programming*, 14(1):265–294, 1978.

[164] A. Nemirovski. Efficient methods in convex programming. Technical Report Lecture notes, Technion, Israel Institute of Technology, 1994.

[165] A. Nemirovski, S. Onn, and U. G. Rothblum. Accuracy certificates for computational problems with convex structure. *Mathematics of Operations Research*, 35(1):52–78, 2010.

[166] A. S. Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in optimization*. John Wiley, 1983.

[167] Y. Nesterov. Complexity estimates of some cutting plane methods based on the analytic barrier. *Mathematical Programming*, 69(1-3):149–176, 1995.

[168] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.

[169] Y. Nesterov. Gradient methods for minimizing composite objective function. Technical report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, 2007.

[170] Y. Nesterov, A. S. Nemirovski, and Y. Ye. *Interior-point polynomial algorithms in convex programming*, volume 13. SIAM, 1994.

[171] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 2nd edition, 2006.

[172] G. Obozinski and F. Bach. Convex relaxation of combinatorial penalties. Technical Report 00694765, HAL, 2012.

[173] G. Obozinski, G. Lanckriet, C. Grant, M. I. Jordan, and W. S. Noble. Consistent probabilistic outputs for protein function prediction. *Genome Biology*, 9(Suppl 1):S6, 2008.

[174] J. B. Orlin. A faster strongly polynomial time algorithm for submodular function minimization. *Mathematical Programming*, 118(2):237–251, 2009.

[175] M. R. Osborne, B. Presnell, and B. A. Turlach. On the Lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–37, 2000.

[176] S. Osher. Level set methods. *Geometric Level Set Methods in Imaging, Vision, and Graphics*, pages 3–20, 2003.

[177] F. Pukelsheim. *Optimal design of experiments*, volume 50. Society for Industrial Mathematics, 2006.

[178] M. Queyranne. Minimizing symmetric submodular functions. *Mathematical Programming*, 82(1):3–12, 1998.

[179] M. Queyranne and A. Schulz. Scheduling unit jobs with compatible release dates on parallel machines with nonstationary speeds. *Integer Programming and Combinatorial Optimization*, 920:307–320, 1995.

[180] N. S. Rao, R. D. Nowak, S. J. Wright, and N. G. Kingsbury. Convex approaches to model wavelet sparsity patterns. In *Proceedings of the International Conference on Image Processing (ICIP)*, 2011.

[181] C. E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[182] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1997.

[183] M. J. Schell and B. Singh. The reduced monotonic regression method. *Journal of the American Statistical Association*, 92(437):128–135, 1997.

[184] M. Schmidt and K. Murphy. Convex structure learning in log-linear models: Beyond pairwise potentials. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

[185] A. Schrijver. A combinatorial algorithm minimizing submodular functions in strongly polynomial time. *Journal of Combinatorial Theory, Series B*, 80(2):346–355, 2000.

[186] A. Schrijver. *Combinatorial optimization: Polyhedra and efficiency*. Springer, 2004.

[187] M. Seeger. On the submodularity of linear experimental design, 2009. `http://lapmal.epfl.ch/papers/subm_lindesign.pdf`.

[188] M. W. Seeger. Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, 9:759–813, 2008.

[189] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[190] D. A. Spielman and S.-H. Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM*, 51(3):385–463, 2004.

[191] P. Sprechmann, I. Ramirez, G. Sapiro, and Y. Eldar. Collaborative hierarchical sparse modeling. In *Proceedings of the Conference on Information Sciences and Systems (CISS)*, 2010.

[192] R. P. Stanley. *Enumerative combinatorics*, volume 49. Cambridge University Press, 2011.

[193] P. Stobbe and A. Krause. Efficient minimization of decomposable submodular functions. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[194] P. Stobbe and A. Krause. Learning Fourier sparse set functions. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

[195] M. Streeter and D. Golovin. An online algorithm for maximizing submodular functions. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[196] R. Tarjan, J. Ward, B. Zhang, Y. Zhou, and J. Mao. Balancing applied to maximum network flow problems. *Algorithms–ESA 2006*, pages 612–623, 2006.

[197] C. H. Teo, S. V. N. Vishwanthan, A. J. Smola, and Q. V. Le. Bundle methods for regularized risk minimization. *Journal of Machine Learning Research*, 11:311–365, 2010.

[198] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

[199] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 67(1):91–108, 2005.

[200] D. M. Topkis. *Supermodularity and complementarity*. Princeton University Press, 2001.

[201] G. Varoquaux, R. Jenatton, A. Gramfort, G. Obozinski, B. Thirion, and F. Bach. Sparse structured dictionary learning for brain resting-state activity modeling. In *NIPS Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions*, 2010.

[202] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

[203] P. Wolfe. Finding the nearest point in a polytope. *Mathematical Programming*, 11(1):128–149, 1976.

[204] L. A. Wolsey. Maximising real-valued submodular functions: Primal and dual heuristics for location problems. *Mathematics of Operations Research*, 7(3):410–425, 1982.

[205] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

[206] M. Yuan and Y. Lin. On the non-negative garrotte estimator. *Journal of The Royal Statistical Society Series B*, 69(2):143–161, 2007.

[207] A. L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15(4):915–936, 2003.

[208] X. Zhang, Y. Yu, and D. Schuurmans. Accelerated training for matrix-norm regularization: A boosting approach. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[209] Z. Zhang and R. W. Yeung. On characterization of entropy function via information inequalities. *IEEE Transactions on Information Theory*, 44(4):1440–1452, 1998.

[210] P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 37(6A):3468–3497, 2009.

[211] P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

[212] S. Zivnı, D. A. Cohen, and P. G. Jeavons. The expressive power of binary submodular functions. *Discrete Applied Mathematics*, 157(15):3347–3358, 2009.