Generalized Conditional Independence and Decomposition Cognizant Curvature: Implications for Function Optimization

Mathias Niepert
Computer Science and Engineering
University of Washington
Seattle, WA, 98195

Pedro Domingos
Computer Science and Engineering
University of Washington
Seattle, WA, 98195

Jeff Bilmes
Department of Electrical Engineering
University of Washington
Seattle, WA, 98195

Abstract

We introduce conditional independence for real-valued set functions which generalizes probabilistic conditional independence. We show that a natural semantics of conditional independence is that of *local modularity*. Generalized conditional independence leads to a spectrum between two extremes: modular functions and functions without any local modularity. We develop a decomposition theory and relate the results to function optimization, an important problem with numerous applications in machine learning. Moreover, we derive a connection between conditional independence and the curvature of a set function. Intuitively, the more conditional independencies hold, the more accurate the solutions of approximation algorithms. Conditional independence for function optimization and approximation problems could have an impact similar to the impact conditional independence has had on probabilistic inference and learning.

1 Introduction

There has been a surge in research on algorithms, both exact and approximate, for submodular function optimization [4]. Submodular functions occur in numerous applications [10, 12, 13, 14, 8] and are discrete functions that have properties analogue to convex and concave functions. Existing optimization algorithms take advantage of the submodularity of the function. Examples include greedy algorithms with good approximation guarantees [15] and graph cut algorithms [9]. Empirical results have also shown that instances of NP-hard optimization problems are often efficiently solvable using these algorithms. Nevertheless, a comprehensive theory that relates notions of conditional independence to function decompositions and approximation algorithms, a theory that has led to impressive algorithmic improvements in the context of probabilistic models, is largely missing. We introduce a generalization of conditional independence and relate it to discrete optimization problems. Maximum a-posteriori inference in probabilistic graphical models corresponds to an optimization problem of a particular real-valued function. This optimization problem is intractable even for restricted model classes such as Bayesian and Markov networks. Despite seemingly daunting intractability results, graphical models have been successfully applied in numerous problem domains and are still subject of intense research activity. To a large extent, problem specific conditional independence relationships among the random variables and the resulting speed-up in inference algorithms are part of the success of graphical models. Similar to probabilistic inference, numerous other optimization problems in the realm of artificial intelligence, data mining, and machine learning, are intractable in their most general form. For instance, it is known that maximizing a submodular function is an intractable, that is, an NP-hard computational problem. In many cases, these problems are defined over graphical structures (max cut, sensor placement, etc.). Hence, we propose to consider a generalized form of conditional independence in the realm of function optimization. We exploit the notion of independence to develop a decomposition theory that generalizes existing decomposition results for submodular functions [3]. Moreover, we show that conditional independence directly influences the curvature of functions and the ability to decompose functions into smaller components.

We hope the presented work is welcomed as a first step towards a deeper understanding of the connections between conditional independence, (graph) decompositions, curvature, and approximation algorithms for real-valued functions on sets.

2 Background

We introduce some basic concepts related to real-valued set functions of the form $f : 2^V \to \mathbb{R}$. We assume, without loss of generality, that $f(\emptyset) = 0$ and that f is not infinite-valued.

2.1 Submodular and Supermodular Functions

Submodular functions (and their dual, supermodular functions) have attracted a considerable amount of attention from the machine learning community. We provide the definitions of important notions used throughout the paper. We often write, for two sets A and B, AB instead of $A \cup B$.

Definition 2.1. Let V be a finite set and let $f : 2^V \to \mathbb{R}$ be a function from subsets of V into the reals. The function f is

- *submodular* if, for all $S, T \subseteq V$, $f(S \cup T) + f(S \cap T) \leq f(S) + f(T)$;
- *supermodular* if, for all $S, T \subseteq V$, $f(S \cup T) + f(S \cap T) \geq f(S) + f(T)$; and
- *modular* if, for all $S, T \subseteq V$, $f(S \cup T) + f(S \cap T) = f(S) + f(T)$.

2.2 Function Optimization

Let V be a finite set and let $f : 2^V \to \mathbb{R}$ be a function from subsets of V into the reals. We are concerned with optimization problems of the form

$$\max_{X \subseteq V} f(X) \quad \text{and} \quad \min_{X \subseteq V} f(X).$$

There are numerous extensions of the optimization problems. Almost all of these involve additional constraints that are imposed on the optimization problem. For instance, one often wants to maximize a submodular function subject to cardinality constraints. These additional constraints often render the problem more difficult. For example, while the minimization of submodular functions is possible in polynomial time, additional lower bound cardinality constraints render the problem intractable.

A concept that is used to assess approximation algorithms for optimization problems (cf. [2, 7]) is the total curvature κ_f of a real-valued function f and the curvature $\kappa_f(S)$ of a function f with respect to a set $S \subseteq V$. Let $f(A \mid B) = f(A \cup B) - f(B)$. Then,

$$\kappa_f = 1 - \min_{j \in V} \frac{f(j \mid V \setminus j)}{f(j)}; \kappa_f(S) = 1 - \min_{j \in S} \frac{f(j \mid V \setminus j)}{f(j)}.$$

3 Generalized Conditional Independence

We define generalized conditional independence for real-valued set functions as *local modularity*.

Definition 3.1 (Generalized Conditional Independence)**.** Let V be a finite set and let $f : 2^V \to \mathbb{R}$. Let $A, B, C \subseteq V$ be mutually disjoint subsets of V. We write that A and B are *conditionally independent* given C, denoted $(A \perp\!\!\!\perp B \mid C)$, if and only if

$$f(ABC) + f(C) = f(AC) + f(BC).$$

$(A \perp\!\!\!\perp B \mid C)$	$\Rightarrow (B \perp\!\!\!\perp A \mid C)$	
$(A \perp\!\!\!\perp BD \mid C)$	$\Rightarrow (A \perp\!\!\!\perp D \mid C)$	$(\{a\} \perp\!\!\!\perp \{b\} \mid CD) \quad \Rightarrow (\{a\} \perp\!\!\!\perp \{b\} \mid C)$
$(A \perp\!\!\!\perp B \mid CD)$ & $(A \perp\!\!\!\perp D \mid C)$	$\Rightarrow (A \perp\!\!\!\perp BD \mid C)$	
$(A \perp\!\!\!\perp BD \mid C)$	$\Rightarrow (A \perp\!\!\!\perp B \mid CD)$	

Figure 1: The semi-graphoid axioms (left) and the Reduction Axiom (right).

Alternatively, we write that f *satisfies the CI statement* $(A \perp\!\!\!\perp B \mid C)$. If $ABC = V$ we say that the CI statement is *saturated*. If $C = \emptyset$ we say that A an B are independent and write $(A \perp B)$. In this case we have that

$$f(AB) = f(A) + f(B).$$

Note that $(A \perp\!\!\!\perp B \mid C)$ is equivalent to f exhibiting local modularity on the sets AC and BC. Intuitively, a generalized CI statement expresses that particular differentials of the function are zero. This simple and intuitive definition of CI turns out to be surprisingly powerful in assessing the hardness of optimization problems. For one, it captures the notion of probabilistic conditional independence. To show this, we leverage a mapping from probability distributions to real-valued set functions using the multi-information function [19]. This mapping was introduced by Studený to connect probabilistic CI statements to the notion of imsets, an algebraic representation of supermodular functions, in an attempt to more thoroughly characterize probabilistic conditional independence. We use $\mathbf{V}$ to denote a set of discrete random variables and $\mathbf{a}$ to denote an assignment to random variables $\mathbf{A} \subseteq \mathbf{V}$.

Theorem 3.2 (Studený [19])**.** *Let* $\Pr$ *be any discrete probability distribution over a set of random variables* $\mathbf{V}$*. Then, there exists a supermodular function* $m_{\Pr} : 2^{\mathbf{V}} \to \mathbb{R}$ *with the following property. For* $\mathbf{A}, \mathbf{B}, \mathbf{C}$ *disjoint subsets of* $\mathbf{V}$ *and possible assignments* $\mathbf{a}, \mathbf{b}$*, and* $\mathbf{c}$ *to these sets of variables, we have that* $\Pr(\mathbf{C} = \mathbf{c}) \Pr(\mathbf{A} = \mathbf{a}, \mathbf{B} = \mathbf{b}, \mathbf{C} = \mathbf{c}) = \Pr(\mathbf{A} = \mathbf{a}, \mathbf{C} = \mathbf{c}) \Pr(\mathbf{B} = \mathbf{b}, \mathbf{C} = \mathbf{c})$ *if and only if* $m_{\Pr}(\mathbf{ABC}) + m_{\Pr}(\mathbf{C}) = m_{\Pr}(\mathbf{AC}) + m_{\Pr}(\mathbf{BC})$.

It is the above connection to probabilistic conditional independence that motivates the generalization of conditional independence. The success-story of probabilistic graphical models has largely been driven by a thorough understanding of the algorithmic and logical properties of CI. We want to explore the extend to which a more general notion of CI as local modularity leads to a deeper understanding of optimization problems.

4 Axioms of Independence

A deep understanding of the algorithmic and logical properties of conditional independence has led to the success story of graphical models such as Bayesian and Markov networks and remains an active research area [18, 19, 17, 16, 5]. Most textbooks on probabilistic graphical models discuss axioms of independence with an emphasis on the semi-graphoid axioms (see Figure 1) and axioms systems that characterize factorizations of a probability distribution with respect to particular graphical models (cf. [18, 11]). The semi-graphoid axioms hold for the probabilistic notion of conditional independence. Axioms are horn rules that express what conditional independencies are implied by a conjunction of given CI statements. We explore axioms of independence for the general notion of conditional independence as local modularity and for general set functions.

Definition 4.1 (Soundness)**.** Let $\mathcal{F}$ be a *class* of real-valued set functions. We say that an axiom $\mathsf{A}_1 \wedge ... \wedge \mathsf{A}_k \Rightarrow \mathsf{C}$ is *sound* relative to $\mathcal{F}$ if, for each $f \in \mathcal{F}$ we have that if f satisfies $\mathsf{A}_1, ..., \mathsf{A}_k$ then f also satisfies C. We say that an axiom system is sound if each of its axioms is sound.

We can now state the following crucial result. The theorem was proved for supermodular functions using imsets, an algebraic representation of supermodular functions [19].

Theorem 4.2. *The semi-graphoid axioms are sound for submodular (supermodular) functions.*

The *Reduction Axiom* depicted in Figure 1 is central to the development of a graph decomposition theory. We state the soundness of said axiom for modular and cut capacity functions.

Lemma 4.3. *The* Reduction Axiom *is sound for modular functions and cut capacity functions.*

The *Reduction Axiom* is also sound for other (submodular) set functions. Moreover, one can show that the *Reduction Axiom* is sound for log-probability functions with respect to probabilistic CI.

Lemma 4.4 (cf. Besag [1]). *Let* $\mathbf{V}$ *be a set of discrete random variables with positive probability distribution* P *and let* H *be its log-probability function. If* P *satisfies the probabilistic CI statement* $(\{a\}\bot\{b\} \mid \mathbf{C})$ *then, for all* $\mathbf{C}' \subseteq \mathbf{C}$, $H(\mathbf{C}' \cup \{a, b\}) + H(\mathbf{C}') = H(\mathbf{C}' \cup \{a\}) + H(\mathbf{C}' \cup \{b\})$ *and, therefore,* $(\{a\}\perp\!\!\!\perp\{b\} \mid \mathbf{C}')$.

5 Decompositions

Based on the notion of generalized conditional independence we develop a decomposition theory that generalizes and complements decomposition results in the literature [3]. The strongest form of a decomposition is possible when the set function is modular. For every set function f, we can relate its modularity, its curvature, and the CI statements it satisfied. We begin by formulating a decomposition theorem for submodular and supermodular functions.

Theorem 5.1. *Let* V *be a finite set and let* $f : 2^V \to \mathbb{R}$ *be a submodular (supermodular) function from subsets of* V *into the reals. Then,* f *satisfies* $(S\perp\!\!\!\perp T \mid U)$ *if and only if* $f(XU) = f((X \cap S)U) + f((X \cap T)U) - f(U)$ *for all* $X \subseteq ST$*. Specifically, we have that* f *satisfies* $(S \perp T)$ *if and only if* $f(X) = f(X \cap S) + f(X \cap T)$ *for all* $X \subseteq ST$.

Theorem 5.1 is a generalization of existing decomposition results where Theorem 5.1 is stated for $ST = V$ [3], that is, for saturated CI statements. It has several noteworthy implications. For one, it paves the way to decompositions of the problem into independent parts. For instance, if for a given set V we know that there is a partition $\mathcal{C}$ of V such that $(C \perp V \setminus \{C\})$ for all $C \in \mathcal{C}$, then the overall optimization problem can be solved by solving the optimization problems $\max_{X \subseteq C} f(X)$ for each $C \in \mathcal{C}$ independently. The solution of the global optimization problem is now simply the sum of the solution of the subproblems. In many realistic scenarios, a decomposition into disconnected components is not possible. Since numerous instances of submodular function optimization problems originate from graph problems, we turn our attention to graph-based decompositions.

5.1 Graph Decompositions

We now define the pairwise Markov property and graph factorization for real-valued set functions.

Definition 5.2 (Pairwise Markov property). Let $f : 2^V \to \mathbb{R}$ be a function from subsets of V into the reals and let G be an undirected graph with node set V and edge set $\mathcal{E}$. We say that f satisfies the pairwise Markov property with respect to G if f satisfies each of the CI statements in $\{(v_1 \perp\!\!\!\perp v_2 \mid V \setminus \{v_1, v_2\}) : \{v_1, v_2\} \notin \mathcal{E}\}$.

Definition 5.3 (Graph factorization). Let $f : 2^V \to \mathbb{R}$ and let G be an undirected graph with node set V and clique set $\mathcal{C}$. We say that f *factorizes according to* G if it can be written as

$$f(X) = \sum_{C \in \mathcal{C}} f_C(X), \tag{1}$$

where $f_C(X)$ depends on X only through $C \cap X$.

Note that if a function factorizes over the cliques of G then it also factorizes over the maximal cliques of G. We are now in the position to prove the Hammersley-Clifford theorem [6, 1] for real-valued set functions for which the *Reduction* axiom is sound.

Theorem 5.4. *Let* $f : 2^V \to \mathbb{R}$ *be a function from subsets of* V *into the reals and let the* Reduction Axiom *be sound relative to* f*. Then* f *satisfies the pairwise Markov property with respect to an undirected graph* G *if and only if it factorizes according to* G.

In conjunction with Lemma 4.4, the theorem allows us to prove a generalization of the Hammersley-Clifford theorem for positive probability distributions.

6 Function Optimization

Can we exploit generalized CI for optimization problems on real-valued set functions? We show that both exact and approximate algorithms can indeed benefit in various ways. If V can be partitioned into independent components, we can solve optimization problems for submodular set functions on V by solving a sequence of smaller problems.

Corollary 6.1. *Let V be a finite set and let $f : 2^V \to \mathbb{R}$ be a submodular (supermodular) function from subsets of V into the reals. Moreover, let $\mathcal{C} = C_1, ..., C_k$ be a partition of V such that, for each $1 \leq i \leq k$, f satisfies $(C_i \perp V \setminus C_i)$. Then,*

$$\max_{X \subseteq V} f(X) = \sum_{\ell=1}^{k} \max_{X \subseteq C_\ell} f(X).$$

Note that the above corollary applies to *submodular functions* and does *not* hold for arbitrary classes of set functions. In the best case, the decomposition of the corollary leads to exponential speed-ups. If f is modular, then f satisfies $(\{j\} \perp V \setminus j)$ for all $j \in V$ and, hence, $\mathcal{C} = \{\{j\} \mid j \in V\}$.

In many applications, the optimization problem features additional constraints such as cardinality or Knapsack constraints which introduce global dependencies between the components. However, non-serial dynamic programming can be applied to combine the local optima into a global one. This is reminiscent of variable elimination in the probabilistic graphical models literature. For instance, for cardinality and Knapsack constraints, the global solution can be computed from the local solutions in polynomial time using dynamic programming. To see this, consider an algorithm which, given a finite set V with $|V| = n$, computes in iteration $i, 1 \leq i \leq n$, a set S_i of size i that it deems maximal among all sets of size i and which satisfies the given constraints. Now, for a given decomposition $\mathcal{C}$, we apply the algorithm to each $C \in \mathcal{C}$ and obtain the sets S_i^C with value $f(S_i^C)$, for $1 \leq i \leq |C|$. We can now apply non-serial dynamic programming to combine the partial solutions, associated with each of the components $C \in \mathcal{C}$, into a global solution that satisfies the constraints.

6.1 Independence and Curvature

We show that approximation algorithms compute more accurate solutions to optimization problems when the problems are decomposable. Approximation bounds are often expressed using the notion of curvature [2, 7]. Here, we define the notion of a curvature *relative to a set $C \subseteq V$*.

Definition 6.2. Let V be a finite set and let $f : 2^V \to \mathbb{R}$ be a submodular function from subsets of V into the reals. We define the curvature of f relative to a subset C of V as

$$\kappa_f^C = 1 - \min_{j \in C} \frac{f(j \mid C \setminus j)}{f(j)}; \quad \kappa_f^C(S) = 1 - \min_{j \in S} \frac{f(j \mid C \setminus j)}{f(j)}.$$

Note that $\kappa_f^C \leq \kappa_f = \kappa_f^V$ and, for all $S \subseteq V$, $\kappa_f^C(S) \leq \kappa_f(S) = \kappa_f^{V(S)}$ by the submodularity of f. We can now show that the curvature of f for V is related to a given decomposition $\mathcal{C}$.

Theorem 6.3. *Let V be a finite set and let $f : 2^V \to \mathbb{R}$ be a function from subsets of V into the reals and let f factorize according to $\mathcal{C}$ or to the graph G which has cliques $\mathcal{C}(G)$. Then,*

$$\kappa_f = 1 - \min_{j \in V} \sum_{C \in \mathcal{C}: j \in C} \frac{f(j \mid C \setminus j)}{f(j)}; \quad \textit{and} \quad \kappa_f(S) = 1 - \min_{j \in S} \sum_{C \in \mathcal{C}: j \in C} \frac{f(j \mid C \setminus j)}{f(j)}.$$

If $\mathcal{C}$ is a partition of V, then we can write

$$\kappa_f = \max_{C \in \mathcal{C}} \kappa_f^C \quad \textit{and} \quad \kappa_f(S) = \max_{C \in \mathcal{C}} \kappa_f^C(S).$$

The theorem establishes a relationship between generalized CI and the curvature of a set function. For instance, if $\mathcal{C}$ is a partition of V, then the curvature of f is bounded by the largest curvature of f on any of the individual components in $\mathcal{C}$. Thus, comparing two functions f_1 and f_2 where f_2 satisfies a superset of the conditional independence statements satisfied by f_1, the curvature can only decrease, leading to better approximation guarantees. In other words, there is a strong relationship between decomposition in the standard sense of a graphical model, and curvature in the sense of a submodular function. This motivates the use of decomposition cognizant curvature for assessing the quality of approximations. The greedy algorithm for maximizing monotone submodular functions has a worst-case approximation bound of $\frac{1}{\kappa_f}(1 - e^{-\kappa_f})$ [2]. Moreover, there is a large body of work on relating the curvature to the problem of approximating a submodular function, learning a submodular function, and minimizing a submodular function [7]. The results translate to these guarantees and provide a link between conditional independence and the goodness of the approximations. For instance, it might be possible to approximately learn submodular functions subject to the curvature being bounded by an ϵ. This would translate to learning a submodular function subject to constraints on the (graphical) decomposition structure.

7 Discussion

We introduced a generalized form of conditional independence based on the notion of local modularity. The well-known concept of probabilistic conditional independence is captured by this notion. In numerous applications, it is possible to extract conditional independence statements. For instance, in a sensor placement scenarios, we might know from the structure of the building that certain conditional independencies hold – the addition of two sensors to two separate rooms are independent events. If the problem is defined via a graph representation (such as the maximum cut problem) we can read the conditional independence statements directly from the graph. Conditional independence and the resulting decompositions have led to a better understanding of algorithmic problems that are generally intractable such as MAP inference in Markov networks. We hope that the notion, generalized to set functions, can have a similar impact on optimization and learning set functions.

References

[1] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B*, 36:192–236, 1974.

[2] M. Conforti and G. Cornuéjols. Submodular set functions, matroids and the greedy algorithm: Tight worst-case bounds and some generalizations of the rado-edmonds theorem. *Discrete Applied Mathematics*, 7(3):251 – 274, 1984.

[3] W. H. Cunningham. Decomposition of submodular functions. *Combinatorica*, 3(1):53–68, 1983.

[4] S. Fujishige. *Submodular Functions and Optimization*. Annals of Discrete Mathematics. Elsevier, 2005.

[5] M. Gyssens, M. Niepert, and D. V. Gucht. On the completeness of the semigraphoid axioms for deriving arbitrary from saturated conditional independence statements. *Information Processing Letters*, 114(11):628–633, 2014.

[6] J. M. Hammersley and P. E. Clifford. Markov random fields on finite graphs and lattices. Unpublished manuscript, 1971.

[7] R. K. Iyer, S. Jegelka, and J. A. Bilmes. Curvature and optimal algorithms for learning and minimizing submodular functions. In *Advances in Neural Information Processing Systems 26*, pages 2742–2750. 2013.

[8] S. Jegelka and J. Bilmes. Submodularity beyond submodular energies: Coupling edges in graph cuts. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1897–1904, 2011.

[9] S. Jegelka, H. Lin, and J. A. Bilmes. On fast approximate submodular minimization. In *Neural Information Processing Society (NIPS)*, pages 460–468, 2011.

[10] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146, 2003.

[11] D. Koller and N. Friedman. *Probabilistic Graphical Models*. The MIT Press, 2009.

[12] A. Krause and C. Guestrin. Near-optimal observation selection using submodular functions. In *National Conference on Artificial Intelligence (AAAI)*, pages 1650–1654, 2007.

[13] A. Krause, J. Leskovec, C. Guestrin, J. VanBriesen, and C. Faloutsos. Efficient sensor placement optimization for securing large water distribution networks. *Journal of Water Resources Planning and Management*, 134(6):516–526, 2008.

[14] R. C. Moore and W. Lewis. Intelligent selection of language model training data. In *Proceedings of the ACL Conference (ACL)*, pages 220–224, 2010.

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

[16] M. Niepert, D. V. Gucht, and M. Gyssens. Logical and algorithmic properties of stable conditional independence. *International Journal of Approximate Reasoning*, 51(5):531–543, 2010.

[17] M. Niepert, M. Gyssens, B. Sayrafi, and D. Van Gucht. On the conditional independence implication problem: A lattice-theoretic approach. *Artificial Intelligence*, 202(0):29–51, 2013.

[18] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., 1988.

[19] M. Studený. *Probabilistic Conditional Independence Structures*. Springer-Verlag, 2005.