Submodular Maximization and Diversity in Structured Output Spaces

Adarsh Prasad
Virginia Tech, UT Austin
adarshprasad27@gmail.com

Stefanie Jegelka
UC Berkeley
stefje@eecs.berkeley.edu

Dhruv Batra
Virginia Tech
dbatra@vt.edu

Abstract

We study the greedy maximization of a submodular set function $F : 2^V \rightarrow \mathbb{R}$ when each item in the ground set V is itself a *combinatorial object*, *e.g.* a configuration or labeling of a base set of variables $\mathbf{z} = \{z_1, ..., z_m\}$. This problem arises naturally in a number of domains, such as Computer Vision or Natural Language Processing, where we want to search for a set of *diverse high-quality* solutions in a structured-output space. Unfortunately, if the space of items is exponentially large, even one linear scan for greedy search is infeasible. We show that when marginal gains of such submodular functions allow structured representations, this enables efficient (sub-linear time) maximization by reducing the greedy augmentation step to an inference problem on a graphical model with appropriately constructed High-Order Potentials (HOPs).

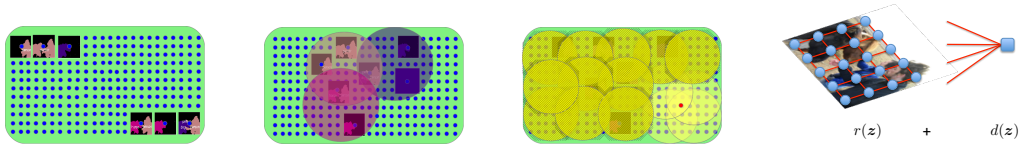
1 Introduction

Many problems in Computer Vision, Natural Language Processing and Computational Biology involve predictions that form a mapping from an input space $\mathcal{X}$ to an exponentially large space $\mathcal{Y}$ of *structured outputs*. For instance, $\mathcal{Y}$ may be the space of all possible segmentations, $|\mathcal{Y}| = L^m$, where m is the number of pixels, and L is the number of object labels that each pixel can take. The factorization of a structured-output into its parts is efficiently exploited by Conditional Random Fields (CRFs) [1], Max-Margin Markov Networks (M^3N) [2], and Structured Support Vector Machines (SSVMs) [3], providing principled models for scoring all solutions $\mathbf{y} \in \mathcal{Y}$ and predicting the highest scoring or maximum *a posteriori* (MAP) configuration.

In a number of scenarios, we seek not only a single prediction but a *set* of good predictions: (1) **Interactive Machine Learning.** Systems like Google Translate (for machine translation) or Photoshop (for interactive image segmentation) solve structured prediction problems; since there is a user in the loop, such systems may produce a small set of relevant outputs and simply let the user pick the best one; (2) **M-Best hypotheses in cascades.** Machine learning algorithms are often cascaded, with the output of one model being fed into another. In such a setting, at the initial stages it is not necessary to make the perfect prediction; the goal is rather to make a set of *plausible* predictions, which may then be re-ranked, combined or processed by a more sophisticated mechanism.

The defining characteristic in many such scenarios is that we need to generate a *diverse set of plausible* structured-outputs $\mathbf{Y} = \{\mathbf{y}^1, \ldots, \mathbf{y}^M\}$, $\mathbf{y}^i \in \mathcal{Y}$ that will be handed to an algorithm or expert downstream.

Submodular Maximization and Diversity. The problem of searching for a diverse but high-quality subset of items in a ground set V of N items has been studied in information retrieval, web search, sensor placement, document summarization, viral marketing and robotics. In many of these works, an effective, theoretically-grounded and practical tool for measuring the diversity of a set $S \subseteq V$ are *submodular* set functions. Submodularity is a property that captures the idea of diminishing



(a) V:All Segmentations. (b) Label Cost Groups. (c) Diminishing Returns. (d) Structured Representation.

Figure 1: Overview: (a) space of all possible object segmentations / labelings of an image (each item is a segmentation); (b) partition of the output space into overlapping groups; (b) 'label cost' groups, where each circle represents a set of segmentations G_ℓ that contain label ℓ; (c) diminishing returns – the number of unused groups becomes smaller as we pick more items; (d) we convert the problem of finding the item with the highest marginal gain $F(a \mid A)$ into a MAP inference problem in a graphical model over base variables $\mathbf{z}$ with an appropriately defined HOP.

marginal returns – a set function $F : 2^V \rightarrow \mathbb{R}$ is submodular if its marginal gains, $F(a \mid S) \equiv F(S \cup a) - F(S)$, are decreasing: $F(a \mid S) \geq F(a \mid T)$ for all $S \subseteq T$ and $a \notin T$. In addition, if F is *monotone*, *i.e.*, $F(S) \leq F(T)$ whenever $S \subseteq T$, then a simple greedy algorithm that iteratively picks the element with the largest marginal gain $F(a \mid S)$ to add to the current set S, achieves the best possible approximation bound of $(1 - \frac{1}{e})$ [4]. This result has had significant practical impact. Unfortunately, if the number $N = |V|$ of items is exponentially large, then even a single linear scan for greedy augmentation is simply infeasible.

In this work, we study conditions under which it is still feasible to maximize a submodular function over an exponentially-large ground set $V = \{v_1, \ldots, v_N\}$ whose elements are *combinatorial objects*, such as configurations or labelings of some *base set* $\mathbf{z} = \{z_1, z_2, \ldots, z_m\}$ of m variables. For instance, in image segmentation, the base variables z_i are pixel labels, and each element $a \in V$ is a particular labeling of these pixels. As another example, each base variable z_e could indicate the presence or absence of an edge e in a graph, and then each configuration may represent a spanning tree or a maximal matching. Our goal will be to efficiently find a set of M plausible and diverse configurations. This means we are aiming for algorithms *sub-linear* in $|V|$ – ideally, scaling as a very low-order polynomial in $\log |V|$. We will mostly assume the submodular function to be monotone, nonnegative and normalized ($F(\emptyset) = 0$) and therefore base our study on the natural greedy algorithm. For illustration, we will focus on pixel labels in this paper, where each base variable takes values in a set $[L] = \{1, \ldots, L\}$ of labels, where $[k]$ is shorthand for $\{1, \ldots, k\}$.

Contributions. First, we show that when marginal gains of submodular functions allow structured representations (*e.g.* as a graphical model over $\mathbf{z}$), this enables efficient submodular maximization over exponentially-large ground sets. As examples, we construct submodular functions for three different, task-dependent definitions of diversity, and demonstrate how the greedy augmentation step can be reduced to a MAP inference problem in a discrete graphical model augmented with a suitably constructed *High-Order Potential* (HOP), for which efficient inference techniques have already been developed in the graphical models community. Thus, our work draws connections between two seemingly disparate but highly related areas in machine learning – submodular maximization and inference in graphical models with structured HOPs. Fig. 1 shows an overview of our approach.

Notation. For clarity, we denote a single item (combinatorial object) by $a \in V$. We use upper case letters for functions over the ground set items $F(a \mid A), R(a \mid A), D(a \mid A)$, and lower case letters for functions over the base set variables $f(\mathbf{z}), r(\mathbf{z}), d(\mathbf{z})$. Let $\phi : V \mapsto [L]^m$ be the bijection that maps any item in V to its corresponding vector representation in terms of the base variables. Conversely, $\phi^{-1} : [L]^m \mapsto V$ maps a labeling $\mathbf{z}$ to its corresponding item $\phi^{-1}(\mathbf{z}) \in V$. With a slight abuse of notation, we may use $\mathbf{z} \in S$ to mean $\phi^{-1}(\mathbf{z}) \in S$, *i.e.* the item corresponding to the labeling $\mathbf{z}$ is present in set S; and $\ell \in \mathbf{z}$ to mean the label ℓ is used in $\mathbf{z}$, *i.e.* $\exists j \;\; s.t. \;\; z_j = \ell$. For a set $c \subseteq [m]$, we use z_c to denote the tuple $\{z_i \mid i \in c\}$.

2 Diversity model and optimization

We measure diversity by a monotone, nondecreasing and normalized submodular function $D : 2^V \rightarrow \mathbb{R}_+$. In addition, a modular scoring function $R(S) = \sum_{a \in S} R(a)$ specifies the quality of a set of configurations by summing the qualities of single configurations. We aim to find a set of

maximizing configurations for the combined score (using a tradeoff parameter $\lambda \geq 0$)

$$F(S) = R(S) + \lambda D(S). \tag{1}$$

Submodular maximization. Our goal is to maximize $F(S)$ subject to a cardinality constraint $|S| \leq M$. If F is monotone, then the optimal algorithm for doing so is the greedy algorithm that starts out with $S^0 = \emptyset$, and iteratively adds the best item:

$$S^{i+1} = S^i \cup a^{i+1}, \quad a^{i+1} \in \arg\max_{a \in V \setminus S^i} F(a \mid S^i). \tag{2}$$

The final solution S^k is within a factor of $(1 - \frac{1}{e})$ of the optimal solution S^*: $F(S^k) \geq (1 - \frac{1}{e})F(S^*)$ [4]. This means in each iteration we must maximize the marginal gain. Clearly, if $|V|$ has exponential size, we can not touch each element even once. Hence, the lazy greedy algorithm [5] does not help here. Another immediate idea might be to take a random sample. Unfortunately, the expected value of a random sample of M elements can be much smaller than the optimal value $F(S^*)$, especially if N is large:

Lemma 1. *Let $S \subseteq V$ be a sample of size M taken uniformly at random. There exist monotone submodular functions where $\mathbb{E}[F(S)] \leq \frac{M}{N} \max_{|S|=M} F(S)$.*

We prove the lemma in the supplement. An alternative option is to use the structure of V, and obtain the element with maximum gain via optimization over the base variables. This is the path we pursue.

Marginal gains in configuration space. To solve the greedy step by optimization, we must transfer the marginal gain from the world of items to the world of base variables, and connect the submodular function on V to the variables $\mathbf{z}$. The configuration with the best marginal gain satisfies $\mathbf{z}^* = \phi(a^*)$ for $a^* \in \arg\max F(a \mid S)$, i.e., it maximizes

$$f(\mathbf{z}; S) \triangleq F(\phi^{-1}(\mathbf{z})) = R(\phi^{-1}(\mathbf{z})) + \lambda D(\phi^{-1}(\mathbf{z}) \mid S). \tag{3}$$

In general, finding the maximizing $\mathbf{z}$ can be a hard combinatorial optimization problem. However, if f has nice structure, then this problem may be efficiently solvable exactly, or at least approximately.

Effect of approximations. Approximate gain maximizers affect the overall solution quality slightly: Lemma 2 gives a generic bound for approximate greedy steps and applies to most examples in Section 3. Parts of Lemma 2 have been observed in previous work [6, 7]; we formally prove the combination in the supplement.

Lemma 2. *If each step of the greedy algorithm uses an approximate gain maximizer b^{i+1} with $F(b^{i+1} \mid S^i) \geq \alpha \max_{a \in V} F(a \mid S^i) - \epsilon^{i+1}$, then $F(S^M) \geq (1 - \frac{1}{e^\alpha}) \max_{|S|=M} F(S) - \sum_{i=1}^{M} \epsilon^i$.*

The properties of f obviously depend on R and the marginal gain of D. In good cases, R and D are "compatible" and have the same type of structure. One example of nice structure is that f is supermodular in $\mathbf{z}$, $\forall S \subseteq V$, thus allowing efficient exact maximization. In our examples, we will study such structured representations of f.

Structured Relevance Function. The relevance score $r(\mathbf{z}) = R(\phi^{-1}(\mathbf{z}))$ can be of any form that allows efficient exact (or provably approximate) search for the highest-scoring configuration, and is compatible with the diversity function $d(\mathbf{z}) = D(\phi^{-1}(\mathbf{z}))$. In this work, we define $R(a)$ to be the log-probability (or negative Gibbs energy) of a Markov Random Field (MRF) over the base variables $\mathbf{z} = \phi(a)$. For a general MRF, the quality of an item $a = \phi^{-1}(\mathbf{z})$ is given by $R(\phi^{-1}(\mathbf{z})) = r(\mathbf{z}) = \sum_{c \in \mathcal{C}} \theta_c(z_c)$, where $\mathcal{C} = \{c \mid c \subseteq \mathcal{V}\}$ be the set of cliques in the graph $G = (\mathcal{V}, \mathcal{E})$ defined over $\{z_1, z_2, \ldots, z_m\}$, and $\theta_c : [L]^{|c|} \mapsto \mathbb{R}$ are the log-potential functions (or factors) for these cliques. Our experiments will utilize both pairwise and high-order MRFs to represent the relevance functions. We can see that finding the single highest quality item (ignoring diversity) involves maximum a posteriori (MAP) inference in the MRF over the base variables.

3 Structured Diversity Functions

One general scheme to achieve diversity is by partitioning the ground set V into *groups* $V = \bigcup_i G_i$. These groups could be defined by task-dependent characteristics – for instance, in image segmentation, G_i can be the set of all segmentations that contain label i. The groups can be overlapping, so

that an item $a \in V$ is a member of more than one group. For instance, if a segmentation y contains pixels labeled "grass" and "cow", then $y \in G_{\text{grass}}$ and $y \in G_{\text{cow}}$.

Group Coverage. Given a partition of V into groups, we measure the diversity of a set S in terms of its *group coverage*:

$$D(S) = \sum_i h\big(|G_i \cap S|\big), \tag{4}$$

where h is any nonnegative concave scalar function. It is easy to show that the above defined function is monotone submodular. One special case of this general definition is *group coverage count* with $h(y) = \min\{1, y\}$, which simply counts the number of groups used ("covered") by items in S. Other natural choices of h are $\sqrt{\cdot}$, or $\log(1 + \cdot)$.

For this general definition of diversity, the marginal gain is given by

$$D(a \mid S) = \sum_{i: a \in G_i} \Big[h\big(1 + |G_i \cap S|\big) - h\big(|G_i \cap S|\big) \Big]. \tag{5}$$

With group coverage count, $D(a \mid S)$ is simply the number of *new* group introduced by a, *i.e.* not present in S.

In each step of the greedy algorithm, we need to maximize $R(\phi^{-1}(\mathbf{z})) + \lambda D(\phi^{-1}(\mathbf{z}) \mid S)$. We have already established a structured representation of $R(a)$ in terms of an MRF on $\mathbf{z}$, resulting in $r(\mathbf{z})$. In the following subsections, we will use three different definitions of groups G_i, which will instantiate three different diversity functions $D(S)$ for which the marginals gains $D(a \mid S)$ can be solved efficiently, i.e., in sub-linear time.

Diversity Function: Label Cost As a first example, we define groups via labels: let group G_ℓ be the set of all items $a = \phi^{-1}(\mathbf{z})$ whose base variables $\mathbf{z}$ contain the label ℓ, *i.e.* $a \in G_\ell$ if and only if $z_j = \ell$ for some $j \in [m]$. Such a diversity function naturally arises in multi-class image segmentation – if the highest scoring segmentation contains "sky" and "grass", we would like to reward other segmentations that contain an unused class label, say "sheep" or "cow".

Let $\text{labelcount}_S(\ell)$ be the number of segmentations in S that contain label ℓ. In the simplest case of count group coverage, the marginal gain provides a constant reward for every as yet unseen label ℓ: For general group coverage, the gain becomes

$$d(\mathbf{z}; S) = D(\phi^{-1}(\mathbf{z}) \mid S) = \sum_{\ell \in \mathbf{z}} \Big[h\big(1 + \text{labelcount}_S(\ell)\big) - h\big(\text{labelcount}_S(\ell)\big) \Big]. \tag{6}$$

Note that h is a concave function and hence $h\big(1 + \text{labelcount}_S(\ell)\big) - h\big(\text{labelcount}_S(\ell)\big)$ decreases as $\text{labelcount}_S(\ell)$ becomes larger. Thus, this diversity function $d(\mathbf{z})$ rewards the presence of a label ℓ by an amount proportional to how rare it is in the segmentations that are part of S. The gain $d(\mathbf{z})$ biases towards using as many labels as possible in a single segmentation $\mathbf{z}$. More realistic on real images is to use few labels in a single segmentation, but different labels across different segmentations. To achieve this, we add a penalty $c(\mathbf{z})$ to $r(\mathbf{z})$ that counts the number of labels used in a single segmentation.

For Equation (6), the greedy step means performing MAP inference in a standard MRF augmented with label reward/cost terms: $\text{argmax}_{\mathbf{z}}\, r(\mathbf{z}) + \lambda d(\mathbf{z}) - c(\mathbf{z})$. If $d(\mathbf{z}) - c(\mathbf{z}) \leq 0$, then an expansion algorithm by Delong *et al.* [8] applies for performing provably approximate MAP inference. Together with Lemma 2, their approximation bounds yield an overall approximation factor.

Diversity Function: Label Transitions. The label cost diversity function can be extended to penalise not just the presence of certain labels, but the presence of certain *label transitions*. Formally, we define one group $G_{\ell,\ell'}$ per label pair ℓ, ℓ', and an item a belongs to $G_{\ell,\ell'}$ if $\mathbf{z} = \phi(a)$ contains two adjacent variables z_i, z_j with labels $z_i = \ell, z_j = \ell'$. This diversity function rewards the presence of a label pair (ℓ, ℓ') by an amount proportional to how rare it is in the segmentations that are part of S. For such functions, the marginal gain $D(a \mid S)$ becomes a HOP called *cooperative cuts* [9].

The inference algorithm in Kohli *et al.* [10] gives a fully polynomial-time approximation scheme (FPTAS) (a multiplicative $\alpha = (1-\epsilon)$-approximation to be used in Lemma 2) for any nondecreasing, nonnegative h, and the exact gain maximizer for the count function $h(y) = \min\{1, y\}$.

Diversity Function: Hamming Ball The previous diversity function simply rewarded the presence of a label l, irrespective of which or how many variables z_i were assigned that label. We will now develop a diversity function that rewards a large *hamming distance* between the base variables of items in a set. Specifically, let $d(\mathbf{z}^1, \mathbf{z}^2) = \sum_{i=1}^{m} [\![z_i^1 \neq z_i^2]\!]$ be the hamming distance between $\mathbf{z}^1$ and $\mathbf{z}^2$, where $[\![\cdot]\!]$ is the Iverson bracket, which is 1 when the input argument is true. Let $\mathcal{B}_k(\mathbf{z}^1)$ denote the k-radius hamming ball centered at $\mathbf{z}^1$, *i.e.* $\mathcal{B}(\mathbf{z}^1) = \{\mathbf{z} \mid d(\mathbf{z}, \mathbf{z}^1) \leq k\}$. We construct one group G_a for each item $a \in V$, which is the set of items "close" to a. Specifically, $G_a = \{b \mid \phi(b) \in \mathcal{B}_k(\phi(a))\}$ is the set of items b whose base variables $\phi(b)$ are in the hamming ball centered at $\phi(a)$.

For hamming ball diversity, the marginal gain $D(a \mid S)$ becomes a HOP called *cardinality potential* [11]. In the simplest case of count group coverage, the marginal gain is

$$d(\mathbf{z}; S) = \Big|\mathcal{B}_k(\mathbf{z})\Big| - \Big|\mathcal{B}_k(\mathbf{z}) \cap \Big[\bigcup_{\mathbf{z}' \in S} \mathcal{B}_k(\mathbf{z}')\Big]\Big|, \tag{7}$$

i.e., the marginal gain of adding $\mathbf{z}$ is the number of *new* configurations $\mathbf{z}'$ covered by the hamming ball centered at $\mathbf{z}$. Unfortunately, estimating the size of intersection of $\mathcal{B}_k(\mathbf{z})$ with a union of other hamming balls does not lend itself to an efficient structured-representation. Thus, we work with a union-bound-like lower bound on $d(\mathbf{z})$:

$$d(\mathbf{z}) \geq \Big|\mathcal{B}_k(\mathbf{z})\Big| - \sum_{\mathbf{z}' \in S} \Big|\mathcal{B}_k(\mathbf{z}) \cap \mathcal{B}_k(\mathbf{z}')\Big|. \tag{8}$$

This lower-bound overcounts the intersection between $\mathcal{B}_k(\mathbf{z})$ and a union of hamming balls, by summing the intersections to each hamming ball in the union. Importantly, (8) depends on $\mathbf{z}$ only via its hamming distance to $\mathbf{z}'$. Such factors can be expressed as *cardinality potentials*, *i.e.* a potential that only depends on the number of variables assigned to a particular state, not which ones.

The greedy step involves performing MAP inference in an MRF augmented with cardinality potentials: $\operatorname{argmax}_{\mathbf{z}} r(\mathbf{z}) + \lambda d(\mathbf{z}; S)$. Tarlow *et al.* [11] showed how message-passing-based MAP inference may be performed in such graphical models, and provided a procedure for exactly computing all outgoing messages from cardinality factors in $O(m \log m)$ time.

A general observation. We next make one general observation that ties properties of D and d with group diversities.

Lemma 3. *Let all G_i be disjoint and $\mathbf{z}$ binary. The gain $d(\mathbf{z})$ of the count group coverage is submodular for all $A \subseteq V$, if and only if each group G_i contains a sublattice of V.*

More complex conditions can be derived for general coverage. This gives us a simple criterion at hand but may be restrictive. We note, however, that this lemma makes strict assumptions on the form of D, and that submodularity is not the only condition for d to be "simple". The label transitions, for example, resulted in a gain that is neither submodular nor supermodular.

4 Experiments

We apply our submodular maximization algorithms to the structured-output space of image segmentations. We performed experiments on category-level object segmentation on the PASCAL VOC 2012 dataset [12], where the goal is to label each pixel with one of 20 object categories or background. We compare all methods on their respective `oracle` accuracies, *i.e.* the accuracy of the *most accurate* segmentation in the set of M diverse segmentations returned by that method. For a small value of $M \sim 5 - 10$, a high `oracle` accuracy indicates that the algorithm has achieved high *recall* and has identified a pool of candidate solutions for further processing in a cascaded pipeline.

Baselines. We compare our proposed methods against DivMBest [13], which greedily produces diverse segmentation by explicitly adding a linear hamming-distance term to the MRF log-potential. Each hamming term is decomposable along the variables z_i and simply modifies the node potentials $\tilde{\theta}(z_i) = \theta(z_i) + \lambda \sum_{\mathbf{z}' \in A} [\![z_i \neq z_i']\!]$. Batra *et al.* [13] extensively tested DivMBest against other techniques like M-Best-MAP [14], which produce high scoring solutions without a focus on diversity, and sampling-based techniques that simply produce diverse solutions without a focus on the relevance term. DivMbest significantly outperformed these methods and thus we do not include them in our comparison. We report experiments with all three diversity functions.

Label Cost				Hamming-Based Diversity				Label Transition			
	MAP	M=5	M=15		MAP	M=5	M=15		MAP	M=5	M=15
$\min\{1, \cdot\}$	42.35	45.43	45.58	DivMBest	43.43	51.21	**52.90**	$\min\{1, \cdot\}$	42.35	44.26	44.78
$\sqrt{(\cdot)}$	42.35	45.72	50.01	Hamming Ball	43.43	51.71	**55.32**	$\sqrt{(\cdot)}$	42.35	45.43	46.21
$\log(1 + \cdot)$	42.35	46.28	**50.39**					$\log(1 + \cdot)$	42.35	45.92	**46.89**

Table 1: VOC 2012 Validation set accuracies for different diversity functions

Model. We construct a multi-label pairwise CRF on superpixels. Our node potentials come from the outputs of category-specific regressors trained by [15], and our edge potentials are multi-label Potts. Inference in the presence of diversity terms is performed with the implementations of Delong *et al*. [8] (for label costs), Tarlow *et al*. [11] (for hamming ball diversity), and Boykov *et al*. [16] (for label transitions).

Results. We evaluated all methods on the VOC 2012 dataset, consisting of `train`, `val` and `test` partitions with $\sim$1450 images each. We train the regressors on `train` and report `oracle` accuracies of different methods on `val`. The accuracy is the standard PASCAL "intersection / union" performance measure, averaged over all categories. For both label cost and transition, we try 3 different concave functions $h(\cdot)$ — $\min\{1, \cdot\}$,$\sqrt{(\cdot)}$ and $\log(1 + \cdot)$. Table 1 shows the `oracle` accuracies for label cost, label transition, hamming ball and DivMbest. We can see that hamming ball diversity performs the best, followed by DivMBest, and label cost/transition perform the worst. Both label costs and transition work well for certain specific scene configurations (we show qualitative examples of these situations in supplementary material), but such scenes are rare, and as a result these two diversity functions perform poorly on average. DivMBest does not have approximation guarantees, but still performs well in practice. It is nice to see that the method with theoretical guarantees (hamming ball) has the best empirical performance.

Acknowledgements. This work was done while AP was an intern at Virginia Tech. This material is based upon work partially supported by the National Science Foundation under Grant No. IIS-1353694.

References

[1] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001. 1

[2] B. Taskar, C. Guestrin, and D. Koller. Max-Margin Markov networks. In *NIPS*, 2003. 1

[3] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 6:1453–1484, 2005. 1

[4] G.L. Nemhauser, L.A. Wolsey, and M.L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294, 1978. 2, 3

[5] M. Minoux. Accelerated greedy algorithms for maximizing submodular set functions. *Optimization Techniques*, pages 234–243., 1978. 3

[6] P.R. Goundan and A.S. Schulz. Revisiting the greedy approach to submodular set function maximization. Manuscript, 2009. 3

[7] M. Streeter and D. Golovin. An online algorithm for maximizing submodular functions. Technical Report CMU-CS-07-171, Carnegie Mellon University, 2007. 3

[8] Andrew Delong, Anton Osokin, Hossam N. Isack, and Yuri Boykov. Fast approximate energy minimization with label costs. In *CVPR*, pages 2173–2180, 2010. 4, 6

[9] S. Jegelka and J. Bilmes. Submodularity beyond submodular energies: Coupling edges in graph cuts. In *CVPR*, 2011. 4

[10] P. Kohli, A. Osokin, and S. Jegelka. A principled deep random field model for image segmentation. In *CVPR*, 2013. 4

[11] D. Tarlow, I. E. Givoni, and R. S. Zemel. HOP-MAP: Efficient message passing with high order potentials. In *AISTATS*, pages 812–819, 2010. 5, 6

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012). 5

[13] D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich. Diverse M-Best Solutions in Markov Random Fields. In *ECCV*, 2012. 5

[14] M. Fromer and A. Globerson. An LP view of the m-best MAP problem. In *NIPS*, 2009. 5

[15] João Carreira, Rui Caseiro, Jorge Batista, and Cristian Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, pages 430–443, 2012. 6

[16] Y. Boykov, O. Veksler, and R. Zabih. Efficient approximate energy minimization via graph cuts. *PAMI*, 20(12):1222–1239, 2001. 6