An Empirical Study of Optimization for Maximizing Diffusion in Networks*

Kiyan Ahmadizadeh, Bistra Dilkina, Carla P. Gomes, and Ashish Sabharwal

Department of Computer Science, Cornell University, Ithaca NY 14853, USA
{kiyan,bistra,gomes,sabhar}@cs.cornell.edu

Abstract. We study the problem of maximizing the amount of stochastic diffusion in a network by acquiring nodes within a certain limited budget. We use a Sample Average Approximation (SAA) scheme to translate this stochastic problem into a simulation-based deterministic optimization problem, and present a detailed empirical study of three variants of the problem: where all purchases are made upfront, where the budget is split but one still commits to purchases from the outset, and where one has the ability to observe the stochastic outcome of the first stage in order to "re-plan" for the second stage. We apply this to a Red Cockaded Woodpecker conservation problem. Our results show interesting runtime distributions and objective value patterns, as well as a delicate trade-off between spending all budget upfront vs. saving part of it for later.

1 Introduction

Many real-world processes are diffusive in nature, giving rise to optimization problems where the goal is to maximize or minimize the spread of some entity through a network. For example, in epidemiology, the spread of infectious diseases in a human or animal network is a diffusion-based process. In ecology, so-called metapopulation models capture the diffusion of species in a fragmented landscape of habitat patches. Similarly, the adoption of a certain marketed product by an individual may trigger his or her friends or fans to adopt that product as well, suggesting viral marketing strategies in human networks. In the social network setting, particularly in Internet-based networks such as Facebook and Twitter, the spread of information between individuals is yet another diffusion process. The stochastic nature of such diffusion processes, or *cascades*, and how best to intervene in order to influence their outcomes, has been the study of several recent papers in these areas [e.g. 2, 4, 6, 8, 10, 11, 13]. A key question in this context is, if one had limited resources to purchase part of the network to use either as the initially "active" nodes or as nodes that may participate in the diffusion process, which nodes should one purchase?

We study this question with a focus on the case where the intervention budget, instead of all being available upfront, is split into two or more time steps.

* Supported by NSF (Expeditions in Computing award for Computational Sustainability, 0832782; IIS grant 0514429) & AFOSR (IISI, grant FA9550-04-1-0151). The authors thank Yahoo! for generously providing access to their M45 compute cloud.

D. Cohen (Ed.): CP 2010, LNCS 6308, pp. 514–521, 2010.

We evaluate our techniques on a specific conservation problem in which the diffusion process of interest is the dispersal and territory establishment of an endangered bird species, the Red-Cockaded Woodpecker (RCW) [3, 5, 15]. Using a stochastic model of the diffusion of RCW through geographic maps, we study the solutions of three optimization variants that differ in when the conservationist is allowed to make purchase decisions. The methodology and results apply not only to RCW conservation but also to many other problems in the field of *computational sustainability* [7], where maximizing or minimizing diffusion is a commonly encountered problem in areas ranging from ecology to poverty mitigation through food networks.

Formally, we pose this optimization task as a stochastic Mixed Integer Programming (MIP) problem. Even extremely simple classes of stochastic linear programs are #P-hard [cf. 15]. However, an effective solution method with stochastic optimality gap guarantees is the Sample Average Approximation (SAA) [14]. The application of SAA to a formulation of the RCW conservation problem has been previously evaluated on a real-life size instance, motivated by the actual data for the RCW population in North Carolina [15]. In order to better understand the effectiveness and scalability of this solution methodology from a computational perspective, we introduce a synthetic problem generator[1] that uses real map data as a basis for generating random instances. This analysis reveals several interesting computational trends: an easy-hard-easy runtime pattern as the budget fraction is varied, an increase in hardness as the "self-colonization" probability is increased, more runtime variation for instances that are harder to solve (for a fixed budget), and a roughly inverse relation between the computational hardness and the solution objective value (again for a fixed budget).

We also study a natural and realistic generalization of the problem where the total budget is actually not available to be spent at the beginning but is rather *split* into two stages. This modification significantly increases the computational difficulty of the problem and makes it non-trivial to apply the SAA methodology which was originally designed for the case where stochastic samples can be drawn right upfront, rather than adaptively. Indeed, the most general approach, a truly *multi-stage stochastic version* of the problem [1], has very limited scalability. Instead, we consider two simpler problem variants that are more scalable: committing to both first and second time step purchase decisions upfront (the *single-stage method for the split-budget problem*) or committing to purchase decisions for the first time step but re-evaluating what is best to purchase in the second time step after observing the stochastic outcome of the first stage (the *two-stage re-planning method*). Our experiments show that the re-planning approach, although computationally more expensive, does pay off—the solution objective obtained through an SAA-style implementation of re-planning is often significantly better than that obtained by the single-stage split-budget with all purchase decisions made upfront; more interestingly, the value of information gained from first stage observations can result in a re-planning objective that is better than spending all budget upfront.

[1] http://www.cs.cornell.edu/∼kiyan/rcw/generator.htm

2 Problem Description, Model, and Solution Methods

A diffusion process can be quite complex; *patch-based models* [8] represent diffusion as occurring on a network of *nodes* which become *active* upon the arrival of the dispersing entity. Given a network of nodes (i.e., an undirected graph), a *network dispersion model* specifies for each pair of nodes (r_1, r_2) the probability that node r_2 will become active at the next time step, given that r_1 is currently active. Similarly, if node r is currently active, it remains so after one time step with a specified *survival probability*. In our optimization setting, we assume that the process can only spread to nodes that have been purchased. We divide the nodes into disjoint *node sets* with an associated cost and assume that a manager making purchase decisions is limited by a budget in each time step. Given the stochastic nature of the dispersion process, the overall goal is to maximize the *expected* number of active nodes at a specified *time horizon*.

For our experiments, we consider a specific instance of this problem in which the diffusing entity is the Red-Cockaded Woodpecker (RCW). Geographic territories suitable for RCW inhabitance represent graph nodes, with territories grouped into real estate parcels available for purchase. Node activity here represents the settlement of a territory by RCW members dispersing from other territories. As in most conservation settings, the geographic territories must be owned and maintained by conservationists for species members to survive there.

We formulate this problem as a stochastic Mixed Integer Program (MIP), shown below. (One can create alternative MIP formulations of this problem as well, using, e.g., network flow.) Let R be the number of nodes in the network, P the number of node sets, H the planning horizon, $C(p)$ the cost of node set $p \in \{1..P\}$, $B(t)$ the budget available at time step $t \in \{0..\mathrm{H}-1\}$, $P(r)$ the node set that node $r \in \{1..R\}$ belongs to, and $I(r)$ the 0-1 indicator of whether node r is initially active (node sets containing an initially active node are assumed to be already owned). Binary variables $\{\mathrm{y}(p,t) \mid p \in \{1..P\}, t \in \{0..\mathrm{H}-1\}\}$ correspond to the action of buying node set p at time step t. Binary variables $\{\mathrm{x}(r,t) \mid r \in \{1..R\}, t \in \{0..\mathrm{H}-1\}\}$ correspond to r being active at time t. The constraints of the MIP encode the basic requirements discussed earlier in the problem description. For lack of space, we refer the reader to [15] for details and mention here only that the budget constraint (2) has been generalized to include a time-step-specific budget and that $\xi_{r',r}^{t-1}$ are the stochastic coefficients that follow the dispersion model probability for r' and r.

In reality, we cannot directly optimize this stochastic MIP. Instead, we use the Sample Average Approximation (SAA) method [14, 18], which uses random samples from the underlying probability distribution of the stochastic parameters to generate a finite number of scenarios and creates a deterministic MIP to optimize the *empirical average* (rather than the true expectation) of the number of active territories over this finite set of sampled scenarios. We will describe this shortly. In this deterministic version of our stochastic MIP defined over a set of k scenarios $S_1, S_2, \ldots, S_k$, we still have one purchase variable for each node set at each time step but k different activity variables for each node at each time step, capturing the different diffusion activity in the k different scenarios. In other

$$\text{maximize} \quad \sum_{\mathrm{r}=0}^{\mathrm{R}} \mathrm{x}(\mathrm{r},\mathrm{H}) \quad \text{such that}$$

$$\mathrm{y}(p,0) = 1 \qquad \forall p \in \text{initial (free) parcels} \qquad (1)$$

$$\sum_{p=1}^{P} \mathrm{C}(p) \times \mathrm{y}(p,t) \leq \mathrm{B}(t) \qquad \forall t \in \{0..\mathrm{H}-1\} \qquad (2)$$

$$\sum_{t=0}^{\mathrm{H}-1} \mathrm{y}(p,t) \leq 1 \qquad \forall p \in \{1..P\} \qquad (3)$$

$$\mathrm{x}(r,t) \leq \sum_{t'=0}^{t} \mathrm{y}(P(r),t') \qquad \forall r \in \{1..R\}, \forall t \in \{1..\mathrm{H}\} \qquad (4)$$

$$\mathrm{x}(r,t) \leq \sum_{r'=1}^{R} \xi_{r',r}^{t-1} \mathrm{x}(r',t-1) \qquad \forall r \in \{1..R\}, \forall t \in \{1..\mathrm{H}\} \qquad (5)$$

$$\mathrm{x}(r,0) = I(r) \qquad \forall r \in \{1..R\} \qquad (6)$$

words, the purchase decisions are synchronized amongst the different scenarios but activity variables differ depending on which nodes were occupied in which scenario. For the objective function, we simply sum up the activity variables at the horizon for all k scenarios, thus optimizing the sum of active territories (or, equivalently, the average activity) over the k scenarios at the horizon. This results in an expanded deterministic formulation very similar to the one above, and we denote it by MIP$(S_1, S_2, \ldots, S_k)$.[2]

While our MIP allows a budget constraint for each time step, in our experiments we consider two variants. In the *upfront* variant, all budget is spent at time step $T = 0$. In the *split* variant, the budget is split into (b_1, b_2) in a given proportion between $T_1 = 0$ and $T_2 < H$, and is set to 0 for other time steps.

2.1 Sample Average Approximation and Re-planning

The SAA approach has been instrumental in addressing large-scale stochastic optimization problems [14, 18]. It provides provable convergence guarantees—it may over-fit for a small number of scenarios, but converges to the true optimum with increasing training samples and provides a statistical bound on the optimality gap. The SAA procedure works in three phases. In the TRAINING phase, we generate N candidate solutions by creating N SAA MIPs with k training scenarios each and solving them to optimality. In the VALIDATION phase, we generate M_1 new validation scenarios, evaluate each of the N candidate solutions on these scenarios, and choose the solution $s*$ with the best validation objective. In the TEST phase, we generate M_2 fresh scenarios to re-evaluate $s*$, thus obtaining and reporting an estimate of the true objective value of $s*$.

[2] In the deterministic MIP, we add redundant constraints that force any variable $\mathrm{x}(r,t)$ of a scenario to be set to 1 whenever the corresponding node set has been bought and there was dispersal from nodes active at the previous time step.

The test objective of $s*$ is a lower bound on the true optimum while the average of the MIP objective of all N candidate solutions is a (stochastic) upper bound on the optimum (see [15] for more details). The above procedure is applied to both the upfront and split budget models. For our experiments we set $N = 100, k = 10, M_1 = 1000, M_2 = 5000$.

In the split budget setting, it is quite pessimistic to assume that decision makers cannot adjust purchase decisions based on first stage observations. The true objective evaluation of a split budget purchase plan needs to be more "dynamic" in nature—at the second decision time step T_2 one can observe the events of the past time steps and accordingly re-plan how to spend b_2. Given a set of purchase decisions for T_1, we describe *how to evaluate the expected objective value under re-planning*, assuming that at the second decision point T_2, one would again apply the SAA solution method to select purchase decisions. The re-planning evaluation of a candidate solution s, representing purchase decisions made at T_1 in the split budget model, is done as follows. We generate a sample set of F "prefix scenarios" over the years $0..T_2 - 1$. For each prefix scenario and considering all nodes sets purchased in s as being available for free at T_2, we perform an SAA evaluation as if we are at time step T_2 and are solving the *upfront* model for the remaining years and with budget b_2. The SAA here is performed for $N = 20$, $k = 10$, $M_1 = 100$ and $M_2 = 500$, for each of $F = 100$ prefix scenarios. Finally, the re-planning objective of s is reported as the average SAA objective over the $F = 100$ prefix scenarios.

3 Experimental Results

We use a graph of nodes derived from a topology of 411 territories grouped into 146 parcels, representative of a region on the coast of North Carolina of interest to The Conservation Fund for RCW preservation. The dispersion model used for this study is based on a habitat *suitability score* (an integer in $[0, 9]$) for each territory as well as known parameters about the ability of birds to disperse between territories at various distances [12]. Suitability scores were estimated using GIS data from the 2001 USGS National Land Cover Dataset (NLCD) [16]. Parcels (corresponding to node sets) were constructed using the US Census Bureau's "census block" divisions [17]. Using the base topology, we created several randomized instances of the problem by (a) perturbing the suitability value by ± 1 and (b) selecting different sets of initially active territories by randomly choosing clusters of territories with high suitability.

We used Yahoo!'s M45 cloud computing platform running Apache Hadoop version 0.20.1 to perform independent parts of our solution methods massively in parallel. IBM ILOG CPLEX v12.1 [9] was used to solve all MIPs involved.

Runtime Distributions and Objective Value of MIP. We study the runtime to solve the SAA MIPs (with $k = 10$ scenarios over $H = 20$ years) under different budgets expressed as a fraction of the total cost of all parcels in the instance. Results are presented in Fig. 1. Each point in these plots corresponds to the average runtime

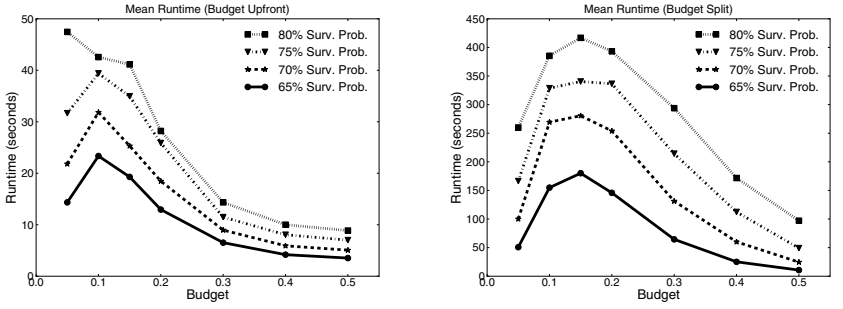

Fig. 1. Runtime (y-axis) as a function of budget (x-axis) for various extinction rates. Left: all budget available upfront. Right: budget split into two time steps.

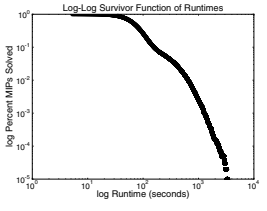

Fig. 2. Runtime distribution for instance map4-30714 exhibits power-law decay (log-log scale)

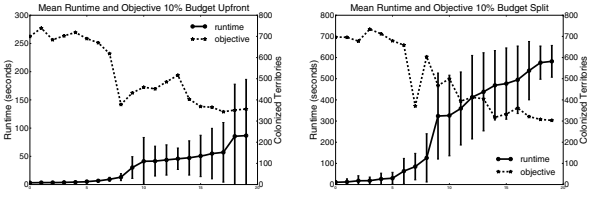

Fig. 3. The high variation in runtime on some instances (lower curve) and the corresponding average MIP objective values (higher curve)

over 100 different samples of $k = 10$ scenarios of each of 20 different variations of the basic map described earlier. The left pane shows the results when all budget is available upfront, while the right pane considers the split-budget case. These curves demonstrate an easy-hard-easy pattern as the budget parameter is varied, and also indicate that the problem becomes harder to solve for higher survival rates. Comparing the left and right plots, we see that the split-budget variant of the problem is roughly 10x harder to solve than when all budget is available upfront (notice the scales on the y-axis).

We evaluate in more detail the performance with 70% survival rate. Fig. 2 shows the distribution of the runtime for one particular variation of the base map, called map4-30714, for 10% budget. All budget is available upfront and the plot is derived from 100,000 runs. This figure demonstrates the typical runtime distribution seen on this suite of instances: a *power-law decay*, indicated by the near-linear (or super-linear) drop in the probability of "failure" or timeout (y-axis) as a function of the runtime (x-axis) when plotted in log-log scale.

We next consider the relation between the running time and the objective value of the SAA MIP, for both the upfront and split budget cases. The lower curves in the plots of Fig. 3 show the average runtime and standard deviation over 100 runs of each of 20 variations of the base map, where the 20 instances

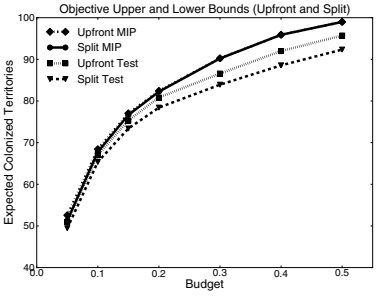

Fig. 4. SAA upper and lower bounds on obj. value for upfront and split budgets

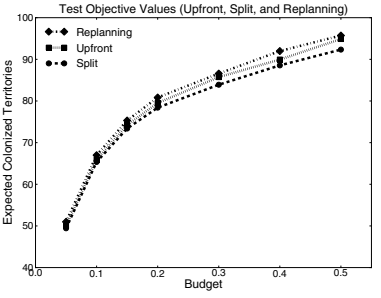

Fig. 5. Objective value of re-planning, compared to upfront and split budgets

are ordered from low to high runtime. The upper curves show the corresponding average objective value achieved for each instance (the variation in the objective value was small). These plots indicate that for our test suite, instances that are hard to solve often have a significantly higher runtime variation than instances that are easy to solve. Moreover, the harder to solve instances typically result in a lower objective value.

Evaluation of Sample Average Approximation and Re-Planning. We evaluate the solution quality of the SAA approach as a function of the budget fraction. Fig. 4 presents results for both the upfront and split budget problems where the budget is divided evenly between $T_1 = 1$ and $T_2 = 10$. The curves marked Upfront MIP and Split MIP present the average MIP objective over the $N = 100$ candidate solutions and are hence a stochastic upper bound on the true optimum. The curves marked Upfront Test and Split Test are the estimated true quality of the solution chosen (and hence provide a lower bound on the quality of the true optimum). The difference between Upfront Test and Split Test measures the penalty of not having all funds available in the first stage. The relatively small gap between the upper and lower bounds confirms that our choice of SAA parameters is good and that the solutions provided are very close to optimal.

Finally, we evaluate the advantage of re-planning in the stochastic setting of our problem. Recall that we would like to understand the tradeoff between spending all available budget upfront vs. re-planning with a portion of investments at a later stage after making stochastic observations. The balance is, in fact, quite delicate. By spending too much money upfront, we leave little room for "adjusting" to the stochastic outcome of the first stage. On the other hand investing too little upfront limits the amount of possible variation in dispersion, thus limiting the worth of stochastic observations. When splitting the budget evenly, making second stage decisions at $T_2 = 10$, re-planning often did not yield as good a result as investing all money upfront. Nonetheless, for other parameters such as a 30-70 split with $T_2 = 5$, we found that re-planning begins to pay off, as is shown in Fig. 5. The top curve in the plot corresponds to re-planning and shows that it

can result in the occupation of more territories in our bird conservation example than spending all budget upfront (the middle curve) or splitting the budget but providing a single-stage style solution that commits to a certain set of purchase decisions at the outset (the lowest curve).

In summary, our experiments have examined the complexity of optimizing stochastic diffusion processes and the value of different planning methodologies. Our results show the considerable benefits of making decisions upfront (e.g. in a single-stage), and the benefits that re-planning based on stochastic observations can have when decisions must be made in multiple stages.

References

[1] S. Ahmed. Introduction to stochastic integer programming. *COSP Stochastic Programming Introduction – http://stoprog.org*, 2004.

[2] Anderson, R., May, R.: Infectious diseases of humans: dynamics and control. Oxford University Press, Oxford (1992)

[3] Conner, R., Rudolph, D., Walters, J., James, F.: The Red-cockaded Woodpecker: surviving in a fire-maintained ecosystem. Univ. of Texas Press (2001)

[4] Domingos, P., Richardson, M.: Mining the network value of customers. In: KDD, pp. 57–66 (2001), ISBN 1-58113-391-X

[5] US Fish and Wildlife Service. Red-cockaded woodpecker recovery plan (2003)

[6] Goldenberg, J., Libai, B., Muller, E.: Talk of the network: A complex systems look at word-of-mouth. Marketing Letters 12(3), 211–223 (2001)

[7] Gomes, C.P.: Computational Sustainability: Computational methods for a sustainable environment, economy, and society. The Bridge, NAE 39(4) (2009)

[8] Hanski, I.: Metapopulation ecology. Oxford University Press, USA (1999)

[9] IBM ILOG, SA. CPLEX 12.1 Reference Manual (2009)

[10] Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: KDD, pp. 137–146 (2003)

[11] Leskovec, J., Adamic, L., Huberman, B.: The dynamics of viral marketing. ACM Transactions on the Web (TWEB) 1(1), 5 (2007)

[12] Letcher, B.H., Priddy, J.A., Walters, J.R., Crowder, L.B.: An individual-based, spatially-explicit simulation model of the population dynamics of the endangered red-cockaded woodpecker, picoides borealis. Biol. Conserv. 86(1), 1–14 (1998)

[13] McDonald-Madden, E., Baxter, P.W., Possingham, H.P.: Making robust decisions for conservation with restricted money and knowledge. Appl. Ecol. 45(9), 1630–1638 (2008)

[14] Shapiro, A.: Monte Carlo sampling methods. In: Ruszczynski, A., Shapiro, A. (eds.) Stochastic Programming. Handbooks in Operations Research and Management Science, vol. 10, pp. 353–426 (2003)

[15] Sheldon, D., Dilkina, B., Elmachtoub, A., Finseth, R., Sabharwal, A., Conrad, J., Gomes, C.P., Shmoys, D., Allen, W., Amundsen, O., Vaughan, B.: Optimal network design for the spread of cascades. Technical report, Cornell University (April 2010), http://hdl.handle.net/1813/14917

[16] The USGS Land Cover Institute (LCI). USGS land cover (2001)

[17] US Census Bereau. Census data: 2009 TIGER/Line shapefiles (2009)

[18] Verweij, B., Ahmed, S., Kleywegt, A., Nemhauser, G., Shapiro, A.: The sample average approximation method applied to stochastic routing problems: a computational study. Computational Optimiz. and Applications 24(2), 289–333 (2003)