The SelMax Planner: Online Learning for Speeding up Optimal Planning*

Carmel Domshlak Technion
Malte Helmert Albert-Ludwigs-Universität Freiburg
Erez Karpas Technion
Shaul Markovitch Technion

Abstract

The SelMax planner combines two state-of-the-art admissible heuristics using an online learning approach. In this paper we describe the online learning approach employed by SelMax, briefly review the Fast Downward framework, and describe the SelMax planner.

Introduction

One of the most prominent approaches to cost-optimal planning is using the A^* search algorithm with an admissible heuristic. Many admissible heuristics have been proposed, varying from cheap to compute yet typically not very informative to expensive to compute but often very informative. Since the accuracy of heuristic functions varies for different problems, and even for different states of the same problem, we can produce a more robust optimal planner by combining several admissible heuristics. The simplest way of doing this is by using their point-wise maximum at each state. Presumably, each heuristic is more accurate, that is, provides a higher estimate, in different regions of the search space, and thus their maximum is at least as accurate as each of the individual heuristics. In some cases it is also possible to use additive (Felner, Korf, and Hanan 2004; Haslum, Bonet, and Geffner 2005; Katz and Domshlak 2008) or mixed additive/maximizing (Coles et al. 2008; Haslum et al. 2007) combinations of admissible heuristics.

An important issue with both max-based and sum-based approaches is that the benefit of adopting them over sticking to just a single heuristic is assured only if the planner is not constrained by time. Otherwise, the time spent on computing numerous heuristic estimates at each state may outweigh the time saved by reducing the number of expanded states. Selective Max (SelMax) is a novel method for combining admissible heuristics that aims at providing the accuracy of their max-based combination while still computing just a single heuristic for each search state.

At a high level, selective max can be seen as a hyper-heuristic (Burke et al. 2003) — a heuristic for choosing between other heuristics. Specifically, selective max is based on a seemingly useless observation that, if we had an oracle indicating the most accurate heuristic for each state, then computing only the indicated heuristic would provide us with the heuristic estimate of the max-based combination. In practice, of course, such an oracle is not available. However, in the time-limited settings of our interest, this is not our only concern: It is possible that the extra time spent on computing the more accurate heuristic (indicated by the oracle) may not be worth the time saved by the reduction in expanded states.

Addressing the latter concern, we first analyze an idealized model of a search space and deduce a decision rule for choosing a heuristic to compute at each state when the objective is to minimize the overall search time. Taking that decision rule as our target concept, we then describe an online active learning procedure for that concept that constitutes the essence of selective max.

Notation

We consider planning in the SAS$^+$ formalism (Bäckström and Nebel 1995); a SAS$^+$ description of a planning task can be automatically generated from its PDDL description (Helmert 2009). A SAS$^+$ task is given by a 4-tuple $\Pi = \langle V, A, s_0, G\rangle$. $V = \{v_1, \ldots, v_n\}$ is a set of *state variables*, each associated with a finite domain $dom(v_i)$. Each complete assignment s to V is called a *state*; s_0 is an *initial state*, and the *goal* G is a partial assignment to V. A is a finite set of *actions*, where each action a is a pair $\langle \mathsf{pre}(a), \mathsf{eff}(a)\rangle$ of partial assignments to V called *preconditions* and *effects*, respectively.

An action a is applicable in a state s iff $\mathsf{pre}(a) \subseteq s$. Applying a changes the value of each state variable v to $\mathsf{eff}(a)[v]$ if $\mathsf{eff}(a)[v]$ is specified. The resulting state is denoted by $s[\![a]\!]$; by $s[\![\langle a_1, \ldots, a_k\rangle]\!]$ we denote the state obtained from sequential application of the (respectively applicable) actions $a_1, \ldots, a_k$ starting at state s. Such an action sequence is a plan if $G \subseteq s_0[\![\langle a_1, \ldots, a_k\rangle]\!]$.

A Model for Heuristic Selection

Given a set of admissible heuristics and the objective of minimizing the overall search time, we are interested in a decision rule for choosing the right heuristic to compute at each search state. In what follows, we derive such a decision rule for a pair of admissible heuristics with respect to an idealized search space model corresponding to a tree-structured

*This paper is strongly based upon Domshlak, Karpas, and Markovitch (2010)

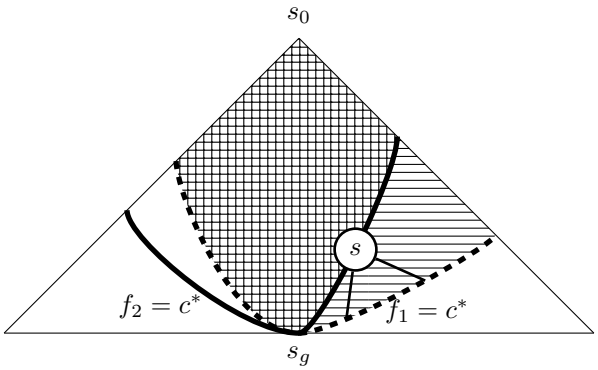

Figure 1: An illustration of the idealized search space model and the f-contours of two admissible heuristics.

search space with a single goal state, constant branching factor b, and uniform cost actions (Pearl 1984). Two additional assumptions we make are that the heuristics are consistent, and that the time t_i required for computing heuristic h_i is independent of the state being evaluated; w.l.o.g. we assume $t_2 \geq t_1$. Obviously, most of the above assumptions do not hold in typical search problems, and later we carefully examine their individual influences on our framework.

Adopting the standard notation, let $g(s)$ be the cost of the cheapest path from s_0 to s. Defining $\max_h(s) = \max(h_1(s), h_2(s))$, we then use the notation $f_1(s) = g(s) + h_1(s)$, $f_2(s) = g(s) + h_2(s)$, and $\max_f(s) = g(s) + \max_h(s)$. The A^* algorithm with a heuristic h expands states in increasing order of $f = g + h$. Assuming the goal state is at depth c^*, let us consider the states satisfying $f_1(s) = c^*$ (the dotted line in Fig. 1) and those satisfying $f_2(s) = c^*$ (the solid line in Fig. 1). The states above the $f_1 = c^*$ and $f_2 = c^*$ contours are those that are surely expanded by A^* with h_1 and h_2, respectively. The states above both these contours (the grid-marked region in Fig. 1), that is, the states $SE = \{s \mid \max_f(s) < c^*\}$, are those that are surely expanded by A^* using $\max_h$ (see Theorem 4, p. 79, Pearl 1984).

Under the objective of minimizing the search time, observe that the optimal decision for any state $s \in SE$ is not to compute any heuristic at all, since all these states are surely expanded anyway. The optimal decision for all other states is a bit more complicated. $f_2 = c^*$ contour that separates between the grid-marked and lines-marked areas. Since $f_1(s)$ and $f_2(s)$ account for the same $g(s)$, we have $h_2(s) > h_1(s)$, that is, h_2 is more accurate in state s than h_1. If we were interested solely in reducing state expansions, then h_2 would obviously be the right heuristic to compute at s. However, for our objective of reducing the actual search time, h_2 may actually be the wrong choice because it might be much more expensive to compute than h_1.

Let us consider the effects of each of our two alternatives. If we compute $h_2(s)$, then s is no longer surely expanded since $f_2(s) = c^*$, and thus whether A^* expands s or not depends on tie-breaking. In contrast, if we compute $h_1(s)$, then s is surely expanded because $f_1(s) < c^*$. Note that not computing h_2 for s and then computing h_2 for one of the descendants s' of s is surely a sub-optimal strategy as we do pay the cost of computing h_2, yet the pruning of A^* is limited only to the search sub-tree rooted in s'. Therefore, our choices are really either computing h_2 for s, or computing h_1 for all the states in the sub-tree rooted in s that lie on the $f_1 = c^*$ contour. Suppose we need to expand l complete levels of the state space from s to reach the $f_1 = c^*$ contour. This means we need to generate order of b^l states, and then invest $b^l t_1$ time in calculating h_1 for all these states that lie on the $f_1 = c^*$ contour. In contrast, suppose we choose to compute $h_2(s)$. Assuming favorable tie-breaking, the time required to "explore" the sub-tree rooted in s will be t_2.

Putting things together, the optimal decision in state s is thus to compute h_2 iff $t_2 < b^l t_1$, or if we rewrite this, if

$$l > \log_b(t_2/t_1).$$

As a special case, if both heuristics take the same time to compute, this decision rule boils down to $l > 0$, that is, the optimal choice is simply the more accurate (for state s) heuristic.

The next step is to somehow estimate the "depth to go" l. For that, we make another assumption about the rate at which f_1 grows in the sub-tree rooted at s. Although there are many possibilities here, we will look at two estimates that appear to be quite reasonable. The first estimate assumes that the h_1 value remains constant in the subtree rooted at s, that is, the additive error of h_1 increases by 1 for each level below s. In this case, f_1 increases by 1 for each expanded level of the sub-tree (because h_1 remains the same, and g increases by 1), and it will take expanding $\Delta_h(s) = h_2(s) - h_1(s)$ levels to reach the $f_1 = c^*$ contour. The second estimate we examine assumes that the absolute error of h_1 remains constant, that is, h_1 increases by 1 for each level expanded, and so f_1 increases by 2. In this case, we will need to expand $\Delta_h(s)/2$ levels. This can be generalized to the case where the estimate h_1 increases by any constant additive factor c, which results in $\Delta_h(s)/(c+1)$ levels being expanded. In either case, the dependence of l on $\Delta_h(s)$ is linear, and thus our decision rule can be reformulated to compute h_2 if

$$\Delta_h(s) > \alpha \log_b(t_2/t_1),$$

where α is a hyper-parameter for our algorithm. Note that, given b, t_1, and t_2, the quantity $\alpha \log_b(t_2/t_1)$ becomes fixed and in what follows we denote simply by *threshold* τ.

Dealing with Model Assumptions

The idealized model above makes several assumptions, some of which appear to be very problematic to meet in practice. Here we examine these assumptions more closely, and when needed, suggest pragmatic compromises.

First, the model assumes that the search space forms a tree with a single goal state and uniform cost actions, and that the heuristics in question are consistent. Although the first assumption does not hold in most planning problems, and the second assumption is not satisfied by some state-of-the-art heuristics, they do not prevent us from using the decision rule suggested by the model. Furthermore, there is some empirical evidence to support our conclusion about exponential growth of the search effort as a function of heuristic

error, even when the assumptions made by the model do not hold. In particular, the experiments of Helmert and Röger (2008) with heuristics with small constant additive errors clearly show that the number of expanded nodes typically grows exponentially as the (still very small and additive) error increases.

The model also assumes that both the branching factor and the heuristic computation times are constant across the search states. In our application of the decision rule to planning in practice, we deal with this assumption by adopting the average branching factor and heuristic computation times, estimated from a random sample of search states. Finally, the model assumes perfect knowledge about the surely expanded search states. In practice, this information is obviously not available. We approach this issue conservatively by treating all the examined search states as if they were on the decision border, and thus apply the decision rule at all the search states. Note that this does not hurt the correctness of our algorithm, but only costs us some heuristic computation time on the surely expanded states. Identifying the surely expanded region during search is the subject of ongoing work, and can hopefully be used to improve search efficiency even further.

Online Learning of the Selection Rule

Our decision rule for choosing a heuristic to compute at a given search state s suggests to compute the more expensive heuristic h_2 when $h_2(s) - h_1(s) > \tau$. However, computing $h_2(s) - h_1(s)$ requires computing in s both heuristics, defeating the whole purpose of reducing search time by selectively evaluating only one heuristic at each state. To overcome this pitfall, we take our decision rule as a target concept, and suggest an *active online learning* procedure for that concept. Intuitively, our concept is the set of states where the more expensive heuristic h_2 is "significantly" more accurate than the cheaper heuristic h_1. According to our model, this corresponds to the states where the reduction in expanded states by computing h_2 outweighs the extra time needed to compute it. In what follows, we present our learning-based methodology in detail, describing the way we select and label training examples, the features we use to represent the examples, the way we construct our classifier, and the way we employ it within A^* search.

To build a classifier, we first need to collect training examples, which should be representative of the entire search space. One option for collecting the training examples is to use the first k states of the search where k is the desired number of training examples. However, this method has a bias towards states that are closer to the initial state, and therefore is not likely to well represent the search space. Hence, we instead collect training examples by sending "probes" from the initial state. Each such "probe" simulates a stochastic hill-climbing search with a depth limit cutoff. All the states generated by such a probe are used as training examples, and we stop probing when k training examples have been collected. In our evaluation, the probing depth limit was set to twice the heuristic estimate of the initial state, that is $2\max_h(s_0)$, and the next state s for an ongoing probe was chosen with a probability proportional to $1/\max_h(s)$.

evaluate(s)
$\langle h, confidence\rangle :=$ CLASSIFY($s, model$)
if ($confidence > \rho$) **then return** $h(s)$
else
  $label := h_1$
  if $h_2(s) - h_1(s) > \alpha \log_b(t_2/t_1)$ **then** $label := h_2$
  update $model$ with $\langle s, label\rangle$
  return $\max(h_1(s), h_2(s))$

Figure 2: The selective max state evaluation procedure.

This "inverse heuristic" selection biases the sample towards states with lower heuristic estimates, that is, to states that are more likely to be expanded during the search. It is worth noting here that more sophisticated procedures for search space sampling have been proposed in the literature (e.g., see Haslum et al. 2007), but as we show later, our much simpler sampling method is already quite effective for our purpose.

After the training examples T are collected, they are first used to estimate b, t_1 and t_2 by averaging the respective quantities over T. Once b, t_1 and t_2 are estimated, we can compute the threshold $\tau = \alpha \log_b(t_2/t_1)$ for our decision rule. We generate a label for each training example by calculating $\Delta_h(s) = h_2(s) - h_1(s)$, and comparing it to the decision threshold. If $\Delta_h(s) > \tau$, we label s with h_2, otherwise with h_1. If $t_1 > t_2$ we simply switch between the heuristics—our decision is always *whether to compute the more expensive heuristic or not*; the default is to compute the cheaper heuristic, unless the classifier says otherwise.

Besides deciding on a training set of examples, we need to choose a set of features to represent each of these examples. The aim of these features is to characterize search states with respect to our decision rule. While numerous features for characterizing states of planning problems have been proposed in previous literature (see, e.g., Yoon, Fern, and Givan (2008); de la Rosa, Jiménez, and Borrajo (2008)), they were all designed for inter-problem learning, and most of them are not suitable for intra-problem learning like ours. In our work we decided to use only elementary features corresponding simply to the actual state variables of the planning problem.

Once we have our training set and features to represent the examples, we can build a binary classifier for our concept. This classifier can then play the role of our hypothetical oracle indicating which heuristic to compute where. However, as our classifier is not likely to be a perfect such oracle, we further consult the confidence the classifier associates with its classification. The resulting state evaluation procedure of selective max is depicted in Figure 2. If state s is to be evaluated by A^*, we use our classifier to decide which heuristic to compute. If the classification confidence exceeds a parameter threshold ρ, then only the indicated heuristic is computed for s. Otherwise, we conclude that there is not enough information to make a selective decision for s, and compute the regular maximum over $h_1(s)$ and $h_2(s)$. However, we use this opportunity to improve the quality of our prediction for states similar to s, and update our classifier. This is done by generating a label based on $h_2(s) - h_1(s)$ and learning from

this new example.[1] This can be viewed as the active part of our learning procedure.

The last decision to be made is the choice of classifier. Although many classifiers can be used here, there are several requirements that need to be met due to our particular setup. First, both training and classification must be very fast, as both are performed during time-constrained problem solving. Second, the classifier must be incremental to allow online update of the learned model. Finally, the classifier should provide us with a meaningful confidence for its predictions. While several classifiers meet these requirements, we found the classical Naive Bayes classifier to provide a good balance between speed and accuracy (Mitchell 1997). One note on the Naive Bayes classifier is that it assumes a very strong conditional independence between the features. Although this is not a fully realistic assumption for planning problems, using a SAS$^+$ formulation of the problem instead of the classical STRIPS helps a lot: instead of many binary variables which are highly dependent upon each other, we have a much smaller set of variables which are less dependent upon each other.

As a final note, extending selective max to use more than two heuristics is rather straightforward—simply compare the heuristics in a pair-wise manner, and choose the best heuristic by a vote, which can either be a regular vote (i.e., 1 for the winner, 0 for the loser), or weighted according to the classifier's confidence. Although this requires a quadratic number of classifiers, training and classification time (at least with Naive Bayes) appear to be much lower than the overall time spent on heuristic computations, and thus the overhead induced by learning and classification is likely to remain relatively low.

The Fast Downward Planning Framework

We have implemented selective max on top of the Fast Downward planning system. In this section we review the relevant (for optimal planning) capabilities of the IPC-2011 version of the Fast Downward planning system. Since Fast Downward incorporates many different algorithms and approaches, which have each been published separately in peer-reviewed conferences and/or journals, we will simply list the available components with pointers to further information for the interested reader.

The Fast Downward planning system (Helmert 2006) is composed of three main parts: the translator, the preprocessor, and the search component, which are run sequentially in this order. The translator (Helmert 2009) is responsible for translating the given PDDL task into an equivalent one in SAS$^+$ representation. This is done by finding groups of propositions which are mutually exclusive and combining them into a single SAS$^+$ variable. The preprocessor performs a relevance analysis and precomputes some data structures that are used by the search and certain heuristics. The search component then searches for a solution to the given SAS$^+$ task.

[1] We do not change the estimates for b, t_1 and t_2, so the threshold τ remains fixed.

Search The search component features three main types of search algorithms: eager best-first search, lazy best-first search (Richter and Helmert 2009), and enforced hill-climbing (Hoffmann and Nebel 2001). For the purposes of optimal planning, only eager search is relevant, since A^* is implemented on top of eager search by using $f = g + h$ and tie-breaking on h.

Heuristics Selective-max can combine arbitrary admissible heuristics from among the following admissible heuristics which are implemented in Fast Downward:

- Blind — 0 for goal states, 1 (or cheapest action cost for non-unit-cost tasks) for non-goal states
- $h^{\max}$ (Bonet, Loerincs, and Geffner 1997; Bonet and Geffner 1999) — the relaxation-based maximum heuristic
- h^m (Haslum and Geffner 2000) — a very slow implementation of the h^m heuristic family
- $h^{\text{M\&S}}$ (Helmert, Haslum, and Hoffmann 2007; 2008) — the merge-and-shrink heuristic
- h_{LA} (Karpas and Domshlak 2009; Keyder, Richter, and Helmert 2010) — the admissible landmark heuristic
- $h^{\text{LM-cut}}$ (Helmert and Domshlak 2009) — the landmark-cut heuristic

Chosen Configuration

Given the number of parameters available for selective-max, as well as the wealth of options of choosing which heuristics to combine, it is difficult to choose one configuration for a submission to the IPC. One option (which was implemented in the FD Autotune planner) is to use some automated algorithm configuration tool (Hutter et al. 2009) to choose a configuration.

In this submission, we chose to combine the two best heuristics available in Fast Downward (according to previous empirical results): $h^{\text{LM-cut}}$ (Helmert and Domshlak 2009) and h_{LA} (Karpas and Domshlak 2009; Keyder, Richter, and Helmert 2010). Since we are using h_{LA}, we also use the *LM-A** search algorithm (rather than regular A^*).

The h_{LA} heuristic uses landmarks generated by two methods: the RHW method (Richter, Helmert, and Westphal 2008) and h^m landmarks with $m = 1$ (Keyder, Richter, and Helmert 2010), which were combined into the same landmark graph (see BJOLP submission paper for details). The parameters for selective-max were chosen based on a limited set of experiments, and are described in the following table:

Parameter	Value
α (heuristic difference bias)	1
ρ (confidence threshold)	0.6
initial sample size	1000
Sampling Method	Probing
Classifier	Naive Bayes

References

Bäckström, C., and Nebel, B. 1995. Complexity results for SAS^+ planning. *Comp. Intell.* 11(4):625–655.

Bonet, B., and Geffner, H. 1999. Planning as heuristic search: New results. In *ECP*, 360–372.

Bonet, B.; Loerincs, G.; and Geffner, H. 1997. A robust and fast action selection mechanism for planning. In *AAAI*, 714–719.

Burke, E.; Kendall, G.; Newall, J.; Hart, E.; Ross, P.; and Schulenburg, S. 2003. Hyper-heuristics: an emerging direction in modern search technology. In *Handbook of meta-heuristics*. chapter 16, 457–474.

Coles, A. I.; Fox, M.; Long, D.; and Smith, A. J. 2008. Additive-disjunctive heuristics for optimal planning. In *ICAPS*, 44–51.

de la Rosa, T.; Jiménez, S.; and Borrajo, D. 2008. Learning relational decision trees for guiding heuristic planning. In *ICAPS*, 60–67.

Domshlak, C.; Karpas, E.; and Markovitch, S. 2010. To max or not to max: Online learning for speeding up optimal planning. In *AAAI*, 1071–1076.

Felner, A.; Korf, R. E.; and Hanan, S. 2004. Additive pattern database heuristics. 22:279–318.

Haslum, P., and Geffner, H. 2000. Admissible heuristics for optimal planning. In *AIPS*, 140–149.

Haslum, P.; Botea, A.; Helmert, M.; Bonet, B.; and Koenig, S. 2007. Domain-independent construction of pattern database heuristics for cost-optimal planning. In *AAAI*, 1007–1012.

Haslum, P.; Bonet, B.; and Geffner, H. 2005. New admissible heuristics for domain-independent planning. In *AAAI*, 1163–1168.

Helmert, M., and Domshlak, C. 2009. Landmarks, critical paths and abstractions: What's the difference anyway? In *ICAPS*, 162–169.

Helmert, M., and Röger, G. 2008. How good is almost perfect? In *AAAI*, 944–949.

Helmert, M.; Haslum, P.; and Hoffmann, J. 2007. Flexible abstraction heuristics for optimal sequential planning. In *ICAPS*, 176–183.

Helmert, M.; Haslum, P.; and Hoffmann, J. 2008. Explicit-state abstraction: A new method for generating heuristic functions. In *AAAI*, 1547–1550.

Helmert, M. 2006. The Fast Downward planning system. *Journal of Artificial Intelligence Research* 26:191–246.

Helmert, M. 2009. Concise finite-domain representations for PDDL planning tasks. 173:503–535.

Hoffmann, J., and Nebel, B. 2001. The FF planning system: Fast plan generation through heuristic search. *Journal of Artificial Intelligence Research* 14:253–302.

Hutter, F.; Hoos, H. H.; Leyton-Brown, K.; and Stützle, T. 2009. ParamILS: an automatic algorithm configuration framework. *Journal of Artificial Intelligence Research* 36:267–306.

Karpas, E., and Domshlak, C. 2009. Cost-optimal planning with landmarks. In *IJCAI*, 1728–1733.

Katz, M., and Domshlak, C. 2008. Optimal additive composition of abstraction-based admissible heuristics. In *ICAPS*, 174–181.

Keyder, E.; Richter, S.; and Helmert, M. 2010. Sound and complete landmarks for and/or graphs. In *ECAI*, 335–340.

Mitchell, T. M. 1997. *Machine Learning*. New York: McGraw-Hill.

Pearl, J. 1984. *Heuristics — Intelligent Search Strategies for Computer Problem Solving*. Addison-Wesley.

Richter, S., and Helmert, M. 2009. Preferred operators and deferred evaluation in satisficing planning. In *ICAPS*, 273–280.

Richter, S.; Helmert, M.; and Westphal, M. 2008. Landmarks revisited. In *AAAI*, 975–982.

Yoon, S.; Fern, A.; and Givan, R. 2008. Learning control knowledge for forward search planning. 9:683–718.