Reviewer 1

I enjoyed reading this paper! It is a nice piece of work on a very interesting topic.

Many thanks for your encouragement!

My main comments are related to i) comparison of algorithmic parameters and ii) extensions to 3D. Comment i) requires some work, but I think it should be relatively fast to do in a revision. Comment ii) can be discussed some more and left for future work.

i) Your paper contains a number of cases, but there are limited comparison of the suggested method using different tuning parameters m, k, alpha. I am guessing by tuning some of these one could have a greedy approach at one end versus a very deep one which is more time consuming at the other end. But I don't see much comparison of using various of these (extreme) inputs as it is now. I am also not sure how easy it would be to compare the suggested approach with ones like Q-learning or other RL / value iteration methods for your case?

The computational expense of the algorithm is primarily controlled by the number of trial trajectories generated m. In general, higher m will result in larger trees with deeper branches and higher computational cost. Changing progressive widening parameters (k α) can also change the computational expense and depth of search (and therefore the greediness of the resultant policy). Overly aggressive widening will tend to result in short-sighted policies that are one-step greedy, since the Monte Carlo estimates for each action will tend to be dominated by very short horizon trajectories. In our problem, this would tend to result in the degenerate behavior of always abandoning the prospect on the first step, since that was the only action with a non-negative expected one-step return.

Reinforcement learning based approaches may also be used to solve the presented mineral exploration POMDP. Without augmentation, they are likely not as well suited as the presented Monte Carlo method. Reinforcement learning methods such as Q learning or policy gradient learning, learn *offline* policies before any actions are taken in the real world. This requires functions to represent policies with functions that can generalize training examples to new experiences. Because these methods learn policies for the entire space of experiences that may be encountered, they tend to require significantly more training data than an online method, such as POMCPOW, that only solves for a single problem being encountered.

Reinforcement learning methods are also formulated for fully-observable domains, without explicitly accounting for state uncertainty. This can be addressed in several ways to allow RL methods to function, however, they are not as efficient as methods developed for domains with state uncertainty. In particular, RL methods tend to use very basic random sampling for exploration. Research has shown that UCB-type exploration (upper confidence

bound) in a tree is significantly more efficient and can make a large impact on problems where information gathering is important.

Value iteration methods may also be used here, with approximations for the continuous state, action, and observation spaces. Approximate value iteration methods, like point-based value iteration, do not allocate computation time as efficiently as MCTS like methods, as the tree tends to be constructed prior to learning, so that learned experiences cannot inform where to grow the tree most efficiently.

We added this additional material in the introduction and discussion section

ii) In practical mining operations, wouldn't there ordinarily be sequential 3D boreholes where one can choose and modify the drilling order / locations? One could also potentially stop data collection (and drilling) in one borehole after a certain depth (before the initial planned depth is reached). Along boreholes one could also have different data collecting frequency. The suggested strategy for collecting data seems a bit restricting in this setting - as it is 2D only in this paper. What more is needed or possible in 3D?

Yes, these are indeed various options. In 3D we would consider

- Location, orientation (azimuth, dip) and depth. All these can be made variable in the approach, but doing so would require extending the problem to continuous parameters. We have added a paragraph in the discussion that addresses the extension to 3D and what extra that would require
- POMCPOW can handle continuous actions as it is currently implemented.Increasing the number of parameters needed to define an action (adding azimuth + dip + depth to the current x, y) tends to increase the amount of computation needed to solve the problem. In such cases, a problem-specific policy can be used to augment the basic UCB exploration method.

iii) Some detail comments:

- Mark a_1 and a_2 on first axis of Figure 1, as well as have 'x' or similar as the axis label.

Agreed

- l175: This is accounts?

Typo: correct to “This ~~is~~ accounts for…”

-Around Table 1, I don't think all these comparison of AI and geo terminology are needed.

We have found it very useful in our own collaboration, so we believe it may in fact be needed to bridge fields

-In Sect 4.1 there is a discussion of "actions", and you state 'the agent may acquire measurements (data)'. But at this point in the presentation there is no observation terminology 'o'. Aren't the action here to mine or abandon?

Line 196. Should we include an additional sentence “The agent may also decide to abandon or proceed to mine the prospect”?

-Not sure L(o_{t+1}|...) is defnied in l 225 expression (it comes much later, I think)?

This is defined on line 214 as Z(o…), we should change it there or at line 225 to be consistent

-Algorithm 1, data line should have d <- d + e, e \sim N(0,\sigma_n^2)

Agreed

-Sunberg and Kochenderfer, 2018 paper is not on the reference list?

Agreed

- \sigma means several different things in the paper and can be a bit confusing.

Agreed

-You often say Figure X below. You don't need the 'below' here.

Agreed

-Would it be possible to color-code the histograms in Fig. 15 (+ similar ones) according to 'mine' or 'abandon' ? Couldn't you also have one bar for each outcome here, rather than bins 0-5, 5-10, etc.?

That is changed

Reviewer 2

Thanks for your interesting contribution. The manuscript is well written and very pleasant to read. The objectives are very clear and the method is rather well explained. I have a few suggestions and questions to clarify some points and facilitate the reproducibility of the work.

Thank you for your encouragement!

Line 56: could you explain what a non-sequential scheme could be in the context of mineral exploration, as it seems to contradict the previous sentence on line 38. It becomes clearer though, when reading the following paragraphs.

Last paragraph of section 1: Which of the mentioned approaches did you select for your demonstration?

Monte Carlo Planning

Section 4.2: how is the state space initialized?

We are a bit confused with this questopn. A state space is not initialized as it is just a space in the mathematical definition. I think the question may be asking how the space is defined/parameterized. In that case, I would say the space is defined by a function and set of parameters that define a massive anomaly (e.g. the center and radius of a circle), and a sample of a Gaussian random field from a Gaussian process. In practice, we represent the combination of these two as a two dimensional array, where each cell of the array represents the mineralization at that location.

Line 289: should it be r(s,a)= -Cost(s,a) or Cost(s,a)=-Cmeasurements to be consistent with the substraction of Cextraction in the profit?

We agree. I prefer the r = - cost representation for clarity.

Table 2: where does d come from? Formatting: should it be an algorithm rather than a table object? See e.g. the example in the latex template

d is the value of a particular particle in the ensemble.

Line 356: ‘At each time step’

Line 357: ‘The full tree is constructed’ – in the case of the POMDP ?

Yes we can see why this is confusing. The “full” tree here does not refer to the complete tree possible under a POMDP. A better wording of what we’re trying to communicate here may be “The tree construction process is completed before …”

Line 359: by trial trajectory, do you mean a branch of the tree or realization of the full tree? How is the (partial) tree generated? What is the prior over the trajectory length , and between the different actions (explore further, mine or abandon)

A trial trajectory is a simulation that starts at the root node and continues until a termination condition is reached (e.g. maximum measurements have been taken or “MINE/ABNANDON” was selected). A portion of each trial trajectory is added to the existing tree structure every step. The logic of how the tree is explored and constructed is defined in the POMCPOW paper. Fully explaining the logic here would require a more in-depth description than is in the scope of this paper.

Line 378: previous visits cumulated over the previous iterations t ?

Agreed

Lines 423 to 425 and figure 9 bottom right panel: can you clarify the stopping criteria as at =5 the mean of the ensemble decreases and is getting smaller than the extraction cost. Can you also clarify how the value of gained information is assessed?

We are not sure we understand the first part of this question. The stopping criteria is not explicitly defined, beyond a maximum number of measurements allowed. The stopping behavior of the agent is learned during the optimization process.

The value of information is not explicitly calculated. An optimal policy would continue to gather information by taking measurements, so long as the value of that information exceeds the measurement’s cost. The value of the information would be calculated by the difference between the expected value of the “MINE/ABANDON” decision with and without the information.

Figure 13: missing scale for the mean average error

Thanks for your careful reading