Firstly, we thank the reviewers for their valuable comments. We will address the comments of each reviewer in what follows.

R1, R4: *The results are very specific to the particular model*: Indeed it is the case that our theoretical results assume that data providers are constrained in ℓ_2-norm, and that both the learner and the data providers are interested in solving linear least squares problems. However, it is fairly easy to see that our theoretical results generalise to kernel ridge regression, and thus, our theoretical results hold for a far wider array of learning models than linear predictors. In addition, we believe that our work is an important first step in relaxing the overly pessimistic assumptions of adversarial machine learning. Whilst it is not reasonable in practice to assume that data is sampled i.i.d. from the distribution of interest, neither is it reasonable to assume that data is provided with the sole intention of hindering learning. By taking the incentives of data providers directly into account during the optimisation process, we hope to better reflect the reality of sampled data in practice. As previously stated, we believe our work forms a first step in achieving this goal.

R1: *It is not clear from the paper how their model differs from previously considered ones*: In this paper, we study a specific subclass of SPGs in which both the learner and data providers are interested in solving least squares linear regression problems with their own data labels. Whilst Brückner and Scheffer (2011) give an algorithm which converges to local optima for general SPGs, we give an algorithm for this specific subclass of SPGs which converges to ***globally optimal solutions***. Note that SPGs are bilevel optimisation problems, which are, in general, NP-hard. With this in mind, we believe our results are novel and significant, as we have provided a practically efficient algorithm for a large subset of bilevel optimisation problems. Algorithms for specific subclasses of SPGs have been considered before, but to the best of our knowledge, our paper is the first to consider SPGs for linear least squares regression. Closely related to our work is the problem of robust regression in which the challenge is to choose a model which performs well in the presence of worst-case noise. The subclass of SPGs we consider allow us to model data providers with more nuanced motivations, which we believe are more likely to arise in practice. We will highlight these differences from previous work in future versions of the paper.

R2: *Why would the learner ever evaluate on both the manipulated and unmanipulated data in practice?*: We believe that our work has applications whenever a learner has access to a reliable, but costly, verification process. In practice, many machine learning models make decisions which affect those who provide data. Thus, data providers may manipulate the data they submit in order to obtain a preferred outcome. In cases in which the learner has access to a verification process, the learner can recover the originally sampled data as well as the goals of the data providers. Unfortunately, this verification process may be too expensive to use on every single data point. However, the learner may be able to verify a sample which can be used to learn the motivations of data providers and select a model which anticipates the manipulations that data providers are likely to make. We believe that our theoretical model captures this dynamic. Note that using verified data points to improve learning has been applied extensively in adversarial machine learning contexts (for examples, see Charikar *et al.* (2017) and Raghavendra and Yau (2020)). One practical example of such a setting is insurance fraud. An insurance company may gather information from a customer to better evaluate potential risk. However, a smart customer, who knows the information they provide could be used to decide the cost of their insurance, may lie when submitting their information. Whilst insurance companies can often verify information regarding their customers, verification is often expensive, requiring a significant deployment of staff and/or resources.

R2: *I am guessing it is the "interior point" line - this seems suspiciously bad. Is this a strong baseline / was the baseline correctly executed with proper hyper-parameter choice?*: The reviewer is correct in their assumption that the interior point line in figure 1 corresponds to the method of Brückner and Scheffer. We shall update the legend of figure 1 to be more clear. This approach involves reformulating SPGs into a single-level optimisation problem which can be solved via conventional optimisation techniques. In their work, Brückner and Scheffer highlight a number of SPGs for which this problem reformulation is simple and can be easily solved. For the family of SPGs we consider, the problem reformulation is nontrivial and nonconvex. The error tolerances used for both our algorithm and the interior point method we use to solve this problem reformulation are the same. We believe that the poor performance of the interior point method for higher values of γ reflects ill-conditioning issues present in the nonconvex objective, but we cannot confirm this.

R2: *Is there any way to contextualize the MSE in Figure 1?*: A data label in the Medical Personal Costs dataset corresponds to the medical charges for a given individual. Since these charges range from \$1000 to \$63,000, we numerically scale the data labels by dividing them by 100 before passing them to each algorithm. For values of $\gamma > 0.4$, we observe that, for modest data providers, our algorithm is at least more accurate by \$45 on average. For severe data providers, we observe an even greater difference. For example, when $\gamma > 0.5$, the predictions made by our algorithm are at least \$120 more accurate on average. Considering the modest data provider only sought to reduce their charges by \$100, and the severe data provider only sought to reduce their charges by \$300, we believe these differences are fairly significant. We shall provide more contextualisation for our empirical experiments in future versions of the paper.