Strategy for Revision

Major Changes

Summary (in somewhat of an order of priority):

1. The lack of overlap should be expected because they are from different parent samples.
2. Selection biases are inherent to the parent samples, not the identification methods.
3. Limitations on spectroscopic identification and mass characteristics of the sample by aperture size need to be re-evaluated.
4. GalaxyZoo sample is not clean enough.
5. GAMA spectroscopy sample is not clean enough.
6. Machine learning is limited by resolution moreso than training set.
7. Comparison between the two imaging methods needs to be more direct.
8. Comparison to other lens searches needs to be re-evaluated.

Details:

1. **The lack of overlap should be expected. Focus away from here.**
    - Considering that the three sets of lens candidates are selected from **three distinctive parent samples, I think such little overlap is well expected.**
    - By design, the spectroscopy method requires at least some portion of the lensing features being close to (and likely blended with) the lensing galaxies, while image-based methods such as machine learning preferentially select lens candidates with lensing features reasonably separated from the lensing galaxies. So the **little overlap between**

these two methods is trivial and straightforward to understand (e.g. see Sonnenfeld et al. 2018).

2. <u>“Intrinsic **selection biases** are *not of the methods themselves*, but of **parent samples**.” We should focus there.</u>
    - **<u>“Compare lens candidates with the parent sample</u> from which they were selected for each individual method independently.”**
        - Bolton et al. 2008 compared SLACS A-grade lenses and lens candidates with a control sample, and **showed using KS tests that the SLACS lenses and lens candidates are statistically consistent with the parent SDSS galaxy population in terms of stellar mass, size, and velocity dispersion**. **<u>Sonnenfeld et al. 2018</u>** used two image-based methods and one spectroscopy method to select lens candidates and compared their performances. **<u>Their comparison is similar to what's done here, but more sensible because their three methods used the same parent sample</u>**. They found that **lenses and lens candidates are slightly biased towards the high-mass end compared to the parent sample, but no particular redshift region is favored.**
        - It's clearly not the methods that "identified galaxies with a distinct range of redshifts", but the parent samples are selected that way.
3. **Spectroscopy limitations based on aperture size** need to be re-evaluated.
        - I don't think "the lensing arc does not intersect with the SDSS aperture" was the main reason that most SLACS strong lenses are at z > 0.1. First of all, the lensing cross section drops rapidly below z~0.5 (e.g. Hilbert et al. 2008, Lapi et al. 2010), so naturally there are fewer lenses at lower redshifts. In addition, the SLACS survey (Bolton et al. 2006) imposed a redshift cut of z > 0.15 for the majority of the lens candidates that were selected from the LRG sample.
        - Again, I don't agree "spectroscopy structurally misses lower redshift lenses". I think this selection bias is very mild, and only at the lowest redshifts (z ~< 0.1).
        - I don't quite understand the statement of "there is a maximum total mass ...". I don't think there is a sharp mass limit for the spectroscopy method. Please clarify.
        - The authors seem to indicate that the spectroscopy method tends to find lower-mass lens candidates compared to the machine learning method, which is not well supported in my opinion. In the literature, Sonnenfeld et al. 2013 actually showed that the stellar mass distribution of the SLACS sample is similar to the SL2S sample, which were selected based on imaging data and should be statistically comparable to a machine learning sample. As I mentioned previously, Bolton et al. 2008 showed that the SLACS sample is statistically consistent with its parent sample in terms of mass. And it's

possible that the majority of the low-mass spectroscopy candidates in Holwerda et al. 2015 are not lenses.

4. Clean up Zoo sample.

- **"It would be useful to provide the separations between the arcs or arcs and galaxies for the GalaxyZoo candidates and their estimated Einstein radii based on velocity dispersion (from spectroscopy) and some fiducial source redshifts."**
    - The fact that there are more overlaps between machine learning and citizen science when the lens score threshold is lowered seems to **suggest that many GalaxyZoo candidates are simply not good lens candidates, perhaps because they were selected with less physical constraints compared to the machine learning method**. For example, **galaxies with obvious, widely-separated arcs would have a high "Lens or Arc" score in GalaxyZoo, but they are simply not lenses given the separations and masses.** Again, ***comparing lens candidates, especially ones that have not been reasonably cleaned, can be biased***!
    - Section 3.3 I don't think Figure 6 shows "The cuts we imposed do not skew the results ...". **I would expect to see all GalaxyZoo candidates in Figure 6 and comparisons of the redshift and stellar mass histograms.**

5. Clean up spectroscopy sample. We have over-emphasized the cleanliness and reliability of the spectroscopic approach.

- **Check the redshift differences** suggested in following note**.**
- "It's again **worth estimating the Einstein radii and lensing cross sections for the spectroscopy candidates from velocity dispersions and redshifts**."
- "The authors can **also estimate the lensing probabilities for the spectroscopy sample following the procedures in Bolton et al. 2004,** which should give a better idea of how many true lenses are expected from this sample."
    - I checked the PG+ELG sample in Holwerda et al. 2015, and found that the **redshift differences between the foreground and background galaxies are small for the majority of the PG+ELG pairs**. **Almost 50% of the PG+ELG sample have redshift differences smaller than 0.1, and almost 70% have redshift differences smaller than 0.2**. I actually **highly doubt that those PG+ELG pairs with small redshift differences could be true lenses given the seemingly small lensing cross sections**. Of course, **it would be very interesting if some of the low-mass candidates in Holwerda et al. 2015 turn out to be true lenses**. But **I suspect that the false positive rate in this spectroscopy sample is much higher towards the low-mass end**. Then it's not surprising to see the spectroscopy candidates appear less massive on average than the machine learning candidates.'
    - The authors claim that "The spectroscopic approach typically results in a small but very clean sample". Given the typical 50% success rate of the SLACS survey, I'm not sure whether I'll call this "very" clean. One can probably

get a similarly small and clean lens sample using the machine learning approach (by adopting a high score).

- The authors seem to imply that ~70% of the spectroscopic lens candidates are true lenses by using "verified", while in fact Chan et al. 2016 only found 10 probable lens systems among those candidates.

6. **Machine Learning limitation** in “imaging resolution rather than training set.”

   - I actually think that **the fundamental limitation of the machine learning approach in Petrillo et al. (2018a) is imaging resolution rather than the training set.** For a particular resolution, even if the training set extends to lower masses, Petrillo et al. (2018a) will still preferentially select candidates with lensing features that can be clearly separated from the lensing galaxies, which naturally results in a sample that is biased towards more massive galaxies.

7. **Comparison between Zoo and Mac** needs to be more *direct and strictly constrained.*

   - “To make a fair comparison, I suggest either **singling out GalaxyZoo candidates that also pass the LRG cuts** used in Petrillo et al. (2018a) *or* **applying the machinery in Petrillo** et al. (2018a) to all GalaxyZoo candidates**.”**
     - **I suspect that many GalaxyZoo candidates are actually low-mass, blue galaxies that were simply skipped (and thus unclassified) in Petrillo et al. (2018a) because Petrillo et al. (2018a) only looked at luminous red galaxies (LRGs)** in the first place. In principle, the machine learning method itself should be able to largely recover candidates from citizen science as they're both based on imaging data.

8. **Comparisons to other lens searches** needs to be re-evaluated.

   - How many SLACS+SLACS for the Masses+BELLS lenses are in the GAMA spectroscopic sample, and how many are recovered in Holwerda et al. 2015?
   - The SLACS and SLACS for the Masses lens samples were selected from the SDSS LRG sample that has a limiting magnitude of r=19.5 (Eisenstein et al. 2001) instead of what's shown in Figure 12. The BELLS lens sample was selected from the BOSS LRG sample that also has a limiting magnitude of r~19.5 (Eisenstein et al. 2011).
   - The limiting r-band magnitude and mean seeing for the DES sample is 24.9 mag and ~1.1 arcsec according to Jacobs et al. 2019. The typical seeing for the DECaLS sample is approximately 1 arcsec according to Huang et al. 2019.
   - It's unclear from the color bar in Figure 12 what the candidates densities are. Please also provide the actual calculations and values in the text. For SLACS, SLACS for the Masses, and BELLS, is the number of confirmed lenses or the number of lens candidates used? I'm surprised to see that GAMA II has

significantly higher candidates density than the other three spectroscopy samples that have similar limiting magnitudes. What makes GAMA II so effective? Or maybe lots of GAMA II candidates are simply not lenses?

*** **Interesting questions** brought up.

- "Why was G136604 not identified in Holwerda et al. 2015? It's worth showing the SDSS and GAMA spectra for this system."
- "How many SLACS+SLACS for the Masses+BELLS lenses are in the GAMA spectroscopic sample, and how many are recovered in Holwerda et al. 2015?"
- "It's worth showing the KiDS image alongside the HST image for G593852 in Figure 10."
- "It's actually interesting to show the KiDS images for the 5 overlapping spectroscopic lens candidates when more relaxed cuts are used, and highlight them in Figure 8. Even though they don't have a high lens score according to imaging data, the fact that they also show higher-redshift emission lines makes them highly probable lenses."
- Are redshifts for the machine learning catalog and citizen science catalog also spectroscopic? What is the typical uncertainty in stellar mass? It would useful to include error bars in the plots involving stellar mass.
    - Uh, yeah? We got them all from the same source.

Strategies to address each problem

1. **The lack of overlap should be expected because they are from different parent samples.**
    a. Organize and review the steps of each method and parent sample to determine a path of processing for each
        i. Directly locate the raw sample of each and determine the exact points at which each bias is introduced.
        ii. GAMA Spectroscopy:
            1. GAMA (obviously) fields G09, G12, and G15
                a. DR-2? 3?
            2. AUTOZ, etc.
        iii. GalaxyZoo:
            1. KiDS postage stamps.
                a. From GZ website "At the start of 2017 we added in images from another recent project that builds upon the

legacy of SDSS, the Galaxy And Mass Assembly Survey (GAMA). Specifically, we have included optical images from the Kilo-Degree Survey (KiDS) over the region of sky covered by GAMA's comprehensive multiwavelength dataset. “

2. GZ voting, etc.

iv. Machine Learning:

1. KiDS
2. LRG cuts, etc.

b. Compare GAMA and KiDS directly.

i. How many of their measurements overlap?

ii. What is different about those two parent samples?

2. **Selection biases are inherent to the parent samples, not the identification methods.**

   a. Compare lens candidates with the parent sample for each method independently.

   i. “Stellar mass, size, velocity dispersion”, etc.

   ii. Could be done through plots that show side-by-side comparisons of stellar mass and redshift for each method relative to their parent sample.

   b. Follows from processing path from Issue 1.

3. **Limitations on spectroscopic identification and mass characteristics of the sample by aperture size need to be re-evaluated.**

   a. Determine quantitative relationship between aperture size and maximum total mass cutoff.

   b. Address redshift-dependence as well.

   i. Lensing cross section drops rapidly at lower z

   ii. SLACS imposed z > 0.15 for LRG sample

   iii. Selection bias is very mild, only at z ~< 0.1

   c. Will be greatly affected by changes in GAMA Spec sample

4. **GalaxyZoo sample is not clean enough.**
   a. Vet more thoroughly
      i. Tighter restrictions?
      ii. Show cutouts?
   b. “Provide separations between arcs or between arcs and galaxies.”
   c. “Estimate Einstein radii based on velocity dispersion (from spectroscopy) and some fiducial source redshifts”
      i. I’m not sure what all this entails.
5. **GAMA spectroscopy sample is not clean enough.**
   a. Vet more thoroughly.
      i. Read Chan et. al 2016 for report on number of probable lenses from this sample
      ii. Look for redshift differences.
         1. Almost 50% had z differences < 0.1
         2. Almost 70% had z differences < 0.2
            a. Referee doubts these are lenses (see above)
      iii. Estimate lensing probabilities following procedures in Bolton et al. 2004.
      iv. Look at stellar masses of the results, if small, then that is interesting.
      v. Follow up with IF spectroscopy of GAMA spec candidates.
         1. Could be our saving grace for validity of that sample.
6. **Machine learning is limited by resolution moreso than training set.**
   a. Discuss resolution as a factor for both GZ and ML, since they do have the same images.

7. **Comparison between the two imaging methods needs to be more direct.**
    a. Along the same lines as above…
    b. Apply ML to GZ candidates?
    c. Analyze GZ candidates that also pass the LRG cuts.
        i. I don't think singling out only these as valid makes sense… We're trying to see what other types of lenses are possibly found, and instituting this bias would only confirm the bias we're trying to point out?
8. **Comparison to other lens searches needs to be re-evaluated.**
    a. Check if SLACS, S4TM, and BELLS are in GAMA fields.
        i. If so, they should be considered here, and compared like all the others.
    b. Completely redo graphs of comparisons.

***

Other things to add…

1. Show all cutouts.
2. Don't be afraid to make smaller samples.
3. Show KiDS images for the 5 overlapping spectroscopic lens candidates when more relaxed cuts are used. (see above)
4. Show all KiDS images next to the HST images.