Dear reviewers, we greatly appreciate your remarks and suggestions. We will address the comments in the following.

1. **Page 4, the ℓ_∞ barrier. The limitations of performing first-order methods over the ℓ_∞ ball go beyond the fact that strongly convex functions grow too fast (Lemma 3.1): there are provable oracle complexity lower bounds for this problem (see Guzman and Nemirovski, "On Lower Complexity Bounds for Large-Scale Smooth Convex Optimization" 2015).**
   Thank you for pointing out the connection to the result by Guzman and Nemirovski, and their lower bound result for smooth convex minimization over ℓ_∞-ball is indeed very interesting in this context. We will include a brief discussion on this point after Lemma 3.1 in the updated version.
2. **Page 6, line 215. "is precisely the primal-dual gap". I don't think this is the case, since the primal-dual gap should be the supremum over $\overline{w}$ variable. Could the authors clarify this?**
   Thanks for spotting this typo. We indeed meant that taking a $\sup$ over the $\overline{w}$ variable gives the primal-dual gap of the solution (x, y, z) as defined earlier in line 109. We will correct this accordingly.
3. **Page 6, definition 4.1. In the same spirit as the previous comment: What is the relation between an eps-optimal solution and the duality gap?**
   If we have an ϵ-optimal solution of Eq(2) (i.e., Definition 4.1), we can read from it a solution (x, y, z) whose duality gap is at most ϵ (as defined in Line 109). We further show (in Lemma 4.3) that such a solution translates to either a ϵ-approximate solution x to the original mixed packing-covering LP, or a certificate (with (y, z)) that the original MPC is infeasible.
4. **Page 7, eqn (3). Shouldn't the supremum be over $\overline{w}$, instead of $\overline{x}$? Page 7, line 253. I believe $\widetilde{\eta}$ is exactly the function defined in (3).**
   Eq(3): Yes, the supremum should be over $\overline{w} \in \mathcal{W}$ instead of $\overline{x}$, and $\widetilde{\eta}$ is the same function defined in (3). We will correct the typo, and clarify the definition in the revised version.
5. **Page 7, algorithm 1. It is mentioned that the algorithm is related to dual extrapolation, but it also looks similar to the extra-gradient method (a.k.a. Mirror-Prox). Are there connections in this direction?**
   As discussed in the dual-extrapolation paper, Nesterov's method was on a high level motivated by Mirror-Prox. The two methods can be the same if the setup is Euclidean and the underlying space is unconstrained, but in general cases they give very different algorithms. Essentially dual-extrapolation carries out extra-gradient steps in the dual space. In that spirit, the connection is similar to the case of gradient descent vs mirror decent.
6. **Page 11, line 381. This is a very minor detail, but after assuming there exists a feasible $\widetilde{x}$, shouldn't the second case start with: Suppose no $\widetilde{x}$ is feasible? (as opposed to just x). Starting from x is formally correct, but it is a bit confusing, given the assumption of the previous paragraph**
   We agree the wording is a bit confusing, and indeed the first half is mostly redundant. The succinct argument should be that if the x we obtained is an ϵ-approximate solution to the MPC instance, then we are done, i.e. case (1). In the second case, i.e. if our x is not an ε-approximate solution, we certify that the original MPC instance is infeasible. The first half is proving the additional statement that if MPC is feasible then x must be ε-approximate solution which is not part of the statement of the lemma. Hence it is confusing the reader. We will make appropriate changes here.
7. **Page 11, proof of Lemma 4.5. This proof needs some polishing. I believe one can directly appeal to Schur complements, and the first two lines are just confusing. Besides, there is a typo on the inverse of $A : a_{22} = a$, and not d.**
   Thanks for pointing this out. We will reword the first couple sentences, and fix the typo accordingly. On the high level, this is supposed to be a fairly elementary (but maybe a bit tedious) technical result.
8. **The paper could have provided a more intuitive explanation of the joint regularization used.**
   We agree that the current approach appears more in a bottom-up way, i.e. starting from the desired algebraic properties of area-convexity and engineering an appropriate function. A more intuitive interpretation of the joint regularization would offer further insight to extend the result. On a high level, comparing to independent regularization (i.e. using $\phi_1(x) + \phi_2(y) + \phi_3(z)$ to regularize), the joint regularization can be viewed as adaptively using a different regularization on x based on the current y, z, which makes the regularization behaves more locally. Although this local nature makes the guarantees weaker, it is sufficient since we only need such guarantees within the local neighborhood where the update step is carried out. This is also the key to why ρ is $O(n^{o(1)})$. Intuitively, we weigh the term involving x_i in the regularization dynamically based on how 'significant' each x_i is at the current step, and the total 'significance' over all variables at any time can be bounded by the infinity norm rather than the number of coordinates n. In passing, we would like to remark that this is still an algebraic interpretation and obtaining a geometric interpretation might be difficult because of high dimensionality of the problem. If one would really like to go in this direction then a starting possibility would be the 2-D toy example in line 183. However, we haven't ventured into this side yet.