We thank all the reviewers [**R1, R2, R3, R4**] for their helpful feedback! We address their comments below.

Clarity and accessibility. We agree with **R1** that more focus on OT makes the paper more accessible: E.g., the counterexample in Remark 1 will be stated directly for OT in a revised version (this works analogously, where the current μ will be one of the marginals in the OT problem). Also, in order to address the request from **R4** for more examples of the form (P), we will add a table of known instances from the literature in the Appendix which will be referred to after Example 1.

R1+2 underline some clarity issues in the notation. We will strive, accordingly, to improve the overall transparency of the paper. More precisely, we will better motivate the use of the general problem (P), describe the role of e_j, h_j, π_j in the definition of $\mathcal{H}$, clarify and adjust non-standard notation (e.g., for the push-forward operator), and add more details on certain points (e.g., the switch from ν to θ_T, meaning of "hidden dimension", meaning of "θ sufficiently rich").

R3 highlights a typo on Line 123. We thank the reviewer for pointing this out and will certainly correct it.

Numerical experiments. As suggested by **R1**, we will add a controlled numerical evaluation in the form of a simple example with OT between Gaussians in the Appendix. Also, as suggested by **R2**, we will include additional details on §5 in the main body, so that the numerical results are self-contained without referring to the Appendix.

References. **R1+3** rightfully suggested some relevant works from existing, related literature that have not been mentioned. In particular, we will discuss the relation to Γ-convergence of MinMax problems and to the recently introduced projection robust Wasserstein distance. Additional references on unbalanced OT will be provided as well.

Theorem 1 (T1), applicability and extensions. **R1** asks whether the results in T1 apply to the experiments in §5. The general requirements for T1 (except for compact support, which is practically given by cutting off the marginals) are satisfied, and thus T1 (i) is applicable. Since $e_3(x, y) = y - x$ for MOT is not non-negative, T1 (ii) is not applicable. For the DCOT problem in §5.1, we believe T1 applies in full (given compactness; for T1 (ii), see the following discussion about the assumption on T_m). **R1** further asks when the hypothesis for T1 (ii) are expected to hold. We agree that more must be said. Note that while the hypothesis on T_m for convergence $(P_\psi^m) \to (P_\psi)$ looks strong, it is still a lot weaker than the requirement for convergence $(P^m) \to (P)$, which (in Remark 1) would be that $T_m = \hat{T}$ for some finite m. Indeed, a strength of T1 (ii) is that $\theta \circ T_m^{-1}$ may put 0 mass to certain small regions. For instance, in the setting of Remark 1 (where T1 (ii) is applicable), one may restrict the map T_m to $[\varepsilon_m, 1]$ (where $\varepsilon_m \to 0$ for $m \to \infty$). In this region, the density $\frac{d\theta \circ T_m^{-1}}{d\theta \circ \hat{T}^{-1}}$ can be approximately 1 for suitable T_m, while in the remaining support $[0, \varepsilon_m)$, the density $\frac{d\theta \circ T_m^{-1}}{d\theta \circ \hat{T}^{-1}}$ is 0, which does not interfere with the assumption of T1 (ii). In a revised version, we will add a related remark.

R3 further asks whether equality may be obtained in T1 (ii) for certain choices of divergence. This is a great point, and indeed we expect equality to hold in some cases. Our attempts to prove this were so far limited by (seemingly necessary and not in the literature) results on the regularity of the functions h_j occurring in the dual formulation of divergences. We will also add a discussion. This will also be related to the comment by **R2** pointing out that no quantitative rates are given, which (in our understanding) also requires knowledge about regularity of optimizers T and h_j.

Further points. **R1** comments on optimizing over MLP intersected with Lip_1. No simple construction that is dense is known to us. The cited results in Poggio et al. (2017) build on highly non-constructive work by Bach (2017). As discussed in §B.3, one utilizes a gradient penalty in practice. We will add a short discussion after introducing (P_L).

R3, Comment 4: We are excited about the mentioned paper and will add a reference and discuss computational complexity in a revised version. An analysis for arbitrary constraints is however beyond the scope of this paper. Aside from the difficulty, a reason is that we want to keep the focus on neural network methods instead of discretization.

R4, Correctness: In Eqs. (5)-(10), the dependence on T is encoded by θ_T. Thus, no term T is missing. **R4, Comment 1**: Eqs. (5) and (6) are special cases of Eqs. (7) and (10) (specified to OT), not the other way around. **R4, Comment 2**: Problems (P) and (P^m) are not equivalent. Note that approximation is not enough, because there are constraints in the optimization problems. The key difficulty is that approximation under constraints is much harder to obtain, and it may indeed simply not hold (see Remark 1). Further, even for the results for (P_L^m), the theory by Yarotsky (2017) is not applicable, because (in arbitrary dimensions) Yarotsky focuses on variable-depth networks, while in our paper the depth is fixed. **R4, on Remark 2:** We agree that the given sentence is misleading. What we want to express is, e.g., that $(P_L^m) \approx (P_L)$ is more justified as an approximation than $(P^m) \approx (P)$ (and not that $(P_L^m) \approx (P)$ is more justified than $(P^m) \approx (P)$). We will clarify this in a revised version.

Novelty of the theoretical contribution. **R3+4** mention that the novelty and contribution of our paper are quite limited and also that the theoretical results seem rather straightforward or even obvious. We hope our arguments above on why (P) and (P^m) are not equivalent showcase that T1 is not obvious. We believe that the statement and proof of T1 (ii) are quite involved, and T1 (i) builds on very strong approximation results for Lipschitz functions. Further, while Remark 1 and T1 are new even for OT, the introduction of the paper cites many problems of practical interest from the literature that the generalized class (P) is necessary for. The corresponding generalized regularizations were not straightforward to obtain. Different plausible generalizations are, e.g., to remove the term $|e_j|$ in the last term of (P_ψ), or require that $h_j \circ \pi_j$ is L-Lipschitz for (P_L). The final statements of (P_L) and (P_ψ) in the paper are the result of extensive numerical experiments and theoretical analysis. Finally, as mentioned above, some theoretical aspects (e.g., verifying the hypothesis for T1 (ii)) will be expanded on in a revision.