We would like to thank all five (!) reviewers for their detailed reviews and their suggestions / questions, which will help to further improve this paper. In the following we will try to address the main points raised.

Experiments (reviewers 2, 3, 4, 5): Given that our main contribution is the theoretical analysis of (emphasized) denoising as a training technique, and its ability/lack of preventing the autoencoder (AE) from overfitting to the identity-function, the space remaining for the experimental section is naturally limited. We nevertheless aim for experimental reproducibility as well as an empirical comparison to other baselines by exactly following the experimental set-up in [19]. Based on the reviews, we will make our paper more self-contained, and add a short review of the experimental protocol in [19]. In the table below, we also added the various models evaluated in [19] for ease of comparison: two linear models (SLIM , WMF), and three deep non-linear AEs (CDAE, MULT-VAE PR, MULT-DAE)–we will also add their citations to the paper. All approaches in this table can be compared to each other: this shows not only that EDLAE (linear model) obtains competitive results compared to the various (non-linear) baselines, but also that the differences among the various types of regularizations can actually be substantial (i.e., possibly larger than the differences between different model-classes). In the table below, we also added Recall @20 and @50, the two metrics we had omitted in the paper, as they largely reflect the same behavior as nDCG@100 does (in more detail, the table shows that EDLAE empirically improves in particular the ranking accuracy in the top-N for *smaller* N). While we limited this paper to *linear* models for reasons of analytical tractability (see paper for the various derived insights), in practice the stochastic version of emphasized denoising is readily applicable to training deep non-linear models, as done in [33], where it was shown that emphasized denoising empirically improves on (standard) denoising.

Identity Function (reviewers 1, 3): We will clarify the motivation/objective at the beginning of this paper in more detail. Due to space constraints, we had unfortunately shortened this part of the paper too much, as we now realize. There are many applications where the data may be noisy or where we want the AE to be able to generalize to unseen data (e.g., in the areas of image processing, information retrieval, etc.). Learning the identity function (i.e., predicting each feature i in the output layer from the *same* feature i in the input layer) is obviously not useful for such prediction tasks. Instead, the AE has to learn all the relevant dependences/interactions among the features, as to achieve maximum prediction accuracy on unseen noisy test-data. Intuitively speaking, when the learned AE makes predictions for a feature i in the output-layer by relying 'too much' on *the same* feature i in the input layer (i.e., identity function), and 'not enough' on the *other* features it depends on, we call this 'overfitting towards the identity function' in this paper. In fact we chose collaborative filtering on implicit feedback data for our experiments exactly because the value 0 in the user-item training-matrix conflates true negative items (which the user would never select) and the true positive items that the user has not selected yet in the observed (training-)data: predicting the positives in the disjoint test-set hence hinges on the AE's ability to predict each feature/item i from the *other* items $j \neq i$, i.e., prediction accuracy immediately suffers in our experiments if the AE overfits to the identity function.

Low-rank models & Denoising (reviewer 2): While fully emphasized denoising (controlled by parameters $a > b = 0$) completely eliminates the 'overfitting toward the identity function', i.e., diagonal of matrix $\mathbf{B}$ (see Section 4 in the paper), note that this is *decoupled* from the amount of L2-norm regularization applied to the off-diagonal entries of $\mathbf{B}$ (which is controlled by the value of dropout-probability p, or Λ), see Eq. 6. In contrast, this decoupling is absent (1) when using (standard) dropout-denoising, which merely induces L2-norm regularization in a linear model (in the asymptotic limit, i.e., when trained to convergence, even on a finite amount of training data), and hence regularizes both the diagonal and off-diagonal entries in the same way (see also Eq. 5); (2) when using low-rank models, where a decrease in the model-rank not only reduces the overfitting towards the identity function, but also the model-capacity in general, possibly leading to under-fitting for small model-ranks. Due to this coupling, the overfitting to the identity can only be prevented partially without suffering from under-fitting the data when using only low-rank and/or denoising, resulting in worse ranking-metrics in the table below (cf. rows 1-4 vs. EDLAE).

We find it remarkable in l. 154-6 (reviewer 4) that training (diagonal removed) differs from prediction (with diagonal).

			ML-20M			***Netflix***			***MSD***		
		model training:	Recall @20	Recall @50	nDCG @100	Recall @20	Recall @50	nDCG @100	Recall @20	Recall @50	nDCG @100
	1.	$\|\|\mathbf{X} - \mathbf{X}\mathbf{U}\mathbf{V}^\top\|\|_F^2 + \lambda \cdot (\|\|\mathbf{U}\|\|_F^2 + \|\|\mathbf{V}\|\|_F^2)$	0.345	0.467	0.376	0.326	0.406	0.357	0.200	0.278	0.249
	2.	$\|\|\mathbf{X} - \mathbf{X}\mathbf{U}\mathbf{V}^\top\|\|_F^2 + \lambda \cdot \|\|\mathbf{U} \cdot \mathbf{V}^\top\|\|_F^2$	0.376	0.508	0.407	0.342	0.423	0.374	0.222	0.303	0.270
	3.	$\|\|\mathbf{X} - \mathbf{X}\mathbf{U}\mathbf{V}^\top\|\|_F^2 + \|\|\tilde{\Lambda}^{1/2} \cdot \mathbf{U} \cdot \mathbf{V}^\top\|\|_F^2$	0.382	0.515	0.417	0.351	0.434	0.384	0.258	0.347	0.311
	4.	DLAE (sampled)	0.383	0.515	0.417	0.351	0.435	0.384	0.257	0.346	0.311
	5.	EDLAE	0.389	0.518	0.420	0.359	0.443	0.392	0.263	0.354	0.320
from [19]		SLIM	0.370	0.495	0.401	0.347	0.428	0.379	–did not finish in [19]–		
		WMF	0.360	0.498	0.386	0.316	0.404	0.351	0.211	0.312	0.257
		CDAE	0.391	0.523	0.418	0.343	0.428	0.376	0.188	0.283	0.237
		MULT-VAE PR	0.395	0.537	0.426	0.351	0.444	0.386	0.266	0.364	0.316
		MULT-DAE	0.387	0.524	0.419	0.344	0.438	0.380	0.266	0.363	0.313