We thank the reviewers for their thoughtful feedback. We are encouraged that they found our approach interesting and elegant [R2, R3], and our algorithm novel and theoretically well developed [R1, R4]. We are pleased that R2 recognized the interest in our approach in the backprop community, and as a new perspective on a classical learning problem that is commonly solved using backprop.

Recap: The goal of our work is to establish a bridge between specific computational tasks and experimentally observed biological phenomena. To this end, we formalize the observation that pyramidal neurons combine instructive and sensory inputs in an optimization problem reflecting the simplest possible computational task. By solving this optimization problem in the online setting we derive a neural network with local learning rules. Interestingly, this simple linear model captures essential aspects of cortical microcircuits including the connectivity structure and the non-Hebbian nature of the learning rules in pyramidal neurons. Furthermore, this approach lets us interrogate which aspects of a detailed model are essential for performing this optimization task. To better highlight the goal of this work, we will change the title to "A simple normative network approximates local non-Hebbian learning in the cortex."

Below we answer some specific comments, but will incorporate all feedback in the final version.

[R1, R2, R4] Relationship to biology and to prior work. The price paid for the clarity of the normative approach is that it does not reproduce every known biological observation. Our results highlight which experimental observations about physiology are important for the circuit to implement this supervisory algorithm. We are grateful to the reviewers for pointing out relevant references [Kampa et al 2007, Urbanczik and Senn 2014, Gidon et al 2020]. We will add a "Related works" section where we will cite and discuss these, as well as [Sacramento et al 2018] in detail and point out the differences and the similarities. We will also amend the experimental evidence section of Sec. 4 to clarify what is speculation, what is hypothesis, and what is fact. We will also clearly delineate realistic and unrealistic features of the model including the lack of apical contribution to the output of the neuron, the slow and binary nature of Ca^{2+}plateaus in rodents, the lack of excitatory/inhibitory distinction among linear neurons.

[R1] The empirical evaluation is one of the weakest aspects of the paper. We reran the numerical experiments of Sec. 6 on five standard datasets of varying difficulties: MNIST, Fashion MNIST, CIFAR10, CIFAR100 and XRMB JW11 (a datset of acoustic and articulation measurements commonly used for testing algorithms for CCA and RRMSE) for ranks $k = 1, 2, 4, 8, 16$. Figure 1 shows a fraction of these results (one rank per dataset and only showing comparison with backprop for clarity and space constraints). In all cases, the performance of Bio-RRR, measured by the objective value in Eq. (3), is comparable to an ANN performing the same task but trained with backprop (as described in Sec. 5). We will include the full results in the supplementary materials.

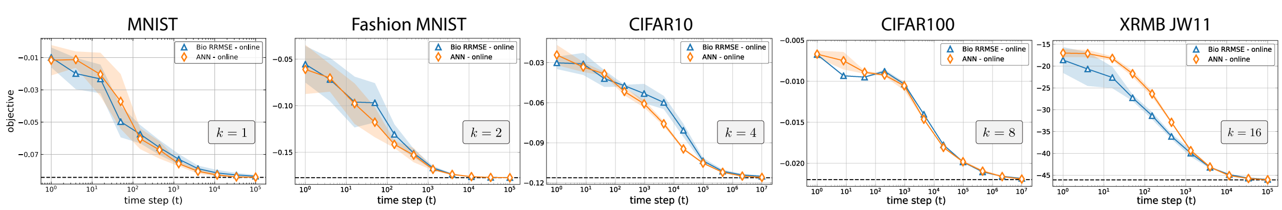

Figure 1: Comparison of the empirical performance of Bio-RRR vs. backprop on an artificial neural network. The dashed black line denote the value of the objective at its global minimum. The figures follow the same conventions as in Fig. 3 of the manuscript.

[R1] Ca^{2+}plateau potentials are generally known to affect distal dendrites. As we understand it, experimental observations regarding plasticity of distal dendrites range from those giving more weight to Ca^{2+}plateaus (in the hippocampus [Golding et al 2002]), to those giving more weight to the back-propagating action potentials (in the neocortex [Sjöström et al 2006]). In our learning rules for distal synapses Eq. (10), the relative significance of the terms $\mathbf{a}_t\mathbf{y}_t$ and $\mathbf{z}_t\mathbf{y}_t$ to learning is determined by the parameter s. The solvability of our normative model allows us to analytically explore the task performed by the circuit as we vary the distal learning rules. Explicitly, we can verify that the two extremes of $s = 0$ and $s = 1$ correspond to the statistical tasks RRMSE and CCA.

[R1, R3, R4] Linearity of model is unrealistic and computationally limiting. As mentioned above, the linearity of the model is a price paid for achieving a simple and analytically tractable model. Because of this, we are limited to considering models that perform dimensionality reduction. However, models of dimensionality reduction are known to be computationally effective and useful for learning. For example, CCA is intimately related to the information bottleneck problem and approaches which find features with the highest amount of mutual information [Chechik et al. 2005]. Previous experience shows that nonlinear extensions of such well understood linear models add useful features like dimension expansion, while retaining many aspects of network structure and of learning.