We thank the reviewers for their comments and constructive feedback. We are happy to see all three reviewers appreciate the novel perspective on BNN training we present. We acknowledge the need for better empirical support of our claims and present further ImageNet experiments below. We then address the concerns of each reviewer.

ImageNet Results. We train the BiReal-Net architecture on ImageNet from scratch using Bop. We train for 200 epochs with a batch size of 1024 and use standard preprocessing with random flip and resize but no further augmentation. We use Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$) for the full-precision weights. We linearly decrease three hyperparameters: τ from 10^{-7} to 10^{-8}, γ from $5 \cdot 10^{-4}$ to $5 \cdot 10^{-7}$ and the Adam learning rate from $2.5 \cdot 10^{-3}$ to $5 \cdot 10^{-6}$. After training we recompute the batch norm statistics over one epoch while keeping the weights fixed. We achieve **56.5%** top-1 and **79.5%** top-5 accuracy, which is comparable to the 56.4% top-1 and 77.2% top-5 accuracy originally reported. Note that the original paper relies on a multi-stage pretraining procedure whereas we train the network from scratch.

Response to Reviewer #2. Tuning of Bop vs. Latent-Weight Methods. Ease of use is an important consideration when selecting an optimizer so this is a valid concern. Bop is a novel optimizer that is disconnected from the vast body of experience that has developed around SGD-based methods. Inevitably, it will take time to develop intuition for the newly introduced hyperparameters. However, we are optimistic about the prospect of developing these intuitions as the introduced hyperparameters can be directly related to the training behavior of the network, as exemplified by Figure 1 in the paper. Furthermore, in latent-weight methods tuning alpha is only sufficient after numerous other hyperparameters have been fixed, including the initialization and clipping of the latent weights, the pseudo-gradient of the weight-binarization and the betas in Adam. Theorem 1 demonstrates the relations between these can be non-trivial and their effect can be non-intuitive. **The Role of Batch Normalization (BN).** The reviewer is correct in pointing out the BN is important for our method to work. Please note the BN is used in latent methods as well and that in BiReal-Net style architectures the BN is also important for the forward pass. **Adaptivitiy of Bop.** The reviewer shares our enthusiasm for the prospect of adaptive variants of Bop. We remark that in terms of training behavior, increasing the threshold is equivalent to lowering the gradients and so second order moments and adaptive thresholds are strongly related concepts. There are many open questions here, such as whether adaptivity should be implemented globally, per layer or per weight. We think it is valuable to have Bop, the simplest possible implementation of a BNN optimizer, as a reference point and leave the exploration of more sophisticated methods to future work.

Response to Reviewer #3. Relation to Existing Methods. The key novelty of Bop is the departure from the ghost network, a persistent feature of existing methods. We are confused by the statement of the reviewer that "the gradient for a particular weight is not defined by its current value" - this seems to ignore the fact that normally weights influence their gradients indirectly by influencing the forward pass, which is not the case for the magnitude of latent weights. Furthermore, although Bop shares the use an exponential moving average (EMA) of gradients with Momentum, here it is motivated by the need for signal consistency (line 135-139), rather than the curvature of a smooth loss landscape, as such a landscape does not exist here. Finally, it is true that the EMA in Bop plays an analogous role to the latent weights in existing methods. However, in previous methods, it is standard to use Momentum or Adam *on top of* latent weights, creating a stacking of effects that is difficult to reason about (as well as creating higher memory requirements during training). **Finetuning Viewpoint.** Although the reviewer agrees with our reinterpretation of the latent variables, he or she later states to perceive the problem as a finetuning issue and critiques our argument in 87-92 as misleading. This argument conclusively demonstrates a closer approximation to the latent weights is not necessarily better. The implications of this observation can be debated. However, we think the observation is relevant in this context and fits very well with the results of Merolla et al. (2016, line 85-86) and the implications of Theorem 1.

Response to Reviewer #4. Understanding Latent Weight Methods. It is interesting that using Momentum and Adam in latent-weight methods appears so important, even though we would argue it is essentially applying "momentum to inertia". Some elaboration on line 96-98 may be illuminating. As long as the sign of the latent weight does not change, spreading out gradients over time as in Momentum does not change the behavior of the network. Therefore, the value of using momentum must lie in the behavior of a latent weight that crosses zero. As the forward pass changes, the gradient may reverse. When using plain SGD, there is a risk of ill-behaved weights that rapidly jump back and forth, generating noisy behavior that harms training. We believe Momentum mitigates this behavior. Similarly, the threshold in Bop prevents rapid flipping of weights even if the gradient reverses. [See also response to Reviewer #3]

Note to Area Chairs. We have presented a novel perspective on BNN optimization and a novel optimizer. Reviewer #3 in particular seems to doubt the value of the presented ideas. Fundamentally, the question at stake here is whether the perspective of a finetuned latent network will prove limiting as a guiding principle for the development of training methods for BNNs. Any answer at this point can only be speculative. Nevertheless we hope the improved empirical results and additional clarifications will convince the reviewers the presented ideas form a promising alternative perspective. We only investigated two architectures, but Bop does show a minor improvement upon existing results in both cases. As we discussed in the original submission, we see Bop as first step along a promising path, and believe the competitive results that have been achieved are very encouraging.