We thank reviewers for their comments and suggestions. Please find below our point-to-point response.

R1, R2, R3; Contextualization, decompression and a concise summary of the present work: We agree with reviewers on splitting the sections into subsections to articulate and present the following crucial ideas and results more clearly. We will update this manuscript accordingly. In order to provide more detailed theoretical and experimental analyses and results, we have been preparing an extended version of the work (e.g. as a technical report/journal paper) with a toolbox/API supporting PyTorch, Tensorflow and MXNet. In this submitted version of the work, we introduce an overview of a unified mathematical and algorithmic framework which can be used to train DNNs employing different constraints on weights, with concrete generalization and convergence properties, and improving accuracy of baselines.

Definition of the proposed major problem: Generalization errors of DNNs are bounded by functions of various norms of weights of the DNNs. Our major goal is improvement of their generalization error by training DNNs according to these norms, under a unified algorithmic framework with precise generalization and convergence properties.
Proposed solutions for the major problem: We propose to (i) learn bounds of norms of weights, and (ii) optimize the weights with bounded norms for training of DNNs with better generalization error/accuracy in theory and practice.
Subproblems: The problem (i) is posed as estimation and learning of bounds of norms using geometry of spaces of feature representations and weights, and statistical properties of data and features, during training. However, weights with varying bounded norms reside on different manifolds, and their geometric properties (i.e. metrics and curvatures) change as bounds are updated while resolving (i) during training. Thus, the problem (ii) is posed as joint optimization of multiple fine-grained weights with different norms residing on products of the corresponding manifolds endowed with dynamically changing geometry with guarantee of convergence to local and global minima (Section 2 and 3).
Proposed solutions for the subproblems and results: To solve (i), we propose a two-stage re-normalization method by first bounding norms of weights to 1.0, and then learning the upper bounds of norms according to dimension of feature spaces and receptive fields determined by weights, and standard deviation of data and features. We also provide bounds and values of norms as functions of these geometric and statistical properties in Table 1 and 2. To solve (ii), we propose the FG-SGD in Section 4, and provide theoretical and experimental results in Sections 4, 5 and supp. mat.

R1, R2; Employment of shallow methods for optimization on product manifolds in DNNs using SGD, and related work: We consider [17,18] as two related works which optimize weights on particular *static product manifolds* to train shallow models. When we apply these methods for optimization on product of two or more *dynamic manifolds* in DNNs using SGD, Hessian of geodesic of the product manifold may not be bounded. In this case, we observe early divergence due to exploding or vanishing gradients. To this end, we first analyze relationship between geometry of product and component manifolds (i.e. metrics and geodesics) in Section 3. Then, we employ these results to bound gradients and Hessian on the product manifolds using those of component manifolds in Section 4, while developing the FG-SGD. Our proposed approach can be used to extend optimization methods proposed in recent related works, such as those proposed by Mishra et al., Sato et al., Huang et al., to apply their methods with dynamic product manifolds in DNNs. We will provide this discussion in the final version of the paper with the additional aforementioned related work.

R1; A sketch of proof idea, and equations (4), (5) and (6): In this work, we develop our algorithms by employing theoretical results using mathematical methods in their implementation. More precisely, the constraints and lemmas/theorems used to prove convergence theorems (Theorem 2 and Corollary 1) are realized and implemented in the algorithms. In order to introduce and explain this approach, we provided an overview of properties of geometry of manifolds used to prove convergence theorems in Section 3. Mathematical assumptions and steps of the proofs of convergence theorems are realized and implemented in the steps (Line 5, 6, 7) of the Algorithm 1. Remark 1 (Lemma 1) given in Section 3 was used to prove Theorem 1 which was used to prove convergence theorems. The equation (6) of Theorem 1 was used to compute functions given in the equations (3) and (5), and the equation (4) is a constraint used to compute learning rate in (3) at Line 6 of the Algorithm 1. These functions were used to prove the convergence theorems. A method used to compute a particular step size in (7) was proposed as a realization of Corollary 2. Therefore, we agree that the overall proof sketch was distributed in different sections of the paper. We consider providing an overview and a graphical sketch of proof of the theorems in Section 3 following suggestion of R1.

R3; Running time: Training time of DNNs for Euc, Sp, and Ob are similar. Training time for the St is affected by running time of matrix decomposition methods used by some projections, depending on numerical library and computer systems. Therefore, we provided a theoretical analysis of their computational complexity. When we apply approximation methods such as singular value bounding or power iteration for projections, then running times for the St approach to those of the other manifolds. For instance, for the experiments given in Table 3, the best running times (images/second) on a Tesla P100 are approximately: 200 (Euc, Sp, Ob), 180 (St), 185 (Sp+Ob+St), 190 (Sp+Ob+St+Euc).

R3; Results on NMT: Thank you for the notification, and we will fix the statement. We removed results obtained for NMT tasks to reduce complexity of presentation of the work and focus on image classification tasks. As an ablation study and a proof of concept, we obtained the following BLEU scores using a transformer network (Vasvani et al., NIPS'17) for English to German translation on the WMT newstest2014: Baseline (27.1), Sp+Ob+St+Euc (28.3).