We first and foremost thank the reviewers for their valuable time and feedback. We will fix all small comments and typos the reviewers note.

Reviewer $\#1$ asks how our results relate to adaptive gradient methods. As the reviewer notes, adaptive gradient methods construct a sequence of linear pre-conditioners to be applied to the stochastic sub-gradient before the update. Our results certify that there exists an optimal linear pre-conditioner for quadratically convex constraint sets. As such, adaptive gradient methods can be minimax (rate) optimal. In online algorithms, the common practice [4, 5, 6, 7, 2] is to measure regret with respect to the "best" post-hoc regularizer (i.e. preconditioner) and fixed predictor; in this context, the regret achieved by AdaGrad is a factor $\sqrt{2}$ away from the regret the best post-hoc linear pre-conditioner achieves over rectangular domains (and may be $\sqrt{d}$ better than standard gradient methods). Our results thus guarantee the minimax optimality of AdaGrad in certain settings. We are perhaps a bit imprecise in that we use "best linear preconditioner" as a shorthand for what adaptive algorithms may achieve; we will make this more precise. The extent to which specific adaptive algorithms find the (optimal) linear pre-conditioner for specific constraint sets remains open.

Reviewer $\#1$ correctly notes that the quadratic convexity of the constraint set is critical via Proposition 4. In the case that Θ is not quadratically convex, one must replace Θ by its quadratic hull when swapping the infimum and supremum, resulting in a (potentially) much larger set (e.g. $\mathsf{QHull}(\mathbf{B}_1(0,1)) = \mathbf{B}_2(0,1)$). We will emphasize this.

Reviewer $\#1$ asks about the origin and meaning of Corollary 3. It follows from Corollary 2 by lower bounding the inequality. To illustrate the corollary, we can observe that when γ is an ℓ_q norm with $q \in [1,2]$, the lower and upper bounds match up to $\sqrt{\log d}$. After the submission of this paper, we derived more precise bounds in the case where the γ-ball is not quadratically convex: we obtained matching lower and upper bounds when the gradients live in any weighted ℓ_q ball i.e. $\gamma(g) = (\sum_{j \le d} a_j |g_j|^q)^{1/q}$ for $a_j > 0$ and $q \in [1, \infty]$. We will include these new results.

Reviewer $\#1$ asks for definitional clarifications for minimax risk and regret. While F_P is a deterministic function, in the definition of the minimax risk, it is applied to $\widehat{\theta}_n(X_1^n)$ which is a (random) estimator based on a sample $X_1^n \sim P$, necessitating an expectation. In the definition of $\mathfrak{M}_n^{\mathsf{S}}$ and $\mathfrak{M}_n^{\mathsf{R}}$, the supremum over the sample space takes place before the infimum as the infimum ranges over all (measurable) functions $\mathcal{X}^n \to \Theta$. This definition accords with the literature on lower bounds in convex optimization, where the supremum is over stochastic oracles [1]. Reviewer $\#2$ asks for a clarification on the definition of X_1^n and x_1^n. The former corresponds to a collection of n i.i.d. random variables $(X_1, \ldots, X_n)$. The latter denotes n (fixed) vectors $(x_1, \ldots, x_n) \in \mathcal{X}^n$. We will clarify all of these definitions.

Reviewer $\#2$ asks for applicability for the non-convex setting. Empirically, even in non-convex settings, AdaGrad tends to outperform vanilla gradient methods when data is sparse (e.g. [3, 8]). Our mathematical results probably do not translate as is beyond convexity. However, deriving similar results in the case, for example, of finding stationary points of non-convex functions is a natural extension and a very interesting future direction.

Reviewer $\#3$ asks for concrete examples where the geometry of the constraint sets matters. In addition to the two deep learning examples of the above paragraph, this work is, for example, applicable in the broad case of linear models. In this setting, the constraint set corresponds to the set of classifiers of interest, and the geometry of the gradients corresponds to the geometry of the features (or covariates). For example, in NLP applications, bag-of-word features are very sparse by nature, so we seek a dense classifier (i.e. a weight for every word). In the terms of our paper, this means that Θ is a weighted ℓ_∞ ball, γ is a weighted ℓ_1 ball, and our theory suggests adaptive scaling is important. (For empirical results, see, e.g. [4].)

References

[1] A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

[2] A. Cutcosky and T. Sarlos. Matrix-free preconditioning in online learning. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[3] J. Dean, G. S. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le, M. Z. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and A. Y. Ng. Large scale distributed deep networks. In *Advances in Neural Information Processing Systems 26*, 2012.

[4] J. C. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[5] B. McMahan and M. Streeter. Adaptive bound optimization for online convex optimization. In *Proceedings of the Twenty Third Annual Conference on Computational Learning Theory*, 2010.

[6] F. Orabona and K. Crammer. New adaptive algorithms for online classification. In *Advances in Neural Information Processing Systems 23*, 2010.

[7] F. Orabona, K. Crammer, and N. Cesa-Bianchi. A generalized online mirror descent with applications to classification and regression. *Machine Learning*, 99(3):411–435, 2015.

[8] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of Empirical Methods for Natural Language Processing*, 2014.