AN ONLINE LEARNING APPROACH TO INTERPOLATION AND EXTRAPOLATION IN DOMAIN GENERALIZATION

Elan Rosenfeld, Pradeep Ravikumar, Andrej Risteski
Machine Learning Department
Carnegie Mellon University
elan@cmu.edu, pradeepr@cs.cmu.edu, aristesk@andrew.cmu.edu

ABSTRACT

A popular assumption for out-of-distribution generalization is that the training data comprises sub-datasets, each drawn from a distinct distribution; the goal is then to "interpolate" these distributions and "extrapolate" beyond them—this objective is broadly known as domain generalization. A common belief is that ERM can interpolate but not extrapolate and that the latter is considerably more difficult, but these claims are vague and lack formal justification. In this work, we recast generalization over sub-groups as an online game between a player minimizing risk and an adversary presenting new test distributions. Under an existing notion of inter- and extrapolation based on reweighting of sub-group likelihoods, we rigorously demonstrate that extrapolation is computationally much harder than interpolation, though their statistical complexity is not significantly different. Furthermore, we show that ERM—or a noisy variant—is *provably minimax-optimal* for both tasks. Our framework presents a new avenue for the formal analysis of domain generalization algorithms which may be of independent interest.

1 INTRODUCTION

Modern machine learning algorithms excel when the training and test distributions match but often fail under even moderate distribution shift (Beery et al., 2018); learning a predictor which generalizes to distributions which differ from the training data—known as out-of-distribution (OOD) generalization—is therefore an important task. One popular assumption is that the training data is comprised of a collection of "environments" (Blanchard et al., 2011; Muandet et al., 2013; Peters et al., 2016) or "groups" (Sagawa et al., 2020), each representing a distinct distribution, where the group identity of each sample is known.

The question of domain generalization thus becomes one of characterizing how future test distributions will relate to these groups and how to ensure good performance on these distributions. Many works assume that the test distribution will be a weighting of the observed group likelihoods (also known as subpopulation shift)—the goal is then to minimize risk over the worst-case combination (Sagawa et al., 2020; Albuquerque et al., 2020; Krueger et al., 2020). More generally, these approaches can be thought of as attempting to learn predictors which are able to "interpolate" the training *distributions* and also "extrapolate" beyond them. A common belief is that Empirical Risk Minimization (ERM) excels at interpolation but not extrapolation; it is also generally held as folklore that extrapolation is a much harder task, which is why generalization is so difficult. However, these claims are somewhat hazy and it remains unclear which of these beliefs, if any, can be formally justified.

Even when evaluating algorithms empirically, existing works often miss the mark for robustly comparing different OOD predictors. For example, Gulrajani & Lopez-Paz (2021) point out that many recent works intentionally evaluate on a single train/test environment split with an unreasonably difficult distribution shift, portraying ERM unfairly. When averaging performance over all possible environment splits, they find that no algorithm outperforms ERM. An intentionally difficult test distribution may be appropriate for quantifying how an algorithm will perform in the worst possible case, but this rarely reflects a predictor's quality in the real world, where the test environments

are usually *not* chosen adversarially. Thus the crucial distinction is that **existing frameworks are minimax because they demand good performance of an algorithm even in the worst case, not because we actually expect the test environments to be chosen adversarially**. This subtle point is critical to our motivation for developing a more realistic metric of success in domain generalization; we discuss it in much greater detail in Appendix C. Evidently, there is a need for a more robust measure of OOD generalization, one which adequately captures the purpose of such algorithms and allows for formal comparison of their computational and statistical properties.

In this work, taking inspiration from the literature of online convex optimization (OCO) (Hazan, 2016), we ask what can be achieved in a game where the learner is allowed to repeatedly refine their predictor upon observing new environments. Our multi-round game directly generalizes existing work on domain generalization, providing new insights into the quantifiable effects of observing different environments as a function of both their number and their geometric diversity. Further, this novel perspective allows for a theoretical analysis of the computational and statistical complexity of interpolation vs. extrapolation, formalizing and verifying the answers to several outstanding questions which until now have only been stated intuitively.

1.1 A Formal Notion of Interpolation Versus Extrapolation for Distributions

In the context of domain generalization, the terms "interpolation" and "extrapolation" do not have an agreed-upon definition. Given a collection of environments, there are many possible ways to consider interpolating them. In this work, we limit our analysis to the notion of likelihood reweighting which has been used previously in several works (Sagawa et al., 2020; Albuquerque et al., 2020; Krueger et al., 2020). We frame the interpolation of a set of environments as all convex combinations (i.e., mixtures) of their distributions. Formally, suppose we have a set of E environments (i.e. domains) $\mathcal{E} = \{e_i\}_{i=1}^{E}$, each of which indexes a probability distribution p^e. Then an interpolation of these distributions is any distribution which can be written $p^\lambda := \sum_{e\in\mathcal{E}} \lambda_e p^e$, where $\lambda \in \Delta_E$ is a vector of convex coefficients (Δ_E is the $(E-1)$-simplex). This is a fairly natural definition, as the space of interpolations is defined as all points in the convex hull of the environments $\mathcal{E}$ in distribution-space.

It is not immediately clear how to extend this concept to include extrapolation. Krueger et al. (2020) suggest allowing for combinations in which the coefficients are affine but may be slightly negative, where the minimum coefficient is given as a hyperparameter α: $\sum_{e\in\mathcal{E}} \lambda_e = 1, \quad \lambda_e \geq -\alpha. \forall e \in \mathcal{E}$. The resulting function is not guaranteed to be a probability distribution, as it could result in negative measure being assigned to some points—they instead frame it as reweighting of the environment *risks*, and we find that such a reframing is helpful for our analysis (see Lemma 1). We refer to such combinations as "bounded affine" combinations, and we study them in Section 3.1.

2 The Online Domain Generalization Game

We consider recasting the task of domain generalization as a continuous game of online learning in which the player is presented with new test environments and must refine their predictor at each round. We would expect that any good learning algorithm will suffer less per distribution as we observe more of them—that is, the *per-round regret* should decrease over time. In particular, we'd like to prove a rate at which our regret goes down as a function of the number of distributions we've observed. Our game allows for a *worst-case* analysis of the *average-case* performance; in Section C we expound upon this idea in detail. The full game can be found in Appendix A. Note that we describe a specific instance where the adversary is limited to group mixtures as described in Section 1.1; **the general game allows for any formally specified action space for the adversary** and we anticipate this will enable future analyses involving rich classes of distribution shift threat models such as f-divergence or $\mathcal{H}$-divergence balls (Bagnell, 2005; Ben-David et al., 2007).

Game Setup Before the game begins, we define a family of predictors parameterized by β lying in a convex set B. For some observation space $\mathcal{X}$ and label space $\mathcal{Y}$, nature provides a fixed loss function $\ell : B \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$, strongly convex in the first argument, as well as a set of E environments $\mathcal{E} = \{e_i\}_{i=1}^{E}$, each of which indexes a distribution p^e over $\mathcal{X} \times \mathcal{Y}$. We assume that B is large enough such that for any $\lambda \in \Delta_E$, the parameter which minimizes risk on p^λ lies in B. We further assume that for any $\beta \in B$ and all $e \in \mathcal{E}$, the expected loss of β under p^e is finite. The game proceeds

as follows: on round t, the player chooses parameters $\hat{\beta}_t \in B$. Next, the adversary chooses a set of coefficients $\lambda_t := \{\lambda_{t,e}\}_{e \in \mathcal{E}}$, which defines the distribution p^{λ_t} as the weighted combination of the likelihoods of environments in $\mathcal{E}$ with coefficients λ_t, as above. At the end of the round, the player suffers loss $f_t(\hat{\beta}_t) = \mathcal{R}^{\lambda_t}(\hat{\beta}_t)$, defined as the risk of the predictor parameterized by $\hat{\beta}_t$ on the adversary's chosen distribution: $\mathcal{R}^{\lambda_t}(\beta) := \mathbb{E}_{(x,y)\sim p^{\lambda_t}}[\ell(\beta, (x, y))]$ (we write $f_e = \mathcal{R}^e$ for the analogous risk on distribution p^e). For clarity, when using any of the above notation we will drop the subscript t when it is not necessary.

We define our objective as minimizing *regret* with respect to the best fixed predictor in hindsight after T rounds. That is, we hope to minimize $\sum_{t=1}^{T} f_t(\hat{\beta}_t) - \min_{\beta \in B} \sum_{t=1}^{T} f_t(\beta)$. This notion of regret straightforwardly generalizes previous work on domain generalization. By allowing $T \to \infty$, we have a more robust measure of success: each time we are presented with a new environment, we update our predictor to improve our average performance. Crucially, this modification allows us to prove a rate at which our regret decreases as a function of the number of environments observed. We again refer the reader to Appendix C for a detailed discussion of the advantages of this objective.

3 Formal Results

Similar to Abernethy et al. (2008), we evaluate the performance of an algorithm by defining the *value* of the game after T timesteps as the player's regret under optimal play by both the player and the adversary: $V_T := \min_{\hat{\beta}_1} \max_{\lambda_1} \dots \min_{\hat{\beta}_T} \max_{\lambda_T} (\sum_{t=1}^{T} f_t(\hat{\beta}_t) - \min_{\beta \in B} \sum_{t=1}^{T} f_t(\beta))$. For a fixed T, this allows us to formalize minimax bounds on the regret. In the more traditional literature on OCO, the adversary is allowed to play losses f_t from a much more general class, such as all strongly convex functions. In this setting, the value of the game in any given round t is known to be exactly $V_t = \sum_{s=1}^{t} \frac{G_s^2}{2s\sigma_{\min}}$, where G_s is the Lipshitz constant of f_s at the parameter chosen by the player. This means the minimax-optimal rate for regret is $\Theta(\log t)$ (Hazan et al., 2007; Bartlett et al., 2007).

In our interpolation game, the adversary is severely restricted, being allowed to play only convex combinations of the risks of each of the E given distributions. We might expect that such a restriction, especially when known to the player, would allow for a faster convergence to zero regret, even if the strategy which attains it is intractable. Our first result demonstrates that this is not the case.

Theorem 1. *Suppose $\sigma_{\max} \geq \sigma_{\min} > 0$ such that $\forall e \in \mathcal{E}$, $\sigma_{\min} I \preceq \nabla^2 f_e \preceq \sigma_{\max} I$. Define g as the minimum gradient norm that is guaranteed to be forceable by the adversary: $g := \min_{\beta \in B} \max_{\lambda \in \Delta_E} \|\nabla f(\beta)\|_2$. Then for all $t \in \mathbb{N}$ it holds that $V_t > \frac{g^2 \sigma_{\min}}{16 \sigma_{\max}^2} \log t$.*

Observe that the minimum forceable gradient norm g encodes a sort of "radius" of the convex hull of loss gradients. The bound does not directly depend on the *number* of environments E; rather it scales quadratically with the size of this region, which appropriately captures the intuition that a smaller regret should be achievable for a collection of sub-distributions that are very similar to one another.

Theorem 1 is somewhat surprising; restricting the adversary to playing within the convex hull of a limited number of functions might be expected to affect the asymptotic statistical complexity. Even more interesting, this rate can be achieved with a very simple algorithm known as Follow-The-Leader (FTL), which just plays the minimizer of the sum of all previously seen functions (Hazan et al., 2007). In our game, this means playing the predictor which minimizes risk over all environments seen so far—*observe that this strategy is precisely ERM!* In other words, ERM is provably minimax-optimal for interpolation. As the adversary's playable region is a strict subset of all strongly convex functions, it is immediate that the regret suffered by playing ERM is upper bounded as $\sum_{s=1}^{t} \frac{G_s^2}{2s\sigma_{\min}} \in O(\log t)$. Theorem 1 also has meaningful implications for the single-round setting. A simple corollary gives the first tight bound on the generalization rate as a function of the number of environments observed.

Corollary 1. *Suppose we've seen t environments so far. Then the additional regret suffered due to one additional round is $\Omega\left(\frac{1}{t}\right)$. This lower bound is attained by ERM.*

3.1 Bounded Affine Combinations

One could argue that allowing the adversary only convex combinations of environments is perhaps too good to hope for. As we've seen, ERM is optimal for such a setting, but it has been widely observed

that ERM fails under minor distribution shift. We might expect that future environments would fall outside of this hull—if combinations within the hull represent a formal notion of "interpolating" the training distributions, then it seems our goal instead should be to "extrapolate" beyond them. As discussed in Section 1.1, Albuquerque et al. (2020); Krueger et al. (2020) consider an adversary playing bounded affine combinations of the environments; this conceptualization of extrapolation seems a natural extension. Clearly, this game is no easier for the player—our results demonstrate that it is in fact *exponentially* harder. We consider a relaxed version of our game with an "oblivious" adversary: this adversary selects the entire sequence of loss functions before the game starts (our lower bounds hold despite this relaxation). For general Lipschitz functions, no deterministic strategy can guarantee sublinear regret, and attaining sublinear regret with a randomized strategy is NP-hard. Further, there is an information-theoretic regret lower bound of $\Omega(\sqrt{T})$ which was recently shown to be achievable with Follow-The-Perturbed-Leader (FTPL), assuming access to an oracle for approximately minimizing a non-convex function (Suggala & Netrapalli, 2020). As in the previous subsection, we extend these results to the task of domain generalization—that is, we demonstrate that despite the (seemingly restrictive) requirement that the adversary play bounded affine combinations of strongly convex losses that are fully known to the player, *the game remains equally hard*.

Theorem 2. *No deterministic algorithm can guarantee sublinear regret against bounded affine combinations of a finite set of strongly convex losses.*

Thus we find that just as in the general non-convex case, a randomized algorithm is necessary. We might hope that the computational requirements of achieving sublinear regret would be lessened—perhaps there would be no need for an optimization oracle. However, Theorem 3 proves otherwise.

Theorem 3. *Even with a randomized algorithm against an oblivious adversary playing bounded affine combinations, achieving sublinear regret is NP-hard.*

Computationally, our game of extrapolation is just as difficult as achieving sublinear regret on arbitrary Lipschitz functions. These results present, for the first time, proof of *an exponential computational complexity gap between interpolation and extrapolation in the domain generalization setting*, formally verifying existing intuition. We now consider the statistical complexity of regret minimization under bounded affine combinations. Recall that for convex combinations, Theorem 1 shows a minimax regret lower bound of $\Omega(\log t)$ which can be achieved with ERM. We again note that for an adversary playing arbitrary Lipschitz functions, Suggala & Netrapalli (2020) demonstrate that FTPL can achieve the minimax lower bound of $\Omega(\sqrt{T})$ with the help of an oracle. The FTPL strategy plays the minimizer of the sum of the observed environments plus a noise term—in our game, then, *FTPL is just a noisy variant of ERM*. The natural next question is if playing against an oblivious adversary limited to bounded affine combinations will allow us to surpass this lower bound. That is, can we outperform ERM in this setting *at all?* Our final result answers this question in the negative. Thus we find once more that ERM remains statistically minimax-optimal, even for extrapolation.

Theorem 4. *Against an oblivious adversary playing bounded affine combinations, the achievable regret is lower bounded as* $\Omega(\sqrt{T})$.

4 Conclusion and Future Directions

This work presents the first formal demonstration of an exponential computational gap between interpolation and extrapolation in domain generalization, a claim which has until now only been given vague intuitive justification. Perhaps more importantly, we've shown that ERM remains statistically minimax-optimal for both tasks, raising questions about the effectiveness of recently proposed alternatives. Due to space constraints, we discuss additional related work in Appendix B.

We see two immediate directions for further research. First, the proposed game serves as a standalone framework for the theoretical analysis of learning algorithms. As discussed in Appendix C, considering regret in the online setting provides a more robust and meaningful signal of an algorithm's expected performance. We hope this new perspective will encourage future work to provide formal OOD generalization guarantees for their proposed methods. Second, there remains significant flexibility in how we define "interpolation" and "extrapolation" with respect to training environments; we consider one specific notion in this work, but it seems likely that different restrictions on the adversary could allow for stronger generalization guarantees. Furthermore, our analysis reveals that the *geometry of the environmental loss functions* is a critical element for generalization. This suggests additional improvements can be achieved with careful representation learning.

REFERENCES

Abernethy, J., Bartlett, P., Rakhlin, A., and Tewari, A. Optimal strategies and minimax lower bounds for online convex games. In *Technical Report No. UCB/EECS-2008-19*, 2008.

Albuquerque, I., Monteiro, J., Darvishi, M., Falk, T. H., and Mitliagkas, I. Generalizing to unseen domains via distribution matching. *arXiv preprint arXiv:1911.00804*, 2020.

Alquier, P., Mai, T. T., and Pontil, M. Regret Bounds for Lifelong Learning. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 261–269, 2017. URL http://proceedings.mlr.press/v54/alquier17a.html.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Bagnell, J. A. Robust supervised learning. In *Proceedings of the 20th national conference on Artificial intelligence-Volume 2*, pp. 714–719, 2005.

Balcan, M.-F., Blum, A., and Vempala, S. Efficient representations for lifelong learning and autoencoding. In *Proceedings of The 28th Conference on Learning Theory*, pp. 191–210, 2015. URL http://proceedings.mlr.press/v40/Balcan15.html.

Bartlett, P. L., Hazan, E., and Rakhlin, A. Adaptive online gradient descent. In *Advances in Neural Information Processing Systems*, pp. 65–72, 2007.

Beery, S., Van Horn, G., and Perona, P. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 456–473, 2018.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2007. URL https://proceedings.neurips.cc/paper/2006/file/b1b0432ceafb0ce714426e9114852ac7-Paper.pdf.

Blanchard, G., Lee, G., and Scott, C. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in Neural Information Processing Systems*, volume 24, pp. 2178–2186. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper/2011/file/b571ecea16a9824023ee1af16897a582-Paper.pdf.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Christiansen, R., Pfister, N., Jakobsen, M. E., Gnecco, N., and Peters, J. A causal framework for distribution generalization. *arXiv preprint arXiv:2006.07433*, 2020.

De Klerk, E. The complexity of optimizing over a simplex, hypercube or sphere: a short survey. *Central European Journal of Operations Research*, 16(2):111–125, 2008.

Didelez, V., Dawid, A. P., and Geneletti, S. Direct and indirect effects of sequential treatments. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pp. 138–146, 2006.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Håstad, J. Clique is hard to approximate within 1- ε. *Acta Mathematica*, 182(1):105–142, 1999.

Hazan, E. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2 (3-4):157–325, 2016.

Hazan, E., Agarwal, A., and Kale, S. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007. URL https://doi.org/10.1007/s10994-007-5016-8.

Heinze-Deml, C. and Meinshausen, N. Conditional variance penalties and domain shift robustness. *Machine Learning*, 2020.

Heinze-Deml, C., Peters, J., and Meinshausen, N. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018.

Hu, S., Zhang, K., Chen, Z., and Chan, L. Domain generalization via multidomain discriminant analysis. In *Uncertainty in Artificial Intelligence*, pp. 292–302. PMLR, 2020.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Priol, R. L., and Courville, A. Out-of-distribution generalization via risk extrapolation (rex). *arXiv preprint arXiv:2003.00688*, 2020.

Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., and Tao, D. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 624–639, 2018.

Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Betteridge, J., Carlson, A., Mishra, B. D., Gardner, M., Kisiel, B., Krishnamurthy, J., Lao, N., Mazaitis, K., Mohamed, T., Nakashole, N., Platanios, E., Ritter, A., Samadi, M., Settles, B., Wang, R., Wijaya, D., Gupta, A., Chen, X., Saparov, A., Greaves, M., and Welling, J. Never-ending learning. In *AAAI Conference on Artificial Intelligence*, 2015. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/10049.

Motzkin, T. S. and Straus, E. G. Maxima for graphs and a new proof of a theorem of turán. *Canadian Journal of Mathematics*, 17:533–540, 1965. doi: 10.4153/CJM-1965-053-6.

Muandet, K., Balduzzi, D., and Schölkopf, B. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18. PMLR, 2013.

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society*, 2016.

Rosenfeld, E., Ravikumar, P. K., and Risteski, A. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=BbNIbVPJ-42.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Suggala, A. S. and Netrapalli, P. Online non-convex learning: Following the perturbed leader is optimal. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, pp. 845–861, 2020. URL http://proceedings.mlr.press/v117/suggala20a.html.

Thrun, S. Lifelong learning algorithms. In *Learning to learn*, pp. 181–209. Springer, 1998.

Tian, J. and Pearl, J. Causal discovery from changes. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pp. 512–521, 2001.

Zhao, H., Combes, R. T. D., Zhang, K., and Gordon, G. On learning invariant representations for domain adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7523–7532. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/zhao19a.html.

A Domain Generalization Game

Domain Generalization Game (group interpolation)

Input: Convex parameter space B, distributions $\{p^e\}_{e\in\mathcal{E}}$ over $\mathcal{X}\times\mathcal{Y}$, strongly convex loss $\ell : B \times (\mathcal{X}\times\mathcal{Y})$

for $t = 1 \dots T$ **do**

1. Player chooses parameters $\hat{\beta}_t \in B$.
2. Adversary chooses convex coefficients

$$\lambda_t = \{\lambda_{t,e}\}_{e\in\mathcal{E}}, \quad \sum_{e\in\mathcal{E}} \lambda_{t,e} = 1,\ \lambda_{t,e} \geq 0\ \forall e \in \mathcal{E}.$$

3. Define $f_t(\beta) := \mathbb{E}_{(x,y)\sim p^{\lambda_t}}[\ell(\beta,(x,y))]$, where $p^{\lambda_t} = \sum_{e\in\mathcal{E}} \lambda_{t,e} p^e$.

end for

Player suffers regret

$$R_T = \sum_{t=1}^{T} f_t(\hat{\beta}_t) - \min_{\beta\in B} \sum_{t=1}^{T} f_t(\beta).$$

B Related Work

Many works provide formal guarantees for OOD generalization by assuming invariances in the causal structure of the data: a series of interventions is assumed to result in separate fixed environments (Peters et al., 2016; Heinze-Deml et al., 2018; Heinze-Deml & Meinshausen, 2020; Christiansen et al., 2020) or distribution shift over time (Tian & Pearl, 2001; Didelez et al., 2006), and the test distribution will likewise represent such an intervention. Under various conditions it is then possible to identify which features have invariant relationships with the target variable. This formalization is typically minimax, in the sense that recovery of these features ensures an invariant classifier with reasonable performance despite arbitrary future interventions on the other variables. However, they typically assume full or partial observation of the covariates, and therefore they do not apply to the setting where the data is a complex function of unobserved latent variables.

Works which eschew a direct causal formalization often still depend upon the intuition of "invariant features" within the context of causality. Arjovsky et al. (2019) propose IRM to learn features for which the optimal predictor is invariant across environments, and Krueger et al. (2020) build upon this idea with their proposed R-Ex which requires the *risks* to be constant. Rosenfeld et al. (2021) formally demonstrated that these and similar objectives often fail to outperform ERM even in simple settings.

Other works attempt to learn "domain invariant" representations (Muandet et al., 2013; Ganin et al., 2016; Li et al., 2018; Hu et al., 2020) for either domain generalization or *unsupervised domain adaptation*, the latter of which assumes access to unlabeled data from the test domain. Zhao et al. (2019) showed that such invariance is often not sufficient for domain adaptation, and recently Gulrajani & Lopez-Paz (2021) gave empirical evidence that many of these algorithms fare no better than ERM in a more realistic testing environment. Rather than considering invariances, Sagawa et al. (2020) frame generalization as a question of performing well on individual sub-populations and give an efficient algorithm for solving the single-round minimax game we describe in Section C. Albuquerque et al. (2020) also consider the convex hull of domains, providing a generalization bound via $\mathcal{H}$-divergence. Unfortunately, this bound scales linearly with both the maximum discrepancy between pairs of training distributions and between the test distribution and training environment hull—it also requires finding in the first place a classifier which does well on all convex combinations of domains.

This work relates the nascent study of domain generalization theory to prior work on online and lifelong learning (Thrun, 1998; Mitchell et al., 2015; Hazan, 2016), for which there already exist provable regret bounds and efficiency guarantees (Balcan et al., 2015; Alquier et al., 2017). The main difference is that those works present new algorithms and give upper bounds, while this work proves lower bounds which match rates already known to be achievable for more general classes

of losses (Hazan et al., 2007; Abernethy et al., 2008; Suggala & Netrapalli, 2020), implying that existing algorithms are already optimal.

C THE BENEFITS OF ONLINE REGRET VS. SINGLE-ROUND LOSS

Our focus on the online setting as opposed to a single round is important; it will be instructive to carefully consider the benefits to such an analysis.

Significance of a baseline Recall the expression for regret introduced in the main body:

$$\sum_{t=1}^{T} f_t(\hat{\beta}_t) - \min_{\beta \in B} \sum_{i=1}^{T} f_t(\beta).$$

The second term in this equation is crucial; the comparison to the best *fixed* parameter prevents the adversary from forcing constant regret at each round and reflects the idea that we hope to eventually perform favorably compared to a single predictor which does reasonably well on all environments. Without this baseline, the player's objective would be to simply minimize the sum of the risks on all environments: $\sum_{t=1}^{T} f_t(\hat{\beta}_t)$. In the adversarial setting,[1] the game therefore reduces to repeated, independent instances of the single-round version; clearly, the best we can do to minimize loss each round is to play the minimax-optimal parameters $\beta^* := \arg\min_{\beta \in B} \max_{\lambda \in \Delta_E} \mathcal{R}^{\lambda}(\beta)$ (Sagawa et al., 2020). In response, the adversary would always choose $\lambda^* := \arg\max_{\lambda \in \Delta_E} \mathcal{R}^{\lambda}(\beta^*)$. This game is uninteresting beyond the first round and does not adequately capture our algorithm's performance in any real-world setting where the environments are *not* chosen adversarially. As mentioned in the introduction, the key observation here is that the single-round minimax framework is used to guarantee good performance even in the worst-case scenario, but we do not actually expect future test environments to be chosen in this way.

As a simple example, if we were to repeatedly play β^* and repeatedly face the test distribution p^{λ^*}, we should consider it more likely that this is representative of future test environments than that nature is actively trying to give us the largest possible loss. Consequently we should switch strategies and play $\arg\min_{\beta \in B} \mathcal{R}^{\lambda^*}(\beta)$. Thus, existing frameworks overemphasize minimax performance in individual rounds—even though in reality, distribution shift is rarely adversarial—while ignoring possible improvements over time. In contrast, our longitudinal analysis allows for an algorithm to occasionally suffer preventable loss in any given turn, so long as the per-turn regret is guaranteed to decrease over time.

Implications of sublinear regret For any sequence of environments, there will be some parameter $\tilde{\beta}$ which *would have* achieved the least possible cumulative loss. Sublinear regret implies that as $T \to \infty$ we will eventually recover the per-round loss of $\tilde{\beta}$, but without committing beforehand and with *no prior knowledge* of the test environment sequence. Thus in the limit we are guaranteeing the lowest possible average loss against a fixed sequence of environments—at the same time, our analysis is minimax so as to guarantee our regret bound holds even against the worst such sequence.

Further, sublinear regret is a very powerful guarantee when the environments are stochastic, as might be expected in any real-world setting. For any prior over environment distributions $\pi(p^e)$, it is easy to see that sublinear regret implies convergence to the performance of the parameter which minimizes loss over the marginal distribution:

$$\arg\min_{\beta \in B} \int_{\mathcal{P}} \pi(p^e)\, \mathbb{E}_{p^e}[\ell(\beta, (x, y))]\, dp^e,$$

where $\mathcal{P}$ is the set of all distributions over $\mathcal{X} \times \mathcal{Y}$. This is because as $T \to \infty$, the π-weighted average of the sum of losses will converge to the loss on the marginal distribution—the baseline will then be whatever parameter minimizes this loss. Observe that this is strictly stronger than the guarantee of ERM, which ensures the same result only in the limit; sublinear regret implies that *for every* T, our regret with respect to the best predictor so far is bounded as $o(T)$. Thus if by chance the distributions

[1] By this we mean the setting where the next environment is always the one which maximizes risk for the parameter chosen by the player.

we've seen are not representative of the prior π (an oft-stated motivation for OOD generalization), we are still ensuring convergence to the loss of the optimal fixed predictor in hindsight, whatever it may be. In particular, if the sequence of environments is so unfavorable to ERM that the optimal predictor in hindsight is an invariant predictor (Peters et al., 2016; Arjovsky et al., 2019; Rosenfeld et al., 2021), which ignores meaningful signal to ensure broad generalization, sublinear regret guarantees that our algorithm's loss converges to this invariant predictor's loss.

D Proofs of the Main Results

D.1 Proof of Theorem 1

Theorem 1. *Suppose* $\sigma_{\max} \geq \sigma_{\min} > 0$ *such that* $\forall e \in \mathcal{E}$, $\sigma_{\min} I \preceq \nabla^2 f_e \preceq \sigma_{\max} I$. *Define* g *as the minimum gradient norm that is guaranteed to be forceable by the adversary:* $g := \min_{\beta \in B} \max_{\lambda \in \Delta_E} \|\nabla f(\beta)\|_2$. *Then for all* $t \in \mathbb{N}$ *it holds that* $V_t > \frac{g^2 \sigma_{\min}}{16 \sigma_{\max}^2} \log t$.

Proof. Define $F_t(z) = \sum_{s=1}^{t} f_s(z)$; since each f is convex, this sum is convex as well. Let β^*_{t-1} be the minimizer of F_{t-1} (by Lemma 2, this will lie in B), and let $z \in B$ be arbitrary. Finally, note that $\nabla^2 F_t \preceq t\sigma_{\max} I$. Then we have the following Taylor expansion:

$$\begin{aligned}
F_t(z) &= F_{t-1}(z) + f_t(z) \\
&= F_{t-1}(\beta^*_{t-1} + (z - \beta^*_{t-1})) + f_t(z) \\
&\leq F_{t-1}(\beta^*_{t-1}) + \nabla F_{t-1}(\beta^*_{t-1})^T (z - \beta^*_{t-1}) + \frac{(t-1)\sigma_{\max}}{2} \|z - \beta^*_{t-1}\|_2^2 + f_t(z) \\
&= F_{t-1}(\beta^*_{t-1}) + \frac{(t-1)\sigma_{\max}}{2} \|z - \beta^*_{t-1}\|_2^2 + f_t(z),
\end{aligned}$$

where we have used the fact that $\nabla F_{t-1}(\beta^*_{t-1}) = 0$ by definition. Thus,

$$\sum_{s=1}^{t} f_s(\hat{\beta}_s) - F_t(z) \geq \left(\sum_{s=1}^{t-1} f_s(\hat{\beta}_s) - F_{t-1}(\beta^*_{t-1}) \right) + (f_t(\hat{\beta}_t) - f_t(z) - \frac{(t-1)\sigma_{\max}}{2} \|z - \beta^*_{t-1}\|_2^2). \tag{1}$$

Then we can write

$$\begin{aligned}
V_t = \min_{\hat{\beta}_1 \in B} \max_{\lambda_1} \dots \min_{\hat{\beta}_t \in B} \max_{\lambda_t, z \in B} & \left(\sum_{s=1}^{t} f_t(\hat{\beta}_t) - F_t(z) \right) \\
\geq \min_{\hat{\beta}_1 \in B} \max_{\lambda_1} \dots \min_{\hat{\beta}_{t-1} \in B} \max_{\lambda_{t-1}} & \left[\left(\sum_{s=1}^{t-1} f_s(\hat{\beta}_s) - F_{t-1}(\beta^*_{t-1}) \right) \right. \\
& \left. + \min_{\hat{\beta}_t \in B} \max_{\lambda_t, z \in B} \left(f_t(\hat{\beta}_t) - f_t(z) - \frac{(t-1)\sigma_{\max}}{2} \|z - \beta^*_{t-1}\|_2^2 \right) \right].
\end{aligned} \tag{2}$$

Thus, by lower bounding the second term, we can unroll the recursion and lower bound the total regret. In particular, showing a bound of $\Omega(\frac{1}{t})$ will result in an overall regret lower bound of $\Omega(\log T)$, which would imply that ERM achieves minimax-optimal rates for OOD generalization (this is also how we prove Corollary 1).

We proceed by lower bounding the inner optimization term. We consider two possibilities for the choice of $\hat{\beta}_t$. Suppose $\|\hat{\beta}_t - \beta^*_{t-1}\|_2^2 \geq \frac{g^2}{8t\sigma_{\max}^2}$. Then by choosing $z = \beta^*_{t-1}$ the inner term can be lower bounded by $\min_{\hat{\beta}_t \in B} \max_{\lambda_t} \left(f_t(\hat{\beta}_t) - f_t(\beta^*_{t-1}) \right)$. Taylor expanding f_t around β^*_{t-1} gives

$$f_t(\hat{\beta}_t) - f_t(\beta^*_{t-1}) \geq \nabla f_t(\beta^*_{t-1})^T (\hat{\beta}_t - \beta^*_{t-1}) + \frac{\sigma_{\min}}{2} \|\hat{\beta}_t - \beta^*_{t-1}\|_2^2.$$

By Lemma 3, the adversary can always play λ_t such that $\nabla f_t(\beta^*_{t-1}) = 0$. So plugging this in we get

$$\min_{\hat{\beta}_t \in B} \max_{\lambda_t} \Big(f_t(\hat{\beta}_t) - f_t(\beta^*_{t-1}) \Big) \geq \frac{\sigma_{\min}}{2} \|\hat{\beta}_t - \beta^*_{t-1}\|_2^2$$
$$\geq \frac{g^2 \sigma_{\min}}{16 t \sigma_{\max}^2}.$$

Now consider the case where $\|\hat{\beta}_t - \beta^*_{t-1}\|_2^2 < \frac{g^2}{8t\sigma_{\max}^2}$. Suppose the adversary plays any λ_t such that $\|\nabla f_t(\hat{\beta}_t)\|_2 \geq g$ (by definition, such a choice is always possible). Here we again split on cases, considering the possible values of $\nabla f_t(\beta^*_{t-1})^T(\hat{\beta}_t - \beta^*_{t-1})$:

Case 1: $\nabla f_t(\beta^*_{t-1})^T(\hat{\beta}_t - \beta^*_{t-1}) \geq \frac{g^2 \sigma_{\min}}{16 t \sigma_{\max}^2}$

Following the same steps as previously, we find the lower bound

$$f_t(\hat{\beta}_t) - f_t(\beta^*_{t-1}) \geq \nabla f_t(\beta^*_{t-1})^T(\hat{\beta}_t - \beta^*_{t-1}) + \frac{\sigma_{\min}}{2} \|\hat{\beta}_t - \beta^*_{t-1}\|_2^2$$
$$\geq \nabla f_t(\beta^*_{t-1})^T(\hat{\beta}_t - \beta^*_{t-1})$$
$$\geq \frac{g^2 \sigma_{\min}}{16 t \sigma_{\max}^2}.$$

Case 2: $\nabla f_t(\beta^*_{t-1})^T(\hat{\beta}_t - \beta^*_{t-1}) < \frac{g^2 \sigma_{\min}}{16 t \sigma_{\max}^2}$

In this case the lower bound follows directly from Lemma 4.

Thus the lower bound is shown in all cases; it follows that

$$V_t \geq \min_{\hat{\beta}_1 \in B} \max_{\lambda_1} \dots \min_{\hat{\beta}_{t-1} \in B} \max_{\lambda_{t-1}} \left[\left(\sum_{s=1}^{t-1} f_s(\hat{\beta}_s) - F_{t-1}(\beta^*_{t-1}) \right) + \frac{g^2 \sigma_{\min}}{16 t \sigma_{\max}^2} \right]$$
$$= \min_{\hat{\beta}_1 \in B} \max_{\lambda_1} \dots \min_{\hat{\beta}_{t-1} \in B} \max_{\lambda_{t-1}} \left[\sum_{s=1}^{t-1} f_s(\hat{\beta}_s) - F_{t-1}(\beta^*_{t-1}) \right] + \frac{g^2 \sigma_{\min}}{16 t \sigma_{\max}^2}$$
$$= V_{t-1} + \frac{g^2 \sigma_{\min}}{16 t \sigma_{\max}^2}.$$

Expanding the recursion finishes the proof. □

D.2 Proof of Theorem 2

Theorem 2. *No deterministic algorithm can guarantee sublinear regret against bounded affine combinations of a finite set of strongly convex losses.*

Proof. We will show that for any deterministic algorithm, there exists a sequence of loss functions for which the regret is bounded as $\Omega(T)$. Assume the adversary can use coefficients greater than $-\alpha$. Define

$$f_{e_1}(x) = x^2, \qquad f_{e_2}(x) = x^4 + \frac{1}{2\alpha} x^2.$$

On round t, our player will choose to play $x \in \mathbb{R}$. We now describe our construction of the tth loss in the sequence: If $|x| < 1$, then we choose

$$f_t = (1 + \alpha) f_{e_1} - \alpha f_{e_2},$$

and if $|x| \geq 1$, we choose

$$f_t = f_{e_1}.$$

In the first case, the player suffers loss $f_t(x) \geq 0$, and in the second case, the player suffers loss ≥ 1. Suppose the player plays the first option a times and the second time b times, for a total of $a + b = T$ rounds, and suffers $\geq b$ loss.

Consider the possible best actions in hindsight. If $a \leq \frac{T}{2}$, then $x^* = 0$ suffers 0 loss, meaning the player's regret is at least $b = T - a \geq \frac{T}{2}$. If, on the other hand, $a > \frac{T}{2}$, then note that for any choice x the loss suffered is

$$\begin{aligned} -a\alpha x^4 + (a/2 + a\alpha + b)x^2 &\leq a\alpha(x^2 - x^4) + (a+b)x^2 \\ &= \left(a\alpha(1-x^2) + T\right)x^2. \end{aligned}$$

Choosing $x^* = \sqrt{1 + \frac{3}{\alpha}}$ results in player regret $\geq \frac{T}{2}$. In either case, the player suffers $\Omega(T)$ regret. □

D.3 Proof of Theorem 3

Theorem 3. *Even with a randomized algorithm against an oblivious adversary playing bounded affine combinations, achieving sublinear regret is NP-hard.*

Proof. Consider the problem of identifying the maximum size of a stable set of a graph on $|V|$ vertices; such a problem is not approximable in polynomial time to within a factor $|V|^{(1/2-\epsilon)}$ for any $\epsilon > 0$ unless $NP = P$ (Håstad, 1999; De Klerk, 2008). We will demonstrate that solving this problem up to a constant factor reduces to achieving sublinear regret on an online strongly convex game with bounded affine coefficients. Let $-\alpha$ represent the minimum negative coefficient allowed for the adversary.

Given the graph G on $|V| > 1$ vertices, denote by A its adjacency matrix. Then the maximum stable set size $\gamma(G)$ can be written

$$\frac{1}{\gamma(G)} = \min_{x \in \Delta_{|V|}} x^T(I + A)x,$$

by a result of Motzkin & Straus (1965). We define a game where the adversary has two functions:

$$f_{e_1}(x) = \frac{1}{1+\alpha}x^T(|V|I + A)x, \;\; f_{e_2}(x) = \frac{|V|-1}{\alpha}\|x\|_2^2.$$

Note the first function is strongly convex because $(|V|-1)I + A$ is diagonally dominant and therefore PSD.

Each round, the player plays some $x \in \Delta_{|V|}$, and the (oblivious) adversary chooses the loss

$$\begin{aligned} (1+\alpha)f_{e_1} - \alpha f_{e_2} &= x^T(|V|I + A)x - (|V|-1)\|x\|_2^2 \\ &= x^T(I + A)x. \end{aligned}$$

Define L_T as the loss suffered by the player after T rounds. Clearly, the optimal choice would be to play x such that $x^T(I + A)x = \frac{1}{\gamma(G)}$ each round, implying that $\frac{T}{\gamma(G)} \leq L_T$ and also that regret can be written $L_T - \frac{T}{\gamma(G)}$. Suppose there exists a polynomial-time strategy with regret growing sublinearly with T. Then by definition, there exists a constant $T_0 \in \text{poly}(|V|)$ such that on all rounds $T > T_0$, the player's regret is upper bounded as

$$L_T - \frac{T}{\gamma(G)} \leq \frac{1}{|V|}T \leq \frac{T}{\gamma(G)} \implies L_T \leq \frac{2T}{\gamma(G)}.$$

Putting these inequalities together,

$$\frac{1}{\gamma(G)} \leq \frac{L_T}{T} \leq \frac{2}{\gamma(G)},$$

which implies

$$\frac{1}{2}\gamma(G) \leq \frac{T}{L_T} \leq \gamma(G).$$

Recall that this holds for all $T > T_0$, so our polynomial-time algorithm has attained a 2-approximation to the maximum stable set size. □

D.4 PROOF OF THEOREM 4

Theorem 4. *Against an oblivious adversary playing bounded affine combinations, the achievable regret is lower bounded as* $\Omega(\sqrt{T})$.

Proof. For a fixed, convex loss ℓ and convex parameter space Θ, predicting with expert advice is known to have an information-theoretic minimax regret lower bound of $\Omega(\sqrt{T})$ (Cesa-Bianchi & Lugosi, 2006, Theorem 3.7). We will give a reduction which demonstrates that the same lower bound holds for bounded affine combinations of strongly convex losses.

Assume a fixed convex loss $\ell : \Theta \times \Theta \mapsto \mathbb{R}$ over convex Θ and fix the adversary's coefficient lower bound as $-\alpha$. Suppose on round t, we are presented with E experts' predictions, which we imagine as an E-dimensional vector $\tilde{\theta}_t$ whose ith entry is the prediction of the ith expert. Define the following functions over elements $\delta \in \Delta_E$:

$$f_{e_1}(\delta, \theta^*) = \frac{1}{1+\alpha}\left[\ell(\delta^T \tilde{\theta}_t, \theta^*) + \|\delta\|_2^2\right],$$
$$f_{e_2}(\delta, \theta^*) = \frac{1}{\alpha}\|\delta\|_2^2.$$

Note that both these functions are both strongly convex in δ. Consider what happens if the adversary plays

$$(1+\alpha) f_{e_1} - \alpha f_{e_2} = \ell.$$

Suppose for the sake of contradiction there exists an algorithm which achieves $o(\sqrt{T})$ regret with respect to δ^*, defined as the best fixed $\delta \in \Delta_E$ in hindsight:

$$\delta^* := \arg\min_{\delta \in \Delta_E} \sum_{t=1}^{T} \ell(\delta^T \tilde{\theta}_t, \theta_t^*).$$

As this represents a convex combination of the experts' predictions, it is clear that the loss suffered by δ^* will be less than or equal to the loss suffered by the best expert. This implies that by taking this algorithm's choice $\hat{\delta}_t$ each round and playing $\hat{\delta}_t^T \tilde{\theta}_t$, we will achieve $o(\sqrt{T})$ regret with respect to the best expert, defying the known lower bound. It follows that the lower bound of $\Omega(\sqrt{T})$ holds even for bounded affine combinations of strongly convex functions. □

E TECHNICAL LEMMAS

Lemma 1. *Recall* $\mathcal{R}^e(\beta)$ *is defined as the risk of* β *on the distribution* p^e. *Then*

$$\mathcal{R}^\lambda(\beta) = \sum_{e \in \mathcal{E}} \lambda_e \mathcal{R}^e(\beta).$$

Proof. The result follows from Fubini's theorem:

$$\begin{aligned}
\mathcal{R}^\lambda(\beta) &= \int_{\mathcal{X}\times\mathcal{Y}} \left[\sum_{e\in\mathcal{E}} \lambda_e p^e(x,y)\right] \ell(\beta, (x,y))\, d(x,y) \\
&= \sum_{e\in\mathcal{E}} \lambda_e \int_{\mathcal{X}\times\mathcal{Y}} p^e(x,y)\ell(\beta,(x,y))\, d(x,y) \\
&= \sum_{e\in\mathcal{E}} \lambda_e \mathcal{R}^e(\beta).
\end{aligned}$$

□

Lemma 2. *For any* $F_t = \sum_{s=1}^{t} f_t$, *there exist convex coefficients* $\hat{\lambda}$ *such that*

$$F_t = t \sum_{e\in\mathcal{E}} \hat{\lambda}_e f_e.$$

Proof. Every loss function f_t can be written as a convex combination of the original environment losses:

$$f_t = \sum_{e \in \mathcal{E}} \lambda_{t,e} f_e.$$

So, write

$$F_t = \sum_{s=1}^{t} f_t = \sum_{s=1}^{t} \sum_{e \in \mathcal{E}} \lambda_{t,e} f_e = \sum_{e \in \mathcal{E}} \left(\sum_{s=1}^{t} \lambda_{t,e} \right) f_e.$$

Clearly, $\sum_{e \in \mathcal{E}} \left(\sum_{s=1}^{t} \lambda_{t,e} \right) = t$. So, defining $\hat{\lambda}_e := \frac{1}{t} \left(\sum_{s=1}^{t} \lambda_{t,e} \right)$ gives the desired result. □

Lemma 3. *For any solution β^*_{t-1} which minimizes the sum of previously seen losses F_{t-1}, there exists a convex combination of losses f_t playable by the adversary for which $\nabla f_t(\beta^*_{t-1}) = 0$.*

Proof. By Lemma 2, we can write $F_{t-1} = (t-1) \sum_{e \in \mathcal{E}} \hat{\lambda}_e f_e$ for some convex coefficients $\hat{\lambda}$. Define $f_t = \sum_{e \in \mathcal{E}} \hat{\lambda}_e f_e = \frac{1}{t-1} F_{t-1}$. Since β^*_{t-1} minimizes F_{t-1} it follows that

$$\nabla f_t(\beta^*_{t-1}) = \frac{1}{t-1} \nabla F_{t-1}(\beta^*_{t-1}) = 0.$$

□

Lemma 4. *Let $\hat{\beta}_t$, λ_t be such that $\|\hat{\beta}_t - \beta^*_{t-1}\|_2^2 < \frac{g^2}{8t\sigma^2_{\max}}$ and $\|\nabla f_t(\hat{\beta}_t)\|_2 \geq g$. Define $z := \beta^*_{t-1} - c\nabla f_t(\hat{\beta}_t)$, where $c := 1/2t\sigma_{\max}$. If $\nabla f_t(\beta^*_{t-1})^T(\hat{\beta}_t - \beta^*_{t-1}) < \frac{g^2\sigma_{\min}}{16t\sigma^2_{\max}}$, then*

$$f_t(\hat{\beta}_t) - f_t(z) - \frac{(t-1)\sigma_{\max}}{2} \|z - \beta^*_{t-1}\|_2^2 \geq \frac{g^2\sigma_{\min}}{16t\sigma^2_{\max}}.$$

Proof. Expanding f_t around $\hat{\beta}_t$,

$$f_t(\hat{\beta}_t) - f_t(z) \geq -\nabla f_t(\hat{\beta}_t)^T(z - \hat{\beta}_t) - \frac{\sigma_{\max}}{2} \|z - \hat{\beta}_t\|_2^2,$$

which gives

$$\begin{aligned} f_t(\hat{\beta}_t) - f_t(z) - \frac{(t-1)\sigma_{\max}}{2} \|z - \beta^*_{t-1}\|_2^2 \\ \geq \nabla f_t(\hat{\beta}_t)^T(\hat{\beta}_t - z) - \frac{\sigma_{\max}}{2} \left(\|z - \hat{\beta}_t\|_2^2 + (t-1)\|z - \beta^*_{t-1}\|_2^2 \right) \end{aligned}$$

$$= \nabla f_t(\hat{\beta}_t)^T(\hat{\beta}_t - \beta^*_{t-1} + c\nabla f_t(\hat{\beta}_t)) - \frac{\sigma_{\max}}{2} \left(\|\beta^*_{t-1} - \hat{\beta}_t - c\nabla f_t(\hat{\beta}_t)\|_2^2 + (t-1)\|c\nabla f_t(\hat{\beta}_t)\|_2^2 \right). \tag{3}$$

By the triangle inequality,

$$\|\beta^*_{t-1} - \hat{\beta}_t - c\nabla f_t(\hat{\beta}_t)\|_2 \leq \|\beta^*_{t-1} - \hat{\beta}_t\|_2 + c\|\nabla f_t(\hat{\beta}_t)\|_2,$$

and therefore

$$\frac{1}{2}\|\beta^*_{t-1} - \hat{\beta}_t - c\nabla f_t(\hat{\beta}_t)\|_2^2 \leq \|\beta^*_{t-1} - \hat{\beta}_t\|_2^2 + c^2\|\nabla f_t(\hat{\beta}_t)\|_2^2.$$

Continuing with the lower bound in Equation 3,

$$\begin{aligned} &\geq \nabla f_t(\hat{\beta}_t)^T(\hat{\beta}_t - \beta^*_{t-1}) + c\|\nabla f_t(\hat{\beta}_t)\|_2^2 - \sigma_{\max} \left(\|\beta^*_{t-1} - \hat{\beta}_t\|_2^2 + c^2\|\nabla f_t(\hat{\beta}_t)\|_2^2 \right) - \frac{(t-1)\sigma_{\max}c^2}{2} \|\nabla f_t(\hat{\beta}_t)\|_2^2 \\ &\geq \nabla f_t(\hat{\beta}_t)^T(\hat{\beta}_t - \beta^*_{t-1}) + \left(c - \frac{1}{8t\sigma_{\max}} - \frac{(t+1)c^2\sigma_{\max}}{2} \right) \|\nabla f_t(\hat{\beta}_t)\|_2^2, \end{aligned}$$

where we've used the upper bound on $\|\beta^*_{t-1} - \hat{\beta}_t\|_2^2$ and simplified. Recalling that $c = \frac{1}{2t\sigma_{\max}}$ and noting that $\frac{t+1}{t^2} \leq \frac{2}{t}$,

$$\begin{aligned}
&= \nabla f_t(\hat{\beta}_t)^T(\hat{\beta}_t - \beta^*_{t-1}) + \left(\frac{1}{2t\sigma_{\max}} - \frac{1}{8t\sigma_{\max}} - \frac{(t+1)}{8t^2\sigma_{\max}}\right)\|\nabla f_t(\hat{\beta}_t)\|_2^2 \\
&\geq \nabla f_t(\hat{\beta}_t)^T(\hat{\beta}_t - \beta^*_{t-1}) + \frac{\|\nabla f_t(\hat{\beta}_t)\|_2^2}{8t\sigma_{\max}} \\
&\geq \nabla f_t(\hat{\beta}_t)^T(\hat{\beta}_t - \beta^*_{t-1}) + \frac{g^2}{8t\sigma_{\max}}.
\end{aligned}$$

By strong convexity,

$$(\nabla f_t(\beta^*_{t-1}) - \nabla f_t(\hat{\beta}_t))^T(\beta^*_{t-1} - \hat{\beta}_t) \geq \sigma_{\min}\|\beta^*_{t-1} - \hat{\beta}_t\|_2^2,$$

and therefore

$$\begin{aligned}
\nabla f_t(\hat{\beta}_t)^T(\hat{\beta}_t - \beta^*_{t-1}) &\geq \sigma_{\min}\|\beta^*_{t-1} - \hat{\beta}_t\|_2^2 - \nabla f_t(\beta^*_{t-1})^T(\hat{\beta}_t - \beta^*_{t-1}) \\
&> -\frac{g^2\sigma_{\min}}{16t\sigma_{\max}^2},
\end{aligned}$$

where the second inequality is due to the assumption in the Lemma statement. Plugging this in above gives

$$\begin{aligned}
\nabla f_t(\hat{\beta}_t)^T(\hat{\beta}_t - \beta^*_{t-1}) + \frac{g^2}{8t\sigma_{\max}} &> -\frac{g^2\sigma_{\min}}{16t\sigma_{\max}^2} + \frac{g^2}{8t\sigma_{\max}} \\
&\geq \frac{g^2\sigma_{\min}}{8t\sigma_{\max}^2} - \frac{g^2\sigma_{\min}}{16t\sigma_{\max}^2} \\
&= \frac{g^2\sigma_{\min}}{16t\sigma_{\max}^2},
\end{aligned}$$

completing the proof. □