An Online Learning Approach to Interpolation and Extrapolation in Domain Generalization

Elan Rosenfeld * Pradeep Ravikumar † Andrej Risteski ‡

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 task 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 is therefore an important task. This objective, broadly referred to as out-of-distribution (OOD) generalization, was classically explored in a setting where there is a single "source" training distribution and a different "target" test distribution. Achieving good performance in this setting is impossible in general, so researchers have formalized several possible frameworks to study. One common choice is to make specific assumptions about covariate or label shift (Widmer & Kubat, 1996; Bickel et al., 2009; Lipton et al., 2018); another approach is Distributionally Robust Optimization (DRO), where the test distribution is assumed to lie in some uncertainty set around the training distribution (Bagnell, 2005; Rahimian & Mehrotra, 2019).

There has been considerable recent interest in moving beyond a single source distribution, instead assuming that the set of training data is comprised of a collection of "environments" (Blanchard et al., 2011; Muandet et al., 2013; Peters et al., 2016) or "groups" (Hu et al., 2018; Duchi et al., 2019; Sagawa et al., 2020), each representing a distinct distribution,[1] where the group identity of each sample may be known. Such a setting is referred to as *domain generalization*. The hope is that by cleverly training on such a collection of groups, one can derive a robust predictor which will better transfer to unseen test data. Previous literature has focused exclusively on worst-case domain generalization, where the test environment is chosen to be the worst choice among a constrained set of possible test environments. It is useful to cast such a task as solving a one-shot min-max game, where the learner selects the predictor and then an adversary selects the test environment. A key specification for this game is how future test distributions depend on the training domains (i.e., the action space for the adversary).

The most immediate choice for the set of possible test environments is simply the set of training environments.

*Machine Learning Department, CMU. Email: elan@cmu.edu
†Machine Learning Department, CMU. Email: pradeepr@cs.cmu.edu
‡Machine Learning Department, CMU. Email: aristesk@andrew.cmu.edu
[1]Throughout this work, we use the terms "domain", "distribution", and "environment" interchangeably.

More broadly, researchers have considered how to perform well when the adversary is allowed to present test distributions which "interpolate" the training distributions or "extrapolate" beyond them, but it is unclear what is the ideal formalization of such interpolations and extrapolations. A popular choice for modeling interpolation is to allow any convex combination of the training environments—this is referred to as *group/sub-population shift*, and the resulting objective is known as *Group Distributionally Robust Optimization* (DRO). Duchi et al. (2019); Sagawa et al. (2020) give efficient algorithms for solving the Group DRO objective, but a key point is that the resulting min-max objective is *exactly equivalent* to when the adversary is limited to playing only the training environments. For modeling extrapolation, Krueger et al. (2020) consider "extrapolating" the training likelihoods (we make this formal in Section 2), but in this game the adversary's choice will still always be a vertex of the playable region. Thus, solving the one-shot min-max game under likelihood reweighting is always equivalent to simply minimizing worst-case risk on a discrete set.

In addition to this, formal analyses of these games are sparse. 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—but these claims are understood intuitively, rather than mathematically. Further, Sagawa et al. (2020) find that when using modern neural networks in the interpolation regime, explicitly solving the Group DRO objective does not yield better solutions than simple ERM with strong regularization. Thus the relative optimality of ERM and other domain generalization algorithms remains unclear. In light of these points, we begin by considering the question: **Is there an alternative to the single-round min-max game which might allow for a more in-depth analysis of the statistical and algorithmic properties of the task of domain generalization?**

One final additional caveat with this line of research is its emphasis on worst-case optimality over all possible test environments, which is often unnecessarily conservative. This is exemplified by empirical evaluations in the OOD literature: these works train a predictor on the source data and then evaluate it on *a single test set* which is chosen adversarially with respect to the predictor. Such a protocol often misses the mark for realistically comparing the expected performance of different algorithms. For example, Gulrajani & Lopez-Paz (2021) point out that many recent works deliberately evaluate on a single train/test environment split with an unreasonably difficult distribution shift. When averaging performance over multiple environment splits, they find that no algorithm outperforms ERM. This adversarial analysis can indeed be appropriate for quantifying how an algorithm will perform in the worst possible case (particularly in safety-critical applications), but this frequently does not reflect a predictor's quality in the real world: when the test environments are *not* chosen adversarially, a reasonable learning algorithm should be able to do significantly better. 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.**[2] This suggests there is room for a more nuanced measure of OOD generalization, one which adequately captures the purpose of such algorithms—to achieve consistently good performance on all possible test distributions—and allows for a formal comparison of their performances.

In this work, we aim to address the two main gaps identified above: formalizing the difference, if any (statistical *and* computational), between ERM and other OOD algorithms in both interpolation and extrapolation group shift settings; and doing so in a framework that allows us to analyze a predictor's performance on potentially non-adversarial (e.g., stochastic) future test environments. To do this, we take inspiration from the literature of online convex optimization (Hazan, 2016) and ask what can be achieved in a game where the learner is allowed to repeatedly refine their predictor upon observing new environments. Our analysis therefore captures an algorithm's ability to *learn and adapt* from multiple training distributions to suffer less under distribution shift and consequently perform better, on average, on future test sets. Our multi-round game 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 new perspective allows for a theoretical analysis of the computational and statistical complexity of interpolation versus extrapolation, formalizing and verifying the answers to several outstanding questions which until now have only been stated intuitively.

[2]This is a subtle point which we discuss in greater detail in Section 3.1.

Concretely, this work makes the following contributions:

- We recast domain generalization as a repeated online game between an adversary presenting test distributions and a player minimizing *cumulative regret*. This framework enables meaningful analysis beyond the single-round minimax setting, and we expect it can serve as a new approach to the formal study of the efficacy of robust OOD generalization algorithms.
- Under an existing notion of inter- and extrapolation, we tightly characterize their respective complexities. Specifically, we prove that i) extrapolation is indeed exponentially more difficult than interpolation in a computational sense, but ii) the statistical complexity of extrapolation is not significantly higher.
- For both inter- and extrapolation, we show that ERM—or a noisy variant—is *provably minimax-optimal* with respect to regret, as a function of the number of environments observed. For minimizing regret over any time horizon, it is impossible to improve over ERM without additional assumptions. This result supplements recent works which support the same idea theoretically (Rosenfeld et al., 2021) and empirically (Gulrajani & Lopez-Paz, 2021) for the single-round setting.

2 The Single-Round Domain Generalization Game

The key assumption of domain generalization is that the training set comprises a set of distinct domains $\mathcal{E} = \{e_i\}_{i=1}^{E}$, each of which indexes a probability distribution p^e, and that the test environment will relate to these domains in some pre-specified way. Let us denote the set of such possible test distributions by $\mathcal{E}_{\text{test}}$. It's common to use a minimax formulation, wherein the learner's goal is to minimize the worst-case error over the possible test distributions $\mathcal{E}_{\text{test}}$. For a set of predictors $\mathcal{F}$ and loss ℓ, our goal is thus to solve the objective

$$\min_{f \in \mathcal{F}} \max_{e \in \mathcal{E}_{\text{test}}} \mathbb{E}_e[\ell(f)].$$

In an adversarial framework, $\mathcal{E}_{\text{test}}$ is the "playable region" of the adversary, similar to the uncertainty set in traditional DRO. A critical ingredient of the game as noted earlier is how this set of test distributions $\mathcal{E}_{\text{test}}$ depends on the training domains $\mathcal{E}$. It is typically presented as belonging to one of two distinct settings: *interpolation* and *extrapolation*. Intuitively, the interpolation setting should consist of environments which do not vary "beyond" the observed training environments, while the extrapolation setting should allow for such variation to some degree. However, these terms do not have a single agreed-upon meaning.

Formally modeling interpolation. 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 (Duchi et al., 2019; Albuquerque et al., 2020; Sagawa et al., 2020).[3] We model the interpolation of a set of domains as all convex combinations (i.e., mixtures) of their likelihoods. Formally, an interpolation of the domains in $\mathcal{E}$ is any distribution which is written

$$p^\lambda := \sum_{e \in \mathcal{E}} \lambda_e p^e, \tag{1}$$

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 the convex hull of the environments $\mathcal{E}$ in distribution-space. We will denote this convex hull $\text{Conv}(\mathcal{E})$.

Observe that this definition is mathematically equivalent to the set of environments which can be generated via group shift, and solving the above min-max objective is precisely Group DRO. However, this notion of single-round interpolation, while perhaps intuitive, does not actually induce a more meaningful playable region for the adversary. This is because for any predictor, the optimal choice for the adversary will be whichever training environment produces the highest risk; that is, the adversary will *always* play a vertex of the simplex. Thus, these two games are equivalent:

[3] As another possibility, we could directly interpolate between two samples, but this is unlikely to be meaningful for highly complex data such as images. If we were to pose a generative model, it would instead be natural to consider interpolations of the generative parameters.

Proposition 1 (Equivalence of interpolation and the discrete one-shot game)**.**

$$\min_{f\in\mathcal{F}} \max_{e\in Conv(\mathcal{E})} \mathbb{E}_e[\ell(f)] \;=\; \min_{f\in\mathcal{F}} \max_{e\in\mathcal{E}} \mathbb{E}_e[\ell(f)].$$

We note that in some prior work on Group DRO, learning models that minimize worst-case sub-population risk is indeed the goal—that is, they only care about test domains that match one of the source domains. In the broader domain generalization literature, however, it does not seem that this form of interpolation provides any additional constraint on OOD learning without additional regularization (Hu et al., 2018).

Generalizing to extrapolation. It is not immediately obvious how to extend this concept to include extrapolation. Krueger et al. (2020) suggest allowing for combinations in which the coefficients are still restricted to sum to 1, but may be slightly negative, where the minimum coefficient is given as a hyperparameter α: $\sum_{e\in\mathcal{E}} \lambda_e = 1,\ \lambda_e \geq -\alpha\ \forall e \in \mathcal{E}$. We refer to such combinations as "bounded affine" combinations, and the objective they induce is equivalent to a fixed linear combination of the average loss plus the worst-case loss. It is immediate that the adversary's optimal choice is still on a vertex, so this game also reduces to minimizing over a discrete set:

Proposition 2 (Equivalence of constraint set for extrapolation and the discrete one-shot game)**.**

$$\min_{f\in\mathcal{F}} \max_{e\in Extr_\alpha(\mathcal{E})} \mathbb{E}_e[\ell(f)] = \min_{f\in\mathcal{F}} \max_{e\in\mathcal{E}} \left[(1+E\alpha)\mathbb{E}_e[\ell(f)] - \alpha \sum_{e'\in\mathcal{E}} \mathbb{E}_{e'}[\ell(f)]\right],$$

where $Extr_\alpha(\cdot)$ is all α-bounded affine combinations.

Thus we find that for a single round, the precise meaning of these objectives is unclear: the adversary is still choosing from a discrete set, and this model does not seem to capture the intuition that extrapolation should be fundamentally "harder" than interpolation. This shortcoming motivates our modified approach based on long-term regret, which we introduce shortly.

For extrapolating likelihoods, note that the resulting function is not guaranteed to be a probability distribution, as it could result in negative measure—one can instead frame it as reweighting of the environment *risks* (thus in Proposition 2 above, $\mathbb{E}[\cdot]$ refers to general Lebesgue integration). We study this reweighting of risks in Section 4.2, and we find that generalizing well over all such combinations is NP-hard. This provable difficulty in extrapolating validates our proposed sequential game, but it also indicates that additional assumptions may be necessary for modeling domain generalization. This raises interesting questions about what is the correct or most useful model of "extrapolation", which we do not address here.

3 The Sequential 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 sequential test domains and must refine their predictor at each round. We're therefore interested in the player's ability to *learn continuously* and improve in 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. Specifically, 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 an analysis of the average loss (over time) of a learning algorithm across all possible test sequences—in order to bound this performance, we consider the worst such sequence. In Section 3.1 we expound upon this idea, comparing in detail our game to existing single-round minimax settings and discussing the benefits it affords.

We now describe the game which will allow a formal analysis of the efficacy of various domain generalization strategies. The full game can be found in the box titled Algorithm 1. Note we describe a specific instance where the adversary is limited to group mixtures as described in Section 2; **the general game allows for any formally specified action space for the adversary** and we expect 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).

Algorithm 1 : Domain Generalization Game
(likelihood reweighting)

Input: Convex parameter space B, distributions $\{p^e\}_{e\in\mathcal{E}}$ over $\mathcal{X}\times\mathcal{Y}$,
strongly convex loss $\ell : B \times (\mathcal{X}\times\mathcal{Y}) \to \mathbb{R}$, playable region Δ.
for $t = 1 \dots T$ **do**
1. Player chooses parameters $\hat{\beta}_t \in B$.
2. Adversary chooses coefficients $\lambda_t \in \Delta$.
3. Define $f_t(\beta) := \mathbb{E}_{(x,y)\sim p^{\lambda_t}}[\ell(\beta,(x,y))] = \sum_{e\in\mathcal{E}} \lambda_{t,e}\mathbb{E}_{(x,y)\sim p^e}[\ell(\beta,(x,y))]$.

end for
Player suffers regret

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

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 all $\beta \in B$ and $e \in \mathcal{E}$, 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 in Equation 1. For now, we assume that every choice of λ by the adversary is a set of convex coefficients—that is, an interpolation—which ensures that p^{λ_t} is a valid probability distribution; we will relax this restriction in Section 4.2. 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 the above notation we will drop the subscript t when it is not necessary.

It's important to note that in this game the player does not begin "training" until the first round; the initial environments $\mathcal{E}$ serve only to define the playable region for the adversary. Thus to recover the existing notion of single-round domain generalization, where the estimator has already seen the source environments $\mathcal{E}$ and next faces an unseen test environment, the online game would actually begin with the adversary playing each of the environment distributions in $\mathcal{E}$ once. As in standard online learning, our goal is to minimize *regret* with respect to the best fixed predictor in hindsight after T rounds. That is, we hope to minimize

$$\sum_{t=1}^{T} f_t(\hat{\beta}_t) - \min_{\beta\in B}\sum_{i=1}^{T} f_t(\beta). \tag{2}$$

Observe that this notion of regret straightforwardly generalizes previous work on single-round domain generalization. By allowing $T \to \infty$, we have a meaningful measure of success: each time we are presented with a new environment, we update our predictor in the hopes of improving our average performance. Crucially, this modification allows us to ask questions about the rate at which our regret decreases as a function of the number of environments observed. It also better reflects the idea that our algorithm's performance should not be evaluated in a vacuum: we aim to perform well relative to how we *could* have performed over all timesteps with a single predictor.

3.1 The Benefits of Online Regret vs. Single-Round Loss

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

Significance of regret with respect to a fixed baseline. The second term in Equation 2 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,[4] the game therefore reduces to repeated, independent instances of the single-round version; clearly, the best we can do to minimize worst-case loss each single round is to play the minimax-optimal parameters $\beta^* := \arg\min_{\beta \in B} \max_{\lambda \in \Delta_E} \mathcal{R}^{\lambda}(\beta)$. 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 an algorithm's performance in a 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^*, we should consider it more likely that this is representative of future test environments (i.e., we will continue to encounter p^*) than that Nature is actively trying to give us the largest possible loss. Consequently we should switch strategies and play $\arg\min_{\beta \in B} \mathcal{R}^{p^*}(\beta)$, which will have better performance if the pattern continues. Thus, existing frameworks overemphasize minimax performance in individual rounds—even though in reality, distribution shift is rarely adversarial—while ignoring possible improvements over time via *adaptation to the changing environments*. 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.

One particular setting where the benefits of this new framework are readily apparent is under gradual distribution shift. The single-round minimax formulation is intended for safety-critical applications where even a tiny mistake is fatal; however, when this is not the case, such an approach is far too conservative, and regret-based analyses provide a much clearer picture of expected performance. Our framework is thus not intended to supplant the single-round setting, but rather to supplement it with a new, more realistic method of formal analysis of domain generalization algorithms.

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 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 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.

[4] 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 emphasize again that while the above example considers a stochastic adversary, **we do not in general assume a prior over environments**. Instead, we perform a minimax analysis to guard against the worst possible sequence of test distributions. We are measuring average regret with respect to *time*.

4 Theoretical Results

Before presenting our main theoretical results, we begin with a lemma which greatly simplifies the analysis by recharacterizing the adversary's playable region.

Lemma 1. *Recall $\mathcal{R}^e(\beta)$ is defined as the risk of β on the distribution p^e. Then for all $\lambda \in \Delta_E$, it holds that $\mathcal{R}^\lambda(\beta) = \sum_{e\in\mathcal{E}} \lambda_e \mathcal{R}^e(\beta)$.*

This reframing allows us to generalize our analysis to extrapolation without worrying that the resulting measure is not a probability distribution. Lemma 1 implies that when the adversary chooses convex coefficients λ_t, they are equivalently choosing a loss function f_t which is a combination of $\{f_e\}_{e=1}^E$, the individual environments' risks. Each choice of λ_t uniquely defines the resulting loss function f_t; moving forward we will drop this explicit dependency in our notation.

4.1 Convex Combinations

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 player and adversary:

$$V_T := \min_{\hat{\beta}_1 \in B} \max_{\lambda_1 \in \Delta_E} \dots \min_{\hat{\beta}_T \in B} \max_{\lambda_T \in \Delta_E} \left(\sum_{t=1}^T f_t(\hat{\beta}_t) - \min_{\beta \in B} \sum_{t=1}^T f_t(\beta) \right).$$

For fixed T, this allows us to formalize minimax bounds on the regret. In the traditional literature, 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 and $\sigma_{\min}$ is the minimum curvature of f.[5] This means the minimax-optimal rate for regret is $\Theta(\log t)$ (Hazan et al., 2007; Bartlett et al., 2007).

In contrast to traditional online learning, where the adversary is free to choose its loss from a large non-parametric class such as all strongly convex functions, our interpolation game severely restricts the adversary, allowing only convex combinations of the risks of the E 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 $V_t > \frac{g^2\sigma_{\min}}{16\sigma_{\max}^2} \log t$.*

Proof Sketch. The general idea of the proof is to lower bound the regret on round t by the optimal regret on round $t-1$ plus some additional loss suffered on round t. This loss depends on the distance from the chosen parameter on round t to the regret minimizer for round $t-1$, as well as the adversary's choice on round t, and it can be bounded as $\Omega(1/t)$. By unrolling the recursion we derive an overall lower bound of order $\sum_{i=1}^t \frac{1}{i} > \log t$. The full proof can be found in Appendix A. □

Theorem 1 provides insight into how the statistical complexity of generalizing to domain interpolations depends on the geometry of the source domains. Observe that the minimum forceable gradient norm g encodes a sort of "radius" of the convex hull of loss gradients—it is easy to see that if a ball of radius r can be embedded in $\text{Conv}(\{\nabla f_e(\beta)\}_{e=1}^E)$ then $g > r$. Thus, the restriction of the adversary to the convex hull of distributions entails a restriction on the geometry of the convex hull of the corresponding loss gradients,

[5] We've omitted some details; see Abernethy et al. (2008) for the full result.

which subsequently determines the regret our player can be forced to suffer. The bound does not directly depend on the *number* of training 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 whose optimal parameters are very similar to one another.

With respect to the asymptotic rate of regret, this theorem provides a somewhat surprising conclusion. Even with full knowledge of the adversary's limited selection, Theorem 1 shows that no algorithm can do asymptotically better than if we were playing against the more powerful adversary playing any strongly convex function. 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—after observing t environments, FTL would therefore play

$$\beta_{\text{FTL}} = \arg\min_{\beta} \sum_{s=1}^{t} f_s(\beta).$$

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 $\sum_{s=1}^{t} G_s^2/2s\sigma_{\min} = O(\log t)$. While Theorem 1 applies to the multi-round game, it has useful implications for the single-round setting. A simple corollary provides a tight bound on the attainable regret as a function of the number of environments seen. To our knowledge, this is the first such bound for single-round domain generalization.

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

4.2 Bounded Affine Combinations

One could argue that allowing the adversary only convex combinations of domains is perhaps too good to hope for. Indeed, 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 2, Krueger et al. (2020) consider allowing the adversary to play bounded affine combinations of the environments; while they provide no formal results for their proposed algorithm, this conceptualization of extrapolation seems a natural extension. Clearly, this game is no easier for the player—in fact, we will demonstrate that it is *significantly* harder. For general Lipschitz functions, it is known that against the worst-case sequence, no deterministic strategy can guarantee sublinear regret, and attaining sublinear regret with a randomized strategy is NP-hard. Further, there is a regret lower bound of $\Omega(\sqrt{T})$ which was recently shown to be achievable with Follow-The-Perturbed-Leader (FTPL), assuming access to an optimization 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.* These results are also surprising, as an adversary that can play arbitrary Lipschitz functions is significantly more powerful than the adversary in our game.

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

Proof. We'll show that for any algorithm, there exists a sequence of loss functions chosen in response by the adversary for which the regret is bounded as $\Omega(T)$. Assume the adversary can use coefficients greater than $-\alpha$. Define

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

On round t, our player will choose to play $\beta \in \mathbb{R}$. We now describe our construction of the tth loss in the sequence: If $|\beta| < 1$, then we choose $f_t = (1+\alpha)f_{e_1} - \alpha f_{e_2}$, and if $|\beta| \geq 1$, we choose $f_t = f_{e_1}$. In the first case, the player suffers loss $f_t(\beta) \geq 0$, and in the second case, the player suffers loss ≥ 1. Suppose the player plays the first option a times and the second option 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 $\beta^* = 0$ 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 β the loss suffered is

$$-a\alpha\beta^4 + (a/2 + a\alpha + b)\beta^2 \leq a\alpha(\beta^2 - \beta^4) + (a+b)\beta^2 = \left(a\alpha(1-\beta^2) + T\right)\beta^2.$$

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

For completeness's sake, in Appendix B we also include a proof of the existence of a regression task and a set of environments which could give rise to such a set of loss functions. □

Thus we find that just as in the general non-convex case, a weaker adversary is necessary. In the following we consider a relaxed version with an "oblivious" adversary: this adversary is forced to select the entire sequence of loss functions at the beginning of the game (our lower bounds hold despite this relaxation). We might hope that against such a restricted adversary, 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. *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 $\frac{1}{\gamma(G)} = \min_{\beta \in \Delta_{|V|}} \beta^T(I + A)\beta$ by a result of Motzkin & Straus (1965). We define a game where the adversary has two functions:

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

Note that f_{e_1} is strongly convex because $(|V|-1)I + A$ is diagonally dominant and therefore PSD. Each round, the player plays some $\beta \in \Delta_{|V|}$, and the (oblivious) adversary chooses the loss

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

Define L_T as the loss suffered by the player after T rounds. Clearly, the optimal choice would be to play β such that $\beta^T(I + A)\beta = \frac{1}{\gamma(G)}$ 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, we get $\frac{1}{\gamma(G)} \leq \frac{L_T}{T} \leq \frac{2}{\gamma(G)}$, 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. □

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 turn our attention to the statistical complexity of regret minimization under bounded affine combinations. Recall that for the case of convex combinations (i.e. interpolations), Theorem 1 shows a minimax lower bound of $\Omega(\log t)$ which can be achieved with standard ERM. Before we consider the bounded affine setting (i.e. extrapolations), we again note that for an adversary playing arbitrary Lipschitz functions, Suggala & Netrapalli (2020) demonstrate that with access to a non-convex optimization oracle, FTPL can achieve the minimax lower bound of $\Omega(\sqrt{T})$. The FTPL strategy is to play the parameter which minimizes the sum of the observed environments plus a noise term—specifically, FTPL takes the sum of existing risks, samples a random linear function of the parameters, and solves for the parameters which minimize this "perturbed" sum. In our game, then, FTPL is just a noisy variant of ERM. Computational limitations notwithstanding, the natural next question is if playing against an oblivious adversary is enough of a relaxation that we can surpass this lower bound. That is, can we outperform ERM in this setting *at all?* Our final result answers this question in the negative:

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 $\delta \in \Delta_E$:

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

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 playing $\hat{\delta}_t$ which achieves $o(\sqrt{T})$ regret with respect to δ^*, defined as the best fixed $\delta \in \Delta_E$ in hindsight:

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

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. □

This theorem implies two crucial points: firstly, that *ERM remains minimax optimal for this model of extrapolation*; and secondly, that *proper regularization is essential for good OOD generalization.* This provides theoretical justification for the empirical findings of Sagawa et al. (2020) and complements existing results on the value of explicit regularization for group shift (Hu et al., 2018). Additionally, we find that even though there is an exponential computational complexity gap between the two tasks, the statistical gap is not too large—$\Theta(\log T)$ versus $\Theta(\sqrt{T})$ regret.

5 Related Work

Many works provide formal guarantees for OOD generalization by assuming invariances in the causal structure of the data: a set 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 sufficiently strong conditions it is then possible to identify which features have invariant relationships with the target variable; recovery of these features ensures reasonable performance despite arbitrary future interventions on the other variables. However, these works 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 "invariance" within the context of causality. The IRM objective (Arjovsky et al., 2019) was designed for such a setting assuming the target variables' causal mechanisms remain invariant, but it lacked serious theoretical justification; Krueger et al. (2020) likewise suggest an algorithm for extrapolation but similarly fail to provide any formal guarantees. Rosenfeld et al. (2021) subsequently showed that, while these and other similar objectives may work under strong conditions in the linear setting, the same cannot be said for more complex data. Albuquerque et al. (2020) theoretically analyze extrapolation beyond the convex hull of domain likelihoods and give generalization bound via $\mathcal{H}$-divergences. Unfortunately, this bound scales linearly with both the maximum discrepancy between pairs of training distributions and between the test distribution and training environment hull.

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—which are for more general online learning—present new algorithms and give upper bounds, while this work focuses on OOD generalization and 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 (ERM and a noisy variant) are already optimal.

6 Conclusion and Future Directions

This work presents the first formal results demonstrating 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—given the observed failure of ERM in practice, this suggests that there is quite a bit more subtlety to distribution shift in the real world. Taken together, our results present strong evidence that the "likelihood reweighting" model of distribution shift, while perhaps appropriate for specific settings involving sub-populations, might not be appropriate for the more general study of extrapolation to new domains. It could instead be beneficial to reconsider existing notions of inter- and extrapolation—particularly those involving linearity or generic likelihood reweighting—in the context of online learning, where the notions of regret and stochastic adversaries allow for more a nuanced study of statistical and algorithmic complexity.

We see two important directions for further research. First, the proposed domain generalization game serves as a standalone framework for the theoretical analysis of learning algorithms. As discussed in Section 3.1, considering regret in the online setting provides a more nuanced signal of an algorithm's expected performance, especially when we are not too worried about the *literal worst case test distribution*. We hope that this new perspective will better enable future work to provide formal OOD generalization guarantees for their proposed methods. We note that this work considers only strongly convex functions, but using the same techniques one could extend the analysis to more general classes such as all convex losses; this setting might eliminate the statistical complexity gap and could lead to additional insight into the differences between inter- and extrapolation.

Second, there still remains significant flexibility in how we define "interpolation" and "extrapolation" with respect to training environments; we consider one specific notion in this work, and we show that ERM remains

optimal—implying that alternative formulations may be preferable. However, 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.

Acknowledgements

We thank Zack Lipton for his feedback on the framing of this work.

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.

Bickel, S., Brückner, M., and Scheffer, T. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(9), 2009.

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.

Duchi, J., Hashimoto, T., and Namkoong, H. Distributionally robust losses for latent covariate mixtures. *arXiv preprint arXiv:2007.13982*, 2019.

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, W., Niu, G., Sato, I., and Sugiyama, M. Does distributionally robust supervised learning give robust classifiers? In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2029–2037. PMLR, 2018.

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.

Lipton, Z., Wang, Y.-X., and Smola, A. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3122–3130. PMLR, 2018. URL http://proceedings.mlr.press/v80/lipton18a.html.

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.

Rahimian, H. and Mehrotra, S. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

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.

Widmer, G. and Kubat, M. Learning in the presence of concept drift and hidden contexts. *Machine learning*, 23(1):69–101, 1996.

A 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,

$$\sum_{s=1}^{t} f_s(\hat{\beta}_s) - F_t(z) \geq \left(\sum_{s=1}^{t-1} f_s(\hat{\beta}_s) - F_{t-1}(\beta_{t-1}^*) \right) + (f_t(\hat{\beta}_t) - f_t(z) - \frac{(t-1)\sigma_{\max}}{2} \|z - \beta_{t-1}^*\|_2^2). \tag{3}$$

Then we can write

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

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

$$\begin{aligned}
\min_{\hat{\beta}_t \in B} \max_{\lambda_t} \left(f_t(\hat{\beta}_t) - f_t(\beta_{t-1}^*) \right) &\geq \frac{\sigma_{\min}}{2} \|\hat{\beta}_t - \beta_{t-1}^*\|_2^2 \\
&\geq \frac{g^2 \sigma_{\min}}{16 t \sigma_{\max}^2}.
\end{aligned}$$

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

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

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

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

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

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

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

Expanding the recursion finishes the proof. □

B Proof of Existence for Theorem 2

We restate Theorem 2 for convenience:

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

In the main body, we prove the primary claim. Here we include proof of the existence of a regression task over a set of distributions which induces the loss functions we construct in our proof.

Proof. Suppose we are regressing labels $y \in \mathbb{R}$ on observations $z \in \mathbb{R}^2$ with squared loss. We'll define our classifier with a parameter β such that given an observation (z_1, z_2) we predict $\beta^2 z_1 + \beta z_2$. This is of course an unusual regression setup, but we're just giving an existence proof for a simple lower bound.

The first environment will assign all its probability mass to a single example $(z_1, z_2, y) = (0, 1, 0)$. Thus, if we choose a parameter β, in this environment we will suffer risk $\mathbb{E}[(\beta z_1^2 + \beta z_2 - y)^2] = \beta^2$. This produces the first environment, loss $f_{e_1}(\beta) = \beta^2$.

We define the second environment as having two possible samples: one is $(z_1, z_2, y) = (0, \sqrt{\frac{2\alpha+1}{2\alpha}}, 0)$ and the other is $(z_1, z_2, y) = (\sqrt{\frac{2\alpha+1}{2\alpha}}, 0, 0)$. Thus, the first sample induces loss $\frac{2\alpha+1}{2\alpha}\beta^2$, and the second induces loss $\frac{2\alpha+1}{2\alpha}\beta^4$. Now for the probabilities: we assign probability $\frac{1}{2\alpha+1}$ to the first point and $\frac{2\alpha}{2\alpha+1}$ to the second point. Clearly these sum to 1, and taking the expectation over losses we see that the overall risk is $\beta^4 + \frac{1}{2\alpha}\beta^2$, as desired. □

C Lemmas

Lemma 1. *Recall $\mathcal{R}^e(\beta)$ is defined as the risk of β on the distribution p^e. Then for all $\lambda \in \Delta_E$, it holds that $\mathcal{R}^\lambda(\beta) = \sum_{e \in \mathcal{E}} \lambda_e \mathcal{R}^e(\beta)$.*

Proof. Using Fubini's theorem, we have

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

□

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

$$\begin{aligned}&f_t(\hat{\beta}_t) - f_t(z) - \frac{(t-1)\sigma_{\max}}{2}\|z - \beta^*_{t-1}\|_2^2 \\ &\geq \nabla f_t(\hat{\beta}_t)^T(\hat{\beta}_t - z) - \frac{\sigma_{\max}}{2}\left(\|z - \hat{\beta}_t\|_2^2 + (t-1)\|z - \beta^*_{t-1}\|_2^2\right) \\ &= \nabla f_t(\hat{\beta}_t)^T(\hat{\beta}_t - \beta^*_{t-1} + c\nabla f_t(\hat{\beta}_t)) - \frac{\sigma_{\max}}{2}\left(\|\beta^*_{t-1} - \hat{\beta}_t - c\nabla f_t(\hat{\beta}_t)\|_2^2 + (t-1)\|c\nabla f_t(\hat{\beta}_t)\|_2^2\right). \qquad (4)\end{aligned}$$

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 4,

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

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. □