Structured Robustness for Distribution Shifts

Erfan Darzi[*]
Harvard University, MIT

Alexander Marx
TU Berlin

Abstract

Out-of-distribution (OOD) data often undermines reliable model deployment in high-stakes domains such as financial markets, where overlooked correlations and unexpected shifts can render predictive systems ineffective. We propose STAR (Structured Transformations and Adversarial Reweighting), a framework that leverages the geometry of distribution shifts by combining *transformation-based* invariances with *divergence-based* robust optimization. Specifically, STAR places an f-divergence ball around each label-preserving transformation of the training sample, empowering an adversary to apply known transformations and reweight the resulting data within a specified divergence radius. This design captures both large, structured shifts and subtle, unmodeled perturbations—a critical step toward mitigating shortcuts and spurious correlations. Notably, STAR recovers standard distributionally robust optimization if no structured transformations are assumed. We establish a uniform-convergence analysis showing that minimizing STAR's empirical nested min–max objective achieves low worst-case error over all admissible shifts with high probability. Our results quantify the additional samples needed to handle the adversary's flexibility, providing theoretical guidance for selecting the divergence radius based on problem complexity. Empirical studies on synthetic and image benchmarks confirm that STAR outperforms baselines, consistent with our theoretical findings.

1 Introduction

Out-of-distribution (OOD) generalization remains a persistent challenge in machine learning: models often fail when test samples deviate from training conditions. One widely used strategy is *transformation-based* learning, which encodes label-preserving transformations $\mathcal{T}$ (e.g., rotations, flips) to protect against known shifts by minimizing worst-case loss over $T \in \mathcal{T}$. Although this approach effectively exploits domain knowledge, it can be restrictive if actual test shifts extend beyond the predefined transformations, leading to overlooked or under-modeled variations.

On the other hand, *divergence-based* distributionally robust optimization (DRO) (Namkoong & Duchi, 2016) explores distributions within an f-divergence ρ of the empirical measure, thereby accommodating unstructured or adversarial reweightings. However, standard DRO does not explicitly leverage geometric or semantic transformations, potentially over-preparing for shifts that simpler invariances would address more directly. Moreover, the lack of explicit structure may cause the model to focus on unrealistic distributions, increasing computational cost and diluting performance in practice.

We propose **STAR** (*Structured Transformations and Adversarial Reweighting*), an adversarial framework designed to address these limitations by simultaneously capturing large, structured shifts and subtler, unmodeled perturbations. STAR places an f-divergence ball around each label-preserving transformation of the training set, allowing the adversary to reweight examples under each transformation. By doing so, STAR adapts to a wider range of potential shifts while retaining the interpretability of transformations and the flexibility of adversarial reweighting.

We establish a uniform-convergence analysis showing that minimizing STAR's empirical min–max objective achieves low worst-case error, with sample-complexity bounds quantifying the data required to handle both geometry and reweighting. Empirical evaluations confirm that STAR outperforms methods relying solely on transformations or on divergence-based DRO.

[*] Corresponding Author. darzi@mit.edu

2 Related Works

Learning robust models that generalize under distribution shifts has been a central focus in recent years. One prominent strategy is transformation-invariant learning, which aims to capture features stable under certain data transformations or across domains. Invariant Risk Minimization (IRM) is a notable example that learns a data representation supporting an optimal predictor invariant across multiple training environments Arjovsky et al. (2019), thereby reducing reliance on spurious correlations. More recently, Montasser et al. investigate theoretical guarantees for transformation-invariant learning in settings where test samples are generated by applying a class of transformations to training data Montasser et al. (2024). They provide learning rules with PAC guarantees and frame the problem as a two-player game between the learner and an adversary choosing transformations, which offers a worst-case (adversarial) view on achieving OOD generalization. These approaches underscore the value of leveraging prior knowledge of invariances to improve out-of-distribution performance.

Another line of research studies adversarial robustness and distributionally robust optimization (DRO) as means to safeguard against worst-case shifts. Namkoong and Duchi introduced an f-divergence–based DRO framework that seeks a worst-case reweighting of the data, effectively up-weighting high-loss examples via a min-max objective Namkoong & Duchi (2016). This formulation provides performance guarantees under shifts and closely relates to adversarial training, which can be seen as DRO over perturbation sets – for example, Sinha et al. show that augmenting updates with worst-case input perturbations (within a Wasserstein ball) confers certifiable robustness to adversarial examples Sinha et al. (2017). Similarly, DRO has been applied to group shifts: Sagawa et al. train models to minimize the worst-case loss across pre-defined data groups, improving worst-group accuracy on challenging subsets Sagawa et al. (2019).

3 Problem setup

We study a learning setting where an m-sample $S = \{(x_i, y_i)\}_{i=1}^m$ is drawn i.i.d. from an unknown distribution $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{Y}$. Alongside this training set, we have a finite family of label-preserving transformations $\mathcal{T} = \{\, T_1, \ldots, T_{|\mathcal{T}|}\}$, each of which acts on $x \in \mathcal{X}$ while leaving the corresponding label y unchanged. To handle shifts beyond these structured transformations, we adopt an f-divergence measure $D_f(\cdot\|\cdot)$ and a radius $\rho \geq 0$, thereby allowing an adversary to reweight or perturb the distribution within ρ divergence of a baseline measure. We measure the performance of any hypothesis $h \in \mathcal{H}$ through a loss function $\ell :\ \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$.

4 Methodology

Let $\mathcal{T}$ be a known set of label-preserving transformations, and let f be any convex function defining an f-divergence D_f. Given an m-sample $S = \{(x_i, y_i)\}_{i=1}^m$ from distribution $\mathcal{D}$, we denote its empirical measure by P_S. Our goal is to learn $h \in \mathcal{H}$ that is robust to applying a transformation $T \in \mathcal{T}$, and reweighting the transformed empirical distribution within $D_f(\cdot\|\ T(P_S)) \leq \rho$.

Our Min–Max Objective. Formally, we solve

$$\min_{h \in \mathcal{H}}\ \max_{T \in \mathcal{T}}\ \max_{Q\,:\,D_f(Q\|\ T(P_S)) \leq \rho}\ \mathbb{E}_{(x,y)\sim Q}\big[\ell\big(h(x), y\big)\big]. \tag{1}$$

Any (T, Q) pair represents a shift: T enacts geometric or semantic changes, while Q further redistributes mass within f-divergence ρ of $T(P_S)$. This nested min-max-max formulation ensures that we minimize the worst-case loss over all transformations and redistributions, rather than an average or sum of losses.

Naïve Augmentation vs. Dual-View. A direct solution is to augment S by all transformations $\{T(x_i)\}$, then optimize $\max_{Q:\,D_f(Q\|\mathbf{u})\leq\rho} \sum p_i\,\ell\big(h,\ \cdot\big)$ over the augmented dataset (with uniform baseline $\mathbf{u}$). However, if $|\mathcal{T}|$ and m are large, the dimension $|\mathcal{T}| \times m$ can be prohibitive. We therefore adopt a *dual-view* strategy with a two-level adversarial approach. This approach maintains a distribution $\boldsymbol{\pi}$ over transformations to identify the worst-case $T \in \mathcal{T}$, while simultaneously assigning a weight vector $\mathbf{p}_T$ for each transformation, constrained by $D_f\big(\mathbf{p}_T \|\frac{1}{m}\mathbf{1}\big) \leq \rho$.

The algorithm proceeds through alternating optimization steps. First, an inner adversary updates each $\mathbf{p}_T$ to maximize the loss under its respective transformation T. Next, an outer adversary updates $\boldsymbol{\pi}$ to concentrate mass on the worst-case transformations. Finally, the learner updates h to minimize the weighted loss $\sum_T \pi_T \sum_i p_{T,i} \ell\big(h, T(x_i), y_i\big)$. This nested structure effectively implements the min-max objective while keeping computation tractable.

Theoretical Guarantees. By standard uniform convergence arguments, our solution (1) guarantees

$$\mathrm{OPT}_\infty = \inf_h \sup_{T \in \mathcal{T}} \sup_{Q:\, D_f(Q \| \, T(\mathcal{D})) \le \rho} \mathbb{E}_Q\big[\ell(h)\big] \approx \min_h \max_{T \in \mathcal{T}} \max_{\substack{Q: \\ D_f(Q \| \, T(P_S)) \le \rho}} \mathbb{E}_Q\big[\ell(h)\big]. \tag{2}$$

provided m is large enough. Section 5 details sample-complexity bounds that combine $\mathrm{VC}(\mathcal{H} \circ \mathcal{T})$ with an additional term $\Psi(\rho)$, reflecting the divergence radius. Notably, when $\rho \to 0$, we recover pure transformation-based learning; while for large ρ, we protect against strong reweighting but need more samples.

4.1 Dual-View Algorithm: Per-Transformation Weights

The naive approach in §3 can be computationally expensive when both $|\mathcal{T}|$ and m are large, since it requires handling an adversarial weight vector of dimension $|\mathcal{T}| \times m$. Here we describe a more efficient "dual-view" algorithm that assigns one weight vector per transformation and updates those weights in parallel, while also maintaining a distribution over transformations to focus on the worst cases. This method extends two-player mirror-descent (Namkoong & Duchi, 2016) to a three-player game with nested adversaries.

Algorithm Sketch. Our algorithm maintains two key distributions throughout the optimization process. First, for each transformation $T_j \in \mathcal{T}$, we maintain a distribution $\mathbf{p}^{(j)} = (p_i^{(j)})_{i=1}^m$ over the training samples, constrained such that $D_f(\mathbf{p}^{(j)} \,\|\, \frac{1}{m}\mathbf{1}) \le \rho$. Second, we maintain a distribution $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_{|\mathcal{T}|})$ over the transformations themselves, which helps identify the worst-case transformations.

Initially, each $\mathbf{p}^{(j)}$ is set to the uniform vector $\frac{1}{m}\mathbf{1}$, and $\boldsymbol{\pi}$ is uniform over $\mathcal{T}$. At iteration $r = 1, 2, \ldots, R$, three updates occur:

(1) Inner Adversary update. For each $T_j \in \mathcal{T}$, run a mirror-ascent step on the loss

$$\sum_{i=1}^m p_i^{(j)} \, \ell\big(h_r,\; T_j(x_i),\; y_i\big),$$

then project back onto the feasible set $\{\mathbf{p}^{(j)} : D_f(\mathbf{p}^{(j)} \| \frac{1}{m}\mathbf{1}) \le \rho\}$. This step increases the weights on points that most expose h_r's weaknesses under transformation T_j. We then compute the total loss for each transformation: $g_j(h_r) = \sum_{i=1}^m p_i^{(j)} \, \ell\big(h_r,\; T_j(x_i),\; y_i\big)$.

(2) Outer Adversary update. Update the distribution $\boldsymbol{\pi}$ over transformations using exponentiated gradient:

$$\pi_j^{(r+1)} \propto \pi_j^{(r)} \exp(\eta_r \cdot g_j(h_r))$$

followed by normalization. This update increases weight on transformations that produce high loss, effectively focusing on the worst-case transformations.

(3) Learner update. Once all weights are updated, perform a gradient-descent step on

$$\sum_{j=1}^{|\mathcal{T}|} \pi_j^{(r+1)} \sum_{i=1}^m p_i^{(j)} \, \ell\big(h,\; T_j(x_i),\; y_i\big),$$

yielding a new hypothesis h_{r+1}. This move counteracts both adversaries by improving h on the transformations and data points that have been emphasized.

After R iterations, the algorithm returns the final hypothesis h_R or an average over all $\{h_r\}_{r=1}^R$. This "dual-view" approach with the nested adversarial structure properly implements the min-max-max objective while keeping computational complexity manageable. Each inner adversary update remains a mirror-ascent step on m coordinates, and the outer adversary update is a simple exponentiated gradient step on $|\mathcal{T}|$ coordinates, making the overhead comparable to standard DRO approaches.

5 Theoretical Analysis

In our theoretical analysis the term $\Psi(\rho, m)$ plays a central role by quantifying the additional complexity introduced by allowing adversarial reweightings within an f-divergence ball of radius ρ. In this section, we provide a formal definition of $\Psi(\rho, m)$ along with explicit bounds under standard regularity assumptions on the divergence function.

5.1 Assumptions on the Divergence Function

Assumption 1 (Regularity of f)**.** *Let $f : \mathbb{R}_{>0} \to \mathbb{R}$ be a convex function with $f(1) = 0$. We assume that f is continuously differentiable and strongly convex in a neighborhood of 1; that is, there exist constants $\mu > 0$ and $\epsilon_0 > 0$ such that for all $u \in (1 - \epsilon_0, 1 + \epsilon_0)$,*

$$f(u) \geq f(1) + f'(1)(u-1) + \frac{\mu}{2}(u-1)^2.$$

This assumption holds for standard divergences such as the Kullback–Leibler divergence and the χ^2-divergence under moderate parameter settings.

5.2 Definition of $\Psi(\rho, m)$

Given a sample $S = \{(x_i, y_i)\}_{i=1}^m$, denote by

$$u = \left(\frac{1}{m}, \frac{1}{m}, \dots, \frac{1}{m}\right)$$

the uniform distribution over the m points. For any probability vector $p \in \Delta^m$ (the m-simplex), define the f-divergence

$$D_f(p \| u) = \sum_{i=1}^m u_i \, f\Big(\frac{p_i}{u_i}\Big).$$

We then consider the set of all reweightings that are within divergence ρ of the uniform distribution:

$$\mathcal{P}(\rho) = \Big\{ p \in \Delta^m \;\Big|\; D_f(p \| u) \leq \rho \Big\}.$$

For a given accuracy $\varepsilon > 0$ (with respect to the ℓ_1-norm), let $\mathcal{N}(\mathcal{P}(\rho), \|\cdot\|_1, \varepsilon)$ denote the covering number of $\mathcal{P}(\rho)$. We define the *divergence complexity* as

$$\Psi(\rho, m) := \inf_{\varepsilon \in (0,1)} \log \mathcal{N}\Big(\mathcal{P}(\rho), \|\cdot\|_1, \varepsilon\Big).$$

This definition captures the effective number of degrees of freedom available to an adversary under the divergence constraint.

5.3 Bounding $\Psi(\rho, m)$

Lemma 1 (Bound on Divergence Complexity)**.** *Under Assumption 1, there exist constants $C_1, C_2 > 0$ (depending on μ and on properties of f) such that for all ρ satisfying $0 \leq \rho \leq \rho_0$ (for some $\rho_0 > 0$),*

$$\Psi(\rho, m) \leq C_1 \, m \, \rho + C_2 \ln m.$$

In particular, when $\rho = o(1)$, the additional complexity $\Psi(\rho, m)$ grows at most sublinearly in m.

This bound shows that for moderate divergence radii, the increase in sample complexity due to adversarial reweighting is controlled.

5.4 Near-Optimality Lower Bound

We now show that the sample complexity bound cannot be substantially improved. In particular, we prove that *any* learner must require on the order of $\Omega\Big(\frac{\mathrm{VC}(\mathcal{H} \circ \mathcal{T}) + \Psi(\rho)}{\varepsilon^2}\Big)$ samples to achieve the same worst-case error bound, thereby demonstrating near-optimality.

Theorem 1 (Near-Optimality Lower Bound)**.** *For every integer m, there exist a distribution $\mathcal{D}$, a hypothesis class $\mathcal{H}$, a transformation family $\mathcal{T}$, and a radius $\rho > 0$ such that, with probability at least $1/4$ over $S \sim \mathcal{D}^m$,*

$$\inf_{\textit{learning rule } \mathcal{A}} \; \sup_{T \in \mathcal{T}} \; \sup_{Q:\, D_f(Q \,\|\, T(\mathcal{D})) \leq \rho} \; \mathbb{E}_Q\big[\ell\big(\mathcal{A}(S),\, x,\, y\big)\big] \;\geq\; \mathrm{OPT}_\infty \;+\; c$$

for some absolute constant $c > 0$. Equivalently, to attain risk less than $\mathrm{OPT}_\infty + c/2$ with high probability, any *rule $\mathcal{A}$ requires $m = \Omega\Big(\frac{\mathrm{VC}(\mathcal{H} \circ \mathcal{T}) + \Psi(\rho)}{\varepsilon^2}\Big)$.*

Sketch of Proof We adapt a standard VC argument to construct $d = \mathrm{VC}(\mathcal{H} \circ \mathcal{T})$ points $\{x_1, \dots, x_d\}$ that can be shattered by $\mathcal{H} \circ \mathcal{T}$. By choosing a random labeling pattern α and an appropriate label-preserving transformation $T_\alpha \in \mathcal{T}$, we form a distribution $\mathcal{D}_\alpha$ that lies within f-divergence ρ of $T_\alpha(\mathcal{D})$, yet forces any single hypothesis $h \in \mathcal{H}$ to misclassify at least one point x_i. If the sample size m is too small, the learner cannot reliably identify which α has been chosen, and so with constant probability it will pick a hypothesis $\hat{h}$ that incurs extra error on $\mathcal{D}_\alpha$. Hence, no algorithm can guarantee near-optimal worst-case risk below $\mathrm{OPT}_\infty + c$, implying the claimed Ω bound on m.

Algorithm 1 Mirror-Descent for Adversarial Reweighting with Transformations

```
Input: A sample $S \sim \mathcal{D}^m$, transformations $\mathcal{T}$, radius $\rho$, loss $\ell$, etc.
Initialize: Choose $h^{(1)} \in \mathcal{H}$. For each $T \in \mathcal{T}$, let $\mathbf{p}_T^{(1)}$ be uniform over $\{T(x_i), y_i\}_{i=1}^m$.
Initialize distribution $\boldsymbol{\pi}^{(1)}$ to be uniform over $\mathcal{T}$.
Fix a step-size schedule, e.g. $\eta_r = c/\sqrt{r}$ for some $c > 0$.
for $r = 1$ to $R$ do
    for $T \in \mathcal{T}$ do
        (Inner Adversary) Update $\mathbf{p}_T^{(r)}$ via mirror-ascent on $\sum_i p_{T,i}^{(r)}\, \ell\big(h^{(r)}, T(x_i), y_i\big)$
        Project onto $\{\mathbf{p}_T^{(r+1)} : D_f(\mathbf{p}_T^{(r+1)} \| T(P_S)) \le \rho\}$.
        Compute transformation loss $g_T(h^{(r)}) = \sum_i p_{T,i}^{(r+1)} \ell(h^{(r)}, T(x_i), y_i)$
    end for
    (Outer Adversary) Update $\boldsymbol{\pi}^{(r+1)}$ via exponentiated gradient:
    $\pi_T^{(r+1)} \propto \pi_T^{(r)} \exp(\eta_r \cdot g_T(h^{(r)}))$ for each $T \in \mathcal{T}$, then normalize
    (Learner step) Update $h^{(r+1)}$ by minimizing $\sum_{T \in \mathcal{T}} \pi_T^{(r+1)} \sum_i p_{T,i}^{(r+1)} \ell(h, T(x_i), y_i)$.
end for
Output: $\overline{h} := \frac{1}{R}\sum_{r=1}^{R} h^{(r)}, \quad \overline{\mathbf{p}}_T := \frac{1}{R}\sum_{r=1}^{R} \mathbf{p}_T^{(r)}, \ \forall T \in \mathcal{T}, \quad \overline{\boldsymbol{\pi}} := \frac{1}{R}\sum_{r=1}^{R} \boldsymbol{\pi}^{(r)}.$
```

Theorem 2 (Approximate Saddle-Point Guarantee)**.** *Under convexity of ℓ in h and appropriate conditions on $D_f(\cdot\|\cdot)$ for mirror-ascent, the final triple $(\overline{h}, \overline{\boldsymbol{\pi}}, \{\overline{\mathbf{p}}_T\})$ produced by Algorithm 1 is an $O\big(R^{-\frac{1}{2}}\big)$-approximate saddle point with high probability (over the sample S and any algorithmic randomness). Concretely, there exists a function $\epsilon(R) = O\big(R^{-\frac{1}{2}}\big)$ such that*

$$
\begin{aligned}
&\max_{T \in \mathcal{T}} \left(\max_{Q:\, D_f(Q \,\|\, T(P_S)) \,\le\, \rho} \mathbb{E}_Q\big[\ell(\overline{h}, x, y)\big] \right) \\
&\qquad \le \min_{h \in \mathcal{H}} \max_{T \in \mathcal{T}} \left(\max_{Q:\, D_f(Q \,\|\, T(P_S)) \,\le\, \rho} \mathbb{E}_Q\big[\ell(h, x, y)\big] \right) + \epsilon(R).
\end{aligned}
\tag{3}
$$

Sketch of Proof Since $\ell(\cdot)$ is convex in h and the feasible sets for $\boldsymbol{\pi}$ and each $\mathbf{p}_T$ are convex under their respective constraints, we can apply a two-level mirror-descent analysis. For the outer level, exponentiated gradient updates on $\boldsymbol{\pi}$ ensure convergence to the distribution that places weight on the worst-case transformations. For the inner level, mirror-descent on each $\mathbf{p}_T$ converges to the worst-case reweighting within the f-divergence ball. Averaging the iterates across all three variables yields an $O(R^{-1/2})$ approximation to the nested min-max-max solution.

6 Experiments and Results

We construct a toy distribution in $\mathbb{R}^2$ consisting of two Gaussian clusters (one per class). To induce structured perturbations, we apply two label-preserving transformations: (i) a mild rotation by $15°$ and (ii) a reflection across the x-axis. Additionally, an adversary can reweight transformed samples within a KL-divergence ball of radius $\rho > 0$. We evaluate three strategies: *(a) Transform-Only*, *(b) Divergence-Only*, and *(c) STAR (Ours)*.

Results. Figure 1 illustrates two key aspects of our analysis: **(a)** a phase-transition plot for sample complexity and **(b)** an empirical error comparison across different sample sizes and divergence radii. As ρ grows, we see a clear boundary between sufficient and insufficient sample regimes, with higher ρ requiring more data to maintain low error. Larger sample sizes ($m = 500 \rightarrow 2000$) reduce error,

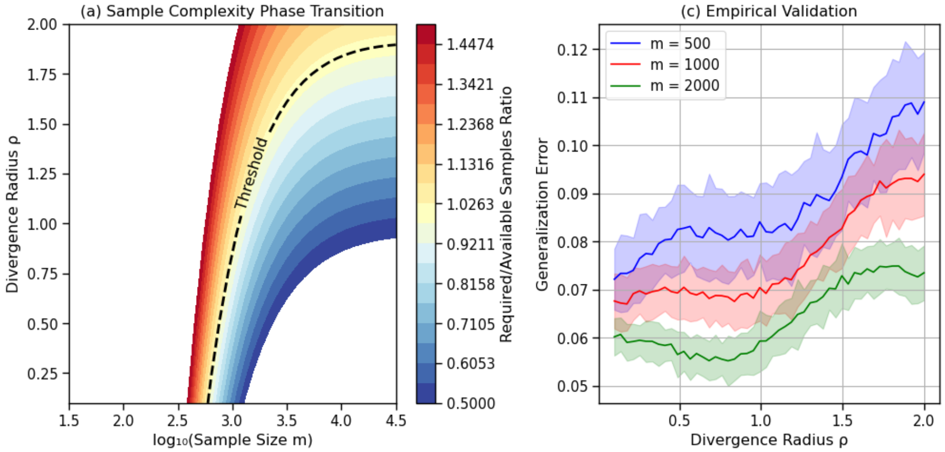

Figure 1: (a) Sample complexity phase transition diagram showing the ratio of required to available samples across different sample sizes and divergence radii. The dashed line indicates the threshold where the ratio equals 1.0. (b) Empirical validation showing generalization error versus divergence radius for different sample sizes (m). Shaded regions represent one standard deviation over multiple trials. The decreasing error with larger sample sizes and graceful degradation with increasing divergence radius validate our theoretical analysis.

but increasing ρ naturally degrades performance. Table 1 shows worst-case errors (over 5 trials) at $\rho = 0.5$ for three methods: *Transform* (label-preserving transformations only), *Divergence* (f-divergence–based reweighting only), and *STAR* (our method). STAR yields the lowest mean error (7.58%) and similar variance, outperforming both baselines.

Table 1: Worst-case test error (%) at $\rho = 0.5$ over 5 trials. Lower is better.

Method	**Mean Error**	**Std. Dev.**
Transform	9.83	0.83
Divergence	8.94	0.77
STAR	7.58	0.79

Key Insights. Our analysis reveals a fundamental trade-off: while nesting transformations within f-divergence balls theoretically requires more samples (as evidenced by the phase transition), it empirically delivers better worst-case performance. The phase transition diagram shows that this increased sample complexity scales predictably with ρ, allowing practitioners to make informed choices about the robustness-complexity trade-off. The empirical results demonstrate that this theoretical cost is justified by the improvements in robustness, particularly as the divergence radius increases.

7 Conclusion

We present a robust learning framework that addresses both structured transformations and adversarial reweightings. By accommodating label-preserving transformations alongside an f-divergence constraint, our approach incorporates domain knowledge in tandem with distributionally robust optimization. The uniform-convergence analysis provides explicit sample-complexity bounds, indicating that this added flexibility may require more data but yields robust generalization over a wider range of potential shifts. Empirical evaluations confirm the method's effectiveness, outperforming baseline techniques.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Omar Montasser, Han Shao, and Emmanuel Abbe. Transformation-invariant learning and theoretical guarantees for ood generalization. *arXiv preprint arXiv:2410.23461*, 2024.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in neural information processing systems*, 29, 2016.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.