Bitrate-Constrained DRO: Beyond Worst Case Robustness To Unknown Group Shifts

Anonymous Author(s)
Affiliation
Address
email

Abstract

Although training machine learning models for robustness is critical for real-world adoption, determining how to best ensure robustness remains an open problem. Some methods (*e.g.,* DRO) are overly conservative, while others (*e.g.,* Group DRO) require domain knowledge that may be hard to obtain. In this work, we address limitations in prior approaches by assuming a more nuanced form of group shift: conditioned on the label, we assume that the true group function is *simple*. For example, we may expect that group shifts occur along high-level features (*e.g.,* image background, lighting). Thus, we aim to learn a model that maintains high accuracy on simple group functions realized by these features, but need not spend valuable model capacity achieving high accuracy on contrived groups of examples. Based on this idea, we formulate a two-player game where conditioned on the label the adversary can only separate datapoints into potential groups using simple features, which corresponds to a bitrate constraint on the adversary's capacity. Our resulting practical algorithm, Bitrate-Constrained DRO (`BR-DRO`), does not require group information on training samples yet matches the performance of Group DRO. Our theoretical analysis reveals that in some settings `BR-DRO` objective can provably yield statistically efficient and less pessimistic solutions than unconstrained DRO.

1 Introduction

A common form of distribution shift is *group* shift, where the source and target differ only in the marginal distribution over finite groups or sub-populations, with no change in group conditionals [43, 18, 46]. Prior works consider various approaches to address group shift. One solution is to ensure robustness to worst case shifts using distributionally robust optimization (DRO) [4, 7, 17], which considers a two-player game where a learner minimizes risk on distributions chosen by an adversary from a predefined uncertainty set. As the adversary is unconstrained in proposing distributions, DRO often yields overly pessimistic solutions [25] and can suffer from statistical challenges [18]. Methods like Group DRO [46] avoid overly pessimistic solutions by assuming knowledge of group membership for each training example. However, these group-based methods provide no guarantees on shifts that deviate from the predefined groups, and are not applicable to problems that lack group knowledge. In this work, we therefore ask: *Can we train non-pessimistic robust models without access to group annotations on training examples?*

We address this question by considering a more nuanced assumption on the structure of the underlying groups. We assume that, conditioned on the label, group boundaries are realized by high-level features that depend on a small set of underlying factors. This leads to simpler group functions with large margins and simple decision boundaries (Figure 1 *(left)*). Invoking the principle of minimum description length [21], restricting our adversary to functions that satisfy this assumption corresponds to a bitrate constraint. In DRO, the adversary upweights points with higher losses under the current

Submitted to the Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

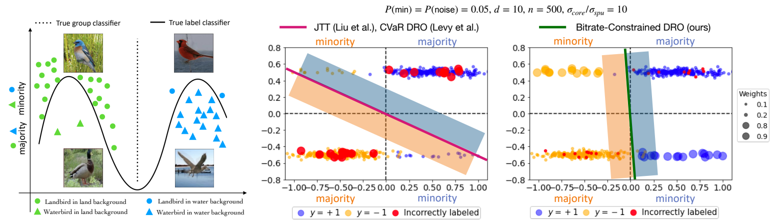


Figure 1: **Bitrate-Constrained DRO**: A method that assumes group shifts along low-bitrate features, and restricts the adversary appropriately so that the solution found is less pessimistic and more robust to group shifts. Our method is also robust to training noise. *(Left)* In Waterbirds [54], the spurious feature background is a large margin simple feature that separates the *majority* and *minority* points in each class. *(Right)* Prior works [31, 35] that upweight arbitrary points with high losses force the model to memorize noisy mislabeled points while our method is robust to noise and only upweights the true minority group without any knowledge of its identity.

learner, which in practice often correspond to examples that belong to a rare group, contain complex patterns, or are mislabeled [14, 53]. Restricting the adversary's capacity prevents it from upweighting individual hard or mislabeled examples (as they cannot be identified with simple features), and biases it towards identifying erroneous data points misclassified by simple features. This also complements the failure mode of neural networks trained with stochastic gradient descent (SGD) that rely on simple spurious features that correctly classify points in the *majority* group but may fail on *minority* groups [10].

The main contribution of this paper is Bitrate-Constrained DRO (`BR-DRO`), a supervised learning procedure that provides robustness to distribution shifts along groups realized by simple functions. Despite not using group information on training examples, we demonstrate that `BR-DRO` can match the performance of methods requiring them. We also find that `BR-DRO` correctly identifies true minority points, whereas DRO without group information does not. This indicates that not optimizing for performance on contrived worst-case shifts can reduce the pessimism inherent in DRO. It further validates: (i) our assumption on the simple nature of group shift; and (ii) that our method of capacity control meaningfully structures the uncertainty set to be robust to such shifts. As a consequence of the constraint, we also find that `BR-DRO` is robust to random noise in the training data [51], since it cannot form "groups" entirely based on randomly mislabeled points with low bitrate features. This is in contrast with existing methods that use the learner's training error to up-weight arbitrary sets of difficult training points [*e.g.,* 35, 31], which we show are highly susceptible to label noise (see Figure 1). Finally, we theoretically analyze our approach—characterizing how the degree of constraint on the adversary can effect worst risk estimation and excess risk (pessimism) bounds, as well as convergence rates for specific online solvers.

2 Bitrate-Constrained DRO

Notation. With covariates $\mathcal{X} \subset \mathbb{R}^d$ and labels $\mathcal{Y}$, the given source P and unknown true target Q_0 are measures over the measurable space $(\mathcal{X} \times \mathcal{Y}, \Sigma)$ and have densities p and q_0 respectively (w.r.t. base measure μ). The learner's choice is a hypothesis $h : \mathcal{X} \mapsto \mathcal{Y}$ in class $\mathcal{H} \subset L^2(P)$, and the adversary's action in standard DRO is a target distribution Q in set $\mathcal{Q}_{P,\kappa} := \{Q : Q \ll P,\ D_f(Q \,||\, P) \leq \kappa\}$. Here, D_f is the f-divergence between Q and P for a convex function f[1] with $f(1) = 0$. An equivalent action space for the adversary is the set of re-weighting functions:

$$\mathcal{W}_{P,\kappa} = \{w : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R} :\ w \text{ is measurable under } P,\ \mathbb{E}_P[w] = 1,\ \mathbb{E}_P f(w) \leq \kappa\} \tag{1}$$

For a convex loss function $l : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}_+$, we denote $l(h)$ as the function over $(\mathbf{x}, \mathrm{y})$ that evaluates $l(h(\mathbf{x}), \mathrm{y})$, and use l_{0-1} to denote the loss function $\mathbb{1}(h(\mathbf{x}) \neq \mathrm{y})$. Given either distribution $Q \in \mathcal{Q}_{P,\kappa}$, or a re-weighting function $w \in \mathcal{W}_{P,\kappa}$, the risk of a learner h is:

$$R(h, Q) = \mathbb{E}_Q\ [l(h)] \qquad R(h, w) = \mathbb{E}_{(\mathbf{x},\mathrm{y})\sim P}\ [l(h(\mathbf{x}), \mathrm{y}) \cdot w(\mathbf{x}, \mathrm{y})] = \langle l(h),\ w \rangle_P \tag{2}$$

Note the overload of notation for $R(h, \cdot)$. If the adversary is stochastic it picks a mixed action $\delta \in \Delta(\mathcal{W}_{P,\kappa})$, which is the set of all distributions over $\mathcal{W}_{P,\kappa}$. Whenever it is clear, we drop P, κ.

Unconstrained DRO [7]. This is a min-max optimization problem understood as a two-player game, where the learner chooses a hypothesis, to minimize risk on the worst distribution that the adversary can choose from its set. Formally, this is given by equation 3. The first equivalence is clear from the definitions and for the second since $R(h, Q)$ is linear in Q, the supremum over $\Delta(\mathcal{W}_{P,\kappa})$ is a Dirac

[1]For *e.g.,* $\mathrm{KL}(Q \,||\, P)$ can be derived with $f(x) = x \log x$ and for Total Variation $f(x) = |x - 1|/2$.

delta over the best weighting in $\mathcal{W}_{P,\kappa}$. In the next section, we will see how a bitrate-constrained adversary can only pick certain actions from $\Delta(\mathcal{W}_{P,\kappa})$.

$$\inf_{h\in\mathcal{H}} \sup_{Q\in\mathcal{Q}_{P,\kappa}} R(h,Q) \equiv \inf_{h\in\mathcal{H}} \sup_{w\in\mathcal{W}_{P,\kappa}} R(h,w) \equiv \inf_{h\in\mathcal{H}} \sup_{\delta\in\Delta(\mathcal{W}_{P,\kappa})} \mathbb{E}_{w\sim\delta}\left[R(h,w)\right] \quad (3)$$

Group Shift. While the DRO framework is broad and addresses any unstructured shift, we focus on the specific case of group shift. First, for a given pair of measures P, Q we define what we mean by the group structure $\mathcal{G}_{P,Q}$ (Definition 2.1). Intuitively, it is a set of sub-populations along which the distribution shifts, defined in a way that makes them uniquely identifiable. For *e.g.,* in the Waterbirds dataset (Figure 1), there are four groups given by combinations of (label, background). Corollary 2.2 follows immediately from the definition of $\mathcal{G}_{P,Q}$. Using this definition, the standard group shift assumption [46] can be formally re-stated as Assumption 2.3.

Definition 2.1 (group structure $\mathcal{G}_{P,Q}$)**.** *For $Q \ll P$ the group structure $\mathcal{G}_{P,Q}=\{G_k\}_{k=1}^K$ is the smallest finite set of disjoint groups $\{G_k\}_{k=1}^K$ s.t. $Q(\cup_{k=1}^K G_k)=1$ and $\forall k$ (i) $G_k \in \Sigma$, $Q(G_k) > 0$ and (ii) $p(\mathbf{x}, \mathrm{y} \mid G_k) = q(\mathbf{x}, \mathrm{y} \mid G_k) > 0$ a.e. in μ. If such a structure exists then $\mathcal{G}_{P,Q}$ is well defined.*

Corollary 2.2 (uniqueness of $\mathcal{G}_{P,Q}$)**.** *$\forall P, Q$, the structure $\mathcal{G}(P,Q)$ is unique if it is well defined.*

Assumption 2.3 (standard group shift)**.** *There exists a well-defined group structure $\mathcal{G}_{P,Q_0}$ s.t. target Q_0 differs from P only in terms of marginal probabilities over all $G \in \mathcal{G}_{P,Q_0}$.*

How expressive is unconstrained adversary? Note that the set $\mathcal{W}_{P,\kappa}$ includes all measurable functions (under P) such that the re-weighted distribution is bounded in f-divergence (by κ). While prior works [48, 17] shrink κ to construct confidence intervals, this *only controls* the total mass that can be moved between measurable sets $G_1, G_2 \in \Sigma$, but *does not restrict* the choice of G_1 and G_2 itself. As noted by Hu et al. [25], such an adversary is highly expressive, and optimizing for the worst case only leads to the solution of empirical risk minimization (ERM) under l_{0-1} loss. Thus, we can conclude that DRO recovers degenerate solutions because the worst target in $\mathcal{W}_{P,\kappa}$ lies far from the subspace of naturally occurring targets. Since it is hard to precisely characterize natural targets we make a nuanced assumption: the target Q_0 only upsamples those rare subpopulations that are misclassified by simple features. We state this formally in Assumption 2.5 after we define the bitrate-constrained function class $\mathcal{W}(\gamma)$ in Definition 2.4. See Appendix A for additional discussion on when/why constraining capacity helps with distribution shift robustness.

Definition 2.4. *A function class $\mathcal{W}(\gamma)$ is bitrate-constrained if there exists a data independent prior π, s.t. $\mathcal{W}(\gamma) = \{\mathbb{E}[\delta] : \delta \in \Delta(\mathcal{W}),\ KL(\delta \mid\mid \pi) \leq \gamma\}$.*

Assumption 2.5 (simple group shift)**.** *Target Q_0 satisfies Assumption 2.3 (group shift) w.r.t. source P. Additionally, for some prior π and a small γ^*, the re-weighting function q_0/p lies in a bitrate-constrained class $\mathcal{W}(\gamma^*)$. In other words, for every group $G \in \mathcal{G}(P, Q_0)$, $\exists w_G \in \mathcal{W}(\gamma^*)$ s.t. $\mathbb{1}((\mathbf{x}, \mathrm{y}) \in G) = w_G$ a.e.. We refer to such a G as a **simple group** that is realized in $\mathcal{W}(\gamma^*)$.*

`BR-DRO` **objective.** According to Assumption 2.5, there cannot exist a target Q_0 such that minority $G_{\text{min}} \in \mathcal{G}(P, Q_0)$ is not realized in bitrate constrained class $\mathcal{W}(\gamma^*)$. Thus, by constraining our adversary to a class $\mathcal{W}(\gamma)$ (for some γ that is user defined), we can possibly evade issues emerging from optimizing for performance on mislabeled or hard examples, even if they were rare. This gives us the objective in Equation 4 where the equalities hold from the linearity of $\langle\cdot,\cdot\rangle$ and Definition 2.4. See Appendix A.1 for details on the practical implementation of `BR-DRO`.

$$\inf_{h\in\mathcal{H}} \sup_{\substack{\delta\in\Delta(\mathcal{W}) \\ \mathrm{KL}(\delta \mid\mid \pi)\leq\gamma}} \mathbb{E}_{w\sim\delta} R(h,w) = \inf_{h\in\mathcal{H}} \sup_{\substack{\delta\in\Delta(\mathcal{W}) \\ \mathrm{KL}(\delta \mid\mid \pi)\leq\gamma}} \langle l(h), \mathbb{E}_\delta[w]\rangle_P = \inf_{h\in\mathcal{H}} \sup_{w\in\mathcal{W}(\gamma)} R(h,w) \quad (4)$$

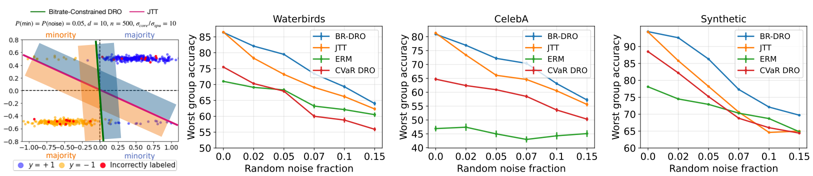
Theoretical Analysis. The main objective of our analysis of `BR-DRO` is to show how adding a bitrate constraint on the adversary can: (i) give us tighter statistical estimates of the worst risk; and (ii) control the pessimism (excess risk) of the learned solution. First, we provide worst risk generalization guarantees using the PAC-Bayes framework [15], along with a result for kernel adversary. Then, we discuss convergence rates and pessimism guarantees for the solution found by our online solver for a specific instance of $\mathcal{W}(\gamma)$. See Appendix B for details.

3 Experiments

We discuss two sets of experiments here on robustness to spurious correlations and random label noise. For more details on these and other experiments please refer to Appendix C.

Method	Waterbirds Avg	Waterbirds WG	CelebA Avg	CelebA WG	CivilComments Avg	CivilComments WG
ERM	97.1 (0.1)	71.0 (0.4)	95.4 (0.2)	46.9 (1.0)	92.3 (0.2)	57.2 (0.9)
LfF [41]	90.7 (0.2)	77.6 (0.5)	85.3 (0.2)	77.4 (0.7)	92.4 (0.1)	58.9 (1.1)
RWY [26]	93.7 (0.3)	85.8 (0.5)	84.9 (0.2)	80.4 (0.3)	91.7 (0.2)	67.7 (0.7)
JTT [35]	93.2 (0.2)	86.6 (0.4)	87.6 (0.2)	81.3 (0.5)	90.8 (0.3)	69.4 (0.8)
CVaR DRO [31]	96.3 (0.2)	75.5 (0.4)	82.2 (0.3)	64.7 (0.6)	92.3 (0.2)	60.2 (0.8)
`BR-DRO` (VIB) (ours)	94.1 (0.2)	86.3 (0.3)	86.7 (0.2)	80.9 (0.4)	90.5 (0.2)	68.7 (0.9)
`BR-DRO` (l_2) (ours)	93.8 (0.2)	86.4 (0.3)	87.7 (0.3)	80.4 (0.6)	91.0 (0.3)	68.9 (0.7)
Group DRO [46]	93.2 (0.3)	91.1 (0.3)	92.3 (0.3)	88.4 (0.6)	88.5 (0.3)	70.0 (0.5)

Table 1: `BR-DRO` **recovers worst group performance gap between CVaR DRO and Group DRO:** On Waterbirds, CelebA and CivilComments we report test average (Avg) and test worst group (WG) accuracies for `BR-DRO` and baselines. In $(\cdot)$ we report the standard error of the mean accuracy across five runs.



(a) (b)

Figure 2: *(Left)* **Visualization (2d) of noisy synthetic data and learned predictors:** We compare the decision boundaries (projected onto core and spurious features) learned by JTT with `BR-DRO` when the adversary is restricted to a sparse predictor. While our method recovers the core feature the baselines memorize the minority points. *(Right)* `BR-DRO` **is robust to random label noise in training data:** Across varying levels of the fraction of noise in training data we compare performance of `BR-DRO` with ERM and methods (JTT, CVaR DRO) that naively up weight high loss datapoints.

Is `BR-DRO` robust to group shifts without training data group annotations? Table 1 compares the average and worst group accuracy for `BR-DRO` with ERM and four group shift robustness baselines: JTT, LtF, SUBY, and CVaR DRO. First, we see that unconstrained CVaR DRO underperforms other heuristic algorithms. This matches the observation made by Liu et al. [35]. Next, we see that adding bitrate constraints on the adversary via a KL term or l_2 penalty significantly improves the performance of `BR-DRO` (VIB) or `BR-DRO` (l_2), which now matches the best performing baseline (JTT). Thus, we see the less conservative nature of `BR-DRO` allows it to recover a large portion of the performance gap between Group DRO and CVaR DRO. Indirectly, this partially validates our Assumption 2.5, which states that the minority group is identified by a low bitrate adversary class. In Section C.3 we discuss exactly what fraction of the minority group is identified, and the role played by the strength of bitrate-constraint.

Bitrate DRO is more robust to random label noise. Several methods for group robustness (*e.g.,* CVaR DRO, JTT) are based on the idea of up weighting points with high training losses. The goal is to obtain a learner with matching performance on every (small) fraction of points in the dataset. However, when training data has mislabeled examples, such an approach will likely yield degenerate solutions. This is because the adversary directly upweights any example where the learner has high loss, including datapoints with incorrect labels. Hence, even if the learner's prediction matches the (unknown) true label, this formulation would force the learner to memorize incorrect labelings at the expense of learning the true underlying function. On the other hand, if the adversary is sufficiently bitrate constrained, it cannot upweight the arbitrary set of randomly mislabeled points, as this would require it to memorize those points. Our Assumption 2.5 also dictates that the distribution shift would not upsample such high bitrate noisy examples. Thus, our constraint on the adversary ensures `BR-DRO` is robust to label noise in the training data and our assumption on the target distribution retains its robustness to test time distribution shifts. In Figure 2b we highlight this failure mode of unconstrained up-weighting methods in contrast to `BR-DRO`. We first induce random label noise [14] of varying degrees into the Waterbirds and CelebA training sets. Then we run each method and compare worst group performance. In the presence of noise, `BR-DRO` significantly outperforms JTT and other approaches on both Waterbirds and CelebA, as it only upsamples the minority examples misclassified by simple features, ignoring the noisy examples for the reasons above. See Appendix C.1 for more details on experiments with synthetic data.

References

[1] Abernethy, J., Lai, K. A., Levy, K. Y., and Wang, J.-K. (2018). Faster rates for convex-concave games. In *Conference On Learning Theory*, pages 1595–1625. PMLR.

[2] Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. (2016). Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*.

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

[4] Bagnell, J. A. (2005). Robust supervised learning. In *AAAI*, pages 714–719.

[5] Bao, Y. and Barzilay, R. (2022). Learning to split for automatic bias detection. *arXiv preprint arXiv:2204.13749*.

[6] Bartlett, P. L., Kulkarni, S. R., and Posner, S. E. (1997). Covering numbers for real-valued function classes. *IEEE transactions on information theory*, 43(5):1721–1724.

[7] Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., and Rennen, G. (2013). Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357.

[8] Bertsimas, D., Gupta, V., and Kallus, N. (2018). Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292.

[9] Blanchet, J. and Murthy, K. (2019). Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600.

[10] Blodgett, S. L., Green, L., and O'Connor, B. (2016). Demographic dialectal variation in social media: A case study of african-american english. *arXiv preprint arXiv:1608.08868*.

[11] Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. (2019). Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.

[12] Boyd, S., Boyd, S. P., and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

[13] Byrd, J. and Lipton, Z. (2019). What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, pages 872–881. PMLR.

[14] Carlini, N., Erlingsson, U., and Papernot, N. (2019). Distribution density, tails, and outliers in machine learning: Metrics and applications. *arXiv preprint arXiv:1910.13427*.

[15] Catoni, O. (2007). Pac-bayesian supervised classification: the thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*.

[16] Creager, E., Jacobsen, J.-H., and Zemel, R. (2021). Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR.

[17] Duchi, J., Glynn, P., and Namkoong, H. (2016). Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*.

[18] Duchi, J. C., Hashimoto, T., and Namkoong, H. (2019). Distributionally robust losses against mixture covariate shifts. *Under review*, 2.

[19] Duchi, J. C. and Namkoong, H. (2021). Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378 – 1406.

[20] Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

[21] Grünwald, P. D. (2007). *The minimum description length principle*. MIT press.

[22] Gulrajani, I. and Lopez-Paz, D. (2020). In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*.

[23] Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. (2018). Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR.

[24] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[25] Hu, W., Niu, G., Sato, I., and Sugiyama, M. (2018). Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR.

[26] Idrissi, B. Y., Arjovsky, M., Pezeshki, M., and Lopez-Paz, D. (2022). Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pages 336–351. PMLR.

[27] Kearns, M., Neel, S., Roth, A., and Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR.

[28] Kirichenko, P., Izmailov, P., and Wilson, A. G. (2022). Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*.

[29] Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. (2021). Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR.

[30] Lee, Y., Yao, H., and Finn, C. (2022). Diversify and disambiguate: Learning from underspecified data. *arXiv preprint arXiv:2202.03418*.

[31] Levy, D., Carmon, Y., Duchi, J. C., and Sidford, A. (2020). Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33:8847–8860.

[32] Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. (2018). Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

[33] Lipton, Z., Wang, Y.-X., and Smola, A. (2018). Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR.

[34] Liu, A. and Ziebart, B. (2014). Robust classification under sample selection bias. *Advances in neural information processing systems*, 27.

[35] Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. (2021). Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR.

[36] Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738.

[37] Lu, Y., Ji, W., Izzo, Z., and Ying, L. (2022). Importance tempering: Group robustness for overparameterized models. *arXiv preprint arXiv:2209.08745*.

[38] Mangoubi, O. and Vishnoi, N. K. (2021). Greedy adversarial equilibrium: an efficient alternative to nonconvex-nonconcave min-max optimization. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 896–909.

[39] McAllester, D. A. (1998). Some pac-bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 230–234.

[40] Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. (2018). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993.

[41] Nam, J., Cha, H., Ahn, S., Lee, J., and Shin, J. (2020). Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684.

[42] Namkoong, H. and Duchi, J. C. (2016). Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in neural information processing systems*, 29.

[43] Oren, Y., Sagawa, S., Hashimoto, T. B., and Liang, P. (2019). Distributionally robust language modeling. *arXiv preprint arXiv:1909.02060*.

[44] Rahimian, H. and Mehrotra, S. (2019). Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*.

[45] Rosenfeld, E., Ravikumar, P., and Risteski, A. (2022). Domain-adjusted regression or: Erm may already learn features sufficient for out-of-distribution generalization. *arXiv preprint arXiv:2202.06856*.

[46] Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. (2019). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

[47] Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. (2020). An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR.

[48] Shafieezadeh Abadeh, S., Mohajerin Esfahani, P. M., and Kuhn, D. (2015). Distributionally robust logistic regression. *Advances in Neural Information Processing Systems*, 28.

[49] Shah, H., Tamuly, K., Raghunathan, A., Jain, P., and Netrapalli, P. (2020). The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33:9573–9585.

[50] Sohoni, N., Dunnmon, J., Angus, G., Gu, A., and Ré, C. (2020). No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352.

[51] Song, H., Kim, M., Park, D., Shin, Y., and Lee, J.-G. (2022). Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

[52] Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. (2018). The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878.

[53] Toneva, M., Sordoni, A., Combes, R. T. d., Trischler, A., Bengio, Y., and Gordon, G. J. (2018). An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*.

[54] Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset. *None*.

[55] Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

[56] Wang, K. A., Chatterji, N. S., Haque, S., and Hashimoto, T. (2021). Is importance weighting incompatible with interpolating classifiers? *arXiv preprint arXiv:2112.12986*.

[57] Wen, J., Yu, C.-N., and Greiner, R. (2014). Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *International Conference on Machine Learning*, pages 631–639. PMLR.

[58] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

[59] Yao, H., Wang, Y., Li, S., Zhang, L., Liang, W., Zou, J., and Finn, C. (2022). Improving out-of-distribution robustness via selective augmentation. *arXiv preprint arXiv:2201.00299*.

[60] Zhai, R., Dan, C., Suggala, A., Kolter, J. Z., and Ravikumar, P. (2021). Boosted cvar classification. *Advances in Neural Information Processing Systems*, 34:21860–21871.

[61] Zhang, Y., Duchi, J., and Wainwright, M. (2013). Divide and conquer kernel ridge regression. In *Conference on learning theory*, pages 592–617. PMLR.

Appendix

A Additional discussion on Bitrate-Constrained DRO

Note on Assumption 2.5. Under the principle of minimum description length [21] any deviation from the prior (*i.e.,* $\mathrm{KL}(\delta \mid\mid \pi)$) increases the *description length* of the encoding $\delta \in \Delta(\mathcal{W})$, thus we refer to $\mathcal{W}(\gamma)$ as being *bitrate-constrained* in the sense that it contains functions (means of distributions) that can be described with a limited number of bits given the prior π. Next we present arguments for why identifiability of simple (satisfy Assumption 2.5) minority groups can be critical for robustness.

Neural networks can perform poorly on simple minorities. For a fixed target Q_0, let's say there exists two groups: G_{min} and $G_{\text{maj}} \in \mathcal{G}(P, Q_0)$ such that $P(G_{\text{min}}) \ll P(G_{\text{maj}})$. By Assumption 2.5, both G_{min} and G_{maj} are simple (realized in $\mathcal{W}(\gamma^*)$), and are thus separated by some simple feature. The learner's class $\mathcal{H}$ is usually a class of overparameterized neural networks. When trained with stochastic gradient descent (SGD), these are biased towards learning simple features that classify a majority of the data [49, 52]. Thus, if the simple feature separating G_{min} and G_{maj} itself correlates with the label y on G_{maj}, then neural networks would fit on this feature. This is precisely the case in the Waterbirds example, where the groups are defined by whether the simple feature background correlates with the label (Figure 1). Thus our assumption on the nature of shift complements the nature of neural networks perform poorly on simple minorities.

The bitrate constraint helps identify simple unfair minorities in $\mathcal{G}(P, Q_0)$. Any method that aims to be robust on Q_0 must up-weight data points from G_{min} but without knowing its identity. Since the unconstrained adversary upsamples any group of data points with high loss and low probability, it cannot distinguish between a rare group that is realized by simple functions in $\mathcal{W}(\gamma^*)$ and a rare group of examples that share no feature in common or may even be mislabeled. On the other hand, the group of mislabeled examples cannot be separated from the rest by functions in $\mathcal{W}(\gamma^*)$. Thus, a bitrate constraint adversary can only identify simple groups and upsamples those that incur high losses – possibly due to the simplicity bias of neural networks.

A.1 Bitrate-Constrained DRO in Practice

`BR-DRO` **in practice.** We parameterize the learner $\boldsymbol{\theta}_h \in \Theta_h$ and adversary $\boldsymbol{\theta}_w \in \Theta_w$ as neural networks[2]. Therefore, the objective in Equation 5 is no longer convex-concave and can have multiple local equilibria or stationary points [38]. The adversary's objective also does not have a strong dual that can be solved through conic programs—a standard practice in DRO literature [42]. Thus, we provide an algorithm where both learner and adversary optimize `BR-DRO` iteratively through stochastic gradient ascent/descent (Algorithm 1). The adversary's action space $\mathcal{W}(\gamma)$ is constrained either with an information bottleneck penalty by setting $\beta_{\text{vib}} \neq 0$ or l_2 norm penalty by setting $\beta_{l_2} \neq 0$ in equation 5 below. While we can choose to constrain the adversary with both forms of constraints simultaneously we find that in practice picking only one of them for a given problem instance helps with tuning the degree of constraint. For more details on the architecture and other details see Appendix E.

$$\min_{\boldsymbol{\theta}_h \in \Theta_h} \langle l(\boldsymbol{\theta}_h), \boldsymbol{\theta}_w^* \rangle_P \quad \text{s.t.} \quad \boldsymbol{\theta}_w^* = \operatorname*{arg\,max}_{\boldsymbol{\theta}_w \in \Theta_w} \; L_{\text{adv}}(\boldsymbol{\theta}_w; \boldsymbol{\theta}_h, \beta_{\text{vib}}, \beta_{l_2}, \eta) \tag{5}$$

$$L_{\text{adv}}(\boldsymbol{\theta}_w; \boldsymbol{\theta}_h, \beta_{\text{vib}}, \beta_{l_2}, \eta) = \langle l(\boldsymbol{\theta}_h) - \eta, \boldsymbol{\theta}_w \rangle_P - \beta_{\text{vib}} \, \mathbb{E}_P \mathrm{KL}(p(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta}_w) \mid\mid \mathcal{N}(\mathbf{0}, \mathbf{I_d})) - \beta_{\mathbf{l_2}} \|\boldsymbol{\theta}_\mathbf{w}\|_\mathbf{2}^\mathbf{2}$$

Training. For each example, the adversary takes as input: (i) the last layer output of the current learner's feature network; and (ii) the input label. The adversary then outputs a weight (in $[0, 1]$). The idea of applying the adversary directly on the learner's features (instead of the original input) is based on recent literature [45, 28] that suggests re-training the prediction head is sufficient for robustness to shifts. The adversary tries to maximize weights on examples with value $\geq \eta$ (hyperparameter) and minimize on others. For the learner, in addition to the example it takes as input the adversary assigned weight for that example from the previous round and uses it to reweigh its loss in a minibatch. Both players are updated in a round (Algorithm 1).

[2] We use θ_h, θ_w and $l(\theta_h)$ to denote $w(\boldsymbol{\theta}_w; (\mathbf{x}, \mathrm{y}))$, $h(\boldsymbol{\theta}_h; \mathbf{x})$ and $l(h(\boldsymbol{\theta}_h; \mathbf{x}), \mathrm{y})$ respectively.

B Theoretical Analysis

The main objective of our analysis of BR-DRO is to show how adding a bitrate constraint on the adversary can: (i) give us tighter statistical estimates of the worst risk; and (ii) control the pessimism (excess risk) of the learned solution. First, we provide worst risk generalization guarantees using the PAC-Bayes framework [15], along with a result for kernel adversary. Then, we provide convergence rates and pessimism guarantees for the solution found by our online solver for a specific instance of $\mathcal{W}(\gamma)$. For both these, we analyze the constrained form of the conditional value at risk (CVaR) DRO objective [31] below.

Bitrate-Constrained CVaR DRO. When the uncertainty set $\mathcal{Q}$ is defined by the set of all distributions Q that have bounded likelihood *i.e.,* $\|q/p\|_\infty \leq 1/\alpha_0$, we recover the original CVaR DRO objective [19]. The bitrate-constrained version of CVaR DRO is given in equation 6 (see Appendix G for derivation). Note that, slightly different from Section 2, we define $\mathcal{W}$ as the set of all measurable functions w: $\mathcal{X} \times \mathcal{Y} \mapsto [0,1]$, since the other convex restrictions in equation 1 are handled by dual variable η. As in Section 2, $\mathcal{W}(\gamma)$ is derived from $\mathcal{W}$ using Definition 2.4. In equation 6, if we replace the bitrate-constrained class $\mathcal{W}(\gamma)$ with the unrestricted $\mathcal{W}$ then we recover the variational form of unconstrained CVaR DRO in Duchi et al. [17].

$$\mathcal{L}^*_{\text{cvar}}(\gamma) = \inf_{h \in \mathcal{H}, \eta \in \mathbb{R}} \sup_{w \in \mathcal{W}(\gamma)} R(h, \eta, w) \;\; \text{where,} \;\; R(h, \eta, w) = (1/\alpha_0)\langle l(h) - \eta, w \rangle_P + \eta \qquad (6)$$

B.1 Worst risk estimation bounds for BR-DRO.

Since we are only given a finite sampled dataset $\mathcal{D} \sim P^n$, we solve the objective in equation 6 using the empirical distribution $\hat{P}_n$. We denote the plug-in estimates as $\hat{h}^\gamma_D, \hat{\eta}^\gamma_D$. This incurs an estimation error for the true worst risk. But when we restrict our adversary to $\Delta(\mathcal{W}, \gamma)$, for a fixed learner h we reduce the worst-case risk estimation error which scales with the bitrate $\text{KL}(\cdot \mid\mid \pi)$ of the solution (deviation from prior π). Expanding this argument to every learner in $\mathcal{H}$, with high probability we also reduce the estimation error for the worst risk of $\hat{h}^\gamma_D$. Theorem B.1 states this generalization guarantee more precisely.

Theorem B.1 (worst-case risk generalization)**.** *With probability $\geq 1 - \delta$ over $\mathcal{D} \sim P^n$, the worst bitrate-constrained α_0-CVaR risk for $\hat{h}^\gamma_D$ can be upper bounded by the following oracle inequality:*

$$\sup_{w \in \mathcal{W}(\gamma)} R(\hat{h}^\gamma_D, \hat{\eta}^\gamma_D, w) \;\lesssim\; \mathcal{L}^*_{cvar}(\gamma) + \frac{M}{\alpha_0}\sqrt{\left(\gamma + \log\left(\frac{1}{\delta}\right) + (d+1)\log\left(\frac{L^2 n}{\gamma}\right) + \log n\right)/(2n-1)},$$

when $l(\cdot,\cdot)$ is $[0, M]$-bounded, L-Lipschitz and $\mathcal{H}$ is parameterized by convex set $\Theta \subset \mathbb{R}^d$.

Informally, Theorem B.1 tells us that bitrate-constraint γ gracefully controls the estimation error $\mathcal{O}(\sqrt{(\gamma + \mathcal{C}(\mathcal{H}))/n})$ (where $\mathcal{C}(\mathcal{H})$ is a complexity measure) if we know that Assumption 2.5 is satisfied. While this only tells us that our estimator is consistent with $\mathcal{O}_p(1/\sqrt{n})$, the estimate may itself be converging to a degenerate predictor, *i.e.,* $\mathcal{L}^*_{\text{cvar}}(\gamma)$ may be very high. For example, if the adversary can cleanly separate mislabeled points even after the bitrate constraint, then presumably these noisy points with high losses would be the ones mainly contributing to the worst risk, and up-weighting these points would result in a learner that has memorized noise. Thus, it becomes equally important for us to analyze the excess risk (or the pessimism) for the learned solution. Since this is hard to study for any arbitrary bitrate-constrained class $\mathcal{W}(\gamma)$, we shall do so for the specific class of reproducing kernel Hilbert space (RKHS) functions.

Special case of bounded RKHS. Let us assume there exists a prior Π such that $\mathcal{W}(\gamma)$ in Definition 2.4 is given by an RKHS induced by Mercer kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, s.t. the eigenvalues of the kernel operator decay polynomially, *i.e.,* $\mu_j \lesssim j^{-2/\gamma}$ ($\gamma < 2$). Then, if we solve for $\hat{h}^\gamma_D, \hat{\eta}^\gamma_D$ by doing kernel ridge regression over norm bounded ($\|f\|_{\mathcal{W}(\gamma)} \leq B \leq 1$) smooth functions f then we can control: (i) the pessimism of the learned solution; and (ii) the generalization error (Theorem B.2). Formally, we refer to pessimism for estimates $\hat{h}^\gamma_D, \hat{\eta}^\gamma_D$ as excess risk defined as:

$$\text{excess risk} := \sup_{w \in \mathcal{W}(\gamma)} |\inf_{h,\eta} R(h, \eta, w) - R(\hat{h}^\gamma_D, \hat{\eta}^\gamma_D, w)|. \qquad (7)$$

Theorem B.2 (bounded RKHS). *For $l, \mathcal{H}$ in Theorem B.1, and for $\mathcal{W}(\gamma)$ described above $\exists\, \gamma_0$ s.t. for all sufficiently bitrate-constrained $\mathcal{W}(\gamma)$ i.e., $\gamma \leq \gamma_0$, w.h.p. $1-\delta$ worst risk generalization error is $\mathcal{O}\left((1/n)\left(\log(1/\delta) + (d+1)\log(nB^{-\gamma}L^{\gamma/2})\right)\right)$ and the excess risk is $\mathcal{O}(B)$ for $\hat{h}^{\gamma}_{D}, \hat{\eta}^{\gamma}_{D}$ above.*

Thus, in the setting described above we have shown how bitrate-constraints given indirectly by γ, R can control both the pessimism and statistical estimation errors. Here, we directly analyzed the estimates $\hat{h}^{\gamma}_{D}, \hat{\eta}^{\gamma}_{D}$ but did not describe the specific algorithm used to solve the objective in equation 6 with $\hat{P}_n$. Now, we look at an iterative online algorithm to solve the same objective and see how bitrate-constraints can also influence convergence rates in this setting.

B.2 Convergence and excess risk analysis for an online solver.

In the following, we provide an algorithm to solve the objective in equation 6 and analyze how bitrate-constraint impacts the solver and the solution. For convex losses, the min-max objective in equation 6 has a unique solution and this matches the unique Nash equilibrium for the generic online algorithm (game) we describe (Lemma B.3). The algorithm is as follows: Consider a two-player zero-sum game where the learner uses a no-regret strategy to first play $h \in \mathcal{H}, \eta \in \mathbb{R}$ to minimize $\mathbb{E}_{w \sim \delta} R(h, \eta, w)$. Then, the adversary plays follow the regularized leader (FTRL) strategy to pick distribution $\delta \in \Delta(\mathcal{W}(\gamma))$ to maximize the same. Our goal is to analyze the bitrate-constraint γ's effect on the above algorithm's convergence rate and the pessimistic nature of the solution found. For this, we need to first characterize the bitrate-constraint class $\mathcal{W}(\gamma)$. If we assume there exists a prior Π such that $\mathcal{W}(\gamma)$ is Vapnik-Chervenokis (VC) class of dimension $O(\gamma)$, then in Theorem B.4, we see that the iterates of our algorithm converge to the equilibrium (solution) in $\mathcal{O}(\sqrt{\gamma \log n / T})$ steps. Clearly, the degree of bitrate constraint can significantly impact the convergence rate for a generic solver that solves the constrained DRO objective. Theorem B.4 also bounds the excess risk (equation 7) on $\hat{P}_n$.

Lemma B.3 (Nash equilibrium). *For convex $l(h)$, $l(h) \in [0, M]$, the objective in equation 6 has a unique solution which is also the Nash equilibrium of the game above when played over compact sets $\mathcal{H} \times [0, M]$, $\Delta(\mathcal{W}, \gamma)$. We denote this equilibrium as $h^*_D(\gamma), \eta^*_D(\gamma), \delta^*_D(\gamma)$.*

Theorem B.4. *At time step t, if the learner plays (h_t, η_t) with no-regret and the adversary plays δ_t with FTRL strategy that uses a negative entropy regularizer on δ then average iterates $(\bar{h}_T, \bar{\eta}_T, \bar{\delta}_T) = (1/T)\sum_{t=1}^{T}(h_t, \eta_t, \delta_t)$ converge to the equilibrium $(h^*_D(\gamma), \eta^*_D(\gamma), \delta^*_D(\gamma))$ at rate $\mathcal{O}(\sqrt{\gamma \log n / T})$. Further the excess risk defined above is $\mathcal{O}((M/\alpha_0)\left(1 - \frac{1}{n^{\gamma}}\right))$.*

C Detailed experiments

Our experiments aim to evaluate the performance of BR-DRO and compare it with ERM and group shift robustness methods that do not require group annotations for training examples. We conduct empirical analyses along the following axes: (i) worst group performance on datasets that exhibit known spurious correlations; (ii) robustness to random label noise in the training data; (iii) average performance on hybrid covariate shift datasets with unspecified groups; and (iv) accuracy in identifying minority groups. See Appendix F for additional experiments and details.

Baselines. Since our objective is to be robust to group shifts without group annotations on training examples, we explore baselines that either optimize for the worst minority group (CVaR DRO [31]) or use training losses to identify specific minority points (LfF [41], JTT [35]). Group DRO [46] is treated as an oracle. We also compare with the simple re-weighting baseline (RWY) proposed by Idrissi et al. [26].

Implementation details. We train using Resnet-50 [24] for all methods and datasets except CivilComments, where we use BERT [58]. For our VIB adversary, we use a 1-hidden layer neural network encoder and decoder (one for each label). As mentioned in Section 2, the adversary takes as input the learner model's features and the true label to generate weights. All implementation and design choices for baselines were adopted directly from Liu et al. [35], Idrissi et al. [26]. We provide model selection methodology and other details in Appendix F.

Datasets. For experiments in the known groups and label noise settings we use: (i) Waterbirds [54] (background is spurious), CelebA [36] (binary gender is spuriously correlated with label "blond"); and

Method	FMoW		Camelyon17
	Avg	W-Reg	Avg
ERM	53.3 (0.1)	32.4 (0.3)	70.6 (1.6)
JTT [35]	52.1 (0.1)	31.8 (0.2)	66.3 (1.3)
LfF [41]	49.6 (0.2)	31.0 (0.3)	65.8 (1.2)
RWY [26]	50.8 (0.1)	30.9 (0.2)	69.9 (1.3)
Group DRO [46]	51.9 (0.2)	30.4 (0.3)	68.5 (0.9)
CVaR DRO [31]	51.5 (0.1)	31.0 (0.3)	66.8 (1.3)
`BR-DRO` (VIB) (ours)	52.0 (0.2)	31.8 (0.2)	70.4 (1.5)
`BR-DRO` (l_2) (ours)	53.1 (0.1)	32.3 (0.2)	71.2 (1.0)

Table 2: Average (Avg) and worst region (W-Reg for FMoW) test accuracies on Camelyon17 and FMoW.

CivilComments (WILDS) [11] where the task is to predict "toxic" texts and there are 16 predefined groups [29]. We use FMoW and Camelyon17 [29] to test methods on datasets that do not have explicit group shifts. In FMoW the task is to predict land use from satellite images where the training/test set comprises of data before/after 2013. Test involves both subpopulation shifts over regions (*e.g.,* Africa, Asia) and domain generalization over time (year). Camelyon17 presents a domain generalization problem where the task is to detect tumor in tissue slides from different sets of hospitals in train and test sets.

C.1 More experiments on robustness to noise.

To further verify our claims, we set up a noisily labeled synthetic dataset (see Appendix F for details). In Figure 2a we plot training samples as well as the solutions learned by `BR-DRO` and and JTT on synthetic data. In Figure 1*(right)* we also plot exactly which points are upweighted by `BR-DRO` and JTT. Using both figures, we note that JTT mainly upweights the noisy points (in red) and memorizes them using $\mathbf{x}_{\text{noise}}$. Without any weights on minority, it memorizes them as well and learns component along spurious feature. On the contrary, when we restrict the adversary with `BR-DRO` to be sparse (l_1 penalty), it only upweights minority samples, since no sparse predictor can separate noisy points in the data. Thus, the learner can no longer memorize the upweighted minority and we recover the robust predictor along core feature.

C.2 How does `BR-DRO` perform on more general covariate shifts?

In Figure 2 we report the average test accuracies for `BR-DRO` and baselines on the hybrid dataset FMoW and domain generalization dataset Camelyon17. In $(\cdot)$ we report the standard error of the mean accuracy across five runs. Given its hybrid nature, on FMoW we also report worst region accuracy. First, we note that on these datasets group shift robustness baselines do not do better than ERM. Some are either too pessimistic (*e.g.,* CVaR DRO), or require heavy assumptions (*e.g.,* Group DRO) to be robust to domain generalization. This is also noted by Gulrajani and Lopez-Paz [22]. Next, we see that `BR-DRO` (l_2 version) does better than other group shift baselines on both both worst region and average datasets and matches ERM performance on Camelyon17. One explanation could be that even though these datasets test models on new domains, there maybe some latent groups defining these domains that are simple and form a part of latent subpopulation shift. Investigating this claim further is a promising line of future work.

C.3 What fraction of minority is recovered by Bitrate-Constrained DRO?

We claim that our less pessimistic objective can more accurately recover (upsample) the true minority group if indeed the minority group is simple (see Assumption 2.5 for our definition of simple). In this section, we aim to verify this claim. If we treat examples in the top 10% (chosen for post hoc analysis) fraction of examples as our predicted minorities, we can check precision and recall of this decision on the Waterbirds and CelebA datasets. Figure 3 plots these metrics at each training epoch for `BR-DRO` (with varying β_{vib}), JTT and CVaR DRO. Precision of the random baseline tells us the true fraction of minority examples in the data. First we note that `BR-DRO` consistently performs much better on this metric than unconstrained CVaR DRO. In fact, as we reduce strength of β_{vib} we recover precision/recall close to the latter. This controlled experiment shows that the bitrate constraint is

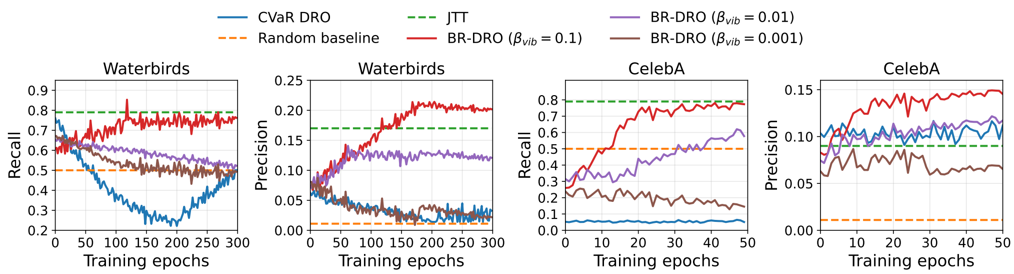


Figure 3: By considering the fraction of points upweighted by our adversary (top 10%) as the positive class we analyze the precision and recall of this class with respect to the minority group. and do the same for JTT, random baseline and CVaR DRO. BR-DRO achieves highest precision and matches recall with JTT asymptotically. We also find that increasing bitrate constraint β_{vib} helps improving precision/recall.

helpful (and very much needed) in practice to identify rare simple groups. In Figure 3 we observe that asymptotically, the precision of BR-DRO is better than JTT on both datasets, while the recall is similar. Since importance weighting has little impact in later stages with exponential tail losses [52, 13], other losses (*e.g.,* polytail Wang et al. [56]) may further improve the performance of BR-DRO as it gets better at identifying the minority classes when trained longer.

D Related Work

Prior works in robust ML [e.g., 32, 33, 20] address various forms of adversarial or structured shifts. We specifically review prior work on robustness to group shifts. While those based on DRO optimize for worst-case shifts in an explicit uncertainty set, the robust set is implicit for some others, with most using some form of importance weighting.

Distributionally robust optimization (DRO). DRO methods generally optimize for worst-case performance on joint $(\mathbf{x}, \mathrm{y})$ distributions that lie in an f-divergence ball (uncertainty set) around the training distribution [7, 44, 8, 9, 40, 17, 19]. Hu et al. [25] highlights that the conservative nature of DRO may lead to degenerate solutions when the unrestricted adversary uniformly upweights all misclassified points. Sagawa et al. [46] proposes to address this by limiting the adversary to shifts that only differ in marginals over predefined groups. However, in addition to it being difficult to obtain this information, Kearns et al. [27] raise "gerrymandering" concerns with notions of robustness that fix a small number of groups apriori. While they propose a solution that looks at exponentially many subgroups defined over protected attributes, our method does not assume access to such attributes and aims to be fair on them as long as they are realized by simple functions. Finally, Zhai et al. [60] avoid conservative solutions by solving the DRO objective over randomized predictors learned through boosting. We consider deterministic and over-parameterized learners and instead constrain the adversary's class.

Constraining the DRO uncertainty set. In the marginal DRO setting, Duchi et al. [18] limit the adversary via easier-to-control reproducing kernel hilbert spaces (RKHS) or bounded Hölder continuous functions [34, 57]. While this reduces the statistical error in worst risk estimation, the size of the uncertainty set (scales with the data) remains too large to avoid cases where an adversary can re-weight mislabeled and hard examples from the majority set [14]. In contrast, we restrict the adversary even for large datasets where the estimation error would be low, as this would reduce excess risk when we only care about robustness to rare sub-populations defined by simple functions. Additionally, while their analysis and method prefers the adversary's objective to have a strong dual, we show empirical results on real-world datasets and generalization bounds where the adversary's objective is not necessarily convex.

Robustness to group shifts without demographics. Recent works [50, 16, 5] that aim to achieve group robustness without access to group labels employ various heuristics where the robust set is implicit while others require data from multiple domains [3, 59] or ability to query test samples [30]. Liu et al. [35] use training losses for a heavily regularized model trained with empirical risk minimization (ERM) to directly identify minority data points with higher losses and re-train on the dataset that up-weights the identified set. Nam et al. [41] take a similar approach. Other methods [26] propose simple baselines that subsample the majority class in the absence of group demographics and the majority group in its presence. Hashimoto et al. [23] find DRO over a χ^2-divergence ball can reduce the otherwise increasing disparity of per-group risks in a dynamical system. Since it does not use features to upweight points (like BR-DRO) it is vulnerable to label noise. Same can be said about some other works (*e.g.,* [35, 41]).

Importance weighting in deep learning. Finally, numerous works [17, 31, 33, 43] enforce robustness by re-weighting losses on individual data points. Recent investigations [52, 13, 37] reveal that such objectives have little impact on the learned solution in interpolation regimes. One way to avoid this pitfall is to train with heavily regularized models [46, 47] and employ early stopping. Another way is to subsample certain points, as opposed to up-weighting [26]. In this work, we use both techniques while training our objective and the baselines, ensuring that the regularized class is robust to shifts under misspecification [57].

E BR-DRO algorithm

If the bitrate constraint is applied via the KL term in equation 5, we implement the adversary as a variational information bottleneck [2] (VIB), where the KL divergence with respect to a standard Gaussian prior controls the bitrate of the adversary's feature set $\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\theta}_w)$. Increasing β_{vib} can be seen as enforcing lower bitrate features *i.e.,* reducing γ in $\mathcal{W}(\gamma)$ (smaller value of KL$(\delta \mid\mid \pi)$ in the primal formulation in Definition 2.4). If the constraint is applied via the l_2 term we implement the adversary as a linear layer. In some cases (*e.g.,* Section 3) we use a sparsity constraint (l_1 norm) on the linear adversary.

Algorithm 1: Bitrate-Constraint DRO (Online Algorithm)

Input: Adversary VIB penalty β_{vib}; Step sizes η_l, η_w; Dataset $\mathcal{D} = (\mathbf{x}_i, \mathrm{y}_i)_{i=1}^n$

Initialize $\theta_h^{(1)}$ and $\theta_w^{(1)}$

for $t = 1, \ldots, T$ **do**

From $\mathcal{D}$, sample $\mathbf{x}, \mathrm{y} \sim \mathcal{D}$ `/* Sample datapoint */`

$\boldsymbol{\theta}_h^{(t+1)} \leftarrow \Pi_{\Theta_h} \left(\boldsymbol{\theta}_h^{(t)} - \eta_h \nabla_{\boldsymbol{\theta}_h} \left[l(\boldsymbol{\theta}_h^{(t)}(\mathbf{x}), \mathrm{y}) \cdot \boldsymbol{\theta}_w(\mathbf{x}, \mathrm{y}) \right] \right)$ `/* Update` $\boldsymbol{\theta}_h$ `*/`

$\boldsymbol{\theta}_w^{(t+1)} \leftarrow \Pi_{\Theta_w} \left(\boldsymbol{\theta}_w^{(t)} + \eta_w \nabla_{\boldsymbol{\theta}_w} L_{\text{adv}}(\boldsymbol{\theta}_w^{(t)}; \boldsymbol{\theta}_h^{(t)}, \beta_{\text{vib}}, \beta_{l_2}, \eta) \right)$ `/* Update` $\boldsymbol{\theta}_w$ `*/`

end

Output: $\bar{\boldsymbol{\theta}}_h = \frac{1}{T} \sum_{t=1}^T \boldsymbol{\theta}_h^{(t)}, \bar{\boldsymbol{\theta}}_w = \frac{1}{T} \sum_{t=1}^T \boldsymbol{\theta}_w^{(t)}$

F Additional empirical results and other experiment details

F.1 Hyper-parameter tuning methodology

There are two ways in which we tune hyperparameters on datasets with known groups (CelebA, Waterbirds, CivilComments): (i) on average validation performance; (ii) worst group accuracy. The former does not use group annotations while the latter does. Similar to prior works [35, 26] we note that using group annotations (on a small validation set) does improve performance. In Table 3 we report our study which varies the the fraction p of group labels that are available at test time. For each setting of p, we do model selection by taking weighted (by p) mean over two entities (i) average validation on all samples, (ii) worst group validation on a fraction p of minority samples. In the case where $p = 0$, we only use average validation. We report our results on CelebA and Waterbirds dataset. For the two WILDS datasets we tune hyper-parameters on OOD Validation set.

F.2 Synthetic dataset details

We follow the explicit-memorization setup in Sagawa et al. [47] which we summarize here briefly. Let input $\mathbf{x} = [\mathbf{x}_{\text{core}}, \mathbf{x}_{\text{spu}}, \mathbf{x}_{\text{noise}}]$ where $\mathbf{x}_{\text{core}} \mid y \sim \mathcal{N}(y, \sigma^2_{\text{core}}),\ \mathbf{x}_{\text{spu}} \mid a \sim \mathcal{N}(a, \sigma^2_{\text{spu}})$ and $\mathbf{x}_{\text{noise}} \sim \mathcal{N}(\mathbf{0}, (\sigma^2_{\text{noise}} \mathbf{I_d})/\mathbf{d})$. Here $a \in \{-1, 1\}$ refers to a spurious attribute, and label is $y \in \{-1, 1\}$, We set $a = y$ with probability $P(_{\text{maj}}) = 1 - P(_{\text{min}})$. The level of correlation between a and y is controlled by $P(_{\text{maj}})$. Additionally, we flip true label with probability $P(_{\text{noise}})$.

Method	Waterbirds				CelebA			
	$p = 0.0$	$p = 0.02$	$p = 0.05$	$p = 0.1$	$p = 0.0$	$p = 0.02$	$p = 0.05$	$p = 0.1$
JTT	62.7	73.9	77.3	84.4	42.1	68.3	80.5	80.3
CVaR DRO	63.9	65.8	72.6	74.1	33.6	40.4	60.4	63.2
LfF	48.6	58.9	70.3	79.5	34.0	58.9	60.0	78.3
BR-DRO (VIB)	**69.3**	77.6	76.1	84.9	**52.4**	71.2	80.3	79.9
BR-DRO (l_2)	**68.9**	75.2	79.4	86.1	**55.8**	63.5	74.6	80.4

Table 3: We check to what extent fraction of group annotations in the training data affect performance. For each dataset and method, we tune its hyper-parameters on the average validation and worst group (only on the small fraction p that is available). We see that while all methods consistently improve as we increase group annotations and tune for worst group accuracy on the annotated samples, BR-DRO does do better that prior works when tuned on just average validation ($p = 0.$). At the same time, we note that this still does not match the performance of BR-DRO when tuned on worst group validation (seen in Table 1).

F.3 Degree of constraint

In Figure 4 we see how worst group performance varies on Waterbirds and CelebA as a function of increasing constraint. We also plot average performance on the Camelyon dataset. We mainly note that for either of the constraint implementations, only when we significantly increase the capacity do we actually see the performance of BR-DRO improve. The effect is more prominent on groups shift datasets with simple groups (Waterbirds, CelebA). Under less restrictive capacity constraints we note that its performance is similar to CVaR DRO (see Figure 3). This is expected since CVaR DRO is the completely unconstrained version of our objective.

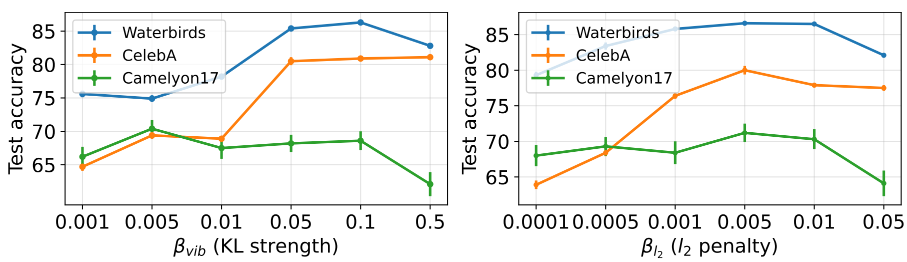


Figure 4: **Optimal bitrate-constraints for robustness to distributions shifts:** For two versions of capacity control: KL, l_2 penalty (see Section 2) we show how worst group performance on Waterbirds, CelebA and average performance on Camelyon test sets improves with increasing constraints under either VIB (β_{vib}) or linear (β_{l_2}) adversaries.

F.4 Hyper-parameter details.

For all hyper-parameters of prior methods we use the ones state in their respective prior works. The implementation Group DRO, JTT, CVaR DRO is borrowed from the implementation made public by authors of Liu et al. [35]. For datasets Waterbirds, CelebA and CivilComments we choose the hyper-parameters (whenever applicable) learning rate, batch size, weight decay on learner, optimizer, early stopping criterion, learning rate schedules used by Liu et al. [35] for their implementation of CVaR DRO method. For datasets FMoW and Camelyon17 we choose values for these hyper-parameters to be the ones used by Koh et al. [29] for the ERM baseline. Details on BR-DRO specific hyper-parameters that we tuned are in Table 4. Also, note that we release our implementation with this submission.

G Omitted Proofs

First we shall state some a couple of technical lemmas that we shall refer to at multiple points. Then, we prove our theoretical claims in our analysis in Appendix B, in the order in which they

Hyper-parameter	Waterbirds	CelebA	CivilComments	FMoW	Camelyon17
learning rate for adversary	0.01	0.05	0.001	0.02	0.01
threshold η	0.05	0.05	0.1	0.1	0.1
β_{vib}	0.1	0.1	0.02	0.005	0.005
β_{l_2}	0.01	0.005	0.005	0.02	0.005

Table 4: Hyperparameters for our method on different datasets (tuned on worst group validation performance). Note, that the threshold η here is the top $x\%$ fraction.

appear. Before we get into those we provide proof for our Corollary 2.2 and the derivation of Bitrate-Constrained CVaR DRO in Equation 6.

Lemma G.1 (Hoeffding bound [55]). *Let $X_1, \ldots, X_n$ be a set of μ_i centered independent sub-Gaussians, each with parameter σ_i. Then for all $t \geq 0$, we have*

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right] \leq \exp\left(-\frac{n^2 t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right). \tag{8}$$

Lemma G.2 (Lipschitz functions of Gaussians [55]). *Let $X_1, \ldots, X_n$ be a vector of iid Gaussian variables and $f : \mathbb{R}^n \mapsto \mathbb{R}$ be L-Lipschitz with respect to the Euclidean norm. Then the random variable $f(X) - \mathbb{E}[f(X)]$ is sub-Gaussian with parameter at most L, thus:*

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2 \cdot \exp\left(-\frac{t^2}{2L^2}\right), \;\; \forall\, t \geq 0. \tag{9}$$

G.1 Proof of Corollary 2.2

Let us recall the definition of a well defined group structure. For a pair of measures $Q \ll P$ we say $\mathcal{G}(P, Q)$ is well defined if given there exists a set of disjoint measurable sets $\mathcal{G}_{P,Q} = \{G_k\}_{k=1}^{K}$ such that $G_k \in \Sigma$, $Q(G_k) > 0$, $Q(\mathcal{G}(P, Q)) = 1$ and we have:

$$K = \min\{|\{G_1, \ldots, G_M\}| : \; p(\mathbf{x}, \mathrm{y} \mid G_m) = q(\mathbf{x}, \mathrm{y} \mid G_m) > 0, \forall(\mathbf{x}, \mathrm{y}) \in G_m \;\; \forall m \in [M]\}$$

Now by definition K is finite. Thus if there exists two well defined group structures $\mathcal{G}_1(P, Q)$ and $\mathcal{G}_2(P, Q)$ for the same pair P, Q then it must be the case that $K = \mathcal{G}_1(P, Q) = \mathcal{G}_2(P, Q)$.

Then, there must exist $G \in \mathcal{G}_1(P, Q)$ such that $Q(G) > 0$ and $G', G'' \in \mathcal{G}_2(P, Q)$ where $Q(G'), Q(G'') > 0$ and $Q(G \cap G'), Q(G \cap G'') > 0$.

Note that since $G, G', G'' \in \Sigma$ that is closed under countable unions, we have that $G \cap G'$ and $G \cap G''$ are two sets where $q(\mathbf{x}, \mathrm{y}) > 0 \; \forall(\mathbf{x}, \mathrm{y}) \in G \cap G', G \cap G''$.

Let $(\mathbf{x}_1, \mathrm{y}_1) \in (G \cap G')$ and $(\mathbf{x}_2, \mathrm{y}_2) \in (G \cap G'')$. From definition we know that $q(\mathbf{x}_2, \mathrm{y}_2), q(\mathbf{x}_1, \mathrm{y}_1) > 0$ and . Since both $(\mathbf{x}_1, \mathrm{y}_1)$ and $(\mathbf{x}_2, \mathrm{y}_2)$ are in G we have that:

$$q(\mathbf{x}_1, \mathrm{y}_1) = \frac{Q(G)}{P(G)} \cdot p(\mathbf{x}_1, \mathrm{y}_1) = \frac{Q(G')}{P(G')} \cdot p(\mathbf{x}_1, \mathrm{y}_1) \tag{10}$$

$$q(\mathbf{x}_2, \mathrm{y}_2) = \frac{Q(G)}{P(G)} \cdot p(\mathbf{x}_2, \mathrm{y}_2) = \frac{Q(G'')}{P(G'')} \cdot p(\mathbf{x}_2, \mathrm{y}_2) \tag{11}$$

Thus, we can conclude that $\frac{Q(G')}{P(G')} = \frac{Q(G'')}{P(G'')}$. This implies that $G' \cup G''$ also satisfies the following that $Q(G' \cup G'') > 0$ and $q(\mathbf{x}, \mathrm{y} \mid G' \cup G'') = p(\mathbf{x}, \mathrm{y} \mid G' \cup G'')$.

Thus, we can construct a new $\mathcal{G}_3(P, Q) = \{G \in \mathcal{G}_2(P, Q) : \ G \notin \{G', G''\}\} \cup \{G' \cup G''\}$. Clearly, $\mathcal{G}_3(P, Q)$ satisfies all group structure properties and is smaller than $\mathcal{G}_2(P, Q)$. Thus, we arrive at a contradiction which proves the claim that $\mathcal{G}(P, Q)$ is indeed unique whenever well defined.

G.2 Derivation of Bitrate-Constrained CVaR DRO in equation 6

Recall that we define $\mathcal{W}$ as the set of all measurable functions $w : \ \mathcal{X} \times \mathcal{Y} \mapsto [0, 1]$, since the other convex restrictions in equation 1 are handled by dual variable η. As in Section 2, $\mathcal{W}(\gamma)$ is derived from the new $\mathcal{W}$ using Definition 2.4. With that let us first state the CVaR objective [31].

$$\begin{aligned} &\mathcal{L}_{\text{cvar}}(h, P) \coloneqq \sup_{q} \int_{\mathcal{X} \times \mathcal{Y}} q(\mathbf{x}, \mathrm{y}) \cdot l(h) \\ &\text{s.t.} \ \ q \geq 0, \ \ \|q/p\|_\infty \leq (1/\alpha_0), \ \ \int_{\mathcal{X} \times \mathcal{Y}} q(\mathbf{x}, \mathrm{y}) = 1 \end{aligned} \tag{12}$$

The objective in q is linear with convex constraints, and has a strong dual (see Duchi et al. [17], Boyd et al. [12] for the derivation) which is given by:

$$\begin{aligned} &\inf_{\eta \in \mathbb{R}} \left\{ \frac{1}{\alpha_0} \mathbb{E}_P (l(h) - \eta)_+ + \eta \right\} \\ &= \inf_{\eta \in \mathbb{R}} \left\{ \frac{1}{\alpha_0} \langle (l(h) - \eta)_+, \mathbb{1} \rangle_P + \eta \right\} \\ &= \inf_{\eta \in \mathbb{R}} \left\{ \frac{1}{\alpha_0} \langle (l(h) - \eta), \mathbb{1}(l(h) - \eta \geq 0) \rangle_P + \eta \right\} \qquad (13) \\ &= \inf_{\eta \in \mathbb{R}} \sup_{w \in \mathcal{W}} \left\{ \frac{1}{\alpha_0} \langle (l(h) - \eta), w \rangle_P + \eta \right\} \qquad (14) \end{aligned}$$

The last equality is true since the set $\mathbb{1}(l(h) - \eta \geq 0)$ is measurable under P (based on our setup in Appendix B). Note that for any h, the objective $\frac{1}{\alpha_0} \langle (l(h) - \eta), w \rangle_P + \eta$ is linear in w, and η. If we further assume the loss $l(h)$ to be the l_{0-1} loss, it is bounded, and thus the optimization over η can be restricted to a compact set. Next, $\mathcal{W}$ is also a compact set of functions since we restrict our solvers to measurable functions that take values bounded in $[0, 1]$.

$$\mathcal{L}_{\text{cvar}}(h, P) = \inf_{\eta \in \mathbb{R}} \sup_{w \in \mathcal{W}} \left\{ \frac{1}{\alpha_0} \langle (l(h) - \eta), w \rangle_P + \eta \right\} \tag{15}$$

The above objective is precisely the Bitrate-Constrained CVaR DRO objective we have in equation 6. Later in the Appendix we shall need an equivalent form of the objective which we shall derive below.

We can now invoke the Weierstrass' theorem in Boyd et al. [12] to give us the following:

$$\begin{aligned} \mathcal{L}_{\text{cvar}}(h, P) &= \inf_{\eta \in \mathbb{R}} \sup_{w \in \mathcal{W}} \left\{ \frac{1}{\alpha_0} \langle (l(h) - \eta), w \rangle_P + \eta \right\} \\ &= \frac{1}{\alpha_0} \sup_{w \in \mathcal{W}} \left\{ \inf_{\eta \in \mathbb{R}} \langle (l(h) - \eta), w \rangle_P + \eta \right\} \end{aligned} \tag{16}$$

Now, the final objective $\inf_{h \in \mathcal{H}} \mathcal{L}_{\text{cvar}}(h, P)$ is given by:

$$\frac{1}{\alpha_0} \inf_{h \in \mathcal{H}} \sup_{w \in \mathcal{W}} \left\{ \inf_{\eta \in \mathbb{R}} \langle (l(h) - \eta), w \rangle_P + \eta \right\} \tag{17}$$

In the above equation we can now replace the unconstrained class $\mathcal{W}$ with our bitrate-constrained class $\mathcal{W}(\gamma)$ to get the following:

$$\frac{1}{\alpha_0} \inf_{h\in\mathcal{H}} \sup_{w\in\mathcal{W}(\gamma)} \left\{ \inf_{\eta\in\mathbb{R}} \langle (l(h)-\eta), w\rangle_P + \eta \right\} \tag{18}$$

G.3 Proof of Theorem B.1

For convenience we shall first restate the Theorem here.

Theorem G.3 ([restated). *worst-case risk generalization] With probability $\geq 1-\delta$ over sample $\mathcal{D} \sim P^n$, the worst risk for $\hat{h}_D^\gamma$ can be upper bounded by the following oracle inequality:*

$$\sup_{w\in\Delta(\mathcal{W},\gamma)} R(\hat{h}_D^\gamma, \hat{\eta}_D^\gamma, w) \; - \; \mathcal{L}^*_{cvar}(\gamma) \; \lesssim \; \frac{M}{\alpha_0}\sqrt{\left(\gamma + \log\left(\frac{1}{\delta}\right) + (d+1)\log\left(\frac{L^2 n}{\gamma}\right) + \log n\right)/(2n-1)},$$

when $l(\cdot,\cdot)$ is $[0,M]$-bounded, L-Lipschitz and $\mathcal{H}$ is parameterized by convex set $\Theta \subset \mathbb{R}^d$.

The overview of the proof can be split into two parts:

- For each learner, first obtain the oracle PAC-Bayes [39] worst risk generalization guarantee over the adversary's action space $\Delta(\mathcal{W},\gamma)$.
- Then, apply uniform convergence bounds using a union bound over a covering of the class $\mathcal{H}$ to get the final result.

Intuition: The only tricky part lies in the fact that oracle PAC-Bayes inequality would not give us arbitrary control over the generalization error for each learner, which we would typically get in Hoeffding type bounds. Hence, we need to ensure that the the worst risk generalization rate decays faster than how the size of the covering would increase for a ball of radius defined by the worst generalization error.

Now, we shall invoke the following PAC-Bayes generalization guarantee stated (Lemma G.4) since $R(h,\eta,w) \in [0, M/\alpha_0]$.

Lemma G.4 (PAC-Bayes [15, 39]). *With probability $\geq 1-\delta$ over choice of dataset $\mathcal{D}$ of size n the following inequality is satisfied*

$$\mathbb{E}_P\mathbb{E}_Q(l_{0-1}(h(\mathbf{x}),\mathrm{y})) \leq \mathbb{E}_{\hat{P}_n}\mathbb{E}_Q(l_{0-1}(h(\mathbf{x}),\mathrm{y})) + \sqrt{\frac{D(Q||P) + \log(1/\delta) + \frac{5}{2}\log n + 8}{2n-1}} \tag{19}$$

A direct application of this gives us that with probability at least $1-\omega$: .

$$\mathbb{E}_{w\sim\delta} R(h,\eta,w) \leq \mathbb{E}_{w\sim\delta}\left[\frac{1}{\alpha_0}\langle l(h)-\eta, w\rangle_{\hat{P}_n}\right] + \eta + \sqrt{\frac{\mathrm{KL}(\delta \mid\mid \pi) + \log(1/\omega) + \frac{5}{2}\log n + 8}{2n-1}}$$

Let $\hat{R}_D(h,\eta,w) = \frac{1}{\alpha_0}\langle l(h)-\eta, w\rangle_{\hat{P}_n} + \eta$ Since the above inequality holds for any data dependent δ:.

$$\sup_{\delta\in\Delta(\mathcal{W},\gamma)} \mathbb{E}_{w\sim\delta} R(h,\eta,w) \leq \sup_{\delta\in\Delta(\mathcal{W},\gamma)} \left[\hat{R}_D(h,\eta,w) + \eta + \sqrt{\frac{\mathrm{KL}(\delta \mid\mid \pi) + \log(1/\omega) + \frac{5}{2}\log n + 8}{2n-1}}\right]$$

Further, we make use of the fact $\mathrm{KL}(\delta \mid\mid \pi) \leq \gamma$.

$$\leq \sup_{\delta_1\in\Delta(\mathcal{W},\gamma)} \left[\hat{R}_D(h,\eta,w)\right] + \sup_{\delta_2\in\Delta(\mathcal{W},\gamma)} \left[\sqrt{\frac{\mathrm{KL}(\delta_2 \mid\mid \pi) + \log(1/\omega) + \frac{5}{2}\log n + 8}{2n-1}}\right]$$

Thus,

$$\sup_{\delta \in \Delta(\mathcal{W},\gamma)} \mathbb{E}_{w\sim\delta} R(h,\eta,w) - \sup_{\delta \in \Delta(\mathcal{W},\gamma)} \mathbb{E}_{w\sim\delta} \hat{R}_D(h,\eta,w) \leq \left[\sqrt{\frac{\gamma + \log(1/\delta) + \frac{5}{2}\log n + 8}{2n-1}} \right]$$

To actually apply this uniformly over h, η, we would first need two sided concentration which we derive below as follows:

Let $a_i = \hat{R}_D(h,\eta,\delta) - R(h,\eta,\delta)$, Since $R(h,\eta,\delta) \leq M/\alpha_0$, we can apply Hoeffding bound with $t = \lambda/n$ in Lemma G.1 on a_i to get:

$$\mathbb{E}_{\mathcal{D}} \exp(\lambda \cdot a_i) \leq \exp \frac{\lambda^2 (M/\alpha_0)^2}{8n} \mathbb{E}_\pi \mathbb{E}_{\mathcal{D}} \exp(\lambda \cdot a_i) \leq \mathbb{E}_\pi \exp \frac{\lambda^2 (M/\alpha_0)^2}{8n}$$

Applying Fubini's Theorem, followed by the Donsker Varadhan variational formulation we get:

$$\begin{aligned} \mathbb{E}_{\mathcal{D}} \mathbb{E}_\pi \left[\exp(\lambda \cdot a_i)\right] &\leq \mathbb{E}_\pi \exp \frac{\lambda^2 (M/\alpha_0)^2}{8n} \\ = \mathbb{E}_{\mathcal{D}} \exp \sup_{\delta \in \Delta(\mathcal{W},\gamma)} \left[(\lambda \cdot a_i) - \mathrm{KL}(\delta \mid\mid \pi)\right] &\leq \exp \frac{\lambda^2 (M/\alpha_0)^2}{8n} \end{aligned}$$

The Chernoff bound finally gives us with probability $\geq 1 - \omega$:

$$\mathbb{E}_{\hat{P}_n} \mathbb{E}_Q((h(\mathbf{x}), \mathrm{y})) \lesssim \mathbb{E}_P \mathbb{E}_Q((h(\mathbf{x}), \mathrm{y})) + \frac{M}{\alpha_0} \sqrt{\frac{\mathrm{KL}(\delta \mid\mid \pi) + \log(1/\omega) + \log n}{2n-1}}$$

Using the reverse form of the empirical PAC Bayes inequality, we can do a derivation similar to the one following the PAC-Bayes bound in Lemma G.4 to get for any fixed $\eta \in [0, M], h \in \mathcal{H}$ we get:

$$\begin{aligned} \left| \sup_{\delta \in \Delta(\mathcal{W},\gamma)} \mathbb{E}_{w\sim\delta} R(h,\eta,w) - \sup_{\delta \in \Delta(\mathcal{W},\gamma)} \mathbb{E}_{w\sim\delta} \hat{R}_D(h,\eta,w) \right| &\lesssim \frac{M}{\alpha_0} \sqrt{\frac{\mathrm{KL}(\delta \mid\mid \pi) + \log(1/\omega) + \log n}{2n-1}} \\ &\lesssim \frac{M}{\alpha_0} \sqrt{\frac{\gamma + \log(1/\omega) + \log n}{2n-1}} \end{aligned}$$

Because we see that in the above bound the dependence on δ, is given by a log term we are essentially getting an "exponential-like" concentration. So we can think about applying uniform convergence bounds over the class $\mathcal{H} \times [0, M]$ to bounds the above with high probability $\forall (h, \eta)$ pairs.

We will now try to get uniform convergence bounds with two approaches that make different assumptions on the class of functions $l(h)$. The first is very generic and we will show why such a generic assumption is not sufficient to get an upper bound on the generalization that is $\mathcal{O}(1/\sqrt{n})$ in the worst case. Then, in the second approach we show how assuming a parameterization will fetch us a rate of that form if we additionally assume that the loss function is L-Lipschitz.

Approach 1:

Assume $l(h)$ lies in a class of $(\alpha, 1)$-Hölder continuous functions Now we shall use the following covering number bound for $(\alpha, 1)$-Hölder continuous functions to get a uniform convergence bound over $\mathcal{H} \times [0, M]$.

Lemma G.5 (Covering number $(\alpha, 1)$-Hölder continuous)**.** *Let $\mathcal{X}$ be a bounded convex subset of $\mathbb{R}^d$ with non-empty interior. Then, there exists a constant K depending only on α and d such that*

$$\log \mathcal{N}(\epsilon, C_1^\alpha(\mathcal{X}), \|\cdot\|_\infty) \leq K \lambda(\mathcal{X}^1) \left(\frac{1}{\epsilon}\right)^{d/\alpha} \tag{20}$$

for every $\epsilon > 0$, where $\lambda(\mathcal{X}^1)$ is the Lebesgue measure of the set $\{x : \|x - \mathcal{X}\| \leq 1\}$. Here, $C_1^\alpha(\mathcal{X})$ refers to the class of $(\alpha, 1)$-Hölder continuous functions.

We assume that $l(h)$ is $(\alpha,1)$-Hölder continuous. And therefore by definition, of $R(h,\eta,\cdot)$, the function is $(\alpha,1)$-Hölder continuous in $(l(h),\eta)$. Similat argument applies for $\sup_{\delta\in\Delta(\mathcal{W},\gamma)} \mathbb{E}_{w\sim\delta} R(h,\eta,w)$ since taking a pointwise supremum for a linear function over a convex set $\Delta(\mathcal{W},\gamma)$ would retain Hölder continuity for some value of α. Applying the above we get:

$$\log \mathcal{N}(\epsilon, \sup_{\delta\in\Delta(\mathcal{W},\gamma)} \mathbb{E}_{w\sim\delta} R(\cdot,\cdot,w), \|\cdot\|_\infty) \lesssim \left(\frac{M}{\alpha_0}\sqrt{\frac{\gamma+\log(1/\omega)+\log n}{2n-1}}\right)^{-(d/\alpha)}$$

Now, we can show that with probability at least $1-\delta$, $\forall h \in \mathcal{H}$ we get:

$$\left|\sup_{\delta\in\Delta(\mathcal{W},\gamma)} \mathbb{E}_{w\sim\delta} R(h,\eta,w) - \sup_{\delta\in\Delta(\mathcal{W},\gamma)} \mathbb{E}_{w\sim\delta} \hat{R}_D(h,\eta,w)\right| \tag{21}$$

$$\lesssim \frac{M}{\alpha_0}\sqrt{\frac{\gamma+\log(\mathcal{N}(\epsilon,R(\cdot,\cdot,w),\|\cdot\|_\infty)/\delta)+\log n}{2n-1}} \tag{22}$$

$$\lesssim \frac{M}{\alpha_0}\sqrt{\frac{\gamma+\left(\left(\frac{M}{\alpha_0}\sqrt{\frac{\gamma+\log(1/\delta)+\log n}{2n-1}}\right)^{-(d/\alpha)}\right)+\log(1/\delta)+\log n}{2n-1}} \tag{23}$$

Note that in the above bound we cannot see if this upper bound shrinks as $n\to\infty$, without assuming something very strong about α. Thus, we need covering number bounds that do not grow exponentially with the input dimension. And for this we turn to parameterized classes, which is the next approach we take. It is more for the convenience of analysis that we introduce the following parameterization.

Approach 2:

Let $l(\cdot,\cdot)$ be a $[0,M]$ bounded L-Lipschitz function in $\|\cdot\|_2$ over Θ where $\mathcal{H}$ be parameterized by a convex subset $\Theta\subset\mathbb{R}^d$. Thus we need to get a covering of the loss function $\sup_\delta \mathbb{E}_{w\sim\delta} R(\theta,\eta,w)$ in $\|\cdot\|_\infty$ norm, for a radius ϵ. A standard practice is to bound this with a covering $\mathcal{N}(\Theta,\frac{\epsilon}{L},\|\cdot\|_2)$, where $\|\cdot\|_2$ is Euclidean norm defined on $\Theta\subset\mathbb{R}^d$.

Lemma G.6 (Covering number for $\mathcal{N}(\Theta\times[0,M], \frac{\epsilon}{L}, \|\cdot\|_2)$ [55])**.** *Let Θ be a bounded convex subset of $\mathbb{R}^d$ with .*

$$\mathcal{N}(\epsilon/L,\Theta,\|\cdot\|) \lesssim \left(1+\frac{L}{\epsilon}\right)^{d+1} \tag{24}$$

We now re-iterate the steps we took previously:

$$\left|\sup_{\delta\in\Delta(\mathcal{W},\gamma)} \mathbb{E}_{w\sim\delta} R(h,\eta,w) - \sup_{\delta\in\Delta(\mathcal{W},\gamma)} \mathbb{E}_{w\sim\delta} \hat{R}_D(h,\eta,w)\right| \tag{25}$$

$$\lesssim \frac{M}{\alpha_0}\sqrt{\frac{\gamma+\log(\mathcal{N}(\epsilon,R(\cdot,\cdot,w),\|\cdot\|_\infty)/\delta)+\log n}{2n-1}} \tag{26}$$

$$\lesssim \frac{M}{\alpha_0}\sqrt{\frac{\gamma+\log\left(1+\frac{L}{\sqrt{\gamma/n}}\right)^{d+1}+\log(1/\delta)+\log n}{2n-1}} \tag{27}$$

$$\lesssim \frac{M}{\alpha_0}\sqrt{\frac{\gamma+(d+1)\log\left(\frac{L^2 n}{\gamma}\right)+\log(1/\delta)+\log n}{2n-1}} \tag{28}$$

Note that the above holds with probability atleast $1-\delta$ and for $\forall h, \eta$. Thus, we can apply it twice:

$$\left| \sup_{\delta \in \Delta(\mathcal{W}, \gamma)} \mathbb{E}_{w \sim \delta} R(\hat{h}_D^\gamma, \hat{\eta}_D^\gamma, w) - \sup_{\delta \in \Delta(\mathcal{W}, \gamma)} \mathbb{E}_{w \sim \delta} \hat{R}_D R(\hat{h}_D^\gamma, \hat{\eta}_D^\gamma, w) \right| \lesssim \frac{M}{\alpha_0} \sqrt{\frac{\gamma + (d+1) \log\left(\frac{L^2 n}{\gamma}\right) + \log(1/\delta) + \log n}{2n-1}}$$

$$\left| \sup_{\delta \in \Delta(\mathcal{W}, \gamma)} \mathbb{E}_{w \sim \delta} R(h^*, \eta^*, w) - \sup_{\delta \in \Delta(\mathcal{W}, \gamma)} \mathbb{E}_{w \sim \delta} \hat{R}_D R(h^*, \eta^*, w) \right| \lesssim \frac{M}{\alpha_0} \sqrt{\frac{\gamma + (d+1) \log\left(\frac{L^2 n}{\gamma}\right) + \log(1/\delta) + \log n}{2n-1}}$$

where h^*, η^* are the optimal for $\mathcal{L}^*_{\text{cvar}}$. Combining the two above proves the statement in Theorem B.1.

G.4 Proof of Theorem B.2

Setup. Let us assume there exists a prior Π such that $\mathcal{W}(\gamma)$ in Definition 2.4 is given by an RKHS induced by Mercer kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, s.t. the eigenvalues of the kernel operator decay polynomially:

$$\mu_j \lesssim j^{-2/\gamma}$$

for ($\gamma < 2$). We solve for $\hat{h}_D^\gamma, \hat{\eta}_D^\gamma$ by doing kernel ridge regression over norm bounded ($\|f\|_{\mathcal{W}(\gamma)} \leq M$) smooth functions f. Thus, $\mathcal{W}(\gamma)$ is compact.

$$\underset{w \in \mathcal{W}(\gamma)), \|w\|_{\mathcal{W}(\gamma)} \leq R}{\arg\max} R(h, \eta, w) = \underset{w \in \mathcal{W}(\gamma)), \|w\|_{\mathcal{W}(\gamma)} \leq R}{\arg\max} \langle l(h) - \eta, w \rangle_P + \eta \tag{29}$$

$$\underset{w \in \mathcal{W}(\gamma), \|w\|_{\mathcal{W}(\gamma)} \leq R}{\arg\max} \mathbb{E}_P \mathbb{1}((l(h) - \eta) \cdot w > 0) \tag{30}$$

We show that we can control: (i) the pessimism of the learned solution; and (ii) the generalization error (Theorem B.2). Formally, we refer to pessimism for estimates $\hat{h}_D^\gamma, \hat{\eta}_D^\gamma$:

$$\text{excess risk or pessimism:} \quad \sup_{w \in \mathcal{W}(\gamma)} |\inf_{h, \eta} R(h, \eta, w) - R(\hat{h}_D^\gamma, \hat{\eta}_D^\gamma, w)|$$

Theorem G.7 ((restated for convenience) bounded RKHS)**.** *For $l, \mathcal{H}$ in Theorem B.1, and for $\mathcal{W}(\gamma)$ described above $\exists\, \gamma_0$ s.t. for all sufficiently bitrate-constrained $\mathcal{W}(\gamma)$ i.e., $\gamma \leq \gamma_0$, w.h.p. $1-\delta$ worst risk generalization error is $\mathcal{O}\left((1/n)\left(\log(1/\delta) + (d+1)\log(nR^{-\gamma} L^{\gamma/2})\right)\right)$ and the excess risk is $\mathcal{O}(M)$ for $\hat{h}_D^\gamma, \hat{\eta}_D^\gamma$ above.*

Generalization error proof:

Note that the objective in equation 30 is a non-parametric classification problem. We can convert this to the following non-parametric regression problem, after replacing the expectation with plug-in $\hat{P}_n$.

$$\inf_{w \in \mathcal{W}(\gamma)), \|w\|_{\mathcal{W}(\gamma)} \leq R} \frac{1}{n} \sum_{i=1}^{n} (w(x_i, y_i) - (l(h(x_i), y_i) - \eta) + \epsilon_i)^2 + \lambda_n \|w\|^2_{\mathcal{W}(\gamma)} \tag{31}$$

where $\lambda_n \to 0$ as $n \to \infty$. Essentially, for non-parametric kernel ridge regression regression the regularization can be controlled to scale with the critical radius, that would give us better estimates and tighter localization bounds as we will see.

Note that in the above problem we add variable ϵ_i which represents random noise $\sim \mathcal{N}(0, \sigma_2)$. Let $\sigma_2 = 1$ for convenience. Since the noise is zero mean and random, any estimator maximizing the above objective on $\hat{P}_n$ would be consistent with the estimator that has a noise free version. We can also thing of this as a form regularization (similar to λ), if we consider the kernel ridge regression problem as the means to obtain the Bayesian predictive posterior under a Bayesian prior that is a Gaussian Process $\mathcal{GP}(\mathbf{0}, \sigma_\mathbf{2}\mathbf{k}(\mathbf{x}, \mathbf{x}))$, under the same kernel as defined above.

First we will show estimation error bounds for the following KRR estimate:

$$\hat{w}_D^\gamma = \underset{w \in \mathcal{W}(\gamma)), \|w\|_{\mathcal{W}(\gamma)} \leq R}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (w(x_i, y_i) - (l(h(x_i), y_i) - \eta) + \epsilon_i)^2 + \lambda_n \|w\|^2_{\mathcal{W}(\gamma)} \tag{32}$$

The estimation error would be measured in terms of $\hat{P}_n$ norm *i.e.,* $\|\hat{w}_D^\gamma - w^*\|_{\hat{P}_n}$ where

$${w_*}^\gamma(x, y) = \underset{w \in \mathcal{W}(\gamma)), \|w\|_{\mathcal{W}(\gamma)} \leq R}{\arg\min} \quad \mathbb{E}_P \mathbb{E}_\epsilon ((l(h(x), y) - \eta) - w(x, y) + \epsilon)^2$$

is the best solution to the optimization objective in population.

Next steps:

- First, we get the estimation error in $\|\hat{w}_D^\gamma - {w_*}^\gamma\|_{\hat{P}_n}$ of $\hat{P}_n$.
- Then using uniform laws [55] we can extend it to $L^2(P)$ norm *i.e.,* $\|\hat{w}_D^\gamma - w^*\|_p$.
- Then we shall prove that if we convert the $\hat{w}_D^\gamma$ and w^* into prediction rules: $\hat{w}_D^\gamma \geq 0$ and ${w_*}^\gamma$, then we can get the estimation error of prdedictor $\hat{w}_D^\gamma \geq 0$ with respect to the optimal decision rule ${w_*}^\gamma \geq 0$ in class $\mathcal{W}(\gamma)$.
- The final step would give us an oracle inequality of the form in Theorem B.1.

Based on the outline above, let us start with getting $\|\hat{w}_D^\gamma - w^*\|_{\hat{P}_n}$. For this we shall use concentration inequalities from localization bounds (see Lemma G.8). Before we use that, we define the quantity δ_n, which is the critical radius (see Ch. 13.4 in [55]). For convenience, we also state it here. Formally, δ_n is the smallest value of δ that satisfies the following inequality (critical condition):

$$\frac{\mathcal{R}_n(\delta)}{\delta} \leq \frac{R}{2} \cdot \delta \tag{33}$$

where,

$$\mathcal{R}_n(\delta) := \mathbb{E}_\epsilon \left[\sup_{g \in (\mathcal{F} - f^*), \|g\|_\mathcal{F} \leq R,\ \|g\|_{\hat{P}_n} \leq \delta} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(x_i, y_i) \cdot l(h(x_i) - y_i) \right| \right]$$

and ϵ is some sub-Gaussian zero mean random variable.

Lemma G.8 ([55])**.** *For some convex RKHS class $\mathcal{F}$ Let $\hat{f}$ be defined as:*

$$\hat{f} \in \underset{f \in \mathcal{F}, \|f\|_\mathcal{F} \leq R}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda_n \|f\|^2_\mathcal{F} \right\}$$

then, with probability $\geq 1 - c_2 \exp\left(-c_3 \frac{nR^2\delta_n^2}{\sigma^2}\right)$ *and when* $\lambda_n \geq \delta_n^2$ *we get:*

$$\|\hat{f} - f^*\|_2^2 \leq c_0 \inf_{\|\mathcal{F}\| \leq R} \|f - f^*\|_n^2 + c_1 R^2 (\delta_n^2 + \lambda_n).$$

Note that it is standard exercise in statistics to derive the following closed form for the problem in equation 32:

$$\hat{w}_D^\gamma(\cdot) = \hat{K}_n(\cdot, Z)(\hat{K}_n^T\hat{K}_n + \lambda_n I)^{-1}\left(l(h)_D - \epsilon_D\right)$$

where $l(h)_D$ is the loss vector and ϵ_D is the noise vector for dataset $\mathcal{D}$ and $\hat{K}_n$ is the empirical kernel matrix given by $\hat{K}_{i,j} = \frac{1}{n}k((x_i, y_i), (x_j, y_j))$, and Z is a matrix of (x, y) pairs in dataset.

Corollary G.9. *[55] Let $\hat{\mu}_j$ be the eigen values $\hat{\mu}_1 \geq \hat{\mu}_2 \ldots \geq \hat{\mu}_n$ for the empirical Kernel matrix $\hat{K}$, then we have for any δ satisfying*

$$\sqrt{\frac{2}{n}\left(\sum_{i=1}^{n} \min(\delta^2, \hat{u}_j)\right)} \;\leq\; \frac{R}{4}\delta^2$$

, it is necessary that δ satisfies the critical condition in equation 33.

To show the above critical condition we shall now use the polynomial decaying property that for the specific kernel induced by $\mathcal{W}(\gamma)$, as stated in our assumption in the beginning of this section. For this we take standard approach taken for polynomial decay kernels [61]. Let $\exists C$ for some large $C > 0$ such that $\hat{\mu}_j \leq Cj^{-2/\gamma}$. Then for some k, such that $\delta^2 \geq ck^{-2/\gamma}$

$$\begin{aligned}
\sqrt{\frac{1}{n}\left(\sum_{j=1}^{n} \min(\delta^2, \hat{\mu}_j)\right)} \quad &\lesssim \quad \sqrt{\frac{2}{n}\left(\sum_{i=1}^{n} \min(\delta^2, Cj^{-2/\gamma})\right)} \\
\lesssim \sqrt{\frac{2}{n}\left(k\delta^2 + C\sum_{j=k+1}^{n} j^{-2/\gamma})\right)} \quad &\lesssim \quad \sqrt{\frac{2}{n}\left(k\delta^2 + C\sum_{j=k+1}^{\infty} j^{-2/\gamma})\right)} \\
\lesssim \sqrt{\frac{2}{n}\left(k\delta^2 + C\int_{j=k+1}^{\infty} z^{-2/\gamma}\,dz)\right)} \quad &\lesssim \quad \sqrt{\frac{2}{n}\left(k\delta^2 + Ck^{-2/\gamma+1}\,dz)\right)} \\
\leq \sqrt{2/n}\left(\sqrt{k}\cdot\delta\right) \leq \frac{1}{\sqrt{n}}\cdot\delta^{1-\gamma/2}
\end{aligned}$$

Now, setting the above into the critical condition equation from Corollary above:

$$\begin{aligned}
\frac{1}{\sqrt{n}}\cdot\delta^{1-\gamma/2} &\leq \frac{R}{4}\delta^2 \\
\implies \delta^{1+\gamma/2} &\geq \frac{1}{\sqrt{n}R}
\end{aligned}$$

This tells us that:

$$\delta_n^2 \gtrsim \left(\frac{1}{nR^2}^{\frac{2}{\gamma+2}}\right) \tag{34}$$

is the critical radius.

We shall later plug this into the bound we have into a uniform bound over the concentration inequality in Lemma G.8. The reason we need a uniform bound over Lemma G.8 is that in its current form, it only bounds $\|\hat{w}_D^\gamma - w_*^\gamma\|^2_{\hat{P}_n}$ for a specific choice of η, h. In order to arrive at the worst risk generalization error of the form we have in Theorem B.1 we need to satisfy that with high probability $1-\delta\ \forall\eta, h$, a critical concentration bound of the form in Lemma G.8 but over $\sup_{\eta,h}\|\hat{w}_D^\gamma - w_*^\gamma\|^2_{\hat{P}_n}$.

Let $\epsilon = c_2\exp\left(c_3 nR^2\frac{\delta_n{}^2}{\sigma^2}\right)$. Since δ_n^2 needs to be large enough (see condition in equation 34), we use Lemma G.8 in the following bound, incorporating δ_n condition we derived.

With high probability $1-\epsilon$:

$$\|\hat{w}_D^\gamma - w_*^\gamma\|_{\hat{P}_n}^2 \lesssim \inf_{w\in\mathcal{W}(\gamma),\|w\|\le R} \|w-w_*^\gamma\|_{\hat{P}_n}^2 + R^2 \max\left(\left(\frac{1}{nR^2}\right)^{\frac{\gamma+2}{\gamma}}, \left(\log(1/\epsilon)\frac{1}{nR^2}\right)\right) \tag{35}$$

To apply uniform convergence argument on the above we would need to apply a union bound on a covering of $\Theta \times [0, M]$, so that we get the probability bound to hold for all η, h.

For this we use the same technique as in the proof of Theorem B.1. First, we shall use Lemma G.6 to get a covering number bound for bounded convex subset Θ of $\mathbb{R}^d$ that parameterizes the learner (Theorem B.2) .

$$\mathcal{N}(\beta/L, \Theta\times[0,M], \|\cdot\|) \lesssim \left(1+\frac{L}{\beta}\right)^{d+1} \tag{36}$$

And we know that a covering of $\Theta \times [0, M]$ in radius β/L, will fetch a covering for $l(h) - \eta$ in β, since we assume $l(\cdot)$ to be Lipschitz in θ. Thus, all we need to prove bound equation 35 holds uniformly is to get a covering in radius $R^2 \max\left(\left(\frac{1}{nR^2}\right)^{\frac{2}{\gamma+2}}, \left(\log(1/\epsilon)\frac{1}{nR^2}\right)\right)$. Thus, a acovering in $R^2\left(\left(\frac{1}{nR^2}\right)^{\frac{2}{\gamma+2}}\right)$. Thus, the number of elements in cover are:

$$J = \left(1+\frac{L}{\left(R^2\left(\frac{1}{nR^2}^{\frac{2}{\gamma+2}}\right)\right)}\right)^{d+1}$$

For union bound we need:

$$\begin{aligned}
&J\epsilon/c_2 = \exp\left(-c_3 nR^2\delta_n^2\right)\\
\implies &\log(\frac{1}{\epsilon}) + \log J \gtrsim c_3 nR^2\delta_n^2\\
\implies &\log(\frac{1}{\epsilon}) + (d+1)\log\left(\frac{L}{\left(R^2\left(\left(\frac{1}{nR^2}\right)^{\frac{2}{2+\gamma}}\right)\right)}\right) \gtrsim c_3 nR^2\delta_n^2\\
\implies &\log(\frac{1}{\epsilon}) + (d+1)\log\left(\left(LR^{-2}\right)^{\frac{\gamma+2}{2}} nR^2\right) \gtrsim c_3 nR^2\delta_n^2
\end{aligned}$$

The uniform convergence bound that we get is $R^2\max\left(\left(\frac{1}{nR^2}\right)^{\frac{\gamma+2}{\gamma}}, \left(\log(J/\epsilon)\frac{1}{nR^2}\right)\right)$. In the above sequence of steps we have shown that, due to the size of J, the second term would be maximum, or at least there exists a γ_0, such that the second term would be higher for all $\gamma \ge \gamma_0$, for any sample size. Thus, we get the following probabilistic uniform convergence. With probability $\ge 1-\epsilon, \forall \eta, h$:

$$\begin{aligned}
\|\hat{w}_D^\gamma - w_*^\gamma\|_{\hat{P}_n}^2 &\lesssim \inf_{w\in\mathcal{W}(\gamma),\|w\|\le R} \|w-w_*^\gamma\|_{\hat{P}_n}^2 && (37)\\
&\lesssim \frac{1}{n}\log(\frac{1}{\epsilon}) + (d+1)\log\left(\left(LR^{-2}\right)^{\frac{\gamma+2}{2}} nR^2\right) && (38)\\
&\lesssim \frac{1}{n}\log(\frac{1}{\epsilon}) + (d+1)\log\left(\left(L^{\gamma/2}R^{-\gamma}\right) n\right) && (39)\\
& && (40)
\end{aligned}$$

Applying the above twice, once one $\hat{w}_D^\gamma$ and another on w_*^γ we prove the generalization bound in Theorem B.2.

<u>Excess risk bound:</u>

In the same setting we shall now prove the excess risk bound. Recall the definition of excess risk:

$$\text{excess risk} := \sup_{w \in \mathcal{W}(\gamma)} |\inf_{h,\eta} R(h, \eta, w) - R(\hat{h}_D^\gamma, \hat{\eta}_D^\gamma, w)|.$$

Let $h^*(w), \eta^*(w) = \inf_{h,\eta} R(h, \eta, w)$, then:

$$\begin{aligned}
\text{excess risk} &= \sup_{w \in \mathcal{W}(\gamma)} |\inf_{h,\eta} R(h, \eta, w) - R(\hat{h}_D^\gamma, \hat{\eta}_D^\gamma, w)| && (41)\\
&\leq \sup_{w \in \mathcal{W}(\gamma)} \Big(R(h^*(w), \eta^*(w), w) - R(\hat{h}_D^\gamma, \hat{\eta}_D^\gamma, w) \Big) && (42)\\
&\leq \sup_{w \in \mathcal{W}(\gamma)} \left(\frac{1}{\alpha_0} \langle l(h^*) - l(\hat{h}_D^\gamma) - (\eta^*(w) - \hat{\eta}_D^\gamma), w \rangle_P \right) && (43)\\
&\leq \frac{M}{\alpha_0} \sup_{w \in \mathcal{W}(\gamma)} \big((\|w\|_{L_2(P)}) \big) && (44)
\end{aligned}$$

Note, that according to our assumption $\|w\|_{\mathcal{W}(\gamma)} \leq B$ *i.e.,* the smooth functions are bounded in RKHS norm. The following lemma relates bounds in RKHS norm to bound in $L_2(P)$ bound for kernels with bounded operator norms:

Lemma G.10. *For an RKHS $\mathcal{H}_k$ with norm $\|\cdot\|_{\mathcal{H}_k}$:*

$$\|f\|_{L^2(P)} = \|T_K^{1/2} f\|_{\mathcal{H}_k} \leq \sqrt{\|T_K^{1/2}\|_{op}} \cdot \|f\|_{\mathcal{H}_k}$$

Proof:

$$\begin{aligned}
\|T_K^{1/2} f\|_{\mathcal{H}_k}^2 &= \langle T_K^{1/2} f, T_K^{1/2} f \rangle_{\mathcal{H}_k} = \langle f, T_K f \rangle_{\mathcal{H}_k}\\
&= \sum_{j=1}^{\infty} \frac{\langle \phi_j, f \rangle_{L^2(P)}, \langle \phi_j, T_K f \rangle_{L^2(P)}}{\lambda_j}\\
&= \|f\|_{L^2(P)}^2
\end{aligned}$$

In the above λ_j are the Eigen values of the kernel and the Eigen functions ϕ_j are orthonormal and span $L^2(P)$/ Thus, $\|f\|_{L^2(P)} \leq \|T_K^{1/2}\|_{op} \cdot \|f\|_{\mathcal{H}_k}$. Since we assume polynomially decaying Eigen values for our kernel, it is easy to see that $\|T_K^{1/2}\|_{op} = \mathcal{O}(1)$.

Applying Lemma G.10 to equation 44, directly gives us the excess risk bound and completes the proof.

$$\text{excess risk} \lesssim \|T_K^{1/2}\|_{op} \cdot B = \mathcal{O}(B)$$

G.5 Proof of Theorem B.4

Setup. The algorithm is as follows: Consider a two-player zero-sum game where the learner uses a no-regret strategy to first play $h \in \mathcal{H}, \eta \in \mathbb{R}$ to minimize $\mathbb{E}_{w \sim \delta} R(h, \eta, w)$. Then, the adversary plays follow the regularized leader (FTRL) strategy to pick distribution $\delta \in \Delta(\mathcal{W}, \gamma)$ to maximize the same. The regularizer used is a negative entropy regularizer. Our goal is to analyze the bitrate-constraint γ's effect on the above algorithm's convergence rate and the pessimistic nature of the solution found. For this, we need to first characterize the bitrate-constraint class $\mathcal{W}(\gamma)$. So we assume there exists a prior Π such that $\mathcal{W}(\gamma)$ is Vapnik-Chervenokis (VC) class of dimension $O(\gamma)$.

Note that $R(h, \eta, w)$ is convex in h and linear in η, l. Thus, as we discuss in the derivation for equation 6 this objective optimized over convex sets has a unique saddle point (Nash equilibrium) by Weierstrass's theorem. Thus, to avoid repetition we only discuss the proofs for the other two claims on convergence and excess risk.

Convergence:

Given that $\mathcal{W}(\gamma)$ is a VC class of dimension $C\gamma$ for some large C, we can use Sauer-Shelah [6] Lemma (stated) below to bound the total number of groups that can be identified by $\mathcal{W}(\gamma)$ in n points.

Lemma G.11 (Sauer's Lemma)**.** *The Vapnik-Chervonenkis dimension of a class $\mathcal{F}$, denoted as VC-dim($\mathcal{F}$), and it is the cardinality of the largest set S shattered by $\mathcal{F}$. Let $d = VC - dim(\mathcal{F})$, then for all m, $C[m] = \mathcal{O}(m^d)$*

Thus, the total number of groups that can be proposed on n points by $\mathcal{W}(\gamma)$ is $\mathcal{O}(n^\gamma)$. A similar observation was made in Kearns et al. [27]. Different from them, our goal is to analyze the algorithm iterates for our solver described above and bound its pessimism.

First, for convergence rate we show that the above algorithm has a low regret—a standard exercise in online convex optimization. Note that any distribution picked by the adversary can be seen as multinomial over a finite set of possible groups that is let's say K, and from discussion above we know that $K = O(n^\gamma)$. Further, the negative entropy regularizer is given as:

$$B(\delta) \coloneqq c \cdot \sum_{i=1}^{K} \delta_i \log \delta_i \tag{45}$$

where the sum is over total possible groups identified by $\mathcal{W}(\gamma)$. Let the probability assigned to group i be denoted as δ_i. The FTRL strategy for adversary is given as:

$$\delta_T = \underset{\delta \in \Delta(\mathcal{W}(\gamma))}{\arg\min} \sum_{t=1}^{T-1} \frac{1}{\alpha_0} \langle l(h_t) - \eta_t, \delta_t \rangle_{\hat{P}_n} + \eta + c \cdot \sum_{i=1}^{K} \delta_i \log \delta_i \tag{46}$$

Then the regret for not having picked a single action δ is given as:

$$\text{REGRET}_T(\delta) \coloneqq \sum_{t=1}^{T} \frac{1}{\alpha_0} \langle l(h_t) - \eta_t, \delta_t - \delta_{t+1} \rangle_{\hat{P}_n} + B(\delta) - B(\delta_1) \tag{47}$$

We bound the two terms in the above bound separately. With $\sum_{k=1}^{k} \delta_k = 1$, we get the strong dual for the FTRL update above as:

$$\sum_{t=1}^{T-1} \frac{1}{\alpha_0} \langle l(h_t) - \eta_t, \delta_t \rangle_{\hat{P}_n} + \eta + c \cdot \sum_{i=1}^{K} \delta_i \log \delta_i + \lambda \cdot (\sum_{i=1}^{K} \delta_i - 1) \tag{48}$$

Solving we get:

$$\delta_t(k) = \frac{\exp\left(\frac{-1}{c}\right) \sum_{t=1}^{t-1} \mathbb{E}_{\hat{P}_n} \frac{1}{\alpha_0} (l(h_t) - \eta_t | G_k) + \eta/K}{\sum_{k=1}^{K} \exp\left(\frac{1}{\alpha_0} \frac{-1}{c}\right) \sum_{k=1}^{t-1} (\mathbb{E}_{\hat{P}_n} \frac{1}{\alpha_0} (l(h_t) - \eta_t | G_k) + \eta/K)} \tag{49}$$

where $\mathbb{E}_{\hat{P}_n}(l(h_t) - \eta_t | G_k)$ is the expected empirical loss in group G_k and $\delta_t(k)$ is the adversary's distribution at time step t for the k^{th} group.

Claim on stability:

$$\frac{1}{\alpha_0} \langle l(h_t) - \eta_t, \delta_t - \delta_{t+1} \rangle_{\hat{P}_n} \leq 1/c \tag{50}$$

The above statement is true because,

$$\delta_{t+1}(i) = \delta_t(i) \cdot \exp\left(\frac{1}{\alpha_0 c} \mathbb{E}[l(h_t) - \eta_t | G_i] + \eta_t / K \right) \tag{51}$$

Thus, if $l(h_t) \in [0, M/\alpha_0]$, *i.e.,* losses are bounded then:

$$\delta_{t+1}(i) \geq \delta_t(i) \cdot e^{-1/c} \geq \delta_t(i) \cdot (1 - 1/c). \tag{52}$$

and our stability claim is easy to see. Thus, we have bounded the first term in our regret bound above. Further, we can to see that $B(x) - B(x_1) \leq c \log K$. Thus, we have bounded both terms in the regret bound above in terms of c.

$$\text{REGRET}_T \leq (T/c) + (c \log K) \tag{53}$$

Setting $c = \sqrt{\frac{T}{\log K}}$, we get:

$$\frac{\text{REGRET}_T}{T} \leq \sqrt{\frac{\log K}{T}} \tag{54}$$

Now, our VC claim gave $K = \mathcal{O}(n^\gamma)$. Hence,

$$\frac{\text{REGRET}_T}{T} = \mathcal{O}\sqrt{\frac{\gamma \log n}{T}} \tag{55}$$

Next, we use Theorem 9 from Abernethy et al. [1] that maps low regret $O(\epsilon)$ algorithms in zero-sum convex-concave games to ϵ-optimal equilibriums.

Let regret be ϵ, then applying their theorem gives us:

$$V^* - \epsilon \leq \inf_{h \in \mathcal{H}, \eta \in \mathbb{R}} R_D(h, \eta, \bar{\delta}_T) \leq V* \leq \sup_{\delta \in \Delta(\mathcal{W}(\gamma))} R_D(\bar{h}_T, \bar{\eta}_T, \delta) \leq V^* + \epsilon \tag{56}$$

where

$$V^* = R_D(h^*_D(\gamma), \eta^*_D(\gamma), \delta^*_D(\gamma)) = \inf_{h \in \mathcal{H}, \eta \in \mathbb{R}} \sup_{\delta \in \Delta(\mathcal{W}(\gamma))} \frac{1}{\alpha_0} \langle l(h) - \eta, \delta \rangle + \eta \tag{57}$$

Excess risk:

For excess risk we need to bound:

$$\frac{1}{\alpha_0} \sup_{h \in \mathcal{H}, \eta \in \mathbb{R}} \left| \sup_{\delta \in \Delta(\mathcal{W}(\gamma))} \langle l(h) - \eta, \delta - \delta^*(\gamma) \rangle \right| \tag{58}$$

$$\leq \frac{M}{\alpha_0} \frac{1}{2} \text{TV}(\delta - \delta^*(\gamma)) \leq \frac{M}{2\alpha_0} (1 - 1/K) = \frac{M}{\alpha_0} \mathcal{O}(1 - 1/n^\gamma) \tag{59}$$

In the above argument we used the fact that at equilibrium, $\delta^*(\gamma)$ would be uniform over all possible distinct group assignments. This completes our proof of Theorem B.4.