APPENDIX OUTLINE

A IMPLEMENTING BR-DRO IN PRACTICE

A.1 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 (Alemi et al., 2016) (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 $\text{KL}(\delta \,||\, \pi)$ in the primal formulation in Definition 4.1). If the constraint is applied via the l_2 term we implement the adversary as a linear layer. In some cases (*e.g.,* Section 6.2) 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: $\bar{\boldsymbol{\theta}}_h = \frac{1}{T}\sum_{t=1}^T \boldsymbol{\theta}_h^{(t)}$, $\bar{\boldsymbol{\theta}}_w = \frac{1}{T}\sum_{t=1}^T \boldsymbol{\theta}_w^{(t)}$

A.2 BR-DRO OBJECTIVE IN EQUATION 5

When describing the actual BR-DRO objective in Equation 5, for brevity we used $\boldsymbol{\theta}_h$ to denote both the parameters of the learner and the learner itself (similarly for $\boldsymbol{\theta}_w$). Here, we describe in detail the parameterized version of the objective in Equation 4, and clarify the modeling of the adversary.

Let us denote the learner as $h(\cdot; \boldsymbol{\theta}_h) : \mathcal{X} \mapsto \mathcal{Y}$, and the class of learners $\mathcal{H} := \{h(\cdot; \boldsymbol{\theta}_h) : \boldsymbol{\theta}_h \in \Theta_h\}$. A learner is composed of two parts: i) a feature extractor $f : \mathcal{X} \mapsto \mathbb{R}^p$ that maps the input into a p-dimensional feature vector and; ii) a classifier $c : \mathbb{R}^p \mapsto \mathcal{Y}$ that maps the features into a predicted label. Similarly, in the bitrate-constrained class $\mathcal{W}(\gamma)$ each adversary is parameterized as $w(\cdot; \boldsymbol{\theta}_w) : \mathbb{R}^p \times \mathcal{Y} \mapsto [0,1]$ where with some overloading of notation we use $w(\mathbf{x}, \mathrm{y})$ to denote $w(f(\mathbf{x}), \mathrm{y})$ *i.e.,* the adversary only operates on the features output by the feature extractor $f(\mathbf{x})$ and the given label y[4]. Since the features are extracted by the deep neural network (and frozen), the adversary is implemented as either a two layer neural network with VIB constraint or a linear layer with l_2 constraint – both of which promote low bitrate functions satisfying our Assumption 4.2. We shall now describe each version of BR-DRO separately.

BR-DRO (VIB): For each label y, $w(f(\mathbf{x}), \mathrm{y})$ is a one hidden layer neural network with ReLU activations (1-layer VIB). The first layer takes as input feature vector $f(\mathbf{x})$ and outputs a $2d$ dimensional vector $\mathbf{u}$. The d dimensional latent encoding $\mathbf{z}_{\boldsymbol{\theta}_w}(\mathbf{x})$ (dependence on $\mathbf{x}$ is made explicit) is sampled from multivariate Normal $\mathbf{z}_{\boldsymbol{\theta}_w}(\mathbf{x}) \sim \mathcal{N}(\mathbf{u}_{\boldsymbol{\theta}_w}[:\mathrm{d}], \text{diag}(\mathbf{u}_{\boldsymbol{\theta}_w}[\mathrm{d}:]))$ where $\mathbf{u}_{\boldsymbol{\theta}_w}[:\mathrm{d}]$ is the mean and $\text{diag}(\mathbf{u}_{\boldsymbol{\theta}_w}[\mathrm{d}:])$ is a diagonal covariance matrix, both parameterized by parameter $\boldsymbol{\theta}_w$ (neural net). Following Alemi et al. (2016), an information bottleneck constraint is applied on the latent variable $\mathbf{z}_{\boldsymbol{\theta}_w}(\mathbf{x})$ in the form of a KL constraint with respect to standard Gaussian prior, with strength given by scalar β_{vib}. Prior works (Tishby & Zaslavsky,

[4]Note that, constraining the output space of the adversary to be bounded ($[0,1]$) does not necessarily deviate from our definition of $\mathcal{W}_{P,\kappa}$ in Equation 1, since we can always re-define the output space by dividing the $w \in \mathcal{W}_{P,\kappa}$ by $\max_{\boldsymbol{x},y} w(\boldsymbol{x}, y), \forall w \in \mathcal{W}_{P,\kappa}$.

2015; Hjelm et al., 2018) have argued why this regularization would bias the adversary to learn low bitrate functions. If we assume the generative model for $\mathbf{x}$ as one defined by latent factors of variation (*e.g.*, orientation, background), then presumably the group identity is a function of these factors. Finally, an output layer with a sigmoid activation maps $\mathbf{z}_{\boldsymbol{\theta}_w}(\mathbf{x})$ into a weight between $[0,1]$. If we believe that the KL constraint on $\mathbf{z}_{\boldsymbol{\theta}_w}(\mathbf{x})$ helps recover some of these factors, then learning a linear transform (with sigmoid activation) over it would amount to learning a simple group function.

BR-DRO (l_2): For each label y, $w(f(\mathbf{x}),\mathrm{y})$ is a linear layer. It takes as input feature vector $f(\mathbf{x})$ and maps it to a scalar which when passed through a sigmoid yields a weight between $[0,1]$. The l_2 constraint over the linear layer parameters is controlled by scalar β_{l_2}. This corresponds to bitrate constraints under certain priors (Polson & Sokolov, 2019). We can incorporate the above parameterizations for VIB and l_2 versions of BR-DRO into the BR-DRO objective in Equation 4 through Equation 8 and Equation 9 respectively to yield the final objective in Equation 5. Note that as we mention in Section 4, we can switch between the two versions of BR-DRO by setting $\beta_{\text{vib}}=0$ (for l_2) or $\beta_{l_2}=0$ (for VIB). 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. Finally, while $\beta_{\text{vib}},\beta_{l_2}$ act as Lagrangian parameters for our bitrate constraint, η is the Lagrangian parameter for the constrains on $w(\mathbf{x},\mathrm{y})$ in the definition of $\mathcal{W}(\gamma)_{P,\kappa}$ in Equation 1.

$$\min_{h(\cdot;\boldsymbol{\theta}_h)\in\mathcal{H}} \mathbb{E}_P[l(h(\mathbf{x};\boldsymbol{\theta}_h),\mathrm{y})\cdot w^*(f(\mathbf{x}),\mathrm{y};\boldsymbol{\theta}^*_w)]$$

$$\text{s.t. } w^* = \operatorname*{argmax}_{w(\cdot;\boldsymbol{\theta}_w)\in\mathcal{W}(\gamma)} L_{\text{adv}}(\boldsymbol{\theta}_w;\boldsymbol{\theta}_h,\beta_{\text{vib}},\beta_{l_2},\eta)$$

$$L_{\text{adv}}(\boldsymbol{\theta}_w,\boldsymbol{\theta}_h,\beta_{\text{vib}},\beta_{l_2},\eta) = \mathbb{E}_P[(l(h(\mathbf{x};\boldsymbol{\theta}_h),\mathrm{y})-\eta)\cdot w(f(\mathbf{x}),\mathrm{y};\boldsymbol{\theta}_w)]$$

$$-\beta_{\text{vib}}\cdot\mathbb{E}_{\mathbf{x}\sim P(\mathbf{x})}[\mathrm{KL}(\mathbf{z}_{\boldsymbol{\theta}_w}(\mathbf{x})\,||\,\mathcal{N}(\mathbf{0},\mathbf{I_d}))] \tag{8}$$

$$-\beta_{l_2}\cdot\|\boldsymbol{\theta}_w\|_2^2 \tag{9}$$

A.3 Connecting bitrate-constrained $\mathcal{W}(\gamma)$ to simple groups (a practical example).

First, let us recall the definition of a simple group in Assumption 4.2. We defined a group G to be simple, if the indicator function $\mathbb{1}((\mathbf{x},\mathrm{y})\in G)$ identifying said group is containing in the bitrate-constrained class of functions $W(\gamma)$. Next, we shall see what this means in terms of a specific prior π and the constraint $\mathrm{KL}(\cdot\,||\,\pi)\leq\gamma$ that defines the class $W(\gamma)$ in Definition 4.1.

Let us assume that the adversary $(\boldsymbol{\theta}_w)$ is parameterized as a linear classifier in $\mathbb{R}^d$ where each parameter $\boldsymbol{\theta}_w$ corresponds to a re-weighting function in $\mathcal{W}_{P,\kappa}$ (Equation 1), and the prior π is designed to have a higher likelihood over low norm solutions. Specifically, π takes the form:

$$\pi(\boldsymbol{\theta}_w)\propto\exp(-\|\boldsymbol{\theta}_w\|_2^2).$$

Now consider the class of densities:

$$\Delta(\Theta_w) := \{\delta_{\boldsymbol{\theta}_w}(\boldsymbol{\theta})\propto\exp(-\|\boldsymbol{\theta}-\boldsymbol{\theta}_w\|_2^2) : \boldsymbol{\theta}_w\in\Theta_w\}.$$

Here, it is easy to verify that:

$$\mathbb{E}_{\boldsymbol{\theta}\sim\delta_{\boldsymbol{\theta}_w}}[\boldsymbol{\theta}] = \boldsymbol{\theta}_w.$$

Note, that while Definition 4.1 concerns with any $\delta\in\Delta(\mathcal{W})$, we restrict ourselves to the subset $\Delta(\Theta_w)\subset\Delta(\mathcal{W})$ for the sake of mathematical convenience, and find that even this set is rich enough to easily violate the bitrate-constraint as we shall see next.

Applying Definition 4.1 on the set $\Delta(\Theta_w)$ we get:

$$\mathcal{W}(\gamma) = \{\boldsymbol{\theta}_w : \mathrm{KL}(\delta_{\boldsymbol{\theta}_w}\,||\,\pi)\leq\gamma\}.$$

Further if we compute $\mathrm{KL}(\delta_{\boldsymbol{\theta}_w}\,||\,\pi)$, we find that $\mathrm{KL}(\delta_{\boldsymbol{\theta}_w}\,||\,\pi) = \|\boldsymbol{\theta}_w\|_2^2 + C$ (for some constant C). Thus, the bitrate constraint $\mathrm{KL}(\delta_{\boldsymbol{\theta}_w}\,||\,\pi)\leq\gamma$ directly transfers into an l_2 norm constraint on the mean parameter $\boldsymbol{\theta}_w$, and $\mathcal{W}(\gamma)$ is simply the set of parameters $\boldsymbol{\theta}_w\in\Theta_w$ that have their l_2 norms bounded above by some constant $\gamma-C$. The objective for the l_2 version of our adversary in Equation 5 reflects this form. Hence, this example connects norm constrained parametric adversaries to bitrate constrained simple group identity functions.

B ADDITIONAL EMPIRICAL RESULTS AND OTHER EXPERIMENT DETAILS

B.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 Liu et al. (2021); Idrissi et al. (2022) 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.

	Waterbirds				CelebA			
Method	$p=0.0$	$p=0.02$	$p=0.05$	$p=0.1$	$p=0.0$	$p=0.02$	$p=0.05$	$p=0.1$
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).

B.2 SYNTHETIC DATASET DETAILS

We follow the explicit-memorization setup in Sagawa et al. (2020) 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 $P_{(\text{noise})}$.

B.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.

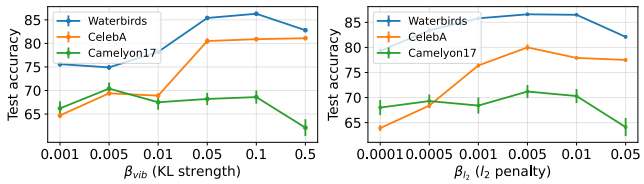

Figure 4: **Optimal bitrate-constraints for robustness to distributions shifts:** For two versions of capacity control: KL, l_2 penalty (see Section 4) we show how worst group performance on Waterbirds, CelebA and average performance on Camelyon test sets improves with increasing constraints under either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

B.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. (2021). 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. (2021) 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. (2021) 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.

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.

B.5 FINE-GRAINED EVALUATION OF WORST-CASE PERFORMANCE ON CIVILCOMMENTS.

The CivilComments dataset Koh et al. (2021) is a collection of comments from online articles, where each comment is rated for its toxicity, In addition, there is information available on 8 demographic identities: *male, female, LGBTQ, Christian, Muslim, other religions, Black and White* for each comment, *i.e.,* a given comment may be attributed to one or more of these demographic identities Koh et al. (2021). There are 133,782 instances in the test set. Each of the 8 identities form two groups based on the toxicity label, for a total of 16 groups. These are the groups used in the training of methods that assume group knowledge like Group DRO and also used in the evaluation. In Table 1 we report the accuracy over the worst group (on the test set) for each of the methods. Here, we do a more fine-grained evaluation of the worst-group performance. In addition to the 16 groups that methods are typically evaluated on, we evaluate the performance over groups created by combinations of two different demographic identities and a label (e.g., (*male*, *christian*, *toxic*) or (*female*, *Black*, *not toxic*)). Thus, the spurious attribute is no longer binary since it is categorical.

In Table 5 we plot the worst-group performance of different methods when evaluated over these $16+2\cdot {}^{8}\text{C}_2=72$ groups. Hence, this evaluation is verifying whether methods that may be robust to group shifts defined by binary attributes, are also robust to shifts when groups are defined by combinations of binary attributes. First, we find that the performance of every method drops, including that of the oracle Group DRO which assumes knowledge of the 8 demographic identities. This observation is in line with some of the findings in prior work Kearns et al. (2018). Next, similar to Table 1, we see that the performance of `BR-DRO` is still significantly better than CVaR DRO where the adversary is unconstrained. Finally, this experiment provides evidence that the bitrate-constraint does not restrict the adversary from identifying groups defined solely by binary attributes. Our simple group shift assumption (Assumption 4.2) is still satisfied when the spurious attribute is not binary as long as the group G corresponds to an indicator function $\mathbb{I}(\mathbf{x},y\in G)$ (e.g., an intersection of hyperplanes) that is realized in a low-bitrate class (e.g., a neural net with a VIB constraint which is how we model the adversary in the VIB version of BR-DRO).

B.6 COMPARING `BR-DRO` WITH OTHER BASELINES THAT DO NOT ASSUME ACCESS TO GROUP LABELS.

In Section 6 we compare `BR-DRO` with baselines JTT Liu et al. (2021), RWY Idrissi et al. (2022), LfF Nam et al. (2020) and CVaR DRO Levy et al. (2020) with regards to their performance on datasets that exhibit known spurious correlations (CelebA, Waterbirds and CivilComments) in Table 1 as well as on domain generalization datasets (FMoW and Camelyon17) that present unspecified covariate shifts in Table 2. None of the above baselines assume access to group annotations on training data. Here, we look at two additional baselines that also do not assume group annotations on training samples: George Sohoni

Method	Worst-group accuracy
ERM	52.8 (0.8)
LfF (Nam et al., 2020)	51.7 (1.0)
RWY (Idrissi et al., 2022)	61.9 (0.8)
JTT (Liu et al., 2021)	60.5 (0.9)
CVaR DRO (Levy et al., 2020)	56.5 (0.7)
BR-DRO (VIB) (ours)	62.9 (0.8)
BR-DRO (l_2) (ours)	62.5 (0.9)
Group DRO (Sagawa et al., 2019)	63.0 (0.8)

Table 5: **BR-DRO is robust to group shifts defined by multiple simple attributes:** On the CivilComments test set we evaluate the worst-group accuracy of BR-DRO and baselines when the groups can comprised of at most two demographic identities and the toxicity label. In $(\cdot)$ we report the standard error of the mean accuracy across five runs.

et al. (2020) and BPA Seo et al. (2022)[5]. Both these baselines employ a two-stage method to learn debiased representations where the first stage involves estimating biased pseudo-attributes using a clustering algorithm based on the observation that for a sufficiently trained model, non-target attributes tend to have similar representations. The second stage trains an unbiased model by optimizing a re-weighted objective where weights assigned to each cluster are updated with an exponential moving average, similar to Group DRO. BPA also accounts for the size of each cluster when assigning the weights.

We evaluate George and BPA on Waterbirds, CelebA and the hybrid dataset FMoW. While these methods were developed with the motivation of tackling group shifts along spurious attributes, following our experiments in Section 6.4, we also test how they do on domain generalization kind of tasks. The comparisons with both versions of BR-DRO are presented in Table 6. On the worst-group accuracy metric, we find that BR-DRO outperforms George on all datasets and BPA on Waterbirds and FMoW, while being comparable with BPA on CelebA. Since both these methods do not up-weight arbitrary points with high losses, we can think of them as having an implicit constraint on their weighting schemes (adversary), thus yielding solutions that are less pessimistic than CVaR DRO. At the same time, unlike the robust set outlined in our Assumption 4.2, and the excess risk results (in Section 5), it is unclear what the precise robust sets are for BPA and George, as well as the excess risk of their learned solutions.

Method	Waterbirds		CelebA		FMoW	
	Avg	WG	Avg	WG	Avg	WG
George (Sohoni et al., 2020)	94.8 (0.4)	77.3 (0.6)	92.8 (0.3)	64.9 (0.7)	50.5 (0.4)	30.6 (0.5)
BPA (Seo et al., 2022)	93.7 (0.5)	85.2 (0.4)	88.0 (0.4)	81.7 (0.5)	51.3 (0.3)	30.7 (0.3)
BR-DRO (VIB) (ours)	94.1 (0.2)	86.3 (0.3)	86.7 (0.2)	80.9 (0.4)	52.0 (0.2)	31.8 (0.2)
BR-DRO (l_2) (ours)	93.8 (0.2)	86.4 (0.3)	87.7 (0.3)	80.4 (0.6)	53.1 (0.1)	32.3 (0.2)

Table 6: We compare the performance of baseline methods George (Sohoni et al., 2020) and BPA (Seo et al., 2022) with BR-DRO on Waterbirds, CelebA and FMoW. We show both average (Avg) and worst group (WG) accuracy. For FMoW, we show the worst-region (W-Reg) accuracy. In $(\cdot)$ we report the standard error of the mean accuracy across five runs.

[5]For both baselines, we use the publicly available implementations from the authors of the original works.

C 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 Section 5, in the order in which they appear. Before we get into those we provide proof for our Corollary 3.2 and the derivation of Bitrate-Constrained CVaR DRO in Equation 6.

Lemma C.1 (Hoeffding bound Wainwright (2019)). *Let $X_1,...,X_n$ be a set of μ_i centered independent sub-Gaussians, each with parameter σ_i. Then for all $t \geq 0$, we have*

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_i)\geq t\right] \leq \exp\left(-\frac{n^2t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right). \tag{10}$$

Lemma C.2 (Lipschitz functions of Gaussians Wainwright (2019)). *Let $X_1,...,X_n$ 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:*

$$\mathbb{P}[|f(X)-\mathbb{E}[f(X)]|\geq t] \leq 2\cdot\exp\left(-\frac{t^2}{2L^2}\right), \forall t \geq 0. \tag{11}$$

C.1 PROOF OF COROLLARY 3.2

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:

$$K = \min\{|\{G_1,...,G_M\}| : p(\mathbf{x},\mathrm{y}\,|\,G_m) = q(\mathbf{x},\mathrm{y}\,|\,G_m) > 0, \forall(\mathbf{x},\mathrm{y}) \in G_m \ \forall m \in [M]\} \tag{12}$$

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:

$$q(\mathbf{x}_1,\mathrm{y}_1) = \frac{Q(G)}{P(G)} \cdot p(\mathbf{x}_1,\mathrm{y}_1) = \frac{Q(G')}{P(G')} \cdot p(\mathbf{x}_1,\mathrm{y}_1) \tag{13}$$

$$q(\mathbf{x}_2,\mathrm{y}_2) = \frac{Q(G)}{P(G)} \cdot p(\mathbf{x}_2,\mathrm{y}_2) = \frac{Q(G'')}{P(G'')} \cdot p(\mathbf{x}_2,\mathrm{y}_2) \tag{14}$$

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 $q(\mathbf{x},\mathrm{y}\,|\,G' \cup G'') = p(\mathbf{x},\mathrm{y}\,|\,G' \cup G'')$.

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.

C.2 Derivation of Bitrate-Constrained CVaR DRO in Equation 6

Recall that we define $\mathcal{W}$ as the set of all measurable functions $w\colon \mathcal{X}\times\mathcal{Y}\mapsto[0,1]$, since the other convex restrictions in Equation 1 are handled by dual variable η. As in Section 4, $\mathcal{W}(\gamma)$ is derived from the new $\mathcal{W}$ using Definition 4.1. With that let us first state the CVaR objective (Levy et al., 2020).

$$\begin{aligned}&\mathcal{L}_{\text{cvar}}(h,P) \coloneqq \sup_{q} \int_{\mathcal{X}\times\mathcal{Y}} q(\mathbf{x},\mathrm{y})\cdot l(h)\\ &\text{s.t. } q\geq 0,\ \|q/p\|_\infty \leq (1/\alpha_0),\ \int_{\mathcal{X}\times\mathcal{Y}} q(\mathbf{x},\mathrm{y}) = 1\end{aligned} \tag{15}$$

The objective in q is linear with convex constraints, and has a strong dual (see Duchi et al. (2016); Boyd et al. (2004) for the derivation) which is given by:

$$\begin{aligned}&\inf_{\eta\in\mathbb{R}}\left\{\frac{1}{\alpha_0}\mathbb{E}_P(l(h)-\eta)_+ + \eta\right\}\\ &= \inf_{\eta\in\mathbb{R}}\left\{\frac{1}{\alpha_0}\langle(l(h)-\eta)_+,\mathbb{1}\rangle_P + \eta\right\}\\ &= \inf_{\eta\in\mathbb{R}}\left\{\frac{1}{\alpha_0}\langle(l(h)-\eta),\mathbb{1}(l(h)-\eta\geq 0)\rangle_P + \eta\right\} \qquad (16)\\ &= \inf_{\eta\in\mathbb{R}}\sup_{w\in\mathcal{W}}\left\{\frac{1}{\alpha_0}\langle(l(h)-\eta),w\rangle_P + \eta\right\} \qquad (17)\end{aligned}$$

The last equality is true since the set $\mathbb{1}(l(h)-\eta\geq 0)$ is measurable under P (based on our setup in Section 3). 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 $[0,1]$.

$$\mathcal{L}_{\text{cvar}}(h,P) = \inf_{\eta\in\mathbb{R}}\sup_{w\in\mathcal{W}}\left\{\frac{1}{\alpha_0}\langle(l(h)-\eta),w\rangle_P + \eta\right\} \tag{18}$$

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. (2004) to give us the following:

$$\begin{aligned}\mathcal{L}_{\text{cvar}}(h,P) &= \inf_{\eta\in\mathbb{R}}\sup_{w\in\mathcal{W}}\left\{\frac{1}{\alpha_0}\langle(l(h)-\eta),w\rangle_P + \eta\right\}\\ &= \frac{1}{\alpha_0}\sup_{w\in\mathcal{W}}\left\{\inf_{\eta\in\mathbb{R}}\langle(l(h)-\eta),w\rangle_P + \eta\right\}\end{aligned} \tag{19}$$

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

$$\frac{1}{\alpha_0}\inf_{h\in\mathcal{H}}\sup_{w\in\mathcal{W}}\left\{\inf_{\eta\in\mathbb{R}}\langle(l(h)-\eta),w\rangle_P + \eta\right\} \tag{20}$$

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:

$$\frac{1}{\alpha_0}\inf_{h\in\mathcal{H}}\sup_{w\in\mathcal{W}(\gamma)}\left\{\inf_{\eta\in\mathbb{R}}\langle(l(h)-\eta),w\rangle_P + \eta\right\} \tag{21}$$

C.3 PROOF OF THEOREM 5.1

For convenience we shall first restate the Theorem here.

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

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

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 (McAllester, 1998) 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 C.4) since $R(h,\eta,w)\in[0,M/\alpha_0]$.

Lemma C.4 (PAC-Bayes (Catoni, 2007; McAllester, 1998)). *With probability $\geq 1-\delta$ over choice of dataset $\mathcal{D}$ of size n the following inequality is satisfied*

$$\mathbb{E}_P\mathbb{E}_Q(l_{0-1}(h(\mathbf{x}),\mathrm{y}))\leq\mathbb{E}_{\hat{P}_n}\mathbb{E}_Q(l_{0-1}(h(\mathbf{x}),\mathrm{y}))+\sqrt{\frac{D(Q||P)+\log(1/\delta)+\frac{5}{2}\log n+8}{2n-1}} \tag{22}$$

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

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

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:.

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

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

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

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 C.1 on a_i to get:

$$\mathbb{E}_{\mathcal{D}} \exp(\lambda \cdot a_i) \leq \exp \frac{\lambda^2 (M/\alpha_0)^2}{8n} \mathbb{E}_\pi \mathbb{E}_{\mathcal{D}} \exp(\lambda \cdot a_i) \leq \mathbb{E}_\pi \exp \frac{\lambda^2 (M/\alpha_0)^2}{8n} \tag{23}$$

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

$$\mathbb{E}_{\mathcal{D}} \mathbb{E}_\pi [\exp(\lambda \cdot a_i)] \leq \mathbb{E}_\pi \exp \frac{\lambda^2 (M/\alpha_0)^2}{8n} \tag{24}$$

$$= \mathbb{E}_{\mathcal{D}} \exp \sup_{\delta \in \Delta(\mathcal{W},\gamma)} [(\lambda \cdot a_i) - \mathrm{KL}(\delta \,||\, \pi)] \leq \exp \frac{\lambda^2 (M/\alpha_0)^2}{8n} \tag{25}$$

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

$$\mathbb{E}_{\hat{P}_n} \mathbb{E}_Q((h(\mathbf{x}),\mathrm{y})) \lesssim \mathbb{E}_P \mathbb{E}_Q((h(\mathbf{x}),\mathrm{y})) + \frac{M}{\alpha_0} \sqrt{\frac{\mathrm{KL}(\delta \,||\, \pi) + \log(1/\omega) + \log n}{2n-1}} \tag{26}$$

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 C.4 to get for any fixed $\eta \in [0,M], h \in \mathcal{H}$ we get:

$$\left| \sup_{\delta \in \Delta(\mathcal{W},\gamma)} \mathbb{E}_{w \sim \delta} R(h,\eta,w) - \sup_{\delta \in \Delta(\mathcal{W},\gamma)} \mathbb{E}_{w \sim \delta} \hat{R}_D(h,\eta,w) \right| \lesssim \frac{M}{\alpha_0} \sqrt{\frac{\mathrm{KL}(\delta \,||\, \pi) + \log(1/\omega) + \log n}{2n-1}} \tag{27}$$

$$\lesssim \frac{M}{\alpha_0} \sqrt{\frac{\gamma + \log(1/\omega) + \log n}{2n-1}} \tag{28}$$

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 C.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*

$$\log \mathcal{N}(\epsilon, C_1^\alpha(\mathcal{X}), \|\cdot\|_\infty) \leq K \lambda(\mathcal{X}^1) \left(\frac{1}{\epsilon}\right)^{d/\alpha} \tag{29}$$

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:

$$\log\mathcal{N}(\epsilon, \sup_{\delta\in\Delta(\mathcal{W},\gamma)} \mathbb{E}_{w\sim\delta} R(\cdot,\cdot,w), \|\cdot\|_\infty) \lesssim \left(\frac{M}{\alpha_0}\sqrt{\frac{\gamma+\log(1/\omega)+\log n}{2n-1}}\right)^{-(d/\alpha)} \tag{30}$$

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

$$\left|\sup_{\delta\in\Delta(\mathcal{W},\gamma)} \mathbb{E}_{w\sim\delta} R(h,\eta,w) - \sup_{\delta\in\Delta(\mathcal{W},\gamma)} \mathbb{E}_{w\sim\delta} \hat{R}_D(h,\eta,w)\right| \tag{31}$$

$$\lesssim \frac{M}{\alpha_0}\sqrt{\frac{\gamma+\log(\mathcal{N}(\epsilon,R(\cdot,\cdot,w),\|\cdot\|_\infty)/\delta)+\log n}{2n-1}} \tag{32}$$

$$\lesssim \frac{M}{\alpha_0}\sqrt{\frac{\gamma+\left(\left(\frac{M}{\alpha_0}\sqrt{\frac{\gamma+\log(1/\delta)+\log n}{2n-1}}\right)^{-(d/\alpha)}\right)+\log(1/\delta)+\log n}{2n-1}} \tag{33}$$

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 C.6 (Covering number for $\mathcal{N}(\Theta\times[0,M],\frac{\epsilon}{L},\|\cdot\|_2)$ Wainwright (2019))**.** *Let Θ be a bounded convex subset of $\mathbb{R}^d$ with .*

$$\mathcal{N}(\epsilon/L,\Theta,\|\cdot\|) \lesssim \left(1+\frac{L}{\epsilon}\right)^{d+1} \tag{34}$$

We now re-iterate the steps we took previously:

$$\left|\sup_{\delta\in\Delta(\mathcal{W},\gamma)} \mathbb{E}_{w\sim\delta} R(h,\eta,w) - \sup_{\delta\in\Delta(\mathcal{W},\gamma)} \mathbb{E}_{w\sim\delta} \hat{R}_D(h,\eta,w)\right| \tag{35}$$

$$\lesssim \frac{M}{\alpha_0}\sqrt{\frac{\gamma+\log(\mathcal{N}(\epsilon,R(\cdot,\cdot,w),\|\cdot\|_\infty)/\delta)+\log n}{2n-1}} \tag{36}$$

$$\lesssim \frac{M}{\alpha_0}\sqrt{\frac{\gamma+\log\left(1+\frac{L}{\sqrt{\gamma/n}}\right)^{d+1}+\log(1/\delta)+\log n}{2n-1}} \tag{37}$$

$$\lesssim \frac{M}{\alpha_0}\sqrt{\frac{\gamma+(d+1)\log\left(\frac{L^2 n}{\gamma}\right)+\log(1/\delta)+\log n}{2n-1}} \tag{38}$$

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

$$\left| \sup_{\delta \in \Delta(\mathcal{W},\gamma)} \mathbb{E}_{w\sim\delta} R(\hat{h}_D^\gamma,\hat{\eta}_D^\gamma,w) - \sup_{\delta \in \Delta(\mathcal{W},\gamma)} \mathbb{E}_{w\sim\delta} \hat{R}_D R(\hat{h}_D^\gamma,\hat{\eta}_D^\gamma,w) \right| \tag{39}$$

$$\lesssim \frac{M}{\alpha_0} \sqrt{\frac{\gamma + (d+1)\log\left(\frac{L^2 n}{\gamma}\right) + \log(1/\delta) + \log n}{2n-1}} \tag{40}$$

$$\left| \sup_{\delta \in \Delta(\mathcal{W},\gamma)} \mathbb{E}_{w\sim\delta} R(h^*,\eta^*,w) - \sup_{\delta \in \Delta(\mathcal{W},\gamma)} \mathbb{E}_{w\sim\delta} \hat{R}_D R(h^*,\eta^*,w) \right| \tag{41}$$

$$\lesssim \frac{M}{\alpha_0} \sqrt{\frac{\gamma + (d+1)\log\left(\frac{L^2 n}{\gamma}\right) + \log(1/\delta) + \log n}{2n-1}} \tag{42}$$

where h^*,η^* are the optimal for $\mathcal{L}^*_{\text{cvar}}$. Combining the two above proves the statement in Theorem 5.1

C.4 Proof of Theorem 5.2

Setup. Let us assume there exists a prior Π such that $\mathcal{W}(\gamma)$ in Definition 4.1 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:

$$\mu_j \lesssim j^{-2/\gamma} \tag{43}$$

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.

$$\operatorname*{argmax}_{w\in\mathcal{W}(\gamma)),\|w\|_{\mathcal{W}(\gamma)}\leq R} R(h,\eta,w) = \operatorname*{argmax}_{w\in\mathcal{W}(\gamma)),\|w\|_{\mathcal{W}(\gamma)}\leq R} \langle l(h)-\eta, w\rangle_P + \eta \tag{44}$$

$$\operatorname*{argmax}_{w\in\mathcal{W}(\gamma),\|w\|_{\mathcal{W}(\gamma)}\leq R} \mathbb{E}_P \mathbb{1}((l(h)-\eta)\cdot w > 0) \tag{45}$$

We show that we can control: (i) the pessimism of the learned solution; and (ii) the generalization error (Theorem 5.2). Formally, we refer to pessimism for estimates $\hat{h}_D^\gamma,\hat{\eta}_D^\gamma$:

$$\text{excess risk or pessimism: } \sup_{w\in\mathcal{W}(\gamma)} \left| \inf_{h,\eta} R(h,\eta,w) - R(\hat{h}_D^\gamma,\hat{\eta}_D^\gamma,w) \right| \tag{46}$$

Theorem C.7 ((restated for convenience) bounded RKHS)**.** *For $l, \mathcal{H}$ in Theorem 5.1, 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 $\mathcal{O}\big((1/n)\big(\log(1/\delta) + (d+1)\log(nR^{-\gamma}L^{\gamma/2})\big)\big)$ 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 45 is a non-parametric classification problem. We can convert this to the following non-parametric regression problem, after replacing the expectation with plug-in $\hat{P}_n$.

$$\inf_{w\in\mathcal{W}(\gamma)),\|w\|_{\mathcal{W}(\gamma)}\leq R} \frac{1}{n} \sum_{i=1}^{n} (w(x_i,y_i) - (l(h(x_i),y_i)-\eta) + \epsilon_i)^2 + \lambda_n \|w\|^2_{\mathcal{W}(\gamma)} \tag{47}$$

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:

$$\hat{w}_D^{\gamma} = \underset{w \in \mathcal{W}(\gamma)), \|w\|_{\mathcal{W}(\gamma)} \leq R}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (w(x_i, y_i) - (l(h(x_i), y_i) - \eta) + \epsilon_i)^2 + \lambda_n \|w\|^2_{\mathcal{W}(\gamma)} \tag{48}$$

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

$$w_*{}^{\gamma}(x, y) = \underset{w \in \mathcal{W}(\gamma)), \|w\|_{\mathcal{W}(\gamma)} \leq R}{\operatorname{argmin}} \mathbb{E}_P \mathbb{E}_{\epsilon}((l(h(x), y) - \eta) - w(x, y) + \epsilon)^2 \tag{49}$$

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 (Wainwright, 2019) 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 5.1.

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 C.8). Before we use that, we define the quantity δ_n, which is the critical radius (see Ch. 13.4 in Wainwright (2019)). For convenience, we also state it here. Formally, δ_n is the smallest value of δ that satisfies the following inequality (critical condition):

$$\frac{\mathcal{R}_n(\delta)}{\delta} \leq \frac{R}{2} \cdot \delta \tag{50}$$

where,

$$\mathcal{R}_n(\delta) := \mathbb{E}_{\epsilon}\left[\sup_{g \in (\mathcal{F} - f^*), \|g\|_{\mathcal{F}} \leq R, \|g\|_{\hat{P}_n} \leq \delta} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(x_i, y_i) \cdot l(h(x_i) - y_i) \right| \right] \tag{51}$$

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

Lemma C.8 ((Wainwright, 2019)). *For some convex RKHS class $\mathcal{F}$ Let $\hat{f}$ be defined as:*

$$\hat{f} \in \underset{f \in \mathcal{F}, \|f\|_{\mathcal{F}} \leq R}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda_n \|f\|^2_{\mathcal{F}} \right\} \tag{52}$$

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:*

$$\|\hat{f} - f^*\|_2^2 \leq c_0 \inf_{\|\mathcal{F}\| \leq R} \|f - f^*\|_n^2 + c_1 R^2 (\delta_n^2 + \lambda_n). \tag{53}$$

Note that it is standard exercise in statistics to derive the following closed form for the problem in Equation 48:

$$\hat{w}_D^{\gamma}(\cdot) = \hat{K}_n(\cdot, Z)(\hat{K}_n^T \hat{K}_n + \lambda_n I)^{-1}(l(h)_D - \epsilon_D) \tag{54}$$

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 C.9. *Wainwright (2019) Let $\hat{\mu}_j$ be the eigen values $\hat{\mu}_1 \geq \hat{\mu}_2 ... \geq \hat{\mu}_n$ for the empirical Kernel matrix $\hat{K}$, then we have for any δ satisfying*

$$\sqrt{\frac{2}{n}\left(\sum_{i=1}^{n} \min(\delta^2, \hat{u}_j)\right)} \leq \frac{R}{4}\delta^2 \tag{55}$$

, it is necessary that δ satisfies the critical condition in Equation 50.

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 Zhang et al. (2013). Let $\exists C$ for some large $C > 0$ such that $\hat{\mu}_j \leq C j^{-2/\gamma}$. Then for some k, such that $\delta^2 \geq c k^{-2/\gamma}$

$$\sqrt{\frac{1}{n}\left(\sum_{j=1}^{n} \min(\delta^2, \hat{\mu}_j)\right)} \lesssim \sqrt{\frac{2}{n}\left(\sum_{i=1}^{n} \min(\delta^2, C j^{-2/\gamma})\right)} \tag{56}$$

$$\lesssim \sqrt{\frac{2}{n}\left(k\delta^2 + C\sum_{j=k+1}^{n} j^{-2/\gamma}\right)} \lesssim \sqrt{\frac{2}{n}\left(k\delta^2 + C\sum_{j=k+1}^{\infty} j^{-2/\gamma}\right)} \tag{57}$$

$$\lesssim \sqrt{\frac{2}{n}\left(k\delta^2 + C\int_{j=k+1}^{\infty} z^{-2/\gamma}\, dz\right)} \lesssim \sqrt{\frac{2}{n}\left(k\delta^2 + Ck^{-2/\gamma+1}\, dz\right)} \tag{58}$$

$$\leq \sqrt{2/n}\left(\sqrt{k}\cdot\delta\right) \leq \frac{1}{\sqrt{n}} \cdot \delta^{1-\gamma/2} \tag{59}$$

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

$$\frac{1}{\sqrt{n}} \cdot \delta^{1-\gamma/2} \leq \frac{R}{4}\delta^2 \tag{60}$$

$$\implies \delta^{1+\gamma/2} \geq \frac{1}{\sqrt{n}R} \tag{61}$$

This tells us that:

$$\delta_n^2 \gtrsim \left(\frac{1}{nR^2}^{\frac{2}{\gamma+2}}\right) \tag{62}$$

is the critical radius.

We shall later plug this into the bound we have into a uniform bound over the concentration inequality in Lemma C.8. The reason we need a uniform bound over Lemma C.8 is that in its current form, it only bounds $\|\hat{w}_D^{\gamma} - w_*^{\gamma}\|_{\hat{P}_n}^2$ for a specific choice of η, h. In order to arrive at the worst risk generalization error of the form we have in Theorem 5.1 we need to satisfy that with high probability $1-\delta$ $\forall \eta, h$, a critical concentration bound of the form in Lemma C.8 but over $\sup_{\eta,h} \|\hat{w}_D^{\gamma} - w_*^{\gamma}\|_{\hat{P}_n}^2$.

Let $\epsilon = c_2 \exp\left(c_3 n R^2 \frac{\delta_n{}^2}{\sigma^2}\right)$. Since δ_n^2 needs to be large enough (see condition in Equation 62), we use Lemma C.8 in the following bound, incorporating δ_n condition we derived.

With high probability $1-\epsilon$:

$$\|\hat{w}_D^\gamma - w_*^\gamma\|_{\hat{P}_n}^2 \lesssim \inf_{w\in\mathcal{W}(\gamma),\|w\|\leq R} \|w-w_*^\gamma\|_{\hat{P}_n}^2 + R^2 \max\left(\left(\frac{1}{nR^2}\right)^{\frac{\gamma+2}{\gamma}}, \left(\log(1/\epsilon)\frac{1}{nR^2}\right)\right) \tag{63}$$

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 5.1. First, we shall use Lemma C.6 to get a covering number bound for bounded convex subset Θ of $\mathbb{R}^d$ that parameterizes the learner (Theorem 5.2) .

$$\mathcal{N}(\beta/L,\Theta\times[0,M],\|\cdot\|) \lesssim \left(1+\frac{L}{\beta}\right)^{d+1} \tag{64}$$

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 63 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:

$$J = \left(1+\frac{L}{\left(R^2\left(\frac{1}{nR^2}^{\frac{2}{\gamma+2}}\right)\right)}\right)^{d+1} \tag{65}$$

For union bound we need:

$$J\epsilon/c_2 = \exp\left(-c_3 nR^2\delta_n^2\right) \tag{66}$$

$$\implies \log(\frac{1}{\epsilon}) + \log J \gtrsim c_3 nR^2\delta_n^2 \tag{67}$$

$$\implies \log(\frac{1}{\epsilon}) + (d+1)\log\left(\frac{L}{\left(R^2\left(\left(\frac{1}{nR^2}\right)^{\frac{2}{2+\gamma}}\right)\right)}\right) \gtrsim c_3 nR^2\delta_n^2 \tag{68}$$

$$\implies \log(\frac{1}{\epsilon}) + (d+1)\log\left(\left(LR^{-2}\right)^{\frac{\gamma+2}{2}} nR^2\right) \gtrsim c_3 nR^2\delta_n^2 \tag{69}$$

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 $\gamma\geq\gamma_0$, for any sample size.

Thus, we get the following probabilistic uniform convergence. With probability $\geq 1-\epsilon$, $\forall\eta,h$:

$$\|\hat{w}_D^\gamma - w_*^\gamma\|_{\hat{P}_n}^2 \lesssim \inf_{w\in\mathcal{W}(\gamma),\|w\|\leq R} \|w-w_*^\gamma\|_{\hat{P}_n}^2 \tag{70}$$

$$\lesssim \frac{1}{n}\log(\frac{1}{\epsilon}) + (d+1)\log\left(\left(LR^{-2}\right)^{\frac{\gamma+2}{2}} nR^2\right) \tag{71}$$

$$\lesssim \frac{1}{n}\log(\frac{1}{\epsilon}) + (d+1)\log\left(\left(L^{\gamma/2}R^{-\gamma}\right)n\right) \tag{72}$$

Applying the above twice, once one $\hat{w}_D^\gamma$ and another on w_*^γ we prove the generalization bound in Theorem 5.2.

Excess risk bound:

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

$$\text{excess risk} := \sup_{w \in \mathcal{W}(\gamma)} |\inf_{h,\eta} R(h,\eta,w) - R(\hat{h}_D^\gamma, \hat{\eta}_D^\gamma, w)|. \tag{73}$$

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

$$\begin{aligned}
\text{excess risk} &= \sup_{w \in \mathcal{W}(\gamma)} |\inf_{h,\eta} R(h,\eta,w) - R(\hat{h}_D^\gamma, \hat{\eta}_D^\gamma, w)| && (74)\\
&\leq \sup_{w \in \mathcal{W}(\gamma)} \Big(R(h^*(w), \eta^*(w), w) - R(\hat{h}_D^\gamma, \hat{\eta}_D^\gamma, w) \Big) && (75)\\
&\leq \sup_{w \in \mathcal{W}(\gamma)} \left(\frac{1}{\alpha_0} \langle l(h^*) - l(\hat{h}_D^\gamma) - (\eta^*(w) - \hat{\eta}_D^\gamma), w \rangle_P \right) && (76)\\
&\leq \frac{M}{\alpha_0} \sup_{w \in \mathcal{W}(\gamma)} \big((\|w\|_{L_2(P)}) \big) && (77)
\end{aligned}$$

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

$$\|f\|_{L^2(P)} = \|T_K^{1/2} f\|_{\mathcal{H}_k} \leq \sqrt{\|T_K^{1/2}\|_{op}} \cdot \|f\|_{\mathcal{H}_k} \tag{78}$$

Proof:

$$\begin{aligned}
\|T_K^{1/2} f\|_{\mathcal{H}_k}^2 &= \langle T_K^{1/2} f, T_K^{1/2} f \rangle_{\mathcal{H}_k} = \langle f, T_K f \rangle_{\mathcal{H}_k} && (79)\\
&= \sum_{j=1}^{\infty} \frac{\langle \phi_j, f \rangle_{L^2(P)}, \langle \phi_j, T_K f \rangle_{L^2(P)}}{\lambda_j} && (80)\\
&= \|f\|_{L^2(P)}^2 && (81)
\end{aligned}$$

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 C.10 to Equation 77, directly gives us the excess risk bound and completes the proof.

$$\text{excess risk} \lesssim \|T_K^{1/2}\|_{op} \cdot B = \mathcal{O}(B) \tag{82}$$

C.5 Proof of Lemma 5.3

First let us re-state the optimization objective which the two players: learner $h \in \mathcal{H}$ and adversary $w \in \mathcal{W}(\gamma)$ are trying to optimize via an online game (Section 5). Next, we use the fact that $R(h,\eta,w)$ is linear in w.

$$\begin{aligned}
\mathcal{L}^*_{\text{cvar}}(\gamma) &= \inf_{h \in \mathcal{H}, \eta \in \mathbb{R}} \sup_{w \in \mathcal{W}(\gamma)} R(h,\eta,w) = \inf_{h \in \mathcal{H}, \eta \in \mathbb{R}} \sup_{w \in \Delta(\mathcal{W},\gamma)} R(h,\eta,w)\\
&\text{where, } R(h,\eta,w) = (1/\alpha_0)\langle l(h) - \eta, w \rangle_P + \eta
\end{aligned} \tag{83}$$

Hence, the adversary is learning a mixed strategy (probabilistic). Now, recall that $l(\cdot)$ is a strictly convex objective over the space of learners $\mathcal{H}$. Under the negative entropy regularizer, the objective above has a unique saddle point since $\mathcal{H}$ and $\mathcal{W}(\gamma)$ are compact sets (Rockafellar (1970)). This saddle point is also exactly the Nash equilibrium of the two player game.

C.6 Proof of Theorem 5.4

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 Bartlett et al. (1997) Lemma (stated) below to bound the total number of groups that can be identified by $\mathcal{W}(\gamma)$ in n points.

Lemma C.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. (2018). 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:

$$B(\delta) \coloneqq c \cdot \sum_{i=1}^{K} \delta_i \log \delta_i \tag{84}$$

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:

$$\delta_T = \operatorname*{argmin}_{\delta\in\Delta(\mathcal{W}(\gamma))} \sum_{t=1}^{T-1} \frac{1}{\alpha_0} \langle l(h_t) - \eta_t, \delta_t \rangle_{\hat{P}_n} + \eta + c \cdot \sum_{i=1}^{K} \delta_i \log \delta_i \tag{85}$$

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

$$\text{REGRET}_T(\delta) \coloneqq \sum_{t=1}^{T} \frac{1}{\alpha_0} \langle l(h_t) - \eta_t, \delta_t - \delta_{t+1} \rangle_{\hat{P}_n} + B(\delta) - B(\delta_1) \tag{86}$$

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:

$$\sum_{t=1}^{T-1}\frac{1}{\alpha_0}\langle l(h_t)-\eta_t,\delta_t\rangle_{\hat{P}_n}+\eta+c\cdot\sum_{i=1}^{K}\delta_i\log\delta_i+\lambda\cdot(\sum_{i=1}^{K}\delta_i-1) \tag{87}$$

Solving we get:

$$\delta_t(k)=\frac{\exp\left(\frac{-1}{c}\right)\sum_{t=1}^{t-1}\mathbb{E}_{\hat{P}_n}\frac{1}{\alpha_0}(l(h_t)-\eta_t|G_k)+\eta/K}{\sum_{k=1}^{K}\exp\left(\frac{1}{\alpha_0}\frac{-1}{c}\right)\sum_{k=1}^{t-1}(\mathbb{E}_{\hat{P}_n}\frac{1}{\alpha_0}(l(h_t)-\eta_t|G_k)+\eta/K)} \tag{88}$$

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:

$$\frac{1}{\alpha_0}\langle l(h_t)-\eta_t,\delta_t-\delta_{t+1}\rangle_{\hat{P}_n}\leq 1/c \tag{89}$$

The above statement is true because,

$$\delta_{t+1}(i)=\delta_t(i)\cdot\exp\left(\frac{1}{\alpha_0 c}\mathbb{E}[l(h_t)-\eta_t|G_i]+\eta_t/K\right) \tag{90}$$

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

$$\delta_{t+1}(i)\geq\delta_t(i)\cdot e^{-1/c}\geq\delta_t(i)\cdot(1-1/c). \tag{91}$$

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.

$$\text{REGRET}_T\leq(T/c)+(c\log K) \tag{92}$$

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

$$\frac{\text{REGRET}_T}{T}\leq\sqrt{\frac{\log K}{T}} \tag{93}$$

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

$$\frac{\text{REGRET}_T}{T}=\mathcal{O}\sqrt{\frac{\gamma\log n}{T}} \tag{94}$$

Next, we use Theorem 9 from Abernethy et al. (2018) 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:

$$V^*-\epsilon\leq\inf_{h\in\mathcal{H},\eta\in\mathbb{R}}R_D(h,\eta,\bar{\delta}_T)\leq V*\leq\sup_{\delta\in\Delta(\mathcal{W}(\gamma))}R_D(\bar{h}_T,\bar{\eta}_T,\delta)\leq V^*+\epsilon \tag{95}$$

where

$$V^* = R_D(h_D^*(\gamma), \eta_D^*(\gamma), \delta_D^*(\gamma)) = \inf_{h\in\mathcal{H},\eta\in\mathbb{R}} \sup_{\delta\in\Delta(\mathcal{W}(\gamma))} \frac{1}{\alpha_0}\langle l(h)-\eta, \delta\rangle + \eta \tag{96}$$

Excess risk:

For excess risk we need to bound:

$$\frac{1}{\alpha_0} \sup_{h\in\mathcal{H},\eta\in\mathbb{R}} \left| \sup_{\delta\in\Delta(\mathcal{W}(\gamma))} \langle l(h)-\eta, \delta-\delta^*(\gamma)\rangle \right| \tag{97}$$

$$\leq \frac{M}{\alpha_0}\frac{1}{2}\text{TV}(\delta-\delta^*(\gamma)) \leq \frac{M}{2\alpha_0}(1-1/K) = \frac{M}{\alpha_0}\mathcal{O}(1-1/n^\gamma) \tag{98}$$

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 5.4.

C.7 Worst-case generalization risk for Group DRO

Recall that the BR-DRO objective in Equation 4 involves an expectation over P, which in practice is replaced by empirical distribution $\hat{P}_n$. This induces errors in estimating the worst-case risk for plug-in estimates $\hat{h}_D^\gamma, \hat{\eta}_D^\gamma$ In Theorem 5.1 we saw how the bitrate-constraint gracefully controls the worst-case generalization guarantee for estimates $\hat{h}_D^\gamma, \hat{\eta}_D^\gamma$ through an oracle inequality. Here, we ask: "How does the worst-case generalization for Group DRO compare with the bound in Theorem 5.1?".

First, let us recall the objective for Group DRO (Equation 99) which assumes the knowledge of ground-truth groups $G_1, G_2, ..., G_K$.

$$\mathcal{L}^*_{\text{gdro}} := \inf_{h\in\mathcal{H}} \sup_{k\in[K]} \mathbb{E}_P[l(h(\mathbf{x}),\mathrm{y}) \,|\, (\mathbf{x},\mathrm{y})\in G_k] \tag{99}$$

Now, let us denote the plug-in estimate as $\hat{h}_D^K$ which solves the above objective using the empirical distribution $\hat{P}_n$. We are now ready to state Theorem C.12 which gives us the worst-risk generalization error for Group DRO.

Theorem C.12 (worst-case risk generalization (Group DRO))**.** *With probability $\geq 1-\delta$ over $\mathcal{D}\sim P^n$, the worst group risk for $\hat{h}_D^K$ can be upper bounded by the following oracle inequality:*

$$\sup_{k\in[K]} \mathbb{E}_P\left[l(\hat{h}_D^K(\mathbf{x}),\mathrm{y}) \,|\, (\mathbf{x},\mathrm{y})\in G_k\right] \lesssim \mathcal{L}^*_{\text{gdro}} + M\sqrt{\left(\log\left(\frac{2K}{\delta}\right) + (d+1)\log(1+L^2 n)\right)/(2n)},$$

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

Proof Sketch.

We can use the covering number result in Lemma C.5, and plug it into the Hoeffding bound in Lemma C.1 to bound the generalization gap for a specific group G_k. This yields the following result with probability at least $1-\delta$:

$$\left|\mathbb{E}_P[l(h(\mathbf{x}),\mathrm{y}) \,|\, (\mathbf{x},\mathrm{y})\in G_k] - \mathbb{E}_{\hat{P}_n}[l(h(\mathbf{x}),\mathrm{y}) \,|\, (\mathbf{x},\mathrm{y})\in G_k]\right| \tag{100}$$

$$\lesssim \sqrt{\frac{1}{2n}\cdot\left(\log\left(\frac{2}{\delta}\right) + (d+1)\log(1+L^2 n)\right)} \tag{101}$$

Next, we use union bound over the K groups to get the final result in Theorem C.12, with probability at least $1-\delta$.

The main difference between the generalization analysis BR-DRO and Group DRO is the order in which we bound the errors for the learner and the adversary (groups). For Group DRO, we begin with the learner

since we can use traditional uniform convergence bounds to bound per group generalization risk and then use union bound over the K groups to bound the worst-case risk. We can do this precisely since we can use the pre-determined groups G_1,G_2,...,G_K assumed by Group DRO. On the other hand, since we do not assume this knowledge for BR-DRO, we first use PAC-Bayes bound (Lemma C.4) to control the generalization error per group through the bitrate constraint γ. The PAC-Bayes bound allows us to reason about the generalization error for a specific learner $h \in \mathcal{H}$. Finally, to get the bound in Thorem 5.1 we relied on a covering number argument for $\mathcal{H}$ parameterized as a convex subset of $\mathbb{R}^d$ (Section C.3).