

being fast and having low memory cost, scaling to much larger models and datasets.

## 5 EXPERIMENTS

**Algorithms:** We compare our approach, IRM with iterative hard thresholding (IRMv1 + IHT), with relevant baselines ERM, sparse ERM, the oracle, and IRM-based methods. For IRM-based methods, we use IRMv1 (Arjovsky et al., 2020), and we provide Proposition 2 to prove it is an acceptable proxy for the minimax formulation in Equation (12). In order, ERM is the standard training loop on the mixture of all environments; and sparse ERM adds IHT (Jain et al., 2014). The oracle trains ERM with spurious features zeroed, upper bounding accuracies for other methods. For the IRM-based methods, we compare with the original IRMv1 (Arjovsky et al., 2020), and IRMv1 with ProbMask (IRMv1+PM) (Zhou et al., 2022, 2021). When comparing sparsity-based methods, we fix the target density of the feature representation to be same across methods.

**Datasets:** We use common invariant representation learning benchmarks, ColoredMNIST (2-CMNIST) is the original binary dataset introduced in Arjovsky et al. (2020), and FullColoredMNIST (10-CMNIST) (Ahmed et al., 2021) is also generated from MNIST, with two environments, 10 labels and 10 colors. MNISTCIFAR concatenates MNIST digits and CIFAR-10 images (Shah et al., 2020). The oracle baseline is constructed per dataset and only has the designated invariant features: the grayscale MNIST for 2- and 10-CMNIST, and the CIFAR image for MNISTCIFAR. Parameters for the dataset configurations, including label noise and environmental correlation, are in Appendix F.

**Hyperparameter selection:** Because we do not know  $d_{\text{inv}}$  at train time, it is common to treat  $s$  in algorithm 1 as a hyperparameter as in e.g. (Wainwright, 2019). Specifically, we take a uniform grid search per dataset. We find also that accuracy is not affected significantly by small perturbations in  $s$ , which is demonstrated by data from additional experiments on MNISTCIFAR in Table 4.

**Evaluation metrics:** Top-1 test accuracy is compared for the three tasks. For ResNet-18 on MNISTCIFAR, we also provide training time results, and the relative timing in comparison to standard ERM.

**Discussion:** We observe that IRM with IHT can match or exceed the performance of competing methods, including IRM with ProbMask sparsity, for larger models and datasets. Sparse ERM, IRMv1+PM, and IRMv1+IHT were computed with 88% weight density in Table 2; this corresponds to 12% of the weights zeroed out by sparsification methods. The  $L_1$  norms of the layer also reflect the sparsification. ProbMask incurs a noticeable computational overhead – an additional 23% over IRMv1. IHT only adds a 4% cost. We expect time savings to scale up with larger models. Additionally,

we provide results for a MLP with two hidden layers of dimension 390, the median configuration of the model used by (Zhou et al., 2022) on these datasets.

## 6 CONCLUSIONS

In this paper, we provide a non-asymptotic analysis of IRM with sparsity constraints. First, we generalize the data model, relaxing the data model to allow for varying correlation between spurious features and the label. Next, we provide the non-asymptotic results for sparse IRM, including a refinement and correction of previous work in sparse IRM, including theoretical guarantees for  $L_1$ - and  $L_0$ -constrained IRM, resulting in a sparse representation that selects invariant features. Finally, we demonstrate that these methods can be computed in a fast and efficient matter using projected gradient descent-based methods, and we provide experimental results that demonstrate improved test accuracy and time savings on domain generalization datasets.

## 7 ACKNOWLEDGEMENTS AND DISCLOSURE OF FUNDING

The work was supported by the National Science Foundation (NSF) through awards IIS 21-31335, OAC 21-30835, DBI 20-21898, as well as a C3.ai research award.

## References

- Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL [https://proceedings.neurips.cc/paper\\_files/paper/2010/file/7cce53cf90577442771720a370c3c723-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2010/file/7cce53cf90577442771720a370c3c723-Paper.pdf).
- Faruk Ahmed, Yoshua Bengio, Harm van Seijen, and Aaron C. Courville. Systematic generalisation with group invariant predictions. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL <https://openreview.net/forum?id=b9PoimzZFJ>.
- Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 145–155. PMLR, 13–18 Jul

Table 2: Top-1 test accuracy of ResNet-18 with timings on MNISTCIFAR. Our method, IRMv1+IHT, bolded, has negligible overhead time cost and the overall best test accuracy.

Method	Test Accuracy	Train Time (s)	% time/ERM	$L_1$ norm of last layer
Oracle	$77.85 \pm 0.14$	$36.38 \pm 0.26$	99%	$19.72 \pm 3.88$
ERM	$44.93 \pm 0.49$	$36.65 \pm 0.25$	- %	$25.05 \pm 2.32$
Sparse ERM	$44.82 \pm 0.42$	$37.31 \pm 0.37$	102%	$18.31 \pm 2.46$
IRMv1	$52.86 \pm 0.53$	$36.51 \pm 0.17$	100%	$18.65 \pm 0.93$
IRMv1+PM	$57.30 \pm 0.45$	$44.98 \pm 1.02$	123%	$21.59 \pm 1.02$
<b>IRMv1+IHT</b>	<b><math>62.44 \pm 0.96</math></b>	<b><math>38.03 \pm 0.51</math></b>	<b>104%</b>	<b><math>9.10 \pm 1.78</math></b>

Table 3: Top-1 train and test accuracy of MLP390.

10-CMNIST Accuracy (%)		
Method	Train	Test
Oracle	$73.06 \pm 0.21$	$71.36 \pm 0.44$
ERM	$90.00 \pm 0.29$	$28.32 \pm 0.10$
Sparse ERM	$87.17 \pm 1.16$	$29.15 \pm 2.14$
IRMv1	$70.77 \pm 0.27$	$58.88 \pm 0.14$
<b>IRMv1+PM</b>	<b><math>92.20 \pm 0.10</math></b>	<b><math>65.16 \pm 0.09</math></b>
<b>IRMv1+IHT</b>	<b><math>80.83 \pm 0.10</math></b>	<b><math>63.03 \pm 0.51</math></b>

2020. URL <https://proceedings.mlr.press/v119/ahuja20a.html>.

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 3438–3450. Curran Associates, Inc., 2021a.

Kartik Ahuja, Jun Wang, Amit Dhurandhar, Karthikeyan Shanmugam, and Kush R Varshney. Empirical or invariant risk minimization? a sample complexity perspective. *ICLR*, 2021b.

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization, 2022.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on*

*Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252. PMLR, 09–15 Jun 2019. URL <https://proceedings.mlr.press/v97/allen-zhu19a.html>.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020.

Arindam Banerjee, Sheng Chen, Farideh Fazayeli, and Vidyashankar Sivakumar. Estimation with norm regularization, 2015.

O E Barndorff-Neils. *Information and exponential families*. Wiley Series in Probability and Statistics. John Wiley & Sons, Nashville, TN, April 2014.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part XVI*, page 472–489, Berlin, Heidelberg, 2018. Springer-Verlag. ISBN 978-3-030-01269-4. doi: 10.1007/978-3-030-01270-0\_28. URL [https://doi.org/10.1007/978-3-030-01270-0\\_28](https://doi.org/10.1007/978-3-030-01270-0_28).

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL [https://proceedings.neurips.cc/paper\\_files/paper/2006/file/b1b0432ceafb0ce714426e9114852ac7-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2006/file/b1b0432ceafb0ce714426e9114852ac7-Paper.pdf).

Thomas Blumensath and Mike E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009. ISSN 1063-5203. doi: <https://doi.org/10.1016/j.acha.2009.04.002>. URL <https://www.sciencedirect.com/science/article/pii/S1063520309000384>.

- Lawrence D. Brown. Fundamentals of statistical exponential families with applications in statistical decision theory. *Lecture Notes-Monograph Series*, 9:i–279, 1986. ISSN 07492170. URL <http://www.jstor.org/stable/4355554>.
- Elliot Creager, Joern-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2189–2200. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/creager21a.html>.
- Peter K Dunn and Gordon K Smyth. *Generalized linear models with examples in R*. Springer texts in statistics. Springer, New York, NY, 1 edition, November 2018.
- Jianqing Fan, Cong Fang, Yihong Gu, and Tong Zhang. Environment invariant linear least squares. *The Annals of Statistics*, 52(5):2268 – 2292, 2024. doi: 10.1214/24-AOS2435. URL <https://doi.org/10.1214/24-AOS2435>.
- Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Training pruned neural networks. *CoRR*, abs/1803.03635, 2018. URL <http://arxiv.org/abs/1803.03635>.
- Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. URL <http://jmlr.org/papers/v17/15-239.html>.
- Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization, 2020.
- Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL <http://arxiv.org/abs/1510.00149>.
- Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015. URL <https://api.semanticscholar.org/CorpusID:7200347>.
- Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression, 2014.
- Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL [https://proceedings.neurips.cc/paper\\_files/paper/2014/file/218a0aefd1d1a4be65601cc6ddc1520e-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2014/file/218a0aefd1d1a4be65601cc6ddc1520e-Paper.pdf).
- Pritish Kamath, Akilesh Tangella, Danica Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 4069–4077. PMLR, 13–15 Apr 2021. URL <https://proceedings.mlr.press/v130/kamath21a.html>.
- Jivat Neet Kaur, Emre Kiciman, and Amit Sharma. Modeling the data-generating process is necessary for out-of-distribution generalization. In *Workshop on Spurious Correlations, Invariance, and Stability, ICML 2022*, July 2022. URL <https://openreview.net/forum?id=KfB7QnuseT9>. Spotlight.
- David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5815–5826. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/krueger21a.html>.
- Zhao-Rong Lai and Weiwen Wang. Invariant risk minimization is a total variation model. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 25913–25935. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/lai24c.html>.
- Bo Li, Yifei Shen, Yezhen Wang, Wenzhen Zhu, Colorado Reed, Dongsheng Li, Kurt Keutzer, and Han Zhao. Invariant Information Bottleneck for Domain Generalization. 36(7):7399–7407. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v36i7.20703. URL <https://ojs.aaai.org/index.php/AAAI/article/view/20703>.
- Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. *ICLR*, abs/1608.08710, 2017. URL <https://api.semanticscholar.org/CorpusID:14089312>.

- Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL [https://proceedings.neurips.cc/paper\\_files/paper/2018/file/54fe976ba170c19ebae453679b362263-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2018/file/54fe976ba170c19ebae453679b362263-Paper.pdf).
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.
- Yong Lin, Hanze Dong, Hao Wang, and Tong Zhang. Bayesian Invariant Risk Minimization. pages 16021–16030. URL [https://openaccess.thecvf.com/content/CVPR2022/html/Lin\\_Bayesian\\_Invariant\\_Risk\\_Minimization\\_CVPR\\_2022\\_paper.html](https://openaccess.thecvf.com/content/CVPR2022/html/Lin_Bayesian_Invariant_Risk_Minimization_CVPR_2022_paper.html).
- Yong Lin, Hanze Dong, Hao Wang, and Tong Zhang. Bayesian invariant risk minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16021–16030, June 2022.
- Po-Ling Loh and Martin J Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL [https://proceedings.neurips.cc/paper\\_files/paper/2013/file/ef0d3930a7b6c95bd2b32ed45989c61f-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2013/file/ef0d3930a7b6c95bd2b32ed45989c61f-Paper.pdf).
- Christos Louizos, Max Welling, and Diederik P. Kingma. Learning sparse neural networks through l0 regularization. *ArXiv*, abs/1712.01312, 2017. URL <https://arxiv.org/pdf/1712.01312>.
- Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. Variational dropout sparsifies deep neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2498–2507. PMLR, 06–11 Aug 2017. URL <https://proceedings.mlr.press/v70/molchanov17a.html>.
- Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. URL [https://proceedings.neurips.cc/paper\\_files/paper/2009/file/dc58e3a306451c9d670adcd37004f48f-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2009/file/dc58e3a306451c9d670adcd37004f48f-Paper.pdf).
- Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal Inference by using Invariant Prediction: Identification and Confidence Intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 10 2016. ISSN 1369-7412. doi: 10.1111/rssb.12167. URL <https://doi.org/10.1111/rssb.12167>.
- Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.
- Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization, 2020.
- Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9573–9585. Curran Associates, Inc., 2020. URL [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/6cfe0e6127fa25df2a0ef2ae1067d915-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/6cfe0e6127fa25df2a0ef2ae1067d915-Paper.pdf).
- Suraj Srinivas, Akshayvarun Subramanya, and R. Venkatesh Babu. Training sparse neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.
- Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*, pages 443–450, Cham, 2016. Springer International Publishing. ISBN 978-3-319-49409-8.
- Xiaoyu Tan, Lin Yong, Shengyu Zhu, Chao Qu, Xihe Qiu, Xu Yinghui, Peng Cui, and Yuan Qi. Provably invariant learning without domain information. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 33563–33580. PMLR, 23–29 Jul

2023. URL <https://proceedings.mlr.press/v202/tan23b.html>.

V. Vapnik. Principles of risk minimization for learning theory. In J. Moody, S. Hanson, and R.P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1991. URL [https://proceedings.neurips.cc/paper\\_files/paper/1991/file/ff4d5fbbafdf976cfdc032e3bde78de5-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/1991/file/ff4d5fbbafdf976cfdc032e3bde78de5-Paper.pdf).

Roman Vershynin. *Introduction to the Non-Asymptotic Analysis of Random Matrices*. arXiv, 2011. URL <http://arxiv.org/abs/1011.3027>.

Roman Vershynin. *High-Dimensional Probability*. Cambridge University Press, 2018.

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS’18*, page 5339–5349, Red Hook, NY, USA, 2018. Curran Associates Inc.

Martin J Wainwright. *High-dimensional statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, England, February 2019.

Haoxiang Wang, Haozhe Si, Bo Li, and Han Zhao. Provable Domain Generalization via Invariant-Feature Subspace Recovery. arXiv. doi: 10.48550/arXiv.2201.12919. URL <http://arxiv.org/abs/2201.12919>.

Yihua Zhang, Pranay Sharma, Parikshit Ram, Mingyi Hong, Kush Varshney, and Sijia Liu. What is missing in irm training and evaluation? challenges and solutions, 2023.

Xiao Zhou, Weizhong Zhang, Hang Xu, and Tong Zhang. Effective sparsification of neural networks with global sparsity constraint. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3598–3607, 2021. doi: 10.1109/CVPR46437.2021.00360.

Xiao Zhou, Yong Lin, Weizhong Zhang, and Tong Zhang. Sparse Invariant Risk Minimization. In *Proceedings of the 39th International Conference on Machine Learning*, pages 27222–27244. PMLR, 2022. URL <https://proceedings.mlr.press/v162/zhou22e.html>.

## A BACKGROUND AND PRELIMINARIES

### A.1 DEFINITIONS AND SETUP

The noise variables are independent sub-Gaussian random variables (or vectors), with 0 mean and (lower-)bounded variance and bounded sub-Gaussian norm. Finally, we have sub-Gaussian norms  $\kappa_{\text{inv}} = \|\epsilon_{\text{inv}}\|_{\psi_2}$ ,  $\kappa_{s,j} = \|\epsilon_{s,j}\|_{\psi_2}$  for spurious features  $j \in [d_s]$  and  $\kappa_{r,j} = \|\epsilon_{r,j}\|_{\psi_2}$  for random features  $j \in [d_r]$ . For simplicity, we will often work with the largest constants  $\kappa_s = \max_{j \in [d_s]} \kappa_{s,j}$  and  $\kappa_r = \max_{j \in [d_r]} \kappa_{r,j}$ .

We have design matrix denoted  $X^e \in \mathbb{R}^{n \times d}$  where  $X^e = [\mathbf{x}_1^e, \mathbf{x}_2^e, \dots, \mathbf{x}_n^e]^\top$ , and  $\mathbf{y}^e \in \mathbb{R}^n$  where  $\mathbf{y} = [y_1^e, y_2^e, \dots, y_n^e]^\top$ . Each feature in  $\mathbf{x}^e$  is a sub-Gaussian random variable, and we assume  $\|\gamma\|_2 = \|\mathbf{x}_{\text{inv}}\|_2 = 1$ .

For the following analysis, we will use the notation

$$\mathbb{E}[f(x^e)] = \int_{x^e \in \mathcal{E}} f(x^e) d\Pr(f(x^e)), \quad \hat{\mathbb{E}}[f(x^e)] = \frac{1}{n} \sum_{i=1}^{n_e} f(x_i^e). \quad (24)$$

Likewise, we assume that the environmental expectations are defined as follows:

$$\mathbb{E}[f(x^e)] = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \mathbb{E}[f(x^e)], \quad \hat{\mathbb{E}}[f(x^e)] = \sum_{e \in \mathcal{E}_{tr}} \frac{n_e}{N} \hat{\mathbb{E}}[f(x^e)] \quad (25)$$

for the total number of training points available  $N = \sum_{e \in \mathcal{E}_{tr}} n_e$ . Note that the empirical expectation over the environment mixture only has access to  $\mathcal{E}_{tr} \subset \mathcal{E}$ . It is common to assume  $n_i = n_j$  for environments  $i \neq j \in \mathcal{E}_{tr}$ , in which case  $\hat{\mathbb{E}}[f(x^e)] = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}_{tr}} \hat{\mathbb{E}}^e[f(x^e)]$ .

This notation extends to modifiers for environment, footprints that are not the invariant footprint, and empirical risk:

$$\begin{aligned} \beta^e &:= \underset{\substack{\mathbf{v} \in \text{Sp}(S_{\text{inv}}) \\ \|\mathbf{v}\|_2 \leq 1}}{\text{argmin}} \mathcal{R}^e(\mathbf{v}), \\ \beta_S^* &:= \underset{\substack{\mathbf{v} \in \text{Sp}(S) \\ \|\mathbf{v}\|_2 \leq 1}}{\text{argmin}} \sum_{e \in \mathcal{E}} \mathcal{R}^e(\mathbf{v}), \\ \hat{\beta}_S &:= \underset{\substack{\mathbf{v} \in \text{Sp}(S) \\ \|\mathbf{v}\|_2 \leq 1}}{\text{argmin}} \sum_{e \in \mathcal{E}} \hat{\mathcal{R}}^e(\mathbf{v}). \end{aligned}$$

For the purposes of our analysis, we assume the population optima  $\beta^e$ ,  $\beta^*$  and  $\beta_S^e$  are normalized.

Then,  $\hat{\Sigma}^e = \hat{\mathbb{E}}^e[\mathbf{x}^e(\mathbf{x}^e)^\top] = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^e(\mathbf{x}_i^e)^\top = \frac{1}{n} (X^e)^\top (X^e)$ . Also, let  $\lambda_{\max}^e := \lambda_1(\Sigma^e)$  for eigenvalues sorted in descending order, and  $\lambda_{\max} := \max_{e \in \mathcal{E}} (\lambda_1(\Sigma^e))$ .

We also include a proof for the Sub-Gaussian design of feature vectors  $\mathbf{x}^e$ , both for Zhou et al. (2022)'s original generative model (Lemma 3) and for our rescaled model (Lemma 4). Notably, the original setup induces a dependency of  $\sqrt{d_s}$ .

**Proposition 2** (Loss difference substitutes gradient norm penalty). *Assuming the environmental risk is RSC, that is,*

$$\mathcal{R}^e(\mathbf{v}') \geq \mathcal{R}^e(\mathbf{v}) + \langle \mathbf{v}' - \mathbf{v}, \nabla_{\mathbf{v}} \mathcal{R}^e(\mathbf{v}) \rangle + \frac{\alpha}{2} \|\mathbf{v}' - \mathbf{v}\|_2^2, \quad (26)$$

*the gradient norm function  $\mathcal{L}_{\text{IRMv1}}$  is an proxy of the loss difference function  $\mathcal{L}_{\text{IRMmm}}$ .*

*Proof.* We restate the RSC condition for the environmental risk  $\mathcal{R}^e(\mathbf{v})$ . For classifiers  $\mathbf{v} \in \mathbb{R}^d$  and  $\mathbf{v}' \in \mathbb{R}^d$ , we have

$$\tilde{\mathcal{R}}^e(\mathbf{v}') := \mathcal{R}^e(\mathbf{v}) + \langle \mathbf{v}' - \mathbf{v}, \nabla_{\mathbf{v}} \mathcal{R}^e(\mathbf{v}) \rangle + \frac{\alpha}{2} \|\mathbf{v}' - \mathbf{v}\|_2^2, \quad (27)$$

for  $\alpha = 2\|\mathbf{x}^e\|_2^2$ , which has bounded sub-Gaussian norm, as shown in Lemma 4.

By the RSC condition,  $\mathcal{R}^e(\mathbf{v}') \geq \tilde{\mathcal{R}}^e(\mathbf{v}')$  for all  $\mathbf{v}' \in \mathbb{R}^d$ . We find  $\inf_{\mathbf{v}} \tilde{\mathcal{R}}^e(\mathbf{v}')$  at the critical point of our function,

$$0 = \nabla_{\mathbf{v}} \mathcal{R}^e(\tilde{\mathbf{v}}) = \nabla_{\mathbf{v}} \mathcal{R}^e(\mathbf{v}') + \alpha(\tilde{\mathbf{v}} - \mathbf{v}),$$

for a minimizer  $\tilde{\mathbf{v}}$ . The environmental risk then evaluates to

$$\inf_{\mathbf{v}} \tilde{\mathcal{R}}^e(\mathbf{v}) = \mathcal{R}^e(\tilde{\mathbf{v}}) = \mathcal{R}^e(\mathbf{v}) - \frac{1}{2\alpha} \|\nabla \mathcal{R}^e(\mathbf{v})\|_2^2, \quad (28)$$

for a reference classifier  $\mathbf{v} \in \mathbb{R}^d$ .

We then get the inequality

$$\mathcal{R}^e(\mathbf{v}') \geq \tilde{\mathcal{R}}^e(\mathbf{v}') \geq \tilde{\mathcal{R}}^e(\tilde{\mathbf{v}}) = \mathcal{R}^e(\mathbf{v}) - \frac{1}{2\alpha} \|\nabla_{\mathbf{v}} \mathcal{R}^e(\mathbf{v})\|_2^2 \quad (29)$$

From this, we can say

$$\mathcal{R}^e(\mathbf{v}) - \mathcal{R}^e(\mathbf{v}') \leq \frac{1}{2\alpha} \|\nabla_{\mathbf{v}} \mathcal{R}^e(\mathbf{v})\|_2^2. \quad (30)$$

Letting  $\mathbf{v}' = \mathbf{v}^e$ , we can see that minimizing the gradient norm penalty, can approximate minimizing the minimax loss  $\mathcal{L}(\mathbf{v})$ . □

**Lemma 3** (Sub-Gaussian Design for Zhou et al. (2022)). *When  $\zeta_s^e = \mathbf{1}^s$ , as in the original generative model introduced by Zhou et al. (2022), we still have  $\kappa_{ones} := \|(\Sigma^e)^{-1/2}(\mathbf{x}^e)^{(i)}\|_{\psi_2} \leq c_1 \sqrt{d_s}$  for all  $i \in n, e \in \mathcal{E}$ .*

*Proof.* Let  $I_s$  be the indices of the spurious features. The invariant features  $\mathbf{x}_{\text{inv}}^e$ , the label noise  $\epsilon_{\text{inv}}$ , and the random features  $\mathbf{x}_r^e$  are all independent, and identically distributed across samples. Then,  $\mathbf{x}_j^e$  for spurious features  $j \in I_s$ , we have

$$\mathbf{x}_j^e = y + \alpha_i^e \epsilon_j = \gamma^\top \mathbf{x}_{\text{inv}}^e + \epsilon_{\text{inv}} + \alpha_i^e \epsilon_j$$

We keep  $\|\epsilon_j\|_{\psi_2} = 1$  for the spurious features. Let  $\mathbf{a} = [\mathbf{a}_{\text{inv}}, \mathbf{a}_s, \mathbf{a}_r]$  satisfy  $\|\mathbf{a}\|_2^2 = 1$ . Then for any  $t \in \mathbb{R}$ ,

$$\mathbb{E}_{\mathbf{x}^e} [\exp(t\mathbf{a}^\top \mathbf{x}^e)] = \mathbb{E}_{\substack{\mathbf{x}_{\text{inv}} \\ \epsilon_{\text{inv}}}} \left[ \exp(t\mathbf{a}_{\text{inv}}^\top \mathbf{x}_{\text{inv}}^e) \mathbb{E}_{\substack{\mathbf{x}_s \\ \epsilon_{\text{inv}}}} [\exp(t\mathbf{a}_s^\top \mathbf{x}_s^e)] \right] \mathbb{E}_{\mathbf{x}_r^e} [\exp(t\mathbf{a}_r^\top \mathbf{x}_r^e)].$$

The random features are bounded with  $\mathbb{E}_{\mathbf{x}_r} [\exp(t\mathbf{a}_r^\top \mathbf{x}_r^e)] \leq \exp(c_3 t^2 \kappa_r^2)$ . From here, we can condition on  $\mathbf{x}_{\text{inv}}, \epsilon_{\text{inv}}$ , getting

$$\mathbb{E}_{\substack{\mathbf{x}_s \\ \epsilon_{\text{inv}}}} [\exp(t\mathbf{a}_s^\top \mathbf{x}_s^e)] = \mathbb{E}_{\substack{\mathbf{x}_s \\ \epsilon_{\text{inv}}}} \left[ \exp \left( \sum_{j \in I_s} t a_j (y + \alpha_j \epsilon_j) \right) \right], \quad (31)$$

$$\leq \exp(ty \mathbf{1}^\top \mathbf{a}_s) \mathbb{E}_{\substack{\mathbf{x}_s \\ \epsilon_{\text{inv}}}} \left[ \exp \left( t \sum_{j \in I_s} a_j \alpha_j \epsilon_j \right) \right], \quad (32)$$

$$\leq \exp(ty \mathbf{1}^\top \mathbf{a}_s) \exp(c_2 t^2 A), \quad (33)$$

for  $A = \max_{j \in I_s} \alpha_j^2$  and  $c_2 > 0$ . Then,

$$\mathbb{E}_{\substack{\mathbf{x}_{\text{inv}} \\ \epsilon_{\text{inv}}}} [\exp(t\mathbf{a}_{\text{inv}}^\top \mathbf{x}_{\text{inv}}^e) \exp(t(\gamma^\top \mathbf{x}_{\text{inv}}^e + \epsilon_{\text{inv}}) \mathbf{1}^\top \mathbf{a}_s) \exp(c_2 t^2 A)], \quad (34)$$

$$= \mathbb{E}_{\substack{\mathbf{x}_{\text{inv}} \\ \epsilon_{\text{inv}}}} [\exp(t(\mathbf{a}_{\text{inv}} + \mathbf{1}^\top \mathbf{a}_s \gamma)^\top \mathbf{x}_{\text{inv}}^e + t(\epsilon_{\text{inv}} \mathbf{1}^\top \mathbf{a}_s)) \exp(c_2 t^2 A)], \quad (35)$$

$$\leq \exp \left( t^2 c_4 \left( (1 + \sqrt{d_s})^2 \|\mathbf{x}_{\text{inv}}^e\|_{\psi_2}^2 + d_s \|\epsilon_{\text{inv}}\|_{\psi_2}^2 \right) \right) \exp(c_2 t^2 A). \quad (36)$$

The above inequality uses  $\mathbf{1}^\top \mathbf{a}_s \leq \sqrt{d_s}$  and  $\gamma_j \leq 1$ . We then have the bound

$$\mathbb{E}_{\mathbf{x}} [\exp(t\mathbf{a}^\top \mathbf{x}^e)] \leq \exp \left( t^2 c_5 (d_s + 2\sqrt{d_s} + 1 + d_s \|\epsilon_{\text{inv}}\|_{\psi_2}^2 + c_2 A + c_3 \kappa_r^2) \right). \quad (37)$$

Taking the square root of the exponent gets  $\|\mathbf{x}^e\|_{\psi_2} = O(\sqrt{d_s} \|\mathbf{x}_{\text{inv}}^e\|_{\psi_2})$ ; the isotropic vector  $(\Sigma^e)^{-1/2}(\mathbf{x}^e)^{(i)}$  then satisfies  $\kappa_{ones} \leq c_1 \lambda_{\max} \sqrt{d_s}$ .

**Remark 9.** This implies that if we are interested in finding the norm for a subset of the features, i.e.  $\mathbf{m} \odot \mathbf{x}^e$  for  $\mathbf{m} \in \{0, 1\}^d$ , this bound scales with the size of the subset  $\|\mathbf{m}\|_1$ . This is pertinent for when we select a smaller (usually  $O(d_{\text{inv}})$ ) subset of features with a sparse predictor under  $L_0$  constraints.

□

**Lemma 4** (Feature vector L2 bound). *We have with probability  $1 - \delta$  the bound*

$$\|\mathbf{x}^e\|_2 \leq 1 + c_s + c_a \kappa_s + c_r \kappa_r + O\left((c_a^2 \kappa_s^2 + c_r^2 \kappa_r^2) \sqrt{\log \frac{1}{\delta}}\right), \quad (38)$$

for positive constants  $c_s, c_a, c_r$  as defined in model generation. The norm itself is a sub-Gaussian RV with  $\|\mathbf{x}^e\|_{\psi_2} = \kappa_{\mathbf{x}} = O(\max\{\kappa_s^2, \kappa_r^2\})$  and mean  $\mathbb{E}[\|\mathbf{x}^e\|_2] = c_s + c_a \kappa_s + c_r \kappa_r$ .

*Proof.* We first apply triangle inequality on the three feature blocks.

$$\|\mathbf{x}^e\|_2 \leq \|\mathbf{x}_{\text{inv}}^e\|_2 + \|\mathbf{x}_s^e\|_2 + \|\mathbf{x}_r^e\|_2$$

We bound the three terms in order. First, we use the assumption that  $\|\mathbf{x}_{\text{inv}}^e\|_2 = 1$ .

Then, to evaluate  $\|\mathbf{x}_s^e\|_2$ , we again use the triangle inequality to separate the label component from the sub-Gaussian noise, getting  $\|\mathbf{x}_s^e\|_2 \leq \|y^e \zeta\|_2 + \|\alpha^e \odot \epsilon_s\|_2$ . With Cauchy-Schwartz, we have  $\|y^e \zeta\|_2 \leq \|\gamma\|_2 \|\mathbf{x}_{\text{inv}}\|_2 \|\zeta_s^e\|_2 = c_s$ .

To bound the second noise component, we apply a variant of Theorem 3.1.1 from (Vershynin, 2018) for zero-mean sub-Gaussian variables with different sub-Gaussian norms on different features.

We define a random variable  $Z = [Z_1, Z_2, \dots, Z_{d_s}]$  with  $Z_i = |\alpha_i^e \epsilon_{s,i}|$ , and we aim to bound  $\|Z\|_2$ . Firstly,  $\mathbb{E}[\|Z\|_2^2] = c_a^2 \kappa_s^2$  and  $\mathbb{E}[Z_i^2] = (\alpha_i^e)^2 \kappa_s^2$ . We know that  $Z_i$  is sub-Gaussian with  $\|Y_i\|_{\psi_2} = \alpha_i^e \kappa_s$ , so it must be that  $Y_i = Z_i^2 - (\alpha_i^e)^2 \kappa_s^2$  is sub-exponential and zero-mean.

Then, let  $K = \max(\|Y_i\|_{\psi_1}) \leq \max(c_1 \|Z_i\|_{\psi_2}^2) \leq c_1 \max_i ((\alpha_i^e)^2) \kappa_s^2 = c_1 c_a^2 \kappa_s^2$  for an absolute constant  $c_1 > 0$ . Note that variables named  $c_1, c_2$ , etc. will also be positive constants going forward.

We apply Bernstein's to get

$$\Pr\left(\left|\sum_{i=1}^{d_s} Z_i^2 - c_a^2 \kappa_s^2\right| \geq u\right) \leq 2 \exp\left(-\frac{c_2 d_s}{K_0} \min(u^2, u)\right), \quad (39)$$

where  $K_0 = \max(K^2, K)$ . Then, using the fact that for non-negative  $z, a$  we have  $|z - a| \geq \delta$  implies  $|z^2 - a^2| \geq \max(\delta, \delta^2)$ . If we let  $u = \max(\delta, \delta^2)$ , we have  $\delta^2 = \min(u^2, u)$ , and

$$\Pr\left(\left|\frac{1}{\sqrt{d_s}} (\|Z\|_2 - c_a \kappa_s)\right| \geq \delta\right) \leq \Pr\left(\left|\frac{1}{d_s} (\|Z\|_2^2 - c_a^2 \kappa_s^2)\right| \geq \max(\delta, \delta^2)\right) \quad (40)$$

$$\leq 2 \exp\left(-\frac{c_2 d_s}{K_0^2} \cdot \delta^2\right). \quad (41)$$

This then gets the bound, with  $t = \delta \sqrt{d_s}$ ,

$$\Pr(\|Z\|_2 - c_a \kappa_s \geq t) \leq 2 \exp\left(-\frac{c_2 t^2}{K_0^2}\right). \quad (42)$$

Then, we have with probability  $1 - \delta_s$  that  $\|\alpha^e \odot \epsilon_s\|_2 \leq c_a \kappa_s + \sqrt{\frac{K_0^2}{c_2} \log \frac{1}{\delta_s}}$ . Together,

$$\|\mathbf{x}_s^e\|_2 \leq c_s + c_a \kappa_s + \sqrt{\frac{K_0^2}{c_2} \log \frac{1}{\delta_s}}.$$

Then, a similar argument gets  $\|\mathbf{x}_r^e\|_2 \leq c_r \kappa_r + \sqrt{\frac{K_1^2}{c_3} \log \frac{1}{\delta_r}}$ . Here, the constant is  $K_1 \leq \max\{K', K'^2\}$  for  $K' = c_3 c_r^2 \kappa_r^2$ . For the final answer, we will assume  $K, K' \geq 1$ . Then, letting  $\delta = \delta_r + \delta_s$ , we get the final bound: with probability  $1 - \delta$ , we have

$$\begin{aligned} \|\mathbf{x}^e\|_2 &\leq 1 + c_s + c_a \kappa_s + c_r \kappa_r + \sqrt{\frac{c_a^4 \kappa_s^4}{c_2} \log \frac{1}{\delta}} + \sqrt{\frac{c_r^4 \kappa_r^4}{c_3} \log \frac{1}{\delta}}, \\ &= O\left(c_s + c_a \kappa_s + c_r \kappa_r + (c_a^2 \kappa_s^2 + c_r^2 \kappa_r^2) \sqrt{\log \frac{1}{\delta}}\right). \end{aligned} \quad (43)$$

This demonstrates that  $\|\mathbf{x}^e\|_2$  is a sub-Gaussian random variable; its sub-Gaussian norm is  $\kappa_{\mathbf{x}} = O(\max\{\kappa_s^2, \kappa_r^2\})$ . Also from the above analysis, we can see that the population mean of the norm is

$$\mathbb{E}[\|\mathbf{x}^e\|_2] = 1 + c_s + c_a \kappa_s + c_r \kappa_r. \quad (44)$$

□

**Lemma 5** (Empirical gap). *Let  $\beta_S^e = \operatorname{argmin}_{\mathbf{v} \in \operatorname{Sp}(S)} \mathbb{E}^e[(y - \mathbf{v}^\top \mathbf{x})^2]$  and  $\hat{\beta}_S^e = \operatorname{argmin}_{\mathbf{v} \in \operatorname{Sp}(S)} \hat{\mathbb{E}}^e[(y - \mathbf{v}^\top \mathbf{x})^2]$  respectively be population and empirical minimizers, with covariance matrix  $\hat{\Sigma}^e = \hat{\mathbb{E}}^e[\mathbf{x}\mathbf{x}^\top]$  and footprint size  $|S| \leq d_{\text{inv}}$ . Then we have with probability  $1 - \delta$ ,*

$$\hat{\mathbb{E}}^e[(y - \beta_S^e{}^\top \mathbf{x})^2] - \hat{\mathbb{E}}^e[(y - \hat{\beta}_S^e{}^\top \mathbf{x})^2] = \|\beta_S^e - \hat{\beta}_S^e\|_{\hat{\Sigma}^e} + \operatorname{err}(1/\delta, n), \quad (45)$$

where  $\operatorname{err}(\frac{1}{\delta}, n)$  depends on assumptions on the generation model, specified in Proposition 7.

*Proof.* We note that  $\mathbf{v}_S^\top \mathbf{x} = \mathbf{v}^\top \Phi(\mathbf{x})$  if  $\Phi$  selects the same features as  $S$ , where  $\Phi(\mathbf{x})$  masks  $\mathbf{x}$  and  $\mathbf{v}_S$  masks  $\mathbf{v}$ . We continue with the set notation  $\beta_S$  for this proof, but the proof applies when working with feature mask  $\Phi$ .

First, we define a noise variable  $\omega_S^e$ :

$$\omega_S^e = y - (\beta_S^e)^\top \mathbf{x}^e = (\beta^* - \beta_S^e)^\top \mathbf{x}^e + \epsilon_{\text{inv}}. \quad (46)$$

We proceed with the algebraic proof.

$$\hat{\mathbb{E}}^e \left[ (y - \mathbf{x}^{e\top} \hat{\beta}_S^e)^2 - (y - \mathbf{x}^{e\top} \beta_S^e)^2 \right] \quad (47)$$

$$= \hat{\mathbb{E}}^e \left[ (\mathbf{x}^{e\top} \hat{\beta}_S^e)^2 \right] - 2\hat{\mathbb{E}}^e \left[ y \mathbf{x}^{e\top} (\hat{\beta}_S^e - \beta_S^e) \right] - \hat{\mathbb{E}}^e \left[ (\mathbf{x}^{e\top} \beta_S^e)^2 \right] \quad (48)$$

$$= \hat{\mathbb{E}}^e \left[ (\mathbf{x}^{e\top} \hat{\beta}_S^e)^2 \right] - 2\hat{\mathbb{E}}^e \left[ (\beta_S^e{}^\top \mathbf{x}^e + \omega_S^e) \mathbf{x}^{e\top} (\hat{\beta}_S^e - \beta_S^e) \right] - \hat{\mathbb{E}}^e \left[ (\mathbf{x}^{e\top} \beta_S^e)^2 \right] \quad (49)$$

$$= \hat{\mathbb{E}}^e \left[ (\mathbf{x}^{e\top} \hat{\beta}_S^e)^2 \right] - 2\hat{\mathbb{E}}^e \left[ (\mathbf{x}^{e\top} \beta_S^e) \mathbf{x}^{e\top} (\hat{\beta}_S^e - \beta_S^e) \right] - \hat{\mathbb{E}}^e \left[ (\mathbf{x}^{e\top} \beta_S^e)^2 \right] + \operatorname{err}(1/\delta, n) \quad (50)$$

$$= \hat{\mathbb{E}}^e \left[ (\mathbf{x}^{e\top} \hat{\beta}_S^e)^2 \right] - 2\hat{\mathbb{E}}^e \left[ \mathbf{x}^{e\top} \hat{\beta}_S^e \mathbf{x}^{e\top} \beta_S^e \right] + \hat{\mathbb{E}}^e \left[ (\mathbf{x}^{e\top} \beta_S^e)^2 \right] + \operatorname{err}(1/\delta, n)$$

$$= \hat{\mathbb{E}}^e \left[ (\mathbf{x}^{e\top} \hat{\beta}_S^e) - (\mathbf{x}^{e\top} \beta_S^e)^2 \right] + \operatorname{err}(1/\delta, n)$$

$$= (\hat{\beta}_S^e - \beta_S^e)^\top \hat{\mathbb{E}}^e \left[ \mathbf{x}^e \mathbf{x}^{e\top} \right] (\hat{\beta}_S^e - \beta_S^e) + \operatorname{err}(1/\delta, n)$$

$$= \left\| \hat{\beta}_S^e - \beta_S^e \right\|_{\hat{\Sigma}^e}^2 + \operatorname{err}(1/\delta, n). \quad (51)$$

We note that step Equation (48) to Equation (50) is where the  $\operatorname{err}(\frac{1}{\delta}, n)$  term is introduced. The proof of Proposition 5 in (Hsu et al., 2014) applies to population error, making use of  $\mathbb{E}[y] = \mathbb{E}[\beta^*{}^\top \mathbf{x}]$ , which does not hold for either empirical risk, nor for classifiers  $\beta^e$  or  $\beta_S^e$  that are not the (invariant optimal) ground truth. □

**Corollary 6** (Empirical gap dominates). *Following definitions introduced in Lemma 5, we have with probability  $1 - \delta$ ,*

$$\|\beta_S^e - \hat{\beta}_S^e\|_{\hat{\Sigma}} \leq |\operatorname{err}(1/\delta, n)|. \quad (52)$$

*Proof.* Recalling that  $\hat{\beta}_S^e = \operatorname{argmin}_{\mathbf{v} \in \operatorname{Sp}(S)} \hat{\mathbb{E}}^e[(y - \mathbf{v}^\top \mathbf{x})^2]$ , we know that

$$\hat{\mathbb{E}}^e \left[ (y - \mathbf{x}^{e\top} \hat{\beta}_S^e)^2 - (y - \mathbf{x}^{e\top} \beta_S^e)^2 \right] \leq 0.$$

□

**Proposition 7** (Empirical gap general). *Given environment  $e \in \mathcal{E}$  and selected features  $S \in 2^d$ , and with probability  $1 - \delta$ ,*

$$\operatorname{err}(1/\delta, n) := O \left( d_{\text{inv}} c_{\text{total}} + \frac{K d_{\text{inv}}}{\sqrt{n}} \log \frac{1}{\delta} \right) \quad (53)$$

for  $c_{\text{total}} = \max\{1 + \kappa_{\text{inv}}^2 + (c_z^2 + \kappa_{\text{inv}}^2 + c_a^2) + c_r^2\}$  and  $K^2 = \max(\kappa_{\mathbf{x}}^4, \kappa_{\text{inv}}^2)$ .

*Proof.* We have the definition

$$\text{err}(\frac{1}{\delta}, n) := -2\hat{\mathbb{E}}^e[\omega_S^e \mathbf{x}^{e\top} (\hat{\beta}_S^e - \beta_S^e)]. \quad (54)$$

We will now upper bound the second term with probability  $1 - \delta$ .

$$\text{err}(\frac{1}{\delta}, n) \leq 2 \left| \hat{\mathbb{E}}^e \left[ \omega_S^e (\mathbf{x}^e)^\top (\hat{\beta}_S^e - \beta_S^e) \right] \right| \quad (55)$$

$$\leq \left| (\beta^* - \beta_S^e)^\top \hat{\mathbb{E}}^e[\mathbf{x}^e (\mathbf{x}^e)^\top] (\hat{\beta}_S^e - \beta_S^e) \right| + \left| \hat{\mathbb{E}}^e[\epsilon_{\text{inv}} \mathbf{x}^e]^\top (\hat{\beta}_S^e - \beta_S^e) \right| \quad (56)$$

$$\leq 2d_{\text{inv}} \left\| \frac{1}{n} (X^e)^\top X^e \right\|_2 + \sqrt{2d_{\text{inv}}} \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_{\text{inv},i} \mathbf{x}_i^e \right\|_2. \quad (57)$$

With constants  $c_1, c_2 > 0$  and  $t_1, t_2 > 0$  we can use Bernstein's inequality to get, with probability  $1 - 2 \exp\left(-c_1 \min\left\{\frac{t_1^2}{\kappa_x^4}, \frac{t_2}{\kappa_x^2}\right\}n\right)$ ,

$$\left\| \frac{1}{n} (X^e)^\top X^e \right\|_2 \leq \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i^e\|_2^2 \leq \mathbb{E}^e[\|\mathbf{x}^e\|_2^2] + t_1. \quad (58)$$

Similarly, with the Hoeffding-type inequality,  $1 - e \cdot \exp\left(-\frac{c_2 t_2^2 n}{K^2}\right)$ ,

$$\left\| \frac{1}{n} \sum_{i=1}^n \epsilon_{\text{inv},i} \mathbf{x}_i^e \right\|_2 \leq \frac{1}{n} \sum_{i=1}^n \|\epsilon_{\text{inv},i} \mathbf{x}_i^e\|_2 \leq \mathbb{E}^e[\|\epsilon_{\text{inv}} \mathbf{x}^e\|_2] + t_2. \quad (59)$$

Combining, we get

$$\text{err}(\frac{1}{\delta}, n) \leq 2d_{\text{inv}}(\mathbb{E}^e[\|\mathbf{x}^e\|_2^2] + t_1) + \sqrt{2d_{\text{inv}}}(\mathbb{E}^e[\|\epsilon_{\text{inv}} \mathbf{x}^e\|_2] + t_2). \quad (60)$$

We let both share the same bound  $t_1 = t_2 = t$ . Then, with constants  $C, c > 0$  and  $K^2 = \max(\kappa_x^4, \kappa_{\text{inv}}^2)$ , we apply union bound to a bound with probability  $1 - c \exp\left(-C \min\left\{\frac{t^2}{K^2}, \frac{t}{\kappa_x^2}\right\}n\right)$ . Due to the mixture of tails resulting from Bernstein's inequality of a sum of sub-exponential variables, we will upper bound the maximum of the two. In other words,  $t \leq \sqrt{\frac{K^2}{Cn}} \log \frac{c}{\delta}$ , and

$$\text{err}(\frac{1}{\delta}, n) \leq 2d_{\text{inv}} \mathbb{E}^e[\|\mathbf{x}^e\|_2^2] + \sqrt{2d_{\text{inv}}} \mathbb{E}^e[\|\epsilon_{\text{inv}} \mathbf{x}^e\|_2] + 2d_{\text{inv}} \sqrt{\frac{K^2}{Cn}} \log \frac{c}{\delta}. \quad (61)$$

We substitute  $\mathbb{E}^e[\|\mathbf{x}^e\|_2^2] \leq 1 + \kappa_{\text{inv}}^2 + (c_z^2 + \kappa_{\text{inv}}^2 + c_A^2 \kappa_A^2) + c_r^2 \kappa_r^2$  and  $\mathbb{E}^e[\|\epsilon_{\text{inv}} \mathbf{x}^e\|_2] \leq \kappa_{\text{inv}}(1 + c_s + c_A \kappa_s + c_r \kappa_r)$  into the above to get the desired result.  $\square$

**Corollary 8** (Empirical gap with ones). *In th original setting in Zhou et al. (2022), where the scaling variables are  $\zeta_s^e = \mathbf{1}^{d_s}$   $\zeta_s^e = \mathbf{1}^{d_s}$ , we have*

$$\text{err}(\frac{1}{\delta}, n) \leq c_{\text{ones}} d_{\text{inv}} + O\left(d_{\text{inv}} \sqrt{\frac{K^2}{n}} \log \frac{1}{\delta}\right). \quad (62)$$

for  $c_{\text{ones}} = d_{\text{inv}}(1 + \kappa_{\text{inv}}^2 + c_A^2 \kappa_s^2)^2$  and  $K^2 = O(d_{\text{inv}})$ .

*Proof.* We modify the bound for the error term introduced in Proposition 7 with  $\zeta_s^e = \mathbf{1}^{d_s}$  and  $\zeta_r = \mathbf{1}^{d_r}$ . That is,  $\mathbf{x}_s = y \mathbf{1}^{d_s} + \alpha^e \odot \epsilon_s$  and  $\mathbf{x}_r = \epsilon_r$ .

**Remark 10.** The spurious features are generated with constant contribution from the label per feature. That is,

$$\mathbb{E}^e[x_{s,i}] = \mathbb{E}^e[y] = \mathbb{E}^e[\gamma^\top \mathbf{x}_{\text{inv}}] = O(1),$$

In this case,  $\mathbb{E}^e[\|\mathbf{x}^e\|_2] \geq c\sqrt{d_s}$  for constant  $c > 0$ . Because this will introduce an undesirable dependency on  $d_s$ , we avoid evaluating  $\|\mathbf{x}^e\|_2$  explicitly.

Again, we want to bound the expression from Equation (56).

$$\text{err}(\frac{1}{\delta}, n) \leq 2 \left| \hat{\mathbb{E}}^e \left[ \omega_S^e(\mathbf{x}^e)^\top (\hat{\beta}_S^e - \beta_S^e) \right] \right| \quad (63)$$

$$\leq \left| \hat{\mathbb{E}}^e [(\beta^* - \beta_S^e)^\top \mathbf{x}^e (\mathbf{x}^e)^\top (\hat{\beta}_S^e - \beta_S^e)] \right| + \left| \hat{\mathbb{E}}^e [\epsilon_{\text{inv}} \mathbf{x}^e]^\top (\hat{\beta}_S^e - \beta_S^e) \right| \quad (64)$$

$$\leq \frac{1}{n} \sum_{i=1}^n A_i B_i + \sqrt{2d_{\text{inv}}} \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_{\text{inv},i} \mathbf{x}_i^e \right\|_2. \quad (65)$$

Above, we apply Cauchy-Schwartz to get  $A_i = \|(\beta^* - \beta_S^e)^\top \mathbf{x}^e\|_2$  and  $B_i = \|(\hat{\beta}_S^e - \beta_S^e)^\top \mathbf{x}^e\|_2$ .  $A_i$  is a sub-Gaussian random variable with mean at least  $\sqrt{d_{\text{inv}}}$ . Likewise for  $B_i$ . Furthermore, let  $K = \max\{\|\mathbf{A}_i\|_{\psi_2}, \|\mathbf{B}_i\|_{\psi_2}\}$ , where the sub-Gaussian norm of both  $A_i$  and  $B_i$  is  $O(\sqrt{d_{\text{inv}}})$  (see final remark in Lemma 3), given that we only sum up to  $2d_{\text{inv}}$  elements of the total features of  $\mathbf{x}^e$ .

Then with probability  $1 - 2 \exp\left(-c \min\{\frac{t_1^2}{K^4}, \frac{t_1}{K^2}\}n\right)$ , constants  $C, c > 0$ , and  $t > 0$  we can say

$$\frac{1}{n} \sum_{i=1}^n A_i B_i \leq cd_{\text{inv}}(1 + \kappa_{\text{inv}}^2 + c_A^2 \kappa_s^2) + t_1. \quad (66)$$

The bound

$$\text{err}(\frac{1}{\delta}, n) \leq cd_{\text{inv}}(1 + \kappa_{\text{inv}}^2 + c_A^2 \kappa_s^2)^2 + \sqrt{2d_{\text{inv}}} \mathbb{E}^e[\|\epsilon_{\text{inv}} \mathbf{x}_i^e\|_2] + 2d_{\text{inv}} \sqrt{\frac{K^2}{Cn}} \log \frac{c}{\delta}.$$

□

**Corollary 9** (Missing empirical term with uniform feature assumption). *When the norms of  $\mathbf{x}_s, \mathbf{x}_r$  are uniformly distributed, i.e.  $\zeta_s^e = \frac{1}{d_s} \cdot \mathbf{1}^{d_s}$  and  $\zeta_r = \frac{1}{d_r} \cdot \mathbf{1}^{d_r}$ , we have with probability  $1 - \delta$ ,*

$$\text{err}(\frac{1}{\delta}, n) \leq O((c_s^2 + c_a^2 + c_r^2)^2 \cdot \frac{d_{\text{inv}}}{\min\{d_s, d_r\}}) + \sqrt{2d_{\text{inv}}} \mathbb{E}^e[\|\epsilon_{\text{inv}} \mathbf{x}_i^e\|_2] + 2d_{\text{inv}} \sqrt{\frac{\kappa_{\text{inv}}^2}{Cn}} \log \frac{c}{\delta}. \quad (67)$$

*Proof.* We can also infer that for a constant  $b > 0$ ,  $\max_{j \in S_{\text{inv}}} x_j^2 \geq \frac{b}{d_{\text{inv}}}$  because  $\|\mathbf{x}_{\text{inv}}\|_2^2 = 1$ .

We again bound Equation (56) by considering  $A_i = \|(\beta^* - \beta_S^e)^\top \mathbf{x}^e\|_2$  and  $B_i = \|(\hat{\beta}_S^e - \beta_S^e)^\top \mathbf{x}^e\|_2$ . With the uniformity assumptions, we can then say that for vectors  $\mathbf{v} \in \text{Sp}(S_1), \mathbf{u} \in \text{Sp}(S_2)$ , for  $|S_1|, |S_2| \leq d_{\text{inv}}$ ,

$$(\mathbf{v} - \mathbf{u})^\top \mathbf{x}^e \leq 2 + \sum_{j \in S_s} (u_j - v_j) x_j^e + \sum_{j \in S_r} (u_j - v_j) x_j^e \leq 2 + \sqrt{d_{\text{inv}}} \left( \frac{c_s^2 + c_a^2}{\sqrt{d_s}} + \frac{c_r^2}{\sqrt{d_r}} \right) \quad (68)$$

We get  $A_i B_i \leq O((c_s^2 + c_a^2 + c_r^2)^2 \cdot \frac{d_{\text{inv}}}{\min\{d_s, d_r\}})$ . Then with probability  $1 - 2 \exp\left(-c \min\{\frac{t_1^2}{K^4}, \frac{t_1}{K^2}\}n\right)$ , constants  $C, c > 0$ , and  $t > 0$  we can say

$$\text{err}(\frac{1}{\delta}, n) \leq O((c_s^2 + c_a^2 + c_r^2)^2 \cdot \frac{d_{\text{inv}}}{\min\{d_s, d_r\}}) + \sqrt{2d_{\text{inv}}} \mathbb{E}^e[\|\epsilon_{\text{inv}} \mathbf{x}_i^e\|_2] + 2d_{\text{inv}} \sqrt{\frac{\kappa_{\text{inv}}^2}{Cn}} \log \frac{c}{\delta}. \quad (69)$$

**Remark 11.** This result is significantly tighter due to the  $\frac{d_{\text{inv}}^2}{\min\{d_s, d_r\}} \leq d_{\text{inv}}$  factor in the first term, generated from the mean of  $\omega_S^e(\mathbf{x}^e)^\top (\hat{\beta}_S^e - \beta_S^e)$ . Because in the overparameterized case  $d_{\text{inv}} \ll d_s + d_r$ , we expect that the predictors  $(\hat{\beta}_S^e - \beta_S^e)$  and  $(\hat{\beta}^* - \beta_S^e)$ , are not likely to pick up the majority of the length of  $\mathbf{x}^e$  with only  $2d_{\text{inv}}$  features. This is the best-case scenario, in which no “heavy hitters” contributing to  $c_s^2$  or  $c_r^2$  are captured by the linear predictors.

□

## A.2 ASSUMPTIONS

**Assumption 1.** The noise on the invariant features is subgaussian. Specifically, there exists  $C > 0$  such that

$$\mathbb{E} [\exp(t\epsilon_{\text{inv}}) \leq \exp(t^2 C^2)] \quad \forall t \in \mathbb{R}.$$

**Assumption 2.** There exist positive constants  $\sigma_0$  and  $\bar{\gamma}$  such that

$$\forall i \in [d_{\text{inv}}], |\gamma_i| \geq \bar{\gamma}, \quad \text{Var}[x_{\text{inv},i}] \geq \sigma_0^2.$$

This states that each invariant feature  $x_{\text{inv},i}$  and its corresponding ground truth weight  $\gamma_i$  must sufficiently contribute to the explanation of the label.

**Assumption 3.** For the  $i$ th spurious feature, let  $\alpha_i^2 = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} (\alpha_i^e)^2$  for spurious features  $i \in [d_s]$ . There exists a constant  $\Delta > 0$  such that the following holds for each spurious feature  $i \in [d_s]$  for all environments  $e \in \mathcal{E}$ :

$$|\alpha_i^2 - (\alpha_i^e)^2| \geq \Delta.$$

**Assumption 4.** Let  $\{\lambda_i^e\}_{i=1}^d$  and  $\{\lambda_i\}_{i=1}^d$  be the eigenvalues of  $\Sigma^e$  and  $\Sigma$  respectively with corresponding eigenvectors  $\{\mathbf{u}_i^e\}_{i=1}^d$  and  $\{\mathbf{u}_i\}_{i=1}^d$ .

Then for those  $\mathbf{u}_i$  such that  $\lambda_i^e - \lambda_i = \alpha_i^2 - (\alpha_i^e)^2 > 0$ , we have a constant  $D$  such that

$$|\mathbb{E}[\mathbf{x}^e \mathbf{y}^e]^\top \mathbf{u}_i| \geq D > 0 \quad (70)$$

**Assumption 5.** The invariant features may not be rank deficient. That is,

$$\min_i \lambda_i \geq c_{\min} > 0 \quad (71)$$

for  $\lambda_i$  (defined in Assumption 4) as an eigenvalue of  $\hat{\Sigma}$ .

**Assumption 6.** The loss function is RSC. Specifically, a function  $\mathcal{L}$  satisfies  $\alpha$ -restricted strong convexity ( $\alpha$ -RSC)

## B PROOF OF THEOREMS 1 AND 2 (INFORMATION-THEORETIC)

This appendix includes the proof and supporting analysis for Theorem 1. We restate it below for clarity.

**Theorem 1** Assume at least  $n$  samples per environment  $e \in \mathcal{E}$ , for a total of  $N = |\mathcal{E}|n$  across the whole training set. If  $n > O\left(\log\left(\frac{d \cdot |\mathcal{E}|}{\delta}\right)\right)$ , together with assumptions in Appendix A.2, with probability at least  $(1 - \delta)$ , we have:

$$\hat{\mathcal{L}}(S_{\text{inv}}) < \hat{\mathcal{L}}(S), \quad \forall |S| \leq d_{\text{inv}}, \quad S \neq S_{\text{inv}}, \quad (72)$$

with constants in sample complexity further specified below.

*Proof.* To sketch the proof, we first analyze  $\hat{\mathcal{L}}(\hat{\mathbf{v}}_S)$  where  $\mathbf{v}_S \in \text{Sp}(S)$  for  $S \in 2^d$ , breaking down its IRM penalty term into three error components  $\mathcal{J}_{\text{IRMmm}}(S) = \xi_a(S) - \xi_b(S) + \xi_c(S)$ . This is then used to bound  $\hat{\mathcal{L}}(\hat{\mathbf{v}}_S) - \hat{\mathcal{L}}(\beta^*)$ . When  $S \neq S_{\text{inv}}$ , we show this gap to be positive.

### B.1 COMPONENTS OF PENALTY

The penalty term from  $\hat{\mathcal{L}}(\hat{\mathbf{v}})$  is

$$\mathcal{J}_{\text{IRMmm}}(S) = \sum_{e \in \mathcal{E}} \left[ \hat{\mathcal{R}}^e(\hat{\beta}_S) - \hat{\mathcal{R}}^e(\hat{\beta}_S^e) \right] = \xi_a(S) - \xi_b(S) + \xi_c(S), \quad (73)$$

with  $\hat{\beta}_S$  and  $\hat{\beta}_S^e$  as defined in Section 3

$$\xi_a(S) = \sum_{e \in \mathcal{E}} \left[ \hat{\mathcal{R}}^e(\hat{\beta}_S) - \mathcal{R}^e(\beta_S^*) \right], \quad (74)$$

$$\xi_b(S) = \sum_{e \in \mathcal{E}} \left[ \hat{\mathcal{R}}^e(\hat{\beta}_S^e) - \mathcal{R}^e(\beta_S^e) \right], \quad (75)$$

$$\xi_c(S) = \sum_{e \in \mathcal{E}} [\mathcal{R}^e(\beta_S^*) - \mathcal{R}^e(\beta_S^e)]. \quad (76)$$

We bound  $|\xi_b(S)|$  in Corollary 11 and  $|\xi_a(S)|$  in Corollary 12, followed by an analysis of  $|\xi_c(S)|$ .

**Lemma 10** (Penalty component). *For a given environment  $e \in \mathcal{E}$  and feature subset  $S \in 2^d$ , we can bound  $|\hat{\mathcal{R}}^e(\hat{\beta}_S^e) - \mathcal{R}^e(\beta_S^e)|$  with probability  $1 - \delta$ , given a sample size per environment of  $n \geq cw^2(A) = O(d_{\text{inv}})$ :*

$$|\hat{\mathcal{R}}^e(\hat{\beta}_S^e) - \mathcal{R}^e(\beta_S^e)| \leq O \left( \kappa_{\text{inv}} \sqrt{\frac{\log(\frac{1}{\delta})}{n}} \right) \quad (77)$$

where  $\lambda_{\max} := \max_{e \in \mathcal{E}} (\lambda_{\max}(\Sigma^e))$  and  $\text{err}(1/\delta, n)$  as defined in Proposition 7.

*Proof.* First, with triangle inequality, we have for all  $S$ ,

$$|\hat{\mathcal{R}}^e(\hat{\beta}_S^e) - \mathcal{R}^e(\beta_S^e)| \leq |\hat{\mathcal{R}}^e(\hat{\beta}_S^e) - \hat{\mathcal{R}}^e(\beta_S^e)| + |\hat{\mathcal{R}}^e(\beta_S^e) - \mathcal{R}^e(\beta_S^e)|. \quad (78)$$

The second term from Lemma 10 can be bounded by generalized Hoeffding's inequality for unbounded sub-Gaussian random variables, as stated in Proposition 5.10 in (Vershynin, 2011). With probability  $1 - O(\delta_b)$ ,

$$|\hat{\mathcal{R}}^e(\beta_S^e) - \mathcal{R}^e(\beta_S^e)| \leq O \left( \kappa_{\text{inv}} \sqrt{\frac{\log(\frac{1}{\delta})}{n}} \right) \quad (79)$$

For the first term with  $\hat{\mathcal{R}}^e, \hat{\beta}_S^e, \beta^e$  as defined above, it is necessarily less than 0 by the definition of  $\hat{\beta}_S^e$ , the minimizer of  $\hat{\mathcal{R}}^e(\beta_S^e)$ .

$$|\hat{\mathcal{R}}^e(\hat{\beta}_S^e) - \hat{\mathcal{R}}^e(\beta_S^e)| \leq 0. \quad (80)$$

□

**Corollary 11** (Second penalty term). *The following bound holds with probability  $1 - \delta_b$  for all  $S \in 2^d$ :*

$$|\xi_b(S)| \leq O \left( |\mathcal{E}| \kappa_{\text{inv}} \sqrt{\frac{\log(\frac{|\mathcal{E}|}{\delta})}{n}} \right) =: |\xi_b| \quad (81)$$

where  $\text{err}(\frac{1}{\delta}, n)$  is as defined in Proposition 7. The RHS is independent of  $S$ , so we name the upper bound  $|\xi_b|$ .

*Proof.* We expand  $\xi_b(S)$  to get

$$|\xi_b(S)| \leq \sum_{e \in \mathcal{E}} |\hat{\mathcal{R}}^e(\hat{\beta}_S^e) - \mathcal{R}^e(\beta_S^e)| \quad (82)$$

Lemma 10 can directly be applied on each of the different environments  $e \in \mathcal{E}$ . Applying the union bound for the environments produces the desired result. □

**Corollary 12** (First penalty component). *The following bound holds with probability  $1 - \delta_a$  for all  $S \in 2^d$ :*

$$|\xi_a(S)| \leq O \left( \kappa_{\text{inv}} \sqrt{\frac{\log(\frac{1}{\delta})}{|\mathcal{E}|n}} \right) =: |\xi_a| \quad (83)$$

where  $\text{err}(\frac{1}{\delta}, n)$  is as defined in Proposition 7. Again, the RHS is independent of  $S$ , and we name the upper bound  $|\xi_a|$ .

*Proof.* Recall that  $\mathcal{R}(\mathbf{v}) = \sum_{e \in \mathcal{E}} \mathcal{R}^e(\mathbf{v})$  for all classifiers  $\mathbf{v} \in \mathbb{R}^d$ . Additionally,  $\beta_S^* = \operatorname{argmin}_{\mathbf{v} \in V_S} \mathcal{R}(\mathbf{v})$  and  $\hat{\beta}_S = \operatorname{argmin}_{\mathbf{v} \in V_S} \hat{\mathcal{R}}(\mathbf{v})$ . We want to bound the following expression with high probability  $(1 - \delta_a)$ :

$$|\xi_a(S)| = \left| \sum_{e \in \mathcal{E}} \left[ \hat{\mathcal{R}}^e(\hat{\beta}_S) - \mathcal{R}^e(\beta_S^*) \right] \right| = \left| \hat{\mathcal{R}}(\hat{\beta}_S) - \mathcal{R}(\beta_S^*) \right| \quad (84)$$

We may apply the analysis of Lemma 10 to the whole dataset, which can be treated as one environment with  $n|\mathcal{E}|$  points as is done in Zhou et al. (2022).  $\square$

**Proposition 13.** For  $D, \Delta$  as defined in Assumption 4,

$$\xi_c(S) \begin{cases} = 0, & \text{if } S \in S_{\text{inv}} \cup S_r \\ \geq \frac{D^2 \Delta^2}{\lambda_c^3}, & \text{otherwise.} \end{cases} \quad (85)$$

$$\lambda_m = \max \left( \bigcup_{e \in \mathcal{E}} \{\lambda_i^e\}_{i=1}^d \cup \{\lambda_i\}_{i=1}^d \right).$$

To bound  $|\xi_c(S)|$ , we reference the analysis of Equation (29) in (Zhou et al., 2022).

Let  $\{\lambda_i^e\}_{i=1}^d$  and  $\{\lambda_i\}_{i=1}^d$  be the eigenvalues of  $\Sigma^e$  and  $\Sigma$  respectively with corresponding eigenvectors  $\{\mathbf{u}_i^e\}_{i=1}^d$  and  $\{\mathbf{u}_i\}_{i=1}^d$ .

Define  $\Sigma_S^e := \mathbb{E}[(\mathbf{x}_S^e)(\mathbf{x}_S^e)^\top]$ , and  $\Sigma_S := \mathbb{E}[(\mathbf{x}_S^e)(\mathbf{x}_S^e)^\top]$ , where  $\mathbf{x}_S^e \in \mathbb{R}^{|S|}$  and is the pruned ‘‘projection’’ of  $\mathbf{x}^e$  keeping only the indices  $i \in S$ . Note that this differs from the analysis in (Zhou et al., 2022), particularly Assumption 1, which involves taking the inverse of the analogous matrix  $\Sigma_{\Phi}^e$  that is not full rank.

$$\xi_c(S) = \sum_{e \in \mathcal{E}} \|(\beta_S^* - \beta_S^e)\|_{\Sigma_S^e} \quad (86)$$

$$= \sum_{e \in \mathcal{E}} \left( \Sigma_S^{-1} \mathbb{E}[\mathbf{x}_S^e y] - (\Sigma_S^e)^{-1} \mathbb{E}^e[\mathbf{x}_S^e y] \right)^\top \Sigma_S^e \left( \Sigma_S^{-1} \mathbb{E}[\mathbf{x}_S^e y] - (\Sigma_S^e)^{-1} \mathbb{E}^e[\mathbf{x}_S^e y] \right) \quad (87)$$

$$= \sum_e \sum_{i \in S} \left( \mathbb{E}^e[\mathbf{x}_S^e y]^\top \mathbf{u}_i \right)^2 \lambda_i \left( \frac{1}{\lambda_i^e} - \frac{1}{\lambda_i} \right)^2 \quad (88)$$

For any classifier footprint  $S \in 2^d$ , we know that  $\Sigma_S^e = \Sigma_S + \text{Diag}(0, \dots, 0, (\alpha_i^e)^2 - \alpha_i^2, 0, \dots, 0)$ . From this, at most  $d_s$  eigenvalues have a nonzero difference  $\lambda_i^e - \lambda_i = (\alpha_i^e)^2 - \alpha_i^2$ . This is bounded by Assumption 3, and  $|\mathbb{E}^e[\mathbf{x}_S^e y]^\top \mathbf{u}_i|$  is bounded in Assumption 4. Note also that

$$\lambda_i \left( \frac{1}{\lambda_i^e} - \frac{1}{\lambda_i} \right)^2 \geq \frac{\lambda_i (\lambda_i - \lambda_i^e)^2}{(\lambda_i^e)^2 \lambda_i^2} \geq \frac{\Delta^2}{\lambda_m^3}, \quad (89)$$

for  $\lambda_m = \max \left( \bigcup_{e \in \mathcal{E}} \{\lambda_i^e\}_{i=1}^d \cup \{\lambda_i\}_{i=1}^d \right)$ . Then the overall bound for  $\xi_c$  for  $S_{\text{inv}}$  is

$$\xi_c \begin{cases} = 0, & \text{if } S \in S_{\text{inv}} \cup S_r \\ \geq \frac{D^2 \Delta^2}{\lambda_c^3}, & \text{otherwise.} \end{cases} \quad (90)$$

This is a simpler and slightly tighter lower bound than provided previously.

## B.2 ANALYZING EMPIRICAL RISK

We want to show that  $\hat{\mathcal{L}}(S_{\text{inv}}) < \hat{\mathcal{L}}(S)$  for footprint  $S \in 2^d$  such that  $\|\mathbf{S}\|_0 \leq d_{\text{inv}}$ . To borrow the analysis in (Zhou et al., 2022), we observe that there are two categories of footprints: those that include at least one spurious feature ( $\mathcal{S}_{\text{with-spu}}$ ) and those that do not ( $\mathcal{S}_{\text{no-spu}}$ ).

We will first show that  $\forall S \subseteq \mathcal{S}_{\text{with-spu}}, \hat{\mathcal{L}}(S_{\text{inv}}) \leq \hat{\mathcal{L}}(S)$ . Then, we will show the same for the no-spurious feature selectors, that is  $\forall S \subseteq \mathcal{S}_{\text{no-spu}}, \hat{\mathcal{L}}(S_{\text{inv}}) \leq \hat{\mathcal{L}}(S)$ . We will use  $\beta_{\text{inv}} = \beta_{S_{\text{inv}}}$  as a shorthand.

First, for  $S \subseteq \mathcal{S}_{\text{with-spu}}$ , we have

$$\begin{aligned}\hat{\mathcal{L}}(S) - \hat{\mathcal{L}}(S_{\text{inv}}) &= \sum_e \left[ \hat{\mathcal{R}}^e(\hat{\beta}_S) + \rho \left( \hat{\mathcal{R}}^e(\hat{\beta}_S) - \hat{\mathcal{R}}^e(\hat{\beta}_S^e) \right) \right] \\ &\quad - \sum_e \left[ \hat{\mathcal{R}}^e(\hat{\beta}_{\text{inv}}) + \rho \left( \hat{\mathcal{R}}^e(\hat{\beta}_{\text{inv}}) - \hat{\mathcal{R}}^e(\hat{\beta}_{\text{inv}}^e) \right) \right]\end{aligned}\quad (91)$$

$$\begin{aligned}&= \sum_e \left[ \hat{\mathcal{R}}^e(\hat{\beta}_S) \right] + \rho \xi_a(S) - \rho \xi_b(S) + \rho \xi_c(S) \\ &\quad - \xi_a(S_{\text{inv}}) - |\mathcal{E}| \cdot \sigma_{\text{inv}}^2 - \rho \xi_a(S_{\text{inv}}) + \rho \xi_b(S_{\text{inv}}) - \rho \xi_c(S_{\text{inv}})\end{aligned}\quad (92)$$

For the second equality, note the definition of  $\xi_a(S_{\text{inv}})$ , using  $\sigma_{\text{inv}}^2$  as the variance of  $\epsilon_{\text{inv}}$ :

$$\xi_a(S_{\text{inv}}) = \sum_e \left[ \hat{\mathcal{R}}^e(\hat{\beta}_{\text{inv}}) - \mathcal{R}^e(\beta^*) \right] = \sum_e \left[ \hat{\mathcal{R}}^e(\hat{\beta}_{\text{inv}}) \right] - \sum_e \sigma_{\text{inv}}^2.$$

By dropping error terms  $\sum_e \hat{\mathcal{R}}^e(\cdot) \geq 0$  and using  $|\xi_a(S)| \leq |\xi_b(S)|$  for all  $S \in 2^d$ , we get a lower bound.

$$\begin{aligned}\hat{\mathcal{L}}(S) - \hat{\mathcal{L}}(S_{\text{inv}}) &\geq -\rho \xi_b(S) + \rho \xi_c(S) \\ &\quad - \xi_a(S_{\text{inv}}) - |\mathcal{E}| \cdot \sigma_{\text{inv}}^2 - \rho \xi_a(S_{\text{inv}}) - \rho \xi_c(S_{\text{inv}})\end{aligned}\quad (93)$$

Additionally, from eq. (85),  $\xi_c(S_{\text{inv}}) = 0$ .

$$\geq -(2\rho + 1)|\xi_b(S)| - |\mathcal{E}| \cdot \sigma_{\text{inv}}^2 + \rho \xi_c(S) \quad (94)$$

We then select  $\rho$  to eliminate the  $\xi_c(S)$  term, while also producing a positive term in the RHS. Specifically, let  $\rho \xi_c(S) \geq 2|\mathcal{E}| \cdot \sigma_{\text{inv}}^2$ , getting

$$\rho = \frac{2|\mathcal{E}| \sigma_{\text{inv}}^2}{D^2 \Delta^2 / \lambda_{\max}^3} \geq \frac{2|\mathcal{E}| \sigma_{\text{inv}}^2}{\xi_c(S)}. \quad (95)$$

Setting the weight  $\rho$  to the LHS, we can write the gap as

$$\begin{aligned}\hat{\mathcal{L}}(S) - \hat{\mathcal{L}}(S_{\text{inv}}) &\geq -(2\rho + 1)O \left( |\mathcal{E}| \sqrt{\frac{\log \frac{|\mathcal{E}|}{\delta}}{n}} \right) + |\mathcal{E}| \cdot \sigma_{\text{inv}}^2.\end{aligned}\quad (96)$$

Finally, we solve for the sample complexity required to differentiate  $S_{\text{inv}}$  and  $S$ .

$$n > \frac{(2\rho + 1)^2 \log \left( \frac{|\mathcal{E}|}{\delta_b} \right)}{\sigma_{\text{inv}}^4} \quad (97)$$

Likewise, to analyze the classifiers that include random features but not spurious features, we now consider  $S \subseteq \mathcal{S}_{\text{no-spu}}$ . Note that  $S_r$  refers to the set of features  $i \in [d]$  such that  $\mathbf{x}_i^e$  is a random feature;  $S \cap S_r$  is the set of random features in  $S$ .

$$\begin{aligned}\hat{\mathcal{L}}(S) - \hat{\mathcal{L}}(S_{\text{inv}}) &= \sum_e \left[ \hat{\mathcal{R}}^e(\hat{\beta}_S) \right] + \rho \xi_a(S) - \rho \xi_b(S) + \rho \xi_c(S) \\ &\quad - \xi_a(S_{\text{inv}}) - |\mathcal{E}| \cdot \sigma_{\text{inv}}^2 - \rho \xi_a(S_{\text{inv}}) + \rho \xi_b(S_{\text{inv}}) - \rho \xi_c(S_{\text{inv}})\end{aligned}\quad (98)$$

From (Zhou et al., 2022), we have  $R^e(\beta_S) - R^e(\beta^*) = \sum_{i \in S_{\text{inv}} \setminus S} \gamma_i x_{\text{inv},i}^2$ , so

$$\xi_a(S) = \sum_e \left[ \hat{\mathcal{R}}^e(\hat{\beta}_S) - R^e(\beta^*) - \mathcal{R}^e(\beta_S) + R^e(\beta^*) \right] = \sum_e \hat{\mathcal{R}}^e(\hat{\beta}_S) - |\mathcal{E}| \cdot \sigma_{\text{inv}}^2 - |\mathcal{E}| \sum_{i \in S_{\text{inv}} \setminus S} \gamma_i x_{\text{inv},i}^2.$$

We also use the definitions in Assumption 2 to lower bound  $\gamma_i$  and  $x_{\text{inv},i}$  with high probability.

**Remark 12.** We need a **lower bound** specifically to provide the sample complexity result after. In other words, we work both with the subgaussian norm to “upper bound” the features, and we need a lower bound on the label noise variance  $\sigma_{\text{inv}}^2$  and the smallest feature variance  $\sigma_0^2$ .

Proceeding to compare the empirical losses,

$$\begin{aligned}\hat{\mathcal{L}}(S) - \hat{\mathcal{L}}(S_{\text{inv}}) &\geq \left( \xi_a(S) + |\mathcal{E}| \cdot \sigma_{\text{inv}}^2 + |\mathcal{E}| \sum_{i \in S_{\text{inv}} \setminus S} \gamma_i x_{\text{inv},i}^2 \right) \\ &\quad + \rho \xi_a(S) - \rho \xi_b(S) + \rho \xi_c(S) \\ &\quad - \xi_a(S_{\text{inv}}) - |\mathcal{E}| \cdot \sigma_{\text{inv}}^2 \\ &\quad - \rho \xi_a(S_{\text{inv}}) + \rho \xi_b(S_{\text{inv}}) - \rho \xi_c(S_{\text{inv}}),\end{aligned}\tag{99}$$

$$\geq -(2\rho + 1)|\xi_b| + |\mathcal{E}| \cdot \bar{\gamma} \cdot \sigma_0^2.\tag{100}$$

Again, we eliminate positive terms; additionally,  $\xi_c(S) = \xi_c(S_{\text{inv}}) = 0$ . The resulting sample complexity for differentiating  $\hat{S}_{\text{inv}}$  from  $S \in \mathcal{S}_{\text{no-spu}}$  is then the following:

$$n > \frac{(2\rho + 1)^2 \log \frac{|\mathcal{E}|}{\delta_b}}{\bar{\gamma}^2 \sigma_0^4}\tag{101}$$

Together, Equation (97) and Equation (101) form the sample complexity; we take the max between the both. We note that both are  $O\left(\frac{|\mathcal{E}|}{\delta_b}\right)$ .

Because  $|\mathcal{S}_{\text{with-spu}} \cup \mathcal{S}_{\text{no-spu}}| = \binom{d}{d_{\text{inv}}} \leq d^{d_{\text{inv}}}$ , we can set  $\delta_b = \frac{\delta}{d^{d_{\text{inv}}}}$  before taking the union bound, incurring a  $\log \frac{|\mathcal{E}|}{\delta_b} = d_{\text{inv}} \log d + \log |\mathcal{E}| - \log \delta \leq d_{\text{inv}} \log \frac{d|\mathcal{E}|}{\delta}$ . Under this sample complexity, we have  $\hat{\mathcal{L}}(S) - \hat{\mathcal{L}}(S_{\text{inv}})$  for all  $|S|_0 \leq d_{\text{inv}}$ ,  $S \neq S_{\text{inv}}$ .  $\square$

### B.3 COMPARISON: EMPIRICAL LOSS WITH POPULATION MINIMA

We provide a proof for Theorem 2 of the main paper, restated below for reference:

**Theorem 2** For population minimizers as defined in Equation (6), and  $n$  samples per environment  $e \in \mathcal{E}$ , for a total of  $N = |\mathcal{E}|n$  across the whole training set, we have

$$\hat{\mathcal{L}}(\beta^*) < \hat{\mathcal{L}}(\beta_S^*), \quad |S| \leq d_{\text{inv}}, S \neq S_{\text{inv}},\tag{102}$$

if  $n > O\left(\text{poly}(d_{\text{inv}}) \log\left(\frac{d|\mathcal{E}|}{\delta}\right)\right)$  with constants specified below.

*Proof.* We want to show that

$$\hat{\mathcal{L}}(\beta_{\text{inv}}^*) < \hat{\mathcal{L}}(\beta_S^*)\tag{103}$$

for  $|S| < d_{\text{inv}}$  and  $S \neq S_{\text{inv}}$ . We will use the notation  $\beta_{\text{inv}}^* := \beta^* = \beta_{S_{\text{inv}}}^*$  for the invariant optimal predictor.

First note that with high probability  $1 - \delta_1$ ,

$$\hat{\mathcal{L}}(\beta_{\text{inv}}^*) = \sum_{e \in \mathcal{E}} \hat{\mathcal{R}}^e(\beta_{\text{inv}}^*) + \rho \sum_{e \in \mathcal{E}} \left[ \hat{\mathcal{R}}^e(\beta_{\text{inv}}^*) - \hat{\mathcal{R}}^e(\beta_{\text{inv}}^e) + \hat{\mathcal{R}}^e(\beta_{\text{inv}}^e) - \hat{\mathcal{R}}^e(\hat{\beta}_{\text{inv}}^e) \right]\tag{104}$$

$$= \sum_{e \in \mathcal{E}} \hat{\mathcal{R}}^e(\beta_{\text{inv}}^*) + \rho(0 + \xi_b(S_{\text{inv}}))\tag{105}$$

$$\leq \sum_{e \in \mathcal{E}} \hat{\mathcal{R}}^e(\beta_{\text{inv}}^*) + \rho |\mathcal{E}| O\left(\kappa_{\text{inv}} \sqrt{\frac{\log(\frac{|\mathcal{E}|}{\delta_1})}{n}}\right)\tag{106}$$

The second equality uses two definitions. First,  $\beta_{\text{inv}}^e = \beta_{\text{inv}}^*$  for  $e \in \mathcal{E}$ , so  $\hat{\mathcal{R}}^e(\beta_{\text{inv}}^e) - \hat{\mathcal{R}}^e(\beta_{\text{inv}}^*) = 0$ .

Then, we use the definition of  $\xi_b(S)$  from Equation (75), which is upper bound in Corollary 11. Also, the equality  $\beta_{\text{inv}}^{e_1} = \beta_{\text{inv}}^{e_2}$  for  $e_1, e_2 \in \mathcal{E}$ , because the subset selects only the invariant features that are shared between all environments.

Next, with high probability  $1 - \delta_2$ ,

$$\hat{\mathcal{L}}(\beta_S^*) = \sum_{e \in \mathcal{E}} \hat{\mathcal{R}}^e(\beta_S^*) + \rho \sum_{e \in \mathcal{E}} \left[ \hat{\mathcal{R}}^e(\beta_S^*) - \hat{\mathcal{R}}^e(\beta_S^e) + \hat{\mathcal{R}}^e(\beta_S^e) - \hat{\mathcal{R}}^e(\hat{\beta}_S^e) \right] \quad (107)$$

$$\geq \sum_{e \in \mathcal{E}} \hat{\mathcal{R}}^e(\beta_S^*) + \rho \xi_c(S) + \rho \sum_{e \in \mathcal{E}} \left[ -2O \left( c_{\text{total}} \sqrt{\frac{\log(|\mathcal{E}|)}{n}} \right) \right] - \rho |\xi_b(S)| \quad (108)$$

$$\geq \sum_{e \in \mathcal{E}} \hat{\mathcal{R}}^e(\beta_S^*) + \rho |\mathcal{E}| O \left( \frac{D^2 \Delta^2}{\lambda_m^3} - c_{\text{total}} \sqrt{\frac{\log(|\mathcal{E}|)}{n}} \right) - \rho |\mathcal{E}| O \left( \kappa_{\text{inv}} \sqrt{\frac{\log(|\mathcal{E}|)}{n}} \right) \quad (109)$$

where  $c_{\text{total}}^2 = c_0 \max\{(c_a^2 \kappa_s^2 + c_r^2 \kappa_r^2)^2, \kappa_{\text{inv}}^4\}$  for some constant  $c_0 > 0$ . To get the first inequality, we want to bound  $|\hat{\mathcal{R}}(\beta_S^*) - \mathcal{R}(\beta_S^*)|$  and  $|\hat{\mathcal{R}}(\beta_S^e) - \mathcal{R}(\beta_S^e)|$ . Hoeffding's inequality may be used, but we need the sub-Gaussian norm of the least squares error; let this be  $Z_j := (y_j - (\beta_S^*)^\top \mathbf{x}_j^e)^2$  for  $j \in [n]$ :

$$Z_j = (y_j - (\beta_S^*)^\top \mathbf{x}_j^e)^2 = ((\beta_{\text{inv}}^* - \beta_S^*)^\top \mathbf{x}_j^e + \epsilon_{\text{inv},j})^2 \leq (\|\beta_{\text{inv}}^* - \beta_S^*\|_2 \|\mathbf{x}_j^e\|_2 + \epsilon_{\text{inv},j})^2. \quad (110)$$

By the assumption that classifiers  $\beta_{\text{inv}}^*, \beta_S^*$  are normalized, we have  $\|\beta_{\text{inv}}^* - \beta_S^*\|_2 \leq 2$ , and  $\|\mathbf{x}^e\|_2$  is sub-Gaussian; from Lemma 4 with probability  $1 - \delta$ , we get

$$\mathbb{E}^e[Z_i] = O(\max\{(c_s + c_a \kappa_s + c_r \kappa_r)^2, \kappa_{\text{inv}}^2\}), \quad (111)$$

and

$$c_{\text{total}} = \|Z_i\|_{\psi_2} \leq O(\max\{4\|\mathbf{x}^e\|_2^2 + 4\|\mathbf{x}^e\|_2 \epsilon_{\text{inv}} + \epsilon_{\text{inv}}^2\}) = O((c_s + c_a \kappa_s + c_r \kappa_r)^2, \kappa_{\text{inv}}^4). \quad (112)$$

Applying Hoeffding's inequality to  $|\hat{\mathcal{R}}^e(\beta_S^*) - \mathcal{R}^e(\beta_S^*)|$  then gets the bound of  $O \left( c_{\text{total}} \sqrt{\frac{\log(|\mathcal{E}|)}{n}} \right)$  with probability

$1 - \delta_1$ . Likewise,  $|\hat{\mathcal{R}}^e(\beta_S^e) - \mathcal{R}^e(\beta_S^e)| \leq O \left( c_{\text{total}} \sqrt{\frac{\log(|\mathcal{E}|)}{n}} \right)$  with probability  $1 - \delta_2$ . We apply these inequalities over  $|\mathcal{E}|$

environments, so we set  $\delta_1 = \delta_2 = \frac{\delta}{2|\mathcal{E}|}$ . To complete the first inequality, we use  $\xi_b(S) = \sum_{e \in \mathcal{E}} [\hat{\mathcal{R}}^e(\beta_S^e) - \hat{\mathcal{R}}^e(\hat{\beta}_S^e)]$  and  $\xi_c(S) = \sum_{e \in \mathcal{E}} [\mathcal{R}^e(\beta_S^*) - \mathcal{R}^e(\beta_S^e)]$ .

We use the upper bound on  $|\xi_b(S)|$  in Corollary 11 and the lower bound on  $\xi_c(S)$  from Proposition 13, noting that  $\xi_c(S)$  is greater than some constant  $c > 0$  with high probability provided that there exists at least one feature  $i$  such that  $\alpha_i^{e_1} \neq \alpha_i^{e_2}$ , for  $e_1, e_2 \in \mathcal{E}, e_1 \neq e_2$ .

Next, we demonstrate that  $\hat{\mathcal{L}}(\beta_S^*) - \hat{\mathcal{L}}(\beta_{\text{inv}}^*) > 0$ .

$$\hat{\mathcal{L}}(\beta_S^*) - \hat{\mathcal{L}}(\beta_{\text{inv}}^*) \geq \sum_{e \in \mathcal{E}} \hat{\mathcal{R}}^e(\beta_S^*) - \sum_{e \in \mathcal{E}} \hat{\mathcal{R}}^e(\beta_{\text{inv}}^*) + \rho |\mathcal{E}| O \left( \frac{D^2 \Delta^2}{\lambda_m^3} - c_{\text{total}} \sqrt{\frac{\log|\mathcal{E}|/\delta_3}{n}} \right) \quad (113)$$

$$= -|\mathcal{E}| \sigma_{\text{inv}}^2 + \rho |\mathcal{E}| |\xi_c(S)| - \rho |\mathcal{E}| c_{\text{total}} \sqrt{\frac{\log|\mathcal{E}|/\delta_3}{n}} \quad (114)$$

We observe that  $\hat{\mathcal{R}}^e(\beta_S^*) - \sum_{e \in \mathcal{E}} \hat{\mathcal{R}}^e(\beta_{\text{inv}}^*) \geq 0 - |\mathcal{E}| \sigma_{\text{inv}}^2$ . First, we set  $\rho$  such that  $\rho |\mathcal{E}| |\xi_c(S)| - |\mathcal{E}| \sigma_{\text{inv}}^2 > \sigma_{\text{inv}}^2 > 0$ . This can be satisfied by setting

$$\rho > \frac{\sigma_{\text{inv}}^2 + |\mathcal{E}| \sigma_{\text{inv}}^2}{|\mathcal{E}| \xi_c(S)}, \quad (115)$$

which can be satisfied with  $\rho > \frac{(1+|\mathcal{E}|)\sigma_{\text{inv}}^2}{|\mathcal{E}| D^2 \Delta^2 / \lambda_{\text{max}}^3}$ . Then, we can guarantee that  $\hat{\mathcal{L}}(\beta_S^*) - \hat{\mathcal{L}}(\beta_{\text{inv}}^*) > 0$ , with the sample complexity computed in the proof of Theorem 1:

$$n > \max \left\{ \frac{(2\rho + 1)^2 c_{\text{total}}^2 \log \frac{|\mathcal{E}|}{\delta_b}}{\sigma_{\text{inv}}^4}, \frac{(2\rho + 1)^2 c_{\text{total}}^2 \log \frac{|\mathcal{E}|}{\delta_b}}{\bar{\gamma}^2 \sigma_0^4} \right\} \quad (116)$$

Again, we need to solve  $\binom{d}{d_{\text{inv}}} \leq d^{d_{\text{inv}}}$  optimization problems, so we set  $\delta_b = \frac{\delta}{d^{d_{\text{inv}}}}$  before taking the union bound.

With  $n > c_{\text{total}}^2 d_{\text{inv}} \log \frac{d|\mathcal{E}|}{\delta}$ , we achieve the desired result.

We note that in the general case,  $c_{\text{total}} = O(\max\{(c_s + c_a \kappa_s + c_r \kappa_r)^2, \kappa_{\text{inv}}^4\})$ . With the original setting in Zhou et al. (2022),  $\zeta_s = \mathbf{1}^{d_s}$ . This leads to  $c_s = \sqrt{d_s}$ . Then, with  $c_{\text{total}}^2 = O(c_s^4) = O(d_s^2)$ , we get a polynomial dependency on  $d_s$  without a more refined analysis.  $\square$

**Remark:** It is also unlikely that spurious features are all equally correlated with the label, as assumed when we take  $\zeta_s^e$  to be all ones. Even with sparse feature selection, it is possible for the predictor to pick up the largest elements of  $\mathbf{x}^e$ . An example of a “heavy-hitter” would be a spurious feature  $x_j^e$  that has a strong correlation with the label through a high  $\zeta_j^e$ , contributing to  $c_s^2$ . In the case that  $j \in S$ , we end up with  $c_{\text{total}} \propto c_s^2$ . This is explored further in Corollary 8.

On the other end of the spectrum, a more evenly distributed feature vector can demonstrate even tighter bounds. This is explored further in Corollary 9.  $\square$

#### B.4 COMPARISON: SPARSE IRM VS ERM AND SPARSE ERM

The following propositions give a characterization of this data generation model to motivate the use of IRM. Proposition 17 shows that ERM with sparsity constraints on the global population is also unable to find the invariant features.

We can represent a given classifier  $\mathbf{v}$  in its three parts  $\mathbf{v} = [\mathbf{v}_{\text{inv}}, \mathbf{v}_s, \mathbf{v}_r]^\top$ .

**Proposition 14** (Invariant Optimal Classifier is ground truth). *In the problem setting defined by Equation (2),  $\beta^* = [\gamma^\top, (\mathbf{0}^{d_s})^\top, (\mathbf{0}^{d_r})^\top]$ , and is also a solution to Equation (1).*

*Proof.* First,  $\mathcal{R}(\beta^*) = \text{Var}(\epsilon_{\text{inv}})$ . Let a comparison be made to candidate parameters  $\beta \in S$  with  $|S| \leq d_{\text{inv}}$ . If any of the parameters in  $\beta$  are random, that is  $\beta_i \neq 0$  for  $i \in S_r$ , we lose information and  $\mathcal{R}^e(\beta) \geq \mathcal{R}^e(\beta^*)$ . Likewise,  $\mathcal{J}^e(\beta) = \mathcal{R}^e(\beta) - \min_{\beta' \in \text{Sp}(S)} \mathcal{R}^e(\beta') \geq 0 = \mathcal{J}^e(\beta)$ . So, any parameter  $\beta$  with random features will not be a solution to Equation (1).

Next, we consider  $\beta_i \neq 0$  for  $i \in S_{\text{sp}}$ , potentially including spurious features. In this case, we want to show that  $\mathcal{L}(\beta^*) - \mathcal{L}(\beta)$  is negative:

$$\begin{aligned} \mathcal{L}(\beta^*) - \mathcal{L}(\beta) &= \sum_{e \in \mathcal{E}} \mathcal{R}^e(\beta^*) - \mathcal{R}^e(\beta) + \rho \mathcal{J}^e(\beta^*) - \rho \mathcal{J}^e(\beta) \\ &\geq \sum_{e \in \mathcal{E}} \text{Var}(\epsilon_{\text{inv}}) - \mathcal{R}^e(\beta) - \rho \mathcal{J}^e(\beta) \\ &= \sum_{e \in \mathcal{E}} \text{Var}(\epsilon_{\text{inv}}) - (1 + \rho) \mathcal{R}^e(\beta) + \rho \min_{\beta' \in \text{Sp}(S)} \mathcal{R}^e(\beta') \end{aligned}$$

We know that  $\mathcal{R}^e(\beta) - \min_{\beta' \in \text{Sp}(S)} \mathcal{R}^e(\beta') = J_S^e \geq 0$ . With an appropriately selected  $\rho$ , we can see that the penalty incurred by a non-invariant  $\beta$  will incur greater IRM population loss than the optimum  $\beta^*$ .  $\square$

**Proposition 15** (ERM does not overfit on random features). *Because  $\mathbf{x}_r^e$  is independent from the other features and zero-mean, we can guarantee that  $\mathbf{v}^e$  does not have elements on the random noise features  $\mathbf{x}_r$ . In other words,*

$$\mathcal{R}^e([\mathbf{v}_{\text{inv}}^\top, \mathbf{v}_s^\top, \mathbf{v}_r^\top]^\top) \leq \mathcal{R}^e([\mathbf{v}_{\text{inv}}^\top, \mathbf{v}_s^\top, \mathbf{0}^\top]^\top) \quad (117)$$

*Proof.* We can see that

$$\begin{aligned} \mathcal{R}^e([\mathbf{v}_{\text{inv}}^\top, \mathbf{v}_s^\top, \mathbf{v}_r^\top]^\top) &= \mathbb{E}^e \left[ (y - [\mathbf{v}_{\text{inv}}, \mathbf{v}_s][\mathbf{x}_{\text{inv}}^e, \mathbf{x}_s^e]) - \mathbf{v}_r^\top \mathbf{x}_r \right)^2] \\ &= \mathbb{E}^e \left[ (y - [\mathbf{v}_{\text{inv}}, \mathbf{v}_s][\mathbf{x}_{\text{inv}}^e, \mathbf{x}_s^e])^2 - 2(y - [\mathbf{v}_{\text{inv}}, \mathbf{v}_s][\mathbf{x}_{\text{inv}}^e, \mathbf{x}_s^e])(\mathbf{v}_r^\top \epsilon_r) + (\mathbf{v}_r^\top \epsilon_r)^2 \right] \\ &= \mathcal{R}^e([\mathbf{v}_{\text{inv}}^\top, \mathbf{v}_s^\top, \mathbf{0}^\top]^\top) + \mathbb{E}[(\mathbf{v}_r^\top \epsilon_r)^2] \\ &\geq \mathcal{R}^e([\mathbf{v}_{\text{inv}}^\top, \mathbf{v}_s^\top, \mathbf{0}^\top]^\top). \end{aligned}$$

The difference  $\mathbb{E}[(\mathbf{v}_r^\top \epsilon_r)^2] = 0$  if and only if  $\mathbf{v}_r^\top \epsilon_r = 0$ , in which case  $\mathbf{v}_R = \mathbf{0}$ .  $\square$

**Proposition 16** (ERM overfits on spurious features). *When there exists spurious feature  $i \in [d]$  such that  $\alpha_i^2 < \sigma_{\text{inv}}^2$ , then*

$$\beta^* = [\gamma, 0, 0]^\top \notin \underset{\mathbf{v} \in \mathbb{R}^d}{\operatorname{argmin}} \mathcal{R}(\mathbf{v}). \quad (118)$$

*Thus, unconstrained ERM on the even mixture of environments will not be able to detect the ground truth.*

*Proof.* This can be seen by setting  $\mathbf{v} = \mathbf{e}_i$  where  $\frac{\mathbf{e}_i}{\zeta_i}$  is the standard basis vectors with a 1 on the  $i$ th feature on the spurious feature  $i$ ,

$$\mathcal{R}\left(\frac{\mathbf{e}_i}{\zeta_i}\right) = \mathbb{E} [\mathbb{E}^e[(y - y - \alpha_i^e \epsilon_{s,i})^2]] = \mathbb{E} [(\alpha_i^e)^2] = \alpha_i^2 < \sigma_{\text{inv}}^2 = \mathcal{R}([\gamma, \mathbf{0}, \mathbf{0}]).$$

While  $\mathbf{e}_i$  is not generally the minimizer of the global population loss, it does show that  $[\gamma, \mathbf{0}, \mathbf{0}]$  does not achieve minimum loss when there are no restrictions on footprint/L0 norm.  $\square$

**Proposition 17** (Sparse ERM doesn't find invariant features). *With population risk, ERM with constrained L0 norm does not find invariant features when  $\alpha_i^2 < \sigma_{\text{inv}}^2$ .*

$$\min_{\|\mathbf{v}\|_0 \leq d_{\text{inv}}} \mathcal{R}(\mathbf{v}) < \mathcal{R}(\beta^*) \quad (119)$$

*Proof.* We can observe that

$$\min_{\|\mathbf{v}\|_0 \leq d_{\text{inv}}} \mathcal{R}(\mathbf{v}) \leq \mathcal{R}\left(\frac{\mathbf{e}_i}{\zeta_i}\right) = \alpha_i^2 < \sigma_{\text{inv}}^2 = \mathcal{R}(\beta^*) \quad (120)$$

In other words, ERM with sparsity constraints is not guaranteed to find the exact invariant footprint in the population case.  $\square$

**Proposition 18** (Sparse IRM finds invariant features). *IRM with sparsity can find the invariant optimal classifier, where*

$$\beta^* = \underset{\mathbf{v}}{\operatorname{argmin}} \mathcal{L}(\mathbf{v}) \text{ s.t. } \|\mathbf{v}\|_0 \leq d_{\text{inv}} \quad (121)$$

*Proof.* With the assumptions in Appendix A.2, we can analyze eq. (76) to see that the penalty term added by  $\mathcal{L}(\mathbf{v})$ , which is  $\mathcal{R}^e(\mathbf{v}) - \mathcal{R}^e(\mathbf{v}^e)$  in the population case, is only zero when the classifier learned has nonzero elements on the invariant features only.  $\square$

## C PROOF OF THEOREM 3 (ALGORITHMIC ERROR BOUND)

### C.1 ITERATIVE HARD THRESHOLDING

Below is the proof for Appendix C.1. We copy the theorem below for reference:

**Theorem 3** Assume  $n$  samples per training environment, for  $n > Q \left( \text{poly}(d_{\text{inv}}) \log(d) \log\left(\frac{|E|}{\delta}\right) \right)$ . Together with assumptions in Appendix A.2, we can say with probability at least  $1 - \delta$ :

$$\tilde{\beta} = \min_{\mathbf{v}} \hat{\mathcal{L}}(\mathbf{v}) \text{ s.t. } \mathbf{v} \in \mathbb{R}^d, \|\mathbf{v}\|_0 \leq d_{\text{inv}},$$

returns a parameter  $\tilde{\beta}$  with low estimation error  $\|\tilde{\beta} - \beta_{\text{inv}}^*\|_2 \leq O(\sqrt{\frac{d_{\text{inv}}}{n}})$ .

*Proof.* We apply Theorem 3 of Jain et al. (2014), which specifically contains an example for Sparse Linear Regression.

**RSS and RSC:** With parameter  $\beta \in \mathbb{R}^d, \beta' \in \mathbb{R}^d$ , we define

$$\delta \mathcal{R}^e(\beta) := \mathcal{R}^e(\beta') - \mathcal{R}^e(\beta) - \langle \nabla_{\beta} \mathcal{R}^e(\beta), \beta' - \beta \rangle, \quad (122)$$

and likewise

$$\delta \mathcal{J}^e(\beta) := \mathcal{J}^e(\beta') - \mathcal{J}^e(\beta) - \langle \nabla_{\beta} \mathcal{J}^e(\beta), \beta' - \beta \rangle. \quad (123)$$

To apply results from Iterative Hard Thresholding, we show that this problems satisfies the Restricted Strong Convexity (RSC) and Restricted Strong Smoothness (RSS) conditions. RSC requires

$$\delta\mathcal{R}^e(\beta) + \rho\delta\mathcal{J}^e(\beta) \geq \frac{\alpha_{\text{IRM}}}{2}\|\Delta\|_2^2. \quad (124)$$

Likewise, for RSS condition,

$$\delta\mathcal{R}^e(\beta) + \rho\delta\mathcal{J}^e(\beta) \leq \frac{L_{\text{IRM}}}{2}\|\Delta\|_2^2. \quad (125)$$

Let  $\Delta = \beta' - \beta$ .

$$\delta\mathcal{J}^e(\beta) = \mathcal{J}^e(\beta + \Delta) - \mathcal{J}^e(\beta) + \langle \nabla_{\beta} \mathcal{J}^e(\beta), \Delta \rangle \quad (126)$$

$$\begin{aligned} &= \left\| \frac{1}{n}(-X^e)^\top (Y - X^e\beta - X^e\Delta) \right\|_2^2 - \left\| \frac{1}{n}(-X^e)^\top (Y - X^e\beta) \right\|_2^2 \\ &\quad + \left\langle \frac{2}{n^2}(-X^e)^\top X^e X^e (Y - X^e\beta), \Delta \right\rangle \end{aligned} \quad (127)$$

$$= \left\| \frac{1}{n}(X^e)^\top X^e \Delta \right\|_2^2 \quad (128)$$

$$= \frac{1}{n^2} \|(X^e)^\top X^e \Delta\|_2^2 \quad (129)$$

If we set  $\alpha_{\text{IRM}} = \alpha_s$  as defined in (Jain et al., 2014), which defines RSC for the least square component  $\delta\mathcal{R}^e(\beta)$ , we recover the RSC property:

$$\delta\mathcal{R}^e(\beta) + \rho\delta\mathcal{J}^e(\beta) = \frac{1}{n}\|X^e\Delta\|_2^2 + \frac{1}{n^2}\|(X^e)^\top X^e \Delta\|_2^2 \geq \frac{1}{n}\|X^e\Delta\|_2^2 \geq \frac{\alpha_{\text{IRM}}}{2}\|\Delta\|_2^2,$$

since  $\frac{1}{n}X^e{}^\top X^e$  is positive semi-definite.

We then want to upper-bound  $\delta\mathcal{J}^e(\beta)$ , and we will also use  $L_s$  as defined in Jain et al. (2014) Let  $X = X^e$ , the data matrix for a single environment. If we write the eigendecomposition  $X^\top X = V\Lambda V^\top$ , with diagonal elements of  $\Lambda$  as  $\lambda_i$  for  $i \in [d]$ , we can also write  $\Delta = V\alpha$  for some coefficients  $\alpha$ . For least squares, we have bounds for  $\|X\Delta\|_2^2 = \|\Lambda^{1/2}\alpha\|_2^2 = \sum_{i=1}^d \lambda_i \alpha_i^2 \leq \frac{L_s}{2}\|\Delta\|_2^2$ .

Define  $L_0 = \frac{L_s}{2}\|\Delta\|_2^2$ . First,  $\|\Delta\|_2^2 = \|\alpha\|_2^2 = 1$ . Furthermore, we note that  $\Delta = \beta - \beta'$  for iterates of the IHT algorithm, and let  $\|\beta\|_0 = s$  and  $\|\beta'\|_0 = s'$ , where  $s + s' < d$ . Then,  $\|\Delta\|_0 \leq s + s'$ . Because  $\Delta$  is low-rank, we can assume there is a set  $T$  of eigenvectors where  $|T| \leq s' + s$ , which defines  $\|X\Delta\|_2^2$ .

$$\sum_{i=1}^d \lambda_i \alpha_i^2 = \sum_{i \in T} \lambda_i \alpha_i^2 \leq L_0, \quad \text{and} \quad \sum_{i=1}^d \lambda_i^2 \alpha_i^2 = \sum_{i \in T} \lambda_i^2 \alpha_i^2. \quad (130)$$

The bounds apply for restricted eigenvectors  $\Delta$  where  $\sum_{i \in T} \lambda_i^2 \alpha_i^2 \leq L_0$  for any  $\|\alpha\|_2^2 = 1$ . Then we can say that  $\sum_{i \in T} \lambda_i^2 \alpha_i^2 = \lambda_i \leq L_0$  for all  $i \in T$ , which means  $\frac{\lambda_i}{L_0} \leq 1$ . Then, it must be that

$$\sum_{i \in T} \frac{\lambda_i^2}{L_0^2} \alpha_i^2 \leq \sum_{i \in T} \frac{\lambda_i}{L_0} \alpha_i^2 \leq 1, \quad (131)$$

and

$$\sum_{i \in T} \lambda_i^2 \alpha_i^2 \leq L_0^2. \quad (132)$$

Then we can set  $L_{\text{IRM}} = L_s^2$ . Following the example of Theorem 3 in (Jain et al., 2014), for sparse linear regression, we apply the same sample complexity  $n > 5c_1 d_{\text{inv}} \log d(\lambda_{\min}^e)$  where  $\lambda_{\min}^e = \min_{i \in d} \lambda(X^e)$  will get us the conditioning constant:

$$K := \frac{(\lambda_{\max}^e)^2}{\lambda_{\min}^e}. \quad (133)$$

We substitute this back into the error bound of IHT. Then with probability at least  $1 - c_1 p^{-c_2}$  for constants  $c_1, c_2 > 0$ , we end up with the bound

$$\|\tilde{\beta} - \beta_{\text{inv}}^*\|_2 \leq c_1 \frac{\lambda_{\max}^2}{9\kappa_s^2} \kappa_s \max_{i \in [d_s]} \alpha_i^e \kappa_s \sqrt{\frac{d_{\text{inv}} \log d}{n}} + 2\sqrt{\frac{\sigma_{\text{inv}}^2}{\kappa_s^2}} = O\left(\lambda_{\max}^2 A \sqrt{\frac{d_{\text{inv}} \log d}{n}} + \frac{\sigma_{\text{inv}}}{\kappa_s}\right). \quad (134)$$

**Remark 13.** The minimax loss in Equation (9), which formulates the IRM penalty as a loss difference  $[\hat{\mathcal{R}}^e(\mathbf{v}_S) - \hat{\mathcal{R}}^e(\mathbf{v}_S^e)]$ , has notable discontinuities between different parameters with different footprints  $S$ :

$$\begin{aligned} \hat{\mathcal{L}}(\mathbf{v}_S) &:= \sum_{e \in \mathcal{E}} \hat{\mathcal{R}}^e(\mathbf{v}_S) + \rho \sum_{e \in \mathcal{E}} \max_{\mathbf{v}_S^e \in \text{Sp}(S)} [\hat{\mathcal{R}}^e(\mathbf{v}_S) - \hat{\mathcal{R}}^e(\mathbf{v}_S^e)] \\ &= (1 + \rho) \sum_{e \in \mathcal{E}} \hat{\mathcal{R}}^e(\mathbf{v}_S) - \rho \sum_{e \in \mathcal{E}} \min_{\mathbf{v}_S^e \in \text{Sp}(S)} \hat{\mathcal{R}}^e(\mathbf{v}_S^e) \\ &= (1 + \rho) \sum_{e \in \mathcal{E}} \hat{\mathcal{R}}^e(\mathbf{v}_S) - \rho M_S. \end{aligned} \quad (135)$$

This presents challenges in applying existing results in linear regression with restricted parameter error, such as by LASSO (Negahban et al., 2009; Banerjee et al., 2015; Wainwright, 2019), or especially IHT (Jain et al., 2014). Instead, we directly analyze the IRMv1 penalty.

□

## D GENERALIZED LINEAR MODEL EXTENSION

We start by restating the original data generation model:

$$\begin{aligned} y &= \gamma^\top \mathbf{x}_{\text{inv}} + \epsilon_{\text{inv}}, \\ \mathbf{x}_s^e &= y \zeta_s + \alpha^e \odot \epsilon_s, \\ \mathbf{x}_r &= \zeta_r \odot \epsilon_r. \end{aligned} \quad (136)$$

GLMs are based on exponential family distributions (Brown, 1986; Barndorff-Neils, 2014; Banerjee et al., 2015), where we assume the conditional distribution of a response  $y_i$  conditioned on covariates  $\mathbf{x}_i^e$  is an exponential density function:

$$P(y_i | \mathbf{x}_i^e, \beta_{\text{inv}}^*) = \exp\{y_i \langle \mathbf{x}_i^e, \beta_{\text{inv}}^* \rangle - \varphi(\langle \mathbf{x}_i^e, \beta_{\text{inv}}^* \rangle)\} = \exp\{y_i \langle \mathbf{x}_{\text{inv},i}, \gamma \rangle - \varphi(\langle \mathbf{x}_{\text{inv},i}, \gamma \rangle)\}, \quad (137)$$

for log-partition function  $\varphi(\langle \mathbf{x}_i^e, \beta_{\text{inv}}^* \rangle) = \log\left(\int_{y_i} \exp\{y_i \langle \mathbf{x}_i^e, \beta_{\text{inv}}^* \rangle\} dy_i\right)$ . For simplicity of notation, we can represent the parameter  $\eta_i = \langle \mathbf{x}_{\text{inv},i}, \gamma \rangle$ . Then the new environmental risk is the negative log-likelihood for the conditional pdf. If we use  $D_n = \bigcup_{e \in \mathcal{E}} \{(\mathbf{x}_i^e, y_i)\}_{i=1}^{n_e}$  to be the entire dataset across different parameters,

$$\begin{aligned} \mathcal{R}^e(\beta, D_n) &= -\frac{1}{n} \sum_{e \in \mathcal{E}} \sum_{i=1}^{n_e} y_i \mathbf{x}_i^e + \frac{1}{n} \sum_{e \in \mathcal{E}} \sum_{i=1}^{n_e} \mathbf{x}_i^e \frac{\partial \varphi(\langle \beta_{\text{inv}}^*, \mathbf{x}_i^e \rangle)}{\partial \eta_i} \\ &= \frac{1}{n} \sum_{e \in \mathcal{E}} \sum_{i=1}^{n_e} \mathbf{x}_i^e (E[y_i | \mathbf{x}_i^e] - y_i) \\ &= \frac{1}{n} X^\top (E[y_i | \mathbf{x}_i^e] - y_i). \end{aligned} \quad (138)$$

We present the conditional Bernoulli distribution example (Banerjee et al., 2015; Dunn and Smyth, 2018). Using parameter  $p_i$  for the conditional mean,

$$\begin{aligned} P(y_i, | p_i) &= p_i^{y_i} (1 - p_i)^{(1-y_i)} \\ &= \exp\left(y_i \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i)\right). \end{aligned} \quad (139)$$

Then the parameter  $\eta_i = \langle \mathbf{x}_{\text{inv},i}, \gamma \rangle = \log\left(\frac{p_i}{1-p_i}\right)$ . We then end up with logistic regression, where

$$p_i = \frac{\exp(\langle \mathbf{x}_{\text{inv},i}, \gamma \rangle)}{1 + \exp(\langle \mathbf{x}_{\text{inv},i}, \gamma \rangle)}. \quad (140)$$

In this case the link function is  $\log(1 - p_i)$ . We emphasize in this setting that this only depends on the invariant features  $\mathbf{x}_{\text{inv}}^e$  and not those of the spurious.

The corresponding IRM penalty term is then

$$\mathcal{J}(\beta) = \sum_{e \in \mathcal{E}} \max_{\beta_S^e \in \text{Sp}(S)} [\mathcal{R}^e(\beta_S) - \mathcal{R}^e(\beta_S^e)]. \quad (141)$$

By showing RSC and RSS for the loss function  $\sum_{e \in \mathcal{E}} \mathcal{R}^e(\beta) + \rho \mathcal{J}(\beta)$ , we can recover Theorem 1.

## E ALGORITHM

---

### Algorithm 2 Sparse IRM with Iterative Hard-Thresholding

---

- 1: **Input:** target nonzero features  $d_{\text{inv}} < d$ ,  $\mathcal{D} = \{\mathcal{D}^e\}_{e \in \mathcal{E}}$  and  $\mathcal{D}^e := \{(\mathbf{x}_i^e, y_i)\}_{i=1}^{n_e}$ .
  - 2: Initialize weights  $(\mathbf{v}, \Phi)$ .
  - 3: **for** training iteration  $t = 1, 2, \dots, T$  **do**
  - 4:    $\mathbf{v}^{t+1} \leftarrow \text{proj}_{d_{\text{inv}}}(\mathbf{v}^t - \eta \nabla_{\mathbf{v}} \hat{\mathcal{L}}(\mathbf{v}^t))$
  - 5:    $\Phi^{t+1} \leftarrow \Phi^t - \eta \nabla_{\Phi} \hat{\mathcal{L}}(\Phi^t)$
  - 6:    $t = t + 1$
  - 7: **end for**
- 

## F EXPERIMENT DETAILS

The hyperparameters used for the experiments in Section 5 are included below. Starred hyperparameters were evaluated via grid search. Remaining hyperparameters are kept from previous experimentation (Arjovsky et al., 2020; Zhou et al., 2022). Hyperparameters for the SparseIRM + PM method, not included in Table 4, are taken from Zhou et al. (2022).

Table 4: Hyperparameter configurations for experiments.

Dataset	2-CMNIST	10-CMNIST	MNISTCIFAR
Model	MLP390	MLP390	ResNet-18
GPUs (NVIDIA A100)	1	1	1
Epochs	1500	1500	50
Optimizer	Adam	Adam	Adam
Learning Rate	0.0004	0.001	0.001
IRMv1 Penalty Weight	10000	10000	10000
IRMv1 Anneal Start Epoch	200	200	13
Learning Rate Scheduler	Cosine	Cosine	Cosine
# Zeroed Weights (last layer)*	40	40	60
IHT starting epoch*	1200	1200	46
Updates between IHT projection*	5	5	5

### F.1 DATASETS

Correlation tuples for the construction of IRM datasets are included below.

Numbers followed by a error bar are 1 standard deviation, i.e., in  $62.44 \pm 0.96$ , 62.44 is the mean, and 0.96 is one standard deviation above and below.

Table 5: Dataset configurations for experiments.

	2-CMNIST	10-CMNIST	MNISTCIFAR
Number of Classes	2	10	2
Correlation Tuple	(0.9, 0.8, 0.1)	(0.999, 0.7, 0.1)	(0.999, 0.7, 0.1)
Noise	25%	20%	10%

## F.2 TUNING THE SPARSITY

In practice, we do not have access to  $d_{\text{inv}}$  when training a model on Sparse IRM.

Perturbation to $d_{\text{inv}}$ (%)	Train Accuracy (%)	Test Accuracy (%)	L1 Norm
-5	$59.61 \pm 0.32$	$56.98 \pm 0.27$	$5.17 \pm 1.43$
-2	$62.11 \pm 0.51$	$59.05 \pm 0.43$	$6.07 \pm 1.05$
+0	$63.39 \pm 0.55$	$60.94 \pm 0.46$	$5.79 \pm 4.05$
+2	$59.41 \pm 0.52$	$57.36 \pm 0.42$	$7.97 \pm 4.05$
+5	$60.34 \pm 0.80$	$58.03 \pm 0.63$	$6.98 \pm 3.18$

Table 6: Performance metrics across different perturbations to  $d_{\text{inv}}$ .