References

Achille, A. and Soatto, S. Emergence of invariance and disentanglement in deep representations. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–9, 2018. doi: 10.1109/ITA.2018.8503149.

Bakry, D., Barthe, F., Cattiaux, P., and Guillin, A. A simple proof of the poincaré inequality for a large class of probability measures. 2008.

Bakry, D., Gentil, I., Ledoux, M., et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

Barrett, D. G. and Dherin, B. Implicit gradient regularization. In *International Conference on Learning Representations*, 2021.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Belkin, M. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. 2021. URL https://arxiv.org/abs/2105.14368.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32), 2019.

Bobkov, S. and Ledoux, M. Poincaré's inequalities and talagrand's concentration phenomenon for the exponential distribution. *Probability Theory and Related Fields*, 107: 383–400, 1997.

Boucheron, S., Lugosi, G., and Bousquet, O. Concentration inequalities. *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003, Tübingen, Germany, August 4-16, 2003, Revised Lectures*, pp. 208–240, 2004.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Bruno, T., Peloso, M. M., and Vallarino, M. Local and non-local poincaré inequalities on lie groups. *Bulletin of the London Mathematical Society*, 54(6):2162–2173, 2022.

Chatterji, N. S., Neyshabur, B., and Sedghi, H. The intriguing role of module criticality in the generalization of deep networks. *ArXiv*, abs/1912.00528, 2019.

Chavel, I. *Eigenvalues in Riemannian geometry*. Academic press, 1984.

Cheeger, J. Differentiability of lipschitz functions on metric measure spaces. *Geometric & Functional Analysis GAFA*, 9:428–517, 1999.

Dherin, B., Munn, M., and Barrett, D. G. The geometric occam's razor implicit in deep learning. *arXiv preprint arXiv:2111.15090*, 2021.

Dherin, B., Munn, M., Rosca, M., and Barrett, D. G. Why neural networks find simple solutions: the many regularizers of geometric complexity. *arXiv preprint arXiv:2209.13083*, 2022.

Dudley, R. M. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

Evans, L. C. *Partial differential equations*, volume 19. American Mathematical Society, 2022.

Ferone, V., Nitsch, C., and Trombetti, C. A remark on optimal weighted poincaré inequalities for convex domains. *Rendiconti Lincei*, 23(4):467–475, 2012.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

Gamba, M., Englesson, E., Björkman, M., and Azizpour, H. Deep double descent via smooth interpolation. *arXiv preprint arXiv:2209.10080*, 2022.

Ghosh, A., Lyu, H., Zhang, X., and Wang, R. Implicit regularization in heavy-ball momentum accelerated stochastic gradient descent. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ZzdBhtEH9yB.

Grant, E. and Wu, Y. Predicting generalization with degrees of freedom in neural networks. In *ICML 2022 2nd AI for Science Workshop*, 2022.

Hajłasz, P. and Koskela, P. *Sobolev met poincaré*, volume 688. American Mathematical Soc., 2000.

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

Hebey, E. *Nonlinear analysis on manifolds: Sobolev spaces and inequalities: Sobolev spaces and inequalities*, volume 5. American Mathematical Soc., 2000.

Heinonen, J. et al. *Lectures on analysis on metric spaces*. Springer Science & Business Media, 2001.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Ledoux, M. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2001.

Long, P. M. and Sedghi, H. Generalization bounds for deep convolutional neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1e_FpNFDr.

Lott, J. and Villani, C. Weak curvature conditions and functional inequalities. *Journal of Functional Analysis*, 245(1):311–333, 2007.

Ma, C. and Ying, L. The sobolev regularization effect of stochastic gradient descent. 2021. URL https://arxiv.org/abs/2105.13462.

Mendelson, S. Geometric methods in the analysis of glivenko-cantelli classes. In *Computational Learning Theory: 14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001 Amsterdam, The Netherlands, July 16–19, 2001 Proceedings*, pp. 256–272. Springer, 2001.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Neyshabur, B. Implicit regularization in deep learning. *arXiv preprint arXiv:1709.01953*, 2017.

Novak, R., Bahri, Y., Abolafia, D. A., Pennington, J., and Sohl-Dickstein, J. Sensitivity and generalization in neural networks: an empirical study. In *International Conference on Learning Representations*, 2018.

Raginsky, M., Sason, I., et al. Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends® in Communications and Information Theory*, 10(1-2):1–246, 2013.

Schlichting, A. Poincaré and log–sobolev inequalities for mixtures. *Entropy*, 21(1):89, 2019.

Smith, S. L., Dherin, B., Barrett, D. G., and De, S. On the origin of implicit regularization in stochastic gradient descent. In *International Conference on Learning Representations*, 2021.

Sokolic, J., Giryes, R., Sapiro, G., and Rodrigues, M. R. D. Robust large margin deep neural networks. *IEEE Transactions on Signal Processing*, 65:4265–4280, 2017.

Tomczak-Jaegermann, N. *Banach-Mazur distances and finite-dimensional operator ideals*, volume 38. Longman Sc & Tech, 1989.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. „understanding deep learning requires rethinking generalization “, iclr 2017. *arXiv preprint arXiv:1611.03530*, 2017.

Zhang, T. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5(Oct):1225–1251, 2004.

Zhang, X., Xu, R., Yu, H., Zou, H., and Cui, P. Gradient norm regularizer seeks flat minima and improves generalization, 2023. URL https://openreview.net/forum?id=z4eslwuymzQ.

A. Comparing theoretical and geometric complexity

Proof of Proposition 3.4. For any dataset $D = \{x_i\}_{i=1}^m$ of $m \geq 1$ points $x_i \in \mathbb{R}^d$ drawn as i.i.d. samples from the continuous probability distribution μ over $\mathbb{R}^d$, the empirical geometric complexity over D is denoted by $\mathrm{GC}(f, D)$. We start by showing that

$$\mathbb{E}_{D\sim\mu^m}\left[\mathrm{GC}(f, D)\right] = \mathrm{GC}(f, \mu).$$

In fact, this follows by computation, keeping in mind that μ is a probability distribution and that the points are independently sampled. Note that,

$$\begin{aligned}
\mathbb{E}_{D\sim\mu^m}\left[\mathrm{GC}(f, D)\right] &= \mathbb{E}_{x_1,\ldots,x_m\sim\mu^m}\left[\frac{1}{m}\sum_{i=1}^m \|\nabla_x f(x_i)\|_F^2\right] \\
&= \frac{1}{m}\int_{\mathbb{R}^{m\times d}} \sum_{i=1}^m \|\nabla_x f(x_i)\|_F^2 d\mu^m(x_1,\ldots,x_m) \\
&= \frac{1}{m}\int_{\mathbb{R}^{m\times d}} \sum_{i=1}^m \|\nabla_x f(x_i)\|_F^2 u(x_1)\cdots u(x_m) dx_1\cdots dx_m \\
&= \frac{1}{m}\sum_{i=1}^m \int_{\mathbb{R}^{(m-1)\times d}} \left[\int_{\mathbb{R}^d} \|\nabla_x f(x_i)\|_F^2 u(x_i) dx_i\right] u(x_1)\cdots \widehat{u(x_i)}\cdots u(x_m) dx_1\cdots \widehat{dx_i}\cdots dx_m \\
&= \frac{1}{m}\sum_{i=1}^m \left[\int_{\mathbb{R}^d} \|\nabla_x f(x_i)\|_F^2 u(x_i) dx_i\right] \\
&= \frac{1}{m}\sum_{i=1}^m \left[\int_{\mathbb{R}^d} \|\nabla_x f(x_i)\|_F^2 d\mu(x_i)\right] \\
&= \frac{1}{m}\sum_{i=1}^m \mathrm{GC}(f, \mu) \\
&= \mathrm{GC}(f, \mu).
\end{aligned}$$

Let D and D' be two samples of size $m \geq 1$ which differ by exactly one point, say x_i in D and x_i' in D'. Then since the map f is L-Lipschitz we have

$$\mathrm{GC}(f, D) - \mathrm{GC}(f, D') = \frac{1}{m}\left(\|\nabla_x f(x_i)\|_F^2 - \|\nabla_x f(x_i')\|_F^2\right) \leq L^2/m,$$

and similarly, $\mathrm{GC}(f, D') - \mathrm{GC}(f, D) \leq L^2/m$. Thus, $|\,\mathrm{GC}(f, D) - \mathrm{GC}(f, D')| \leq L^2/m$ and by applying McDiarmid's inequality (e.g. (Mohri et al., 2018)), we have that for any $\epsilon > 0$,

$$\mathbb{P}\left[\mathrm{GC}(f, D) - \mathbb{E}_{D\sim\mu^m}[\mathrm{GC}(f, D)] \leq \epsilon\right] \geq 1 - \exp(-2m\epsilon^2/L^2). \tag{6}$$

Thus, since $\mathbb{E}_{D\sim\mu^m}[\mathrm{GC}(f, D)] = \mathrm{GC}(f, \mu)$ and setting $\delta/2 = \exp(-2m\epsilon^2/L^2)$ and substituting for ϵ in (6), we get that for any $\delta > 0$ with probability as least $1 - \delta/2$ the following holds:

$$\mathrm{GC}(f, \mu) \leq \mathrm{GC}(f, D) + L\sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

This completes the proof. □

B. Proof of Theorem 4.2

Let us restate the theorem and provide the proof:

Theorem B.1. *Given a_1, a_2 be positive reals. Let $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ be i.i.d. input-output pairs in $\mathbb{R}^d \times \{\pm 1\}$ and suppose the distribution μ of the x_i satisfies the Poincaré inequality with constant $\rho > 0$. Then, for any $\delta > 0$, with probability at least $1-\delta$, every margin $\gamma > 0$ and network $f : \mathbb{R}^d \to \mathbb{R}$ which satisfies $\mathrm{GC}(f, \mu) \le a_1$ and $|\mathbb{E}_\mu(f)| \le a_2$*

$$\mathbb{P}\left[yf(x) \le 0\right] \le \widehat{\mathcal{R}}_{S,\gamma}(f) + \frac{12\tilde{C}\sqrt{\pi}}{\gamma m} + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$

where $\widehat{\mathcal{R}}_{S,\gamma}(f) = m^{-1}\sum_i \mathbb{1}_{y_i f(x_i) \le \gamma}$ and $\tilde{C} = a_2 + \sqrt{a_1 \rho / \delta}$.

The proof follows by combining fairly standard arguments in the literature. We include the full details here for completeness.

Proof. Let $\mathcal{F}$ denote the class of differentiable maps

$$\mathcal{F} := \{f : \mathbb{R}^d \to \mathbb{R} \mid \mathrm{GC}(f, \mu) \le a_1, |\mathbb{E}_\mu[f]| \le a_2\}.$$

and let $\widetilde{\mathcal{F}} = \{z = (x, y) \mapsto yf(x) \mid f \in \mathcal{F}\}$. For any $\gamma > 0$, define

$$\widetilde{\mathcal{F}}_\gamma := \{(x, y) \mapsto \ell_\gamma(-yf(x)) \mid f \in \mathcal{F}\}.$$

Since ℓ_γ has range $[0, 1]$, it follows classic generalization bounds based on the Rademacher complexity (see, for example Theorem 3.3 in (Mohri et al., 2018)) that, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. sample S of size m, we have for all $f \in \widetilde{\mathcal{F}}_\gamma$:

$$\mathbb{E}[\ell_\gamma(-yf(x))] \le \frac{1}{n}\sum_{i=1}^{n} \ell_\gamma(-y_i f(x_i)) + 2\widehat{\mathfrak{R}}_S(\widetilde{\mathcal{F}}_\gamma) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}. \tag{7}$$

We can further simplify the term $\widehat{\mathfrak{R}}_S(\widetilde{\mathcal{F}}_\gamma)$ here. Namely, $\widehat{\mathfrak{R}}_S(\widetilde{\mathcal{F}}_\gamma) = \widehat{\mathfrak{R}}_S(\ell_\gamma \circ \widetilde{\mathcal{F}})$ and, since the ramp loss ℓ_γ is $1/\gamma$-Lipschitz, by Talagrand's lemma (e.g. see Lemma 5.7 of (Mohri et al., 2018)), the empirical Rademacher complexity of $\ell_\gamma \circ \widetilde{\mathcal{F}}$ can be bounded in terms of the empirical Rademacher complexity of the original hypothesis set $\widetilde{\mathcal{F}}$; that is,

$$\widehat{\mathfrak{R}}_S(\ell_\gamma \circ \widetilde{\mathcal{F}}_\gamma) \le \frac{1}{\gamma}\widehat{\mathfrak{R}}_S(\widetilde{\mathcal{F}}). \tag{8}$$

Since the $y_i \in \{\pm 1\}$, by computing the empirical Rademacher complexity of $\widetilde{\mathcal{F}}$ over the set S, we also have $\widehat{\mathfrak{R}}_S(\widetilde{\mathcal{F}}) = \widehat{\mathfrak{R}}_S(\mathcal{F})$. Therefore, and by recalling the definition of $\widehat{\mathcal{R}}_{S,\gamma}(f)$, (7) becomes

$$\mathbb{E}_\mu[\ell_\gamma(-yf(x))] \le \widehat{\mathcal{R}}_{S,\gamma}(f) + \frac{2}{\gamma}\widehat{\mathfrak{R}}_S(\mathcal{F}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

Focusing now on the left hand side of (7), note that by definition of the ramp loss, since $\mathbb{1}_{-yf(x)\ge 0} \le \ell_\gamma(-yf(x))$, we have

$$\mathbb{E}[\mathbb{1}_{-yf(x)\ge 0}] \le \mathbb{E}[\ell_\gamma(-yf(x))]$$

and $\mathbb{P}\left[yf(x) \le 0\right] = \mathbb{E}[\mathbb{1}_{-yf(x)\ge 0}]$. Therefore,

$$\mathbb{P}\left[yf(x) \le 0\right] \le \widehat{\mathcal{R}}_{S,\gamma}(f) + \frac{2}{\gamma}\widehat{\mathfrak{R}}_S(\mathcal{F}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

Furthermore, by definition of the ramp loss, we have that

$$\begin{aligned}\mathbb{P}\left[yf(x) \le 0\right] &= \mu(-yf(x) \ge 0) \\ &= \mathbb{E}_\mu[\mathbb{1}_{-yf(x)\ge 0}] \\ &\le \mathbb{E}_\mu[\ell_\gamma(-yf(x))].\end{aligned}$$

Therefore,

$$\mathbb{P}\left[yf(x) \leq 0\right] \leq \widehat{\mathcal{R}}_{S,\gamma}(f) + \frac{2}{\gamma}\widehat{\mathfrak{R}}_S(\mathcal{F}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}. \tag{9}$$

To complete the proof we can use a form of the Dudley entropy integral to deduce an upper bound on $\widehat{\mathfrak{R}}_S(\mathcal{F})$. The Dudley entropy integral lemma (see Lemma A.5 of (Bartlett et al., 2017)) states that

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) \leq \inf_{\alpha>0}\left(\frac{4\alpha}{\sqrt{m}} + \frac{12}{m}\int_{\alpha}^{\sqrt{m}}\sqrt{\log\mathcal{N}\left(\mathcal{F}|_S, \epsilon, \|\cdot\|_2\right)}d\epsilon\right).$$

Examining the integral term above, note that $\mathcal{F}|_S = \{f(X) \mid f \in \mathcal{F}\}$ where $X = S_X$ is the projection of the sample S onto the inputs, so $\mathcal{N}(\mathcal{F}|_{\mathcal{S}}, \epsilon, \|\cdot\|_2) = \mathcal{N}(\{f(X) \mid f \in \mathcal{F}\}, \epsilon, \|\cdot\|_2)$ and, as in Lemma 4.1 taking $\tilde{C} = \tilde{C}(a_1, a_2, \rho, \delta) := a_2 + \sqrt{a_1\rho/\delta}$, it follows that $\mathcal{N}(\{f(X) \mid f \in \mathcal{F}\}, \epsilon, \|\cdot\|_2) = 1$ for all $\epsilon \geq \tilde{C}$ since it requires only one ball of radius greater than or equal to $\tilde{C}$ to cover a ball of radius $\tilde{C}$. Thus, the integrand above is zero for any $\epsilon \geq \tilde{C}$. We can further upper bound this integral by swapping the integral limit $\sqrt{m}$ with $\tilde{C}$ since the integral of a positive function is no greater than the integral of that function over a potentially larger domain. Therefore, we get,

$$\begin{aligned}\widehat{\mathfrak{R}}_S(\mathcal{F}) &\leq \inf_{\alpha>0}\left\{\frac{4\alpha}{\sqrt{m}} + \frac{12}{m}\int_{\alpha}^{\min(\sqrt{m},\tilde{C})}\sqrt{\log\mathcal{N}(\mathcal{F}|_{\mathcal{S}}, \epsilon, \|\cdot\|_2)}d\epsilon\right\} \\ &\leq \inf_{\alpha>0}\left\{\frac{4\alpha}{\sqrt{m}} + \frac{12}{m}\int_{\alpha}^{\tilde{C}}\sqrt{\log\mathcal{N}(\mathcal{F}|_{\mathcal{S}}, \epsilon, \|\cdot\|_2)}d\epsilon\right\}\end{aligned}$$

To simplify this, let's first compute the integral term. By Lemma 4.1,

$$\begin{aligned}\int_{\alpha}^{\tilde{C}}\sqrt{\log\mathcal{N}\left(\mathcal{F}|_S, \epsilon, \|\cdot\|_2\right)}d\epsilon &\leq \int_{\alpha}^{\tilde{C}}\sqrt{\log\left(\tilde{C}/\epsilon\right)}d\epsilon \\ &= \epsilon\sqrt{\log\left(\tilde{C}/\epsilon\right)}\Bigg|_{\epsilon=\alpha}^{\epsilon=\tilde{C}} - \frac{\tilde{C}\sqrt{\pi}}{2}\operatorname{erf}\left(\sqrt{\log\left(\tilde{C}/\epsilon\right)}\right)\Bigg|_{\epsilon=\alpha}^{\epsilon=\tilde{C}}\end{aligned}$$

where erf denotes the error function

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt.$$

Evaluating the right hand side fully, we get

$$\begin{aligned}\int_{\alpha}^{\tilde{C}}\sqrt{\log\mathcal{N}\left(\mathcal{F}|_S, \epsilon, \|\cdot\|_2\right)}d\epsilon &\leq \epsilon\sqrt{\log\left(\tilde{C}/\epsilon\right)}\Bigg|_{\epsilon=\alpha}^{\epsilon=\tilde{C}} - \frac{\tilde{C}\sqrt{\pi}}{2}\operatorname{erf}\left(\sqrt{\log\left(\tilde{C}/\epsilon\right)}\right)\Bigg|_{\epsilon=\alpha}^{\epsilon=\tilde{C}} \\ &= -\alpha\sqrt{\log(\tilde{C}/\alpha)} + \frac{\tilde{C}\sqrt{\pi}}{2}\operatorname{erf}\left(\sqrt{\log\left(\tilde{C}/\alpha\right)}\right) \\ &\leq \frac{\tilde{C}\sqrt{\pi}}{2} - \alpha\sqrt{\log(\tilde{C}/\alpha)},\end{aligned}$$

where in the last inequality we simply used the fact that for any $z > 0$ we have $\operatorname{erf}(z) \leq 1$. Therefore, substituting this back into the entropy bound for $\widehat{\mathfrak{R}}_S(\mathcal{F})$ above, and bounding the $\inf$ by taking the limit α goes to zero; we get,

$$\begin{aligned}\widehat{\mathfrak{R}}_S(\mathcal{F}) &\leq \inf_{\alpha>0}\left\{\frac{4\alpha}{\sqrt{m}} + \frac{12}{m}\left(\frac{\tilde{C}\sqrt{\pi}}{2} - \alpha\sqrt{\log(\tilde{C}/\alpha)}\right)\right\} \\ &\leq \lim_{\alpha\to0}\left\{\frac{4\alpha}{\sqrt{m}} + \frac{12}{m}\left(\frac{\tilde{C}\sqrt{\pi}}{2} - \alpha\sqrt{\log(\tilde{C}/\alpha)}\right)\right\} \\ &= \frac{6\tilde{C}\sqrt{\pi}}{m}\end{aligned}$$

Note that in the inequalities above we are not finding the optimal or tightest upper bounds for $\widehat{\mathfrak{R}}_S(\mathcal{F})$ that are possible. However, given the nature of these expressions it is possible to determine bounds on how sharp these inequalities are. We simply note for the time being that, although these bounds are not sharp, they are not gross overestimates of the true infimum. Finally, substituting this bound on $\widehat{\mathfrak{R}}_S(\mathcal{F})$ into (9) we get.

$$\mathbb{P}\left[yf(x) \leq 0\right] \leq \widehat{\mathcal{R}}_{S,\gamma}(f) + \frac{12\tilde{C}\sqrt{\pi}}{\gamma m} + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

which completes the proof.

□

C. Proof of Theorem 1.1

One of the key components of our proof is the Poincaré inequality, originally stated for real-valued functions as in (Evans, 2022). Under similar assumption the Poincaré inequality naturally extends to vector valued maps. We include the proof here.

Let's now detail the proof of the main covering lemma behind Theorem 1.1. As mentioned previously, the proof follows the same logic as the idea as case $k = 1$ only here we need to be a bit more careful about multivariate norms. Note also, the final ball counting argument on the image in $\mathbb{R}^k$ incurs an additional cost resulting in an exponent k which ultimately incurs a cost of a factor $\sqrt{k}$ in our final bound; c.f. (Zhang, 2004).

Proof of Lemma 4.3. Given $f \in \mathcal{F}$, let f^i denote the component functions of f for $i \in [k]$ and define $\tilde{f} : \mathbb{R}^d \to \mathbb{R}^k$ by

$$\tilde{f} := (f^1 - \mathbb{E}_\mu[f^1], \ldots, f^k - \mathbb{E}_\mu[f^k]).$$

Thus, $\mathbb{E}_\mu[\tilde{f}] = 0 \in \mathbb{R}^k$ and $\mathrm{GC}(\tilde{f}, \mu) = \mathrm{GC}(f, \mu) \leq a_1$. Futhermore, by extending Chebyshev's inequality to this multivariate setting, we get that, for any $t \in \mathbb{R}_+$,

$$\mathbb{P}\left[\|\tilde{f}\| \leq t\right] \geq 1 - \frac{\sum_i \mathrm{Var}_\mu(\tilde{f}^i)}{t^2}.$$

Note that by the definition of $\tilde{f}$ and since μ satisfies $\mathrm{PI}(\rho)$, for each $i \in [k]$,

$$\mathrm{Var}_\mu(\tilde{f}^i) = \int |\tilde{f}^i|^2 d\mu \leq \rho \int \|\nabla \tilde{f}^i\|^2 d\mu = \rho\, \mathrm{GC}(\tilde{f}^i, \mu).$$

Furthermore, by the definition of $\mathrm{GC}(f, \mu)$ it follows that

$$\mathrm{GC}(\tilde{f}, \mu) = \int \|\nabla_x \tilde{f}\|_F^2 d\mu = \int \sum_{i,j} \left|\frac{\partial \tilde{f}^i}{\partial x_j}\right|^2 d\mu = \sum_i \mathrm{GC}(\tilde{f}^i, \mu).$$

Using this simplification of $\mathrm{GC}(\tilde{f}, \mu)$ and substituting for $\mathrm{Var}_\mu(\tilde{f}^i)$ in the application of Chebyshev's inequality above, we get

$$\mathbb{P}\left[\|\tilde{f}\| \leq t\right] \geq 1 - \frac{\rho\, \mathrm{GC}(\tilde{f}, \mu)}{t^2} \geq 1 - \frac{a_1 \rho}{t^2}.$$

As before, taking $\delta = a_1\rho/t^2$ and solving for t, we get $t = \sqrt{a_1\rho/\delta}$; thus, for any $\delta \in (0, 1)$, it follows that

$$\mathbb{P}\left[\|\tilde{f}\| \leq \sqrt{a_1\rho/\delta}\right] \geq 1 - \delta.$$

Therefore, since $\|\mathbb{E}_\mu[f]\| \leq a_2$, for any $f \in \mathcal{F}$ with high probability we can bound the image of f within a ball in $\mathbb{R}^k$; namely,

$$\mathbb{P}\left[\|f\| \leq a_2 + \sqrt{a_1\rho/\delta}\right] \geq 1 - \delta.$$

The rest of the argument follows from a standard ball counting argument in $\mathbb{R}^k$. Given $\epsilon > 0$, let $r := a_2 + \sqrt{a_1\rho/\delta}$ and take a maximal set of points $p_i \in B(r)$ such that $\mathrm{dist}(p_i, p_j) > \epsilon$ for $i \neq j$. It follows that $B_{p_i}(\epsilon/2) \cap B_{p_j}(\epsilon/2) = \emptyset$ and

$\cup_i B_{p_i}(\epsilon/2) \subset B(r(1+\epsilon/2))$. Thus, by construction and taking volumes on both sides, $\sum_i |B_{p_i}(\epsilon/2)| \leq |B(r(1+\epsilon/2))|$. Let N denote the number of points p_i and since $|B_{p_i}(\epsilon/2)| = |B(\epsilon/2)|$ for all $i \in [N]$, we get

$$N \leq \frac{|B(r(1+\epsilon/2))|}{|B(\epsilon/2)|} = r^k(1+2/\epsilon)^k.$$

Therefore, for small $\epsilon < 1$, $N \leq r^k(3/\epsilon)^k$ and thus

$$\mathcal{N}(\{f(X) \mid f \in \mathcal{F}\}, \epsilon, \|\cdot\|_2) \leq \frac{3^k}{\epsilon^k}\left(a_2 + \sqrt{\frac{a_1\rho}{\delta}}\right)^k.$$

This completes the proof. □

Using this covering lemma we can now prove our main Theorem 1.1. In fact, the argument follows the same lines as the case $k = 1$ with only slight modification to account for margin operator in the multi-class setting and the application of the Dudley entropy formula when bounding the empiricial Rademacher complexity.

Let us restate the theorem and provide the proof:

Theorem C.1. *Given a_1, a_2 be positive reals. Let $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ be i.i.d. input-output pairs in $\mathbb{R}^d \times \{1, \cdots, k\}$ and suppose the distribution μ of the x_i satisfies the Poincaré inequality with constant $\rho > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, every margin $\gamma > 0$ and network $f : \mathbb{R}^d \to \mathbb{R}^k$ which satisfies $\mathrm{GC}(f, \mu) \leq a_1$ and $\|\mathbb{E}_\mu(f)\| \leq a_2$ satisfy*

$$\mathbb{P}\left[\arg\min_j f(x)_j \neq y\right] \leq \widehat{\mathcal{R}}_{S,\gamma}(f) + \frac{36\tilde{C}\sqrt{k\pi}}{\gamma m} + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

where $\widehat{\mathcal{R}}_{S,\gamma}(f) = m^{-1}\sum_i \mathbb{1}_{y_i f(x_i) \leq \gamma}$ and $\tilde{C} = a_2 + \sqrt{a_1\rho/\delta}$.

Proof of Theorem 1.1. Let $\mathcal{F}$ denote the class of differentiable maps

$$\mathcal{F} := \{f : \mathbb{R}^d \to \mathbb{R}^k \mid \mathrm{GC}(f, \mu) \leq a_1, \|\mathbb{E}_\mu[f]\| \leq a_2\}$$

and for any $\gamma > 0$ define

$$\widetilde{\mathcal{F}}_\gamma := \{(x, y) \mapsto \ell_\gamma(-\mathcal{M}(f(x), y) \mid f \in \mathcal{F}\}$$

where $\mathcal{M}(\cdot,\cdot)$ denotes the margin operator $\mathcal{M} : \mathbb{R}^k \times \{1, \ldots, k\} \to \mathbb{R}$ defined by $\mathcal{M}(v, y) = v_y - \max_{i \neq y} v_i$ and $\ell_\gamma : \mathbb{R} \to \mathbb{R}^+$ denotes the usual ramp loss.

Similar to the proof of Theorem 4.2, since ℓ_γ has range $[0, 1]$ and it follows from classic generalization bounds based on the Rademacher complexity (e.g., Theorem 3.3 in (Mohri et al., 2018)) that, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of an i.i.d. sample S of size m, we have for all $f \in \widetilde{\mathcal{F}}_\gamma$:

$$\mathbb{E}[\ell_\gamma(-\mathcal{M}(f(x), y))] \leq \widehat{\mathcal{R}}_{S,\gamma}(f) + 2\widehat{\mathfrak{R}}_S(\widetilde{\mathcal{F}}_\gamma) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}} \tag{10}$$

where now $\widehat{\mathcal{R}}_{S,\gamma}(f) = m^{-1}\sum_i \ell_\gamma(-\mathcal{M}(f(x_i), y_i))$.

We can lower bound the left hand side of (10) (see Lemma A.4 of (Bartlett et al., 2017)) so that $\mathbb{P}[\arg\max_i f(x)_i \neq y] \leq \mathbb{E}[\ell_\gamma(-\mathcal{M}(f(x), y))]$ and, via Talagrand's lemma, we can also upper bound $\widehat{\mathfrak{R}}_S(\widetilde{\mathcal{F}}_\gamma)$ on the right hand side to get

$$\mathbb{P}\left[\arg\max_i f(x)_i \neq y\right] \leq \widehat{\mathcal{R}}_{S,\gamma}(f) + \frac{2}{\gamma}\widehat{\mathfrak{R}}_S(\mathcal{F}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

It remains to bound $\widehat{\mathfrak{R}}_S(\mathcal{F})$ which we can again accomplish through the Dudley entropy integral, as in the proof of Theorem 4.2, with only a very slight modification when using the covering number bound afforded by Lemma 4.3. Namely, taking as before $\tilde{C} = \tilde{C}(a_1, a_2, \rho, \delta) := a_2 + \sqrt{a_1\rho/\delta}$, then $\mathcal{N}\left(\mathcal{F}|_S, \epsilon, \|\cdot\|_2\right) \leq (3\tilde{C}/\epsilon)^k$. Following the same argument to evaluate the integral we can obtain a comparable bound on the empirical Rademacher complexity of $\mathcal{F}$ over S, but now paying a cost of $\sqrt{k}$; i.e.,

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) \leq \inf_{\alpha>0} \left\{ \frac{4\alpha}{\sqrt{m}} + \frac{12\sqrt{k}}{m} \left(\frac{3\tilde{C}\sqrt{\pi}}{2} - \alpha\sqrt{\log(3\tilde{C}/\alpha)} \right) \right\} \leq \frac{18\tilde{C}\sqrt{k}\sqrt{\pi}}{m}.$$

Thus, collecting terms we get

$$\mathbb{P}\left[\arg\min_j f(x)_j \neq y\right] \leq \widehat{\mathcal{R}}_{S,\gamma}(f) + \frac{36\tilde{C}\sqrt{k\pi}}{\gamma m} + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

□

D. Experiment details

We trained a ResNet18 (He et al., 2016) with SGD on CIFAR10 and CIFAR-100 with both original and random labels. During training we trained with batch size 256 for 100000 steps with learning rate 0.05. Here we plot the curves for the excess risk (test accuracy - train accuracy) and compare with the geometric complexity during training.

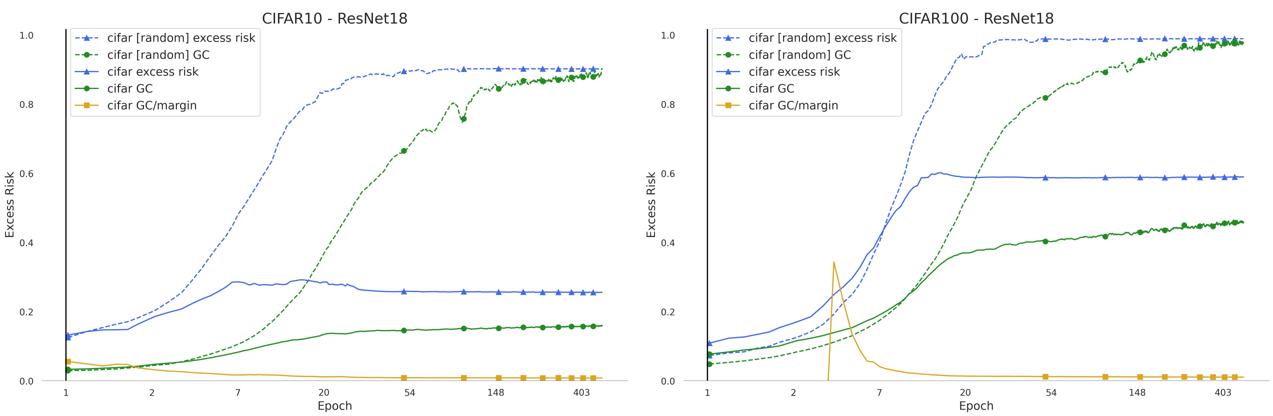

Figure 2. Analysis of ResNet-18 (He et al., 2016) trained with SGD on CIFAR-10 (left) and CIFAR-100 (right) with both original and with random labels. The triangle-marked curves plot the excess risk across training epochs (on a log scale). Circle-marked curves track the geometric complexity (GC). Note that the GC tightly correlates with excess risk in both settings. Normalizing the GC by the margin neutralizes growth across epochs.