Group Excess Risk Bound of Overparameterized Linear Regression with Constant-Stepsize SGD

Anonymous Author(s)
Affiliation
Address
email

Abstract

It has been observed that machine learning models trained using stochastic gradient descent (SGD) exhibit poor generalization to certain groups within and outside the population from which training instances are sampled. This has serious ramifications for the fairness, privacy, robustness, and out-of-distribution (OOD) generalization of machine learning. Hence, we theoretically characterize the inherent generalization of SGD-learned overparameterized linear regression to intra- and extra-population groups. We do this by proving an excess risk bound for an arbitrary group in terms of the full eigenspectra of the data covariance matrices of the group and population. We additionally provide a novel interpretation of the bound in terms of how the group and population data distributions differ and the effective dimension of SGD, as well as connect these factors to real-world challenges in practicing trustworthy machine learning. We further empirically validate the tightness of our bound on simulated data.

1 Introduction

Much recent work has sought to better understand the inductive biases of stochastic gradient descent (SGD), such as benign overfitting and implicit regularization in overparameterized settings [1, 2, 3]. However, this line of literature has overwhelmingly focused on bounding the excess risk of an SGD-learned model over the entire population $\mathcal{P}$ from which training instances are sampled, and has not investigated how SGD (e.g., its assumption that training data are IID) yields poor model generalization to intra-population groups $\mathcal{G}_{\text{intra}}$ (i.e., subsets of the population) and extra-population groups $\mathcal{G}_{\text{extra}}$ (i.e., instances that fall outside the population). We illustrate these concepts in Figure 1.

Establishing the theory behind this phenomenon is critical, as it provides provable guarantees about the trustworthiness (e.g., fairness, privacy, robustness, and out-of-distribution generalization) of SGD-learned models. As an example, let's consider an automated candidate screening system that a company trains on the data (e.g., number of years previously worked, relevant skills) of a sample from the population $\mathcal{P}$ of past job applicants [4]. In the context of fairness, many works have observed that for models trained using SGD, group-imbalanced data distributions translate to generalization disparities [5, 6, 7, 8]. Hence, the candidate screening system may generalize poorly for minoritized groups $\mathcal{G}_{\text{intra}}$ (i.e., not satisfy equal opportunity [9]), which can yield hiring discrimination. This phenomenon also has implications for privacy, as adversaries, against the desire of a job applicant, can infer whether the applicant's data were used to train the candidate screening system, based on the system's loss on the applicant [10]. When considering robustness, we may be interested in how well the candidate screening system generalizes to a target group $\mathcal{G}_{\text{intra}}$ of applicants when $\mathcal{P}$ is noisy or corrupted [11, 12]. Finally, in the context of out-of-distribution generalization, models deployed in the real world often have to deal with data distributions that differ from the training distribution [13]; for instance, the candidate screening system, if trained prior to a recession, may generalize poorly to

Submitted to 2022 Trustworthy and Socially Responsible Machine Learning (TSRML 2022). Do not distribute.

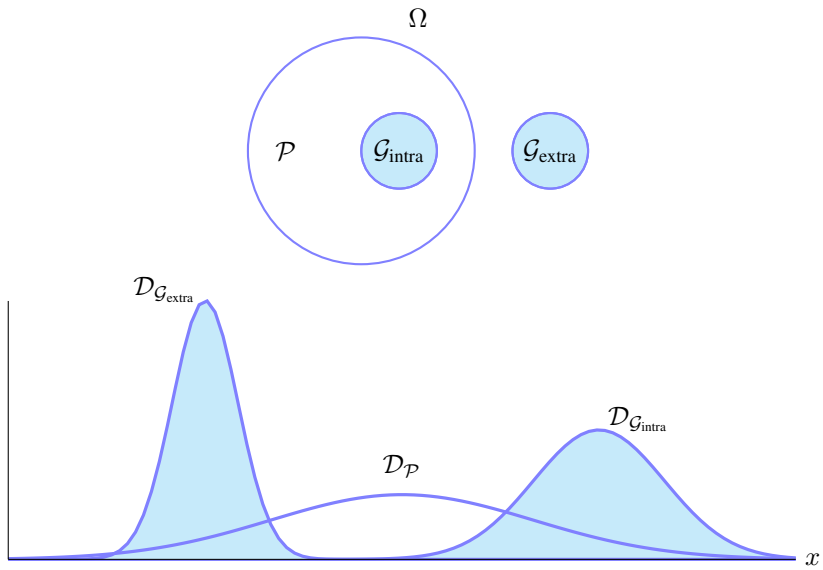

Figure 1: Euler diagram of a population $\mathcal{P}$, an intra-population group $\mathcal{G}_{\text{intra}}$, and an extra-population group $\mathcal{G}_{\text{extra}}$, as well as a visual depiction of their respective possible data distributions $\mathcal{D}_{\mathcal{P}}, \mathcal{D}_{\mathcal{G}_{\text{intra}}}, \mathcal{D}_{\mathcal{G}_{\text{extra}}}$ for a feature x (which are noticeably distinct).

stellar job applicants $\mathcal{G}_{\text{extra}}$ who were laid off during the recession for reasons beyond their control. Therefore, it is paramount to understand how SGD-learned models generalize to extra-population groups [14], especially in terms of how the properties of the data distributions for these groups differ from those of the training distribution.

Towards bolstering the theoretical foundations of trustworthy machine learning, we characterize the inherent generalization of constant-stepsize SGD (with iterate averaging) to groups within and outside the population from which training instances are sampled, for an arguably simple setting: overparameterized linear regression. We prove an excess risk bound for an arbitrary group which can be decomposed into bias and variance factors that depend on the full eigenspectra of the data covariance matrices of the group and population. We then re-express the excess risk bound in terms of: 1) how the group and population data distributions diverge, 2) how the distributions' feature variances differ, and 3) the effective dimension of SGD. We connect these three components to real-world challenges in practicing trustworthy machine learning, such as limited features and sample size disparities for minoritized groups. Finally, we empirically validate the tightness of our bound on simulated data. As a whole, we set the stage for future research to extend our results to deep neural networks and models trained with other variants of gradient descent (e.g., minibatch GD, SGD with learning rate scheduling).

2 Related work

Imbalanced learning The tendency of machine learning models to overpredict the majority class in the presence of class-imbalanced training samples [6, 15, 7, 16, 8, 17] and underperform for minoritized groups [18, 19] has been extensively empirically and theoretically studied. Some papers have theoretically investigated worst-case group generalization in the overparameterized regime [20, 21]. However, these works have not examined how SGD in particular (e.g., its assumption that training data are IID) causes poor generalization for a minoritized group, even in the arguably simple case of overparameterized linear regression. We do so in terms of the data covariance matrices of the group and population, rather than representation dimension or information, which affords greater interpretability.

Inductive biases of SGD Theoretically analyzing the inductive biases of stochastic gradient descent (e.g., implicit regularization, benign overfitting), especially in the overparameterized regime, is a

nascent area of research [2, 1, 3, 22, 23, 24] and strengthens our understanding of how deep learning works. In this paper, we make novel contributions to learning theory by analyzing constant-stepsize SGD with iterate averaging when training instances are sampled from a different distribution than the evaluation distribution. We do so by extending the analysis of [1], which only analyzes the excess risk of SGD-learned linear regression over the *entire* population from which training instances are sampled, and not over a particular group.

Trustworthy machine learning Numerous works in fair machine learning have explored the implications of generalization disparities among groups (known as equal opportunity [9]) for model-induced harms faced by minoritized groups [25, 5]. For example, in the case of automated loan approval, if white men enjoy better model generalization, their loan applications will be less likely to be incorrectly rejected compared to women and people of color. Prior research has also theoretically and empirically studied worst-case group generalization in the context of fairness without demographics and distributionally robust optimization [26, 27, 28]. In this work, we prove a group excess risk bound for overparameterized linear regression and contextualize the bound in terms of real-world challenges in practicing trustworthy machine learning.

3 Problem setup

The linear regression problem of interest is $\min_w L_{\mathcal{D}}(w)$, where $L_{\mathcal{D}}(w) = \frac{1}{2}\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[(y - \langle w, x\rangle)^2\right]$. In this equation, $x \in \mathcal{H}$ is the feature vector (we assume $\dim(\mathcal{H}) = \infty$ to model the overparameterized regime), $y \in \mathbb{R}$ is the response, $w \in \mathcal{H}$ is the weight vector to be optimized, and $\mathcal{D}$ is the arbitrary population distribution over x and y. Furthermore, suppose an arbitrary group m (within or outside the population) has the arbitrary distribution $\mathcal{D}_m$ over x and y. Now, assume the optimal parameters w_m^* for group m satisfy the first-order optimality condition $\nabla L_{\mathcal{D}_m}(w_m^*) = \mathbb{E}_{(x,y)\sim\mathcal{D}_m}\left[(y - \langle w_m^*, x\rangle)\, x\right] = 0$.

In this paper, we consider constant stepsize SGD with iterate averaging; at each iteration t, a training instance $(x_t, y_t) \sim \mathcal{D}$ is independently observed and the weight is updated as follows:

$$w_t := w_{t-1} - \gamma\left(\langle w_{t-1}, x_t\rangle - y_t\right) x_t, t = 1, \ldots, N$$

where $\gamma > 0$ is a constant stepsize, N is the number of samples observed, and the weights are initialized as $w_0 \in \mathcal{H}$. Following [1], the final output is the average of the iterates $\bar{w}_N := \frac{1}{N}\sum_{t=0}^{N-1} w_t$.

4 Main result

We now introduce relevant notation and our assumptions (which are similar to those in [1]), as well as state our main result.

Assumption 1 (Regularity conditions) For each group m, assume $H_m := \mathbb{E}_{(x,y)\sim\mathcal{D}_m}\left[xx^T\right]$ (i.e., the data covariance matrix[1] of $\mathcal{D}_m$) and $\mathbb{E}_{(x,y)\sim\mathcal{D}_m}\left[y^2\right]$ exist and are finite. Furthermore, assume that $\operatorname{tr}(H_m)$ is finite (i.e., H_m is trace-class) and H_m is symmetric positive semi-definite (PSD). Let $\{\lambda_i\}_{i=1}^{\infty}(H_m)$ be the eigenvalues of H_m sorted in decreasing order.

Additionally, denote the population data covariance matrix $H := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[xx^T\right]$ Let $\{\lambda_i\}_{i=1}^{\infty}(H)$ be the eigenvalues of H sorted in decreasing order. Furthermore, suppose the eigendecomposition of $H = \sum_i \lambda_i v_i v_i^T$; then, $H_{k:\infty} := \sum_{i>k} \lambda_i v_i v_i^T$. Similarly, the head of the identity matrix $I_{0:k} := \sum_{i=1}^{k} v_i v_i^T$.

We also define the following linear operators (which we assume to exist and be finite): $\mathcal{I} = I \otimes I$, $\mathcal{M} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[x \otimes x \otimes x \otimes x\right]$ (where $\otimes$ is the tensor product), $\mathcal{T} = H \otimes I + I \otimes H - \gamma\mathcal{M}$. All the results about these operators from Lemma 4.1 in [1] hold.

Assumption 2 (Fourth moment conditions) Assume that there exists a positive constant $\alpha > 0$, such that for any PSD matrix A, it holds that $\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[xx^T A xx^T\right] \preceq \alpha \operatorname{tr}(HA)\, H$. This assumption

[1] We refer to H_m as the data covariance matrix (following [1]), but it is in fact the raw second moment. We do **not** assume $\mathbb{E}[x] = 0$.

is satisfied for Gaussian distributions by $\alpha = 3$, and it is further implied if the distribution over $H^{-\frac{1}{2}}x$ has sub-Gaussian tails (see Lemma A.1 in [1]).

Assumption 3 (Noise conditions) Suppose that $\Sigma_{\text{noise}} := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[(y - \langle w_m^*, x\rangle)^2 xx^T\right]$ (i.e., the covariance matrix of the gradient noise at ${w_m^*}^2$) and $\sigma^2 := \left\|H^{-\frac{1}{2}}\Sigma_{\text{noise}}H^{-\frac{1}{2}}\right\|_2$ (i.e., the additive noise) exist and are finite.

Assumption 4 (Learning rate condition) Assume $\gamma \leq \frac{1}{\alpha \text{tr}(H)}$.

The excess risk of a trained model for group m quantifies how much worse the model performs for group m than the optimal model for m (i.e., the model parameterized by w_m^*) does. In Theorem 1, we present an excess risk bound of overparameterized linear regression with constant-stepsize SGD (with iterate averaging) for group m in terms of the full eigenspectra of the data covariance matrices of the group and population.

Under the assumptions above, we are ready for the statement of the main theorem:

Theorem 1 We can bound the excess risk $\mathcal{E}_m$ for group m as:

$$\mathcal{E}_m = \mathbb{E}_{\mathcal{D}}\left[L_{\mathcal{D}_m}\left(\bar{w}_N\right)\right] - L_{\mathcal{D}_m}\left(w_m^*\right) \leq 2 \cdot \text{EffectiveBias} + 2 \cdot \text{EffectiveVar},$$

where:

$$\text{EffectiveBias} \leq \begin{cases} \frac{\lambda_1(H_m)\|w_0 - w_m^*\|_2^2}{\gamma^2 N^2 \lambda_1^2(H)}, & \lambda_1(H) \geq \frac{1}{\gamma N} \\ \lambda_1\left(H_m\right)\left\|w_0 - w_m^*\right\|_2^2, & \text{otherwise} \end{cases}$$

$$\text{EffectiveVar} \leq \frac{\frac{2\alpha}{N\gamma}\left(\left\|w_0 - w_m^*\right\|_{I_{0:k^*}}^2 + N\gamma\left\|w_0 - w_m^*\right\|_{H_{k^*:\infty}}^2\right) + \sigma^2}{1 - \gamma\alpha\text{tr}\left(H\right)} \cdot \left(\underbrace{\frac{1}{N}\sum_{i\leq k^*}\frac{\lambda_i\left(H_m\right)}{\lambda_i\left(H\right)}}_{\text{head}} + N\gamma^2\underbrace{\sum_{i>k^*}\lambda_i\left(H_m\right)\lambda_i\left(H\right)}_{\text{tail}}\right),$$

where $k^* = \max\left\{k : \lambda_k(H) \geq \frac{1}{\gamma N}\right\}$. As the bound suggests, to obtain a vanishing bound, we need:

1. A sufficiently large sample from the population: $\lambda_1(H) \geq \frac{1}{\gamma N}$
2. The head to converge in N: $\sum_{i\leq k^*}\frac{\lambda_i(H_m)}{\lambda_i(H)} = o(N)$
3. The tail to converge in N: $\sum_{i>k^*}\lambda_i\left(H_m\right)\lambda_i\left(H\right) = o(1/N)$

We prove Theorem 1 in Section A. In the following section, we contextualize the group excess risk bound through the lens of real-world challenges in practicing trustworthy machine learning.

4.1 Interpreting the group excess risk bound

In interpreting the bound for $\mathcal{E}_m$, we consider the case where the eigenspectrum of H rapidly decays and thus focus on the head of the bound.[3] We first re-express the head in terms of the variance and mean of $\mathcal{D}_m$ and $\mathcal{D}$. We denote $\mu := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[x\right]$, $\Sigma := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[(x-\mu)(x-\mu)^T\right]$, $\mu_m := \mathbb{E}_{(x,y)\sim\mathcal{D}_m}\left[x\right]$, and $\Sigma_m := \mathbb{E}_{(x,y)\sim\mathcal{D}_m}\left[(x-\mu_m)(x-\mu_m)^T\right]$. Without loss of generality,

[2] Recall that w_m^* is an optimum of $L_{\mathcal{D}_m}(w)$.

[3] We leave rigorously analyzing the tail of the bound as future work. If the eigenspectrum of H doesn't decay rapidly (i.e., there exist many high-variance features), then the variance error of the group excess risk will be higher.

we assume $\mu = 0$. Then:

$$\begin{aligned}
\sum_{i \leq k^*} \frac{\lambda_i(H_m)}{\lambda_i(H)} &= \sum_{i \leq k^*} \frac{\lambda_i(\Sigma_m) + \lambda_i\left(\mu_m \mu_m^T\right)}{\lambda_i(\Sigma) + 0} \\
&= \frac{\lambda_1(\Sigma_m) + \|\mu_m\|_2^2}{\lambda_1(\Sigma)} + \sum_{2 \leq i \leq k^*} \frac{\lambda_i(\Sigma_m)}{\lambda_i(\Sigma)} \\
&= 2 \underbrace{\left[\mathrm{KL}(p_1, q_1) + \sum_{2 \leq i \leq k^*} \mathrm{KL}(p_i, q_i) \right]}_{\text{distributional difference}} + \underbrace{\sum_{i \leq k^*} \log \frac{\lambda_i(\Sigma_m)}{\lambda_i(\Sigma)}}_{\text{relative feature variance}} + \underbrace{k^*}_{\text{effective dimension}},
\end{aligned}$$

where $p_1 = \mathcal{N}\left(\|\mu_m\|_2, \lambda_1(\Sigma_m)\right)$, $p_i = \mathcal{N}\left(0, \lambda_i(\Sigma_m)\right)$, and $q_i = \mathcal{N}\left(0, \lambda_i(\Sigma)\right)$ [29]. This result shows that the excess risk for group m can be minimized (and thus generalization to group m can be improved) by:

1. Making the distributional difference between group m and the population smaller. This result corroborates findings in the fairness literature that *randomly* oversampling or increasing training data from minoritized groups (thereby boosting the representation of group m in the population) improves worst-case group generalization [19].
2. Minimizing the variance of feature values in group m relative to the variance of feature values in the population. High relative feature variance can occur when group m has sparse or noisy data, which poses a challenge in the real world because minoritized groups are often sidelined in data collection [30] and data are only partially observed [31]. This finding is also consistent with the literature on SGD's implicit bias to rely less on high-variance features [8].
3. Reducing the effective dimension k^* of SGD. Recall that $k^* = \max\left\{k : \lambda_k(H) \geq \frac{1}{\gamma N}\right\}$ and $\gamma \leq \frac{1}{\alpha \mathrm{tr}(H)} = \frac{1}{\alpha \sum_i \lambda_i(H)}$; therefore, k^* can be reduced by: 1) decreasing the variance of feature values in the population $\sum_i \lambda_i(H)$ and 2) increasing the number of training samples N. This is consistent with intuition, as 1) SGD implicitly relies less on high-variance features [8] and 2) increasing the number of *randomly*-sampled training instances can improve the representation of minoritized groups in the training data.

While, theoretically, it seems that increasing the representation of minoritized groups in the training data and better including them in data collection improves generalization to such groups, it is important to not engage in predatory inclusion and exploitative data collection practices[4]. We further emphasize that simply increasing the sheer number of samples, especially without analyzing the *randomness* or validity of sampling strategies, does not imply increasing the representation of minoritized groups in the training data [32]. Overall, we believe that one of the first steps of socially conscientious data work is to consider how data collection practices reinforce and contribute to the power relations and complex social inequality experienced by minoritized groups [33].

5 Empirical results

To investigate the tightness of our group excess risk bound, we empirically examine how well our bound aligns with the real group excess risk in a simulated setting, wherein we have control over $\mathcal{D}$ and $\mathcal{D}_m$. In particular, we assume $\mathcal{D} := p\mathcal{D}_m + (1-p)\mathcal{D}_{\text{rest}}$ is a mixture distribution that interpolates $\mathcal{D}_m := \mathcal{N}(\mu_m, \Sigma_m)$ and $\mathcal{D}_{\text{rest}} := \mathcal{N}(\mu_{\text{rest}}, \Sigma_{\text{rest}})$ for $p \in [0, 1]$. If $p << 0.5$, m could be considered a minoritized group, and the excess risk for group m would have implications for fairness (Section 1). If $p >> 0.5$, $\mathcal{D}_{\text{rest}}$ could be viewed as noise, so a model's excess risk for group m would offer insight into the robustness of the model (Section 1). $p = 0$ models a privacy-risk or OOD setting, as $\mathcal{D}_m$ would be an extra-population group (Section 1). In our experiments, we compare the group excess risk and our bound thereof for various values of $p \in [0, 1]$.

Inspired by [1] (Section 6), we consider two over-parameterized linear regression problem instances with different rates of eigenspectrum decay for H that satisfy our assumptions from Section 4:

[4] `https://slideslive.com/38955136/beyond-the-fairness-rhetoric-in-ml`

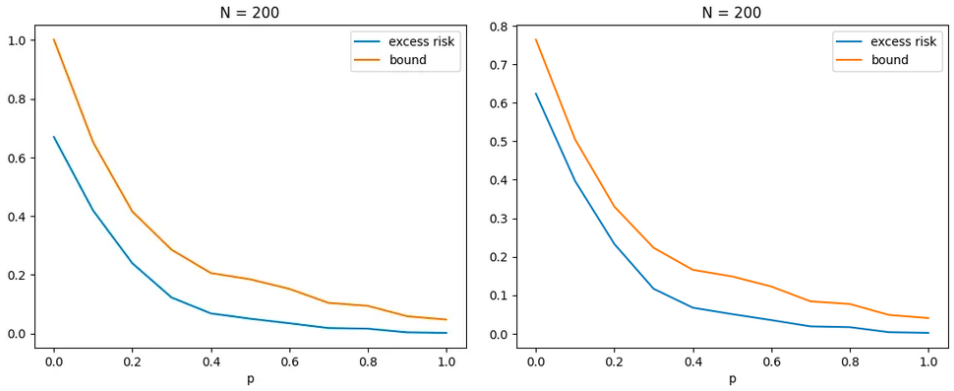

Figure 2: True group excess risk and our bound thereof for $p \in \{0.0, 0.1, 0.2, \ldots, 0.9, 1.0\}$ for problem instances 1 (left) and 2 (right). Each data point in the plots is averaged over 10 independent runs, and the true group excess risk data points are approximated over 10^5 samples from $\mathcal{D}_m$.

1. $\mu_{\text{rest}} := \mathbf{0}, \mu_m[i] = \lambda_i(\Sigma_m) = \lambda_i(\Sigma_{\text{rest}}) := (i+1)^{-1}\log(i+1)^{-2}$
2. $\mu_{\text{rest}} := \mathbf{0}, \mu_m[i] = \lambda_i(\Sigma_m) = \lambda_i(\Sigma_{\text{rest}}) := i^{-2}$

For both problem instances, $w_m^*[i] = \frac{i^{-1}}{2}$ and $w_{\text{rest}}^*[i] = i^{-1}$ (where w_{rest}^* are the optimal parameters relating $(x, y) \sim \mathcal{D}_{\text{rest}}$). Additionally, $\sigma = 0.1$. We choose sufficiently different μ_m, μ_{rest} and w_m^*, w_{rest}^* and smaller σ than [1] to enlarge the total variation distance between $\mathcal{D}_m$ and $\mathcal{D}_{\text{rest}}$, towards stress-testing our group excess risk bound. Otherwise, we follow the same experimental settings as [1]: $N = 200$, $d = 2000 >> N$ (to simulate overparametrization), $\alpha = 6$, and $\gamma = \frac{1}{\alpha tr(H)}$.

Our results are displayed in Figure 2. In both plots, the group excess risk decreases (i.e., generalization improves) when p increases, as $\mathcal{D}_m$ is sampled at a higher rate during training. This finding corroborates our commentary about fairness, privacy, robustness, and OOD generalization. Furthermore, the plots demonstrate that our group excess risk bound closely captures the true excess risk across various $p \in [0, 1]$, suggesting that our bound is tight (empirically).

6 Discussion and conclusion

In this paper, we characterize the inherent generalization of overparameterized linear regression with constant-stepsize SGD (with iterate averaging) to groups within and outside the population from which training instances are sampled. We do so by proving the excess risk bound for an arbitrary group in terms of the full eigenspectra of the data covariance matrices of the group and population. We additionally present a novel interpretation of the group excess risk bound through the lens of real-world challenges in practicing trustworthy machine learning. Finally, we empirically validate the tightness of our bound on simulated data.

This paper offers numerous promising future directions for research. We encourage proving a lower bound on the group excess risk to determine if our upper bound is tight (theoretically). We also suggest proving group excess risk bounds for tail averaging and last-iterate SGD with learning rate decay [1, 2]. It would further be interesting to extend this work to prove group excess risk bounds for logistic regression, 2-layer neural networks, and 1-layer graph convolutional networks [34].

References

[1] Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham M. Kakade. Benign Overfitting of Constant-Stepsize SGD for Linear Regression. *ArXiv*, abs/2103.12692, 2021.

[2] Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham M. Kakade. Last Iterate Risk Bounds of SGD with Decaying Stepsize for Overparameterized Linear Regression. *ArXiv*, abs/2110.06198, 2021.

[3] Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, Dean P. Foster, and Sham M. Kakade. The Benefits of Implicit Regularization from SGD in Least Squares Problems. *ArXiv*, abs/2108.04552, 2021.

[4] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 666–677, New York, NY, USA, 2021. Association for Computing Machinery.

[5] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[6] Marcus A. Maloof. Learning When Data Sets are Imbalanced and When Costs are Unequal and Unknown. In *ICML-2003 Workshop on Learning from Imbalanced Data Sets II*, 2003.

[7] Thomas Oommen, Laurie G. Baise, and Richard M. Vogel. Sampling Bias and Class Imbalance in Maximum-likelihood Logistic Regression. *Mathematical Geosciences*, 43:99–120, 2011.

[8] Klas Leino, Matt Fredrikson, Emily Black, Shayak Sen, and Anupam Datta. Feature-Wise Bias Amplification. *ArXiv*, abs/1812.08999, 2019.

[9] Sahil Verma and Julia Rubin. Fairness Definitions Explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, page 1–7, New York, NY, USA, 2018. Association for Computing Machinery.

[10] R. Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks Against Machine Learning Models. *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017.

[11] Baharan Mirzasoleiman, Kaidi Cao, and Jure Leskovec. Coresets for Robust Training of Neural Networks against Noisy Labels. *ArXiv*, abs/2011.07451, 2020.

[12] Scott Pesme and Nicolas Flammarion. Online Robust Regression via SGD on the l1 loss. *ArXiv*, abs/2007.00399, 2020.

[13] Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Out-of-Distribution Generalization in Kernel Regression. In *NeurIPS*, 2021.

[14] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. An online learning approach to interpolation and extrapolation in domain generalization. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 2641–2657. PMLR, 28–30 Mar 2022.

[15] Haibo He and Edwardo A. Garcia. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[16] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks : the official journal of the International Neural Network Society*, 106:249–259, 2018.

[17] Martín Arjovsky, Kamalika Chaudhuri, and David Lopez-Paz. Throwing away data improves worst-class error in imbalanced classification. *ArXiv*, abs/2205.11672, 2022.

[18] Aditya Krishna Menon, Ankit Singh Rawat, and Sanjiv Kumar. Overparameterisation and worst-case generalisation: friend or foe? In *International Conference on Learning Representations*, 2021.

[19] Badr Youbi Idrissi, Martín Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *CLeaR*, 2022.

[20] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An Investigation of Why Overparameterization Exacerbates Spurious Correlations. *ArXiv*, abs/2005.04345, 2020.

[21] Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis. Label-Imbalanced and Group-Sensitive Classification under Overparameterization. In *NeurIPS*, 2021.

[22] Daniel Soudry, Elad Hoffer, Suriya Gunasekar, and Nathan Srebro. The Implicit Bias of Gradient Descent on Separable Data. *ArXiv*, abs/1710.10345, 2018.

[23] Stéphane d'Ascoli, Marylou Gabrié, Levent Sagun, and Giulio Biroli. On the interplay between data structure and loss function in classification problems. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8506–8517. Curran Associates, Inc., 2021.

[24] Peiyan Li, Xu Liang, and Haochen Song. A Survey on Implicit Bias of Gradient Descent. In *2022 14th International Conference on Computer Research and Development (ICCRD)*, pages 108–114, 2022.

[25] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning. *ArXiv*, abs/1610.02413, 2016.

[26] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. *ArXiv*, abs/1911.08731, 2019.

[27] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. Fairness without Demographics through Adversarially Reweighted Learning. *ArXiv*, abs/2006.13114, 2020.

[28] Natalia L Martinez, Martin A Bertran, Afroditi Papadaki, Miguel Rodrigues, and Guillermo Sapiro. Blind Pareto Fairness and Subgroup Robustness. 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 7492–7501. PMLR, 18–24 Jul 2021.

[29] Joram Soch, The Book of Statistical Proofs, Thomas J. Faulkenberry, Kenneth Petrykowski, and Carsten Allefeld. StatProofBook/StatProofBook.github.io: StatProofBook 2020, December 2020.

[30] Solon Barocas and Andrew D. Selbst. Big Data's Disparate Impact. *California Law Review*, 104:671, 2016.

[31] Nicolò Cesa-Bianchi, Shai Shalev-Shwartz, and Ohad Shamir. Efficient learning with partially observed attributes. In *ICML*, 2010.

[32] Xiao-Li Meng. Statistical paradises and paradoxes in big data (i) law of large populations, big data paradox, and the 2016 us presidential election. *The Annals of Applied Statistics*, 12(2):685–726, 2018.

[33] Patrícia Hill Collins. Black feminist thought in the matrix of domination from. 2017.

[34] Jiaqi Ma, Junwei Deng, and Qiaozhu Mei. Subgroup Generalization and Fairness of Graph Neural Networks. In *NeurIPS*, 2021.

[35] Jean B. Lasserre. A trace inequality for matrix product. *IEEE Trans. Autom. Control.*, 40:1500–1501, 1995.

A Proof of the main result

A.1 Bias-variance decomposition

Towards proving our main result, we first decompose the excess risk for group m into bias and variance errors. We define the centered SGD iterate $\eta_t := w_t - w_m^*$; similarly, $\bar{\eta}_N := \frac{1}{N}\sum_{t=0}^{N-1}\eta_t$.

[1] (Equation 4.2) shows the bias-variance decomposition of the iterate:

$$\eta_t = \eta_t^{\text{bias}} + \eta_t^{\text{var}}$$
$$\eta_t^{\text{bias}} = \left(I - \gamma x_t x_t^T\right) \eta_{t-1}^{\text{bias}}, \eta_0^{\text{bias}} = \eta_0$$
$$\eta_t^{\text{var}} = \left(I - \gamma x_t x_t^T\right) \eta_{t-1}^{\text{var}} + \gamma \xi_t x_t, \eta_0^{\text{var}} = 0,$$

where $\xi_t := y_t - \langle w_m^*, x_t \rangle$ is the inherent noise. [1] (Equations 4.4 and 4.5) then proves the recursive forms:

$$B_t := \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\eta_t^{\text{bias}} \otimes \eta_t^{\text{bias}}\right] = (\mathcal{I} - \gamma\mathcal{T}) \circ B_{t-1}, B_0 = \eta_0 \otimes \eta_0$$
$$C_t := \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\eta_t^{\text{var}} \otimes \eta_t^{\text{var}}\right] = (\mathcal{I} - \gamma\mathcal{T}) \circ C_{t-1}, C_0 = 0$$

We also define $\bar{\eta}_N^{\text{bias}} := \frac{1}{N}\sum_{t=0}^{N-1} \eta_t^{\text{bias}}$ and $\bar{\eta}_N^{\text{var}} := \frac{1}{N}\sum_{t=0}^{N-1} \eta_t^{\text{var}}$, and see that $\bar{\eta}_t = \bar{\eta}_t^{\text{bias}} + \bar{\eta}_t^{\text{var}}$. Without loss of generality, we present a bias-variance decomposition of the excess risk for group m:

$$\begin{aligned}
\mathcal{E}_m &= \mathbb{E}_{\mathcal{D}} \left[L_{\mathcal{D}_m}(\bar{w}_N) - L_{\mathcal{D}_m}(w_m^*)\right] \\
&= \mathbb{E}_{\mathcal{D}} \left[\frac{1}{2}\mathbb{E}_{(x,y)\sim\mathcal{D}_m} \left(y - \bar{w}_N^T x\right)^2 - \frac{1}{2}\mathbb{E}_{(x,y)\sim\mathcal{D}_m} \left(y - {w_m^*}^T x\right)^2\right] \\
&= \frac{1}{2}\mathbb{E}_{\mathcal{D}} \left[\mathbb{E}_{(x,y)\sim\mathcal{D}_m} \left(\left(w_m^* - \bar{w}_N\right)^T x + \xi\right)^2 - \mathbb{E}_{(x,y)\sim\mathcal{D}_m} \xi^2\right] \\
&= \frac{1}{2}\mathbb{E}_{\mathcal{D}} \left[\mathbb{E}_{(x,y)\sim\mathcal{D}_m} \left[\left(\bar{w}_N - w_m^*\right)^T x x^T \left(\bar{w}_N - w_m^*\right) - 2\left(\bar{w}_N - w_m^*\right)^T (\xi x)\right]\right] \\
&= \frac{1}{2} \left\langle H_m, \mathbb{E}_{\mathcal{D}} \left[\bar{\eta}_N \otimes \bar{\eta}_N\right]\right\rangle
\end{aligned}$$

Then, by Lemma B.2 from [1] and Young's inequality, $\mathbb{E}_{\mathcal{D}} \left[L_{\mathcal{D}_m}(\bar{w}_N) - L_{\mathcal{D}_m}(w_m^*)\right] \leq \left(\sqrt{\text{bias}} + \sqrt{\text{var}}\right)^2 \leq 2 \cdot \text{bias} + 2 \cdot \text{var}$, where $\text{bias} := \frac{1}{2}\left\langle H_m, \mathbb{E}_{\mathcal{D}} \left[\bar{\eta}_N^{\text{bias}} \otimes \bar{\eta}_N^{\text{bias}}\right]\right\rangle$ and $\text{var} := \frac{1}{2}\left\langle H_m, \mathbb{E}_{\mathcal{D}} \left[\bar{\eta}_N^{\text{var}} \otimes \bar{\eta}_N^{\text{var}}\right]\right\rangle$. The bias and var errors provide a decomposition of a bound for $\mathcal{E}_m$.

A.2 Bounding the bias error

Now that we have decomposed a bound for $\mathcal{E}_m$ into bias and variance errors, we will separately bound these errors in terms of the full eigenspectra of the data covariance matrices of group m and the population, and then combine these bounds to achieve the desired bound for $\mathcal{E}_m$. In Theorem 2, we focus on the bias error.

Theorem 2 We can bound the bias error as:

$$\text{bias} \leq \frac{\alpha \text{tr}\left(B_{0,N}\right)}{N\gamma\left(1 - \gamma\alpha\text{tr}\left(H\right)\right)} \left(\frac{1}{N}\sum_{i \leq k^*} \frac{\lambda_i\left(H_m\right)}{\lambda_i\left(H\right)} + N\gamma^2 \sum_{i > k^*} \lambda_i\left(H_m\right)\lambda_i\left(H\right)\right)$$
$$+ \begin{cases} \frac{\lambda_1(H_m)\|w_0 - w_m^*\|_2^2}{\gamma^2 N^2 \lambda_1^2(H)}, & \lambda_1(H) \geq \frac{1}{\gamma N} \\ \lambda_1\left(H_m\right)\|w_0 - w_m^*\|_2^2, & \text{otherwise} \end{cases}$$

Proof of Theorem 2 [1] (Lemma B.3) shows that:

$$\mathbb{E}_{\mathcal{D}} \left[\bar{\eta}_N^{\text{bias}} \otimes \bar{\eta}_N^{\text{bias}}\right] \preceq \frac{1}{N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1} \left(I - \gamma H\right)^{k-t} \mathbb{E}\left[\eta_t^{\text{bias}} \otimes \eta_t^{\text{bias}}\right] + \mathbb{E}\left[\eta_t^{\text{bias}} \otimes \eta_t^{\text{bias}}\right]\left(I - \gamma H\right)^{k-t}$$

Because H and H_m are PSD and symmetric, and H commutes with $(I-\gamma H)^{k-t}$:

$$
\begin{aligned}
\text{bias} &= \frac{1}{2}\left\langle H_m, \mathbb{E}\left[\bar{\eta}_N^{\text{bias}} \otimes \bar{\eta}_N^{\text{bias}}\right]\right\rangle \\
&\leq \frac{1}{2N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left\langle H_m, (I-\gamma H)^{k-t}\,\mathbb{E}\left[\eta_t^{\text{bias}} \otimes \eta_t^{\text{bias}}\right] + \mathbb{E}\left[\eta_t^{\text{bias}} \otimes \eta_t^{\text{bias}}\right](I-\gamma H)^{k-t}\right\rangle \\
&= \frac{1}{2N^2}\sum_{t=0}^{N-1}\sum_{k=t}^{N-1}\left\langle H_m (I-\gamma H)^{k-t} + (I-\gamma H)^{k-t} H_m, \mathbb{E}\left[\eta_t^{\text{bias}} \otimes \eta_t^{\text{bias}}\right]\right\rangle \\
&= \frac{1}{2\gamma N^2}\sum_{t=0}^{N-1}\left\langle H_m H^{-1}\left(I-(I-\gamma H)^{N-t}\right) + \left(I-(I-\gamma H)^{N-t}\right)H^{-1}H_m, B_t\right\rangle \\
&\leq \frac{1}{2\gamma N^2}\left\langle H_m H^{-1}\left(I-(I-\gamma H)^{N}\right) + \left(I-(I-\gamma H)^{N}\right)H^{-1}H_m, \sum_{t=0}^{N-1} B_t\right\rangle
\end{aligned}
$$

[1] (Lemma B.10) shows that:

$$
\sum_{t=0}^{N-1} B_t \preceq \sum_{k=0}^{N-1}(I-\gamma H)^k\left(\frac{\gamma\alpha\text{tr}\left(B_{0,N}\right)}{1-\gamma\alpha\text{tr}\left(H\right)}\cdot H + B_0\right)(I-\gamma H)^k,
$$

where $B_{0,N} = B_0 - (I-\gamma H)^N B_0 (I-\gamma H)^N$. Furthermore, [1] (Lemma B.11) shows that $\text{tr}\left(B_{0,N}\right) \leq 2\left(\|w_0 - w_m^*\|^2_{I_{0:k^*}} + N\gamma\|w_0 - w_m^*\|^2_{H_{k^*:\infty}}\right)$. Therefore:

$$
\begin{aligned}
\text{bias} \leq \frac{1}{2\gamma N^2}\sum_{k=0}^{N-1}\Big\langle & H_m H^{-1}\left(I-(I-\gamma H)^N\right) + \left(I-(I-\gamma H)^N\right)H^{-1}H_m, \\
& (I-\gamma H)^k\left(\frac{\gamma\alpha\text{tr}\left(B_{0,N}\right)}{1-\gamma\alpha\text{tr}\left(H\right)}\cdot H + B_0\right)(I-\gamma H)^k\Big\rangle \\
\leq \frac{1}{2\gamma N^2}\sum_{k=0}^{N-1}\Big\langle & (I-\gamma H)^k - (I-\gamma H)^{N+k}, \frac{\gamma\alpha\text{tr}\left(B_{0,N}\right)}{1-\gamma\alpha\text{tr}\left(H\right)}\cdot HH_mH^{-1} + B_0H_mH^{-1} \\
& + H^{-1}H_mH\cdot\frac{\gamma\alpha\text{tr}\left(B_{0,N}\right)}{1-\gamma\alpha\text{tr}\left(H\right)} + H^{-1}H_mB_0\Big\rangle,
\end{aligned}
$$

where we use that $(I-\gamma H)^k \preceq I$. We now define the following terms:

$$
\begin{aligned}
I_1 &= \frac{\alpha\text{tr}\left(B_{0,N}\right)}{2N^2\left(1-\gamma\alpha\text{tr}\left(H\right)\right)}\sum_{k=0}^{N-1}\left\langle (I-\gamma H)^k - (I-\gamma H)^{N+k}, HH_mH^{-1}\right\rangle \\
I_2 &= \frac{1}{2\gamma N^2}\sum_{k=0}^{N-1}\left\langle (I-\gamma H)^k - (I-\gamma H)^{N+k}, B_0H_mH^{-1}\right\rangle
\end{aligned}
$$

We first focus on bounding I_1. If two matrices A and B are symmetric and PSD, then, $\text{tr}\left(AB\right) \leq \sum_i \lambda_i(A)\lambda_i(B)$, where $\{\lambda_i(A)\}_{i=1}^\infty$ and $\{\lambda_i(B)\}_{i=1}^\infty$ are the eigenvalues of A and B, respectively

[35]. Thus:

$$
\begin{aligned}
I_1 &= \frac{\alpha \mathrm{tr}\left(B_{0,N}\right)}{2\gamma N^2\left(1-\gamma\alpha\mathrm{tr}\left(H\right)\right)}\left\langle \left(I-\left(I-\gamma H\right)^N\right)^2 H^{-1}, HH_mH^{-1}\right\rangle \\
&\leq \frac{\alpha \mathrm{tr}\left(B_{0,N}\right)}{2\gamma N^2\left(1-\gamma\alpha\mathrm{tr}\left(H\right)\right)}\cdot\sum_i \frac{\lambda_i\left(H_m\right)\cdot\left(1-\left(1-\gamma\lambda_i\left(H\right)\right)^N\right)^2}{\lambda_i\left(H\right)} \\
&\leq \frac{\alpha \mathrm{tr}\left(B_{0,N}\right)}{2\gamma N^2\left(1-\gamma\alpha\mathrm{tr}\left(H\right)\right)}\cdot\sum_i \frac{\lambda_i\left(H_m\right)\cdot\min\left\{1,\gamma^2N^2\lambda_i^2\left(H\right)\right\}}{\lambda_i\left(H\right)} \\
&= \frac{\alpha \mathrm{tr}\left(B_{0,N}\right)}{2N\gamma\left(1-\gamma\alpha\mathrm{tr}\left(H\right)\right)}\left(\frac{1}{N}\sum_{i\leq k^*}\frac{\lambda_i\left(H_m\right)}{\lambda_i\left(H\right)}+N\gamma^2\sum_{i>k^*}\lambda_i\left(H_m\right)\lambda_i\left(H\right)\right),
\end{aligned}
$$

where $k^* = \max\left\{k : \lambda_k(H) \geq \frac{1}{\gamma N}\right\}$. Because $\mathrm{tr}(H)$ is finite, $\sum_{i>k^*}\lambda_i(H)$ converges. Furthermore, $\{\lambda_i(H_m)\}_{i\geq k^*}$ decreases monotonically and is bounded by $[0, \lambda_{k^*}(H_m)]$. Therefore, by Abel's Lemma, $\sum_{i>k^*}\lambda_i\left(H_m\right)\lambda_i\left(H\right)$ converges. Furthermore, we see that the bound for I_1 reduces to that in [1] when $\forall i, \lambda_i(H_m) = \lambda_i(H)$.

Now, we bound I_2. Because $B_0 = \left(w_0 - w_m^*\right)\left(w_0 - w_m^*\right)^T$, the largest and only non-zero eigenvalue of B is $\lambda_1(B) = \|w_0 - w_m^*\|_2^2$. Therefore:

$$
\begin{aligned}
I_2 &\leq \frac{1}{2\gamma^2N^2}\left\langle \left(I-\left(I-\gamma H\right)^N\right)^2 H^{-1}, B_0H_mH^{-1}\right\rangle \\
&\leq \frac{1}{2\gamma^2N^2}\cdot\frac{\lambda_1\left(H_m\right)\|w_0-w_m^*\|_2^2}{\lambda_1^2\left(H\right)}\cdot\begin{cases}1, & \lambda_1(H)\geq\frac{1}{\gamma N}\\ \gamma^2N^2\lambda_1^2\left(H\right), & \text{otherwise}\end{cases} \\
&= \begin{cases}\frac{\lambda_1(H_m)\|w_0-w_m^*\|_2^2}{2\gamma^2N^2\lambda_1^2(H)}, & \lambda_1(H)\geq\frac{1}{\gamma N}\\ \frac{1}{2}\lambda_1\left(H_m\right)\|w_0-w_m^*\|_2^2, & \text{otherwise}\end{cases}
\end{aligned}
$$

Now that we have bounds for I_1 and I_2:

$$
\begin{aligned}
\text{bias} &\leq 2I_1 + 2I_2 \\
&\leq \frac{\alpha \mathrm{tr}\left(B_{0,N}\right)}{N\gamma\left(1-\gamma\alpha\mathrm{tr}\left(H\right)\right)}\left(\frac{1}{N}\sum_{i\leq k^*}\frac{\lambda_i\left(H_m\right)}{\lambda_i\left(H\right)}+N\gamma^2\sum_{i>k^*}\lambda_i\left(H_m\right)\lambda_i\left(H\right)\right) \\
&\qquad + \begin{cases}\frac{\lambda_1(H_m)\|w_0-w_m^*\|_2^2}{\gamma^2N^2\lambda_1^2(H)}, & \lambda_1(H)\geq\frac{1}{\gamma N}\\ \lambda_1\left(H_m\right)\|w_0-w_m^*\|_2^2, & \text{otherwise}\end{cases}
\end{aligned}
$$

Thus, we have successfully bounded the bias error in terms of the full eigenspectra of the data covariance matrices of group m and the population.

A.3 Bounding the variance error

Proceeding in our journey to prove Theorem 1, we will bound the variance error.

Theorem 3 We can bound the variance error as:

$$
\text{var} \leq \frac{\sigma^2}{1-\gamma\alpha\mathrm{tr}\left(H\right)}\left(\frac{1}{N}\sum_{i\leq k^*}\frac{\lambda_i\left(H_m\right)}{\lambda_i\left(H\right)}+N\gamma^2\sum_{i>k^*}\lambda_i\left(H_m\right)\lambda_i\left(H\right)\right)
$$

Similar to for the bias error:

$$\text{var} \leq \frac{1}{2\gamma N^2} \sum_{t=0}^{N-1} \left\langle H_m H^{-1} \left(I - (I-\gamma H)^{N-t} \right) + \left(I - (I-\gamma H)^{N-t} \right) H^{-1} H_m, C_t \right\rangle$$

[1] (Lemma B.5) shows $C_t \preceq \frac{\gamma\sigma^2}{1-\gamma\alpha\text{tr}(H)} \left(I - (I-\gamma H)^t \right)$. Therefore:

$$\begin{aligned}
\text{var} &\leq \frac{\sigma^2}{2N^2 \left(1-\gamma\alpha\text{tr}\left(H\right)\right)} \sum_{t=0}^{N-1} \left\langle H_m H^{-1} \left(I - (I-\gamma H)^{N-t} \right) + \left(I - (I-\gamma H)^{N-t} \right) H^{-1} H_m, \right. \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \left. I - (I-\gamma H)^t \right\rangle \\
&= \frac{\sigma^2}{2N^2 \left(1-\gamma\alpha\text{tr}\left(H\right)\right)} \left\langle \sum_{t=0}^{N-1} \left(I - (I-\gamma H)^{N-t} \right) \left(I - (I-\gamma H)^t \right), H_m H^{-1} + H^{-1} H_m \right\rangle \\
&\leq \frac{\sigma^2}{2N \left(1-\gamma\alpha\text{tr}\left(H\right)\right)} \left\langle \left(I - (I-\gamma H)^N \right)^2, H_m H^{-1} + H^{-1} H_m \right\rangle \\
&\leq \frac{\sigma^2}{1-\gamma\alpha\text{tr}\left(H\right)} \left(\frac{1}{N} \sum_{i \leq k^*} \frac{\lambda_i\left(H_m\right)}{\lambda_i\left(H\right)} + N\gamma^2 \sum_{i > k^*} \lambda_i\left(H_m\right) \lambda_i\left(H\right) \right)
\end{aligned}$$

Having appropriately bounded both the bias and variance errors, we have proved Theorem 1.