OPT2025: 17th Annual Workshop on Optimization for Machine Learning

Optimal Implicit Bias in Linear Regression

K Nithin Varma NKANUMUR@CALTECH.EDU
Babak Hassibi HASSIBI@CALTECH.EDU
California Institute of Technology, USA

Abstract

Most modern learning problems are over-parameterized, where the number of learnable parameters is much greater than the number of training data points. In this over-parameterized regime, the training loss typically has infinitely many global optima that completely interpolate the data with varying generalization performance. The particular global optimum we converge to depends on the implicit bias of the optimization algorithm. The question we address in this paper is, "What is the implicit bias that leads to the best generalization performance?". To find the optimal implicit bias, we provide a precise asymptotic analysis of the generalization performance of interpolators obtained from minimizing convex functions/potentials for over-parameterized linear regression with non-isotropic Gaussian data. In particular, we obtain a tight lower bound on the best generalization error possible among this class of interpolators in terms of the over-parameterization ratio, the variance of the noise in the labels, the eigenspectrum of the data covariance, and the underlying distribution of the parameter to be estimated. Finally, we find the optimal convex implicit bias that achieves this lower bound under certain sufficient conditions involving the log-concavity of the distribution of a Gaussian convolved with the prior of the true underlying parameter.

1. Introduction

Classical statistical learning theory primarily focuses on problems in data-rich regimes, where the number of data points is significantly greater than the number of unknown parameters to be learned/estimated. In contrast, most modern learning problems like deep learning are typically highly overparameterized, i.e., the problem dimension n is much greater than the number of training data points m. Due to this over-parameterization, these models possess the capacity to completely fit any set of training data (even possibly random) [59]. Despite this overfitting, these models surprisingly generalize well on unseen data, and this so-called *double descent* phenomenon was observed, for example, in [10, 39]. In this so-called *interpolating* regime [37] of over-parameterized models, there generally exist (infinitely) many global optima of weights that interpolate the data with varying generalization properties. The particular convergent global optima are dependent on the implicit bias of the optimizer used in training. Due to the empirical success of stochastic gradient descent (SGD) and its variants, the implicit bias solution of SGD, which is the minimum ℓ_2 interpolator, has been a subject of great interest for linear models [7, 39, 43]. A generalization of SGD is stochastic mirror descent (SMD) [41], which generalizes the implicit bias to an arbitrary convex potential [25, 26, 50], as shown below

$$\hat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}\in\mathbb{R}^n}{\arg\min}\, \Psi(\boldsymbol{\beta}) \quad \text{s.t} \quad y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} \text{ for } i \in [m], \tag{1}$$

where the minimum ℓ_2-norm solution is a special case. The recent works of [31, 40, 57, 60] provide a statistical analysis of low-norm interpolators in the highly overparameterized regime. We provide a more in-depth discussion on related works in the Appendix.

Contributions: In this work, **(i)** we characterize the precise asymptotic limit of the generalization error of interpolators obtained through (1) for separable Ψ; see Theorem 1 in the proportional regime similar to [27, 39] where the number of data points scales at a proportional rate to problem dimension. We show that the value of this limit is obtained as a solution of a system of two nonlinear equations with two unknowns. **(ii)** We establish a tight lower bound on the achievable generalization error for a broad class of interpolators obtained through (1), and computing this lower bound involves solving a scalar non-linear equation; see Theorem 2. We also provide a slightly weaker but simplified version of this lower bound for isotropic data, and we show that this bound is indeed tight when the true underlying parameter has a Gaussian density (Corollary 3). **(iii)** Under certain conditions, we provide a construction of the optimal convex potential, whose asymptotic limit of the generalization error matches the lower bounds, indeed confirming that the lower bounds are tight (Theorem 4). We also describe the special case when SGD or the minimum ℓ_2-norm interpolator is optimal (Corollary 6).

2. Problem Setup

We consider the problem of linear regression in the over-parameterized regime. We model the data $\{\boldsymbol{x}_i, y_i\}_{i=1}^m$ be generated from an additive noisy linear model:

$$y_i = \boldsymbol{x_i}^T \boldsymbol{\beta}^* + z_i \quad \text{for} \quad i = 1, \cdots, m \tag{2}$$

where labels/outputs $y_i \in \mathbb{R}$ are a linear function of the covariate vectors $\boldsymbol{x}_i \in \mathbb{R}^n$ perturbed by unknown noise $z_i \in \mathbb{R}$. Here, $\boldsymbol{\beta}^* \in \mathbb{R}^n$ is the true unknown model/weights to be estimated through learning. The goal of the learner is to come up with an estimate $\hat{\boldsymbol{\beta}}$ that minimizes the population risk/generalization error

$$r(\hat{\boldsymbol{\beta}}) := \mathbb{E}_{\tilde{y},\tilde{\boldsymbol{x}}}[(\tilde{y} - \tilde{\boldsymbol{x}}^T \hat{\boldsymbol{\beta}})^2] \tag{3}$$

or equivalently the excess risk $r(\hat{\boldsymbol{\beta}}) - r(\boldsymbol{\beta}^*)$. Here, the expectation is over an independent realization of the $(\tilde{y}, \tilde{\boldsymbol{x}})$, which are related by (2). In the overparameterized regime $n > m$, the minimizer obtained from linear regression on the dataset is not unique, and the goal of this work is to study the dependence of the generalization error on the structure of the underlying signal $\boldsymbol{\beta}^*$, covariates, and the choice of global optima from the interpolating subspace. To this end, we assume the following in our analysis.

Assumption 1 *(High-dimensional asymptotics) Throughout this paper, we consider the asymptotic proportional limit where both $m, n \to \infty$ at a fixed ratio $\delta = m/n$, where $0 < \delta < 1$.*

Assumption 2 *(Gaussian features and noise) Given the feature covariance $\boldsymbol{\Sigma}$ the data/feature vectors $\boldsymbol{x}_i := \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{g}_i$ where $\boldsymbol{g}_i \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, \frac{1}{n}\boldsymbol{I}_n)$, $i \in [m]$ and the noise $z_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, $i \in [m]$.*

Assumption 3 *(True parameter and eigen-spectrum distribution) We consider a fixed diagonal covariance $\boldsymbol{\Sigma}$ and unknown signal parameter $\boldsymbol{\beta}^*$ such that the pair $\{\Sigma_{i,i}, \beta_i^*\}$ of the eigen-values and the coordinates of the true underlying signal parameter are sampled i.i.d. from $\{\Lambda^2, B\}$ with distribution $P_{\Lambda,B}$. Additionally, we assume that the marginal distribution P_B has a finite, non-zero second moment and Λ is a strictly positive, with bounded support.*

Under these assumptions, the generalization error in (3) simplifies to the following limit.

$$r(\hat{\boldsymbol{\beta}}) = \lim_{n\to\infty} \frac{1}{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T \boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \sigma^2. \tag{4}$$

We denote $\boldsymbol{X} := (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m)^T \in \mathbb{R}^{m\times n}$, label vector $\boldsymbol{y} := (y_1, y_2, \cdots, y_m)^T \in \mathbb{R}^m$ and $\boldsymbol{z} := (z_1, z_2, \cdots, z_m)^T \in \mathbb{R}^m$. We assume that the data and noise satisfy the conditions in Assumption 2 and that $\boldsymbol{\Sigma}$ is fixed. Performing empirical risk minimization (ERM) on the dataset in the overparameterized regime leads to interpolation, and we denote the subspace of global optima as $\tilde{\mathcal{B}} := \{\boldsymbol{\beta} : \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}\}$. The global optimum we converge to is given by the implicit bias of the algorithm, and in this work, we consider separable convex potentials, i.e. $\Psi(\boldsymbol{\beta}) := \sum_{i=1}^n \psi_i(\beta_i)$ which satisfy $\lim_{\|\boldsymbol{\beta}\|\to\infty} \Psi(\boldsymbol{\beta}) = \infty$. Further, we consider the general case of $\psi_i(.) = \psi(., \Sigma_{i,i})$ where the learner has access to the diagonal entries of $\boldsymbol{\Sigma}$. For cases where $\boldsymbol{\Sigma}$ is unknown to the learner, one can simply take $\Psi(\boldsymbol{\beta}) := \sum_{i=1}^n \psi(\beta_i)$. Our analysis is general enough to cover both these cases. Under these assumptions, (1) can be reformulated as

$$\hat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}\in\mathbb{R}^n}{\arg\min} \sum_{i=1}^n \psi(\beta_i, \Sigma_{i,i}) \quad \text{s.t} \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}. \tag{5}$$

This formulation includes the minimum ℓ_2 norm interpolator as a special case when $\Psi(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2$. When $\boldsymbol{\Sigma}$ is diagonal, we can also pick $\Psi(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T\boldsymbol{\Sigma}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$ which recovers the theoretical optimal interpolator shown in [39], although this is not achievable since $\boldsymbol{\beta}^*$ is unknown to the learner.

3. Main Results

We first characterize the precise asymptotic of problem (5) in terms of the solution to a system of non-linear equations (6). Next, we leverage this system of equations to drive tight lower bounds on the generalization error and provide the construction of the optimal convex potential that achieves these bounds.

3.1. System of Non-Linear Equations

For a given overparameterization ratio $0 < \delta < 1$ and noise variance σ^2, we have the following system of non-linear equations in α, u

$$\mathbb{E}\left[\frac{H}{\Lambda}\mathcal{M}'_{\psi,1}(B + \frac{\alpha}{\sqrt{\delta}\Lambda}H; \frac{\alpha}{u\sqrt{\delta}\Lambda^2})\right] = u(1-\delta) \tag{6a}$$

$$\mathbb{E}\left[\left(\frac{1}{\Lambda}\mathcal{M}'_{\psi,1}(B + \frac{\alpha}{\sqrt{\delta}\Lambda}H; \frac{\alpha}{u\sqrt{\delta}\Lambda^2})\right)^2\right] = u^2(1-\delta) - \frac{\delta\sigma^2u^2}{\alpha^2} \tag{6b}$$

Here, the expectation is over the random variables B, H, and Λ, where Λ, B as defined in Assumption 3 denote the distribution of the eigenspectrum and the underlying signal, respectively, and H is a standard Gaussian.

3.2. Precise Asymptotics

Theorem 1 *(Generalization error) Let Assumptions 1, 2, and 3 hold and $\hat{\boldsymbol{\beta}}$ be the solution of (5). if we assume that α_ψ, u_ψ are the unique solutions to (6), then in the asymptotic limit $n, m \to \infty$, $\frac{m}{n} \to \delta$, we have that*

$$r(\hat{\boldsymbol{\beta}}) \xrightarrow{P} \alpha_\psi^2. \tag{7}$$

The statements in Theorem 1 hold for a general class of separable convex potentials in (5). In the special cases of ℓ_1 and ℓ_2 potentials, our analysis recovers many of the results from [15, 27, 35]. The details of the proof are given in Appendix E.

3.3. Fundamental Limits

In this section, we study the fundamental limits on the generalization error of interpolators obtained through (5). To be precise, we consider the following class of convex potentials defined below

$$\mathcal{C}_\psi := \left\{ \Psi \mid \Psi(\boldsymbol{\beta}) = \sum_{i=1}^{n} \psi(\beta_i, \Sigma_{i,i}) \quad \text{s.t} \quad \psi(., \Sigma_{i,i}) \text{ is convex } \forall i \in [n] \right\}. \tag{8}$$

Therefore $\mathcal{C}_\psi$ contains a much broader class of separable convex functions that can have a dependence on the eigenspectrum of the data source, which is assumed to be known a priori. We provide a lower bound on the generalization error, which is valid for every potential in $\mathcal{C}_\psi$. Additionally, under certain conditions, we show that this lower bound is tight by constructing the optimal convex potential, which achieves this bound.

Theorem 2 *(Lower bound on α_ψ^2) Let Assumptions 1,2, and 3 hold. Define the random variable $V_\alpha = B + \frac{\alpha}{\sqrt{\delta}\Lambda} H$ where $H \sim \mathcal{N}(0,1)$ and B, Λ as defined in Assumption 2 and 3. Let α_* be the unique solution of the following non-linear equation*

$$\alpha^2 = \frac{\delta \sigma^2}{1-\delta} + \frac{\delta(1-\delta)}{\mathcal{I}_\Lambda(V_\alpha)} \tag{9}$$

where $\mathcal{I}_\Lambda(V_\alpha)$ is the weighted Fisher information of V_α defined as

$$\mathcal{I}_\Lambda(V_\alpha) := \mathbb{E}\left[\left(\frac{\xi_{V_\alpha}(V_\alpha|\Lambda)}{\Lambda} \right)^2 \right] \tag{10}$$

where $\xi_{V_\alpha}(v|\Lambda) := \frac{p'_{V_\alpha}(v|\Lambda)}{p_{V_\alpha}(v|\Lambda)}$ is the conditional score function of V_α. Then for every $\Psi \in \mathcal{C}_\psi$, with α_ψ^2 as the asymptotic limit of the generalization error as in (7), we have $\alpha_\psi^2 \geq \alpha_*^2$.

The proof of Theorem 2 is deferred to Appendix F, and it also involves showing the existence of a unique solution to the non-linear equation (9). The weighted Fisher information (10) can be computed more generally, even when P_B is not a differentiable potential, since adding a Gaussian smoothens out the density. Next, we provide a slightly weaker lower bound which avoids solving a non-linear equation like (9) in the special case of isotropic data, i.e, $\boldsymbol{\Sigma} = \boldsymbol{I}_n$.

Corollary 3 *Let Assumptions 1,2 and 3 hold and α_* be defined as the solution to (9), if $\Lambda = 1$ almost surely then*

$$\alpha_*^2 \geq \frac{\sigma^2}{1-\delta} + \frac{1-\delta}{\mathcal{I}(B)} \tag{11}$$

whenever, the Fisher information $\mathcal{I}(B)$ is well defined. The inequality becomes an equality if and only if B is a Gaussian.

The proof of corollary 3 is found in Appendix F and involves the application of Stam's inequality for Fisher information to $\mathcal{I}(V_\alpha)$, which makes it possible to solve (9) analytically.

Looking at (11) closely, the first term $\frac{\sigma^2}{1-\delta}$ represents the theoretical lower bound on the generalization error for all possible interpolating solutions [39] and it shows that overfitting the noise can benign when $\delta \ll 1$ i.e in the highly over-parameterized regime. The second term $\frac{1-\delta}{\mathcal{I}(B)}$ ends up being equal to the Bayes risk when B is Gaussian, and the variance of the noise is zero. Therefore (11) can be interpreted as the sum of the error from overfitting the noise and the error of the best possible estimator in the absence of noise. Contrary to the first term, $\frac{1-\delta}{\mathcal{I}(B)}$ is minimized when δ approaches 1, i.e., we have an equal number of equations and unknowns to fully recover the underlying unknown parameter. So there is an inherent tension between the two terms, and increasing δ, although recovers more of the signal, amplifies the noise due to overfitting.

Next, we will argue that these lower bounds obtained from Theorem 2 and Corollary 3 are indeed tight by constructing optimal convex potentials that achieve these lower bounds.

3.4. Optimal implicit Bias

Theorem 4 *(Optimal Ψ) Let Assumptions 1, 2, and 3 hold and α_* be defined as the solution to (9). Consider the following function $\psi_* : \mathbb{R}^2 \to \mathbb{R}$*

$$\psi_*(v, \lambda) := -\mathcal{M}_{\log(P_{V_{\alpha_*}}(v|\lambda))}\left(v; \frac{\alpha_*^2(1-\delta) - \delta\sigma^2}{\delta(1-\delta)\lambda^2}\right), \tag{12}$$

if $P_{V_{\alpha_*}}(v|\lambda)$ is log-concave in v and we define $\Psi_*(\boldsymbol{\beta}) = \sum_{i=1}^n \psi_*(\beta_i, \Sigma_{i,i})$, then

1. *$\Psi_*(\boldsymbol{\beta}) \in \mathcal{C}_\psi$*
2. *α_* is a solution to the system of equations (6) obtained using $\psi^*(v, \lambda)$ and is therefore the optimal convex implicit bias.*

Theorem 4 provides a construction of the optimal potential (12) that satisfies the system of equations (6). When Ψ_* obtained using (12) belongs to $\mathcal{C}_\psi$, then the $\alpha_{\psi_*}^2$ obtained from (6) denotes the asymptotic limit of the generalization error of Ψ_* and by Theorem 4, we have that $\alpha_{\psi_*}^2 = \alpha_*^2$, thereby achieving the lower bound. The proof of this deferred to Appendix F and involves verifying that the construction in (12) satisfies the system of equations (6) and characterizing the sufficient conditions for the convexity of Ψ_*.

Remark 5 *The log-concavity of $P_{V_{\alpha_*}}(v|\lambda)$ must be verified on a case-by-case basis by first solving for α_*. A sufficient condition that always ensures log-concavity of $P_{V_{\alpha_*}}(v|\lambda)$ is by letting P_B be a log-concave density since convolving a Gaussian density with P_B preserves its log-concavity. Even when P_B is not log-concave, it is possible that if α_* obtained from solving (9) is greater than a*

certain threshold value, it smoothens out P_B enough such that the resulting $P_{V_{\alpha_*}}(v|\lambda)$ is log-concave. This, in turn, characterizes a region of δ and σ^2, where the optimal convex potential is achievable and the lower bound is tight.

We'll observe in numerically in Section A, where we consider the cases when P_B is sparse-Gaussian and Rademacher both of which are not even differentiable but the optimal potential construction obtained from Theorem 4 is indeed convex for the values of δ and σ^2 chosen. Next, we look at the special case when the underlying parameter density P_B is a Gaussian density.

Corollary 6 *(Ψ_* for Gaussian B) Let Assumptions 1, 2, and 3 hold and $B \sim \mathcal{N}(0,1)$, then the optimal implicit bias is given as*

$$\Psi_*(\boldsymbol{\beta}) = \boldsymbol{\beta}^T \boldsymbol{\Sigma}^{1/2} \left(\frac{\sigma^2}{1-\delta} \boldsymbol{I}_n + \boldsymbol{\Sigma} \right)^{-1} \boldsymbol{\Sigma}^{1/2} \boldsymbol{\beta}. \tag{13}$$

Corollary 6 comes from the direct application of Theorem 4 while ignoring the constant terms in the potential since they don't affect the outcome of the optimization. In fact, we observe that (13) is quadratic, which is convex, and therefore, it achieves the lower bound α_*^2 on the generalization error. If $\frac{\sigma^2}{1-\delta} \gg \Lambda$, which happens either when the noise variance σ^2 is large or we are slightly over-parameterized, i.e., δ approaches 1, then the optimal potential $\Psi_*(\boldsymbol{\beta}) \approx \boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta}$. In the other case, when $\frac{\sigma^2}{1-\delta} \ll \Lambda$, the optimal potential is $\Psi_*(\boldsymbol{\beta}) \approx \|\boldsymbol{\beta}\|_2^2$. In the special case when Σ is isotropic or when the variance of the noise $\sigma^2 = 0$, the optimal convex potential is exactly the Euclidean norm squared, the implicit bias of SGD.

Remark 7 *(Comparison with [44]) Although we consider a bigger class of interpolators in (8), the set of interpolators considered in [44] is not completely inclusive in our class. The key difference is that our optimal convex potential construction doesn't depend on the observed labels data X, y. In contrast, the optimal linear response interpolator considered in [44] involves a quadratic potential that can depend on X, y, and this dependence makes analysis quite difficult outside of the quadratic case. This seems to be a subtle difference, and we, in fact, see similar qualitative trends as discussed under Corollary 6.*

4. Conclusion

This work provides a precise asymptotic analysis of the generalization performance of interpolators for linear models obtained as a minimization of a convex potential, which is characterized by the implicit bias of the optimization algorithm. Additionally, we also derive the fundamental lower bounds on the achievable generalization error of interpolators obtained from the minimization of convex potentials and characterize the optimal convex potential that achieves these bounds. Additional numerical simulations of the derived results can be found in Appendix A. Extending these results to non-asymptotic settings and characterizing the optimal implicit bias in this setting are important future directions. It would also be interesting to generalize these results to non-linear models and even study the role of implicit bias in classification problems.

References

[1] Ehsan Abbasi, Christos Thrampoulidis, and Babak Hassibi. General performance metrics for the lasso. In *2016 IEEE Information Theory Workshop (ITW)*, pages 181–185. IEEE, 2016.

[2] Madhu Advani and Surya Ganguli. Statistical mechanics of optimal convex inference in high dimensions. *Physical Review X*, 6(3):031034, 2016.

[3] Shun-ichi Amari, Jimmy Ba, Roger Grosse, Xuechen Li, Atsushi Nitanda, Taiji Suzuki, Denny Wu, and Ji Xu. When does preconditioning help or hurt generalization? *arXiv preprint arXiv:2006.10732*, 2020.

[4] Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. Living on the edge: A geometric theory of phase transitions in convex optimization. 2013.

[5] Per Kragh Andersen and Richard D Gill. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120, 1982.

[6] Navid Azizan, Sahin Lale, and Babak Hassibi. Stochastic mirror descent on overparameterized nonlinear models. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12): 7717–7727, 2021.

[7] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[8] Mohsen Bayati and Andrea Montanari. The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2011.

[9] Derek Bean, Peter J Bickel, Noureddine El Karoui, and Bin Yu. Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences*, 110(36): 14563–14568, 2013.

[10] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[11] Nelson Blachman. The convolution inequality for entropy powers. *IEEE Transactions on Information theory*, 11(2):267–271, 1965.

[12] Zhiqi Bu, Jason Klusowski, Cynthia Rush, and Weijie Su. Algorithmic analysis and statistical estimation of slope via approximate message passing. *Advances in Neural Information Processing Systems*, 32, 2019.

[13] Michael Celentano and Andrea Montanari. Fundamental barriers to high-dimensional regression with convex penalties. *The Annals of Statistics*, 50(1):170–196, 2022.

[14] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6): 805–849, 2012.

[15] Xiangyu Chang, Yingcong Li, Samet Oymak, and Christos Thrampoulidis. Provable benefits of overparameterization in model compression: From double descent to pruning neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6974–6983, 2021.

[16] Niladri S Chatterji and Philip M Long. Foolish crowds support benign overfitting. *Journal of Machine Learning Research*, 23(125):1–12, 2022.

[17] Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

[18] Konstantin Donhauser, Nicolo Ruggeri, Stefan Stojanovic, and Fanny Yang. Fast rates for noisy interpolation require rethinking the effect of inductive bias. In *International Conference on Machine Learning*, pages 5397–5428. PMLR, 2022.

[19] David Donoho and Andrea Montanari. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166:935–969, 2016.

[20] David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[21] David L Donoho, Arian Maleki, and Andrea Montanari. The noise-sensitivity phase transition in compressed sensing. *IEEE Transactions on Information Theory*, 57(10):6920–6941, 2011.

[22] Noureddine El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170:95–175, 2018.

[23] Melikasadat Emami, Mojtaba Sahraee-Ardakan, Parthe Pandit, Sundeep Rangan, and Alyson Fletcher. Generalization error of generalized linear models in high dimensions. In *International Conference on Machine Learning*, pages 2892–2901. PMLR, 2020.

[24] Ky Fan. Minimax theorems. *Proceedings of the National Academy of Sciences*, 39(1):42–47, 1953.

[25] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1827–1836, 2018.

[26] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Implicit bias of gradient descent on linear convolutional networks. *arXiv preprint arXiv:1806.00468*, 2018.

[27] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.

[28] Hong Hu and Yue M Lu. Slope for sparse linear regression: asymptotics and optimal regularization. *IEEE Transactions on Information Theory*, 68(11):7627–7664, 2022.

[29] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[30] Noureddine El Karoui. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*, 2013.

[31] Frederic Koehler, Lijia Zhou, Danica J Sutherland, and Nathan Srebro. Uniform convergence of interpolators: Gaussian width, norm bounds and benign overfitting. *Advances in Neural Information Processing Systems*, 34:20657–20668, 2021.

[32] Gil Kur, Pedro Abdalla, Pierre Bizeul, and Fanny Yang. Minimum norm interpolation meets the local theory of banach spaces. In *Forty-first International Conference on Machine Learning*, 2024.

[33] Lihua Lei, Peter J Bickel, and Noureddine El Karoui. Asymptotics for high dimensional regression m-estimates: fixed design results. *Probability Theory and Related Fields*, 172: 983–1079, 2018.

[34] Mingchen Li, Yahya Sattar, Christos Thrampoulidis, and Samet Oymak. Exploring weight importance and hessian bias in model pruning. *arXiv preprint arXiv:2006.10903*, 2020.

[35] Yue Li and Yuting Wei. Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent. *arXiv preprint arXiv:2110.09502*, 2021.

[36] Panagiotis Lolas. Regularization in high-dimensional regression and classification via random matrix theory. *arXiv preprint arXiv:2003.13723*, 2020.

[37] Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pages 3325–3334. PMLR, 2018.

[38] Léo Miolane and Andrea Montanari. The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *The Annals of Statistics*, 49(4):2313–2335, 2021.

[39] Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 2020.

[40] Jeffrey Negrea, Gintare Karolina Dziugaite, and Daniel Roy. In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors. In *International Conference on Machine Learning*, pages 7263–7272. PMLR, 2020.

[41] Arkadii Nemirovski and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

[42] Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

[43] Behnam Neyshabur, Ryota Tomioka, Ruslan Salakhutdinov, and Nathan Srebro. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071*, 2017.

[44] Eduard Oravkin and Patrick Rebeschini. On optimal interpolation in linear regression. *Advances in Neural Information Processing Systems*, 34:29116–29128, 2021.

[45] Samet Oymak and Babak Hassibi. Sharp mse bounds for proximal denoising. *Foundations of Computational Mathematics*, 16:965–1029, 2016.

[46] Samet Oymak and Joel A Tropp. Universality laws for randomized dimension reduction, with applications. *Information and Inference: A Journal of the IMA*, 7(3):337–446, 2018.

[47] Samet Oymak, Christos Thrampoulidis, and Babak Hassibi. The squared-error of generalized lasso: A precise analysis. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1002–1009. IEEE, 2013.

[48] R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970.

[49] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[50] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19 (70):1–57, 2018.

[51] Mihailo Stojnic. Various thresholds for ℓ_1-optimization in compressed sensing. *arXiv preprint arXiv:0907.3666*, 2009.

[52] Mihailo Stojnic. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013.

[53] Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Sharp asymptotics and optimal performance for inference in binary models. In *International Conference on Artificial Intelligence and Statistics*, pages 3739–3749. PMLR, 2020.

[54] Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Fundamental limits of ridge-regularized empirical risk minimization in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 2773–2781. PMLR, 2021.

[55] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pages 1683–1709. PMLR, 2015.

[56] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

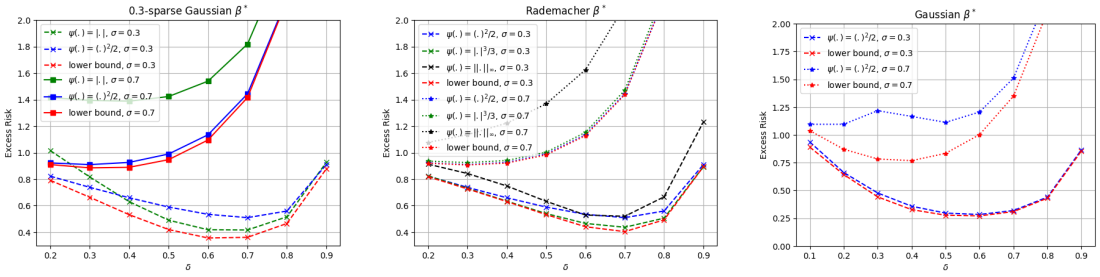
[57] Guillaume Wang, Konstantin Donhauser, and Fanny Yang. Tight bounds for minimum ℓ_1-norm interpolation of noisy data. In *International Conference on Artificial Intelligence and Statistics*, pages 10572–10602. PMLR, 2022.

[58] Shuaiwen Wang, Haolei Weng, and Arian Maleki. Does slope outperform bridge regression? *Information and Inference: A Journal of the IMA*, 11(1):1–54, 2022.

[59] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[60] Lijia Zhou, Danica J Sutherland, and Nati Srebro. On uniform convergence and low-norm interpolation learning. *Advances in Neural Information Processing Systems*, 33:6867–6877, 2020.

Appendix A. Numerical Simulations

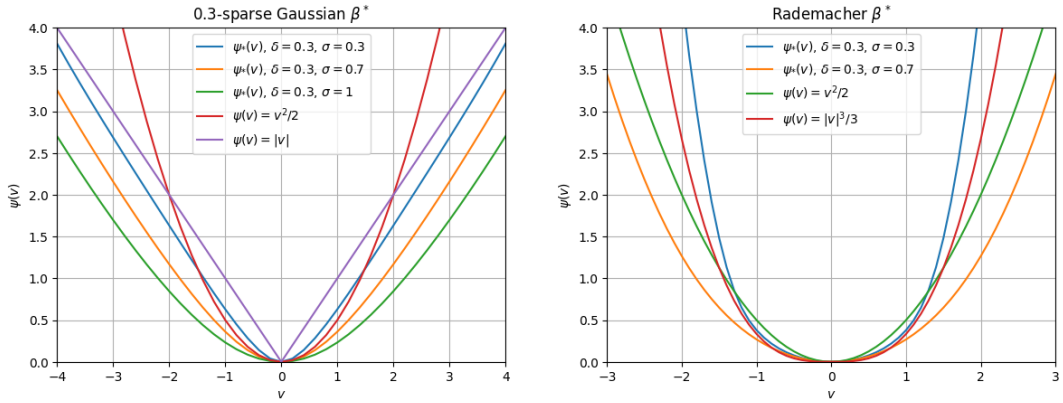


(a) (b) (c)

Figure 1: Comparison of excess risk of different interpolators (a) when P_B has a 0.3-sparse Gaussian density and $\Lambda = 1$ a.s (b) when P_B has a Rademacher density and $\Lambda = 1$ a.s (c) when P_B has a Gaussian density and Λ^2 takes the value 4 with probability 0.3 and 0.1 with probability 0.7.

In this section, we provide numerical simulations of the results derived in section 3.2 and 3.3 and provide insights on the implications of these results for certain special cases. In particular, we study the cases when the prior distribution of the underlying signal parameter P_B is a sparse Gaussian (Figure 1 (a)), Rademacher (Figure 1 (b)) and Gaussian (Figure 1 (c)). Figures 2 (a) and 2 (b) show the structure of the optimal convex potentials when the underlying signal parameter P_B is a sparse Gaussian and Rademacher, respectively. We normalize the underlying prior signal such that $\mathbb{E}[B^2] = 1$ and define the signal-to-noise ratio (SNR) as $\frac{1}{\sigma^2}$ and consider the regimes of low SNR ($\sigma = 0.7$) and high SNR ($\sigma = 0.3$). All the above-mentioned plots evaluate the asymptotic theoretical limits of the results derived and are obtained from solving a system of nonlinear equations involving expectations of certain quantities. We use standard packages like Scipy to compute these expectations using numerical integrals and solve the system of non-linear equations. Additional plots demonstrating the concentration of non-asymptotics are deferred to the Appendix E.

Sparse Gaussian. In Figure 1 (a), we consider the case when the underlying unknown parameter B is 0.3-sparse Gaussian, i.e., with probability 0.3 behaves like a Gaussian and is zero otherwise. We consider isotropic data and B is scaled appropriately such that its variance is one. The y-axis represents the excess risk of the interpolating solution obtained, and the x-axis sweeps across the overparameterization ratio δ. In the absence of noise, it's well known that we get perfect recovery for the ℓ_1 norm at approximately two times the sparsity of the signal [14], but in the presence of noise, interpolation can't recover the signal. In the high SNR regime ($\sigma = 0.3$), we observe that the minimum ℓ_1 interpolator does, in fact, outperform the minimum ℓ_2 interpolator for certain regions of δ. But in the low SNR regime ($\sigma = 0.7$), observe that the ℓ_1 interpolator performs significantly worse than ℓ_2, which suggests that imposing structure while interpolating noisy labels can amplify the noise more than the recovery of the signal. We additionally observe that ℓ_2 is, in fact, very close to the optimal performance characterized by the lower bound. Figure 2 (a) shows the structure of the optimal convex potentials that achieve these lower bounds at $\delta = 0.3$ at different SNRs. We see that the optimal potential behaves like a smoothened version of the ℓ_1-norm, and as SNR decreases, the



(a) (b)

Figure 2: Comparison of the structure of the optimal convex potentials when (a) P_B has a 0.3-sparse Gaussian density and $\Lambda = 1$ a.s (b) P_B has a Rademacher density and $\Lambda = 1$ a.s

optimal potential approaches the Euclidean norm squared, supporting our previous observations that the ℓ_2-norm interpolator is, in fact, close to optimal at low SNR.

Rademacher. In Figure 1 (b), we consider Rademacher distributed B, where it take the values $\{\pm 1\}$ with equal probability with isotropic Gaussian data. The recovery threshold in the absence of noise for the ℓ_∞-norm was shown to be $\delta = 0.5$ [14]. In the presence of noise, even in the high SNR regime ($\sigma = 0.3$), we observe that ℓ_∞-norm is outperformed by both the minimum ℓ_2 and ℓ_3 norm interpolators. In the low SNR regime ($\sigma = 0.7$), this gap in performance only grows wider as expected. Looking at the structure of the optimal convex potential in Figure 2 (b), we see that for high SNRs, the optimal potential is much flatter around the origin and increases steeply at around 1 compared to the square and cubic potentials and as we move to the low SNR regime, the optimal potential smoothens out.

Gaussian. Finally, we consider the Gaussian prior in Figure 1 (c). For isotropic data, we have established in Corollary 6 that the optimal potential is indeed the minimum ℓ_2-norm interpolator. So now, we consider non-isotropic data with a bi-level eigen-spectrum where the diagonal entries of covariance $\mathbf{\Sigma}$ are set to 4 with probability 0.3 and 0.1 with probability 0.7. In the high SNR regime ($\sigma = 0.3$), we observe that the minimum ℓ_2-norm interpolator is quite close to the lower bound, which is achieved by the optimal potential, which has access to the elements of $\mathbf{\Sigma}$. This is not surprising since from (13), we observe that for high SNRs, the optimal potential approaches the Euclidean norm squared. In contrast, for the low SNR regime ($\sigma = 0.7$), we see a significant gap in the performance of the minimum ℓ_2 norm interpolator and the optimal achievable interpolator. This again can be explained from (13), where for low SNRs, we see that the optimal convex potential $\Psi_*(\boldsymbol{\beta}) \approx \boldsymbol{\beta}^T \mathbf{\Sigma} \boldsymbol{\beta}$. Therefore, having access to the eigen-spectrum of the data source makes the difference when the noise variance is large.

Appendix B. Related Work

There has been extensive literature studying the precise asymptotics of different convex regularized estimators for linear models [1, 4, 8, 12–15, 17, 19, 21–23, 27, 28, 30, 33, 36, 38, 45–47, 51, 52, 55, 56, 58]. The convex Gaussian minimax Theorem (CGMT) [52, 55] provides a framework to do this precise asymptotic analysis in many of these afore-mentioned works and will also be the primary tool for analysis in our work. Another popular approach used in precise asymptotic analysis is approximate message passing (AMP) [20, 21], and it is conceivable that results obtained in our work can also be derived using AMP analysis. In the context of interpolation, under the proportional regime of Assumption 1 [15, 27] do a precise asymptotic analysis of the minimum ℓ_2-norm interpolators and [35] studied the precise asymptotics of the minimum ℓ_1-norm interpolator for isotropic Gaussian data using AMP. Our results extend this analysis to general separable convex potentials on non-isotropic Gaussian data, which include ℓ_1 and ℓ_2 norms as special cases, and we recover the previous results.

There is a rich body of work on non-asymptotic analysis showing consistent rates of minimum-norm interpolators [7, 16, 18, 31, 32, 40, 57, 60]. These works typically consider the highly overparameterized regime, i.e., $\delta \ll 1$, which is a necessary condition for consistency. In contrast, we consider the proportional regime, where $m, n \to \infty$ and $m/n \to \delta$, where consistency is not possible [39], and therefore a sharp characterization of the asymptotic generalization error is of interest. The recent works of [31, 40, 57, 60] also use CGMT in their analysis, where Gaussian comparison lemmas are used to obtain bounds on the risk using uniform convergence arguments. This approach is different from ours, where CGMT is used directly on the objective (1), which simplifies the objective to a scalar optimization in the asymptotic limit similar to [15].

In terms of characterizing fundamental limits, [2, 9, 19] were the first works to derive lower bounds and optimal loss functions for high-dimensional linear regression problems in the absence of regularization, and therefore, they consider the under-parameterized regime with a unique global optimum. More recently, [13] studied the fundamental limits of convex regularized linear regression, where they considered the square loss and derived lower bounds on the prediction error obtained from an optimally tuned convex regularizer. Similar results on the lower bounds of the prediction error were also studied for binary classification [53] and for ridge-regularized regression for linear and binary models in [54]. None of these prior works consider the case of interpolation in over-parameterized models, and our analysis extends these ideas to derive fundamental lower bounds on the generalization error of interpolating solutions on linear models obtained from minimizing convex potentials, improving upon previous results of [3, 44].

Appendix C. Deep Neural Networks

In this section, we discuss the applicability of our results to non-linear models, in particular, deep networks. Linearity plays a key role throughout our analysis; therefore, our results presented may not directly translate to general non-linear settings, but in certain regimes, it can be shown that neural networks are well-approximated by linear models. The neural tangent kernel (NTK) [29] framework has been one the main tools to theoretically understand the optimization of infinitely wide neural networks. In this infinite width limit, training using gradient descent becomes equivalent to optimizing linear functions in the infinite-dimensional Hilbert space defined by the NTK, which is in line with our problem setting of letting $n \to \infty$. Another key assumption we make is the Gaussianity of the data. Although it is conceivable that due to Gaussian universality, the analysis shown holds more generally for non-Gaussian data, this remains to be shown. In terms of a practical algorithm

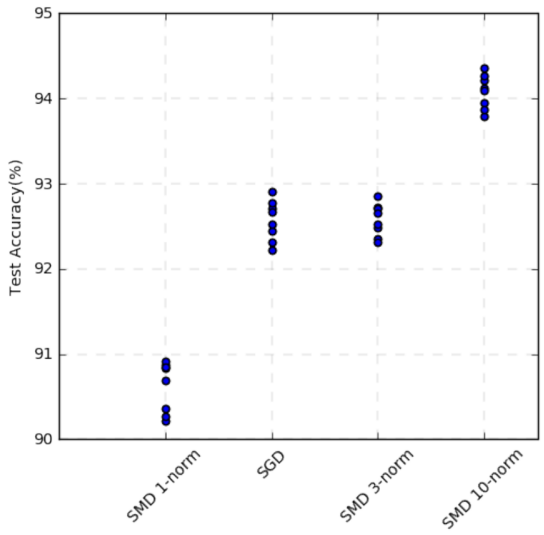


Figure 3: Generalization performance of ResNet-18 on the CIFAR-10 using different SMD algorithms. [6].

to arrive at these minimum convex interpolators, [6] shows both experimentally and theoretically the validity of the implicit regularization property of SMD if the initialization is close enough to the manifold of global minima (something that comes for free in the highly overparameterized case). The optimal convex potentials derived in this work can be implemented directly using the SMD update rule if the derivative of the convex potential is invertible, and the separability makes this implementation efficient. Clearly, the choice of potential plays an important role in determining the generalization performance, as seen in Figure C; therefore, it should be treated as a hyper-parameter during optimization. An extensive survey of empirical experiments by varying the choice of potential on different model architectures and problem domains would be useful in guiding the choice of potential for a given learning problem.

Appendix D. Useful facts

D.1. Properties of Moreau Envelope

In this section, we provide some properties of Moreau Envelope, which will be used in our analysis later. These results are mainly borrowed from [49] and also previously used in [53]

Proposition 8 *Let $\psi : \mathbb{R} \to \mathbb{R}$ be a proper lower semi-continuous convex function.*

1. *Then $\lim_{\alpha \to 0_+} \mathcal{M}_\psi(x;\alpha) \to \psi(x)$ and $\lim_{\alpha \to \infty} \mathcal{M}_\psi(x;\alpha) \to \min_{y \in \mathbb{R}} \psi(y)$.*
2. *The derivatives of the Moreau Envelope satisfy the following*

$$\mathcal{M}'_{\psi,1}(x;\alpha) := \frac{\partial \mathcal{M}_\psi(x;\alpha)}{\partial x} = \frac{1}{\alpha}(x - prox_\psi(x;\alpha)), \tag{14}$$

$$\mathcal{M}'_{\psi,2}(x;\alpha) := \frac{\partial \mathcal{M}_\psi(x;\alpha)}{\partial \alpha} = -\frac{1}{2\alpha^2}(x - prox_\psi(x;\alpha))^2. \tag{15}$$

Proposition 9 *For $\alpha > 0$ and a function h, we have that*

$$\mathcal{M}_{\psi}(x;\alpha) = \frac{x^2}{2\alpha} - \frac{1}{\alpha}\left(\frac{x^2}{2} + \alpha\psi(x)\right)^* \tag{16}$$

where $\left(\frac{x^2}{2} + \alpha\psi(x)\right)^*$ is the convex conjugate of $\frac{x^2}{2} + \alpha\psi(x)$.

Proposition 10 *(Inverse of Moreau envelope)[[2], result 23 in appendix] For $\alpha > 0$ and ψ a convex, lower semi-continuous function such that $g(x) = \mathcal{M}_{\psi}(x;\alpha)$, the Moreau envelope can be inverted so that $\psi(x) = -\mathcal{M}_{-g}(x;\alpha)$.*

D.2. Properties of Fisher Information

We now state some standard properties of the Fisher information of location which are used in our analysis and the additional details of which can be found in [11].

Proposition 11 *Let X be a zero mean random variable with a differentiable probability density P_X such that $P_X(x) > 0, -\infty < x < \infty$ and the following integral is well-defined*

$$\mathcal{I}(X) := \int_{-\infty}^{\infty} \frac{(P_X^{'}(x))^2}{P_X(x)} dx. \tag{17}$$

The Fisher information of location $\mathcal{I}(X)$ defined as above satisfies the following properties.

(a) For any $c \in \mathbb{R}$, $\mathcal{I}(cX) = \mathcal{I}(X)/c^2$

(b) (Cramer-Rao bound) $\mathcal{I}(cX) \leq \frac{1}{\mathbb{E}[X^2]}$, with equality if and only if X has a Gaussian.

(c) Let X_1, X_2 be independent random variables with well-defined $\mathcal{I}(X_1), \mathcal{I}(X_2)$ and $\alpha \in [0,1]$. Then it holds that

$$\mathcal{I}(X_1 + X_2) \leq \alpha^2 \mathcal{I}(X_1) + (1-\alpha)^2 \mathcal{I}(X_2) \tag{18}$$

(d) (Stam's inequality)For the two independent random variables X_1, X_2 defined above, the following holds

$$\mathcal{I}(X_1 + X_2) \leq \frac{\mathcal{I}(X_1) \cdot \mathcal{I}(X_2)}{\mathcal{I}(X_1) + \mathcal{I}(X_2)} \tag{19}$$

The inequality is obtained from optimizing the upper bound over α in (c), and the inequality becomes equality if and only if X_1, X_2 are independent Gaussians.

Proposition 12 *Consider the random variables $H \sim \mathcal{N}(0,1)$ and B such that $V_\alpha := B + \alpha H$ satisfies the conditions in Proposition 11 for $\alpha \in \mathbb{R}_{>0}$. Then we have that*

(a) $\lim_{\alpha \to 0_+} \alpha^2 \mathcal{I}(V_\alpha) \to 0$

(b) $\lim_{\alpha \to \infty} \alpha^2 \mathcal{I}(V_\alpha) \to 1$

The proof of the above proposition involves using property (c) of proposition 11, and a similar result was also shown in [54].

Appendix E. Precise Asymptotics

In this section, we provide the proof for the precise asymptotics of the interpolating solutions obtained from (5) and also study the special cases of the minimum ℓ_1, ℓ_2, ℓ_3 and ℓ_∞-norm interpolator.

E.1. Proof of Theorem 1

Proof Consider the following problem given in (1)

$$\min_{\boldsymbol{\beta}} \Psi(\boldsymbol{\beta}) \quad \text{subject to} \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} \tag{20}$$

Doing a change of variable $\boldsymbol{w} := \frac{1}{\sqrt{n}}\Sigma^{\frac{1}{2}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$, we the following constrained minimization problem

$$\min_{\boldsymbol{w}} \Psi(\boldsymbol{\beta}^* + \sqrt{n}\Sigma^{-1/2}\boldsymbol{w}) \quad \text{subject to} \quad \boldsymbol{G}\boldsymbol{w} = \boldsymbol{z} \tag{21}$$

Boundedness of solution Now, we assume that $\boldsymbol{w}$ can be restricted to a large enough bounded set $\boldsymbol{w} \in \mathcal{W} := \{\boldsymbol{w} \text{ s.t } \|\boldsymbol{w}\|_2 \leq B_+\}$ without changing the optimization problem. This is more of a technicality required for the application of CGMT. In [15], it was explicitly shown that for the minimum ℓ_2-norm interpolator, this assumption is true. Since the value of ℓ_p-norms is less than ℓ_2-norm for $p > 2$, it can be argued that this bounded set construction for ℓ_2 is also valid for ℓ_p-norms bigger than ℓ_2. But more generally, for separable convex functions, this condition must be verified on a case-by-case basis. Alternatively, if we assume that α_ψ is bounded, then letting $B_+ = 2\alpha_\psi$ also obviates this issue. Taking this into consideration, we now have the following primary optimization (PO) problem.

$$\Phi(\boldsymbol{G}) = \min_{\boldsymbol{w}\in\mathcal{W}} \Psi(\boldsymbol{\beta}^* + \sqrt{n}\Sigma^{-1/2}\boldsymbol{w}) \quad \text{subject to} \quad \boldsymbol{G}\boldsymbol{w} = \boldsymbol{z} \tag{22}$$

Using constrained CGMT formulation [34], the Auxiliary optimization (AO) is given as

$$\phi(\boldsymbol{g}, \boldsymbol{h}) = \min_{\boldsymbol{w}\in\mathcal{W}} \Psi(\boldsymbol{\beta}^* + \sqrt{n}\Sigma^{-1/2}\boldsymbol{w}) \quad \text{subject to} \quad \|\boldsymbol{g}\|\sqrt{\|\boldsymbol{w}\|^2 + \sigma^2} - \boldsymbol{h}^T\boldsymbol{w} - \sigma h \leq 0 \tag{23}$$

where $\boldsymbol{g}$ and $\boldsymbol{h}$ are random vector with iid standard Gaussian entries and h is an iid standard Gaussian scalar. Bringing the constraint into the objective of the AO using Lagrange multiplier $u \geq 0$ and normalizing the constraint with $\sqrt{n}$, we get

$$(\hat{\boldsymbol{w}}_n^{AO}, u_n) := \arg\min_{\boldsymbol{w}\in\mathcal{W}} \max_{u\geq 0} \psi(\boldsymbol{\beta}^* + \sqrt{n}\Sigma^{-1/2}\boldsymbol{w}) + u(\|\boldsymbol{g}_e\|\sqrt{\delta}\sqrt{\|\boldsymbol{w}\|^2 + \sigma^2} - \boldsymbol{h}_e^T\boldsymbol{w} - \frac{\sigma h}{\sqrt{n}}) \tag{24}$$

where $\boldsymbol{h}_e := \frac{\boldsymbol{h}}{\sqrt{n}}$ and $\boldsymbol{g}_e := \frac{\boldsymbol{g}}{\sqrt{m}}$. Interchange min max using [24] since the objective is convex in $\boldsymbol{w}$ on a compact set $\mathcal{W}$ and concave in u. Next, using the square root trick on $\|\boldsymbol{g}_e\|\sqrt{\|\boldsymbol{w}\|^2 + \sigma^2}$, we have that

$$\|\boldsymbol{g}_e\|\sqrt{\|\boldsymbol{w}\|^2 + \sigma^2} = \min_{\alpha\in\mathcal{A}} \frac{\alpha\|\boldsymbol{g}_e\|^2}{2} + \frac{\|w\|^2 + \sigma^2}{2\alpha}. \tag{25}$$

where $\mathcal{A} = [\sigma, \sqrt{\sigma^2 + B_+^2}]$. Plugging back into the AO, we get

$$\begin{aligned}(\hat{\boldsymbol{w}}_n^{AO}, u_n, \alpha_n) := \arg\max_{u\geq 0} \min_{\boldsymbol{w}\in\mathcal{W},\alpha\in\mathcal{A}} &\frac{u\alpha\sqrt{\delta}\|\boldsymbol{g}_e\|^2}{2} + \frac{u\sqrt{\delta}\sigma^2}{2\alpha} \\ &- \frac{u\sigma h}{\sqrt{n}} + \frac{u\sqrt{\delta}}{2\alpha}\|\boldsymbol{w}\|^2 - u\boldsymbol{h}_e^T\boldsymbol{w} + \Psi(\boldsymbol{\beta}^* + \sqrt{n}\Sigma^{-1/2}\boldsymbol{w})\end{aligned} \tag{26}$$

Using separability of Ψ and appropriate re-scaling, we let $\Psi(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\psi(x_i, \Sigma_{i,i})$, and moving the minimization over $\boldsymbol{w}$ inside the objective, we get

$$(\hat{\boldsymbol{w}}_n^{AO}, u_n, \alpha_n) := \arg\max_{u\geq 0}\min_{\alpha\in\mathcal{A}} \frac{u\alpha\sqrt{\delta}}{2} + \frac{u\sqrt{\delta}\sigma^2}{2\alpha} - \frac{u\sigma h}{\sqrt{n}} + \min_{\boldsymbol{w}\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n}\{\frac{u\sqrt{\delta}n}{2\alpha}w_i^2 - u\sqrt{n}h_i w_i + \psi(\beta_i^* + \sqrt{n}\Sigma_{i,i}^{-1/2}w_i, \Sigma_{i,i})\}. \quad (27)$$

Doing a change of variable back to the original weight vector $\boldsymbol{\beta} = \boldsymbol{\beta}^* + \sqrt{n}\Sigma^{-1/2}\boldsymbol{w}$, we get

$$(\hat{\boldsymbol{\beta}}_n^{AO}, u_n, \alpha_n) := \arg\max_{u\geq 0}\min_{\alpha\in\mathcal{A}} \frac{u\alpha\sqrt{\delta}}{2} + \frac{u\sqrt{\delta}\sigma^2}{2\alpha} - \frac{u\sigma h}{\sqrt{n}} + \min_{\boldsymbol{\beta}\in\mathcal{B}_\beta} \frac{1}{n}\sum_{i=1}^{n}\{\frac{u\sqrt{\delta}}{2\alpha}\Sigma_{i,i}(\beta_i - \beta_i^*)^2 - uh_i\Sigma_{i,i}^{1/2}(\beta_i - \beta_i^*) + \psi(\beta_i, \Sigma_{i,i})\} \quad (28)$$

where $\mathcal{B}_\beta := \{\boldsymbol{\beta} \text{ s.t } \frac{\sqrt{\Sigma}}{\sqrt{n}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \in \mathcal{W}\}$. Completing the squares for β_i and identifying the Moreau Envelope gives us $D_n(\alpha, u)$ defined as the optimal value of the objective of the optimization defined below.

$$(u_n, \alpha_n) := \arg\max_{u\geq 0}\min_{\alpha\in\mathcal{A}} \frac{u\alpha\sqrt{\delta}}{2} + \frac{u\sqrt{\delta}\sigma^2}{2\alpha} - \frac{u\sigma h}{\sqrt{n}} + \frac{1}{n}\sum_{i=1}^{n}\{\mathcal{M}_\psi(\beta_i^* + \frac{\alpha}{\sqrt{\delta}\Sigma_{i,i}^{1/2}}h_i; \frac{\alpha}{u\sqrt{\delta}\Sigma_{i,i}}) - \frac{u\alpha}{2\sqrt{\delta}}h_i^2\} \quad (29)$$

Here $\hat{\boldsymbol{\beta}}_{n,i}^{AO} = \text{prox}_\psi(\beta_i^* + \frac{\alpha_n}{\sqrt{\delta}\Sigma_{i,i}^{1/2}}h_i; \frac{\alpha_n}{u_n\sqrt{\delta}\Sigma_{i,i}})$ is always unique given α_n, u_n, since proximal is the solution to a strongly convex optimization. The above optimization is strictly convex in α, so the saddle point solutions (u_n, α_n) have unique α_n. For u_n to be unique, we need to assume $\frac{1}{n}\sum_{i=1}^{n}\{\mathcal{M}_\psi(\beta_i^* + \frac{\alpha}{\sqrt{\delta}\Sigma_{i,i}^{1/2}}h_i; \frac{\alpha}{u\sqrt{\delta}\Sigma_{i,i}})\}$ is strictly concave with probability approaching 1.

Asymptotic limits We consider, the proportional asymptotic limit $n, m \to \infty$, $\frac{m}{n} \to \delta < 1$. In this limit, $\frac{u\sigma h}{\sqrt{n}} \xrightarrow{P} 0$ and we also have that $\frac{1}{n}\sum_{i=1}^{n}\frac{u\alpha}{2\sqrt{\delta}}h_i^2 \xrightarrow{P} \frac{u\alpha}{2\sqrt{\delta}}$. Next,

$$\frac{1}{n}\sum_{i=1}^{n}\{\mathcal{M}_\psi(\beta_i^* + \frac{\alpha}{\sqrt{\delta}\Sigma_{i,i}^{1/2}}h_i; \frac{\alpha}{u\sqrt{\delta}\Sigma_{i,i}})\} \xrightarrow{P} \mathbb{E}[\mathcal{M}_\psi(B + \frac{\alpha}{\sqrt{\delta}\Lambda}H; \frac{\alpha}{u\sqrt{\delta}\Lambda^2})] \quad (30)$$

As a consequence, we have point wise convergence of $D_n(\alpha, u) \xrightarrow{P} D(\alpha, u)$ which is the following scalar optimization problem

$$\arg\min_{\alpha\in\mathcal{A}}\max_{u\geq 0} D(\alpha, u) := \arg\min_{\alpha\in\mathcal{A}}\max_{u\geq 0} -\frac{u\alpha(1-\delta)}{2\sqrt{\delta}} + \frac{u\sqrt{\delta}\sigma^2}{2\alpha} + \mathbb{E}[\mathcal{M}_\psi(B + \frac{\alpha}{\sqrt{\delta}\Lambda}H; \frac{\alpha}{u\sqrt{\delta}\Lambda^2})] \quad (31)$$

By [5], since both D_n and D are convex, concave in α, u, the convergence is uniform, and we have that the objective of converges

$$\phi(\boldsymbol{g}, \boldsymbol{h}) \xrightarrow{P} \min_{\alpha\in\mathcal{A}}\max_{u\geq 0} D(\alpha, u) \quad (32)$$

If $\mathbb{E}[\mathcal{M}_\psi(B+\frac{\alpha}{\sqrt{\delta}\Lambda}H;\frac{\alpha}{u\sqrt{\delta}\Lambda^2})]$ is strictly concave in u, then $D(\alpha,u)$ is strictly convex and strictly concave, we have parameter convergence by [42] (Lemma 7.75), therefore

$$(\alpha_n,u_n)\xrightarrow{P}(\alpha^*,u^*):=\arg\min_{\alpha\in\mathcal{A}}\min_{u\geq 0}D(\alpha,u) \tag{33}$$

In the absence of strong concavity of u, we only have the convergence of $\alpha_n\xrightarrow{P}\alpha^*$. Typically, distributional convergence requires strict concavity. Generalization error analysis doesn't.

$$r(\hat{\boldsymbol{\beta}}^{AO})=\frac{1}{n}(\hat{\boldsymbol{\beta}}^{AO}-\boldsymbol{\beta}^*)^T\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}^{AO}-\boldsymbol{\beta}^*)+\sigma^2=\|\hat{\boldsymbol{w}}^{AO}\|^2+\sigma^2=\|\bar{\boldsymbol{g}}\|\alpha_n^2\xrightarrow{P}\alpha_* \tag{34}$$

Next, we derive the first-order optimality conditions for the scalar minimax problem.

First-order optimality conditions

$$\begin{aligned}\frac{\partial D(\alpha,u)}{\partial\alpha}=&-\frac{u(1-\delta)}{2\sqrt{\delta}}-\frac{u\sqrt{\delta}\sigma^2}{2\alpha^2}+\frac{1}{\sqrt{\delta}}\mathbb{E}[\frac{H}{\Lambda}\mathcal{M}'_{\psi,1}(B+\frac{\alpha}{\sqrt{\delta}\Lambda}H;\frac{\alpha}{u\sqrt{\delta}\Lambda^2})]\\&+\frac{1}{u\sqrt{\delta}}\mathbb{E}[\frac{1}{\Lambda^2}\mathcal{M}'_{\psi,2}(B+\frac{\alpha}{\sqrt{\delta}\Lambda}H;\frac{\alpha}{u\sqrt{\delta}\Lambda^2})]=0\end{aligned} \tag{35}$$

$$\frac{\partial D(\alpha,u)}{\partial u}=-\frac{\alpha(1-\delta)}{2\sqrt{\delta}}+\frac{\sqrt{\delta}\sigma^2}{2\alpha}-\frac{\alpha}{u^2\sqrt{\delta}}\mathbb{E}[\frac{1}{\Lambda^2}\mathcal{M}'_{\psi,2}(B+\frac{\alpha}{\sqrt{\delta}\Lambda}H;\frac{\alpha}{u\sqrt{\delta}\Lambda^2})]=0 \tag{36}$$

Using properties of Moreau Envelope (Proposition 8), we have

$$\mathbb{E}[\frac{1}{\Lambda^2}\mathcal{M}'_{\psi,2}(B+\frac{\alpha}{\sqrt{\delta}\Lambda}H;\frac{\alpha}{u\sqrt{\delta}\Lambda^2})]=-\frac{1}{2}\mathbb{E}[\left(\frac{1}{\Lambda}\mathcal{M}'_{\psi,1}(B+\frac{\alpha}{\sqrt{\delta}\Lambda}H;\frac{\alpha}{u\sqrt{\delta}\Lambda^2})\right)^2] \tag{37}$$

Using the above inequality, we can derive the following optimality conditions

$$\mathbb{E}[\frac{H}{\Lambda}\mathcal{M}'_{\psi,1}(B+\frac{\alpha}{\sqrt{\delta}\Lambda}H;\frac{\alpha}{u\sqrt{\delta}\Lambda^2})]=u(1-\delta) \tag{38}$$

$$\mathbb{E}[\left(\frac{1}{\Lambda}\mathcal{M}'_{\psi,1}(B+\frac{\alpha}{\sqrt{\delta}\Lambda}H;\frac{\alpha}{u\sqrt{\delta}\Lambda^2})\right)^2]=u^2(1-\delta)-\frac{\delta\sigma^2u^2}{\alpha^2} \tag{39}$$

Distributional convergence

Next, to show distributional convergence, we first assume weak convergence of the solution of the AO, i.e.

$$\hat{\boldsymbol{\beta}}^{AO}_{n,i}=\text{prox}_\psi(\beta_i^*+\frac{\alpha_n}{\sqrt{\delta}\Sigma_{i,i}^{1/2}}h_i;\frac{\alpha_n}{u_n\sqrt{\delta}\Sigma_{i,i}})\xrightarrow{D}\text{prox}_\psi(B+\frac{\alpha_\psi}{\sqrt{\delta}\Lambda}H;\frac{\alpha_\psi}{u_\psi\sqrt{\delta}\Lambda^2}) \tag{40}$$

and we want to show that

$$\hat{\boldsymbol{\beta}}^{PO}_{n,i}\xrightarrow{D}\text{prox}_\psi(B+\frac{\alpha_\psi}{\sqrt{\delta}\Lambda}H;\frac{\alpha_\psi}{u_\psi\sqrt{\delta}\Lambda^2}) \tag{41}$$

Define

$$F_n(\hat{\boldsymbol{\beta}}_n, \boldsymbol{\beta}^*, \Sigma) := \frac{1}{n}\sum_{i=1}^{n} f(\hat{\beta}_{n,i}, \beta_i^*, \Sigma_{i,i}) \tag{42}$$

where f is any bounded Lipschitz function. Also, define the limit

$$\kappa := \mathbb{E}[f(X(B, \Lambda, H), B, \Lambda^2)] \tag{43}$$

For any fixed $\epsilon > 0$, define the set

$$\mathcal{S}_\epsilon = \mathcal{S}_\epsilon(\boldsymbol{\beta}^*, \boldsymbol{\Sigma}) = \{\boldsymbol{w} = \frac{1}{\sqrt{n}}\Sigma^{\frac{1}{2}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \in \mathcal{W} \text{ s.t } |F_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*, \Sigma) - \kappa| > 2\epsilon\} \tag{44}$$

Consider the following perturbed PO and AO problems

$$\Phi_{\mathcal{S}_\epsilon}(\boldsymbol{G}) = \min_{\boldsymbol{w}\in\mathcal{S}_\epsilon} \psi(\boldsymbol{\beta}^* + \sqrt{n}\Sigma^{-1/2}\boldsymbol{w}) \quad \text{subject to} \quad \boldsymbol{G}\boldsymbol{w} = \sigma\boldsymbol{z} \tag{45}$$

and

$$\phi_{\mathcal{S}_\epsilon}(\boldsymbol{g}, \boldsymbol{h}) = \min_{\boldsymbol{w}\in\mathcal{S}_\epsilon} \psi(\boldsymbol{\beta}^* + \sqrt{n}\Sigma^{-1/2}\boldsymbol{w}) \quad \text{subject to} \quad \|\boldsymbol{g}\|\sqrt{\|\boldsymbol{w}\|^2 + \sigma^2} - \boldsymbol{h}^T\boldsymbol{w} - \sigma h \le 0 \tag{46}$$

Using [56] Theorem 6.1(iii), it is sufficient to show existence of constants $\bar{\phi}, \bar{\phi}_{\mathcal{S}_\epsilon}$ and $\eta > 0$ satisfying

1. $\bar{\phi}_{\mathcal{S}_\epsilon} \ge \bar{\phi} + 3\eta$
2. $\phi_(\boldsymbol{g}, \boldsymbol{h}) \le \bar{\phi} + \eta$ with probability approaching 1.
3. $\phi_{\mathcal{S}_\epsilon}(\boldsymbol{g}, \boldsymbol{h}) \ge \bar{\phi}_{\mathcal{S}_\epsilon} - \eta$ with probability approaching 1.

to prove that $\hat{\boldsymbol{w}}_n \notin \mathcal{S}_\epsilon$ wpa 1.

Condition 2 Choose $\bar{\phi} = D(\alpha_*, u_*)$, we have shown that $\phi(\boldsymbol{g}, \boldsymbol{h}) \xrightarrow{P} \bar{\phi}$. So for any $\eta > 0$, we have that

$$\bar{\phi} + \eta \ge \phi(\boldsymbol{g}, \boldsymbol{h}) \ge \bar{\phi} - \eta \tag{47}$$

Condition 3 Let $c(\boldsymbol{w}) := \|\boldsymbol{g}\|\sqrt{\|\boldsymbol{w}\|^2 + \sigma^2} - \boldsymbol{h}^T\boldsymbol{w} - \sigma h$, clearly c is strictly convex in $\boldsymbol{w}$. At the optimum $\hat{\boldsymbol{w}}_n^{AO}$, we have that

$$-\lambda\nabla_{\boldsymbol{w}} c(\hat{w}_n^{AO}) \in \partial_{\boldsymbol{w}}\Psi(\boldsymbol{\beta}^* + \sqrt{n}\Sigma^{-1/2}\hat{\boldsymbol{w}}_n^{AO}) \tag{48}$$

for $\lambda \ge 0$ and also by feasibility, we have $c(\hat{\boldsymbol{w}}_n^{AO}) = 0$; otherwise, we can always move along the negative gradient of the objective to reduce the objective value of the objective assuming that $\partial_{\boldsymbol{w}}\Psi(\boldsymbol{\beta}^* + \sqrt{n}\Sigma^{-1/2}\hat{\boldsymbol{w}}_n^{AO})\backslash\{\boldsymbol{0}\}$ is non-empty, which is true when ψ has an unique minimizer. Next, we argue that in the new constrained formulation $\mathcal{S}_\epsilon \cap \{\boldsymbol{w} \text{ s.t } c(\boldsymbol{w}) \le 0\}$, if $\|\boldsymbol{w} - \hat{\boldsymbol{w}}_n^{AO}\| \ge \tilde{\epsilon}$, then the value of objective increases. By convexity of objective and optimality of $\hat{\boldsymbol{w}}_n^{AO}$, we have that

$$\Psi(\boldsymbol{\beta}^* + \sqrt{n}\Sigma^{-1/2}\boldsymbol{w}) \ge \phi(\boldsymbol{g}, \boldsymbol{h}) - \lambda\nabla_{\boldsymbol{w}} c(\hat{w}_n^{AO})^T(\boldsymbol{w} - \hat{\boldsymbol{w}}_n^{AO}) \tag{49}$$

Case 1 If $\nabla_{\boldsymbol{w}} c(\hat{w}_n^{AO})^T(\boldsymbol{w} - \hat{\boldsymbol{w}}_n^{AO}) < 0$, then objective increases.
Case 2 If $\nabla_{\boldsymbol{w}} c(\hat{w}_n^{AO})^T(\boldsymbol{w} - \hat{\boldsymbol{w}}_n^{AO}) \ge 0$, then

$$c(\boldsymbol{w}) > c(\hat{\boldsymbol{w}}_n^{AO}) + \nabla_{\boldsymbol{w}} c(\hat{w}_n^{AO})^T(\boldsymbol{w} - \hat{\boldsymbol{w}}_n^{AO}) \tag{50}$$

and the inequality is strict due to strict convexity if the constraint function and therefore $c(\boldsymbol{w}) > 0$, therefore its not feasible. So the feasible option is Case 1 and there exists a constant $\lambda \nabla_{\boldsymbol{w}} c(\hat{w}_n^{AO})^T (\boldsymbol{w} - \hat{\boldsymbol{w}}_n^{AO}) > q(\tilde{\epsilon}) > 0$ such that

$$\Psi(\boldsymbol{\beta}^* + \sqrt{n}\Sigma^{-1/2}\boldsymbol{w}) > \phi(\boldsymbol{g}, \boldsymbol{h}) + q(\tilde{\epsilon}) \tag{51}$$

which implies that with probability approaching 1, we have

$$\Psi(\boldsymbol{\beta}^* + \sqrt{n}\Sigma^{-1/2}\boldsymbol{w}) > \bar{\phi} + q(\tilde{\epsilon}) - \eta \tag{52}$$

Next, we argue that for small enough η, we have **Condition 1**, which is equivalent to showing that

$$\bar{\phi} + q(\tilde{\epsilon}) - \eta - (\bar{\phi} + \eta) \geq \eta \tag{53}$$

Choosing $\eta = \frac{q(\tilde{\epsilon})}{4}$ and $\bar{\phi}_{\mathcal{S}_\epsilon} = \bar{\phi} + q(\tilde{\epsilon})$, satisfies the above inequality.

Next, we show $\|\boldsymbol{w} - \hat{w}_n^{AO}\| \geq \tilde{\epsilon}$. By definition, we have

$$|F_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*, \Sigma) - \kappa| > 2\epsilon \tag{54}$$

We have already shown that

$$|F_n(\hat{\boldsymbol{\beta}}_n^{AO}, \boldsymbol{\beta}^*, \Sigma) - \kappa| \leq \epsilon \tag{55}$$

By Lipschitzness and Cauchy-Schwarz, we get

$$|F_n(\hat{\boldsymbol{\beta}}_n^{AO}, \boldsymbol{\beta}^*, \Sigma) - F_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*, \Sigma)| \leq C\|\boldsymbol{w} - \hat{w}_n^{AO}\| \tag{56}$$

$$\begin{aligned} 2\epsilon &\leq |F_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*, \Sigma) - \kappa| + |F_n(\hat{\boldsymbol{\beta}}_n^{AO}, \boldsymbol{\beta}^*, \Sigma) - F_n(\boldsymbol{\beta}, \boldsymbol{\beta}^*, \Sigma)| \\ &\leq \epsilon + C\|\boldsymbol{w} - \hat{w}_n^{AO}\| \end{aligned}$$

which implies $\|\boldsymbol{w} - \hat{w}_n^{AO}\| \geq \tilde{\epsilon}$

■

Appendix F. Fundamental limits

F.1. Proof of Theorem 2

Theorem 13 *(Lower bound on α_ψ^2) Let Assumptions 1,2, and 3 hold. Define the random variable $V_\alpha = B + \frac{\alpha}{\sqrt{\delta}\Lambda} H$ where $H \sim \mathcal{N}(0,1)$ and B, Λ as defined in assumption 2 and 3. Let α_* be the unique solution of the following non-linear equation*

$$\alpha^2 = \frac{\delta\sigma^2}{1-\delta} + \frac{\delta(1-\delta)}{\mathcal{I}_\Lambda(V_\alpha)} \tag{57}$$

where $\mathcal{I}_\Lambda(V_\alpha)$ is the weighted fisher information of V_α defined as

$$\mathcal{I}_\Lambda(V_\alpha) := \mathbb{E}\left[\left(\frac{\xi_{V_\alpha}(V_\alpha|\Lambda)}{\Lambda}\right)^2\right] \tag{58}$$

where $\xi_{V_\alpha}(v|\Lambda) := \frac{p'_{V_\alpha}(v|\Lambda)}{p_{V_\alpha}(v|\Lambda)}$ is the conditional score function of V_α. Then for every $\Psi \in \mathcal{C}_\psi$, with α_ψ^2 as the asymptotic limit of the generalization error as in (7), we have $\alpha_\psi^2 \geq \alpha_*^2$.

Proof Recall the system of non-linear equations and let (α_Ψ, u_Ψ) be a solution

$$\mathbb{E}\left[\frac{H}{\Lambda}\mathcal{M}'_{\psi,1}(B+\frac{\alpha}{\sqrt{\delta}\Lambda}H;\frac{\alpha}{u\sqrt{\delta}\Lambda^2})\right] = u(1-\delta) \tag{59a}$$

$$\mathbb{E}\left[\left(\frac{1}{\Lambda}\mathcal{M}'_{\psi,1}(B+\frac{\alpha}{\sqrt{\delta}\Lambda}H;\frac{\alpha}{u\sqrt{\delta}\Lambda^2})\right)^2\right] = u^2(1-\delta) - \frac{\delta\sigma^2u^2}{\alpha^2} \tag{59b}$$

and let α_* be the solution of (57). First, we argue that the solution α_* always exists when $\sigma > 0$. Consider the function

$$h(\alpha) = \frac{\delta\sigma^2}{\alpha^2(1-\delta)} + \frac{\delta(1-\delta)}{\alpha^2\mathcal{I}_\Lambda(V_\alpha)}$$

Therefore, at α_*, we have $h(\alpha_*) = 1$. Note that $h(\alpha)$ is continuous on $\mathbb{R}_{>0}$ and when $\sigma > 0$, we have that $\lim_{\alpha\to 0^+} h(\alpha) = \infty$ and $\lim_{\alpha\to\infty} h(\alpha) = 1-\delta < 1$ using the fact that $\lim_{\alpha\to\infty}\alpha^2\mathcal{I}_\Lambda(V_\alpha) = \delta$ using properties from Proposition 12. Therefore, by the mean value theorem, we can argue that the existence of α_*. To show the uniqueness of α_*, we need to show that $h(\alpha)$ is monotonically decreasing. Using properties of Fisher information $\alpha^2\mathcal{I}_\Lambda(V_\alpha) = \mathcal{I}_\Lambda(\frac{V_\alpha}{\alpha})$ and one can verify that $\mathcal{I}_\Lambda(\frac{V_\alpha}{\alpha})$ is monotonically increasing using Proposition 11 (c), as a consequence $h(\alpha)$ is monotonically decreasing. Now that we have established the existence and uniqueness of α_*, we next show that α_* is a lower on α_Ψ.

Consider the following integral

$$\frac{\sqrt{\delta}}{\alpha}\mathbb{E}[\frac{\alpha H}{\sqrt{\delta}\Lambda}\mathcal{M}'_{\psi,1}(B+\frac{\alpha}{\sqrt{\delta}\Lambda}H;\frac{\alpha}{u\sqrt{\delta}\Lambda^2})] = \frac{\sqrt{\delta}}{\alpha}\iiint g\mathcal{M}'_{\psi,1}(b+g;\frac{\alpha}{u\sqrt{\delta}\lambda^2})p_B(b)p_G(g|\lambda)p_\Lambda(\lambda)\mathrm{d}b\mathrm{d}g\mathrm{d}\lambda \tag{60}$$

where G is a conditionally Gaussian random variable $p_G(.|\lambda) \sim \mathcal{N}(0, \frac{\alpha^2}{\delta\lambda^2})$. Next, we use the following using the property of Gaussian density

$$p'_G(g|\lambda) = -g\frac{\delta\lambda^2}{\alpha^2}p_G(g|\lambda) \tag{61}$$

Plugging back, we get

$$-\frac{\alpha}{\sqrt{\delta}}\iiint\frac{1}{\lambda^2}\mathcal{M}'_{\psi,1}(b+g;\frac{\alpha}{u\sqrt{\delta}\lambda^2})p'_G(g|\lambda)p_B(b)p_\Lambda(\lambda)\mathrm{d}b\mathrm{d}g\mathrm{d}\lambda \tag{62}$$

If consider, the following change of variable $v = b+g$, then $\mathrm{d}v = \mathrm{d}b$ and

$$-\frac{\alpha}{\sqrt{\delta}}\iiint\frac{1}{\lambda^2}\mathcal{M}'_{\psi,1}(v;\frac{\alpha}{u\sqrt{\delta}\lambda^2})p'_G(g|\lambda)p_B(v-g)p_\Lambda(\lambda)\mathrm{d}v\mathrm{d}g\mathrm{d}\lambda \tag{63}$$

Integrating g out using Gaussian Integration of parts, one can verify that

$$\int_{-\infty}^{\infty}p'_G(g|\lambda)p_B(v-g) = p'_V(v|\lambda) \tag{64}$$

Plugging the back, we have

$$-\frac{\alpha}{\sqrt{\delta}}\iint \frac{1}{\lambda^2}\mathcal{M}'_{\psi,1}(v;\frac{\alpha}{u\sqrt{\delta}\lambda^2})p'_V(v|\lambda)p_\Lambda(\lambda)\mathrm{d}v\mathrm{d}g\mathrm{d}\lambda = -\frac{\alpha}{\sqrt{\delta}}\mathbb{E}[\frac{1}{\lambda}\mathcal{M}'_{\psi,1}(v;\frac{\alpha}{u\sqrt{\delta}\lambda^2})\cdot\frac{p'_V(v|\lambda)}{\lambda p_V(v|\lambda)}] \tag{65}$$

Therefore, we have essentially proved the following identity

$$\frac{\sqrt{\delta}}{\alpha}\mathbb{E}[\frac{\alpha H}{\sqrt{\delta}\Lambda}\mathcal{M}'_{\psi,1}(B+\frac{\alpha}{\sqrt{\delta}\Lambda}H;\frac{\alpha}{u\sqrt{\delta}\Lambda^2})] = -\frac{\alpha}{\sqrt{\delta}}\mathbb{E}[\frac{1}{\lambda}\mathcal{M}'_{\psi,1}(v;\frac{\alpha}{u\sqrt{\delta}\lambda^2})\cdot\frac{p'_V(v|\lambda)}{\lambda p_V(v|\lambda)}] \tag{66}$$

Using Cauchy Schwartz inequality, we have that

$$\mathbb{E}[\frac{1}{\lambda}\mathcal{M}'_{\psi,1}(v;\frac{\alpha}{u\sqrt{\delta}\lambda^2})\cdot\frac{p'_V(v|\lambda)}{\lambda p_V(v|\lambda)}]^2 \le \mathbb{E}[(\frac{1}{\lambda}\mathcal{M}'_{\psi,1}(v;\frac{\alpha}{u\sqrt{\delta}\lambda^2}))^2]\mathbb{E}[(\frac{p'_V(v|\lambda)}{\lambda p_V(v|\lambda)})^2] \tag{67}$$

If we let $\mathcal{I}_\Lambda(V_\alpha) := \mathbb{E}\left[\left(\frac{\xi_{V_\alpha}(V_\alpha|\Lambda)}{\Lambda}\right)^2\right]$. Using the optimality conditions, we can show that the following inequality holds

$$\frac{u^2\delta(1-\delta)^2}{\alpha^2} \le \left(u^2(1-\delta) - \frac{\delta\sigma^2u^2}{\alpha^2}\right)\mathcal{I}_\Lambda(V_\alpha) \tag{68}$$

Note that the inequality is independent of Ψ and is true for every (α_ψ, u_ψ) for which the system of equations is satisfied. Simplifying the above inequality, we get

$$u^2\left(\alpha^2(1-\delta)\mathcal{I}_\Lambda(V_\alpha) - \delta\sigma^2\mathcal{I}_\Lambda(V_\alpha) - \delta(1-\delta)^2\right) \ge 0 \tag{69}$$

One can verify that $u > 0$ since it's a Lagrange multiplier of an active constraint. Eliminating u gives the following inequality

$$\alpha^2(1-\delta)\mathcal{I}_\Lambda(V_\alpha) - \delta\sigma^2\mathcal{I}_\Lambda(V_\alpha) - \delta(1-\delta)^2 \ge 0 \tag{70}$$

Writing the inequality in terms of $h(\alpha)$, we arrive at

$$1 \ge h(\alpha) \tag{71}$$

As we have previously established, $h(\alpha)$ is monotonically decreasing and α_* is the unique solution of $h(\alpha_*) = 1$. Therefore α_* is the smallest α that satisfies the above inequality, and since the above inequality holds for every convex potential Ψ whose optimality conditions are given by (59), we have that $\alpha_\Psi \ge \alpha_*$.

■

F.2. Proof of Corollary 3

Corollary 14 *Let Assumptions 1,2 and 3 hold and α_* be defined as the solution to (9), if $\Lambda = 1$ almost surely then*

$$\alpha_*^2 \ge \frac{\sigma^2}{1-\delta} + \frac{1-\delta}{\mathcal{I}(B)} \tag{72}$$

whenever, the Fisher information $\mathcal{I}(B)$ is well defined. The inequality becomes an equality if and only if B is a Gaussian.

Proof From the previous theorem, we have that

$$\alpha_*^2 = \frac{\delta\sigma^2}{1-\delta} + \frac{\delta(1-\delta)}{\mathcal{I}_\Lambda(V_{\alpha_*})} \tag{73}$$

If $\Lambda = 1$ a.s and $\mathcal{I}(B)$ is well defined, then

$$\mathcal{I}_\Lambda(V_{\alpha_*}) = \mathcal{I}(B + \frac{\alpha_*}{\sqrt{\delta}}H) \leq \frac{\mathcal{I}(B)}{1 + \frac{\alpha_*^2}{\delta}\mathcal{I}(B)} \tag{74}$$

where the inequality is obtained from Stam's inequality and is strict when B is a Gaussian. Plugging back, we have that

$$\alpha_*^2 \geq \frac{\delta\sigma^2}{1-\delta} + \frac{(1-\delta)(\delta + \alpha_*^2\mathcal{I}(B))}{\mathcal{I}(B)} \tag{75}$$

Re-arranging the terms gives us the desired lower bound

$$\alpha_*^2 \geq \frac{\sigma^2}{1-\delta} + \frac{1-\delta}{\mathcal{I}(B)} \tag{76}$$

■

F.3. Proof of Theorem 4

Theorem 15 *(Optimal Ψ) Let Assumptions 1, 2, and 3 hold and α_* be defined as the solution to (9). Consider the following function $\psi_* : \mathbb{R}^2 \to \mathbb{R}$*

$$\psi_*(v,\lambda) := -\mathcal{M}_{\log(P_{V_{\alpha_*}}(v|\lambda))}\left(v; \frac{\alpha_*^2(1-\delta) - \delta\sigma^2}{\delta(1-\delta)\lambda^2}\right), \tag{77}$$

if $P_{V_{\alpha_*}}(v|\lambda)$ is log-concave in v and we define $\Psi_*(\boldsymbol{\beta}) = \sum_{i=1}^n \psi_*(\beta_i, \Sigma_{i,i})$, then

1. *$\Psi_*(\boldsymbol{\beta}) \in \mathcal{C}_\psi$*
2. *α_* is a solution to the system of equations (6) obtained using $\psi^*(v,\lambda)$ and is therefore the optimal convex implicit bias.*

Proof By Proposition 9, we have that

$$-\mathcal{M}_{\log(P_{V_{\alpha_*}}(v|\lambda))}\left(v; \frac{\alpha_*^2(1-\delta) - \delta\sigma^2}{\delta(1-\delta)\lambda^2}\right) = \frac{\delta(1-\delta)\lambda^2}{\alpha_*^2(1-\delta) - \delta\sigma^2}\left(\left(\frac{v^2}{2} + \frac{\alpha_*^2(1-\delta) - \delta\sigma^2}{\delta(1-\delta)\lambda^2}\log P_{V_{\alpha_*}}(v|\lambda)\right)^* - \frac{v^2}{2}\right) \tag{78}$$

Showing $\psi_*(v,\lambda)$ is convex is equivalent to showing $\left(\left(\frac{v^2}{2} + \frac{\alpha_*^2(1-\delta)-\delta\sigma^2}{\delta(1-\delta)\lambda^2}\log P_{V_{\alpha_*}}(v|\lambda)\right)^* - \frac{v^2}{2}\right)$ is convex. First, we will verify that $\frac{v^2}{2} + \frac{\alpha_*^2(1-\delta)-\delta\sigma^2}{\delta(1-\delta)\lambda^2}\log P_{V_{\alpha_*}}(v|\lambda)$ is convex. By definition

$$\log P_{V_{\alpha_*}}(v|\lambda) = -\frac{\delta\lambda^2 v^2}{2\alpha_*^2} + \log\int_{-\infty}^{\infty}\exp\left(\delta\lambda^2(2vb - b^2)/2\alpha_*^2\right)P_B(b)db + c \tag{79}$$

for some constant c. One can verify the convexity of $\log \int_{-\infty}^{\infty} \exp\left(\delta\lambda^2(2vb - b^2)/2\alpha_*^2\right) P_B(b)db$ by double differentiation with respect to v. Therefore, it sufficient if $\frac{v^2}{2} - \frac{\alpha_*^2(1-\delta)-\delta\sigma^2}{\alpha_*^2(1-\delta)}\frac{v^2}{2}$ is convex, which is trivially true. Therefore, we have now verified the convexity of $\frac{v^2}{2} + \frac{\alpha_*^2(1-\delta)-\delta\sigma^2}{\delta(1-\delta)\lambda^2}\log P_{V_{\alpha_*}}(v|\lambda)$.

Next, we'll use the following property of the derivatives of convex conjugates from [48] (Cor. 23.5.1), which says that if $f(x)$ is convex, then

$$(f^*)'(x) = (f')^{-1}(x) \tag{80}$$

Using the property of the derivative of an inverse, we further have that

$$(f^*)''(x) = \frac{1}{f''\left((f')^{-1}(x)\right)} \tag{81}$$

Using the above property, taking double derivative of $\psi_*(v,\lambda)$ gives us

$$\psi_*''(v,\lambda) = c_*\left(\frac{1}{1 + c_*(\log P_{V_{\alpha_*}}(v|\lambda))''(g(v))} - 1\right) \tag{82}$$

where $g(v) := (v + c_*(\log P_{V_{\alpha_*}}(g(v)|\lambda))')^{-1}(v)$. Since $P_{V_{\alpha_*}}(v|\lambda)$ is log-concave by assumption, $(\log P_{V_{\alpha_*}}(v|\lambda))''(v) \leq 0$ for all v. This implies that $\psi_*''(v,\lambda) \geq 0$, finishing the proof on the sufficient condition.

Next, we verify optimality conditions (59) to prove that the optimal convex potential is given by Theorem 4. Let α_* be the solution to (57). Now, consider the following candidate for optimal potential

$$\mathcal{M}'_{\psi_*,1}(v; \frac{\alpha_*^2(1-\delta) - \delta\sigma^2}{\delta(1-\delta)\lambda^2}) = -\frac{P'_{V_{\alpha_*}}(v|\lambda)}{P_{V_{\alpha_*}}(v|\lambda)} \tag{83}$$

We now show that α_* and $u_* := \frac{\alpha_*\sqrt{\delta}(1-\delta)}{\alpha_*^2(1-\delta)-\delta\sigma^2}$ satisfy, the optimality conditions for the above candidate. Plugging in (59)(b), we get

$$\mathbb{E}\left[\left(\frac{1}{\Lambda}\mathcal{M}'_{\psi_*,1}(B + \frac{\alpha_*}{\sqrt{\delta}\Lambda}H; \frac{\alpha_*}{u_*\sqrt{\delta}\Lambda^2})\right)^2\right] = \mathcal{I}_\Lambda(V_{\alpha_*}) \tag{84}$$

$$= u_*^2\frac{(\alpha_*^2(1-\delta) - \delta\sigma^2)^2}{\alpha_*^2\delta(1-\delta)^2}\mathcal{I}_\Lambda(V_{\alpha_*}) \tag{85}$$

$$= u_*^2(1-\delta) - \frac{\delta\sigma^2 u_*^2}{\alpha_*^2} \tag{86}$$

where the second equality is from the definition of u_* and the third equality is from (57). Next, we verify (59)(a)

$$\mathbb{E}\left[\frac{H}{\Lambda}\mathcal{M}'_{\psi,1}(B + \frac{\alpha_*}{\sqrt{\delta}\Lambda}H; \frac{\alpha_*}{u_*\sqrt{\delta}\Lambda^2})\right] = -\frac{\alpha_*}{\sqrt{\delta}}\mathbb{E}[\frac{1}{\lambda}\mathcal{M}'_{\psi,1}(v; \frac{\alpha_*}{u_*\sqrt{\delta}\lambda^2}) \cdot \frac{p'_V(v|\lambda)}{\lambda p_V(v|\lambda)}] \tag{87}$$

$$= \frac{\alpha_*}{\sqrt{\delta}}\mathcal{I}_\Lambda(V_{\alpha_*}) \tag{88}$$

$$= u_*(1-\delta) \tag{89}$$

where the first equality is due to the identity (66) and the rest follows from definitions of u_* and $\mathcal{I}_\Lambda(V_{\alpha_*})$. Now that we have shown that (83) satisfies optimality conditions, taking anti derivative of (83) gives $\mathcal{M}_{\psi_*}(v; \frac{\alpha_*^2(1-\delta)-\delta\sigma^2}{\delta(1-\delta)\lambda^2}) = -\log P_{V_{\alpha_*}}(v|\lambda)$ which can be inverted to give us the optimal convex potential whenever the $P_{V_{\alpha_*}}$ is log-concave by Proposition 10.

■

F.4. Proof Corollary 6

Corollary 16 *(Ψ_* for Gaussian B) Let Assumptions 1,2 and 3 hold and $B \sim \mathcal{N}(0,1)$, then the optimal implicit bias is given as*

$$\Psi_*(\boldsymbol{\beta}) = \boldsymbol{\beta}^T \boldsymbol{\Sigma}^{1/2} \left(\frac{\sigma^2}{1-\delta} \boldsymbol{I}_n + \boldsymbol{\Sigma} \right)^{-1} \boldsymbol{\Sigma}^{1/2} \boldsymbol{\beta}. \tag{90}$$

Proof When $B \sim \mathcal{N}(0,1)$, then $P_{V_{\alpha_*}}(v|\lambda)$ is Gaussian density

$$P_{V_{\alpha_*}}(v|\lambda) = c_0 \exp\left(-\frac{v^2}{2(1+\frac{\alpha_*^2}{\delta\lambda^2})} \right) \tag{91}$$

where c_0 is a constant independent of v. Note that $P_{V_{\alpha_*}}(v|\lambda)$ is log-concave, by Theorem 4, we have that the optimal potential is given as

$$\psi_*(v,\lambda) = -\mathcal{M}_{\log(P_{V_{\alpha_*}}(v|\lambda))}\left(v; \frac{\alpha_*^2(1-\delta)-\delta\sigma^2}{\delta(1-\delta)\lambda^2} \right) \tag{92}$$

Solving the Moreau envelope gives us

$$\psi_*(v,\lambda) = \frac{\lambda^2 v^2}{2(\lambda^2 + \frac{\delta\sigma^2}{1-\delta})} + c_1 \tag{93}$$

Here, c_1 is independent of v, so we can ignore it. In the vectorized form, $\Psi_*(\boldsymbol{\beta}) = \sum_{i=1}^n \psi_*(\beta_i, \Sigma_{i,i})$ gives the desired result.

$$\Psi_*(\boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\beta}^T \boldsymbol{\Sigma}^{1/2} \left(\frac{\sigma^2}{1-\delta} \boldsymbol{I}_n + \boldsymbol{\Sigma} \right)^{-1} \boldsymbol{\Sigma}^{1/2} \boldsymbol{\beta}. \tag{94}$$

■