Generalization analysis on linear regression

Eva Pradhan

Benign Overfitting of Constant-Stepsize SGD for Linear Regression

- Difan Zou : Department of Computer Science University of California, Los Angeles
- Jingfeng Wu : Department of Computer Science Johns Hopkins University, Baltimore
- Vladimir Braverman : Department of Computer Science Johns Hopkins University, Baltimore
- Quanquan Gu : Department of Computer Science University of California, Los Angeles
- Sham M. Kakade : Department of Computer Science University of Washington, Seattle & Microsoft Research, Seattle

GRADIENT DESCENT

“Gradient descent is an iterative algorithm, that starts from a random point on a function and travels down its slope in steps until it reaches the lowest point of that function.”

- Gradient Descent is a generic optimization algorithm capable of finding optimal solutions to a wide range of problems
- The general idea is to tweak parameters iteratively in order to minimize the cost function

STEPS OF THE GRADIENT DESCENT ALGORITHM IN MACHINE LEARNING

- Find the slope of the objective function **with respect to each parameter/feature**

OR

Compute the gradient of the function

- Pick a random initial value for the parameters. (E.g. In a parabola, differentiate “y” with respect to “x”. If we have more than one features like x1, x2 etc., we take the partial derivative of “y” with respect to each of the features.)
- Update the gradient function by plugging in the parameter values.
- Calculate the step sizes for each feature as : **step size = gradient * learning rate.**
- Calculate the new parameters as : **new params = old params -step size**
- Repeat above 3 steps until gradient is almost 0.

Ideally learning rate should be small so that it doesn’t jump over the minima E.g. 0.01, but also not too large that convergence takes long

TYPES OF GRADIENT DESCENT

- Batch Gradient Descent
- Stochastic Gradient Descent
- Mini-batch Gradient Descent

STOCHASTIC GRADIENT DESCENT

- Gradient Descent can end up involving expensive amount of calculations
- ‘*stochastic*‘ : a system or process linked with a random probability.
- This problem is solved by Stochastic Gradient Descent
- SGD uses only a single sample, i.e., a batch size of one, to perform each iteration. The sample is randomly shuffled and selected for performing the iteration.
- We can induce randomness while selecting data points at each step to calculate the derivatives. SGD randomly picks one data point from the whole data set at each iteration to reduce the computations enormously.
- It is also common to sample a small number of data points instead of just one point at each step and that is called “mini-batch” gradient descent. Mini-batch tries to strike a balance between the goodness of gradient descent and speed of SGD.

Introduction

- Explores the issue of overfitting in machine learning
- Proposes constant step size Stochastic Gradient Descent (SGD) for linear regression in overparameterized settings.
- While this phenomenon has been observed in many modern machine learning models, it is not well understood in the context of SGD.
- The authors provide a sharp excess risk bound that reveals a bias-variance decomposition, characterizing when generalization is possible.
- They demonstrate the sharpness of the established bound by proving a matching lower bound for SGD with iterate averaging and show the advantage of SGD with tail averaging over iterate averaging.
- The authors also reflect on the differences between the algorithmic regularization afforded by SGD compared to ordinary least squares and ridge regression
- Provide experimental results on synthetic data to support their findings.

Introduction Continued

- The paper use constant stepsize SGD with iterate averaging to investigate this phenomenon and provide evidence that benign overfitting occurs in overparameterized linear regression, even with a constant stepsize.
- They show that this occurs even for simple linear models and provide experimental results to support their claims.

ITERATE AVERAGING

- A variant of the classic Polyak–Ruppert averaging scheme, broadly used in stochastic gradient methods. Rather than a uniform average of the iterates, considers a weighted average, with weights decaying in a geometric fashion.
- In the context of linear least-squares regression, it shows that this averaging scheme has the same regularizing effect, and indeed is asymptotically equivalent, to ridge regression.
- In particular, derives finite-sample bounds for the proposed approach that match the best known results for regularized stochastic gradient methods.

TAIL AVERAGING

- Tail averaging consists in averaging the last examples in a stream.
- Common techniques either have a memory requirement which grows with the number of samples to average, are not available at every timestep or do not accommodate growing windows.

EXCESS RISK

- The difference between the risk of a given function and the minimum possible risk over a function class.

BIAS VARIANCE DECOMPOSITION

- The bias–variance decomposition is a way of analyzing a learning algorithm's expected generalization error with respect to a particular problem as a sum of three terms, the bias, variance, and a quantity called the irreducible error, resulting from noise in the problem itself.

Introduction And Related Work

- The paper discusses the phenomenon of modern machine learning models achieving near-zero training error while still being able to generalize effectively.
- This is observed even in overparameterized and under-regularized settings. The paper aims to understand this effect in the context of stochastic gradient descent (SGD) for least squares regression in the overparameterized regime.
- While there is a growing amount of work studying generalization in basic linear models, the algorithmic aspects of generalization for SGD are not well understood.
- In the under parameterized case, iterate averaged SGD has been shown to achieve the optimal rate, but there is very less work in the overparameterized case.
- The paper seeks to fill this gap

SGD FOR LINEAR REGRESSION

$$\min_{\mathbf{w}} L(\mathbf{w}), \text{ where } L(\mathbf{w}) = \frac{1}{2}\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(y - \langle \mathbf{w}, \mathbf{x}\rangle)^2\right],$$

$\mathbf{x} \in \mathcal{H}$, is the feature vector, $y \in \mathbb{R}$ is the response; $\mathbf{w} \in \mathcal{H}$ is the weight vector to be optimized.

$\mathcal{H}$ Some finite d-dimensional or infinite dimensional Hilbert space

$\mathcal{D}$ Unknown distribution over x and y

overparameterized regime, where $d \gg N$ (or possibly countably infinite).

Weight is updated according to SGD as follows

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \gamma\left(y_t - \langle \mathbf{w}_{t-1}, \mathbf{x}_t\rangle\right)\mathbf{x}_t, \qquad t = 1, \ldots, N,$$

$\gamma > 0$ is a constant stepsize, N is the number of samples observed,

Weight initialized as $\mathbf{w}_0 \in \mathcal{H}.$

The final output is the average of the iterates

$$\overline{\mathbf{w}}_N := \frac{1}{N}\sum_{t=0}^{N-1} \mathbf{w}_t.$$

THE OUTPUTS

- The bound is stated in a general manner, in terms of the full eigen spectrum of the data covariance matrix along with a functional dependency on the initial iterate
- The paper shows how the benign-overfitting phenomenon can be observed for SGD, provided certain spectrum conditions of decay on the data covariance are met

Introduction And Related Work Continued

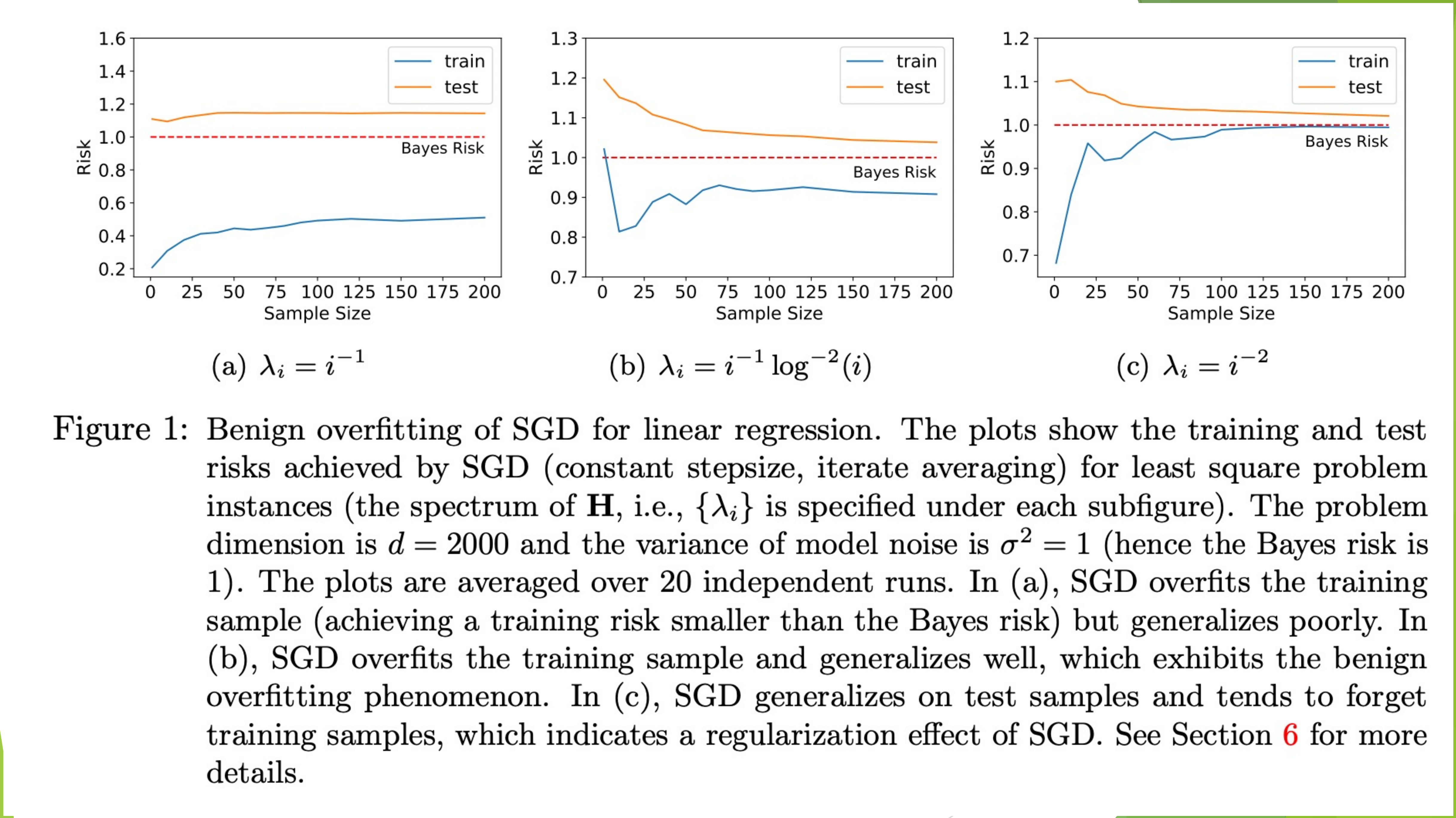
- Stochastic gradient descent (SGD) is a popular optimization method for this problem, particularly in the overparameterized regime where the dimensionality of the weight vector is greater than the number of training samples. The step size of the update is a fixed constant, and the final weight vector is the average of the iterates.
- In the under parameterized case, where the dimensionality of the weight vector is less than or equal to the number of training samples, the optimal risk is achieved by SGD for sufficiently large training samples.
- In the overparameterized case, where the dimensionality of the weight vector is greater than the number of training samples, it has been observed that SGD can overfit the training data, achieving a training error smaller than the Bayes error, but still generalizing well to test data, which is known as the benign overfitting phenomenon.
- This is because in the overparameterized regime, there exist multiple solutions that achieve the same training error, and SGD can find a solution that generalizes well.
- Experimental results have shown that the regularization effect of SGD increases as the dimensionality of the weight vector increases. Hence, overparameterization can act as regularization for SGD.

Introduction And Related Work Continued

- The paper shows a sharp excess risk bound that shows how unregularized SGD can generalize in the infinite-dimension case.
- The bound is stated in terms of the full eigen spectrum of the data covariance matrix, with a functional dependency on the initial iterate.
- The lower bound of the characterization is also shown to be tight.
- The paper also experiments with SGD and tail-averaging
- The paper says that benign overfitting occurs in SGD if certain spectrum decay conditions on the data covariance are met.
- It also shows that SGD with iterate averaging also gives good generalization in the overparameterized setting for linear regression

Introduction And Related Work Continued

- The sharpness of the bounds derived from SGD allows for comparisons with the minimum-norm interpolator ordinary least squares (OLS) and ridge regression.
- It shows that the inductive bias of SGD results in better generalization than minimum-norm interpolator with no regularization.
- The variance in SGD contributes to the final excess risk bound.
- This defines a "bias process" in SGD and compares it with gradient descent.



(a) $\lambda_i = i^{-1}$

(b) $\lambda_i = i^{-1}\log^{-2}(i)$

(c) $\lambda_i = i^{-2}$

Figure 1: Benign overfitting of SGD for linear regression. The plots show the training and test risks achieved by SGD (constant stepsize, iterate averaging) for least square problem instances (the spectrum of $\mathbf{H}$, i.e., $\{\lambda_i\}$ is specified under each subfigure). The problem dimension is $d = 2000$ and the variance of model noise is $\sigma^2 = 1$ (hence the Bayes risk is 1). The plots are averaged over 20 independent runs. In (a), SGD overfits the training sample (achieving a training risk smaller than the Bayes risk) but generalizes poorly. In (b), SGD overfits the training sample and generalizes well, which exhibits the benign overfitting phenomenon. In (c), SGD generalizes on test samples and tends to forget training samples, which indicates a regularization effect of SGD. See Section 6 for more details.

MAIN

- The paper presents upper and lower bounds on the excess risk of iterate averaged stochastic gradient descent (SGD) for linear regression.
- The paper compares these rates to those of ordinary least squares (OLS) and ridge regression then compare similarities and differences.
- The paper introduces several assumptions, including mild regularity conditions on the moments of the data distribution, a fourth moment condition, and a noise condition
- It is observed that the assumptions are weaker than those commonly used for iterate averaged SGD in the under parameterized case.

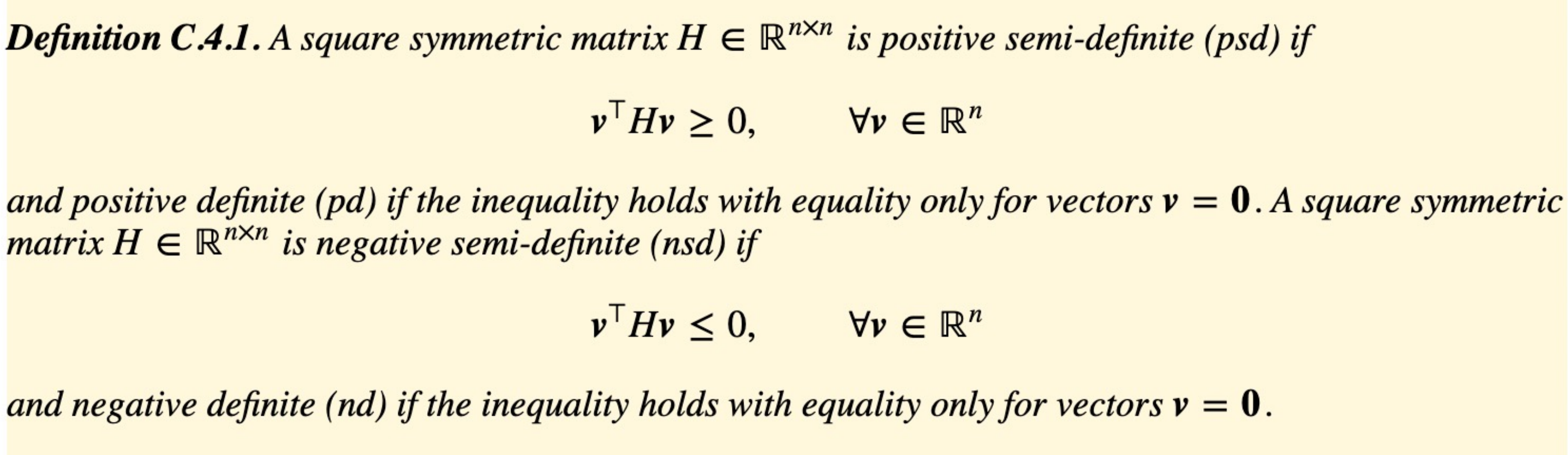
ASSUMPTION 1 : REGULARITY CONDITIONS

- Assumption 2.1 : Certain mathematical properties (regularity conditions) hold for the variables x and y in a mathematical model.
- Assumes that certain statistical moments exist and are finite, and that the second moment of x satisfies certain conditions.
- Assumes that optimization problem has a unique global optimum.
- The assumption deals with the behaviour of the fourth moment as a linear operator on PSD matrices.
- Mild regularity conditions on the moments of the data distribution

$$\mathbf{H} := \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top],$$

ASSUMPTION 2 : FOURTH MOMENT CONDITION

- Assumption 2.2 in a regression analysis states that there is a positive constant α > 0 such that for any positive semidefinite matrix A,



Definition C.4.1. *A square symmetric matrix $H \in \mathbb{R}^{n\times n}$ is positive semi-definite (psd) if*

$$\boldsymbol{v}^\top H \boldsymbol{v} \geq 0, \qquad \forall \boldsymbol{v} \in \mathbb{R}^n$$

and positive definite (pd) if the inequality holds with equality only for vectors $\boldsymbol{v} = \mathbf{0}$. A square symmetric matrix $H \in \mathbb{R}^{n\times n}$ is negative semi-definite (nsd) if

$$\boldsymbol{v}^\top H \boldsymbol{v} \leq 0, \qquad \forall \boldsymbol{v} \in \mathbb{R}^n$$

and negative definite (nd) if the inequality holds with equality only for vectors $\boldsymbol{v} = \mathbf{0}$.

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top \mathbf{A}\mathbf{x}\mathbf{x}^\top] \preceq \alpha\, \mathrm{tr}(\mathbf{H}\mathbf{A})\mathbf{H}.$$

- It is the behaviour of the fourth moment, when viewed as a linear operator on PSD (positive semidefinite) matrices
- For Gaussian distributions, α can be taken as 3. This assumption is weaker than assuming sub-Gaussian tails over H−1/2x which is standard assumption in regression analysis. The assumption can be relaxed to only require that A is PSD and commutable with H.

ASSUMPTION 3 : NOISE CONDITION

- Assumption 3 is a noise condition, where y - <w∗, x> is interpreted as additive noise

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(y - \langle \mathbf{w}^*, \mathbf{x}\rangle)\mathbf{x}] = \nabla L(\mathbf{w}^*) = \mathbf{0}.$$

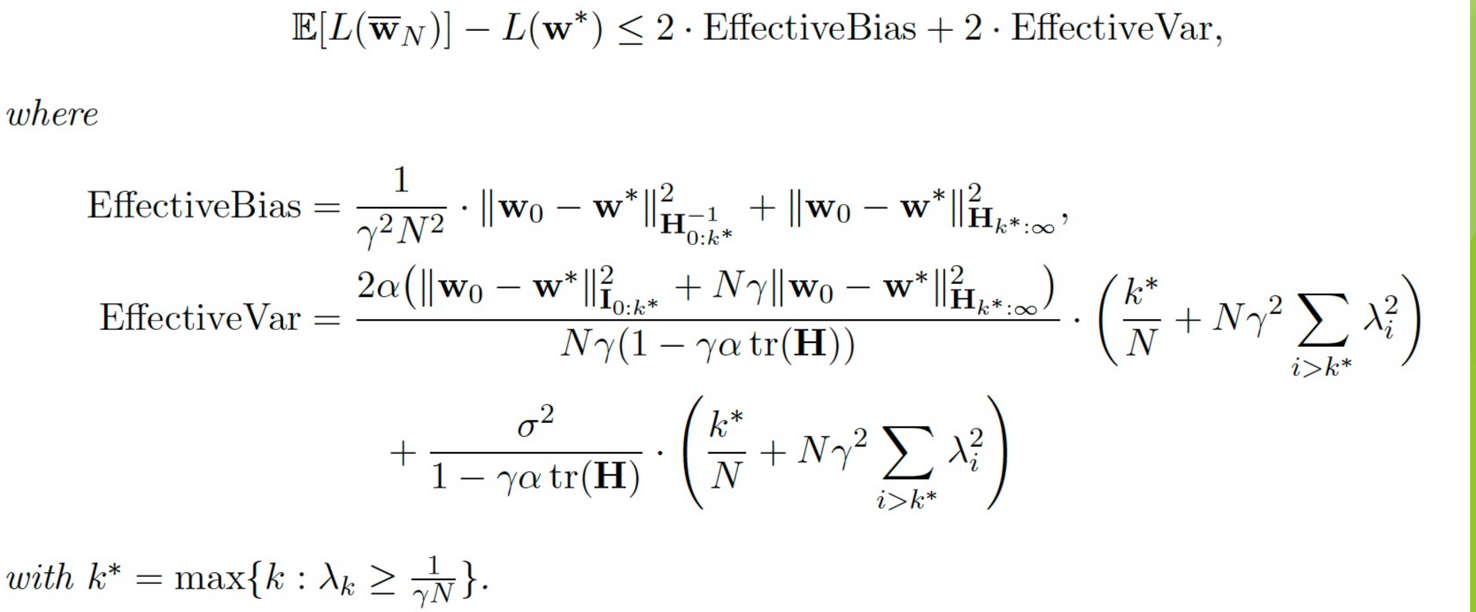
- It states that the covariance matrix of the gradient noise at w should exist and be finite.
- This assumption can be relaxed to permit model mis-specification
- Includes the eigen decomposition of the Hessian.

THEOREM 1 : BENIGN OVERFITTING OF SGD

- Providesa bound on the excess risk of constant-stepsize stochastic gradient descent for linear regression under certain conditions.
- The excess risk is bounded by the sum of the "effective bias" and the "effective variance" terms.
- ”Effective bias" : Convergence rate of gradient descent on the true loss
- ”Effective variance" : The noise in the data and the difference between SGD and GD.
- The bound depends on the "effective dimension" which should be small relative to the sample size. The lower bound is based on lower bound on the fourth moment.
- Step size

$$\gamma < 1/(\alpha\,\mathrm{tr}(\mathbf{H})).$$

- Excess risk upper bound



$$\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) \leq 2 \cdot \mathrm{EffectiveBias} + 2 \cdot \mathrm{EffectiveVar},$$

where

$$\mathrm{EffectiveBias} = \frac{1}{\gamma^2 N^2} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}},$$

$$\mathrm{EffectiveVar} = \frac{2\alpha\big(\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}} + N\gamma\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\big)}{N\gamma(1 - \gamma\alpha\,\mathrm{tr}(\mathbf{H}))} \cdot \left(\frac{k^*}{N} + N\gamma^2 \sum_{i>k^*} \lambda_i^2\right)$$

$$+ \frac{\sigma^2}{1 - \gamma\alpha\,\mathrm{tr}(\mathbf{H})} \cdot \left(\frac{k^*}{N} + N\gamma^2 \sum_{i>k^*} \lambda_i^2\right)$$

with $k^* = \max\{k : \lambda_k \geq \frac{1}{\gamma N}\}$.

ASSUMPTION 4 : FOURTH MOMENT CONDITION, LOWER BOUND

- States that for any positive semi-definite matrix A, there exists a constant $\beta \geq 0$, such that the expected value of the fourth moment of a random variable x, which follows a distribution D, is greater than or equal to the product of the matrix A and β times the trace of the product of matrix A and H.
- In Gaussian distributions, β can be 2. It implies that the upper bound on the noise is not improvable except for absolute constants when the noise is well-specified.

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top\mathbf{A}\mathbf{x}\mathbf{x}^\top] - \mathbf{HAH} \succeq \beta\,\mathrm{tr}(\mathbf{HA})\mathbf{H}.$$

THEOREM 2 : EXCESS RISK LOWER BOUND

- Provides a lower bound for excess risk in supervised learning, where the data distribution is well-specified and meets certain assumptions. The excess risk lower bound is expressed in terms of the Effective Bias and Effective Variance
- Effective Variance is given by model noise and variance in stochastic gradient descent (SGD)
- Step size $\gamma < 1/\lambda_1,$

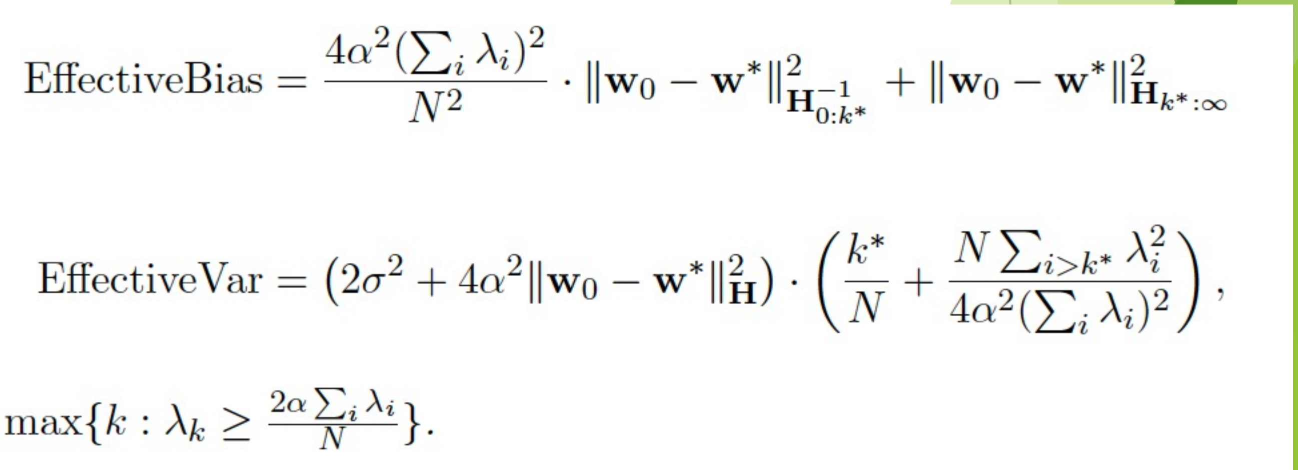
$$\begin{aligned}
\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) \geq\ & \frac{1}{100\gamma^2 N^2} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \frac{1}{100} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}} \\
& + \frac{\beta\left(\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{I}_{0:k^*}} + N\gamma\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}\right)}{1000N\gamma} \cdot \left(\frac{k^*}{N} + N\gamma^2 \sum_{i>k^*} \lambda_i^2\right) \\
& + \frac{\sigma^2_{\text{noise}}}{50} \cdot \left(\frac{k^*}{N} + N\gamma^2 \sum_{i>k^*} \lambda_i^2\right)
\end{aligned}$$

$\textit{with } k^* = \max\{k : \lambda_k \geq \frac{1}{N\gamma}\}.$

COROLLARY 1 : BENIGN OVERFITTING WITH LARGE STEP SIZES

- Corollary to Theorem 2.1 : Provides expressions for effective bias and effective variance, which decay at different rates in different subspaces
- In the "head" eigenspace, the bias error decays faster at a rate of O(1/N^2)
- In the "tail" eigenspace, the decay rate is slower at O(1/N)
- Provides a crude bias bound which means that bias never decays slower than O(1/N).
- Step size is large

$$\gamma = 1/(2\alpha \textstyle\sum_i \lambda_i).$$



$$\mathrm{EffectiveBias} = \frac{4\alpha^2(\sum_i \lambda_i)^2}{N^2} \cdot \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}$$

$$\mathrm{EffectiveVar} = (2\sigma^2 + 4\alpha^2\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}}) \cdot \left(\frac{k^*}{N} + \frac{N\sum_{i>k^*}\lambda_i^2}{4\alpha^2(\sum_i \lambda_i)^2}\right),$$

$$\textit{where } k^* = \max\{k : \lambda_k \geq \tfrac{2\alpha\sum_i \lambda_i}{N}\}.$$

COROLLARY 2 : CRUDE BIAS BOUND

- Provides an upper bound on the expected difference between the loss function of the learned model and optimal model under certain assumptions and a specific step size.
- The excess risk achieved by stochastic gradient descent (SGD) depends on the covariance matrix's spectrum.
- It provides a crude bias bound, showing that bias never decays ore slowly than O (1/N)

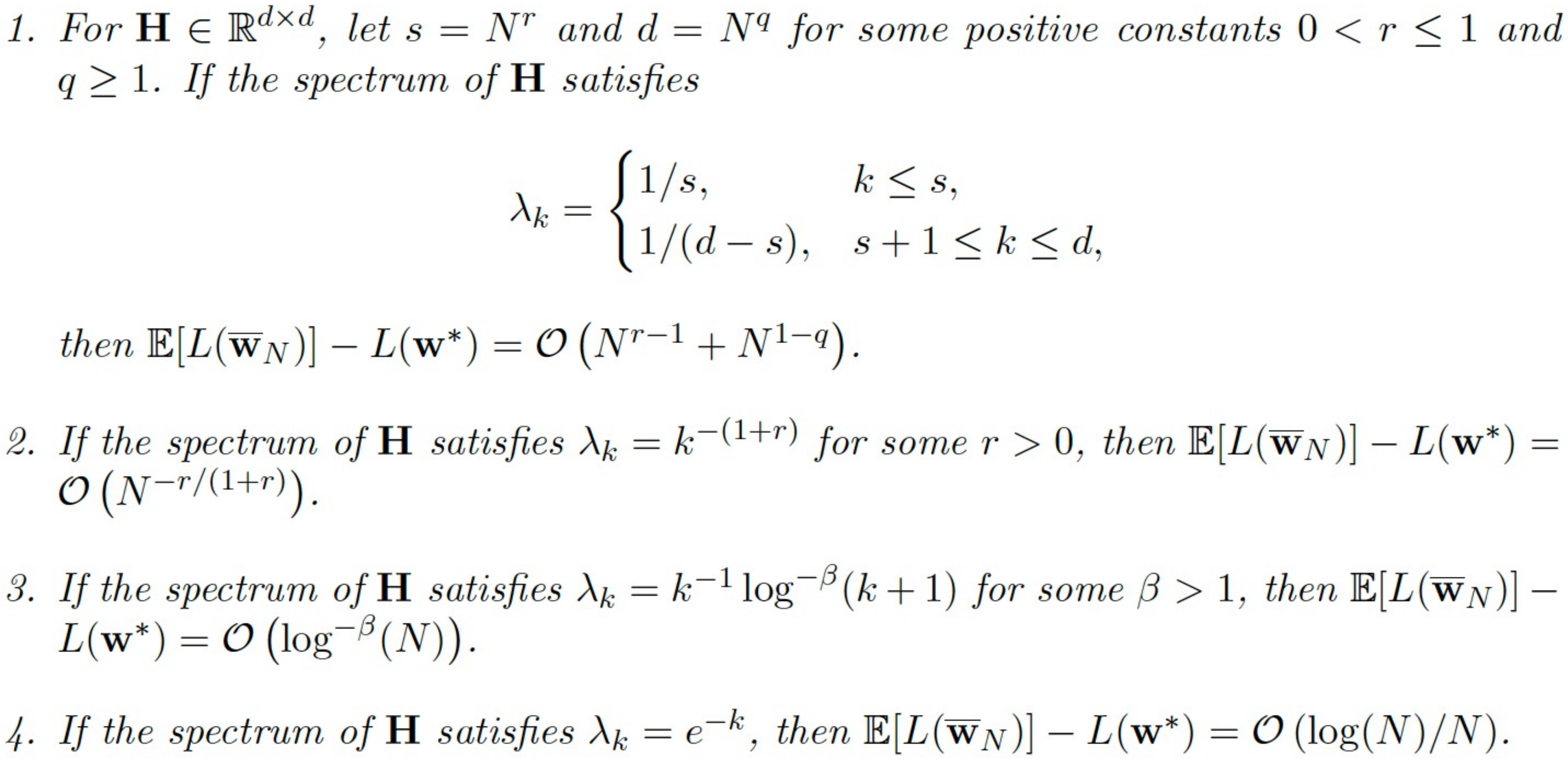
- Step Size

$$\gamma = 1/(2\alpha \textstyle\sum_i \lambda_i).$$

$$\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) \le \frac{8\alpha\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2 \cdot \sum_i \lambda_i}{N} + 4\sigma^2 \cdot \left(\frac{k^*}{N} + \frac{N \sum_{i>k^*} \lambda_i^2}{4\alpha^2(\sum_i \lambda_i)^2}\right),$$

$$\textit{where } k^* = \max\{k : \lambda_k \ge \tfrac{2\alpha \sum_i \lambda_i}{N}\}.$$

COROLLARY 3 : EXAMPLE DATA DISTRIBUTIONS



1. *For* $\mathbf{H} \in \mathbb{R}^{d\times d}$*, let* $s = N^r$ *and* $d = N^q$ *for some positive constants* $0 < r \leq 1$ *and* $q \geq 1$*. If the spectrum of* $\mathbf{H}$ *satisfies*

$$\lambda_k = \begin{cases} 1/s, & k \leq s, \\ 1/(d-s), & s+1 \leq k \leq d, \end{cases}$$

then $\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) = \mathcal{O}\left(N^{r-1} + N^{1-q}\right)$.

2. *If the spectrum of* $\mathbf{H}$ *satisfies* $\lambda_k = k^{-(1+r)}$ *for some* $r > 0$*, then* $\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) = \mathcal{O}\left(N^{-r/(1+r)}\right)$.

3. *If the spectrum of* $\mathbf{H}$ *satisfies* $\lambda_k = k^{-1}\log^{-\beta}(k+1)$ *for some* $\beta > 1$*, then* $\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) = \mathcal{O}\left(\log^{-\beta}(N)\right)$.

4. *If the spectrum of* $\mathbf{H}$ *satisfies* $\lambda_k = e^{-k}$*, then* $\mathbb{E}[L(\overline{\mathbf{w}}_N)] - L(\mathbf{w}^*) = \mathcal{O}\left(\log(N)/N\right)$.

COMPARISON WITH OLS (ORDINARY LEAST SQUARES) AND RIDGE REGRESSION

- The paper compares the rates of benign overfitting achieved by OLS, ridge regression, and constant-stepsize stochastic gradient descent (SGD) for linear regression.
- The paper shows that while OLS and ridge regression require slow decaying rates of the data spectrum to achieve benign overfitting
- SGD can achieve vanishing excess risk for any decay rate including a fast decaying spectrum.
- The effectiveness of SGD is attributed to its ability to control the tail summation of the data spectrum while achieving a small effective dimension.

SGD VS MINIMUM-NORM SOLUTION OF OLS

- Previous papers show that the minimum l2 norm interpolator for the linear regression can reach vanishing excess risk when the data spectrum decays in a specific rate and under certain conditions.
- SGD can achieve vanishing excess risk for any alpha > 1 and beta >= 0, as well as when alpha = 1 and beta >= 1.
- It is due to the fact that fast decaying spectrum can have both small effective dimension and small tail summation

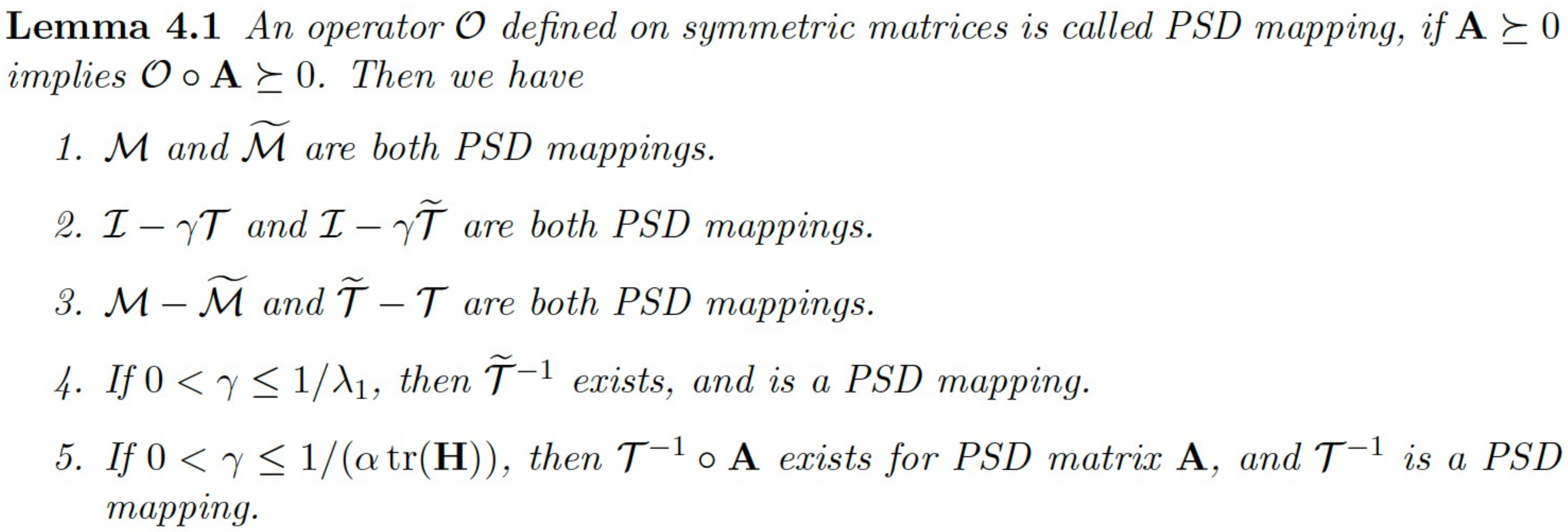
SGD VS RIDGE REGRESSION

- Previous papers show that lower bound and upper bound on the excess risk for ridge regression and compares it with the excess risk for stochastic gradient descent (SGD).
- The lower bound for ridge regression closely matches the upper bound
- It implies that SGD with a constant step size and iterate averaging can perform similarly to ridge regression with a constant regularization parameter
- However, further study is needed

MORE RELATED WORK

- The paper discusses previous research on iterate averaging in both under parameterized and overparameterized cases.
- In the finite dimension case constant step size SGD with iterate or tail average has been studied properly.
- In the overparameterized case previous work only covered specific data covariance spectrum.
- However, this paper's bounds apply to least square instances with any data covariance spectrum
- Previous papers discuss dimension-independent bounds for averaged SGD, but their excess risk bounds for linear regression are not as sharp as those provided in this paper.
- The bias error rate can be improved by considering tail averaging
- The variance error rate has a convergence rate of O(d/N)

PRELIMINARIES : SOME TECHNICAL PROPERTIES



Lemma 4.1 *An operator* $\mathcal{O}$ *defined on symmetric matrices is called PSD mapping, if* $\mathbf{A} \succeq 0$ *implies* $\mathcal{O} \circ \mathbf{A} \succeq 0$. *Then we have*

1. $\mathcal{M}$ *and* $\widetilde{\mathcal{M}}$ *are both PSD mappings.*
2. $\mathcal{I} - \gamma\mathcal{T}$ *and* $\mathcal{I} - \gamma\widetilde{\mathcal{T}}$ *are both PSD mappings.*
3. $\mathcal{M} - \widetilde{\mathcal{M}}$ *and* $\widetilde{\mathcal{T}} - \mathcal{T}$ *are both PSD mappings.*
4. *If* $0 < \gamma \leq 1/\lambda_1$, *then* $\widetilde{\mathcal{T}}^{-1}$ *exists, and is a PSD mapping.*
5. *If* $0 < \gamma \leq 1/(\alpha\, \mathrm{tr}(\mathbf{H}))$, *then* $\mathcal{T}^{-1} \circ \mathbf{A}$ *exists for PSD matrix* $\mathbf{A}$, *and* $\mathcal{T}^{-1}$ *is a PSD mapping.*

BIAS VARIANCE DECOMPOSITION

- The bias-variance decomposition is a tool for analysing averaged SGD in the under parameterized regime.
- The bias error is captured by the "bias iterates," which is stochastic process of SGD on a consistent linear system
- The variance error is give n by the "variance iterates," which is stochastic process of SGD initialized from the optimum.
- The excess risk can be decomposed into a bias error term and a variance error term
- Upper bounds on these terms can be obtained using the Kronecker product and the Cauchy-Schwarz inequality
- The analysis uses finite summations instead of taking the inner summation to infinity

BOUNDING THE VARIANCE ERROR

- In the analysis of the variance error, a weaker assumption can be used instead of Assumption 2.2
- The paper gives a proof under the original assumption, shows that the crude upper bound on Ct obtained from Lemma 5 cannot give a sharp rate in the overparameterized setting.
- The paper refines the upper bound of Ct
- It plugs this refined upper bound into the equation for variance
- It contributes to part of Effective Variance in Theorem 2.1

BOUNDING THE BIAS ERROR

- A similar bound to variance error cannot be applied to the bias error as the sequence of bias is contracting
- The sequence of partial sum of bias is increasing in the PSD sense
- A recursive form is derived to express it
- A tight upper bound for the bias sequence is obtained in the same way as variance error
- The upper bound for the bias sequence consists of two terms
- The first contributing to the Effective Bias
- The second is merged with the variance error to contribute to the Effective Variance term in Theorem 2.1

EFFECT OF TAIL AVERAGING

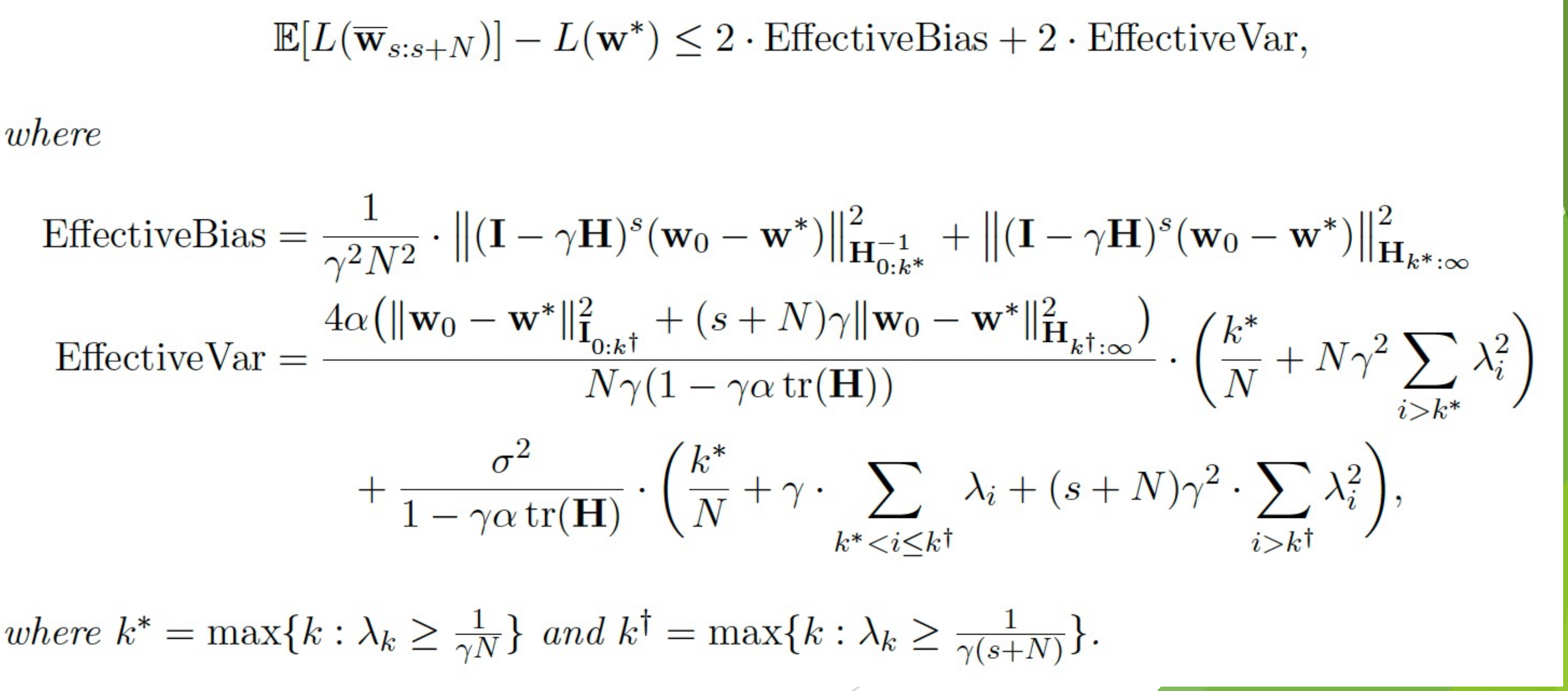
- The paper discusses the effect of tail-averaging on benign overfitting of SGD.
- They present a theorem as a counterpart to Theorem 2.1

THEORM 1 : BENIGN OVERFITTING OF SGD WITH TAIL AVERAGING

It states that SGD with tail-averaging and a specific step size can upper bound the excess risk as the sum of effective bias and effective variance.

It also shows that tail-averaging is better than iterate-averaging, especially for under parameterized and strongly convex cases.

The paper also provides a corresponding lower bound on the excess risk for SGD with tail-averaging, This implies that upper bound is nearly tight.



$$\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N})] - L(\mathbf{w}^*) \leq 2 \cdot \mathrm{EffectiveBias} + 2 \cdot \mathrm{EffectiveVar},$$

where

$$\mathrm{EffectiveBias} = \frac{1}{\gamma^2 N^2} \cdot \left\|(\mathbf{I} - \gamma\mathbf{H})^s(\mathbf{w}_0 - \mathbf{w}^*)\right\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \left\|(\mathbf{I} - \gamma\mathbf{H})^s(\mathbf{w}_0 - \mathbf{w}^*)\right\|^2_{\mathbf{H}_{k^*:\infty}}$$

$$\mathrm{EffectiveVar} = \frac{4\alpha\left(\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{I}_{0:k^\dagger}} + (s+N)\gamma\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^\dagger:\infty}}\right)}{N\gamma(1 - \gamma\alpha\,\mathrm{tr}(\mathbf{H}))} \cdot \left(\frac{k^*}{N} + N\gamma^2 \sum_{i>k^*} \lambda_i^2\right)$$

$$+ \frac{\sigma^2}{1 - \gamma\alpha\,\mathrm{tr}(\mathbf{H})} \cdot \left(\frac{k^*}{N} + \gamma \cdot \sum_{k^* < i \leq k^\dagger} \lambda_i + (s+N)\gamma^2 \cdot \sum_{i > k^\dagger} \lambda_i^2\right),$$

where $k^* = \max\{k : \lambda_k \geq \frac{1}{\gamma N}\}$ *and* $k^\dagger = \max\{k : \lambda_k \geq \frac{1}{\gamma(s+N)}\}$.

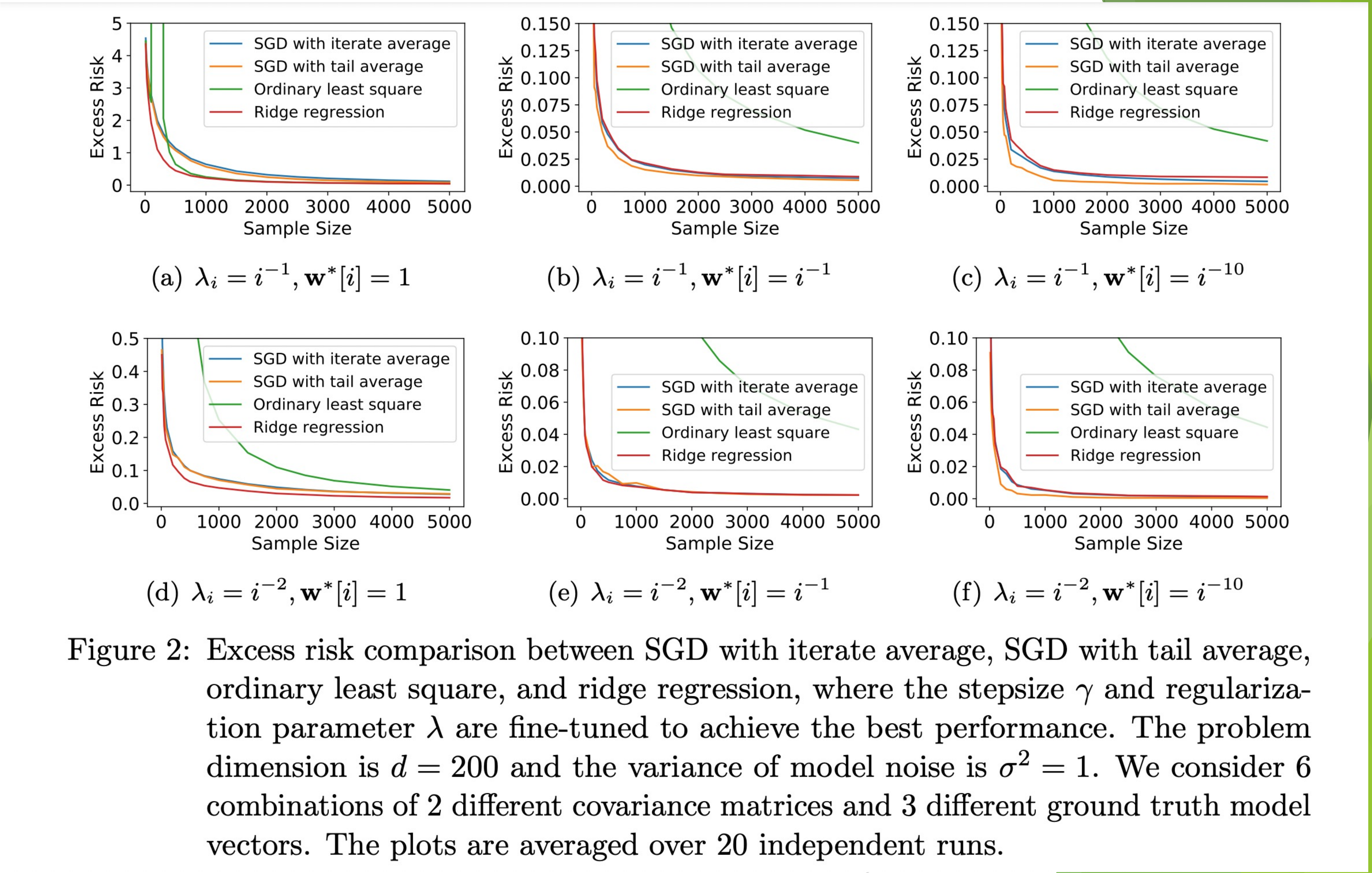
THEOREM 2 : EXCESS RISK LOWER BOUND, TAIL AVERAGING

- It states that SGD with tail-averaging under certain assumptions and suitable step size

$$
\begin{aligned}
\mathbb{E}[L(\overline{\mathbf{w}}_{s:s+N})] - L(\mathbf{w}^*) \geq{}& \frac{1}{100\gamma^2 N^2} \cdot \|(\mathbf{I} - \gamma\mathbf{H})^s(\mathbf{w}_0 - \mathbf{w}^*)\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \frac{\|(\mathbf{I} - \gamma\mathbf{H})^s(\mathbf{w}_0 - \mathbf{w}^*)\|^2_{\mathbf{H}_{k^*:\infty}}}{100} \\
&+ \frac{\beta\|\mathbf{w}_0 - \mathbf{w}^*\|^2_{\mathbf{H}_{k^\dagger:\infty}}}{10^4}\left(\frac{k^*}{N} + N\gamma^2 \sum_{i>k^*} \lambda_i^2\right) \\
&+ \frac{\sigma^2_{\text{noise}}}{600}\left(\frac{k^*}{N} + \gamma \sum_{k^*<i\leq k^\dagger} \lambda_i + (s+N)\gamma^2 \sum_{i>k^\dagger} \lambda_i^2\right),
\end{aligned}
$$

$\textit{where } k^* = \max\{k : \lambda_k \geq \frac{1}{N\gamma}\} \textit{ and } k^\dagger = \max\{k : \lambda_k \geq \frac{1}{(s+N)\gamma}\}.$

- The upper and lower bounds match for most terms, except for the first effective variance term.
- Proper matching upper and lower bounds needs more research



(a) $\lambda_i = i^{-1}, \mathbf{w}^*[i] = 1$

(b) $\lambda_i = i^{-1}, \mathbf{w}^*[i] = i^{-1}$

(c) $\lambda_i = i^{-1}, \mathbf{w}^*[i] = i^{-10}$

(d) $\lambda_i = i^{-2}, \mathbf{w}^*[i] = 1$

(e) $\lambda_i = i^{-2}, \mathbf{w}^*[i] = i^{-1}$

(f) $\lambda_i = i^{-2}, \mathbf{w}^*[i] = i^{-10}$

Figure 2: Excess risk comparison between SGD with iterate average, SGD with tail average, ordinary least square, and ridge regression, where the stepsize γ and regularization parameter λ are fine-tuned to achieve the best performance. The problem dimension is $d = 200$ and the variance of model noise is $\sigma^2 = 1$. We consider 6 combinations of 2 different covariance matrices and 3 different ground truth model vectors. The plots are averaged over 20 independent runs.

EXPERIMENTS

- Experiments are conducted to observe the benign overfitting phenomenon of SGD (Stochastic Gradient Descent) in Gaussian least square problems and to verify the generalization performance of SGD.
- Three over-parameterized linear regression problem instances with different spectral properties of H are considered
- The ground truth is fixed to be w∗[i] = i-1.
- The experiments show that benign overfitting of SGD can happen when the spectrum of H decays neither fast nor slow.
- The results show that SGD with iterate averaging and SGD with tail averaging achieve a smaller excess risk compared to ordinary least square and ridge regression when the step size and regularization parameter are fine-tuned.
- The experiments suggest that the benign overfitting of SGD can happen in practice and that SGD can achieve better generalization performance than traditional methods.

Thank You!