Implicit Regularization or Implicit Conditioning? Exact Risk Trajectories of SGD in High Dimensions

Anonymous Author(s)
Affiliation
Address
email

Abstract

Stochastic gradient descent (SGD) is a pillar of modern machine learning, serving as the go-to optimization algorithm for a diverse array of problems. While the empirical success of SGD is often attributed to its computational efficiency and favorable generalization behavior, neither effect is well understood and disentangling them remains an open problem. Even in the simple setting of convex quadratic problems, worst-case analyses give an asymptotic convergence rate for SGD that is no better than full-batch gradient descent (GD), and the purported implicit regularization effects of SGD lack a precise explanation. In this work, we study the dynamics of multi-pass SGD on high-dimensional convex quadratics and establish an asymptotic equivalence to a stochastic differential equation, which we call homogenized stochastic gradient descent (HSGD), whose solutions we characterize explicitly in terms of a Volterra integral equation. These results yield precise formulas for the learning and risk trajectories, which reveal a mechanism of implicit conditioning that explains the efficiency of SGD relative to GD. We also prove that the noise from SGD negatively impacts generalization performance, ruling out the possibility of any type of implicit regularization in this context. Finally, we show how to adapt the HSGD formalism to include streaming SGD, which allows us to produce an exact prediction for the excess risk of multi-pass SGD relative to that of streaming SGD (bootstrap risk).

1 Introduction

Stochastic gradient descent (SGD) is the algorithm of choice for optimization in modern machine learning and has been hailed as a major reason for deep learning's success [9, 15]. Explanations for the effectiveness of SGD typically refer to its computational efficiency and to its favorable generalization properties, but theoretical understanding of these purported benefits is far from complete.

The efficiency of SGD has been the subject of extensive research, dating back to the original work of Robbins and Monro [43] and extending to modern large-scale machine learning applications (see e.g. [8, 10]). However, despite its widespread adoption and algorithmic simplicity, surprisingly little is known about how SGD performs in the types of high-dimensional optimization problems that occur in practice. Part of the challenge in deriving robust high-level conclusions about the efficiency of SGD is simply that those conclusions can depend on precisely which quantities are measured and what assumptions are leveraged. For example, in the extreme setting where the samples are one-hot vectors, running SGD on a quadratic function is actually identical to running full-batch gradient descent; as such, any statements about the two algorithms' relative efficiency must be data-dependent. Furthermore, the majority of prior analyses focus on the streaming or single-pass setting, where each sample is seen a single time. While this setting is appropriate when the number of samples

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

n is much larger than the dimensionality d, it does not adequately describe the practically-relevant overparameratized or high-dimensional settings where $d \gtrsim n$.

Moreover, the practical success of SGD has been so remarkable in recent years that a growing body of literature has suggested that its benefit to generalization extends beyond what any improved efficiency might reasonably afford [50, 22, 11, 46, 49]. Some of the myriad explanations for SGD's favorable generalization properties include the local geometry of minimizers [23, 19, 53, 17], connections to approximate Bayesian inference [30], and the regularization properties of noise [48], among many others. While some of the these perspectives are intuitive and compelling, they are often difficult to rigorously establish from either an empirical or a theoretical perspective. Empirically, simulations at large scale command significant computational resources, and it can be challenging to push to sufficiently late times or sufficiently large batches to establish the appropriate baselines [46, 48]. Theoretically, the strongest existing results are again in the single-pass setting, for which a number of works have established excess risk bounds for quadratic problems [6, 12, 14, 52]. Much less is known in the multi-pass setting, though stability results were established by [18], and some recent works have begun examining generalization [28].

In this work, we study the dynamics of multi-pass SGD on high-dimensional convex quadratic functions and derive exact asymptotic predictions for the learning and risk trajectories. Our analysis establishes an asymptotic equivalence to a stochastic differential equation, which we call homogenized stochastic gradient descent (HSGD), whose solutions we characterize explicitly in terms of a Volterra integral equation. These results allow us to define a precise data-dependent implicit-conditioning ratio (ICR) that determines whether SGD is more efficient than its full-batch cousins. The ICR favors SGD for many practical datasets, providing some explanation for the observed superior efficiency of SGD; interestingly, we also highlight settings for which SGD is less efficient than full-batch momentum gradient descent, underscoring the data-dependence of the conclusions. Moreover, our results also show that SGD does not improve generalization performance, whether measured in-distribution or out-of-distribution, and therefore that SGD does not offer any form of implicit regularization in this setting. We emphasize that our results do not rule out possible benefits for non-convex problems, but they do provide some of the first explicit negative results in the convex quadratic case.

1.1 Contributions

Our primary contributions are to:

1. Establish the equivalence of quadratic statistics computed on the iterates of SGD and on a particular stochastic Langevin diffusion process called homogonized SGD (Theorem 1);
2. Exactly characterize the asymptotic training and risk trajectories as the solutions of a deterministic Volterra integral equation (Theorem 2);
3. Prove that the noise from SGD negatively impacts generalization performance, both in- and out-of-distribution (Section 3.1), but explain why the impact is often minimal in practice;
4. Introduce the implicit-conditioning ratio that describes when and by how much SGD accelerates convergence relative to the best full-batch methods (Section 3.2);
5. Analyze the limit of streaming SGD to show its inability to capture many salient features of the dynamics of multi-pass SGD (Section 4).

2 Preliminaries and background

Problem setting. We consider high-dimensional ℓ^2-regularized least squares problems defined by,

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \left\{ f(\boldsymbol{x}) \stackrel{\text{def}}{=} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2 + \frac{\delta}{2}\|\boldsymbol{x}\|^2 = \sum_{i=1}^{n} \underbrace{\frac{1}{2}\left((\boldsymbol{a}_i\boldsymbol{x} - b_i)^2 + \frac{\delta}{n}\|\boldsymbol{x}\|^2 \right)}_{\stackrel{\text{def}}{=} f_i(\boldsymbol{x})} \right\}, \tag{1}$$

where $\delta \geq 0$ is the ridge-regularization parameter. We denote the ridgeless empirical risk as

$$\mathcal{L}(\boldsymbol{x}) \stackrel{\text{def}}{=} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2. \tag{2}$$

On the problem (1), the steps taken by gradient decent (GD) can be written recursively as

$$\boldsymbol{x}_{k+1}^{\text{m-gd}} = \boldsymbol{x}_k^{\text{m-gd}} - \gamma_k \nabla f(\boldsymbol{x}_k^{\text{m-gd}}) = \boldsymbol{x}_k^{\text{m-gd}} - \gamma_k \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x}_k^{\text{m-gd}} - \boldsymbol{b}) - \gamma_k \delta \boldsymbol{x}_k^{\text{m-gd}} + \Delta(\boldsymbol{x}_k^{\text{m-gd}} - \boldsymbol{x}_{k-1}^{\text{m-gd}}), \quad (3)$$

where $\Delta > 0$ is the momentum parameter, γ_k is the learning rate schedule, and $\boldsymbol{x}_0 \in \mathbb{R}^d$ is an initial vector assumed to be independent of all other randomness and having norm at most 1. When A is large, computing these updates can be expensive, so an unbiased estimator for the true gradient is often used, where a subset of the data points are selected uniformly at random. We focus on the setting with batch size equal to one and without momentum, which we refer to as stochastic gradient descent (SGD), and for which the iterates can be written recursively as

$$\boldsymbol{x}_{k+1}^{\text{sgd}} = \boldsymbol{x}_k^{\text{sgd}} - \gamma_k \nabla f_{i_k}(\boldsymbol{x}_k^{\text{sgd}}) = \boldsymbol{x}_k^{\text{sgd}} - \gamma_k \boldsymbol{A}^T \boldsymbol{e}_{i_k} \boldsymbol{e}_{i_k}^T (\boldsymbol{A}\boldsymbol{x}_k^{\text{sgd}} - \boldsymbol{b}) - \tfrac{\gamma_k \delta}{n} \boldsymbol{x}_k^{\text{sgd}}, \quad (4)$$

where the $i_k \sim \text{Unif}([n])$ iid. While it would also be possible to consider mini-batch SGD, previous work has shown that batch sizes that are vanishingly small as a fraction of the number of samples are equivalent to the single-batch analysis, after appropriately adjusting the time by a factor of the batch size [37, Theorem 1]; similarly, we do not consider high-dimensional SGD with momentum as it degenerates to SGD [36]. See also [21].

Diffusion approximations and homogenized SGD. A common paradigm for understanding SGD is through stochastic Langevin diffusions (SLD), i.e. solutions of equations of the form

$$\mathrm{d}\boldsymbol{X}_t = -\gamma \nabla f(\boldsymbol{X}_t) + \sqrt{\boldsymbol{\Sigma}_t} d\boldsymbol{B}_t\,, \quad (5)$$

where γ is the step size of SGD, f is the loss function, $\boldsymbol{B}_t$ is a d-dimensional standard Brownian motion, and the matrix $0 \preceq \boldsymbol{\Sigma}_t \in \mathbb{R}^{d\times d}$ models the noise covariance. In many analyses, no concrete connection between SGD and SLD is developed, and the diffusion is merely used to build intuition. A common example is the isotropic case ($\boldsymbol{\Sigma}_t \propto \mathbf{I}_d$), for which the Fokker-Planck equation implies that the dynamics are reversible with respect to a density proportional to $e^{-f(x)}$. Consequently, the process can escape local minima, exhibiting a trade-off between the entropy and depth of minima and thereby highlighting a possible mechanism of implicit regularization. In the general anisotropic case, describing the stationary distribution is more difficult; nonetheless, the local geometry near minima of f can be analyzed, see [11, 25].

While this type of implicit entropic regularization might ultimately underlie the generalization benefits of SGD for nonconvex problems, currently we lack a precise connection between a concrete SLD and a practical nonconvex learning problem. As such, the implicit regularization effects of SGD on nonconvex losses remains a largely unsolved problem.

For convex quadratics, however, the implications of Eq. (5) are quite clear: there is no notion of implicit regularization as the noise in SLD negatively impacts generalization performance. Note that because the noise is mean zero, any SLD is centered around *gradient flow* (GF) $\mathfrak{X}_t^{\text{gf}}$, which solves

$$\mathrm{d}\mathfrak{X}_t^{\text{gf}} = -\nabla f(\mathfrak{X}_t^{\text{gf}})\,, \quad (6)$$

leading to the following conclusion for generalization:

Lemma 1. *Suppose the objective function is $f(\boldsymbol{x}) = \frac{1}{2}\big(\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 + \delta\|\boldsymbol{x}\|^2\big)$. Suppose $(\boldsymbol{X}_t : t \in [0,\infty))$ is an SLD (i.e. $\boldsymbol{X}_t$ solves (5)) with $\|\boldsymbol{\Sigma}_t\|_{op}$ almost surely bounded by some $C < \infty$. Suppose the population risk $\mathcal{R} : \mathbb{R}^d \to \mathbb{R}$ is a convex function and denote $\boldsymbol{x}_* \stackrel{def}{=} \lim_{t\to\infty} \mathfrak{X}_t^{gf}$, then*

$$\underbrace{\mathbb{E}[\mathcal{R}(\boldsymbol{X}_t)]}_{\text{pop. risk of SLD}} \geq \underbrace{\mathcal{R}(\mathfrak{X}_{\gamma t}^{gf})}_{\text{pop. risk of gradient flow}} \quad \text{for all } t \geq 0 \text{ and hence,} \quad \underbrace{\liminf_{t\to\infty} \mathbb{E}[\mathcal{R}(\boldsymbol{X}_t)]}_{\text{limiting pop. risk of SLD}} \geq \underbrace{\mathcal{R}(\boldsymbol{x}_*)}_{\text{limiting pop. risk gradient flow}}.$$

If in addition $\mathcal{R}$ is strictly convex, and $\boldsymbol{\Sigma}_t \to \boldsymbol{\Sigma}_\infty$ with $\boldsymbol{\Sigma}_\infty \succ 0$, then the inequality is strict.

Proof. The mean $\mathbb{E}[\boldsymbol{X}_t]$, by the linearity of the gradient ∇f, is GF. Under the conditions given, the law of $\boldsymbol{X}_t$ converges to a Gaussian variable centered at $\boldsymbol{x}_*$. Hence by Fatou's lemma and Jensen's inequality, the inequality follows. □

We emphasize that this conclusion applies even under general distribution shifts, so long as the risk remains a convex function. Still, the utility of Lemma 1 may not be immediately clear, as it pertains

to SLD and we have not yet established any concrete connection between SLD and the process of interest, SGD. Nor is it evident what form such a connection should take—the agreement between SGD and an SLD cannot occur at the level of individual states since the randomness from each process is not assumed to be coupled. Instead, the most we can hope for is that statistics of the processes agree. Specifically, we might hope that matching the noise structure of SGD with a careful choice of SLD will cause relevant statistics, like the population risk, to be equal.

It turns out that such a choice of SLD exists for convex quadratic problems in high dimensions, and is given by *homogenized SGD* (HSGD), introduced simultaneously in [33, 36]. Both the empirical and population risks ($\mathcal{L}$, $\mathcal{R}$ resp.) of HSGD agree with the same of SGD in the high-dimensional limit (see Thm. 1). Mathematically, HSGD is the strong solution of the stochastic differential equation:

$$\mathrm{d}\boldsymbol{X}_t \stackrel{\text{def}}{=} -\gamma(t)\nabla\mathcal{L}(\boldsymbol{X}_t)\,\mathrm{d}t + \gamma(t)\sqrt{\tfrac{2}{n}\mathcal{L}(\boldsymbol{X}_t)\nabla^2\mathcal{L}(\boldsymbol{X}_t)}\,\mathrm{d}\boldsymbol{B}_t, \quad \text{for quadratic } \mathcal{L}, \tag{7}$$

where again $\boldsymbol{B}_t$ is a d-dimensional standard Brownian motion, $\gamma(t)$ is the learning rate schedule, and the initial condition is $\boldsymbol{X}_0 = \boldsymbol{x}_0$. Roughly, HSGD is a diffusion approximation to SGD that gains explanatory power when the *dimensionality* is large. It does not require the step size γ to be small. Note that as with other universality results, the details of the noise distribution are not relevant and only the second-order correlations contribute, which are carefully matched by HSGD to SGD.

The precise sense of the comparison requires us to evaluate low-dimensional statistics of the high-dimensional dynamics; "low-dimensional" must be effective, in that the univariate statistics of the SGD iterates concentrate around the same statistic evaluated on HSGD. For understanding generalization or implicit regularization properties, a important statistic is the population risk, $\mathcal{R}$.

Assumptions. For all parts of our analysis to hold, the pair $(\boldsymbol{A}, \boldsymbol{b})$ of the data matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ and target vector $\boldsymbol{b} \in \mathbb{R}^n$ must satisfy some *quasi-random* assumptions—a set of deterministic conditions on the pair $(\boldsymbol{A}, \boldsymbol{b})$ that are satisfied with high probability by natural classes of random matrix-vector pairs (see Appendix A for specifics). We use the convention that the target and initialization vectors are bounded independent of n, $\|\boldsymbol{b}\|_2^2 \le C$ and $\|\boldsymbol{x}_0\|_2^2 \le C$, respectively.

We illustrate some examples below that we have shown to satisfy the quasi-random assumptions.

- Gaussian linear regression. Here the rows of $\boldsymbol{A}$ are iid and drawn from a Gaussian with norm-bounded covariance Σ and the target $\boldsymbol{b}$ is drawn from a generative model, $\boldsymbol{b} = \boldsymbol{A}\widetilde{\boldsymbol{x}} + \boldsymbol{\eta}$ for some unknown signal $\widetilde{\boldsymbol{x}} \in \mathbb{R}^d$ and independent noise $\boldsymbol{\eta} \in \mathbb{R}^n$.
- Subgaussian linear designs. In the example above, we can relax the Gaussian assumption to be of the form $\boldsymbol{x}\Sigma^{1/2}$ for $\boldsymbol{x}$ a vector of iid centered subgaussian random variables [55, 21].
- Gaussian random features with a linear ground truth [32, 3, 42, 4]. Suppose $\boldsymbol{A}$ is given by $\sigma(\boldsymbol{X}\boldsymbol{W})$ for an iid standard Gaussian weight matrix $\boldsymbol{W}$ and Gaussian data matrix $\boldsymbol{X}$. Suitable assumptions on the activation function σ and the covariance Σ of $\boldsymbol{X}$ added.

Assumption 1. *The population risk $\mathcal{R} : \mathbb{R}^d \to \mathbb{R}$ is a quadratic, that is, it is a degree-2 polynomial or, equivalently, can be represented by*

$$\mathcal{R}(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{T}\boldsymbol{x} + \boldsymbol{u}^T\boldsymbol{x} + c$$

for some $d \times d$ symmetric matrix $\boldsymbol{T}$, vector $\boldsymbol{u} \in \mathbb{R}^d$, and scalar $c \in \mathbb{R}$. We further assume that $\|\nabla^2\mathcal{R}\|_{op} \le C$, $\|\nabla\mathcal{R}(0)\|_2^2 \le C$, and $|\mathcal{R}(0)| \le C$.

A natural population risk is given by $\mathcal{R}(\boldsymbol{x}) = \frac{1}{2}\mathbb{E}\left[(\boldsymbol{a}\cdot\boldsymbol{x} - b)^2\right]$ where $(\boldsymbol{a}, b) \sim \mathcal{D}$. This distribution $\mathcal{D}$ may or may not be the same as the distribution that generated the data $[\boldsymbol{A} \mid \boldsymbol{b}]$ used in training.

As we work in the high-dimensional limit, we suppose that $\gamma_k = \gamma(k/n)$ for a smooth, bounded function $\gamma(\cdot)$ such that $\gamma(t) \to \gamma \in [0, \infty)$ and $\widehat{\gamma} \stackrel{\text{def}}{=} \sup_{t\ge 0}\gamma(t) < \infty$.

3 Main results

Our main results are analyzable (non-asymptotic) expressions for the empirical risk $\mathcal{L}$ and the population risk $\mathcal{R}$ of SGD at any time t for the high-dimensional least squares problem (1). To begin, we first establish the following equivalence between SGD and HGSD.

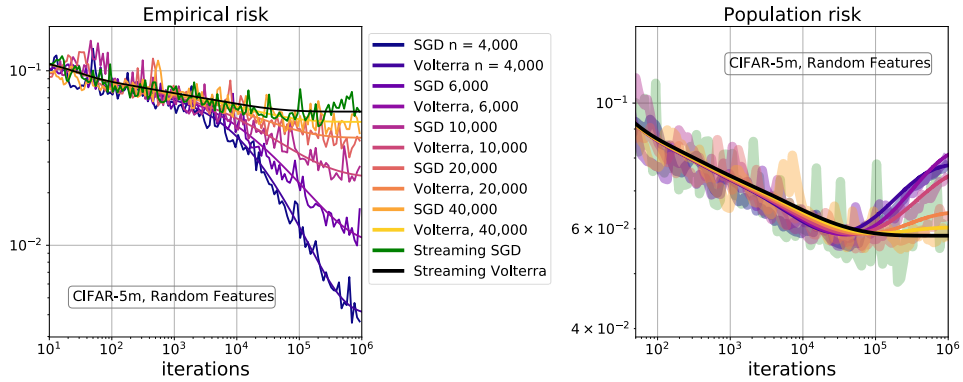

Figure 1: **Single runs of SGD vs. HSGD (Volterra) in streaming** on standarized CIFAR-5M [34] with car/plane class vector $(1,000,000$ samples); a standarized ReLu (71) random features model (see Appendix A.1) was applied with increasing number of samples n and fixed $d = 6000$. The predicted behavior from HSGD (denoted by Volterra) matches the performance of single runs of SGD for finite n and streaming $(n = \infty)$. Shaded region (right) is the moving average of a single run of SGD. Empirical risk (left) increases monotonically with n to its limit while population risk generally decreases with n. Streaming corresponds to $n = \infty$ (see Sec. 4). For consistency across sample sizes, time is measured in iterations. Additional details in App E.

Theorem 1 (Equivalence of SGD and HSGD)**.** *Suppose the pair* $(\boldsymbol{A}, \boldsymbol{b}) \in \mathbb{R}^{n\times d} \times \mathbb{R}^d$ *satisfy the quasi-random assumptions with* $d^{\varepsilon} \le n \le d^{1/\varepsilon}$ *for some* $\varepsilon \in (0, 1]$. *Let the iterates* $\boldsymbol{x}_t = \boldsymbol{x}^{sgd}_{\lfloor t \rfloor}$ *be generated from multi-pass SGD Eq.* (4) *and* $\boldsymbol{X}_t$ *be the solution of Eq.* (7). *Then for any deterministic* $T > 0$ *and any* $D > 0$, *there is a* $C > 0$ *such that*

$$\Pr\left[\sup_{0\le t\le T} \left\| \begin{pmatrix} \mathcal{L}(\boldsymbol{x}_{\lfloor tn \rfloor}) \\ \mathcal{R}(\boldsymbol{x}_{\lfloor tn \rfloor}) \end{pmatrix} - \begin{pmatrix} \mathcal{L}(\boldsymbol{X}_t) \\ \mathcal{R}(\boldsymbol{X}_t) \end{pmatrix} \right\|_2 > d^{-\varepsilon/2}\right] \le C d^{-D}.$$

As a result, for the rest of this paper, we can use homogenized SGD to analyze the behavior of multi-pass SGD. We will establish a similar conjecture for streaming SGD in Section 4.

While the comparison of SGD to HSGD requires relatively strong assumptions on $\boldsymbol{A}$ and $\boldsymbol{b}$, the analysis of HSGD can be performed under weaker assumptions (no quasirandomness assumptions are needed). It suffices to suppose the problem is high dimensional in the following sense:

Assumption 2. *The empirical risk* $\mathcal{L}$ *satisfies* $\operatorname{tr} \nabla^2 \mathcal{L} = n$ *and* $0 \preceq \nabla^2 \mathcal{L} \preceq n d^{-\epsilon}$ *for some* $\epsilon > 0$.

This corresponds to the normalization where $\nabla^2 \mathcal{L} = \boldsymbol{A}^T \boldsymbol{A}$ and each row of $\boldsymbol{A}$ is length 1 and hence $\operatorname{tr} \nabla^2 \mathcal{L} = n$.

Under Assumptions 1 and 2, the dynamics of the empirical and population risk under HSGD concentrate around a deterministic dynamical system driven by a Volterra integral equation:

Volterra Dynamics, Multi-pass. The following deterministic dynamical system is the high-dimensional limit for $\mathcal{L}(\boldsymbol{X}_t)$ and $\mathcal{R}(\boldsymbol{X}_t)$, respectively

$$\Psi_t = \mathcal{L}\big(\mathfrak{X}^{\mathrm{gf}}_{\Gamma(t)}\big) + \int_0^t K(t, s; \nabla^2 \mathcal{L}) \Psi_s \, \mathrm{d}s \qquad \text{(Empirical risk)} \tag{8}$$

$$\Omega_t = \mathcal{R}\big(\mathfrak{X}^{\mathrm{gf}}_{\Gamma(t)}\big) + \int_0^t K(t, s; \nabla^2 \mathcal{R}) \Psi_s \, \mathrm{d}s \qquad \text{(Population risk)} \tag{9}$$

where the *integrated learning rate* Γ and *kernel* K, for any $d \times d$ matrix $\boldsymbol{P}$, respectively are

$$\Gamma(t) = \int_0^t \gamma(s) \, \mathrm{d}s, \;\; K(t, s; \boldsymbol{P}) = \tfrac{\gamma^2(s)}{n} \operatorname{tr}\big((\nabla^2 \mathcal{L}) \boldsymbol{P} \exp\big(-2(\nabla^2 \mathcal{L} + \delta \mathbf{I}_d)(\Gamma(t) - \Gamma(s))\big)\big). \tag{10}$$

Theorem 2 (Concentration of HSGD around Volterra dynamics)**.** *Under Assumptions 1 and 2, for any $T > 0$ and for any $D > 0$ there exists sufficiently large $C > 0$ such that for all $d > 0$*

$$\Pr\left[\sup_{0\le t\le T}\left\|\begin{pmatrix}\mathcal{L}(\boldsymbol{X}_t)\\ \mathcal{R}(\boldsymbol{X}_t)\end{pmatrix}-\begin{pmatrix}\Psi_t\\ \Omega_t\end{pmatrix}\right\| > d^{-\epsilon/2}\right] \le Cd^{-D},$$

where Ψ_t and Ω_t solve (8) and (9).

We give a formal proof of the concentration result in Appendix B.1 in Theorem 11.

3.1 No implicit regularization from SGD

From (9), for convex $\mathcal{R}$ we observe immediately that the population risk Ω_t is only larger than the population risk of GF. Moreover, we have an explicit formula for the excess risk due to SGD noise,

$$\underbrace{\Omega_t - \mathcal{R}\big(\mathfrak{X}^{\text{gf}}_{\Gamma(t)}\big)}_{\text{excess risk due to SGD}} \stackrel{\text{def}}{=} \int_0^t K(t,s;\nabla^2\mathcal{R}) \times \underbrace{\Psi_s}_{\text{limiting loss } \mathcal{L}} \,\mathrm{d}s.$$

Note that the population risk of SGD tracks that of GF. If GF overfits, SGD overfits as well; there is no statistical regularization due to the noise of SGD applied to empirical risk minimization (ERM).

We can further analyze the long-time behavior of SGD with exact limiting values for this excess risk.

Theorem 3 (Time infinity risk values)**.** *If $\gamma(t) \to 0$ as $t \to \infty$ but $\Gamma(t) \to \infty$ (i.e. the usual Robbins-Monro setting), then the excess population risk of SGD over GF tends to 0. If on the other hand $\gamma(t) \to \gamma \in (0, 2(\frac{1}{n}\operatorname{tr}\{\frac{(\boldsymbol{A}^T\boldsymbol{A})^2}{\boldsymbol{A}^T\boldsymbol{A}+\delta\boldsymbol{I}_d}\})^{-1}$, then with Ψ_∞ given by the limiting empirical risk,*

$$\Psi_\infty = \mathcal{L}\big(\mathfrak{X}^{gf}_\infty\big) \times \left(1 - \frac{\gamma}{2n}\operatorname{tr}\left\{\frac{(\nabla^2\mathcal{L})^2}{\nabla^2\mathcal{L}+\delta\boldsymbol{I}_d}\right\}\right)^{-1}$$

the excess risk due to SGD converges to

$$\Omega_t - \mathcal{R}\big(\mathfrak{X}^{gf}_{\Gamma(t)}\big) \to \frac{\gamma\Psi_\infty}{2n} \times \operatorname{tr}\left\{\frac{(\nabla^2\mathcal{R})(\nabla^2\mathcal{L})}{\nabla^2\mathcal{L}+\delta\boldsymbol{I}_d}\right\}.$$

There are a few conclusions to draw directly from this. In the interpolation regime, that is where $\Psi_\infty = 0$, there is no excess risk due to SGD and there is no need to send γ to 0. Moreover, if the empirical risk Ψ_∞ is small, the excess risk due to SGD is proportional to $\gamma\Psi_\infty$, and hence it is frequently orders of magnitude smaller than other potential sources of error. Furthermore, the excess risk is affected by how similar the population and empirical risks are, in the large directions. Ridge regularization can substantially reduce excess risk due to SGD in cases where population risk has many small eigenvalues. In summary, either by sending $\gamma \to 0$, working in the interpolation regime, or otherwise in a regime Ψ_∞ is small, *the excess risk incurred by running SGD is minimal*.

3.2 Implicit conditioning of SGD

In contrast, the algorithmic advantages of SGD are substantial. To simplify the discussion, we consider only the case of constant learning rate γ. In this case, the kernel in (8) and (9) simplifies to a convolution kernel, which has a much simpler theory. To characterize the rates, we define $\lambda_{\min}$ as the smallest non-zero eigenvalue of $\nabla^2\mathcal{L}$. Then for generic initial conditions, (in particular almost surely if $\boldsymbol{X}_0$ is nonzero isotropic), GF has the following convergence rate

$$\lim_{t\to\infty}\left(\mathcal{L}(\mathfrak{X}^{\text{gf}}_{\gamma t}) - \mathcal{L}(\mathfrak{X}^{\text{gf}}_\infty)\right)^{1/t} = \begin{cases} e^{-\gamma(\lambda_{\min}(\nabla^2\mathcal{L})+\delta)}, & \text{if } \delta > 0,\\ e^{-2\gamma\lambda_{\min}(\nabla^2\mathcal{L})}, & \text{otherwise.}\end{cases}$$

Here we use the notation that $\lambda_{\min}(\boldsymbol{H})$ and $\lambda_{\max}(\boldsymbol{H})$ are the smallest and largest eigenvalues of the matrix $\boldsymbol{H}$. The rate of convergence of Ψ_t to Ψ_∞ can be no faster than the underlying GF, given by the rate above. On the other hand, for larger γ the Volterra term in (8) can frustrate the convergence. The *Malthusian exponent* of the convolution Volterra equation is given by

$$\lambda_* = \inf\left\{x : 1 = \int_0^\infty e^{xt}K(t;\nabla^2\mathcal{L})\,\mathrm{d}t \stackrel{\text{def}}{=} \gamma^2\int_0^\infty e^{xt}\operatorname{tr}\big(\big(\nabla^2\mathcal{L}\big)^2\exp\big(-2\gamma(\nabla^2\mathcal{L}+\delta\mathbf{I}_d)t\big)\big)\,\mathrm{d}t\right\}. \tag{11}$$

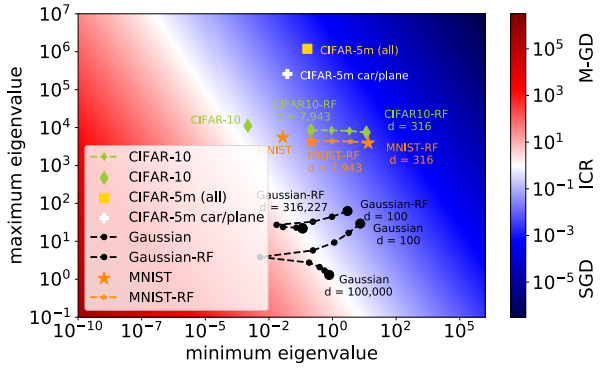

Figure 2: ICR as function of largest and smallest eigenvalues of $\boldsymbol{A}$ with trace normalized to be 1; blue (smaller, SGD favored) and red (larger, full batched M-GD favored). Points indicate ICR for image datasets (MNIST, CIFAR-10, CIFAR-5m) as well as their images under ReLu random feature maps of various dimensions (lines). Gaussian random features ($n = 2000$, $n_0 = 100$, various d) and Gaussian data ($n = 2000$, various d) shows ICR for over- and under-parameterized models. Natural datasets tend to favor SGD.

As $\nabla^2\mathcal{L}$ is finite dimensional, we have that $\lambda_* \leq 2\gamma(\lambda_{\min}(\nabla^2\mathcal{L}) + \delta)$, owing to the divergence of the integral as x approaches this value from below. Note that in principal the Malthusian exponent can be negative, in which case SGD is *divergent*. The Malthusian exponent gives the effective rate of convergence of constant learning rate SGD. Define

$$\Xi(\gamma) \stackrel{\text{def}}{=} \begin{cases} \min\{\gamma(\lambda_{\min}(\nabla^2\mathcal{L}) + \delta), \lambda_*(\gamma)\} & \text{if } \delta > 0, \\ \lambda_*(\gamma) & \text{if } \delta = 0. \end{cases} \tag{12}$$

Theorem 4 (SGD convergence rates, average-case)**.** *Then the rates of convergence of both the empirical and population risk are controlled by this parameter*

$$\lim_{t\to\infty} \left(\Psi_t - \Psi_\infty\right)^{1/t} = e^{-\Xi(\gamma)} = \lim_{t\to\infty} \left(\Omega_t - \Omega_\infty\right)^{1/t}.$$

Furthermore, when $\gamma = \frac{\lambda_{\min}(\nabla^2\mathcal{L})n}{2\operatorname{tr}(\boldsymbol{A}^T\boldsymbol{A})}$*, we have the rate guarantee* $\Xi(\gamma) \geq \frac{\lambda_{\min}(\nabla^2\mathcal{L})+\delta}{2}$.

The major difference between SGD and full batch methods such as momentum gradient descent (M-GD; see Appendix D.2 for definitions) is that they have different sensitivities to the Hessian spectrum of the empirical risk $\mathcal{L}(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2$. Define the condition numbers

$$\kappa \stackrel{\text{def}}{=} \frac{\lambda_{\max}(\nabla^2\mathcal{L}) + \delta}{\lambda_{\min}(\nabla^2\mathcal{L}) + \delta} \quad \text{and} \quad \overline{\kappa} \stackrel{\text{def}}{=} \frac{\frac{1}{n}\operatorname{tr}(\nabla^2\mathcal{L})}{\lambda_{\min}(\nabla^2\mathcal{L}) + \delta}.$$

The first of these is the classical condition number of the ridge problem, while the second is the averaged condition number that regulates the behavior of SGD in the high-dimensional limit. M-GD has been long established to have a rate of convergence, with proper tuning, controlled by the square root of the condition number [41], which is known to be optimal amongst first order algorithms.

Theorem 5 (Convergence rates for M-GD)**.** *For isotropic random initialization* $\boldsymbol{x}_0$ *or noisy* $\boldsymbol{b}$*,* $\delta > 0$*, and strictly convex population risk* $\mathcal{R}$

$$\left(\mathcal{L}(\boldsymbol{x}_k^{m\text{-}gd}) - \mathcal{L}(x_*)\right)^{1/k} \xrightarrow[k\to\infty]{\text{a.s.}} \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right) \quad \textit{and} \quad \left(\mathcal{R}(\boldsymbol{x}_k^{m\text{-}gd}) - \mathcal{R}(x_*)\right)^{1/k} \xrightarrow[k\to\infty]{\text{a.s.}} \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right).$$

See Appendix D.2 for elaboration.

In light of Theorems 4 and 5, we can define the *implicit-conditioning ratio* as

$$\text{ICR} \stackrel{\text{def}}{=} \frac{\overline{\kappa}}{\sqrt{\kappa}} \approx \log\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)\overline{\kappa},$$

which measures the efficiency of SGD over M-GD in that SGD with constant learning rate $n/\operatorname{tr}(\nabla^2\mathcal{L})$ trains in an ICR-multiple of the number of epochs that M-GD requires (lower is better for SGD).

Problems favor SGD when there are large outlier eigenvalues, a common feature of Hessian spectra in practice [44, 45, 1]. Indeed, if the largest eigenvalues are on the same order as the *unnormalized trace*, individual SGD iterates are as effective as full-batch gradient. In contrast, when the Hessian spectrum is tightly packed, which is less common in practice but can occur after some preprocessing techniques or e.g. for uncorrelated Gaussian samples, then M-GD is favored. See Fig. 2.

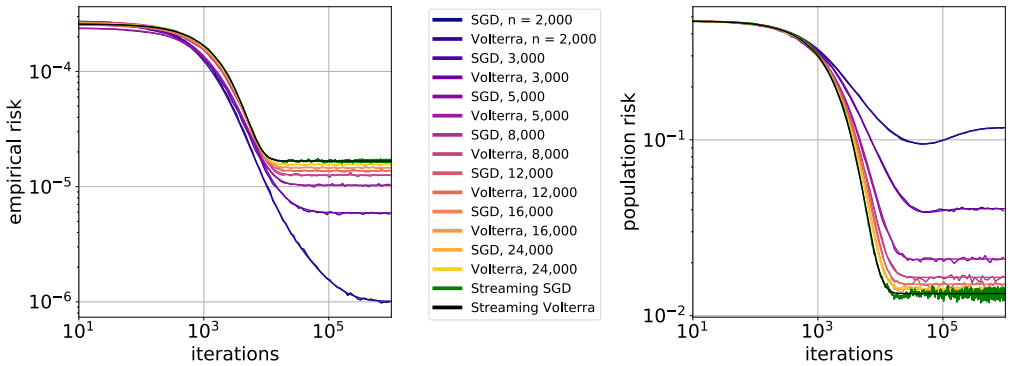

Figure 3: **Single runs of SGD vs. HSGD (Volterra) for a simple Gaussian linear regression problem**, with increasing number of samples n and $d = 2000$. Empirical risk (left) increases monotonically with n to a limit while population risk (right) decreases monotonically in n. Streaming corresponds to $n = \infty$. Covariance of Gaussian samples is $\mathbf{I}_d$, with target given by $b = a \cdot \beta + \eta Z$ for $\eta = 0.2$ and β a unit vector and $Z \sim N(0, 1)$. For consistency across sample sizes, time is measured in iterations.

4 Streaming SGD

In this section we introduce constant learning rate ($\gamma(t) \equiv \gamma$) streaming SGD. Let $(\boldsymbol{a}_k, b_k)_{k=1}^{\infty}$ be iid samples from a $\mathbb{R}^d \times \mathbb{R}$-dimensional distribution $\mathcal{D}$. Streaming SGD using data from $\mathcal{D}$, which we denote $\mathcal{D}$-SGD, is the algorithm

$$\boldsymbol{s}_{k+1} = \boldsymbol{s}_k - \gamma \boldsymbol{a}_k(\boldsymbol{a}_k \cdot \boldsymbol{s}_k - b_k) \quad \text{for} \quad k \in \{0, 1, 2, \dots\}. \tag{13}$$

This naturally describes one-pass SGD, in which data points are used only once. If $\mathcal{R}$ is the expected risk (i.e. population risk), and if $\mathcal{R}$ is given by $\frac{1}{2}\mathbb{E}\,(\boldsymbol{a} \cdot \boldsymbol{x} - b)^2$ with $(\boldsymbol{a}, b) \sim \mathcal{D}$, then $\mathcal{D}$-SGD is directly solving the population risk minimization. This is an idealized situation as one does not have access to infinite data in practice.

Deterministic behavior of streaming risks and comparison to HSGD. $\mathcal{D}$-SGD can encompass multi-pass SGD by letting $(\boldsymbol{a}_k, b_k)_{k=1}^{n}$ be the first n iid samples from $\mathcal{D}$ and considering $\widehat{\mathcal{D}}$-SGD for

$$\widehat{\mathcal{D}}_n \stackrel{\text{def}}{=} \frac{1}{n} \sum_{k=1}^{n} \delta_{(\boldsymbol{a}_k, b_k)}. \tag{14}$$

This means that a sample from $\widehat{\mathcal{D}}_n$ (*conditionally* on the dataset) is distributed like $(\boldsymbol{a}_I, b_I)$, where I is an uniformly random choice of index.

To enable a comparison of streaming SGD to multi-pass SGD, we suppose for some n the matrix $[\boldsymbol{A} \,|\, \boldsymbol{b}]$ is $n \times (d+1)$ matrix whose rows are iid samples from $\mathcal{D}$. We construct the empirical risk $\mathcal{L}$ from these n samples as in (2). We then define the streaming loss

$$\mathcal{S}(\boldsymbol{x}) \stackrel{\text{def}}{=} \tfrac{1}{2}\,\mathbb{E}(\boldsymbol{a} \cdot \boldsymbol{x} - b)^2 = \tfrac{1}{2n}\,\mathbb{E}\,\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 = \tfrac{1}{n}\,\mathbb{E}\,\mathcal{L}(\boldsymbol{x}), \quad \text{where} \quad (\boldsymbol{a}, b) \sim \mathcal{D}.$$

The associated homogenized SGD representation, which we call homogenized $\mathcal{D}$-SGD, is

$$\mathrm{d}\boldsymbol{Y}_t \stackrel{\text{def}}{=} -\gamma(\nabla \mathcal{S}(\boldsymbol{Y}_t))\,\mathrm{d}t + \gamma\sqrt{2\mathcal{S}(\boldsymbol{Y}_t)(\nabla^2 \mathcal{S})}\,\mathrm{d}\boldsymbol{B}_t \quad \text{for} \quad t \geq 0. \tag{15}$$

This naturally leads to Volterra dynamics in which $\nabla^2 \mathcal{L}$ is replaced by $\nabla^2 \mathcal{S}$ in Eqs. (8)-(10) whose solution we denote with Ψ_t^s and Ω_t^s (see Appendix B).

We prove a weak equivalence between HSGD and homogenized $\mathcal{D}$-SGD in the following theorem.

Theorem 6. *Suppose $\boldsymbol{X}_0 = \boldsymbol{Y}_0$. After rescaling time,*

$$\lim_{n \to \infty} \boldsymbol{X}_{t/n} = \boldsymbol{Y}_t \quad \textit{uniformly on compact sets of time a.s.}$$

Furthermore, GF converges, $\mathcal{X}^{gf}_{\gamma t/n} \to \mathcal{X}^{s\text{-}gf}_{\gamma t}$, and the Volterra equations converge, $\Psi_{t/n} \to \Psi_t^s$ and $\Omega_{t/n} \to \Omega_t^s$, uniformly on compact sets of time.

We expect this to hold in much greater generality as suggested by Figures 1 and 3. This is an immediate consequence of the law of large numbers due to which $\frac{1}{n}\mathcal{L}$ converges almost surely to $\mathcal{S}$; the applications in which we are typically interested would take time large as a function of n (the numerical results are extremely strong, see Figures 3 and 1), and we leave a deeper mathematical investigation of this point as an open question. We also note that for t very large with n, there is likely another behavior that takes hold. For example, when $t = n$, observe that $\boldsymbol{X}_{t/n}$ will have used approximately $(1 - 1/e)n$ samples, whereas $\boldsymbol{Y}_n$ will have used n, and hence we expect a breakdown in the connection. Other works also examined the expected risk of streaming including [7, 54] but not as the limit of the dynamics of multi-pass SGD as we have done.

Features of generalization. While there are connections between streaming and multi-pass SGD, certain behavior is only accessible in the multi-pass setting. For example, early stopping can be a useful ingredient in avoiding overfitting when learning overparameterized models. However, late-time overfitting is only observable with multi-pass SGD and does not occur for streaming SGD (see the nonmonotonicity in population risk in Fig. 3).

Similarly, comparing streaming and multi-pass SGD has been suggested as a method for analyzing generalization. The bootstrap risk was introduced in [34] as an intepretable component in a decomposition of the population risk. It is defined as the excess risk of SGD for ERM $\mathcal{L}$ with n iid samples from $\mathcal{D}$ when compared to $\mathcal{D}$-SGD, i.e.

$$\underbrace{\Omega_{t/n} - \Omega_t^{\mathrm{s}}}_{\text{bootstrap risk}} = \underbrace{\mathcal{R}\big(\mathfrak{X}^{\mathrm{gf}}_{\gamma t/n}\big) - \mathcal{R}\big(\mathfrak{X}^{\text{s-gf}}_{\gamma t}\big)}_{\text{model risk}} + \underbrace{\int_0^{t/n} K(t/n, u; \nabla^2\mathcal{R})\Psi_u \,\mathrm{d}u - \int_0^t K^{\mathrm{s}}(t, u; \nabla^2\mathcal{R})\Psi_u^{\mathrm{s}} \,\mathrm{d}u}_{\text{SGD bootstrap risk}}. \tag{16}$$

Thus the Volterra equations can be used to give an exact expression for the bootstrap risk. In particular, we can predict the iteration at which the bootstrap risk becomes large as the streaming and multi-pass risks bifurcate. This allows for quantitative prediction of the stopping time in Claim 1 of [34], a central conjecture of their paper.

On taking time to infinity, we can further evaluate the SGD bootstrap risk.

Theorem 7. *If $\gamma \le \min\{2(\frac{1}{n}\operatorname{tr}(\boldsymbol{A}^T\boldsymbol{A}))^{-1}, 2(\operatorname{tr}\nabla^2\mathcal{S})^{-1}\}$, then the limiting bootstrap risk is given by*

$$\lim_{t\to\infty} \Omega_{t/n} - \Omega_t^s = \mathcal{R}\big(\mathfrak{X}^{gf}_\infty\big) - \mathcal{R}\big(\mathfrak{X}^{s\text{-}gf}_\infty\big) + \frac{\gamma\Psi_\infty}{2n} \times \operatorname{tr}\left\{\frac{(\nabla^2\mathcal{R})(\nabla^2\mathcal{L})}{\nabla^2\mathcal{L} + \delta\boldsymbol{I}_d}\right\}$$

Here Ψ_∞ is the limiting training losses given by Theorem 3, i.e.

$$\Psi_\infty = \mathcal{L}\big(\mathfrak{X}^{gf}_\infty\big) \times \left(1 - \frac{\gamma}{2n}\operatorname{tr}\left\{\frac{(\nabla^2\mathcal{L})^2}{\nabla^2\mathcal{L} + \delta\boldsymbol{I}_d}\right\}\right)^{-1}.$$

5 Conclusion.

Using a specific type of SLD (called HSGD) that matches the second-order correlations in the noise of SGD, we demonstrated that their empirical and population risks match in the high-dimensional limit. Moreover, the risks of HSGD behavior deterministically, as described by a Volterra equation. With this connection, we investigated the benefits of SGD on a convex objective. While there is no statistical benefit to generalization from the noise of SGD, in overparameterized, interpolating settings little is lost compared to GD. Moreover, when computational restrictions are imposed, SGD can be radically faster than GD because of its dependence on a different condition number of the Hessian. We characterized this speed up using the ICR, which when calculated for datasets common in deep learning clearly favors SGD. This should highlight the difficulty in studying implicit regularization for SGD empirically: any experiment necessarily has a finite computational budget and may find lower population risks with SGD simply via its improved conditioning. Finally, we demonstrated limitations in using streaming SGD alone as a tool for studying generalization.

As future work, a proper theoretical connection between streaming and multi-pass SGD remains to be made. A major outstanding problem (both theoretically and empirically) is extending the analysis above to non-quadratic losses, both train and test, and especially to other high-dimensional problems not in the kernel regime. Finally, data augmentation can naturally be considered by randomly augmenting each sample from $\widehat{\mathcal{D}}_n$ in Eq. (14).

References

[1] Ahmed Abdel-Gawad and Simon Ratner. Adaptive optimization of hyperparameters in l2-regularised logistic regression. *Technical report*, 2007.

[2] B. Adlam and J. Pennington. The Neural Tangent Kernel in High Dimensions: Triple Descent and a Multi-Scale Theory of Generalization. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 74–84. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/adlam20a.html.

[3] B. Adlam and J. Pennington. Understanding Double Descent Requires A Fine-Grained Bias-Variance Decomposition. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 11022–11032, 2020.

[4] Ben Adlam, Jake A Levinson, and Jeffrey Pennington. A random matrix perspective on mixtures of nonlinearities in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 3434–3457. PMLR, 2022.

[5] S. Asmussen. *Applied probability and queues*, volume 51 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 2003. Stochastic Modelling and Applied Probability.

[6] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). *Advances in neural information processing systems*, 26, 2013.

[7] B. Bordelon and C. Pehlevan. Learning Curves for SGD on Structured Features. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=WPI2vbkAl3Q.

[8] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[9] Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, 2012.

[10] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. *Advances in neural information processing systems*, 20, 2007.

[11] Pratik Chaudhari and Stefano Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, 2018.

[12] Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18 (1):3520–3570, 2017.

[13] M. Engeli, Th. Ginsburg, H. Rutishauser, and E. Stiefel. Refined iterative methods for computation of the solution and the eigenvalues of self-adjoint boundary value problems. *Mitt. Inst. Angew. Math. Zürich*, 8:107, 1959.

[14] Rong Ge, Sham M Kakade, Rahul Kidambi, and Praneeth Netrapalli. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. *Advances in Neural Information Processing Systems*, 32, 2019.

[15] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[16] G. Gripenberg. On the resolvents of nonconvolution Volterra kernels. *Funkcial. Ekvac.*, 23(1):83–95, 1980. ISSN 0532-8721. URL http://www.math.kobe-u.ac.jp/~fe/xml/mr0586277.xml.

[17] M. Gurbuzbalaban, U. Simsekli, and L. Zhu. The Heavy-Tail Phenomenon in SGD. *arXiv preprint arXiv:2006.04740*, 2020.

[18] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48 of *Proceedings of Machine Learning Research*, pages 1225–1234. PMLR, 20–22 June 2016.

[19] S. Hochreiter and J. Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

[20] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems (NeurIPS)*, 2018.

[21] P. Jain, S. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Accelerating Stochastic Gradient Descent for Least Squares Regression. In *Proceedings of the 31st Conference On Learning Theory (COLT)*, volume 75 of *Proceedings of Machine Learning Research*, pages 545–604. PMLR, 2018.

[22] Stanisław Jastrzębski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

[23] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[24] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, U. of Toronto, 2009.

[25] Daniel Kunin, Javier Sagastuy-Brena, Lauren Gillespie, Eshed Margalit, Hidenori Tanaka, Surya Ganguli, and Daniel LK Yamins. Rethinking the limiting dynamics of sgd: modified loss, phase space oscillations, and anomalous diffusion. *arXiv preprint arXiv:2107.09133*, 2021.

[26] Y. LeCun, C. Cortes, and C. Burges. "mnist" handwritten digit database, 2010. URL http://yann.lecun.com/exdb/mnist.

[27] J. Lee, Y. Bahri, R. Novak, S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep Neural Networks as Gaussian Processes. In *International Conference on Learning Representations (ICLR)*, 2018.

[28] Yunwen Lei, Ting Hu, and Ke Tang. Generalization performance of multi-pass stochastic gradient descent with convex loss functions. *J. Mach. Learn. Res.*, 22:25–1, 2021.

[29] Z. Liao, R. Couillet, and M. Mahoney. A Random Matrix Analysis of Random Fourier Features: Beyond the Gaussian Kernel, a Precise Phase Transition, and the Corresponding Double Descent. *arXiv preprint arXiv:2006.05013*, 2020.

[30] Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *arXiv preprint arXiv:1704.04289*, 2017.

[31] Vladimir A Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1967.

[32] S. Mei and A. Montanari. The generalization error of random features regression: precise asymptotics and the double descent curve. *Comm. Pure Appl. Math.*, 75(4):667–766, 2022.

[33] T. Mori, L. Ziyin, K. Liu, and M. Ueda. Logarithmic landscape and power-law escape rate of sgd. *arXiv preprint arXiv:2105.09557*, 2021.

[34] P. Nakkiran, B. Neyshabur, and H. Sedghi. The Deep Bootstrap Framework: Good Online Learners are Good Offline Generalizers. In *International Conference on Learning Representations (ICLR)*, 2021.

[35] R.M. Neal. *Priors for Infinite Networks*, pages 29–53. Springer New York, New York, NY, 1996. doi: 10.1007/978-1-4612-0745-0_2. URL https://doi.org/10.1007/978-1-4612-0745-0_2.

[36] C. Paquette and E. Paquette. . In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

[37] C. Paquette, B. van Merriënboer, and F. Pedregosa. Halting Time is Predictable for Large Models: A Universality Property and Average-case Analysis. *arXiv preprint arXiv:2006.04299*, 2020.

[38] C. Paquette, K. Lee, F. Pedregosa, and E. Paquette. SGD in the Large: Average-case Analysis, Asymptotics, and Stepsize Criticality. *arXiv preprint arXiv:2102.04396*, 2021.

[39] E. Paquette, C. Paquette, B. Adlam, and J. Pennington. Homogenization of SGD in high-dimensions: exact dynamics and generalization properties. *arXiv preprint arXiv:2205.07069*, 2022.

[40] F. Pedregosa. A Hitchhiker's Guide to Momentum, 2021. URL http://fa.bianp.net/blog/2021/hitchhiker/.

[41] B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 04, 1964.

[42] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1177–1184, 2008.

[43] H. Robbins and S. Monro. A Stochastic Approximation Method. *Ann. Math. Statist.*, 1951.

[44] Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*, 2016.

[45] Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

[46] Christopher J Shallue, Jaehoon Lee, Joseph Antognini, Jascha Sohl-Dickstein, Roy Frostig, and George E Dahl. Measuring the effects of data parallelism on neural network training. *Journal of Machine Learning Research*, 20:1–49, 2019.

[47] V. Shankar, A. Fang, W. Guo, S. Fridovich-Keil, J. Ragan-Kelley, L. Schmidt, and B. Recht. Neural Kernels Without Tangents. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 8614–8623. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/shankar20a.html.

[48] Samuel Smith, Erich Elsen, and Soham De. On the generalization benefit of noise in stochastic gradient descent. In *International Conference on Machine Learning*, pages 9058–9067. PMLR, 2020.

[49] Samuel L Smith and Quoc V Le. A bayesian perspective on generalization and stochastic gradient descent. *arXiv preprint arXiv:1710.06451*, 2017.

[50] Samuel L Smith, Benoit Dherin, David GT Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. *arXiv preprint arXiv:2101.12176*, 2021.

[51] N. Tripuraneni, B. Adlam, and J. Pennington. Covariate Shift in High-Dimensional Random Feature Regression. *arXiv preprint arXiv:2111.08234*, 2021.

[52] Jingfeng Wu, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham M Kakade. Last iterate risk bounds of sgd with decaying stepsize for overparameterized linear regression. *arXiv preprint arXiv:2110.06198*, 2021.

[53] Lei Wu, Chao Ma, and Weinan E. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/6651526b6fb8f29a00507de6a49ce30f-Paper.pdf.

[54] L. Ziyin, K. Liu, T. Mori, and M. Ueda. Strength of Minibatch Noise in SGD. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=uorVGbWV5sw.

[55] D. Zou, J. Wu, V. Braverman, Q. Gu, and S. M. Kakade. Risk Bounds of Multi-Pass SGD for Least Squares in the Interpolation Regime. *arXiv preprint arXiv:2203.03159*, 2022.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We make sure to clearly state our assumptions particular that the results only hold for the ℓ^2 regularized least squares problem.
   (b) Did you describe the limitations of your work? [Yes] We ensure the reader is aware that many of our results follow from another paper. We also make clear which results are proven and which results speculative, but have numerical support.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We address the potential negative societal impacts in the appendix. We do not anticipate any ethical or societal issues. The results presented in this paper concern the analysis of existing methods.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] The work presented is purely theoretical.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] The full set of assumptions on the data matrix $\boldsymbol{A}$, signal $\boldsymbol{\beta}$, initialization $\boldsymbol{x}_0$, and target vector $\boldsymbol{b}$ are listed in the theorems statements with a complete description in Appendix A. Moreover we state our setting, namely that we are working on the ℓ^2 regularized least squares problem, clearly in the introduction of the main paper.
   (b) Did you include complete proofs of all theoretical results? [Yes] All proofs of theoretical results are presented and, if not, referenced to where the proof can be found. We also clearly state conjectures which we believe should be true (see e.g., streaming SGD in Section 4).
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We provide code in our supplemental material and describe our experimental simulations in detail in Appendix E. We state all parameters in the captions of the figures.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] All parameter choices are clearly stated under each figure and in Appendix E. For example, we give the exact dimensions of the matrices used as well as the learning rates on SGD.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Variance across the runs of SGD were reported.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We indicated the type of resources in Appendix E. All our experiments were done using Colabs and only required a single standard GPU.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We use three data sets to verify our claims. All three data sets (MNIST [26], CIFAR-10 [24], and CIFAR-5m are open source and available in TensorFlow. We cited the creators in the main text.
   (b) Did you mention the license of the assets? [Yes] All three data sets used are open source and available on TensorFlow.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include the code used our simulations as part of the supplemental material.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] The data sets are open source.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] The data sets used do not contain any personally identifiable information or offensive content. The data sets are standard data sets used in benchmarking.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] The work is purely theoretical and does not require any crowdsourcing or human subjects.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] The work is purely theoretical and does not require any crowdsourcing or human subjects.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] The work is purely theoretical and does not require any crowdsourcing or human subjects.