Content:

- Inductive biases of gradient descent: low rank bias in DLNs, edge of stability and other phenomena
- Langevin model of SGD: SGD as approximate Bayesian inference
- Anisotropic noise introduced implicit regularization that enables to learn sparse solutions
- Stochasticity plays a role: degeneracies are sticky for SGD
- Comparing different models of SGD: what statistics to track?
- Escape time around local mimima: Kramers
- SGD on low loss converges toward more degenerate solutions
- Metastability: no work on this yet for saddle to saddle dynamics and degenerate models?
- Mixing times: how to study it?
- Limits to SDE models of SGD (large learning rate)
- Alternative approaches to SGD:
- Q. what do we want from SGD, especially for safety?

Some related work on inductive biases of SGD (emphasis on anisotropic noise)

This is a shallow review where I write the results that seem most salient, it's not a comprehensive review. First, let's start with the inductive biases of gradient descent flow. In the continuous limit, gradient flow writes:

$$\dot{\theta} = -\nabla L_N(\theta) \tag{1}$$

One central result comes from the exact study of gradient flow on deep linear networks (DLNs). In such a setting it has been shown that gradient descent has an implicity bias towards low rank solutions [Saxe et al., 2013]. It has also been shown that in deep learning, GD works at the edge of stability [Cohen et al., 2021]. The edge of stability means that GD is implicitly aware of the curvature by escaping sharp minima with curvature above $\frac{2}{\eta}$, where η is the learning rate. This can be seen by looking at the maximum eigenvalue of the Hessian during the training process. In other words, there is some flatness bias during gradient descent. In the continuous limit, the edge of stability behaviour can be modelled with a central flow term [Cohen et al., 2024]. We can write the continuous limit as:

$$\boxed{\frac{d\theta}{dt} = -\eta\left(\nabla L(\theta) \; + \; \frac{1}{2}\nabla_\theta \left\langle H(\theta), \Sigma(t)\right\rangle\right)}$$

Where H is the Hessian matrix, $\Sigma(t) = \mathbb{E}[(\theta_t - \theta(t))(\theta_t - \theta(t))^\top]$ is the covariance of fast oscillations. Another interesting bias that has been noted in deep linear networks it a bias toward aligning the weight matrix of neural networks: during GD, layers become aligned [Ji and Telgarsky, 2018]. This can be seen by showing that the scalar product between left and right singular vectors tend to 1 in the long time limit. Another interesting bias is a bias toward learning low frequency solutions first [Xu et al., 2024].

Beyond the inductive biases of GD, the stochasticity in SGD induces other inductive biases. By modelling SGD as a Langevin dynamics with anisotropic noise (ALD):

$$\dot{\theta}(t) = -\nabla L_N(\theta(t)) + \sqrt{2\eta\Sigma(\theta(t))}dW(t) \tag{2}$$

where $W(t)$ is a Brownian motion. We can see that noise anisotropy can shrink parameters on degenerate critical directions.

Some previous work suggest that SGD has a bias toward solution that generalize better than GD. In diagonal linear networks, it has been shown that SGD induces sparser solutions than GD by minimizing an implicit objective function inducing sparser parameters [Pesme et al., 2021]. Some other work in rank 1 over-parameterized deep linear networks suggests that SGD is less dependent on initialization than GD [Lyu and Zhu, 2023]. Modelling the noise of ALD as a labelled noise in 2 layers LNs, it has been shown that SGD induces a further bias toward low rank solutions[Varre et al., 2024]:

$$\Theta_{t+1} \;=\; \Theta_t \;-\; \eta\,\Theta_t J_t \;-\; \eta\sqrt{\delta}\,\Theta_t\,\xi_t, \tag{3}$$

where

$$J_t = \begin{bmatrix} \mathbf{0}_{p\times p} & X^\top\big(Y - XW_{1,t}W_{2,t}\big) \\ \big(X^\top\big(Y - XW_{1,t}W_{2,t}\big)\big)^\top & \mathbf{0}_{k\times k} \end{bmatrix}, \qquad \xi_t = \begin{bmatrix} \mathbf{0}_{p\times p} & X^\top\varepsilon_t \\ \varepsilon_t^\top X & \mathbf{0}_{k\times k} \end{bmatrix},$$

and the label-noise satisfies $\varepsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{n\times k})$.

$$\mathrm{d}\Theta \;=\; \Theta\Big[\,J\,\mathrm{d}t + \sqrt{\eta\delta}\,\mathrm{d}\xi\Big], \tag{4}$$

with $\mathrm{d}\xi$ sharing the same block structure as ξ_t but driven by Brownian motion.

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\det\big(\Theta^\top\Theta\big) \;=\; 0, \tag{5}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\det\big(\Theta^\top\Theta\big) \;=\; -\,2\eta\delta\,k\;\mathrm{tr}\big(X^\top X\big)\;\det\big(\Theta^\top\Theta\big), \tag{6}$$

whose explicit solution is

$$\det\big(\Theta(t)^\top\Theta(t)\big) \;=\; \det\big(\Theta_0^\top\Theta_0\big)\;\exp\Big[-2\eta\delta\,k\;\mathrm{tr}(X^\top X)\,t\Big].$$

The noise in ALD diminishes the determinant along the trajectory, leading to a simplification of the network over time. The larger the noise and the stepsize, the faster the determinant vanishes. The vanishing of the determinant suggests that the rank of the parameters decreases by at least one, effectively eliminating some components. Note: we could look into sub-determinants of the dynamics. Even though we have a stochastic process, the determinant satisfies a deterministic differential equation.

SGD as GD plus label noise. In another paper, it has been shown that SGD can induces a bias toward sparse feature in diagonal linear networks using a labelled noise model of the noise [Andriushchenko et al., 2023]. For a mini-batch of size 1, the standard update

$$\theta_{t+1} = \;\theta_t \;-\; \eta\,\big(h_{\theta_t}(x_{i_t}) - y_{i_t}\big)\,\nabla_\theta h_{\theta_t}(x_{i_t}) \tag{7}$$

can be rewritten exactly as *full–batch* GD on *corrupted labels* $\tilde{y}_{t,i} = y_i + \xi_{t,i}$:

$$\theta_{t+1} = \;\theta_t \;-\; \frac{\eta}{n}\sum_{i=1}^{n}\big(h_{\theta_t}(x_i) - \tilde{y}_{t,i}\big)\,\nabla_\theta h_{\theta_t}(x_i), \qquad \xi_{t,i} \;=\; \big(h_{\theta_t}(x_i) - y_i\big)\,\big(1 - n\,\mathbf{1}_{\{i=i_t\}}\big). \tag{8}$$

The label-noise is mean-zero and its variance scales with the instantaneous empirical loss, $\mathbb{E}\big[\|\xi_t\|^2\big] = 2n(n-1)L(\theta_t)$, so the noise naturally *vanishes at convergence* and is largest when the model is still far from fitting the data [**?**].

Effective SDE on the loss plateau. When a large learning-rate causes the training loss to stabilise at some plateau $L(\theta_t) \simeq \delta > 0$, the discrete dynamics (8) is well approximated (in the Stratonovich sense) by

$$\boxed{\mathrm{d}\theta_t \;=\; -\nabla_\theta L(\theta_t)\,\mathrm{d}t \;+\; \sqrt{\eta\,\delta}\;\Phi_\theta(X)^\top \mathrm{d}B_t} \tag{9}$$

where

- $L(\theta) = \frac{1}{2}\sum_{i=1}^{n}\big(h_\theta(x_i) - y_i\big)^2$ is the empirical loss,
- $\Phi_\theta(X) = \big[\nabla_\theta h_\theta(x_1), \ldots, \nabla_\theta h_\theta(x_n)\big]^\top \in \mathbb{R}^{n\times \dim(\theta)}$ is the Jacobian (a.k.a. NTK feature matrix),
- B_t is an n-dimensional Brownian motion,
- $\delta \simeq L(\theta_t)$ is (approximately) constant along the plateau.

Because the noise lies in the *same span* as the gradients, it systematically pushes the parameters toward directions that also reduce the loss, yielding an implicit preference for "simpler" (sparser) representations.

Diagonal linear network: explicit calculation. Consider the two-layer *diagonal linear* model $h_{u,v}(x) = \langle u \odot v, x\rangle$ with $\theta = (u, v) \in \mathbb{R}^d \times \mathbb{R}^d$. On the plateau, (9) becomes coordinate-wise

$$\mathrm{d}u_t \;=\; -\nabla_u L(u_t, v_t)\,\mathrm{d}t \;+\; \sqrt{\eta\,\delta}\;v_t \odot X^\top \mathrm{d}B_t, \qquad \mathrm{d}v_t \;=\; -\nabla_v L(u_t, v_t)\,\mathrm{d}t \;+\; \sqrt{\eta\,\delta}\;u_t \odot X^\top \mathrm{d}B_t, \tag{10}$$

where $X = [x_1, \ldots, x_n]^\top$. Lochrie *et al.* show that:

1. if feature j is *off-support* ($u_j^\star v_j^\star = 0$ in the ground truth), then $|u_{t,j}|$, $|v_{t,j}|$ shrink exponentially fast;
2. for on-support coordinates, $(u_{t,j}, v_{t,j})$ stay inside an $O\big(\sqrt{\eta\,\delta}\big)$ band and converge (in $O(\delta^{-1})$ time) to the *sparsest* interpolator.

Thus, the multiplicative noise operates as a *soft* ℓ_0 *penalty* that prunes irrelevant features during the plateau. The multiplicative structure $\phi_\theta(X)^{!\top}\mathrm{d}B_t$ makes the noise amplitude self-proportional to each feature's strength. That converts the stochastic term into a geometric Brownian motion acting on every column norm, whose intrinsic negative drift drives unnecessary features toward zero while leaving only those columns that the gradient term must keep alive to fit the data. The result is the systematic sparsification observed across architectures.

Conjecture — general sparse-feature learning. Because the stochastic term in (9) always lives in $\mathrm{span}\big\{\nabla_\theta h_\theta(x_i)\big\}_{i=1}^{n}$, it is expected to minimise the ℓ_2 column norms of $\Phi_\theta(X)$, thereby promoting *Jacobian sparsity* irrespective of architecture. Empirically, Lochrie *et al.* observe a sharp drop in both *feature-sparsity coefficient* and the rank of $\Phi_\theta(X)$ across diagonal linear nets, shallow ReLU MLPs, ResNets and DenseNets once training enters the loss plateau. We conjecture that this *multiplicative-noise sparsification* is a universal inductive bias of SGD at large learning rates.

Implicit simplification by stochastic collapse [Chen et al., 2023]

1. **SGD as SGF:**
$$\boxed{\mathrm{d}\theta_t = -\nabla L(\theta_t)\,\mathrm{d}t + \sqrt{\tfrac{\eta}{\beta}}\,\Sigma(\theta_t)\,\mathrm{d}B_t} \tag{SC 1}$$
where Σ is the square-root covariance of per-sample gradients.

2. **Invariant sets.** Sign sets $w_{\text{in}} = w_{\text{out}} = 0$ (sparsity) and permutation sets $w_{\text{in},p} = w_{\text{in},q},\ w_{\text{out},p} = w_{\text{out},q}$ (rank-reduction) satisfy $Q\theta = \theta$ for a reflection matrix Q.

3. **Attractivity criterion (1-D):** Near $A = \{0\}$ $\mathrm{d}\theta_t \simeq -L''(0)\theta_t\,\mathrm{d}t + \sqrt{D''(0)}\theta_t\,\mathrm{d}B_t$; stochastic attraction occurs iff
$$\boxed{L''(0) + \tfrac{1}{2}D''(0) > 0} \tag{SC 2}$$
so stronger noise broadens the set of attractive curvatures.

4. **Sign-set example:** For $f(x) = w_2\sigma(w_1 x)$ under label-noise (ζ) SGD,
$$A = \{w_1 = w_2 = 0\} \text{ is attractive if } \frac{\sigma'(0)\eta\zeta^2}{2} > \frac{|\sum_i x_i y_i|}{\sum_i x_i^2}. \tag{SC 3}$$

5. **Rank-collapse SDE (teacher–student):**
$$\boxed{\mathrm{d}s_i = 2s_i\big((\tilde{s}_i + \tfrac{1}{2}\eta\zeta^2) - s_i\big)\mathrm{d}t + 2\sqrt{\eta\zeta^2}\,s_i\,\mathrm{d}B_t} \tag{SC 4}$$
Modes with $\tilde{s}_i < \eta\zeta^2/2$ shrink to zero, yielding a low-rank bias.

SGD $\Longrightarrow$ degenerate SDE from [Barbieri et al., 2025] Mini-batch SGD with learning-rate η and batch-size b is the Euler–Maruyama scheme of the *state-dependent* SDE
$$dX_t = -\nabla L(X_t)\,dt + \sqrt{\tfrac{\eta}{b}}\,Q(X_t)^{1/2}\,dW_t, \qquad \varepsilon^2 := \frac{\eta}{2b}, \tag{11}$$
where $Q(x) = \operatorname{cov}\big[\nabla L_i(x)\big] \succeq 0$ is typically **degenerate** (rank $\le N - 1 < d$ in the over-parameterised regime).

Fokker–Planck formulation. The transition density $\rho(t, x)$ obeys
$$\partial_t\rho = \nabla\cdot\big(\varepsilon^2\nabla\cdot(Q\rho) + \rho\nabla L\big). \tag{12}$$

1. **Drift regime – Local mass concentration.** If L is λ–convex in $B_{(1+\delta)R_0}$ and $0 \le Q \preceq \sigma I$, then [, Thm. 1.2]
$$\boxed{\int_{|x|\le R_0 e^{-\frac{\lambda}{2}(t-t_0)}} \rho(t,x)\,dx \;\ge\; \int_{|x|\le R_0} \rho(t_0,x)\,dx \;-\; \beta \quad \text{for } 0 < t < t_0 + T_\varepsilon}$$
with $T_\varepsilon = \frac{2}{\lambda}\log\big(\frac{R_0}{a\varepsilon^\alpha}\big)$. Hence small effective noise keeps most probability mass near the nearest minimum for an *exponentially long* time in ε.

2. **Diffusion regime – Mean Exit Time (MET).** For a ball $B_{R_0}(x_0)$ and any $r < R_0$:
$$\underline{\text{Lower bound:}} \quad \mathbb{E}[\tau_x^{B_{R_0}}] \;\ge\; \frac{R_0^2 - r^2}{2\varepsilon^2\sigma d}, \tag{13}$$
$$\underline{\text{Upper bound:}} \quad \mathbb{E}[\tau_x^{\Omega}] \;\le\; \frac{2}{\Lambda}\left(e^{\frac{\Lambda R_0^2}{2\beta\varepsilon^2}} - e^{\frac{\Lambda r^2}{2\beta\varepsilon^2}}\right) \quad \big(\text{if } v^\top Q v \ge \beta > 0 \text{ and } (v\cdot\nabla L)(v\cdot x) \le \Lambda(v\cdot x)^2\big). \tag{14}$$
Even very degenerate noise gives finite escape times as soon as it is non-zero in *one* direction.

3. **Steady-state existence for degenerate Q.** If Q is globally bounded and Lipschitz (9) and L is coercive and one-sided Lipschitz (10)–(11), there exists an invariant probability ρ_∞ solving

$$\boxed{\nabla\cdot\big(\varepsilon^2\nabla\cdot(Q\rho_\infty)+\rho_\infty\nabla L\big)=0}$$

4. **Convergence to equilibrium.** *(i) Noisy-SGD:* adding an isotropic δI to Q yields a uniformly elliptic SDE whose law converges (qualitatively) to the unique steady state via the duality method. *(ii) Constant–Q_0, quadratic L:* if C is positive stable and $\ker Q_0$ satisfies Hörmander's condition, the relative entropy decays exponentially

$$\mathcal{E}\big(\rho(t)\mid\rho_\infty\big)\ \le\ c\,e^{-2\gamma t}\,\mathcal{E}\big(\rho_0\mid\rho_\infty\big),$$

giving quantitative rates. *(iii) Non-Hörmander block-degenerate case:* a 2-Wasserstein convergence still holds, with explicit moment decay (Theorem 1.6).

SGD behaves like gradient descent plus very anisotropic, state-dependent noise. Early on, that noise is too small to matter; later it becomes the key factor that lets the optimiser hop out of bad basins. How quickly it can do so—and whether it will eventually mix—depends critically on the learning-rate–to-batch-size ratio and on whether any extra, more isotropic noise is present. There are two regimes, drift and diffusion regimes. Pb is that Q is degenerate, but can add some small isotropic noise to break the degeneracy then we have a convergence to an invariant distribution. Convergence to a stationary distribution is otherwise difficult.

Escape times

Def (non-strict saddle point): A non-strict saddle point is a critical point such that the Hessian has a zero eigenvalue and at least one negative eigenvalue (i.e. strict saddle point with a flat direction).
Def (escape time): We want to find the escape time around a non-strict saddle point. Mathematically, let θ^* be a non-strict saddle point. We are interested in the escape time i.e. the time it takes for the trajectory to escape from the saddle point toward lower losses regions.

$$\tau_r = \inf\{t>0 : L(\theta^*)-L(\theta(t))>\epsilon\} \tag{15}$$

For the Langevin SDE, the escape time is given by the Kramers law where:

$$<\tau>\sim \exp[-\beta\frac{\Delta L}{\sigma^2}] \tag{16}$$

From the diffusion theory of SGD [Xie et al., 2020], we can derive the escape time of SGD between non degenerate minima

Hessian-aligned gradient noise. Near a critical point,

$$C(\theta)\ \approx\ \frac{1}{B}\,H(\theta),\qquad D(\theta)=\frac{\eta}{2B}\,H(\theta). \tag{6–7}$$

Continuous-time SGD.

$$d\theta_t=-\nabla L(\theta_t)\,dt+\sqrt{2D(\theta_t)}\,dW_t.$$

Assumptions.

1. *Second-order Taylor* around minima and saddles.
2. *Quasi-equilibrium* inside each valley.
3. *Low temperature* $\varepsilon^2=\eta/2B\ll 1$.

Mean escape time (SGD). For a single most-probable path $a \to b$,

$$\boxed{\tau_{a\to b} = 2\pi \, \frac{1}{|H_{b,e}|} \, \exp\Big[\frac{2B\,\Delta L}{\eta}\Big(\frac{s}{H_{a,e}} + \frac{1-s}{|H_{b,e}|}\Big)\Big]} \qquad \text{(Thm 3.2)}$$

with path factor $s \in (0,1)$.

Corollaries.

- *Flat-minimum bias:* τ grows like $\exp\big[\frac{2B\,\Delta L}{\eta H_{a,e}}\big]$, so flatter minima (small $H_{a,e}$) are exponentially more stable.
- *Batch / LR trade-off:* increasing B or decreasing η multiplies τ by $\exp\big[(B/\eta)\cdot(\ldots)\big]$.
- *Effective dimension:* only directions with large Hessian eigenvalues matter, so minima with few such "outliers" survive longer.

In the power law escape time of SGD paper [Mori et al., 2022], the authors derive an expression for the escape time around a local minimum and saddle point under a few assumptions: parameters obey the quasi-stationary distribution around critical points, escape from critical point is dominated by most probable escape path (whose direction is given by some eigen value of the Hessian), SGD dynamics is restricted to the non zero eigen value of the Hessian (which forms a lower subspace of the loss). In such case the escape rate obeys a power law given by:

$$\tau_r \sim \frac{\sqrt{h_e^* \, |h_e^s|}}{2\pi} \left[\frac{L(\theta^s)}{L(\theta^*)}\right]^{\left(\frac{B}{\eta\, h_e^*} + 1 - \frac{n}{2}\right)}$$

Multiplicative noise ⇒ loss-proportional covariance.

$$\Sigma(\theta) \;\approx\; \frac{2\,L(\theta)}{B}\, H(\theta_\star), \tag{6}$$

where $H(\theta_\star) = \nabla^2 L(\theta_\star)$ and B is the batch size.

Random-time change ⇒ log-loss SDE. Let $\tau = \displaystyle\int_0^t L(\theta_{t'})\, dt'$ and define $U(\theta) = \log L(\theta)$. Then mini-batch SGD is the Euler–Maruyama scheme of the additive-noise SDE

$$d\theta_\tau \;=\; -\nabla_\theta U(\theta_\tau)\, d\tau + \sqrt{\frac{2\eta}{B}}\; H(\theta_\star)^{1/2}\, dW_\tau, \tag{14}$$

with learning rate η.

Escape rate from a loss basin. Under the assumptions

1. quasi-stationarity of the basin probability mass, and
2. an "outlier" Hessian spectrum of effective dimension n,

the transition rate to the neighbouring saddle θ_s is

$$\boxed{\kappa \;\sim\; \sqrt{\frac{h_e^\star |h_e^s|}{2\pi\, L(\theta_s)}}\, \left(\frac{L(\theta_s)}{L(\theta_\star)}\right)^{-\left(\frac{B}{\eta\, h_e^\star} + \frac{1-n}{2}\right)}} \tag{15}$$

where $h_e^\star$ is the largest positive eigenvalue of $H(\theta_\star)$, $h_e^s < 0$ the negative curvature at the saddle, and n the number of such outlier directions.

Stability threshold for effective dimension. The basin remains stable only while

$$n \;<\; n_{\mathrm{c}} \;:=\; 2\Big(\tfrac{B}{\eta\, h_e^\star} + 1\Big). \tag{16}$$

Important gap: Escape time at non-strict saddle or degenerate minima. What is the quasi-stationary distribution in that case? Also what happens when the noise is anisotropic? Another question is about mixing time in degenerate subspaces. Why do we care about escape times for safety? Empirically, we could check the escape time of SGD, Langevin and power law by estimating the time it takes for the trajectory to go from one saddle point to another. In DLNs, Find the first time at which the loss starts to plateau and the time it takes to escape the plateau. Take the same seed and average over many runs. We can plot the log of escape time with respect to $\log \Delta L$ and with respect to $\Delta \log L$. If anisotropy is present we should observe a power law as opposed to a linear relationship.

We could also look at the mixing time around a non-strict saddle point. Mathematically, let θ^* be a non-strict saddle point. We are interested in the mixing time i.e. the time it takes for the trajectory to mix around the saddle point.

References

Maksym Andriushchenko, Aditya Vardhan Varre, Loucas Pillaud-Vivien, and Nicolas Flammarion. Sgd with large step sizes learns sparse features. In *International Conference on Machine Learning*, pages 903–925. PMLR, 2023.

Davide Barbieri, Matteo Bonforte, and Peio Ibarrondo. Is stochastic gradient descent effective? a pde perspective on machine learning processes. *arXiv preprint arXiv:2501.08425*, 2025.

Feng Chen, Daniel Kunin, Atsushi Yamamura, and Surya Ganguli. Stochastic collapse: How gradient noise attracts sgd dynamics towards simpler subnetworks. *Advances in Neural Information Processing Systems*, 36:35027–35063, 2023.

Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *arXiv preprint arXiv:2103.00065*, 2021.

Jeremy M Cohen, Alex Damian, Ameet Talwalkar, Zico Kolter, and Jason D Lee. Understanding optimization in deep learning with central flows. *arXiv preprint arXiv:2410.24206*, 2024.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018.

Bochen Lyu and Zhanxing Zhu. Implicit bias of (stochastic) gradient descent for rank-1 linear neural network. *Advances in Neural Information Processing Systems*, 36:58166–58201, 2023.

Takashi Mori, Liu Ziyin, Kangqiao Liu, and Masahito Ueda. Power-law escape rate of sgd. In *International Conference on Machine Learning*, pages 15959–15975. PMLR, 2022.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230, 2021.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Aditya Vardhan Varre, Margarita Sagitova, and Nicolas Flammarion. Sgd vs gd: Rank deficiency in linear networks. *Advances in Neural Information Processing Systems*, 37:60133–60161, 2024.

Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. *arXiv preprint arXiv:2002.03495*, 2020.

Zhi-Qin John Xu, Yaoyu Zhang, and Tao Luo. Overview frequency principle/spectral bias in deep learning. *Communications on Applied Mathematics and Computation*, pages 1–38, 2024.