(S)GD dynamics – stochasticity and stepsize effects

Suriya Gunasekar

Joint work with

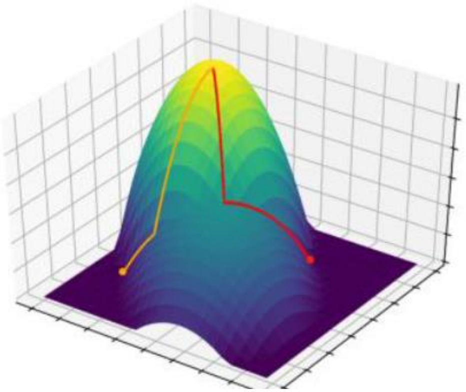
Mathieu Even, Inria, Paris

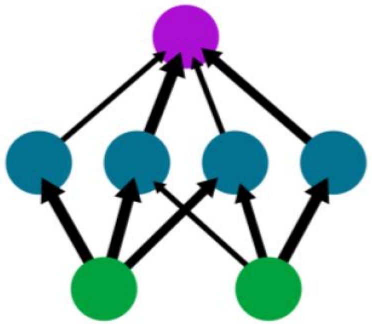
Scott Pesme, EPFL

Nicholas Flammarion, EPFL

Implicit regularization from optimization algorithms

- In overparametrized problems, trajectory of optimization algorithm implicitly introduces structure in the learned solutions!
- and in some cases, the structure is useful for the learning problem
- Understanding generalization from implicit regularization effects is in itself non-trivial and depends on many other factors



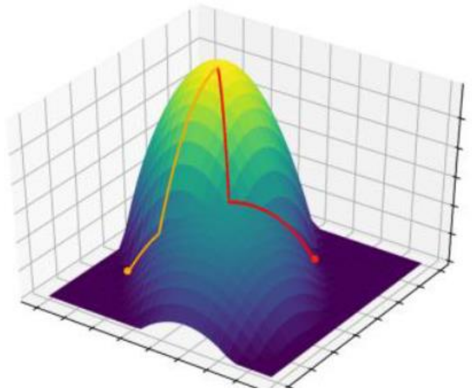


architecture
$\phi: \mathbf{w} \to f_\phi(\mathbf{w}, .)$

dataset $(\mathbf{x}, y) \sim \mathcal{D}$
+augmentations

Alternatively – fun results about optimization

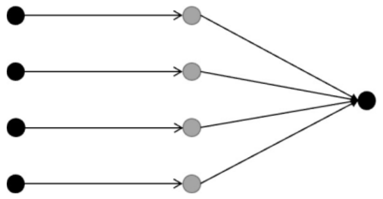
- Overparametrization
- Mirror descent
- Sparse regression



Linear diagonal networks

$$\bar{L}(w) = \|Xw - y\|_2^2 \text{ where } X \in \mathbb{R}^{N \times d} \text{ and } N \ll d$$

$$L(u, v) = \|X(u \odot v) - y\|_2^2$$



- We will look at GD and mini-batch SGD updates $u(t), v(t)$

$$u(t+1) = u(t) - \gamma \nabla_u L_B^{stoc.}\big(u(t), v(t)\big)$$
$$v(t+1) = v(t) - \gamma \nabla_v L_B^{stoc.}\big(u(t), v(t)\big)$$

- and resulting trajectory $w(t) = u(t) \odot v(t)$
- Focus on regression setting for today - many more interesting results when looking at asymptotic minimization with logistic/exponential loss

Some early results...

$$L(u,v) = \|X\,(u \odot v) - y\|_2^2$$

- $w = u \odot v \;\; \equiv \;\; w = w_+^2 - w_-^2$

All (S)GD trajectories coincide with initialization $\begin{Bmatrix} u(0) = \sqrt{2}\alpha \\ v(0) = 0 \end{Bmatrix} = \begin{Bmatrix} w_+(0) = \alpha \\ w_-(0) = \alpha \end{Bmatrix}$

- Gradient flow limit of the algorithm $\gamma \to 0$
- $\boldsymbol{u_\alpha(0) = \alpha 1, v_\alpha(0) = 0}$
- $w_\alpha(t) = u_\alpha(t) \odot v_\alpha(t)$
- As $\alpha \to 0$, if $\lim_{\alpha \to 0} w_\alpha(t)$ if exists, converges to

$$\lim_{\alpha \to 0} w_\alpha(t) \to \operatorname*{argmin}_{Xw=y} \|w\|_1$$

not a kernel regression

- (also, $\alpha \to \infty$, $\lim_{\alpha \to \infty} w_\alpha(t) \to \operatorname*{argmin}_{Xw=y} \|w\|_2$)

Corollaries from G, Woordworth, Bhojanapalli, Neyshabur, Srebro 2017; Chizat, Oyallun, Bach 2019

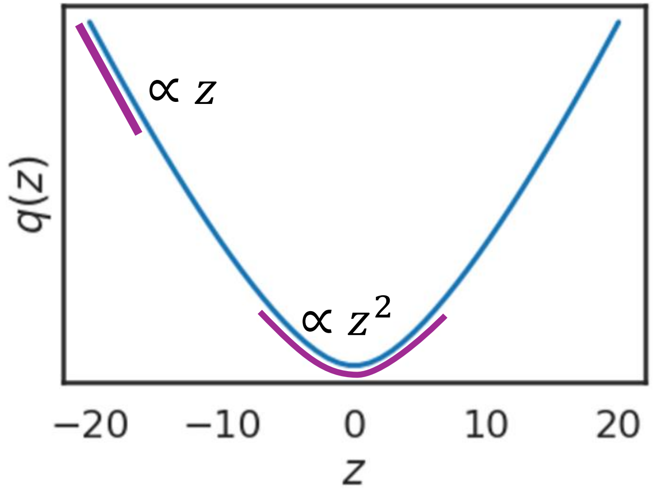
Changing scale of initialization

- Gradient flow limit of the algorithm $\eta \to 0$
- $\boldsymbol{u_{\vec{\alpha}}(0) = \vec{\alpha}, v_{\vec{\alpha}}(0) = 0}$
- For all $\vec{\alpha}$,

$$w_\alpha(t) \to \underset{Xw=y}{\operatorname{argmin}}\, Q_{\vec{\alpha}}(w) := \Sigma_{i\in[d]}\, q\left(\frac{w_i}{\vec{\alpha}_i^2}\right)$$

$\equiv$

Mirror descent on $L(w) = \|Xw - y\|_2^2$
w.r.t hyperentropy potential $Q_{\vec{\alpha}}(w)$



For $\vec{\alpha} = \alpha 1$,

$$Q_\alpha(w) \to \begin{cases} \|w\|_2^2 \; as\ \alpha \to \infty & \#kernel\ regime \\ \|w\|_1 \; as\ \alpha \to 0 & \#sparse\ recovery \end{cases}$$

Woordworth, G, Lee, Moroshko, Savarese, Golan, Soudry, Srebro 2019

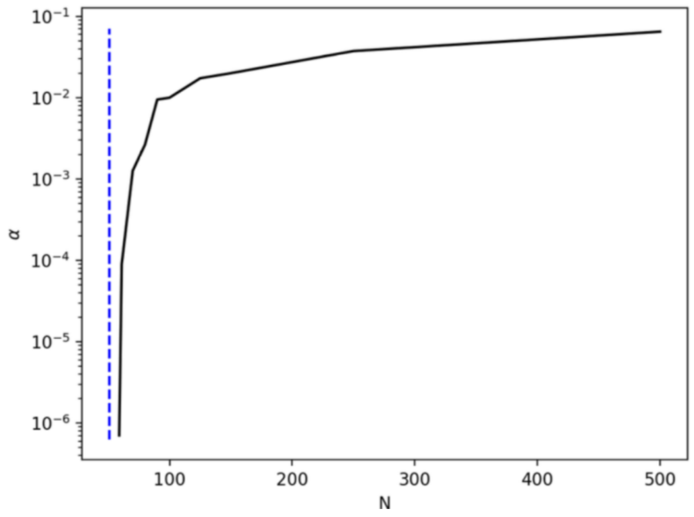
Sparse recovery?

$$w_\alpha(t) \to w_\alpha^\infty = \underset{Xw=y}{\operatorname{argmin}}\, Q_{\vec{\alpha}}(w) \coloneqq \Sigma_{i\in[d]}\, q\left(\frac{w_i}{\vec{\alpha}_i^2}\right) \xrightarrow{\vec{\alpha}\to 0} \propto \|w\|_1$$

- Problem 1.
  - How small an α do we need?

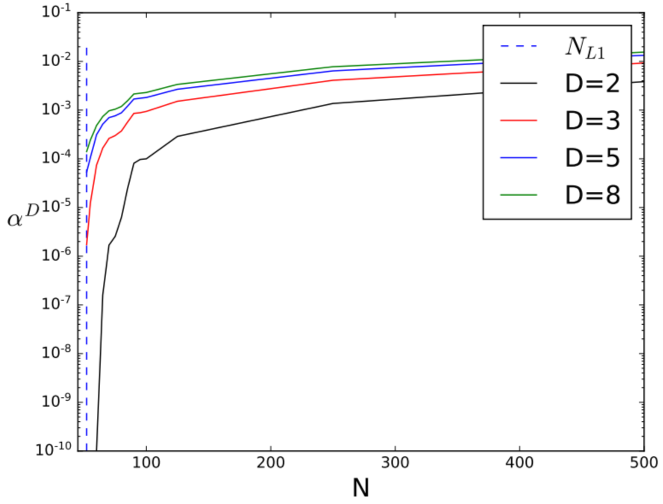
$$if\ \underbrace{\alpha \le \exp\left(-\tilde{O}(\epsilon)\right)}_{\text{BAD!!}} \Rightarrow \|w_\alpha^\infty\|_1 \le (1+\epsilon) \min_{Xw=y} \|w\|_1$$

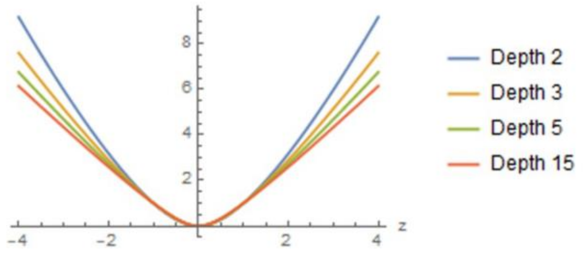
Sparse recovery!



Increase number of samples. Sparse regression simulation with N gaussian measurements for $d = 1000$ dimentional r$= 10$ sparse recovery.

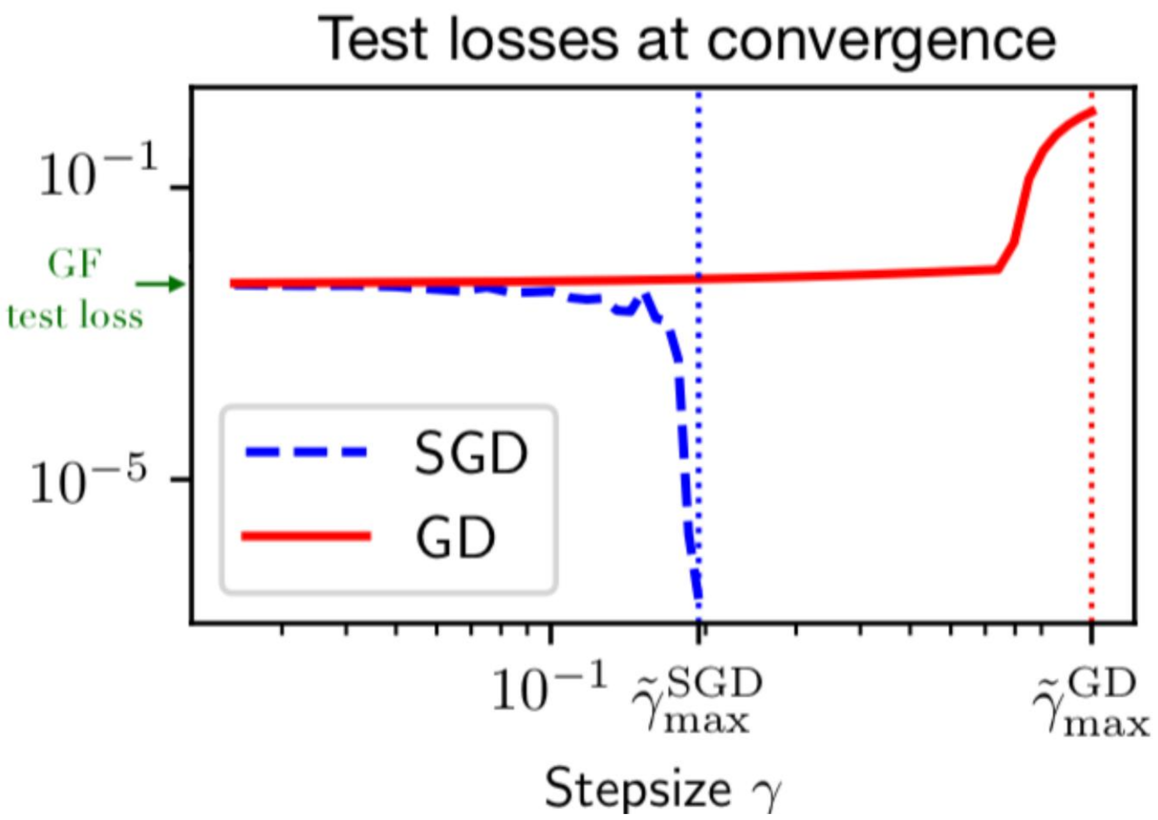
Plot shows smallest α such that recovery error < 0.025 for different sample sizes N.





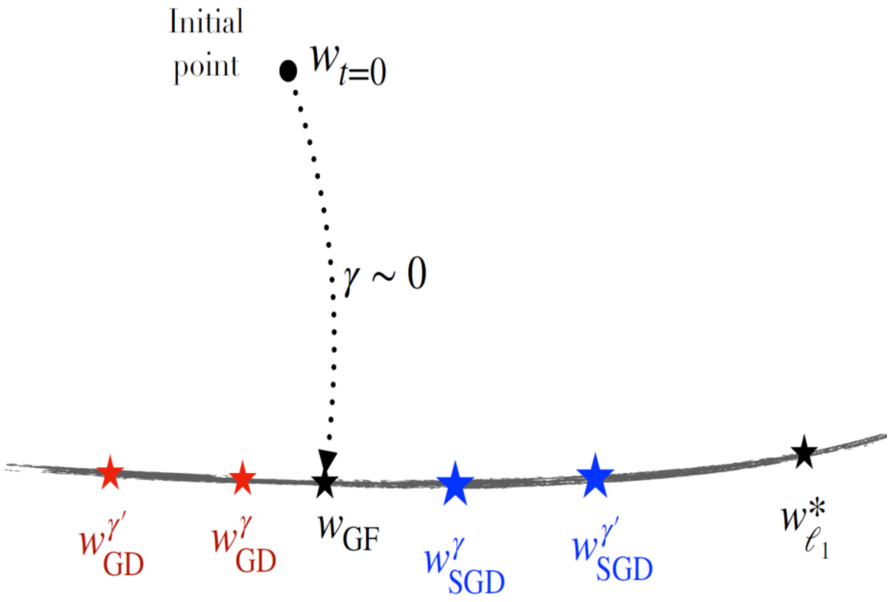
Higher depth. $w = w_+^D - w_-^D$ has effective regularization Q_α^D that has better dependence on α for approximating minimum ℓ_1 norm solution

Today…



Sparse regression simulation: $\tilde{\gamma}_{\max}^{(S)GD}$ are the maximum stepsizes until which (S)GD empirically converges to a valid solution

Role of stochasticity and step-size



Many related work (small subset, sorry for missed references)…

- Vaškevičius, Kanade, Rebeschini
- Nacson, Ravichandran, Srebro, and Soudry
- Andriushchenko, Varre, Pillaud-Vivien, Flammarion
- HaoChen, Wei, Lee, and Ma
- Pillaud-Vivien, Reygner, Flammarion
- Pesme, Pillaud-Vivien, Flammarion

Role of stochasticity and step-size

Recall: $(u, v) \leftarrow (u, v) - \gamma_t \nabla L_B(u, v)$

Setup

- Initialization: $u_\alpha(0) = \alpha 1, v_\alpha(0) = 0$
- $w_\alpha(t) = u_\alpha(t) \odot v_\alpha(t)$
- Stepsize**: $\forall \gamma_t : w_\alpha(t) \to w_\alpha^\infty$

Characterization

$$w_\alpha^\infty = \underset{Xw=y}{\operatorname{argmin}}\, D_{\boldsymbol{Q}_{\alpha_\infty}}(w, \widetilde{w}_0)$$

- D_ϕ is Bregmman divergence
- $\widetilde{\boldsymbol{w}}_0 \leq \alpha^2 \leftarrow$ small in our case and ignorable
- $\boldsymbol{\alpha}_\infty \leftarrow$ **effective initialization**

$$w_\alpha^\infty \approx \underset{Xw=y}{\operatorname{argmin}}\, Q_{\alpha_\infty}(w)$$

$$\boldsymbol{\alpha}_\infty = \alpha \odot \exp\left(-g\left(\Sigma_t \gamma_t \nabla \bar{L}_{B_t}(w(t))\right)\right)$$

where $g(x) = -\frac{1}{2}\log\left(\left(1 - x^2\right)\right)^2 \geq 0\ \forall |x| \leq \sqrt{2}$

Also, can show convergence for $\gamma_t \leq \frac{c}{LB}$ where L, B are problem dependent and fixes scaling.

Why is this characterization interesting?

$w_\alpha^\infty \approx \underset{Xw=y}{\operatorname{argmin}}\, Q_{\alpha_\infty}(w)$, with $\boldsymbol{\alpha}_\infty = \alpha \odot \exp\left(-g\left(\Sigma_t \gamma_t \nabla \bar{L}_{B_t}(w(t))\right)\right)$ such that $g(x) \geq 0$ in region of interest

- α_∞ depends on the trajectory → not really an implicit regularization result
- but, can give useful insights compared to trivial characterization
  - *E.g., effective initialization $\boldsymbol{\alpha}_\infty \leq \boldsymbol{\alpha}$ algorithmic initialization*
  - α_∞ has dependence on important parameters in an analyzable way
  - empirical tracking of partial sums in α_∞ can indicate properties of eventual converged solution
- Look at effects on α_∞ from learning rate, stochasticity …

“gain” from (stochastic) gradient descent

$w_\alpha^\infty \approx \underset{Xw=y}{\operatorname{argmin}}\, Q_{\alpha_\infty}(w)$, with $\boldsymbol{\alpha}_\infty = \alpha \odot \exp\left(-g\left(\Sigma_t \gamma_t \nabla \bar{L}_{B_t}(w(t))\right)\right)$ such that $g(x) \geq 0$ in region of interest

Can capture deviation of (S)GD from GF using “gain”

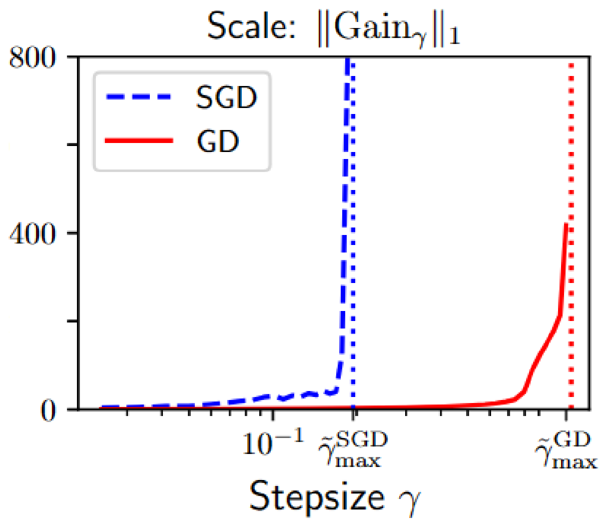
$$Gain_\gamma = \log\left(\frac{\alpha^2}{\alpha_\infty^2}\right) \geq 0, \qquad \in \mathbb{R}_+^d$$

larger $Gain_\gamma \Rightarrow$ larger deviation from GF

larger $Gain_\gamma \Rightarrow$ smaller is the effective initialization

Large stepsizes help sparse recovery?

$Gain_\gamma = \log\left(\frac{\alpha^2}{\alpha_\infty^2}\right) \geq 0, \in \mathbb{R}_+^d$ (larger $Gain_\gamma \Rightarrow$ smaller is the effective initialization)



$\tilde{\gamma}_{\max}^{(S)GD}$ are the maximum stepsizes until which (S)GD empirically converges to a valid solution (edge of stability)

$$\lambda_b \gamma^2 \sum_t \bar{L}(w(t)) \leq \mathbb{E}\left[\left\|Gain_\gamma\right\|_1\right] \leq \Lambda_b \gamma^2 \sum_t \bar{L}(w(t))$$

$\Lambda_b, \lambda_b > 0$ are data-dependent constants for mini-batch size b

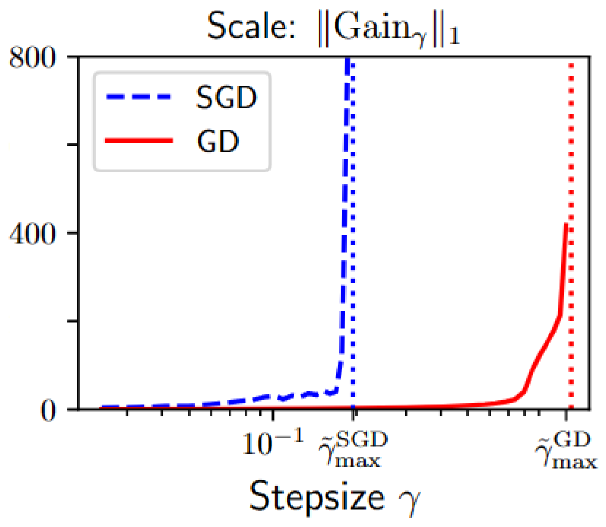
s.t. $\lambda_b H \leq \mathbb{E}_{B_k}[H_{B_k}] \leq \Lambda_b H$

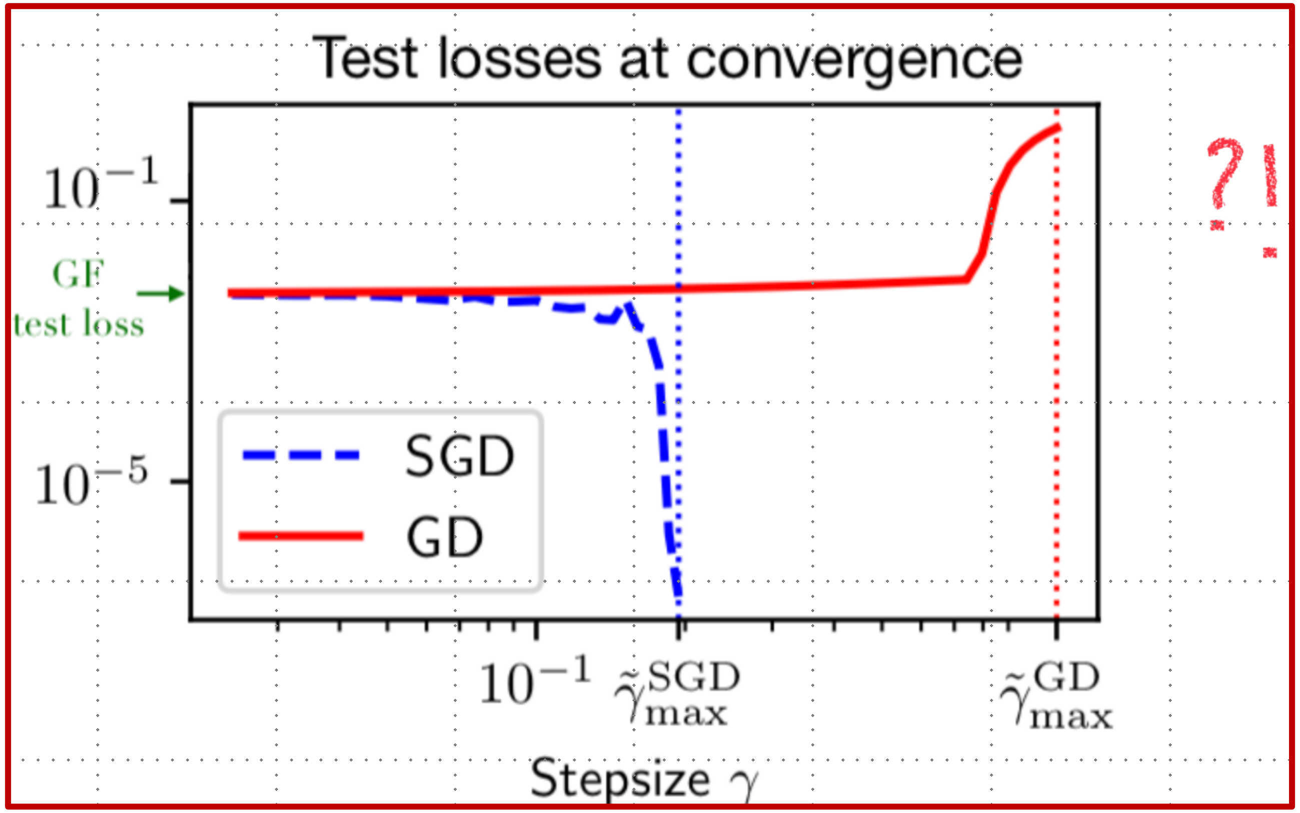
If X sampled iid $\mathcal{N}(0, \sigma^2)$, then w.h.p.

$$\mathbb{E}\left[\left\|Gain_\gamma\right\|_1\right] = \Theta\left(\frac{\boldsymbol{\gamma}}{\boldsymbol{b}}\, \sigma^2 d \log\frac{1}{\alpha}\, \left\|w_{\ell_1}^*\right\|_1\right)$$

Large stepsizes help sparse recovery?

$Gain_\gamma = \log\left(\frac{\alpha^2}{\alpha_\infty^2}\right) \geq 0, \in \mathbb{R}_+^d$ (larger $Gain_\gamma$ smaller is the effective initialization)





Sparse recovery?

$$w_\alpha(t) \to w_\alpha^\infty = \underset{Xw=y}{\operatorname{argmin}}\, Q_{\vec{\alpha}}(w) \coloneqq \Sigma_{i\in[d]}\, q\left(\frac{w_i}{\vec{\alpha}_i^2}\right) \xrightarrow{\vec{\alpha}\to 0} \propto \|w\|_1$$

- Problem 1.
  - How small an α do we need?

$$if\ \alpha \le \exp\left(-\tilde{O}(\epsilon)\right) \Rightarrow \|w_\alpha^\infty\|_1 \le (1+\epsilon)\min_{Xw=y}\|w\|_1$$

- Problem 2.
  - $Q_{\vec{\alpha}}(w) \to \propto \|w\|_1$ if and only if **if $\vec{\alpha} \to 0$ uniformly on all coordinates**
  - If say, $\vec{\alpha} = \bar{\alpha}\exp(-h)$ where h is constant and $\bar{\alpha} \to 0$ uniformly, then

$$Q_{\vec{\alpha}}(w) \xrightarrow{\bar{\alpha}\to 0} \propto \underbrace{\sum_i h_i\, |w_i|}_{\text{can be bad for sparse recovery}}$$

Shape of gain

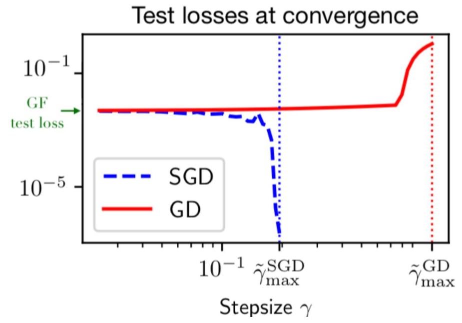
if $\vec{\alpha} = \bar{\alpha}\exp(-h)$, then $Q_{\vec{\alpha}}(w) \xrightarrow{\bar{\alpha}\to 0} \propto \sum_i h_i\,|w_i|$

$$Gain_\gamma = \log\left(\frac{\alpha^2}{\alpha_\infty^2}\right)$$

$\Rightarrow$ in case of non-uniform gains on coordinates,

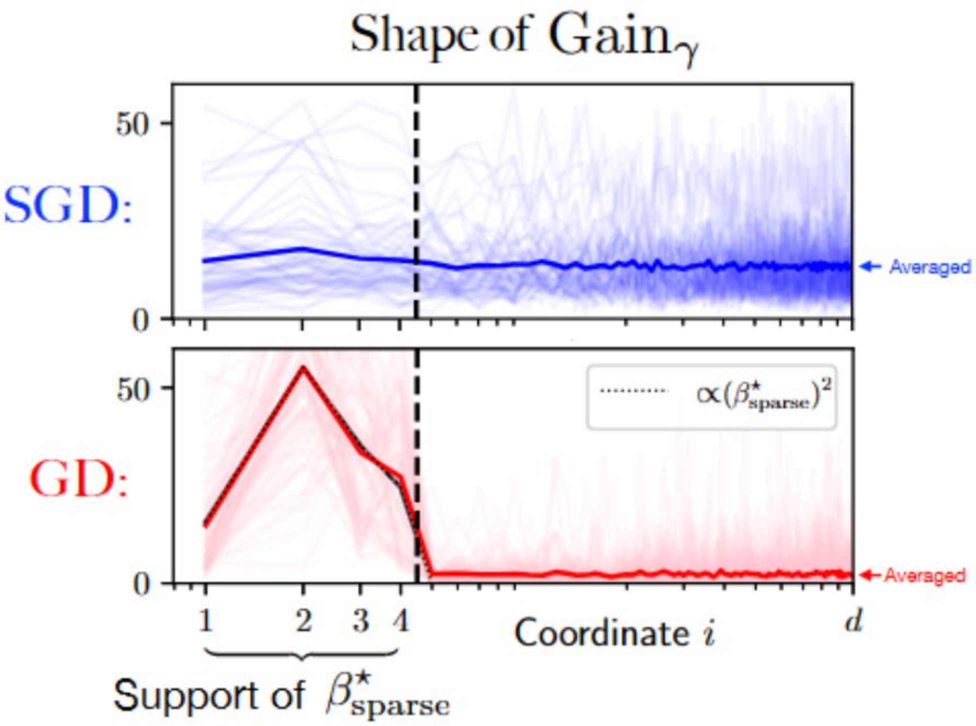
coordinates with large gain → coordinates with lower effective initialization → coordinates with higher penalty on weighted ℓ_1 bias

- Both SGD and GD have high gain magnitude,
- but if the gain is non-uniformly, that would explain



Shape of gain

coordinates with large gain → coordinates with lower effective initialization
→ coordinates with higher penalty on weighted ℓ_1 bias



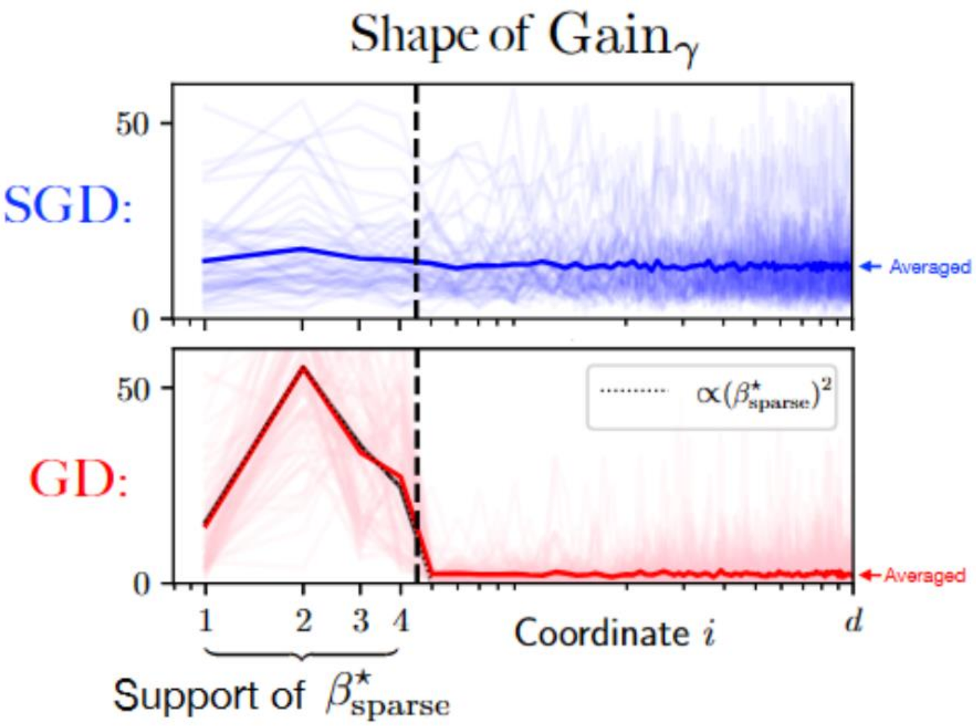
- At large learning rates, gain of GD is high exactly on support of the sparse solution

⇒ Effective regularization is weighted ℓ_1 norm with high weights on support of w^*_{sp}

- Very bad for sparse recovery!

Shape of gain

coordinates with large gain → coordinates with lower effective initialization
→ coordinates with higher penalty on weighted ℓ_1 bias



For Gaussian measurement matrices (and generalizations), given a sparse solution w_{sp}^*

$$\nabla\bar{L}\big(w(0)\big)^2 = {w_{sp}^*}^2 + \epsilon$$

$$\mathbb{E}\left[\nabla L_{i_0}\big(w(0)\big)^2\right] = \Theta\left(\|w_{sp}^*\|_2^2 1\right)$$

Gradient dynamics in linear diagonal networks

This simple network already exhibits many layers (pun intended) of complexity and qualitatively different behavior wrt different hyperparameters

- effect of initialization scale in infinitesimal stepsize regime
- effective “reinitialization” with macroscopic step sizes using (S)GD
- show qualitatively distingt behavior between GD and SGD with large stepsizes
- gain broader insights on the effects of stepsize and of stochasticity