Effect of Macroscopic Stepsizes on (S)GD over Diagonal Linear Networks

Theoretical Deep Learning Evaluation
Nafissa BENALI & Angel REYERO

M2 Mathematics & Artificial Intelligence
Institut de Mathématiques d’Orsay (IMO)
Paris-Saclay University

Teacher:
Hédi HADIJI

February 13, 2024

Mathématiques Orsay

Contents

Setting

- **Overparametrized regression:** input $X \in \mathbb{R}^{dn}$, output $y \in \mathbb{R}^n$ with $d \gg n$. Then, infinite number of interpolators:

$$\mathscr{S} := \left\{ \beta^\star \in \mathbb{R}^d \text{ st } \langle \beta^\star, x_i \rangle = y_i, \forall i \in [n] \right\}.$$

- **2-Layer diagonal network:** $x \to \langle u \odot v, x \rangle = \langle u, \sigma(\mathrm{diag}(v)x) \rangle$ with $w = (u, v) \in \mathbb{R}^{2d}$ the weights.
- **Mini-Batch quadratic loss** for $\mathscr{B} \subset [n]$ of size b,

$$F_{\mathscr{B}}(w) := \mathscr{L}_{\mathscr{B}}(u \odot v) = \frac{1}{2b} \sum_{i \in \mathscr{B}} (y_i - \langle u \odot v, x_i \rangle)^2.$$

- **Mini-Batch SGD** $w_0 = (u_0, v_0)$, $w_{k+1} = w_k - \gamma_k \nabla F_{\mathscr{B}_k}(w_k)$.

Motivation

Why does SGD generalize better with large step sizes? Why does GD not?

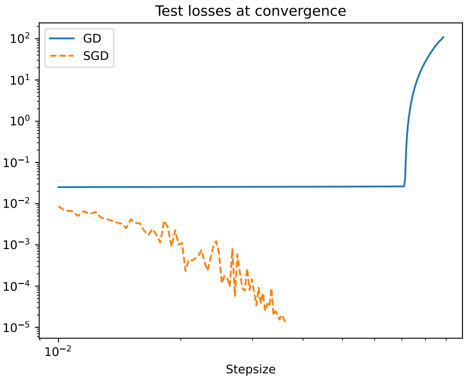

Figure 1: Generalization difference between GD and SDG for DLN. **Setting:** $(x_i)_{i\in[n]} \sim \mathcal{N}(0, \mathrm{I_d})$, $y_i = \langle \beta^\star, x_i \rangle$ for some s-sparse vector $\beta^\star$,uniform initialisation $\boldsymbol{\alpha} = \alpha \mathbf{1}$ and $(n, d, s, \alpha) = (20, 30, 3, 0.1)$.

Preliminaries: Gradient Flow on DLN

Theorem 1 (Gradient Flow on DLN)

The limit $\beta^\star_\alpha$ of the GF $\mathrm{d}w_t = -\nabla F(w_t)\mathrm{d}t$ initialised at $(u_0, v_0) = (\sqrt{2}\boldsymbol{\alpha}, 0)$ is the solution of the minimal interpolation problem:

$$\beta^\star_\alpha = \underset{\beta^\star \in \mathscr{S}}{\operatorname{argmin}} \psi_\alpha(\beta^\star)$$

where ψ_α is the hyperbolic entropy function.

Generalization properties depend on

- **Scale** of $\boldsymbol{\alpha}$: if $\boldsymbol{\alpha} = \alpha 1$, $\psi_\alpha \sim \ln(1/\alpha)\|\cdot\|_1$ as $\alpha \to 0$.
- **Shape** of $\boldsymbol{\alpha}$: $\psi_\alpha(\beta) \sim \sum_{i=1}^{d} \ln(1/\alpha_i)|\beta_i|$ as $\boldsymbol{\alpha} \to 0$.

Main result

Theorem 2 (Implicit bias and convergence)

Assume that $(\beta_k)_{k\geq 0} = (u_k \odot v_k)_{k\geq 0}$ converge to some interpolator $\beta^\star_\infty \in \mathscr{S}$. Then,

$$\beta^\star_\infty = \underset{\beta^\star \in \mathscr{S}}{\operatorname{argmin}} \mathscr{D}_{\psi_{\alpha_\infty}} \left(\beta^\star, \widetilde{\beta}_0\right),$$

with $\widetilde{\beta}_0$ a small perturbation term and $\mathscr{D}_{\psi_{\alpha_\infty}}$ is the Bregman divergence with hypentropy ψ_{α_∞} of the effective initialization

$$\alpha^2_\infty = \alpha^2 \odot \exp\left(-\sum_{k=0}^{\infty} q(\gamma_k \nabla \mathscr{L}_{\mathscr{B}_k}(\beta_k))\right)$$

*and $q(x) = -\frac{1}{2}\ln\left((1-x^2)^2\right)$. Moreover, with a small enough stepsize, the iterates **converge** to $\beta^\star_\infty$.*

If $(\gamma_k)_{k\geq 0} = \gamma$, we denote $\mathrm{Gain}_\gamma := \ln\left(\frac{\alpha^2}{\alpha^2_\infty}\right) = \sum_{k=0}^{\infty} q(\gamma \nabla \mathscr{L}_{\mathscr{B}_k}(\beta_k))$.

Interpretation

- $\widetilde{\beta}_0$ can be ignored:

$$\beta^\star_\infty = \underset{\beta^\star \in \mathscr{S}}{\operatorname{argmin}} \mathscr{D}_{\psi_{\alpha_\infty}}\left(\beta^\star, \widetilde{\beta}_0\right) \approx \underset{\beta^\star \in \mathscr{S}}{\operatorname{argmin}} \mathscr{D}_{\psi_{\alpha_\infty}}(\beta^\star, 0) = \underset{\beta^\star \in \mathscr{S}}{\operatorname{argmin}} \psi_{\alpha_\infty}(\beta^\star).$$

- As the stepsize shrinks to 0, $\alpha_\infty \to \alpha$ and $\beta^\star_\alpha \approx \beta^\star_\infty$(ie, GF$\approx$ (S)GD).
- On the Gain (with constant stepsize and uniform α):
  - **Scale:**
    - As $\|\mathrm{Gain}_\gamma\|_1 \approx 0$, then $\alpha_\infty \approx \alpha$.
    - The larger the **stepsize**, the larger the Gain_γ.
    - The larger the **batch size**, the smaller the Gain_γ.
  - **Shape:** $\psi_{\alpha_\infty}(\beta) \sim \ln\left(\frac{1}{\alpha}\right)\|\beta\|_1 + \sum_{i=1}^d \mathrm{Gain}_\gamma(i)|\beta_i|$.
    - **Heterogeneous** for GD.
    - **Homogeneous** for SGD.

Conclusions

In the context of Diagonal Linear Networks, we are able to prove:

- Convergence of (S)GD.
- For small stepsizes between (S)GD $\approx$ GF.
- The scale of Gain_γ explains the differences between (S)GD and GF.
- The shape of Gain_γ explains the differences between GD and SGD.

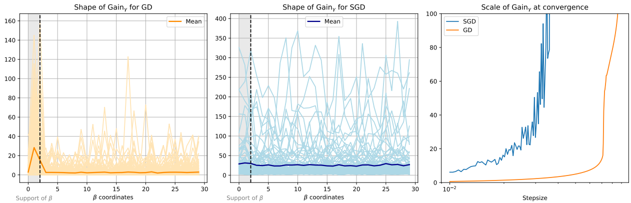

Figure 2: On the left, the heterogeneous shape of Gain_γ for GD, on the center the homogeneous shape of Gain_γ for SGD, and on the right the arbitrarily large Gain_γ in the EoS regime.

References

[Even et al.(2023)Even, Pesme, Gunasekar, and Flammarion] Mathieu Even, Scott Pesme, Suriya Gunasekar, and Nicolas Flammarion. (s)gd over diagonal linear networks: Implicit regularisation, large stepsizes and edge of stability, 2023.

[Woodworth et al.(2020)Woodworth, Gunasekar, Savarese, Moroshko, Go
Blake Woodworth, Suriya Gunasekar, Pedro Savarese, Edward Moroshko, Itay Golan, Jason Lee, Daniel Soudry, and Nathan Srebro.
Kernel and rich regimes in overparametrized models, 2020.

Thank You, Questions?

Definition of functions for Theorem 1

The hyperbolic entropy function:

$$\psi_\alpha(\beta) := \frac{1}{2}\sum_{i=1}^{d}\left(\beta_i \operatorname{arcsinh}\left(\frac{\beta_i}{\alpha_i^2}\right) - \sqrt{\beta_i^2 + \alpha_i^4} + \alpha_i^2\right)$$

Definition of functions for Theorem 2

The small perturbation term $\widetilde{\beta}_0 \in \mathbb{R}^d$ introduced in Theorem 2 is given by $\widetilde{\beta}_0 = \frac{1}{2}(\alpha_+^2 - \alpha_-^2)$ where $q_\pm(x) = \mp 2x - \ln\left((1 \mp x)^2\right)$ and $\alpha_\pm^2 = \alpha^2 \odot \exp\left(-\sum_{k=0}^{\infty} q_\pm\left(\gamma_k \nabla \mathscr{L}_{\mathscr{B}_k}(\beta_k)\right)\right)$.

Estimated parameters with SGD and GD

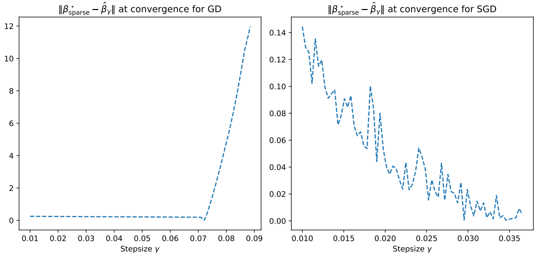

Figure 3: With a macroscopic step size, the result obtained by Gradient Descent (GD) differs substantially from the true parameter, while Stochastic Gradient Descent (SGD) benefits from this step size.