SGD: Role of Randomness and Sharpness

ATML Track 1

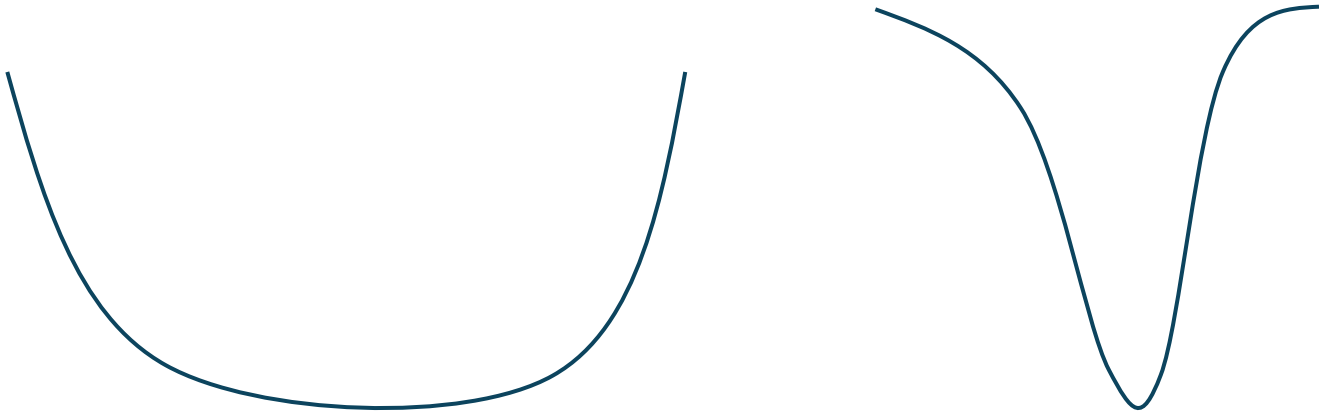
Rik Sarkar

Today: Why does SGD work well?

- Sharp and flat minima: which is better?
- Why does SGD find Flat minima?
- How to measure sharpness
- How to use sharpness

Sharp and Flat minima

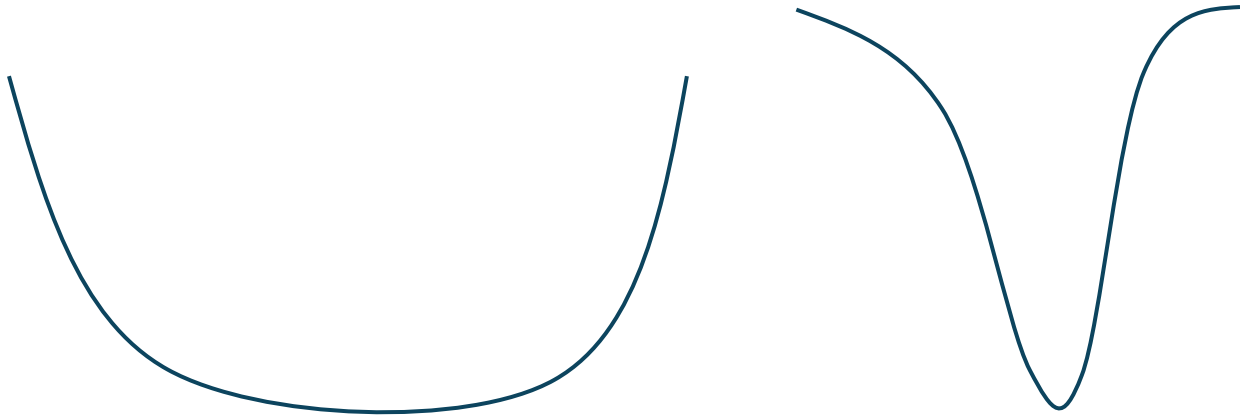
- A minimum could be flat or sharp.
  - Which is better?
  - And why?



Sharp and Flat minima

- A minimum could be flat or sharp.
  - Which is better?
  - And why?

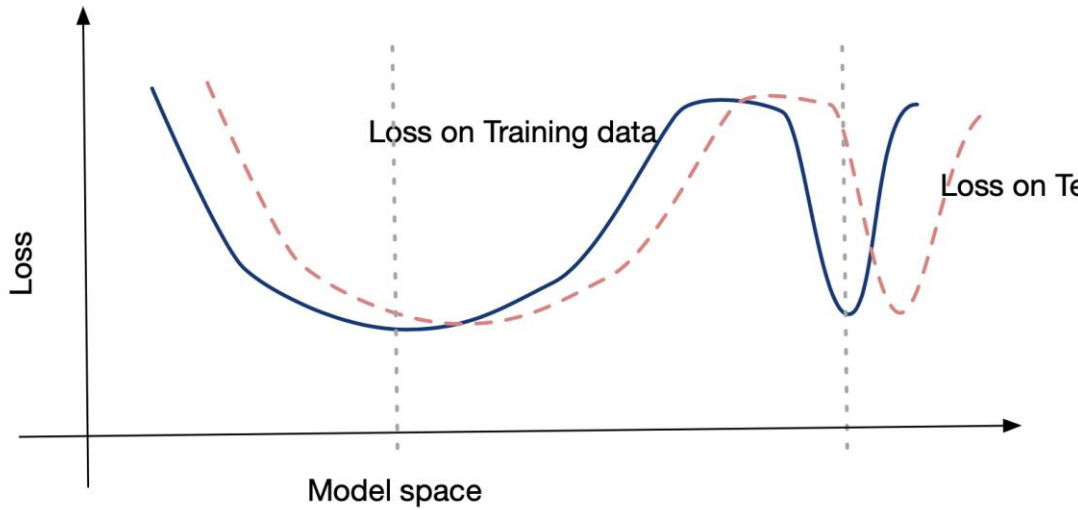
- Sharp min could represent overfitting
  - E.g. one point fits really well at that model point. But not in general.



Flat and sharp minima

- Flat minima generalize better
- Sharper minima likely to represent overfitting
  - If we take a slightly different model or slightly different data
  - The loss will jump

- Flat minima have better generalization
  - And better stability
  - Even away from the min



SGD prefers flat minima

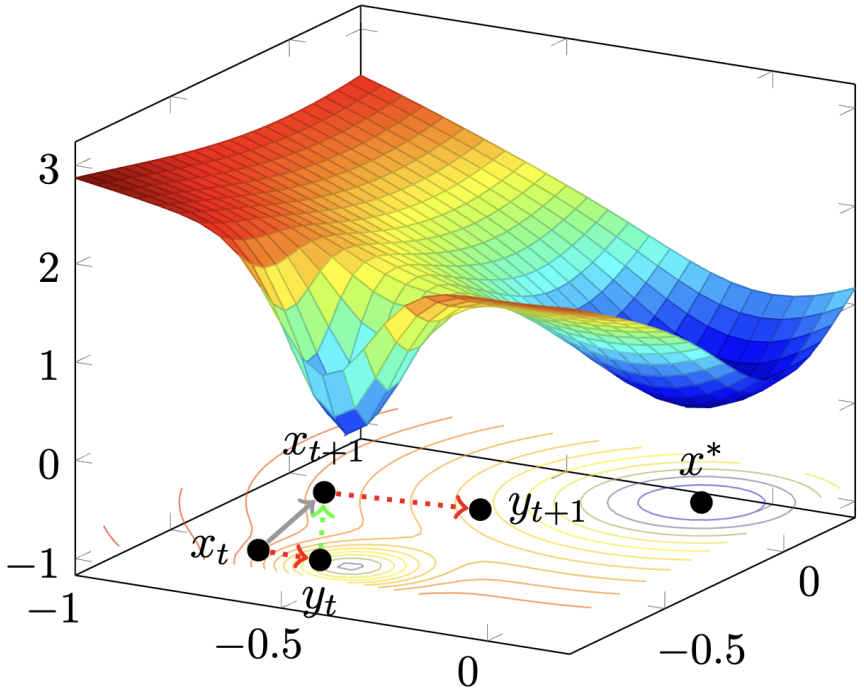
- SGD includes randomness
  - At the same position w, suppose the gradient is $\nabla f_b(w)$ for batch b
  - This depends on the batch b selected.

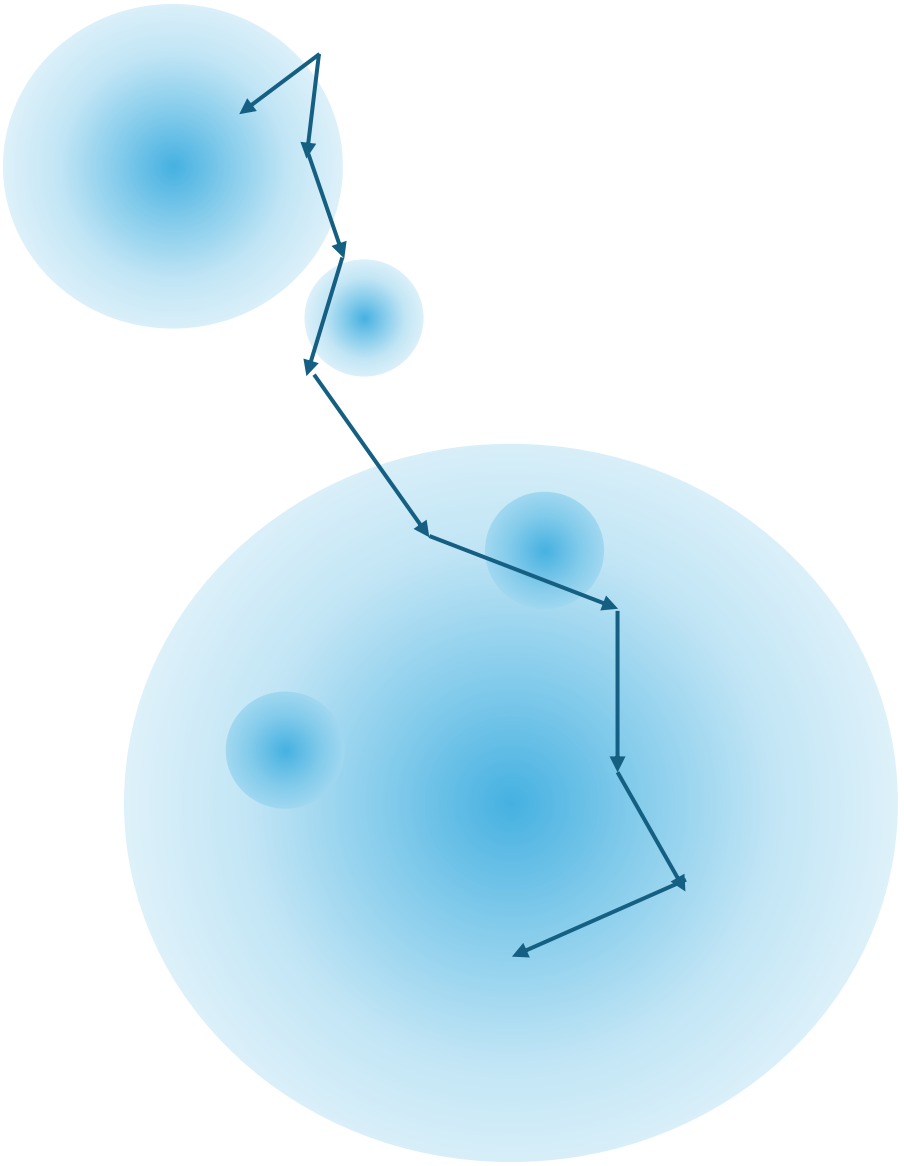
- Pure gradient descent (GD) computes $\nabla f_S(w)$ with whole S
  - Does not depend on any randomness

- We can view SGD as GD + noise
  - I.e. compute gradient vector as in GD, add a noise representing the random selection in SGD

SGD escapes sharp minima

- Converges to large flat basins

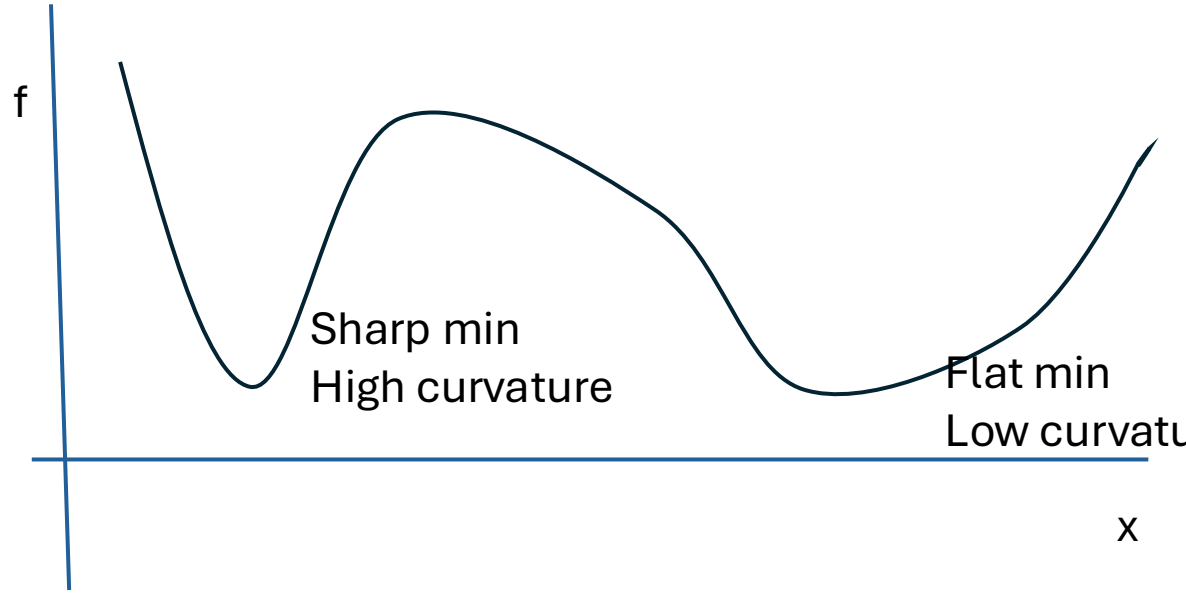




Kleinberg et al. 2018

How do we measure sharpness?

- Using curvature
- Curvature in 1-D
- Measured by 2nd derivative
  - $\frac{d^2 f}{dx^2}$
  - The rate of change of the derivative $\frac{df}{dx}$



Hessian: Curvature in high dimensions

- But our models are in high dimensions
  - Instead of a single x, we have $x_1, x_2, \ldots$
  - For any function f, rate of change of f is a vector:
    - $[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots \frac{\partial f}{\partial x_n}]$ Q: what is this vector called?
  - Now, let us measure how fast each of these components change:

Hessian: Curvature in high dimensions

- But our models are in high dimensions
  - Instead of a single x, we have $x_1, x_2, \ldots$
  - For any function f, rate of change of f is a vector:
    - $[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots \frac{\partial f}{\partial x_n}]$ Q: what is this vector called?
  - Now, let us measure how fast each of these components change:

$$\mathbf{H}_f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1\,\partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1\,\partial x_n} \\ \frac{\partial^2 f}{\partial x_2\,\partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2\,\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n\,\partial x_1} & \frac{\partial^2 f}{\partial x_n\,\partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$

Simpler example: 2D case

- For the Loss function L:

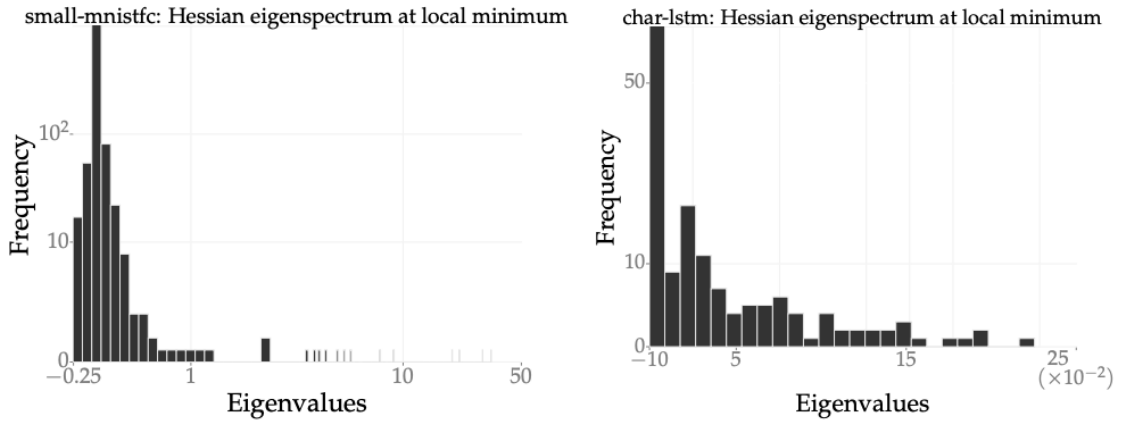
- $H = \begin{pmatrix} \frac{\partial^2 L}{\partial w_1^2} & \frac{\partial^2 L}{\partial w_1 \partial w_2} \\ \frac{\partial^2 L}{\partial w_2 \partial w_1} & \frac{\partial^2 L}{\partial w_2^2} \end{pmatrix}$

- Overall, larger numbers in Hessian matrix means larger curvature, sharper loss

Sharpness measures

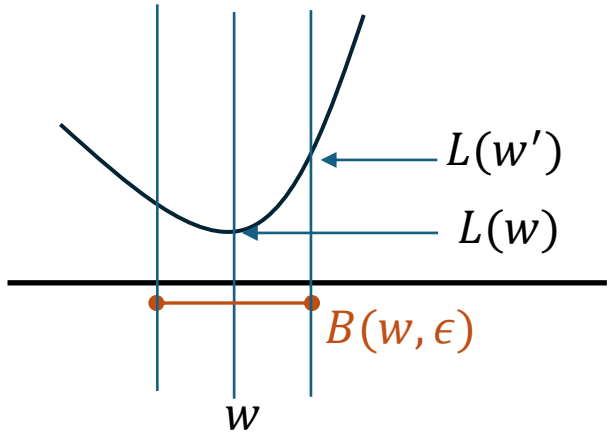
- Frobenius norm of the hessian
  - $\left|\left|H\right|\right|_F = \sqrt{\sum_i \sum_j \left|x_{ij}\right|^2}$

- Distribution of Eigen values
  - If there are more of large eigen values that implies a sharper min



ϵ-Sharpness

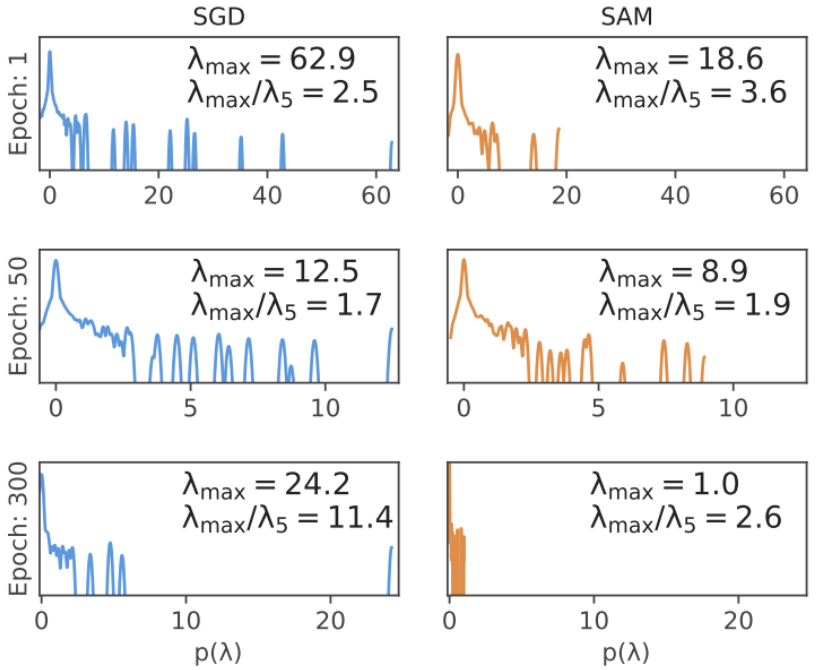
- At min w take ball $B(w, \epsilon)$ of radius ϵ
  - Set of all points within a distance ϵ of w
- Sharpness is:
  - $$\frac{\max_{w' \in B(\epsilon, w)} (L(w') - L(w))}{1 + L(w)}$$



- Single number, easier measure
  - But requires a reasonable ϵ

Sharpness aware minimization

- Use ϵ sharpness
  - Minimize $L(w) + [L(w + \epsilon') - L(w)]$
- Suppose we write $\boldsymbol{\alpha}$ is a vector such that $w + \boldsymbol{\alpha} \in B(w, \epsilon)$
- Theorem:
  - $L_{\mathcal{D}} \leq \max_{|\boldsymbol{\alpha}|_2 \leq \epsilon} L_S(w + \boldsymbol{\alpha}) + h(||w||_2^2 / \epsilon^2)$
  - Observe that RHS =
  - $[\max_{|\boldsymbol{\alpha}|_2 \leq \epsilon} L_S(w + \boldsymbol{\alpha}) - L_S(w)] + L_S(w) + h(||w||_2^2 / \epsilon^2)$
  - Sharpness + Loss + L2 regulariser



Other algorithms

- Using Hessian for regularization

- Stochastic weight averaging
  - Average weights of last c models
  - Produces flat minima

Observations about SGD

- Noise & randomness is actually good!
  - See simulated annealing
- We have seen two ways randomness benefits SGD
- Randomness causes SGD to escape sharp local min and converge to large flat basins
- Randomness improves stability
  - We saw in last lecture: randomness helps stability to small changes
  - Therefore improves generalisation

Next week (likely)

- Exam review class Monday, Tuesday + some extra (non-examinable topic) on tuesday
- Look out for announcements on Piazza.