Reminders

↳ Quiz

↳ Pset

↳ Office Hours

↳ Discord

↳ Events

- Conference Oct 3-5
- Redistricting Sep 16 Lunch @ Ath
- Trees Sep 16 7pm @ Mudd

Review

$\underline{x}^{(i)} \in \mathbb{R}^d \qquad y^{(i)} \in \mathbb{R}$

$\underline{X} \in \mathbb{R}^{n \times d} \qquad \underline{y} \in \mathbb{R}^n$

$$\begin{bmatrix} - \underline{x}^{(i)T} - \end{bmatrix} \qquad \begin{bmatrix} y^{(i)} \end{bmatrix}$$

Goal: $f(x^{(i)}) \approx y^{(i)}$

1. Model — Linear! $f(x) = w^T x$
2. Loss — MSE! $\mathcal{L}(w) = \sum_{i=1}^{n} \left[f(x^{(i)}) - y^{(i)} \right]^2 \frac{1}{n}$
3. Optimizer — Exact! $w^* = (X^T X)^+ X^T y$

Why pseudo inverse?

"Invert" non-invertible matrices

e.g. $\underline{X} \in \mathbb{R}^{n \times d}$

$$\underline{X} = \sum_{i=1}^{d} \sigma_i \, \underline{u}_i \, \underline{v}_i^T$$

$$\underline{X}^+ \underline{X} = \sum_{j=1}^{d} \frac{1}{\sigma_j} \underline{v}_j \, \underline{u}_j^T \sum_{i=1}^{d} \sigma_i \, \underline{u}_i \, \underline{v}_i^T$$

$$= \sum_{i=1}^{d} \underline{v}_i \, \underline{v}_i^T = I_d$$

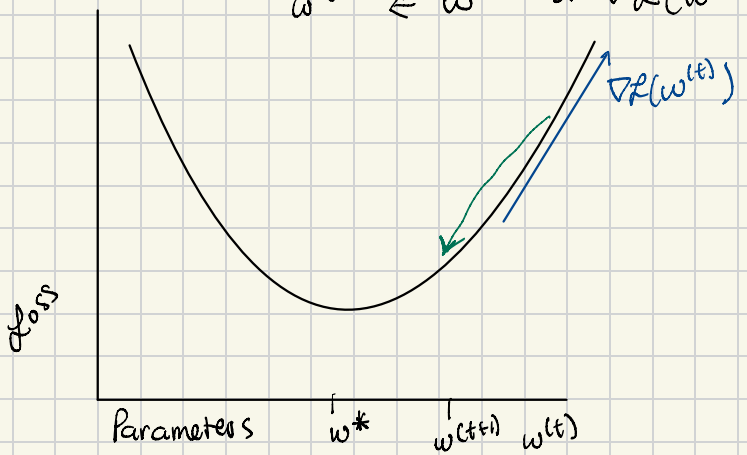
Greatest advice: "Go sit on your rock"

Two issues:

- ↳ Time to compute w^*
- ↳ What if data is not linear?

Gradient Descent

$$w^{(t+1)} \leftarrow w^{(t)} - \alpha \nabla \mathcal{L}(w^{(t)})$$



Intuition: Move away from steepest ascent

α = "step size" or "learning rate"

The Math (as promised)

$w \in \mathbb{R}$

$$\frac{\partial \mathcal{L}(w)}{\partial w} = \lim_{\Delta \to 0} \frac{\mathcal{L}(w+\Delta) - \mathcal{L}(w)}{\Delta}$$

$$\Rightarrow \quad \mathcal{L}(w+\Delta) - \mathcal{L}(w) \approx \frac{\partial \mathcal{L}(w)}{\partial w} \Delta$$

Choose Δ so $\mathcal{L}(w+\Delta) - \mathcal{L}(w)$ is negative ....

$$\Delta = -\frac{\partial \mathcal{L}(w)}{\partial w}$$

$w \in \mathbb{R}^d$

$$\frac{\partial \mathcal{L}(w)}{\partial w_i} = \lim_{\Delta \to 0} \frac{\mathcal{L}(w + \Delta e_i) - \mathcal{L}(w)}{\Delta}$$

$$\Rightarrow \mathcal{L}(w + \Delta e_i) - \mathcal{L}(w) \approx \Delta \langle \nabla_w \mathcal{L}(w), e_i \rangle$$

$$\mathcal{L}(w + v) - \mathcal{L}(w) \approx \Delta \langle \nabla_w \mathcal{L}(w), v \rangle$$

Choose $\Delta v = -\alpha \nabla_w \mathcal{L}(w)$

$$\mathcal{L}(w + v) - \mathcal{L}(w) \approx -\alpha \| \nabla_w \mathcal{L}(w) \|_2^2$$
$$= -\alpha \| \nabla_w \mathcal{L}(w) \|_2 \| \nabla_w \mathcal{L}(w) \|_2$$

Recall $\langle a, b \rangle = \|a\|_2 \|b\|_2 \cos(\theta)$

$\max_\theta \cos\theta = 1 \Rightarrow$ achieve "best" update!

For linear regression,

$$\nabla_w \mathcal{L}(w) = \frac{2}{n} X^T (Xw - y)$$

Time to compute:

Even Faster

n and d can be prohibitively large

Enter: STOCHASTIC gradient descent

"~~sto~~khos"

$S \subseteq \{1, \ldots, n\}$

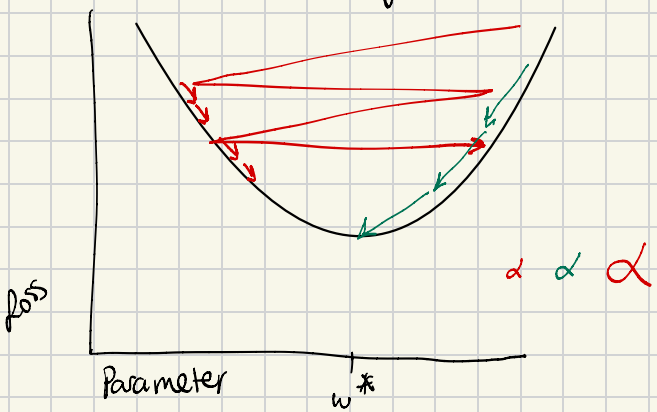
$$\mathcal{L}_S(w) = \frac{1}{|S|} \sum_{i \in S} [f(x^{(i)}) - y^{(i)}]^2$$

For linear regression,

$$\nabla_w \mathcal{L}_S(w) = \frac{2}{|S|} X_S^T (X_S w - y_S)$$

Time to compute:

Learning Rate



α α α

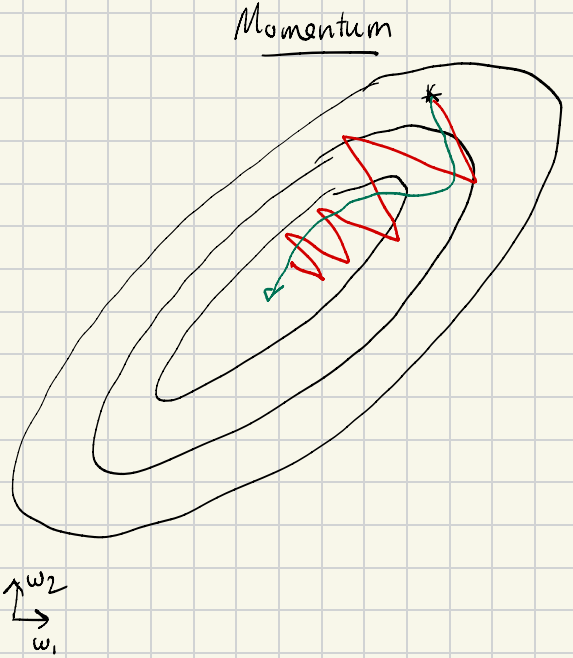
heuristic strategies:

↳ If loss unstable, decrease ($\alpha/2$)
If loss slowly decreasing, increase (2α)

↳ "scheduler" start big → small, cycle

↳ Adaptive $\alpha^{(t+1)} \leftarrow \frac{\alpha^{(t)}}{(\nabla_w \mathcal{L}(w^{(t)}))^2}$

Momentum



$m^{(t+1)} \leftarrow \beta m^{(t)} + (1-\beta) \nabla_w \mathcal{L}_s(w^{(t)})$

$w^{(t+1)} \leftarrow w^{(t)} - \alpha m^{(t+1)}$