Let $f: \mathbb{R}^d \to \mathbb{R}$

$$x^{(t+1)} = x^{(t)} - \alpha \nabla f(x^{(t)})$$

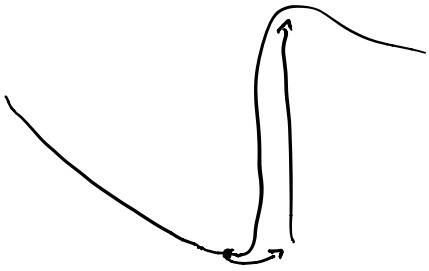
α was too big

Can $f(x^{(t+1)}) > f(x^{(t)})$?

If $\alpha \gg \frac{1}{M}$, $\Rightarrow$ divergence

$$x^{(t+1)} = (1 - \alpha M) x^{(t)}$$

$$\alpha > \frac{2}{M}$$



What morally is gradient descent?

$$\min_x f(x)$$

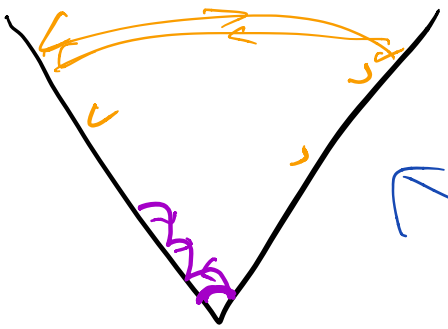
$$x^{(t+1)} = x^{(t)} - \alpha \nabla f(x^{(t)})$$

Continuous variant: gradient flow

$$\frac{dx}{dt} = -\nabla f(x(t)) \leftarrow$$

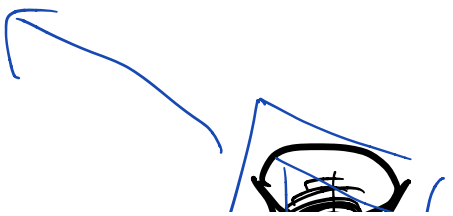
GD on $f(x) = \|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_d^2}$

with fixed step size?

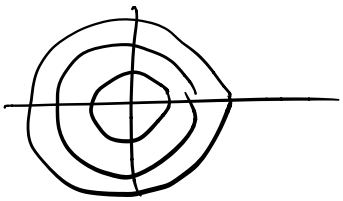


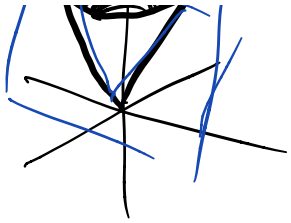
— α too big

— α not too big



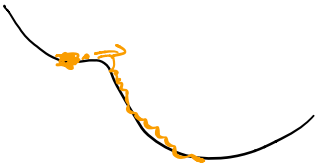
$f(x) = \sqrt{x_1^2 + x_2^2}$





Virtues of SGD

- Can't see all data at once (online learning)
- computationally cheaper to compute approx gradient per update cost is lower
- regularization (Stochasticity can lead us to jump out of minima/



$$\min_\theta \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(x_i), y_i)$$

$$\frac{1}{|B|} \sum_{i \in B} \ell(f_\theta(x_i), y_i)$$

How big to choose B?

Largest we can afford w/ our GPU

$\{1, 2, \ldots\ldots, n\}$

Π

$$G(\theta) = \nabla_\theta \ell(f_\theta(x_j), y_j) \quad \text{for random } j$$

$$\mathbb{E}_j G(\theta) = \sum_j \frac{1}{n} \nabla_\theta \ell(f_\theta(x_j), y_j)$$

EM & Estimating Gaussian Mixtures

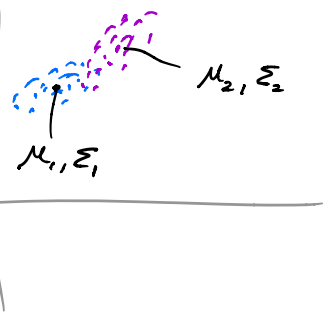
Data: $\{X_i\}_{i=1\cdots n}$

Model: Mixture of 2 Gaussians

$$Z_i = \begin{cases} 1 & \text{w/ prob } \tau_1 \\ 2 & \text{w/ prob } \tau_2 = 1-\tau_1 \end{cases}$$

latent variable

$$X_i \mid Z_i=1 \sim N(\mu_1, \Sigma_1)$$
$$X_i \mid Z_i=2 \sim N(\mu_2, \Sigma_2)$$



Given $\{X_i\}$, find: $\theta = \{\tau, \mu_1, \Sigma_1, \mu_2, \Sigma_2\}$

Likelihood of data under the model

$$L(\theta; X, Z) = P(X, Z \mid \theta) = \prod_{i=1}^{n} \prod_{j=1}^{2} \left(f(x_i; \mu_j, \Sigma_j) \tau_j \right)^{\mathbb{1}_{z_i=j}}$$

likelihood under normal dist

$P(X \mid Z, \theta)\, P(Z \mid \theta)$

Challenge: don't know Z

① Estimate dist over Z given $\theta = \hat{\theta}$

$$p(Z_i = j \mid X_i, \theta) = \frac{P(X_i \mid Z_i=j, \theta)\, P(Z_i=j \mid \theta)}{P(X_i \mid \theta)}$$

$$= \frac{f(x_i; \mu_j, \Sigma_j)\, \tau_j}{f(x_i; \mu_1, \Sigma_1)\, \tau_1 + f(x_i; \mu_2, \Sigma_2)\, \tau_2}$$

E-step Expectation

We can rewrite log likelihood function:

$$Q(\theta \mid \hat{\theta}) = \mathbb{E}_{Z \mid X, \hat{\theta}} \log L(\theta; X, Z)$$

$$= \sum_{i=1}^{n} \mathbb{E}_{Z_i \mid X_i, \hat{\theta}} \log L(\theta; X_i, Z_i)$$

$$= \sum_{i=1}^{n} \sum_{j} P(Z_i=j \mid X_i, \hat{\theta}) \log L(\theta; X_i, Z_i)$$

M-step Maximization

② $$\theta_{new} = \operatorname{argmax} Q(\theta \mid \hat{\theta})$$

Crowd Sourcing Problem

Majority vote?

	P_1	P_2	P_3	P_4	...	P_m	
X_1	$\hat{y}$		$\hat{y}$		$\hat{y}$	$\hat{y}$	$\rightarrow y_1$
X_2		$\hat{y}$	$\hat{y}$		$\hat{y}$	$\hat{y}$	$\rightarrow y_2$
X_3							
X_4	$\hat{y}$		$\hat{y}$				
⋮		$\hat{y}$					
X_n	$\hat{y}$			$\hat{y}$	$\hat{y}$		$\rightarrow y_n$

Why can EM beat MV?

Estimate quality of each rater

True class	i	P_1	P_2	P_3	P_4	P_5
(A)	1	A	A	C	C	C
(B)	2	B	B	D		
(C)	3	C	C		D	
(A)	4	A	A			B
	5					

If each rater rated 1 image only?