L16: Stochastic GD

Applied on Data

Mar 4, 2026

FoDA

Jeff M. Phillips

Gradient Descent $f: \mathbb{R}^d \to \mathbb{R}$

Goal $\underset{\alpha \in \mathbb{R}^d}{\operatorname{argmin}} f(\alpha)$

Input $(X, y) = \{(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)\} \subset \mathbb{R}^q \times \mathbb{R}$

"loss" function $L((X, y), M_\alpha)$

$$f(\alpha) = L((X, y), M_\alpha) = SSE((x, y), M_\alpha)$$

$$= \sum_{(x_i, y_i) \in (X, y)} \left(M_\alpha(x_i) - y_i\right)^2$$

Model M_α for polynomial regression

$d=1$ $\quad x_i \in \mathbb{R}$ $\quad p=2$

$$M_\alpha(x_i) = \langle \alpha, (1, x_i, x_i^2) \rangle = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2$$

$\alpha \in \mathbb{R}^3 \quad \alpha = (\alpha_0, \alpha_1, \alpha_2)$

GD: $\alpha = \alpha - \gamma \nabla f(\alpha)$

consider $n=1$

$j = 0, 1, 2$

$$\frac{\partial}{\partial \alpha_j} f(\alpha) = \frac{\partial}{\partial \alpha_j} (M_\alpha(x_1) - y_1)^2 = 2 (M_\alpha(x_1) - y_1) \frac{\partial}{\partial \alpha_j} (M_\alpha(x_1) - y_1)$$

$$= 2 (M_\alpha(x_1) - y_1) \frac{\partial}{\partial \alpha_j} \left(\sum_{i=0}^{2} \alpha_i x_1^i - y_1 \right)$$

$$= 2 (M_\alpha(x_1) - y_1) x_1^j$$

$$\frac{\partial}{\partial \alpha_j} f_1(\alpha) = 2\left(M_\alpha(x_1) - y_1\right) x_1^j$$

$n = 1$

$$\nabla f_1(\alpha) = \left(\frac{\partial}{\partial \alpha_0} f_1(\alpha), \frac{\partial}{\partial \alpha_1} f_1(\alpha), \frac{\partial}{\partial \alpha_2} f_1(\alpha)\right)$$

$$= 2\underbrace{\left(M_\alpha(x_1) - y_1\right)}_{\text{scalar}} \underbrace{\left(1, x_1, x_1^2\right)}_{\text{data point}}$$

error = residual of (x_1, y_1)

LMS update rule / Widrow-Hoff update

$$f(\alpha) = \sum_{i=1}^{n} f_i(\alpha)$$

$n \geq 1$

decomposable function

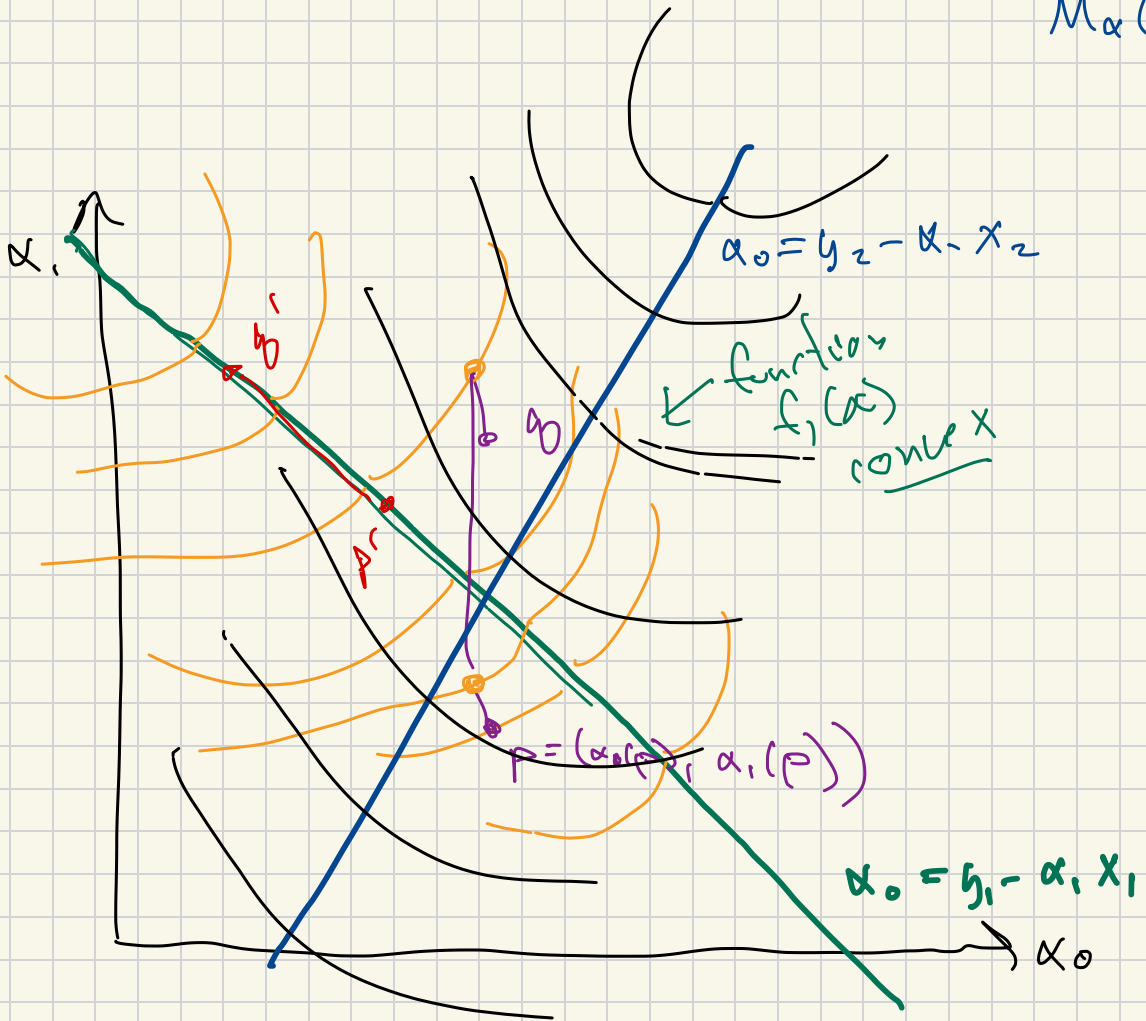
$\sum$ for 1 data point (x_i, y_i)

$$f_i(\alpha) = \left(M_\alpha(x_i) - y_i\right)^2$$

$$= \left(\alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 - y_i\right)^2$$

if each f_i is convex, then

$f = \sum_i f_i$ is also convex



$M_\alpha(x_i) = \alpha_0 + \alpha_1 x_i$

(x_1, y_1)

$y_1 = \alpha_0 + \alpha_1 x_1$

$f(\alpha) = 0$

$\Leftrightarrow$

$\alpha_0 = y_1 - \alpha_1 x_1$

← function $f_1(\alpha)$ convex

(x_2, y_2)

$f_2(\alpha) = 0 \Rightarrow \alpha_0 = y_2 - \alpha_1 x_2$

↳ $\sum_i f_i$ usually strongly convex

Batch Gradient Descent

n = large

$$\alpha = \alpha - \gamma \nabla f(\alpha)$$

$$\nabla f(\alpha) = \nabla \sum_{i=1}^{n} f_i(\alpha) = \sum_{i=1}^{n} \nabla f_i(\alpha)$$

$$\frac{\partial}{\partial \alpha_j} f(\alpha) = \sum_{i=1}^{n} \frac{\partial}{\partial \alpha_j} f_i(\alpha) = 2 \sum_{i=1}^{n} \left(M_\alpha(x_i) - y_i\right) x_i^j$$

$$\nabla f(\alpha) = 2 \sum_{i=1}^{n} \left(M_\alpha(x_i) - y_i\right) \left(1, x_i, x_i^2\right)$$

Incremental Gradient Descent

$$\nabla f(\alpha) = \sum \nabla f_i(\alpha) = 2\left(M_a(y_i) - y_i\right)\begin{pmatrix} 1 \\ x_i \\ x_i^2 \end{pmatrix}$$

$\alpha^{(0)} = \alpha^{start}$ $\quad i = 1$

repeat

$$\alpha^{(k+1)} = \alpha^{(k)} - \gamma_k \nabla f_i(\alpha^{(k)})$$

$i = (i+1) \bmod n$

until $\left(\| \nabla f_i(\alpha^{(k)}) \| < \tau \right)$

return $\alpha^{(k)}$

Σ average over <= 10 steps

Stochastic Gradient Descent

Init $\alpha^{(0)} = \alpha^{start}$

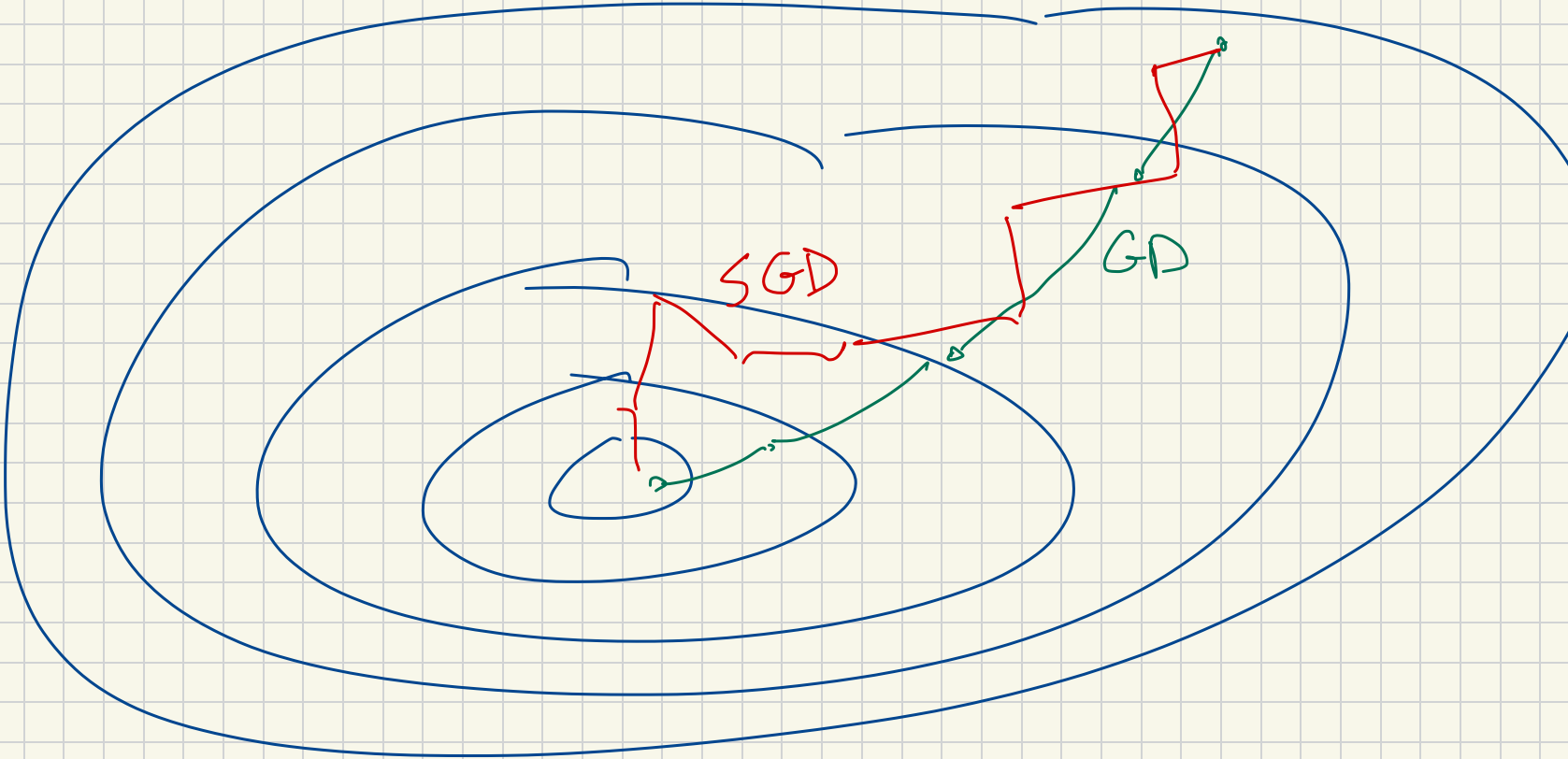
repeat

- choose $i \leftarrow [1 \dots n]$ at random
- $\alpha^{(k+1)} = \alpha^{(k)} - \gamma_k \nabla f_i(\alpha^{(k)})$

until $\left(\| \nabla f_i(\alpha^{(k)}) \| \leq \tau \right)$

← average

return $\alpha^{(k)}$


SGD
GD