FoDA - L17

Gradient Descent

SGD, on data

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

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

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

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

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

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

e.g. $\langle \alpha, x_i \rangle$

Model M_α for polynomial regression

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

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

$$= \sum_{j=0}^{2} \alpha_j x_i^j$$

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

$n = 1$

single data point

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

$j = \{0, 1, 2\}$

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

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

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

$$\frac{\partial}{\partial \alpha_j} f(\alpha) = 2\left(M_\alpha(x_1) - y_1\right) x_1^j \qquad n=1$$

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

$$= 2\left(M_\alpha(x_1) - y_1\right) \cdot \left(1, x_1, x_1^2\right)$$

LMS update rule, Widrow-Hoff update

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

decomposable function

$n > 1$

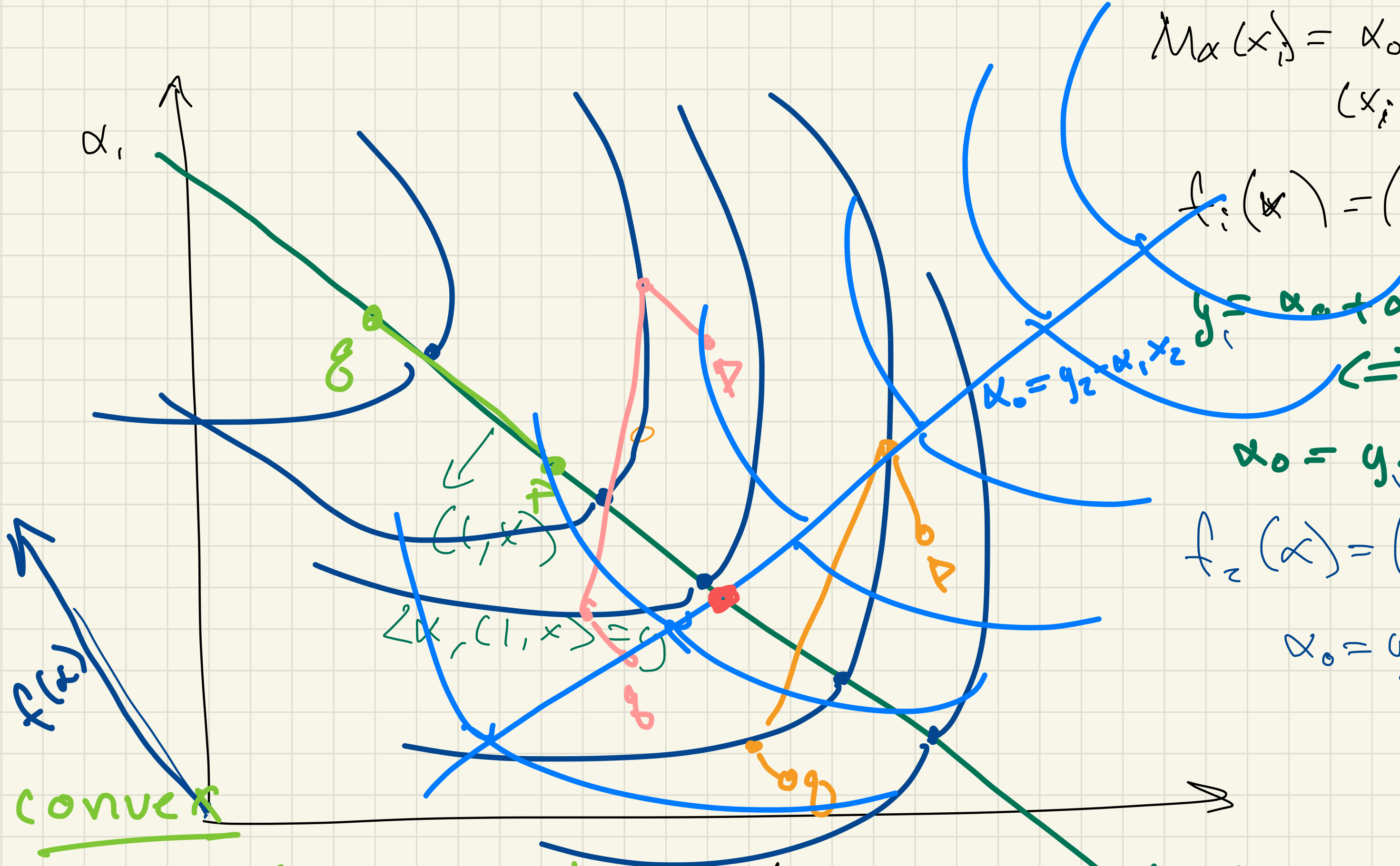
each f_i one observation (x_i, y_i)

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

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

if each f_i is (strongly) convex, then

$f = \sum_i f_i$ is also (strongly) convex



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

(x_i, y_i)

$f_i(\alpha) = (y_i - \alpha_0 - \alpha_1 x_i)^2$

$y_1 = \alpha_0 + \alpha_1 x_1$

$\iff f(\alpha) = 0$

$\alpha_0 = y_1 - \alpha_1 x_1$

$f_2(\alpha) = (y_2 - \alpha_0 - \alpha_1 x_2)^2$

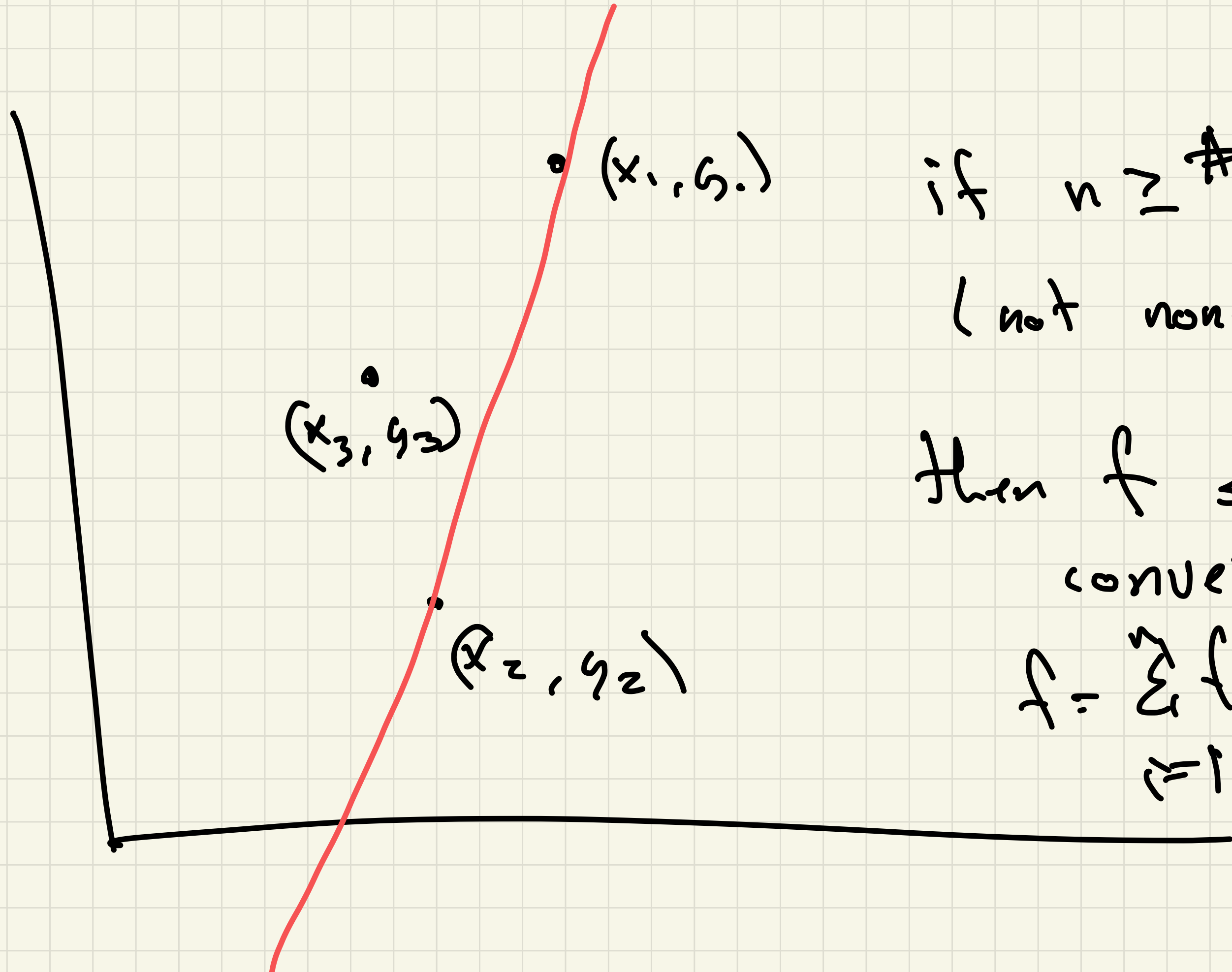
$\alpha_0 = y_2 - \alpha_1 x_2$

$\alpha_0 = y - \alpha_1 x$

$\alpha = (\alpha_0, \alpha_1)$

convex

not strongly convex



if $n \geq$ # parameters

(not non-general case)

then f strongly convex

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

Batch Gradient Descent (Batch size $= n$)

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

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

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

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

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

Incremental Gradient Descent

$$\nabla f(\alpha) \approx \nabla f_i(\alpha) = 2(M_\alpha(x_i) - y_i)(1, x_i, x_i^2)$$

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

repeat

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

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

until ($\|\nabla f_i(\alpha^{(k)})\| \leq T$)

return $\alpha^{(k)}$

$O(1)$ time to compute

take average over B steps

$B = 10$?

Stochastic Gradient Descent (SGD)

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

repeat

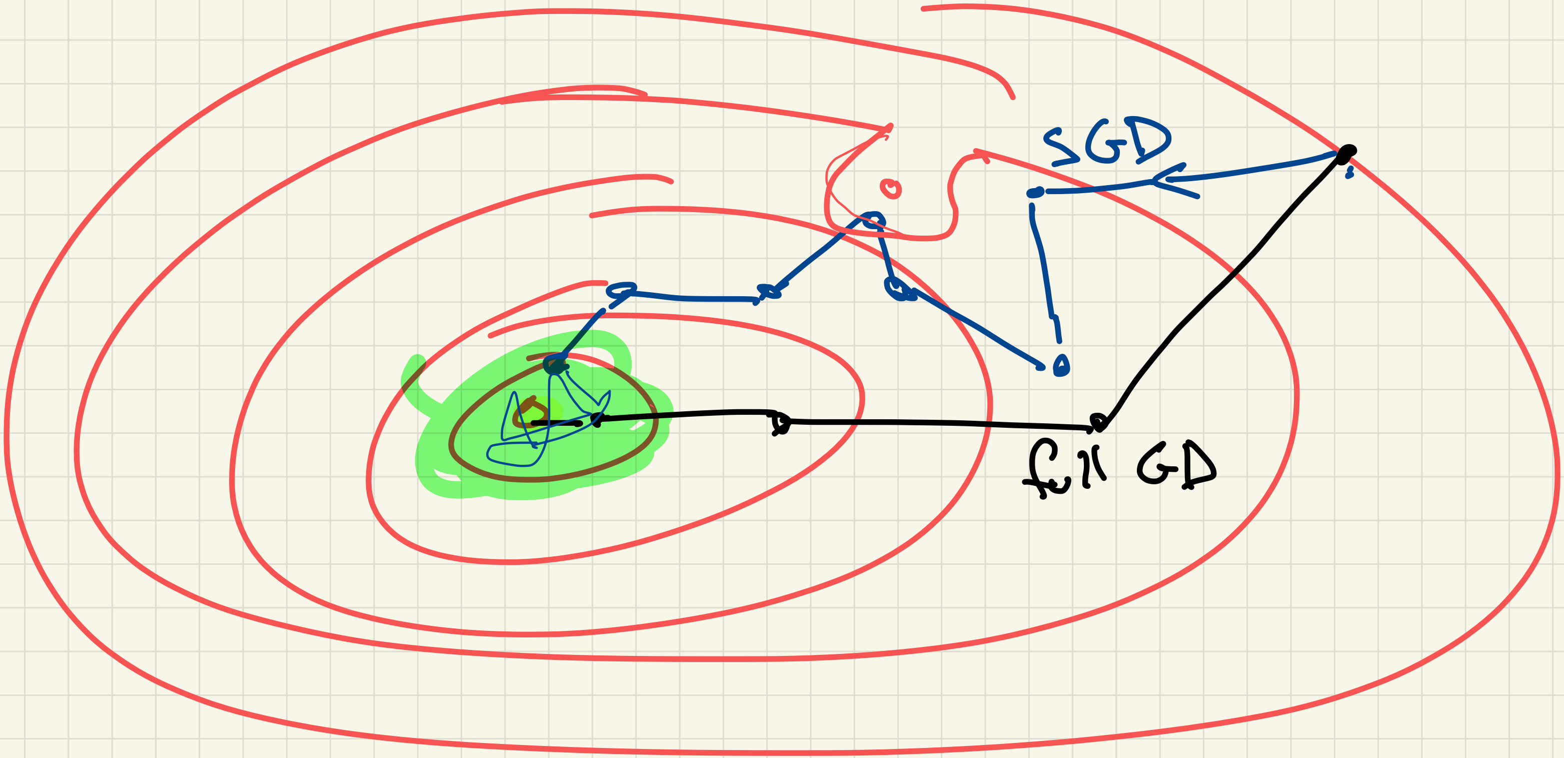
choose $i \leftarrow [1 \ldots n]$ at random

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

until $(\|\nabla f(\alpha^{(k)})\| \leq T)$

average

return $\alpha^{(k)}$



SGD: more steps
smaller steps | but each step is much, much faster