FoDA

L17

- Gradient Descent
- Fitting Models to Data

data set $(X, y) = \{(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)\}$

$\subset \mathbb{R}^d \times \mathbb{R}$

Goal: Model M_α

$\alpha = (\alpha_1, \dots, \alpha_k)$

minimize loss function

$$f(\alpha) = L((X, y), M_\alpha) \qquad f(\alpha): \mathbb{R}^k \to \mathbb{R}$$

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

$(X, y) \qquad X \in \mathbb{R}^n \qquad y \in \mathbb{R}^n$

$$(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$$

model

quadratic

$$M_\alpha(x_i) = \langle \alpha, (1, x_i, x_i^2) \rangle$$

$$= \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2$$

Gradient Descent $\quad f(\alpha) = \sum_{i=1}^{n} (M_\alpha(x_i) - y_i)^2$

\nabla

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

Single Data Point

$(x, y) \overset{\boxed{n=1}}{=} (x_1, y_1)$

$$f(\alpha) = f_1(\alpha) = (\alpha_0 + \alpha_1 x_1 + \alpha_2 x_1^2 - y_1)^2$$

convex in α

$$\nabla f(\alpha) = \left(\frac{\partial}{\partial \alpha_0} f(\alpha), \frac{\partial}{\partial \alpha_1} f(\alpha), \frac{\partial}{\partial_2} f(\alpha) \right)$$

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

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

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

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

LMS update rule

Widrow-Hoff learning rule

$$\nabla f(\alpha) = \left(\frac{\partial}{\partial \alpha_0} f(\alpha), \frac{\partial}{\partial \alpha_1} f(\alpha), \frac{\partial}{\partial \alpha_2} f(\alpha) \right) = 2 (M_\alpha(x_1) - y_1) (1, x_1, x_1^2)$$

Decomposable Functions

$(n > 1)$ Data Points

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

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

$$f(\alpha) = \sum_{i=1}^{n} f_i(\alpha) = SSE((x, y), M_\alpha)$$

Batch Gradient Descent

$$\nabla f(\alpha) = \sum_{i=1}^{n} \nabla f_i(\alpha) = \left(\sum_{i=1}^{n} \frac{\partial}{\partial \alpha_0} f_i(\alpha), \ldots, \sum_{i=1}^{n} \frac{\partial}{\partial \alpha_k} f_i(\alpha) \right)$$

LMS Update

$$= \sum_{i=1}^{n} 2(M_\alpha(x_i) - y_i)(1, x_i, x_i^2, \ldots)$$

LMS Update

$$f(\alpha) = \sum_{i=1}^{n} \left(\alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 - y_i\right)^2$$

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

LMS Update $\quad \nabla f_i(\alpha) = 2\left(\alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 - y_i\right) \cdot (1, x_i, x_i^2)$

residual $(M_\alpha(x_i) - y_i)$

$$\nabla f(\alpha) = \sum_{i=1}^{n} 2\left(\alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 - y_i\right)\left(1, x_i, x_i^2\right)$$

expanded explanatory data point

$x_i \rightsquigarrow (1, x_i, x_i^2)$

Since $f_i(\alpha)$ convex

$$\longrightarrow \sum_{i=1}^{n} f_i(\alpha) \quad \text{convex}$$

strongly convex?

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

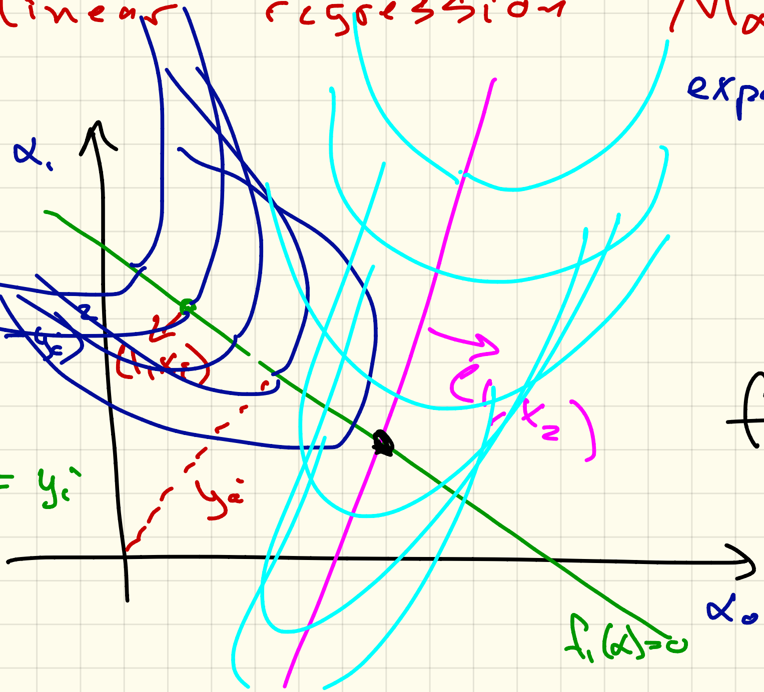
expand $x_i \to (1, x_i) \in \mathbb{R}^2$

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

$f_1(\alpha) = (\langle \alpha, (1, x_i)\rangle - y_i)^2$

$= 0$ at

$\langle \alpha, (1, x_i)\rangle = y_i$



$$f(\alpha) = f_1(\alpha) + f_2(\alpha)$$

strongly convex if $\begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ & \dots & \end{bmatrix}$ full rank

What about Big Data? n very big

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

take $\Omega(n)$

---

approximate $\nabla f(\alpha)$ in constant time

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

(x_i, y_i)

Incremental Gradient Descent

0. Initialize $\alpha^{(0)} \in \mathbb{R}^d$, $i=1$, $t=0$
1. repeat

   $\alpha^{(t+1)} = \alpha^{(t)} - \gamma \nabla f_i(\alpha^{(t)})$ ← Grad at 1 data point.

   maybe make smaller than full GD

   $i := (i+1) \bmod n$
2. Until $(\|\nabla f(\alpha^{(t)})\| \leq T)$ ← take sliding average over B step

Stochastic Gradient Descent (SGD)

0. $x^{(0)} = x \in \mathbb{R}^d$

1. <u>repeat</u>

    a. Randomly choose $i \in \{1, 2, \ldots n\}$

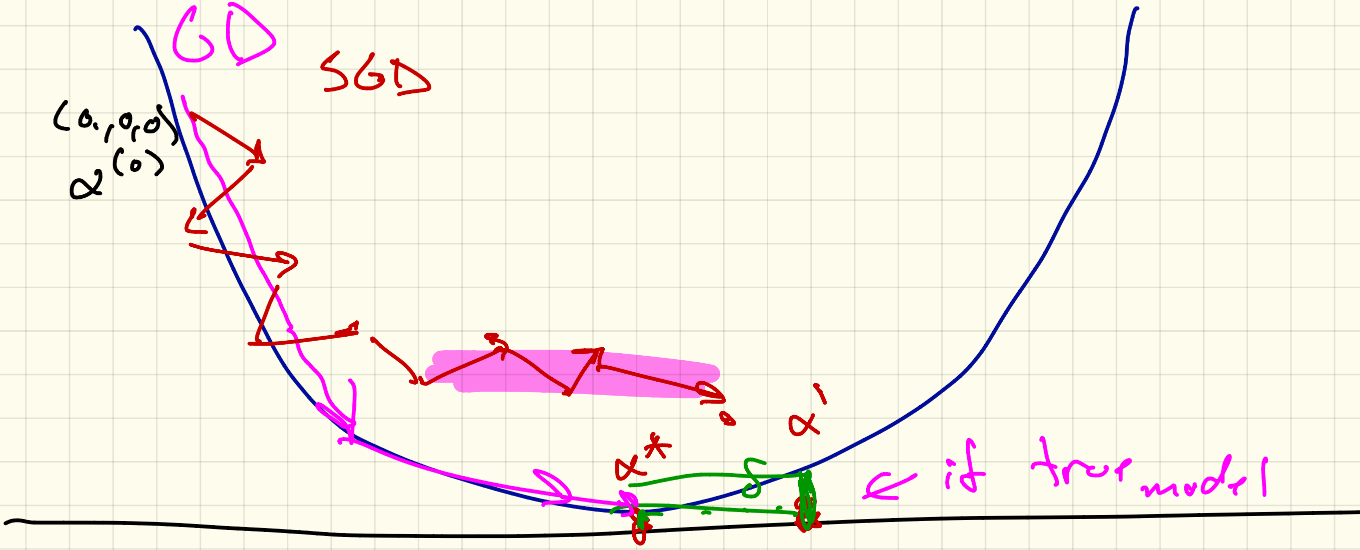
    b. $x^{(k+1)} = x^{(k)} - \gamma \nabla f_i(x^{(k)})$

2. <u>until</u> $(\|\nabla f_i(x^{(k)})\| \leq T)$

3. return $x^{(k)}$

... **Recently**

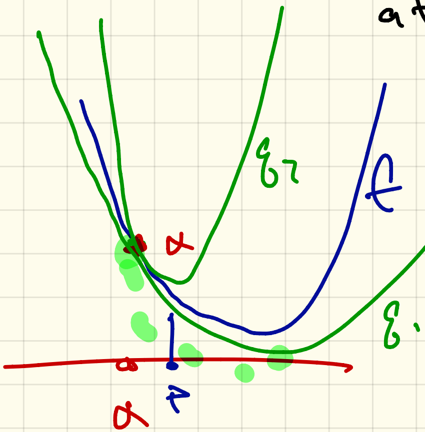
SGD tends to generalize better than full / batch GD.



Strongly Convex

$f(\alpha)$ is st. convex if

at each $\alpha \in \mathbb{R}^d$ $\exists$ quadratic functions $g_1, g_2 : \mathbb{R}^d \to \mathbb{R}$



so all $p \in B_r(\alpha)$ have

Grad ∇f L-Lipschitz

$$g_1(p) \le f(p) \le g_2(p)$$

$$g_1(\alpha) = f(\alpha) = g_2(\alpha)$$

$$g_1(p) = f(\alpha) + \langle \nabla f(\alpha), p - \alpha \rangle + \frac{\eta}{2} \| p - \alpha \|^2$$

Decomposable function

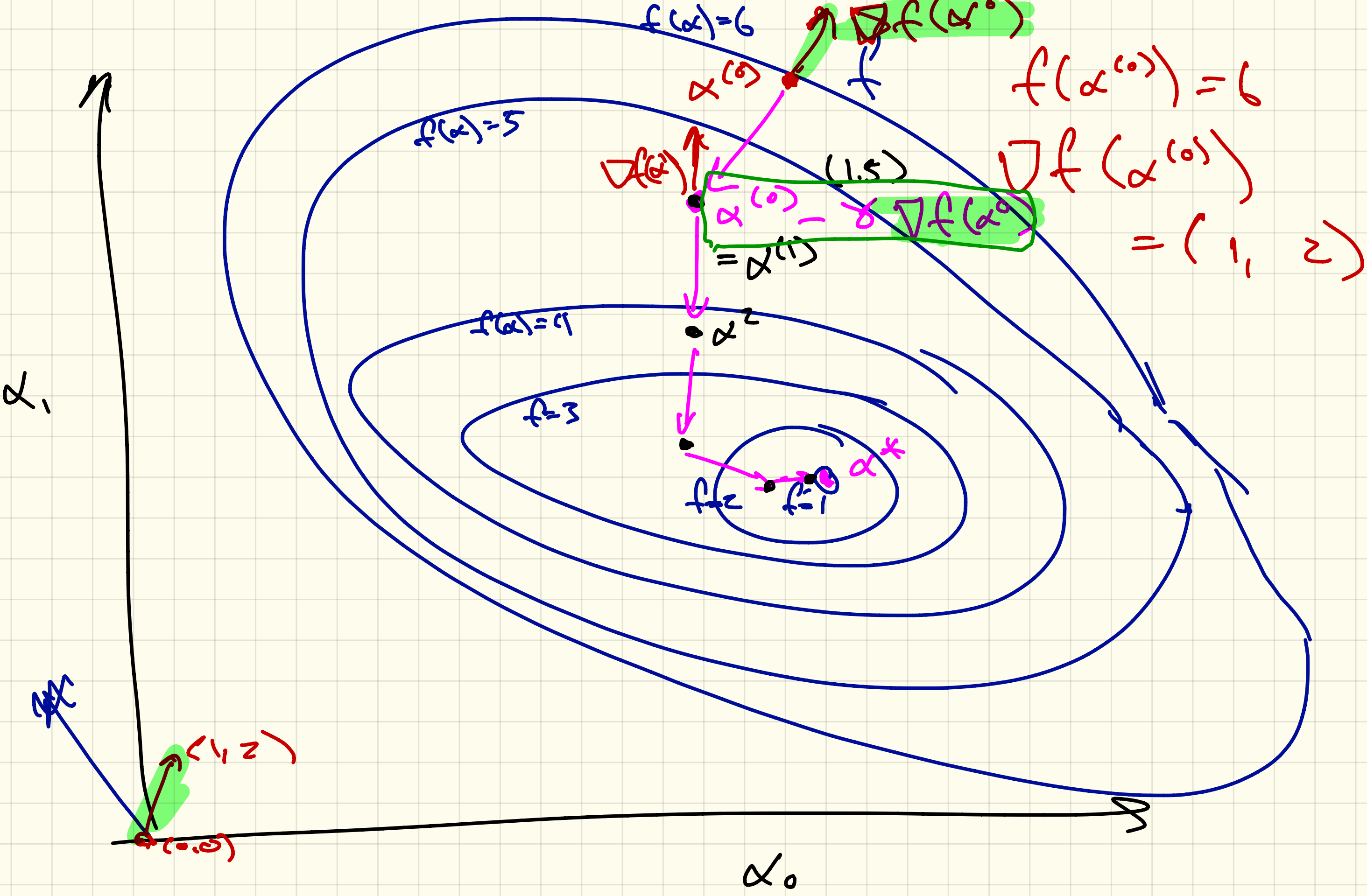
$$f(\alpha) = \sum_{i=1}^{n} f_i(\alpha)$$

simple form

usually each $f_i \Leftarrow$ 1 data point (x_i, y_i)

Non-decomposable fxn

$$f(\alpha) = (\alpha_1 + 1 + \alpha_2\alpha_3)^2 (\alpha_1 - \alpha_2)(\alpha_3^2)$$



$f(\alpha)=6$
$\nabla f(\alpha^{0})$
f
$\alpha^{(0)}$
$f(\alpha)=5$
$\nabla f(\alpha^{1})$
$(1,5)$
$\alpha^{(0)} - \delta \nabla f(\alpha^{0})$
$= \alpha^{(1)}$
α^{2}
$f(\alpha)=4$
$f=3$
α^{*}
$f=2$
$f=1$
α_1
α_0
$(1, 2)$
$(0,0)$
$f(\alpha^{(0)}) = 6$
$\nabla f(\alpha^{(0)}) = (1, 2)$