Today:
onvergence rates

Assignment

2 marked assign due now

3 marked assign due Monday

4 due now

5 out this weekend

Last week: neural networks:

instead of $\sum_{i=1}^{N} f_i(\bar{w}^T \bar{x}_i)$, we have $\sum_{i=1}^{N} f_i(\bar{w}_1^T g(\bar{w}_2^T \bar{x}_i))$

or $\sum_{i=1}^{N} f_i(\bar{w}_1^T g(\bar{w}_2^T h(\bar{w}^T \bar{x}_i$

Where do subgradients exist?

- everywhere for convex functions

Why focus on gradient methods?

- Newton, linprog, quadprog, etc cost $O(d^3)$ or $O(n^3)$ or larger

Convergence rate of gradient descent

Gradient method: $x^{t+1} = x^t - \alpha_t \nabla f(x^t)$

How fast does this converge?

- to answer this you need assumptions
  (function or gradient can't change too quickly)

We'll assume $$\underline{\mu I \preccurlyeq \nabla^2 f(x) \preccurlyeq L I}$$

$(\mu > 0)$ "strongly convex"

$(L < \infty)$ "strongly smooth"

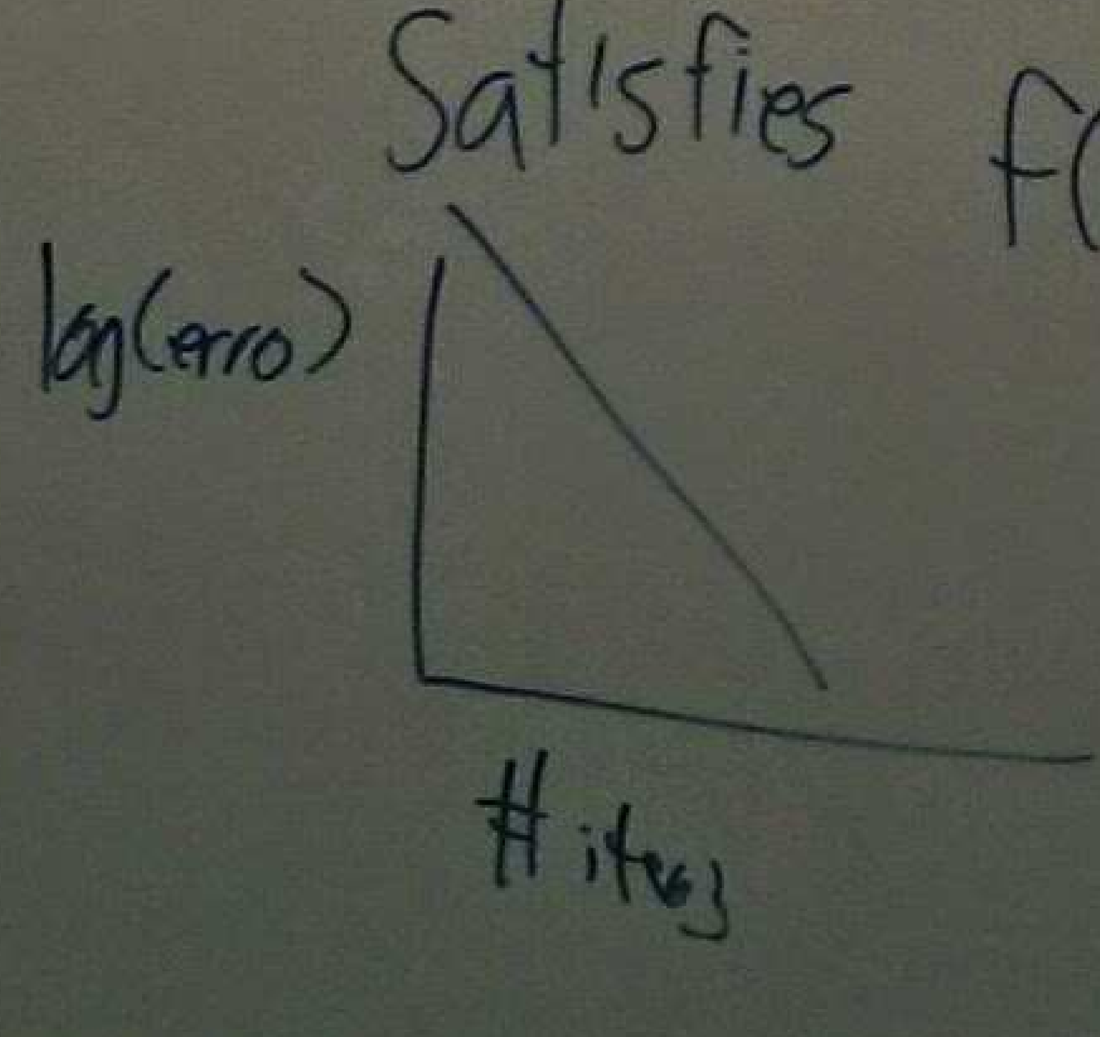
⊛ 0^{th}-order: $(f(x) - \frac{\mu}{2}\|x\|^2)$ is convex

⊛ 1^{st}-order: $\nabla f(x)$ is L-Lipschitz continuous
$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x-$$

→ $$\forall_x \forall_v \; v^T \nabla^2 f(x) v \leq v^T (LI) v$$
$$= L(v^T I v) = L v^T v = L\|v\|^2$$

Proposition: With $\alpha_t = 1/L$, gradient method

Satisfies $$f(x^t) - f(x_*) \leq \left(1 - \frac{\mu}{L}\right)^t [f(x^0) - f(x_*)]$$



Proof:

One v

$\forall_{x,y,?} \exists_z$ st

"guaranteed progress"

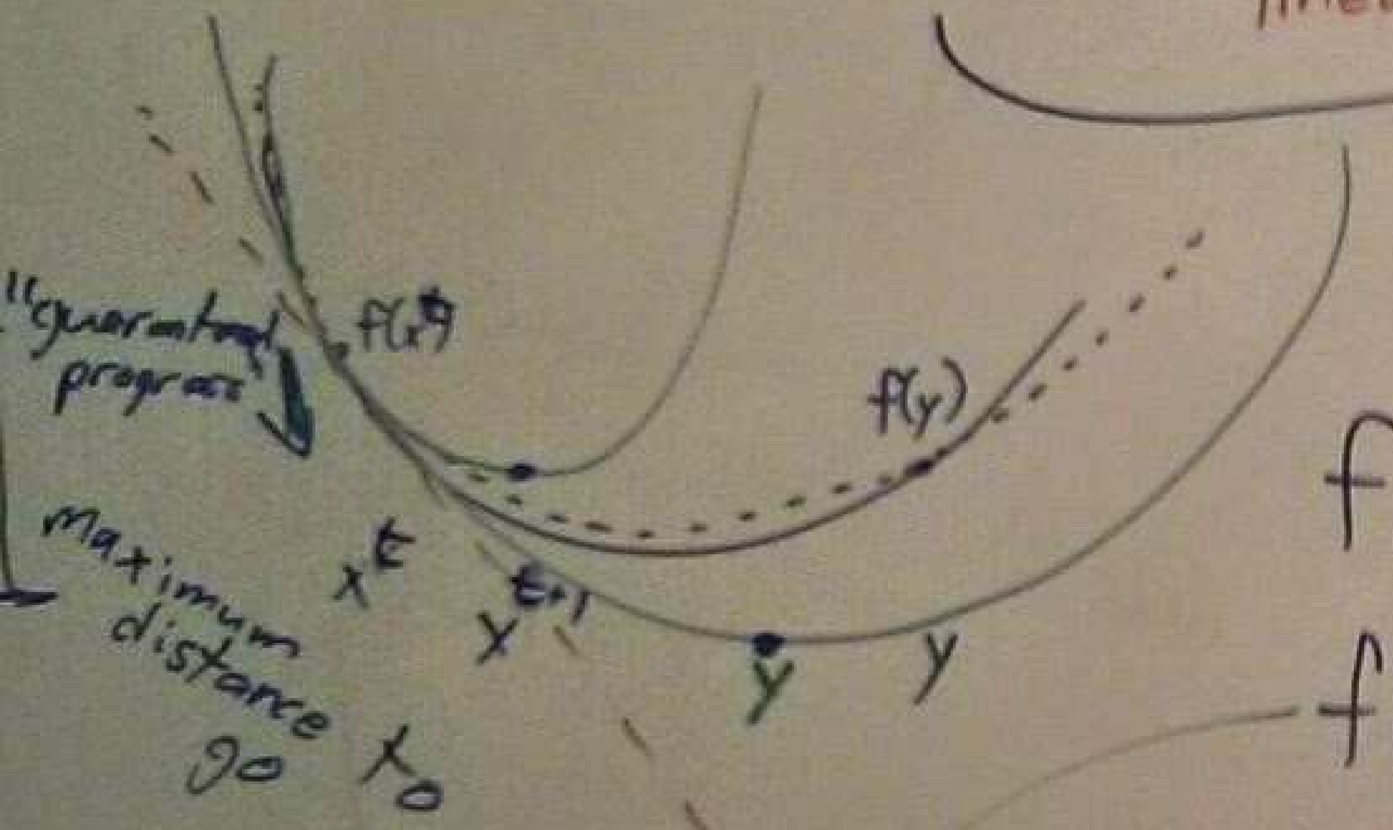
maximum distance to go

$f(x_*) \geq$

Proof:

One version of Taylor expansion:

$$\forall_{x,y}\ \exists_z \text{ st } f(y) = \underbrace{\underbrace{f(x) + \nabla f(x)^T(y-x)}_{\text{linearization}} + \frac{1}{2}(y-x)^T \nabla^2 f(z)(y-x)}_{\text{quadratic function of } y}$$



$$f(y) \le f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|^2$$

$$f(y) \ge f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}\|y-x\|^2$$

minimize in terms of 'y' $(y = x - \frac{1}{\mu}\nabla f(x))$

$$f(x_*) \ge f(x^t) - \frac{1}{2\mu}\|\nabla f(x^t)\|^2$$

Set $x = x^t$, $y = x^{t+1} = x^t - \frac{1}{L}\nabla f(x^t)$

$$f(x^{t+1}) \le f(x^t) + \nabla f(x^t)^T(x^{t+1} - x^t) + \frac{L}{2}\|x^{t+1} - x^t\|^2$$

use $x^{t+1} - x^t = -\alpha_t \nabla f(x^t)$:

$\| \le L\|x-y\|$

$$-\|\nabla f(x^t)\|^2 \le -2\mu[f(x^t) - f(x_*)] \quad (*)$$

$$= f(x^t) + \nabla f(x^t)^T\left(-\left(\tfrac{1}{L}\right)\nabla f(x^t)\right) + \frac{L}{2}\left\|-\left(\tfrac{1}{L}\right)\nabla f(x^t)\right\|^2$$

$$\overset{(*)}{=} f(x^t) - \frac{1}{2L}\|\nabla f(x^t)\|^2 \quad (*)$$

$x_*)]$

$$f(x^t) - f(x_*) \leq \left(1 - \frac{\mu}{L}\right)\left[f(x^{t-1}) - f(x_*)\right]$$

Add $-f(x_*)$ to $\circledast$

$$f(x^{t+1}) - f(x_*) \leq f(x^t) - f(x_*) - \frac{1}{2L}\|\nabla f(x^t)\|^2$$

Use $\circledast$

$$\leq f(x^t) - f(x_*) - \frac{2\mu}{2L}\left[f(x^t) - f(x_*)\right]$$

$$= \left(1 - \frac{\mu}{L}\right)\left[f(x^t) - f(x_*)\right]$$

$$\leq \left(1 - \frac{\mu}{L}\right)\left(1 - \frac{\mu}{L}\right)\left[f(x^{t-1}) - f(x_*)\right]$$

$\vdots$

$$f(x^t) - f(x_*) \leq \left(1 - \frac{\mu}{L}\right)^t\left[f(x_0) - f(x_*)\right]$$

$$\left(1 - \frac{\mu}{L}\right)^t \leq \exp\left(-\frac{t\mu}{L}\right)$$

$f(x^t) - f(x_*) \leq \epsilon$

$t = O(\log(1/\epsilon))$

total cost: $O(Nd\log(1/\epsilon))$

Marked Assign 3

Assign 4

Today:
Convergence rates

Deterministic Gradient Complexity Zoo

Strongly Smooth

+ strongly convex

+ convex

"linear" $O((1-\frac{\mu}{L})^t)$

$O(1/t)$ "sublinear"

+ Lipschitz Hessian, Newton's method

extrapolation

extrapolation (heavy-ball, momentum, Nesterov's accelerated gradient)

"superlinear" $O(\prod_{i=1}^{t} \rho_i), \rho_i \to 0$

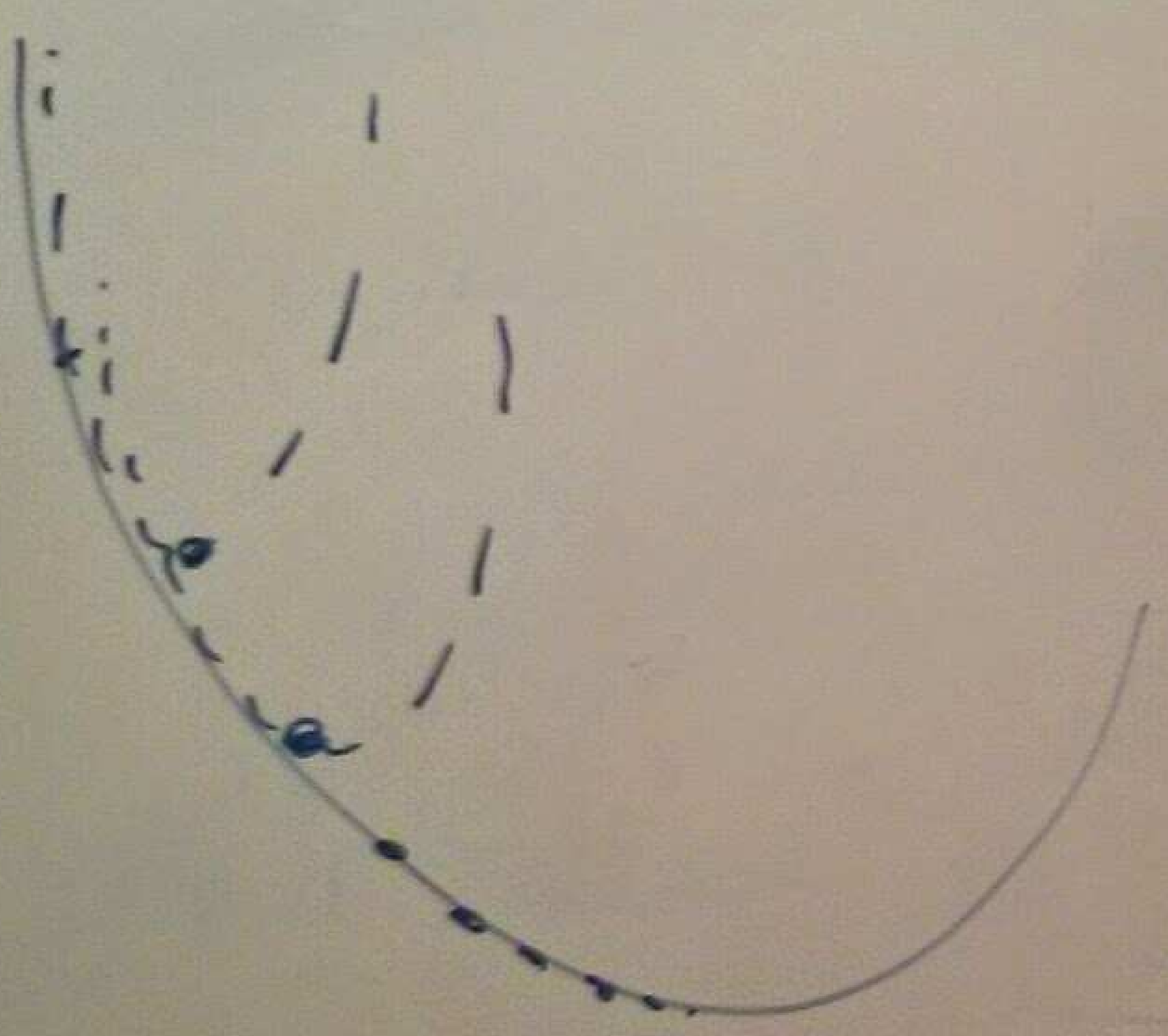
$O((1-\sqrt{\frac{\mu}{L}})^t)$

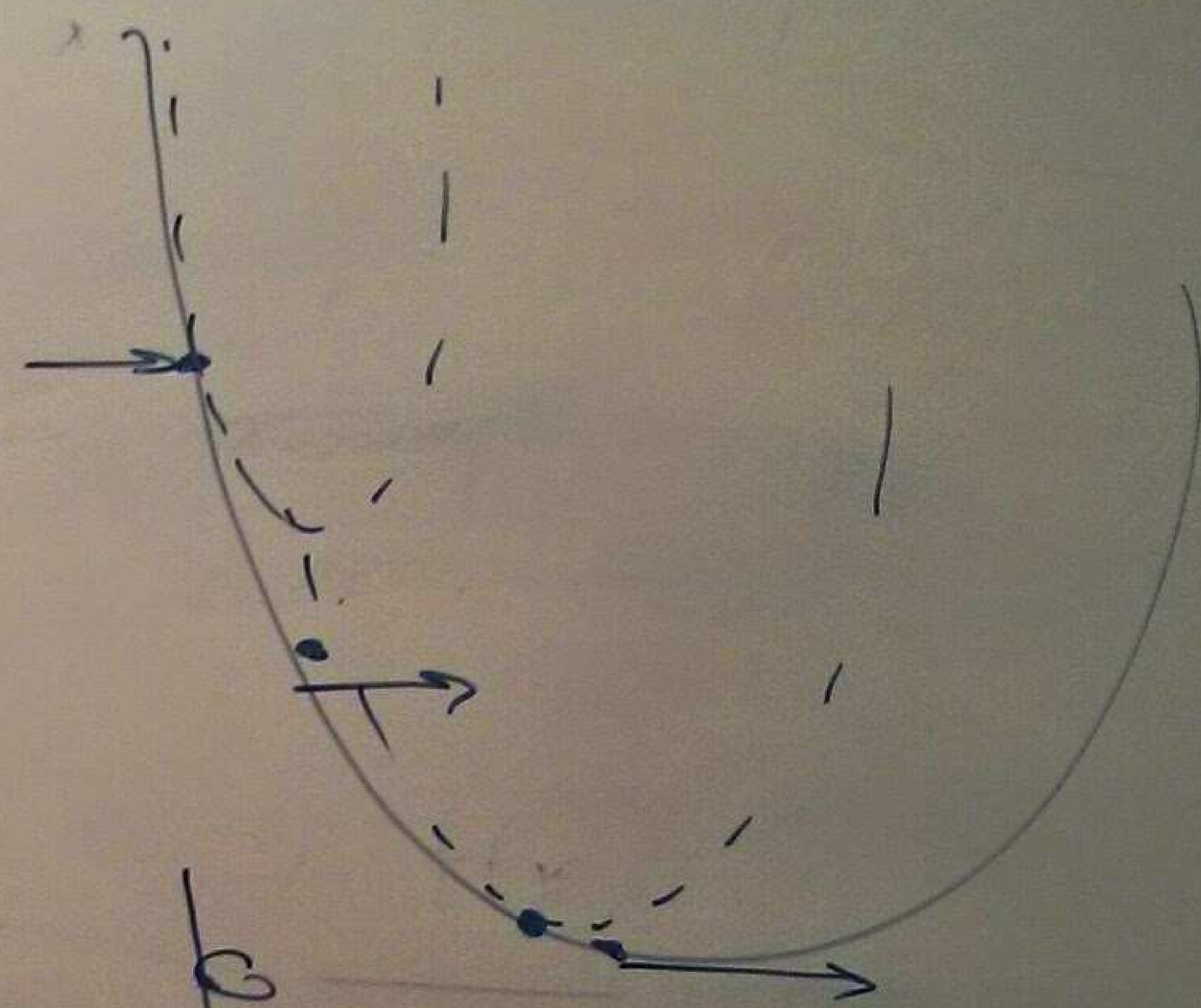
$O(1/t^2)$

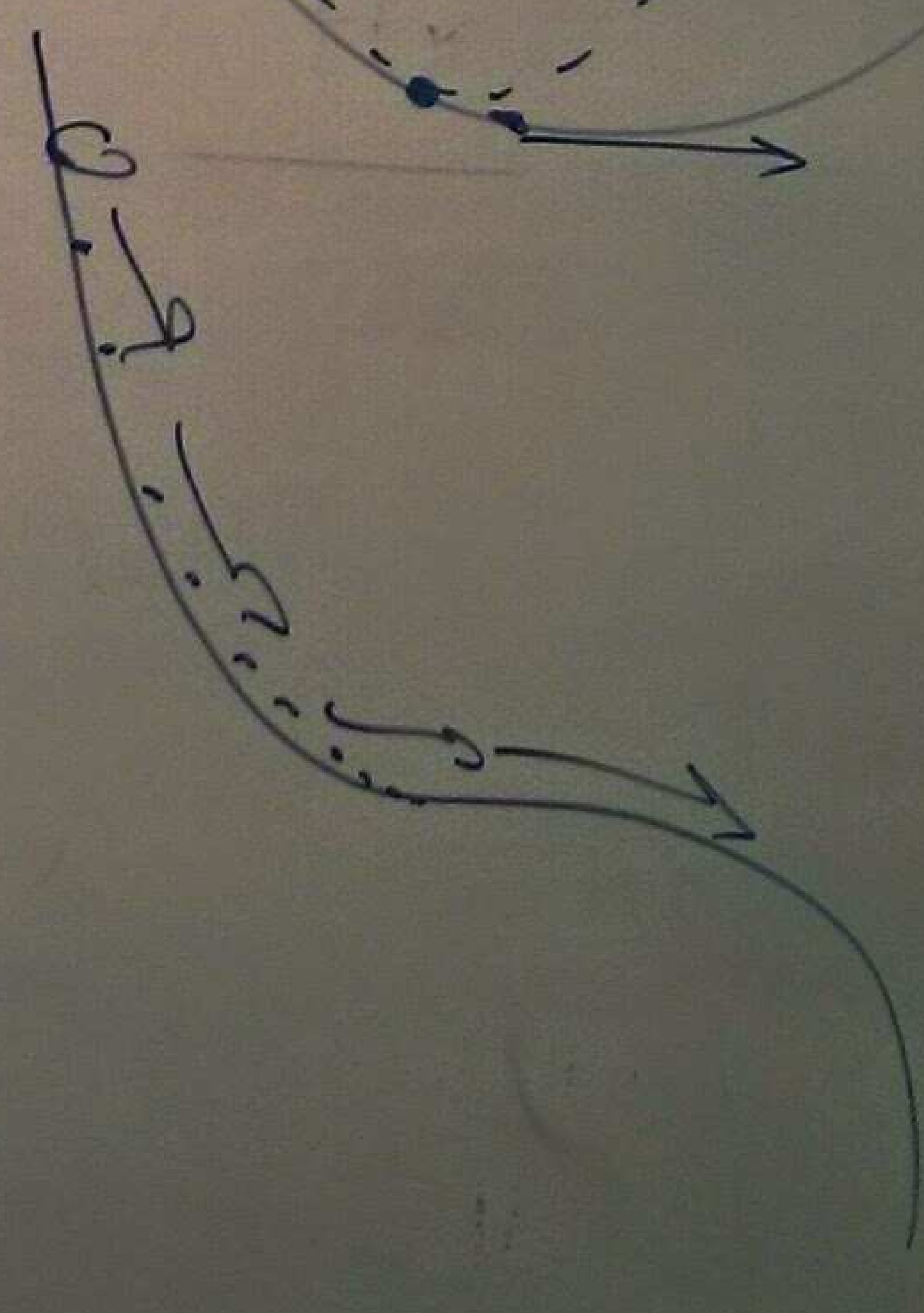
Approximate Newton:

- Barzilai-Borwein
- Quasi-Newton (L-BFGS)
- Hessian-free Newton

$$x^{t+1} = x^t - \alpha_t \nabla f(x_t) + \beta_t (x_t - x_{t-1})$$







* Line-search
  - important in practice
  - no effect on theory

Sub-gradient and/or Stochastic Gradient

x_* is not a fixed-point

$$x_* \neq x_* - \alpha g(x_*)$$

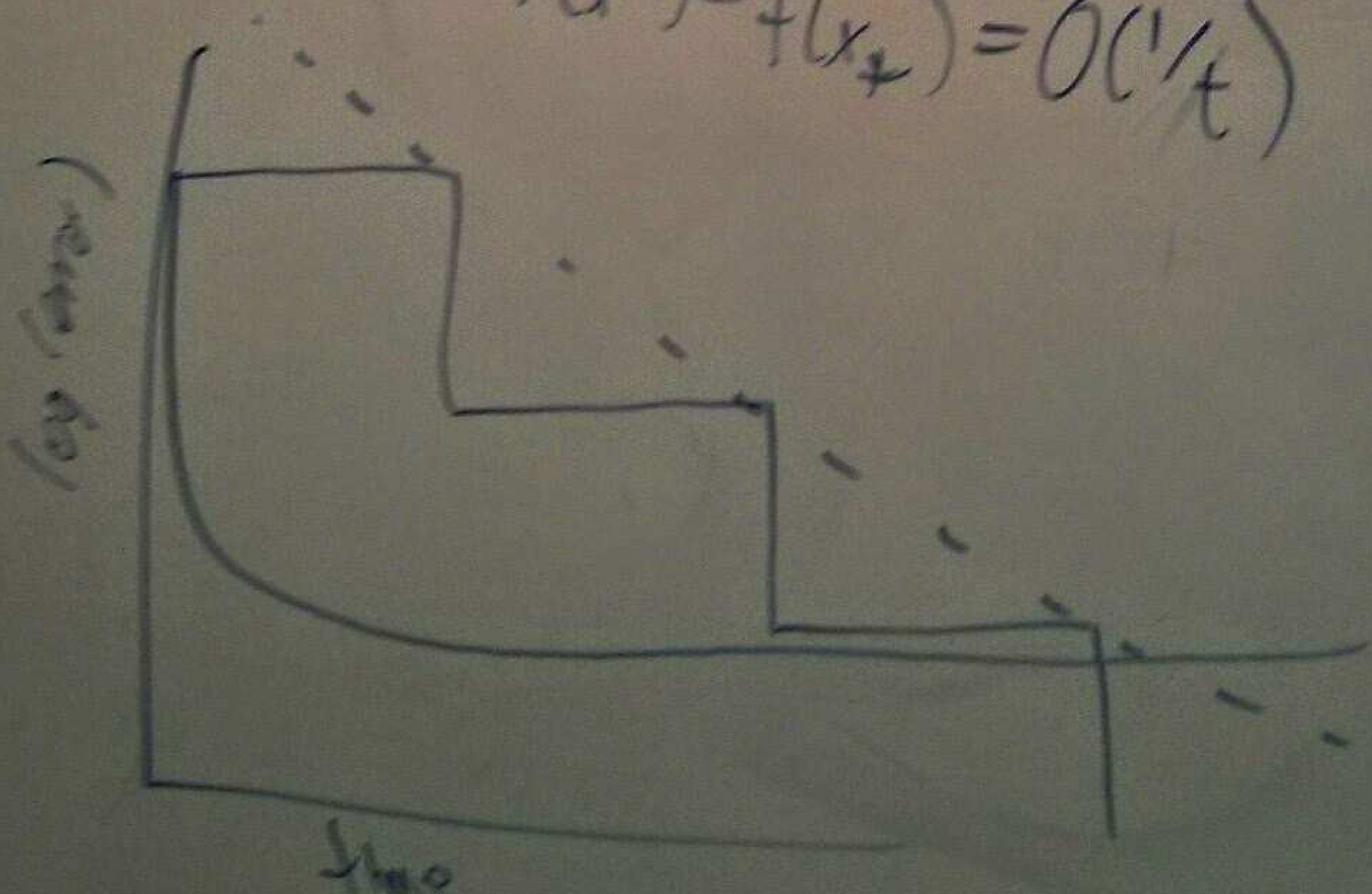
to get around this, need $\alpha_t \longrightarrow 0$

Robbins-Monro: $\sum_{t=1}^{\infty} \alpha_t = \infty$

$$\sum_{t=1}^{\infty} (\alpha_t)^2 < \infty$$

E.g., $\alpha_t = \frac{1}{\mu t}$

gives $f(x^t) - f(x_*) = O(1/t)$



Subgradient/Stoch. Grad. Complexity Zoo

Lipschitz f

- $\xrightarrow{\text{strongly}}$ $O\left(\frac{\log(t)}{t}\right)$
  - $\xrightarrow{\text{tail averaging}}$ $O(1/t)$
  - $\xrightarrow{\text{strongly smooth}}$ $O(1/t)$
    - $\xrightarrow{\text{constant step-size}}$ $O((1-2\mu\alpha)^t)+O(\alpha)$
    - $\xrightarrow{\text{finite SAG}}$ $O(\rho^t)$
- $\xrightarrow{\text{convex}}$ $O(1/\sqrt{t})$
  - $\xrightarrow{\text{smooth approximation}}$ $O(1/\sqrt{t})$
    - $\xrightarrow{\text{(exact gradient) extrapolation}}$ $O(1/t)$
  - $\xrightarrow{\text{finite (strongly smooth) SAG}}$ O(

finite

(strongly smooth)

SAG

$\rightarrow O(1/t)$

radient)

tion

$(1-\frac{\mu}{L})^t \leq \exp(-t\frac{L}{\mu})$

$f(x^t)-f(x_*) \leq \epsilon$

$t = O(\log(1/\epsilon))$

total cost: $O(Nd\log(1/\epsilon))$

Marked
Assign 3

Assign. 4