Course Projects
Proximal-Gradient
Fenche Duality

A2: pick at end of class

A3: marked version due now

A4: marked version due next Monday

A5: due Wednesday of next week.

Proposal: due Nov 3.

- groups of 1-3
- maximum 2 pages (shorter is ok)
- written for me (no need for background on ML)
- this is just a sanity check

urse Project Suggestions

Apply ML to new domain (Supervised problem from your research? Kaggle competition?)

compare/contrast ML methods from class.

Compare different extensions of a method from clas
on a set of test problems.
(you can add your own extensions, too)

Explore scaling methods up to BIG data sets (e.
(you can add your own strategies, too)

Prove a theoretical result.

Do a larger coding project (implement a bunch of ML methods)

Last time:

today composite, proximal-gradient

gradient: linear convergence

Subgradient: sublinear   stoch grad: sublinear convergence

finite SAG

12:30 tomorrow at ESB 4133

Gradient Method:

- We want to solve $\min_{x \in \mathbb{R}^d} f(x)$ (smooth)

- Minimize quadratic approximation

$$x^{t+1} = \operatorname*{argmin}_{y \in \mathbb{R}^d} f(x^t) + \langle \nabla f(x^t), y - x \rangle + \frac{1}{2\alpha_t} \|y - x\|^2$$

- Solution is gradient method:

$$x^{t+1} = x^t - \alpha_t \nabla f(x^t)$$ (Guaranteed decrease if $\alpha_t < \frac{2}{L}$

rnels)

Projected Gradient Method:

- We want to solve $\min_{x \in C} f(x)$ (smooth)
- Minimize quadratic approximation

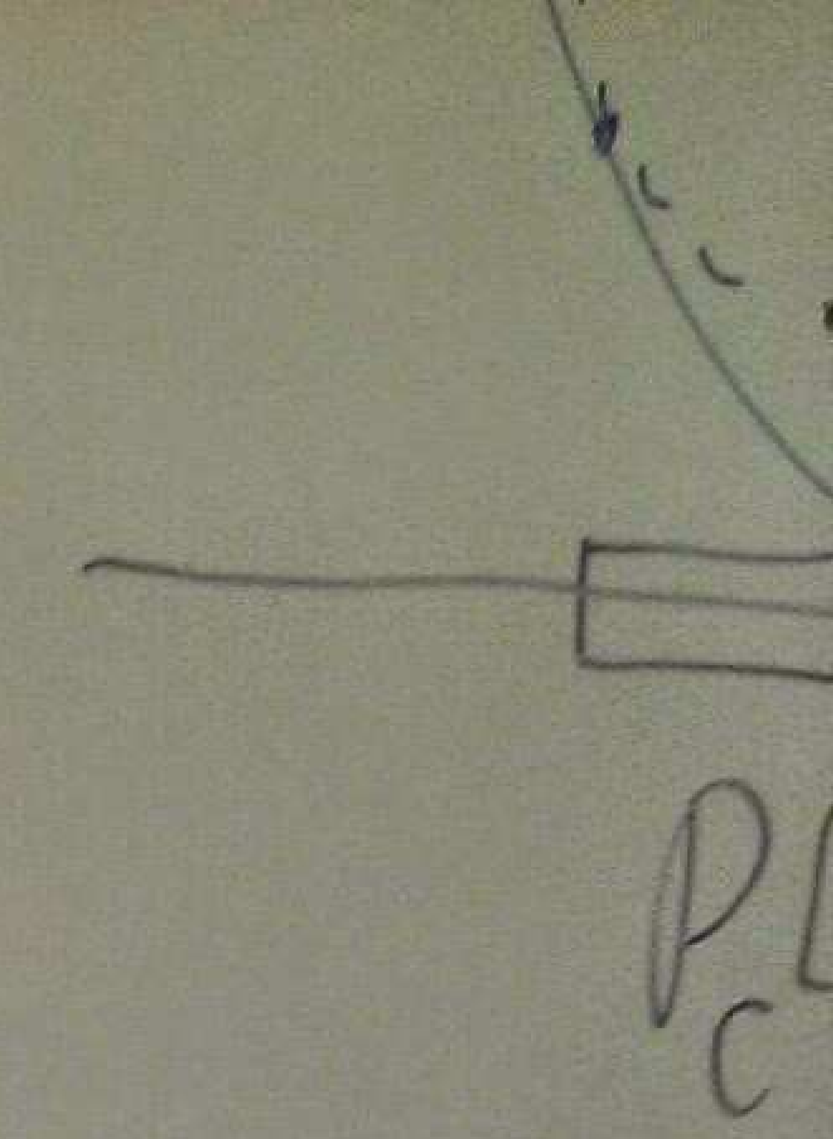
$$x^{t+1} = \operatorname*{argmin}_{y \in C} f(x^t) + \langle \nabla f(x^t), y - x \rangle + \frac{1}{2\alpha_t} \|y - x\|^2$$

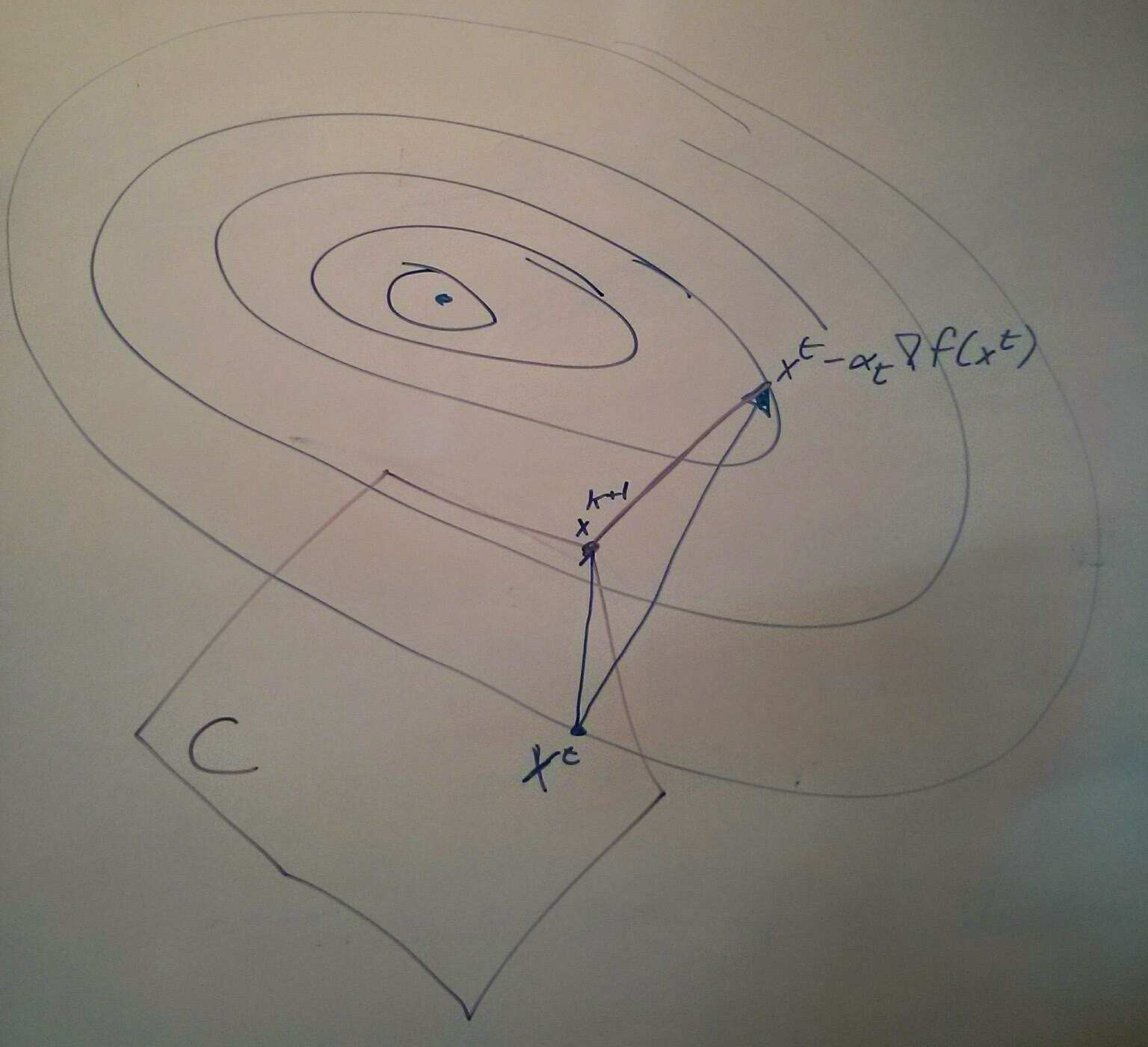
- Solution is

$$x^{t+1} = \underbrace{P_C}_{\text{"projection"}}\Big[\overbrace{x^t - \alpha_t \nabla f(x^t)}^{\text{gradient method:}}\Big]$$

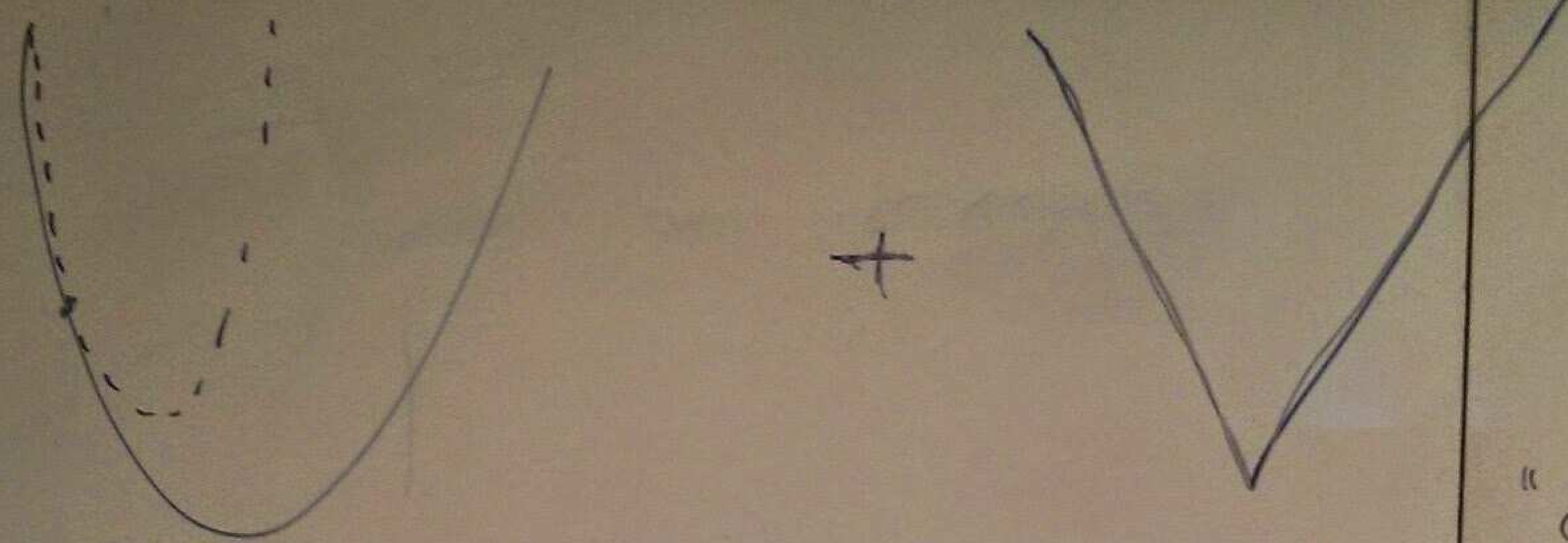
(Guaranteed decrease if $\alpha_t < \frac{2}{L}$)

- Linear convergence rate.





$x^t - \alpha_t \nabla f(x^t)$
x^{t+1}
x^t
C



$r(x) = \lambda \|x\|$

$x^{t+1/2} = x^t -$

"$\alpha r(\cdot)$"

$x^{t+1} = \mathrm{sign}(x$

Proximal Gradient Method:

- We want to solve $\min_{x \in \mathbb{R}^d} \underset{\text{(smooth)}}{f(x)} + \underset{\text{(non-smooth)}}{r(x)}$

  $\mathrm{prox}_{\alpha r}[x] = \mathrm{argmi}$
  $y \in \mathbb{R}$

- Minimize quadratic approximation

$$x^{t+1} = \underset{y \in \mathbb{R}^d}{\mathrm{argmin}}\ f(x^t) + \langle \nabla f(x^t), y - x \rangle + \frac{1}{2\alpha_t}\|y - x\|^2 + r(x)$$

- Solution is gradient method:

$$x^{k+1} = \mathrm{prox}_{\alpha r}\left[x^t - \alpha_t \nabla f(x^t)\right]$$

"proximal operator"

(Guaranteed decrease if $\alpha_t < \frac{2}{L}$)

- Linear convergence rate.

- You c

x^*

$$\underbrace{r(x) = \lambda \|x\|_1}$$

$$r(x) = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{if } x \notin C \end{cases}$$

$$x^{t+1/2} = x^t - \alpha_t \nabla f(x^t)$$

"$\alpha r(\cdot)$"

$$x^{t+1} = \text{sign}(x^{t+1/2}) \max\{0, |x^{t+1/2}| - \alpha_t \lambda\}$$

"soft-t

non-smooth)

$$\text{prox}_{\alpha r(\cdot)}[x] = \underset{y \in \mathbb{R}^d}{\text{argmin}} \frac{1}{2}\|y - x\|^2 + \alpha r(y)$$

(x)"

– You can show x^* is optimal iff

$$x^* = P_C[x^* - \nabla f(x^*)]$$

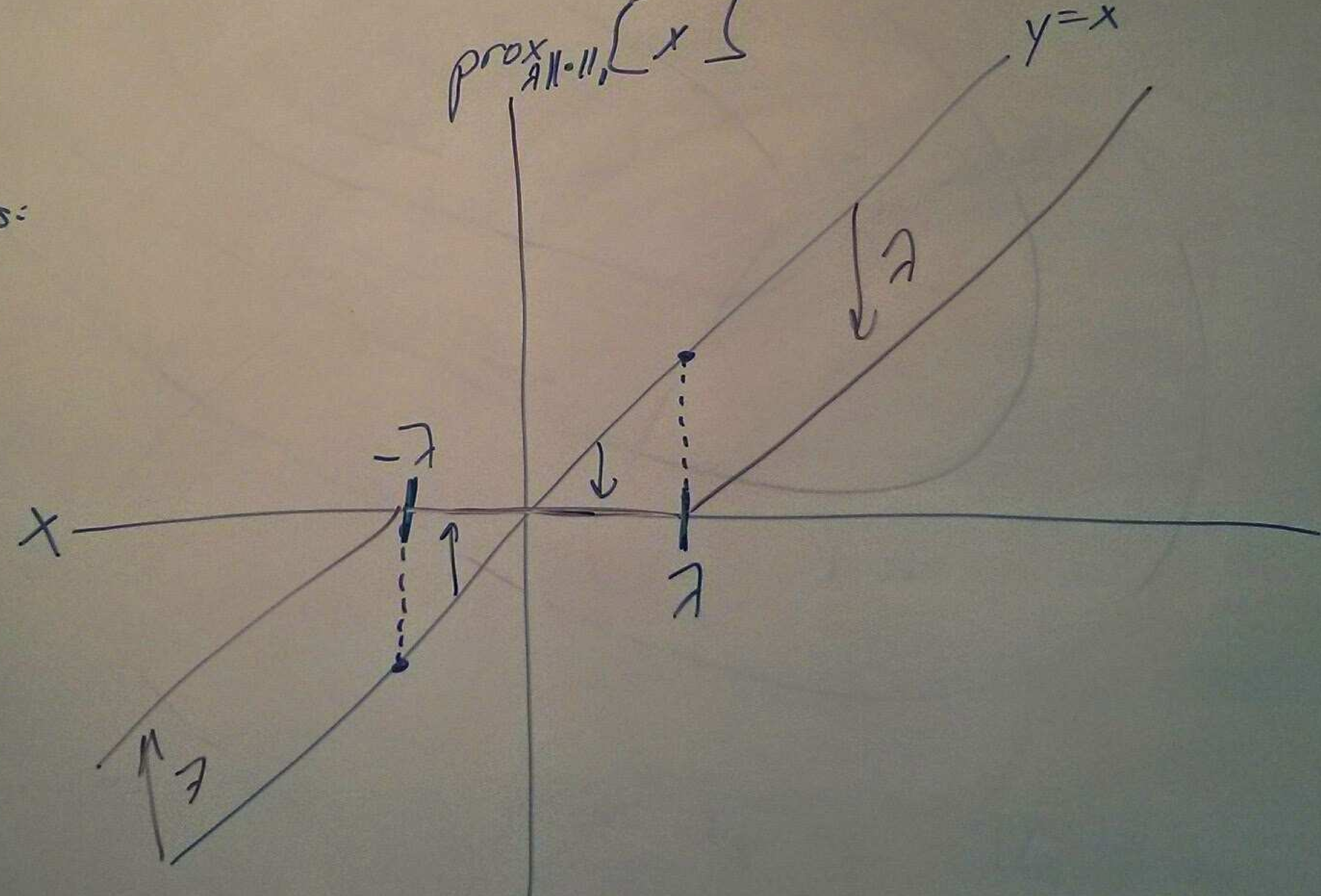
$$+ \frac{1}{2\alpha_t}\|y - x\|^2 + r(x)$$

Marked

Ass 2

(wait until

smooth optimizers:



nal- Gradient

Duality

(*) Other constrained/non-smooth optimizers:

- penalty methods
- log-barrier "
- augmented Lagrangian
- ADMM
- interior-point

Fenchel Duality

s:

Why?

- generic smoothing of non-smooth convex functions
- dual may have fewer variables
- generic kernel trick
- formulate non-convex problems
  as convex problems.

Primal problem: $\min_{x \in \mathbb{R}^d} P(x) = f(Ax) + g(x)$,

Fenchel dual: $\max_{y \in \mathbb{R}^n} D(y) = -f^*(y) - g^*(A^T y)$

f^* is the "convex conjugate" of f.

(*) $\min_{x_1, x_2} f(x_1) + g(x_2)$, s.t. $x_1 = A x_2$

s:

Primal problem: $\min_{x \in \mathbb{R}^d} P(x) = f(Ax) +$

Fenchel dual: $\max_{y \in \mathbb{R}^n} D(y) = -f^*(y)$

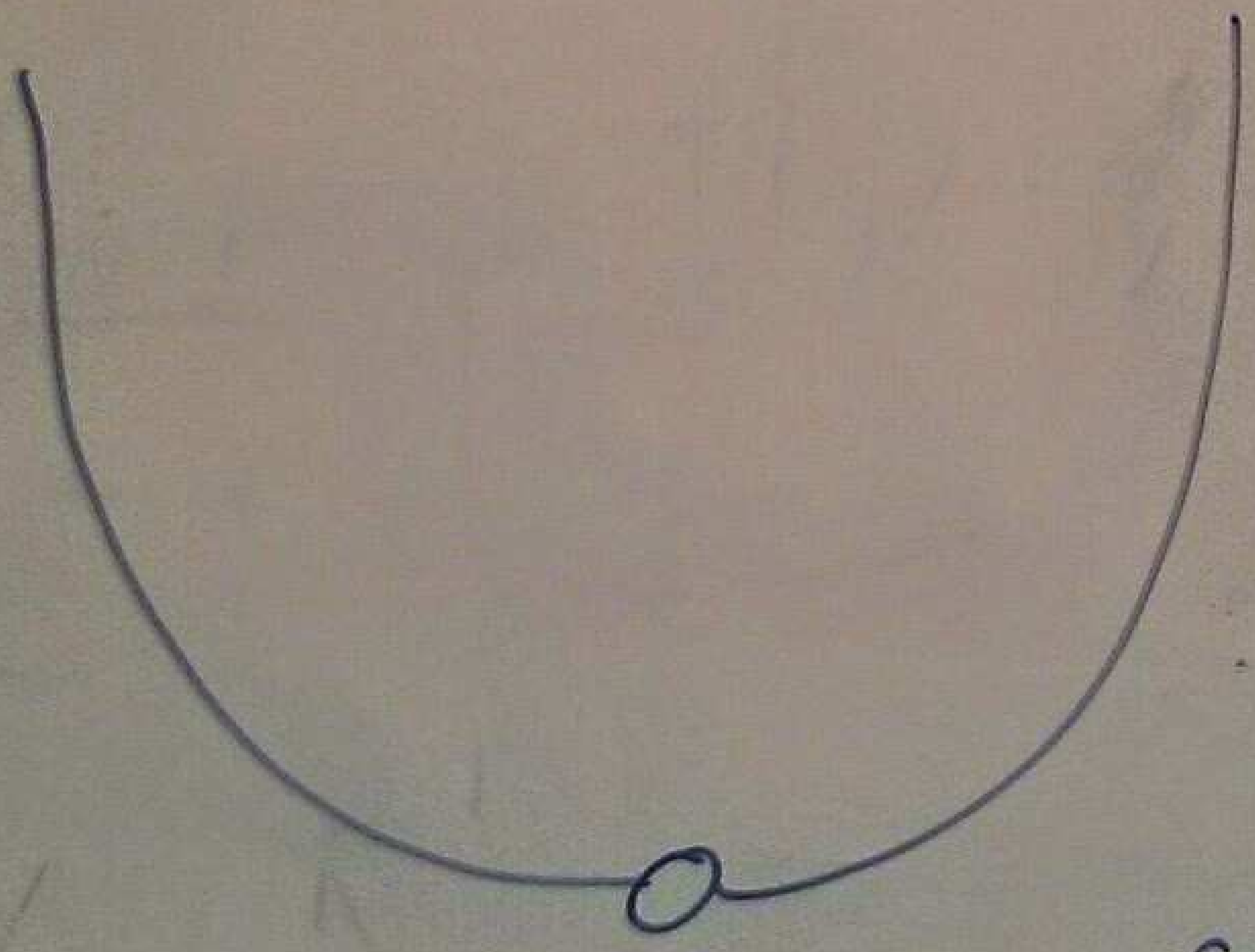
f^* is the

(*) $\min_{x_1, x_2} f(x_1) + g(x_2)$, s.t. $x_1 = Ax_2$

Notes:

1. $y \in \mathbb{R}^n$
2. $D(y) \leq P(x)$ (weak duality)
3. $D(y^*) = P(x^*)$ (strong duality)
   under certains conditions (f, g are convex)
4. $-D(y)$ is always convex.
5. $P(x)$ strongly-convex $\Longrightarrow D(y)$ is smooth.
6. $f^{**} = f$, if f is convex and 'closed'

$\inf_x \{e^x\} = 0 \quad \min_x \{e^x\}$ ONE



$y = x^2, \quad x \neq 0$

$y = \infty, \quad x = 0$

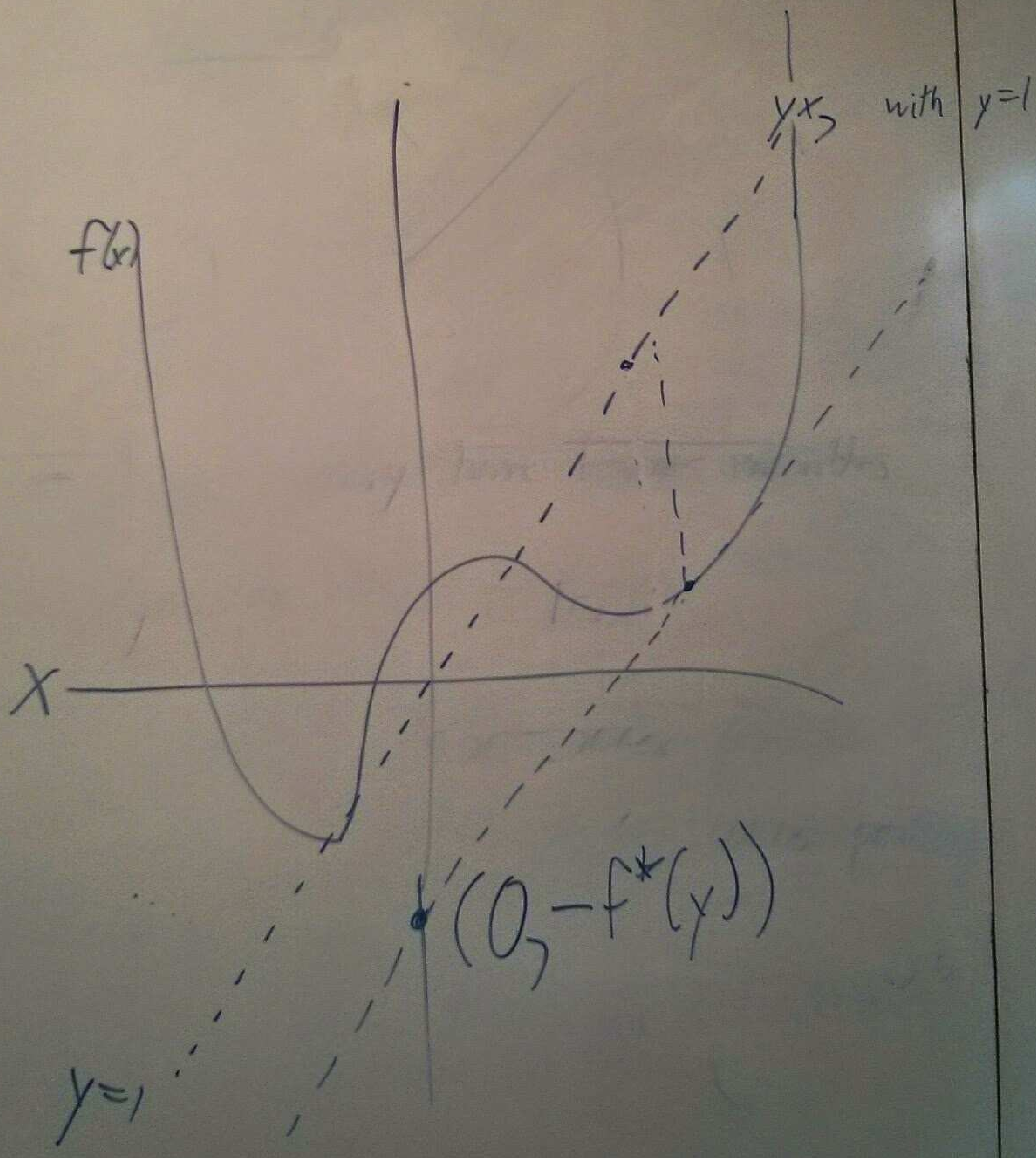
$\inf\{f(x)\} = 0$

Convex Conjugate

$$f^*(y) = \sup_x \{ y^T x - f(x) \}$$

$$= -\inf_x \{ -y^T x + f(x) \}$$

e^x



yx, with y=1
f(x)
X
(0, -f*(y))
y=1

Examples

1. $f(x) = \frac{1}{2}\|x\|^2$

$$f^*(y) = \sup_x \{ y^T x - \frac{1}{2}\|x\|^2 \}$$

(take deriv, set to 0)

$$0 = y - x$$

$$x = y.$$

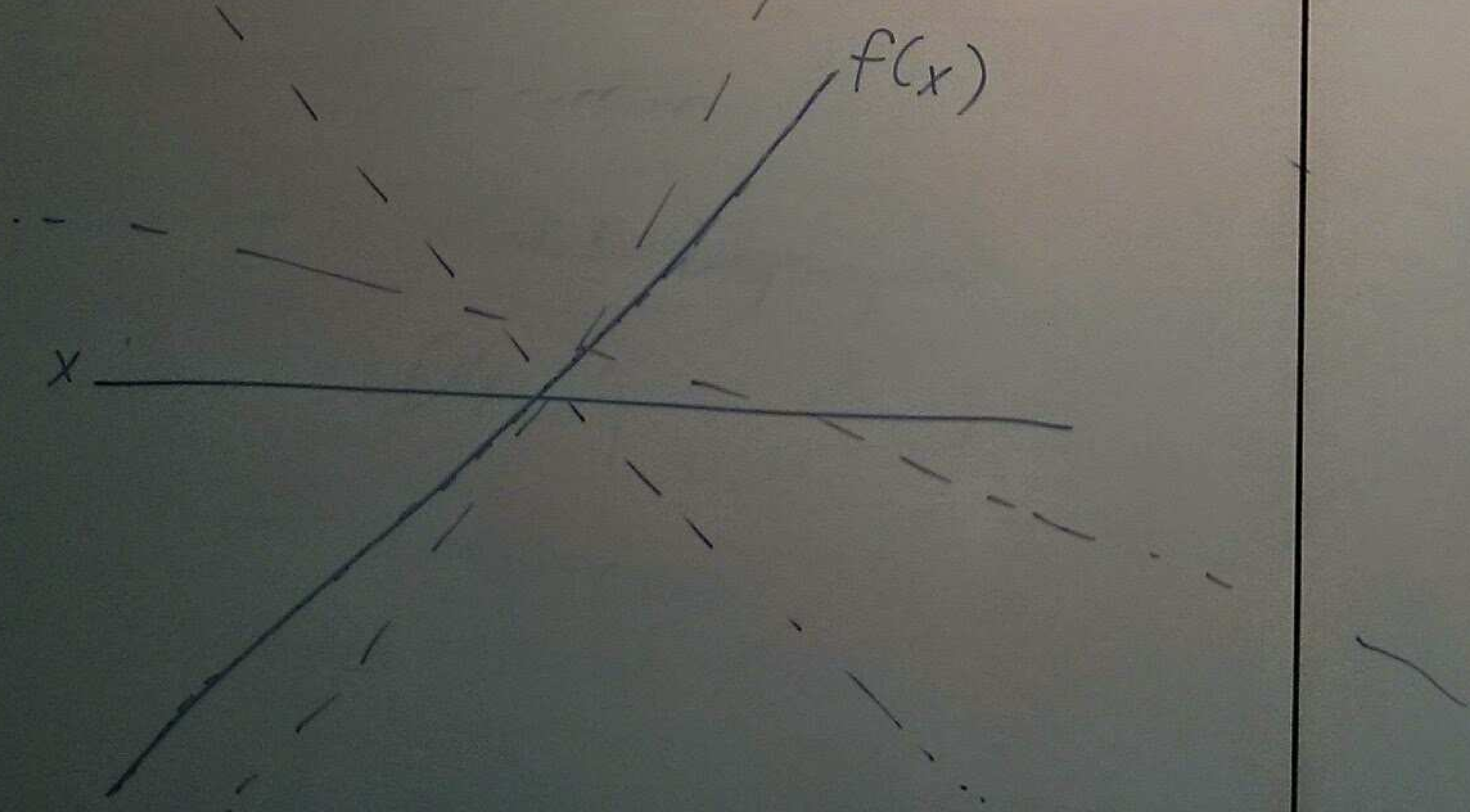
$$f^*(y) = y^T y - \frac{1}{2}\|y\|^2$$

$$= \frac{1}{2}\|y\|^2$$

e Projects
al- Gradient
Duality

$$f(x) = a^T x$$

$$f^*(y) = \sup_x \{ y^T x - a^T x \} = \begin{cases} 0 & \text{if } y = a \\ \infty & \text{otherwise} \end{cases}$$



Project i

- parrallo
- apply
- converg

if $y = a$

0 otherwise

$$P(x) = f(Ax)$$

$$-D(y) = f^*(y)$$

Project ideas

- parrallel/distributed stochastic gradient
- apply SAG/SVRG to a non-linear, non-convex problem.
- convergence rates
- fenchel dual of a non-linear, non-convex problem

* Kernel trick for any L_2-reg. linear model

Fenchel dual w/ L_2-regularization

$$P(x) = f(Ax) + \frac{\lambda}{2}\|x\|^2$$

$$-D(y) = f^*(y) + \underbrace{\frac{1}{2\lambda}\|A^T y\|^2}_{= \frac{1}{2\lambda} y^T (AA^T) y}$$

Gram matrix
(replace this w/
kernel matrix)

⊛ Kernel trick
for <u>any</u> L_2-reg.
linear model

non-linear,
non-convex problem.

non-convex problem