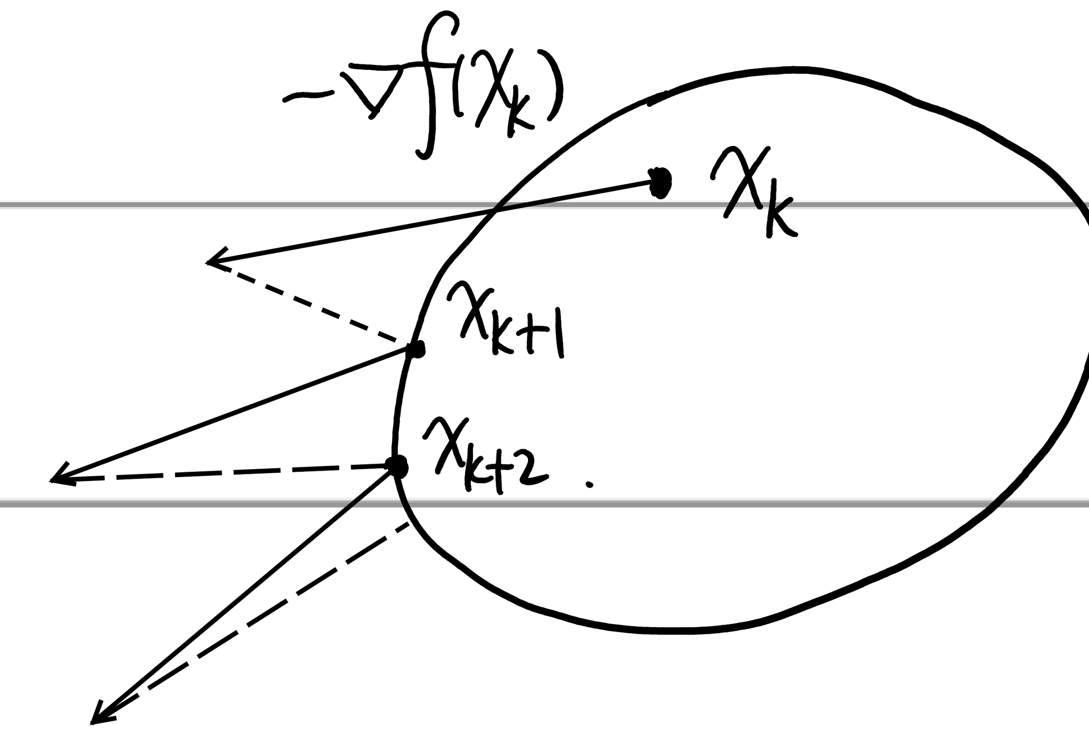
Lecture 22. Projected gradient descent; intro to IPM.

Recall the gradient descent method: $x_{k+1} = x_k + t_k d_k$. $-\nabla f(x_k)$

Now we consider $\min_{x \in X} f(x)$.



If $x_k - t\nabla f(x_k)$ is infeasible.

project it onto X.

Projection operator: $P_X(y) = \arg\min_{x \in X} \|x-y\|$.

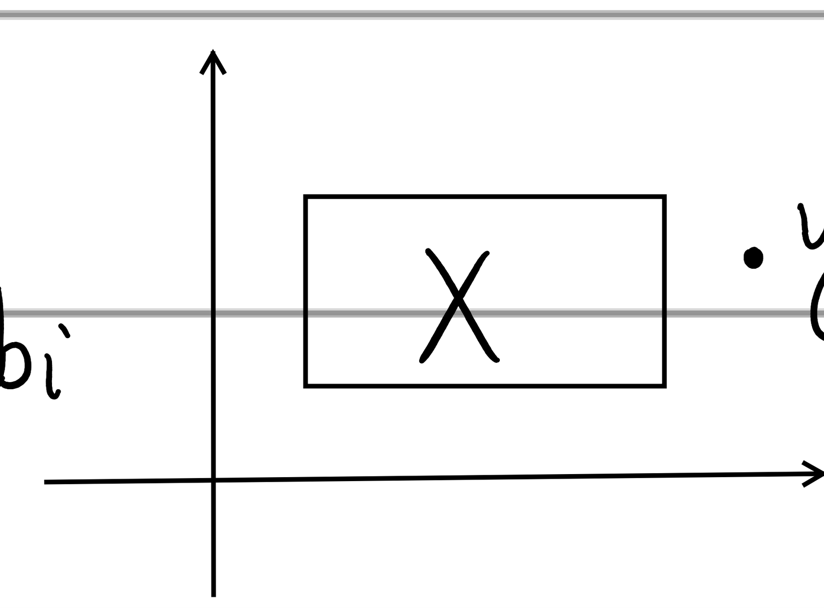
Projected gradient descent: $x_{k+1} = x_k - P_X(x_k - t\nabla f(x_k))$.

In general, projection is an optimization which is hard to solve.

Here we discuss some examples where projection can be efficiently computed.

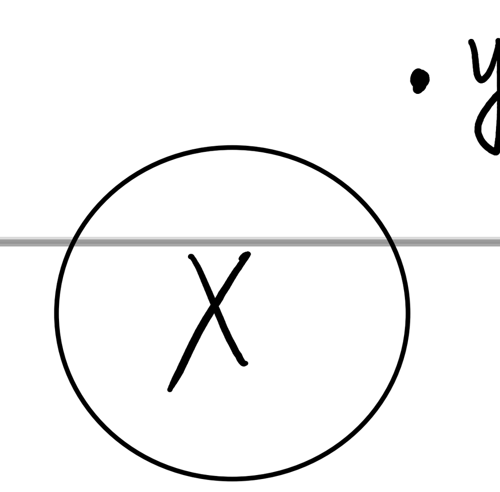
Example. Box constraints. $X = \{x : a_i \le x_i \le b_i,\ i=1,2,\dots,n\}$.

$$(P_X(y))_i = \min\{b_i, \max\{a_i, y_i\}\} = \begin{cases} a_i & y_i < a_i \\ y_i & a_i \le y_i \le b_i \\ b_i & y_i > b_i \end{cases}$$



Example. l_2 constraints. $X = \{x : \|x\|_2 \le t\}$ (ridge regression).

$P_X(y) = \min\{1, \frac{t}{\|y\|_2}\} y$. why?



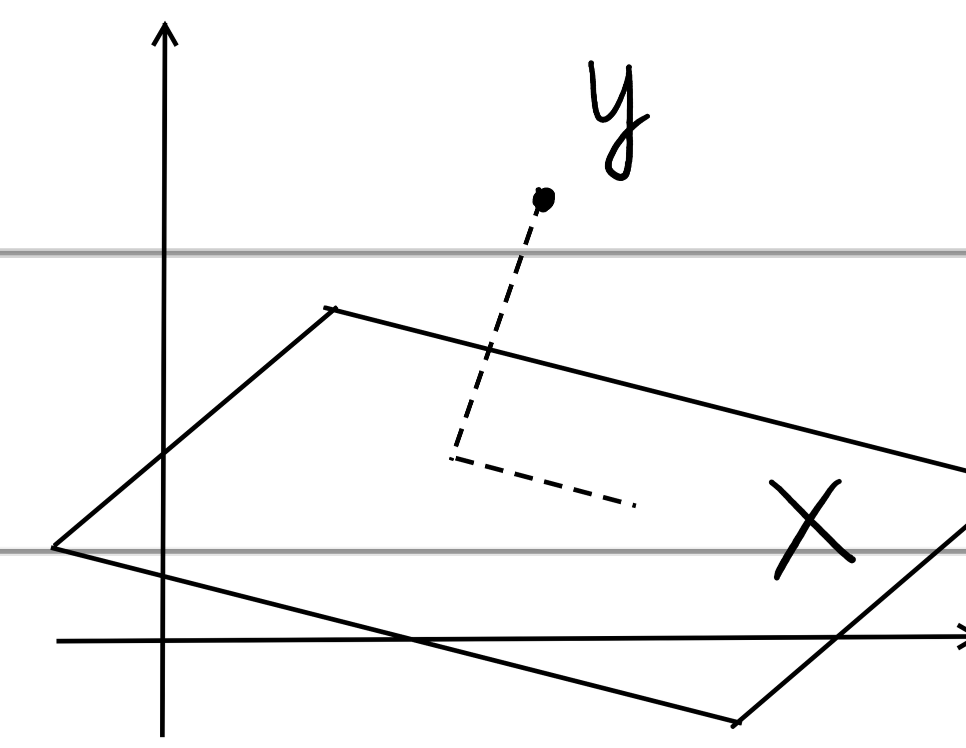
$\min \|x-y\|^2$. s.t. $\|x\|_2^2 \le t^2$. By KKT condition.

$\exists \mu \ge 0$. s.t. $2(x-y) + 2\mu x = 0$. $\Rightarrow y = (1+\mu)x \propto x$.

$\mu(\|x\|^2 - t) = 0$. either $\mu = 0$ or $\|x\|_2 = t$.

Example. Affine constraints. $X = \{x : Ax = b\}$. $A \in \mathbb{R}^{m\times n}$. rank $= m$.

$\min \|x-y\|^2$. s.t. $Ax = b$. By Lagrange



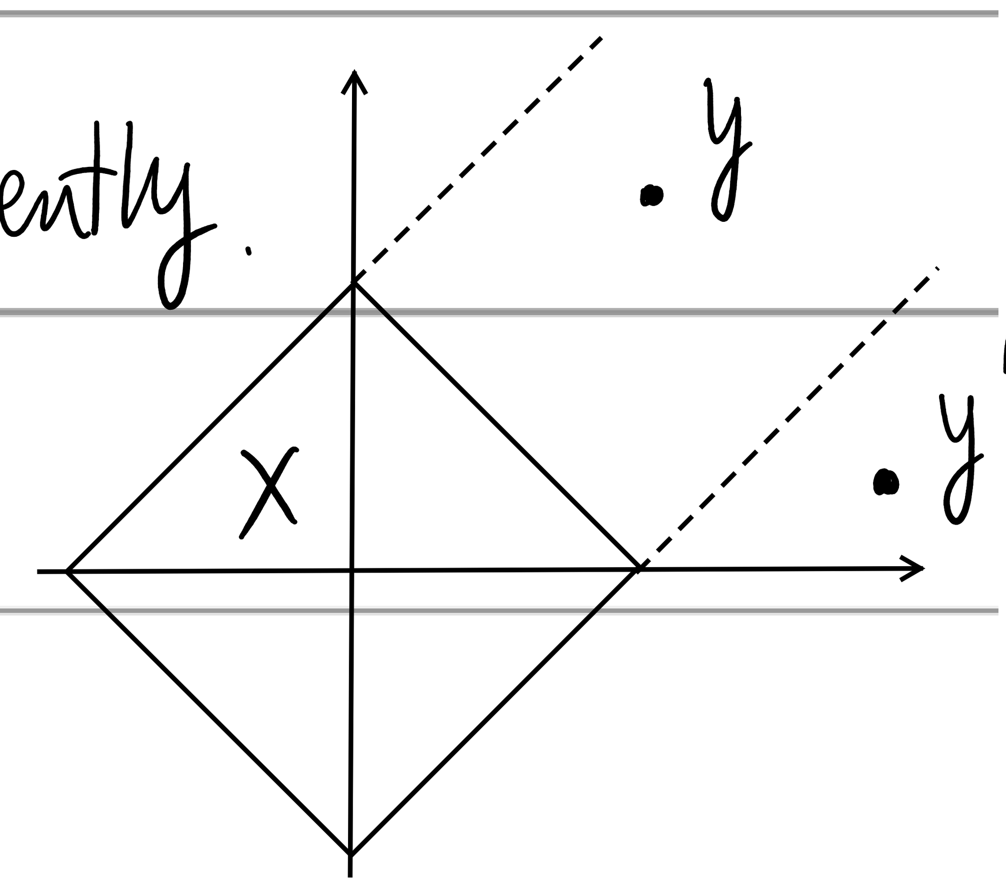
$\exists \lambda$. s.t. $2(x-y) + A^T\lambda = 0 \Rightarrow x = y - \frac{1}{2}A^T\lambda$

Since $Ax = b$. $Ay - \frac{1}{2}AA^T\lambda = b \Rightarrow$

$\lambda = 2(AA^T)^{-1}(Ay-b) \quad \Rightarrow \quad P_X(y) = y - A^T(AA^T)^{-1}(Ay-b)$.

Example: ℓ_1 constraints. $X = \{x : \|x\|_1 \le t\}$ (lasso).

no closed form. but can be computed efficiently.



By symmetry. only need to consider $y \ge 0$.

$\min \|x-y\|^2$. s.t. $\sum x_i \le t$. $x_i \ge 0 \ \forall i$.

By KKT conditions. $\exists \mu_0, \mu_1, \ldots, \mu_n \ge 0$. $\begin{cases} x_i - y_i + \mu_0 - \mu_i = 0. \\ \mu_0(\sum x_i - t) = 0 \\ \mu_i x_i = 0. \end{cases}$

Case 1. $\|y\|_1 \le t \Rightarrow \begin{matrix}\mu_0 = 0 \\ \mu_i = 0.\end{matrix}$ $x = y$.

Case 2. $\|y\|_1 > t. \Rightarrow \mu_0 > 0$. $\quad x_i = y_i - \mu_0 + \mu_i$. $\quad \mu_i x_i = 0$.

if $\mu_i = 0 \Rightarrow x_i = y_i - \mu_0 \ge 0$. if $x_i = 0 \Rightarrow \mu_i = \mu_0 - y_i \ge 0$.

$\Rightarrow x_i = \begin{cases} y_i - \mu_0 & \text{if } y_i \ge \mu_0 \\ 0 & \text{o.w.} \end{cases}$ s.t. $\sum x_i = t$.

w.l.o.g. assume $y_1 \ge y_2 \ge \cdots \ge y_n$. binary search. or.

let $\xi_k = \frac{1}{k}\left(\sum_{i=1}^{k} y_i - t\right)$. $\mu_0 = \xi_{k_0}$. where $k_0 = \min\{k : \xi_k \ge y_{k+1}\}$

assume $y_{n+1} = 0$.

Convergence analysis. Recall proximal gradient descent.

$\min f(x) = g + h$. g convex. differentiable. h convex.

$\text{prox}_{h,t}(x) = \underset{y}{\text{argmin}} \frac{1}{2t}\|x-y\|^2 + h(y)$. $x_{k+1} = \text{prox}(x_k - t\nabla g(x_k))$.

projected: $P_X(x) = \underset{y\in X}{\text{argmin}} \|x-y\|^2$. $x_{k+1} = P_X(x_k - t\nabla f(x_k))$.

let $h(x) = I_X(x) = \begin{cases} 0 & \text{if } x \in X \\ \infty & \text{o.w.} \end{cases}$ $h(x)$ convex. if X convex.

$\text{prox}_{h,t}(x) = \text{prox}_{I_X}(x) = \underset{y\in X}{\text{argmin}} \|x-y\|^2$

$\underset{x\in X}{\min} f(x) = \underset{x}{\min} (f(x) + I_X(x))$. the same as proximal GD.

The convergence result is the same as proximal GD as well.

Theorem: Let X be a nonempty convex set. f is L-smooth and m-strongly convex. (m is allowed to be 0). function over X. Suppose x^* is the optimal solution to $\underset{x\in X}{\min} f(x)$. and $\{x_k\}$ is the sequence produced by projected GD with step size $t \le 1/L$. Then.

① $f(x_{k+1}) \le f(x_k)$

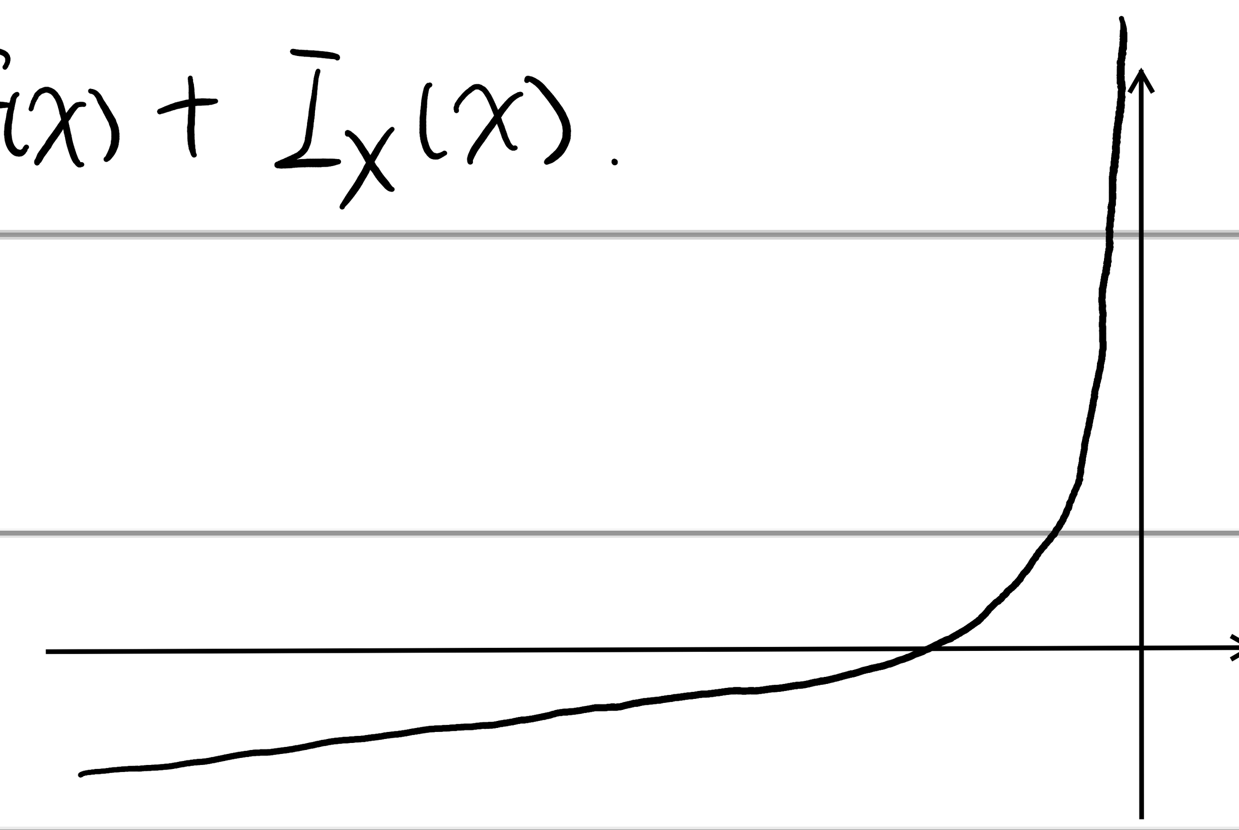
② $\|x_{k+1} - x^*\|_2 \le \|x_k - x^*\|_2$.

③ $f(x_k) - f(x^*) \le \frac{L}{2k}\|x_0 - x^*\|^2$.

④ if $m > 0$. $\|x_k - x^*\|^2 \le (1 - m/L)^k \|x_0 - x^*\|^2$.

Introduction to the interior points method. (IPM).

$\min_{x \in X} f(x)$ is equivalent to $\min_{x} f(x) + I_X(x)$.



X is given by $\{x : h(x) \le 0\}$.

define $I_t(x) = -\frac{1}{t} \log(-h(x))$.

$I_t(x)$ is called the log-barrier. I_t convex. and $\lim_{t\to\infty} I_t \to I_X$.

$x_t^* = \operatorname{argmin} f(x) - \frac{1}{t} \Sigma \log(-h_i(x))$ let $t \uparrow$ continuously.

$= \operatorname{argmin} t f(x) - \Sigma \log(-h_i(x))$. $\{x_t^*\}$: central path.

$\lim_{t\to\infty} x_t^* \to x^*$. solved by Newton. x_t^* as intial point of next t.

IPM is the first algorithm efficient in both theory and practice to solve LP. (Karmarkar's algorithm in 1984).