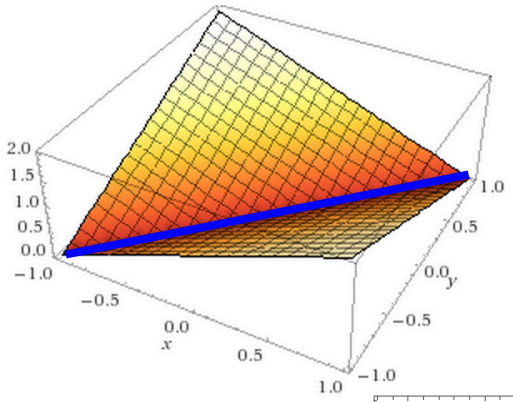
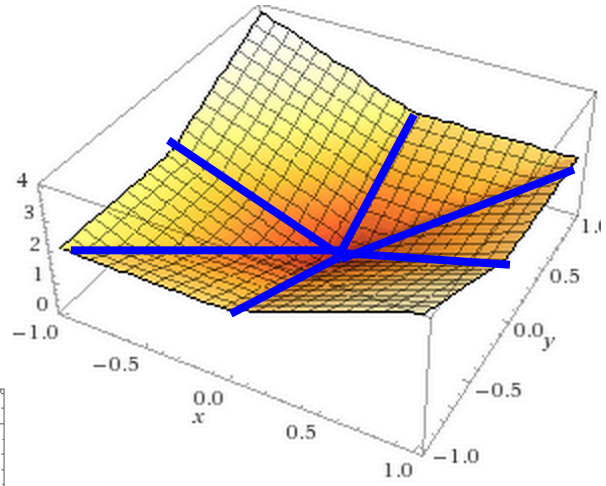
CMSC 510 – L11
Regularization Methods for Machine Learning

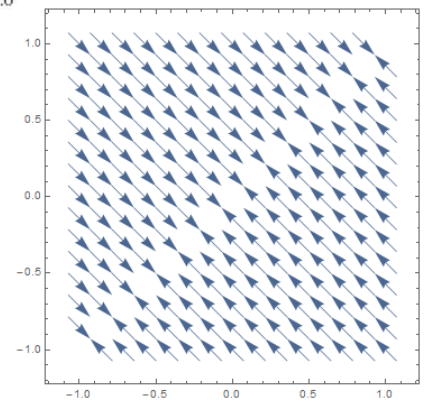
Instructor:
Dr. Tom Arodz

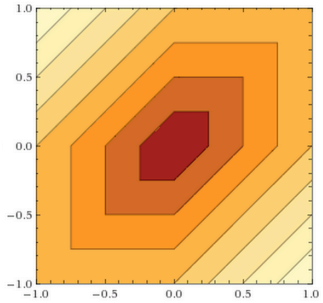
Recap: L_1 / Non-smooth optimization

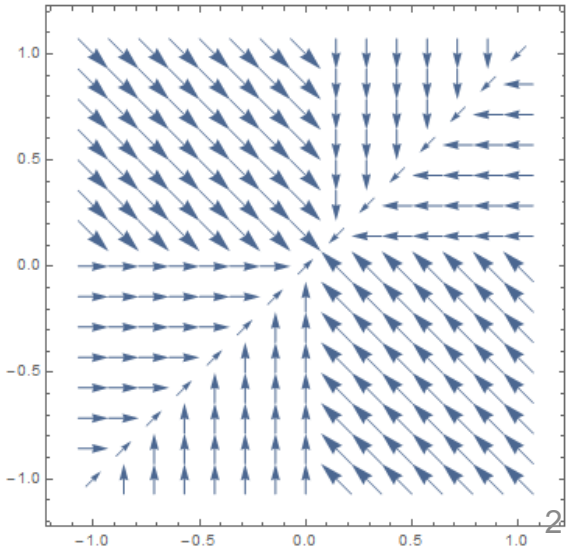
- L_1 regularization leads to a non-differentiable objective function
- For non-smooth functions, there will be points for which we can't calculate gradient (some partial derivatives don't exist)





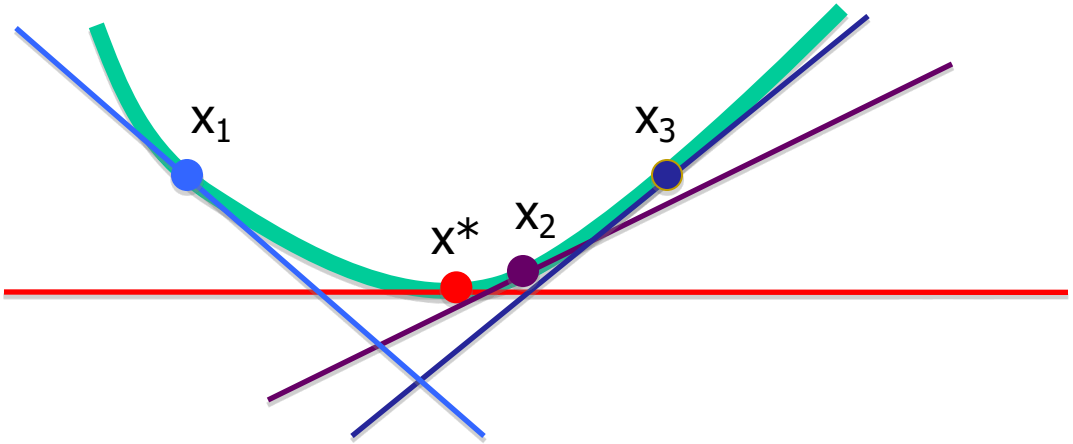






Recap: Convex optimization

For all **convex** functions f:



f is **convex** if all its values f(y) are (weakly) above any gradient-based linear approximation of f

For all x,y:

$f(y) >= f(x) + <\nabla f(x), y-x>$

For all convex functions f:

if $\nabla f(x^*) = 0$, then x^* is global optimum

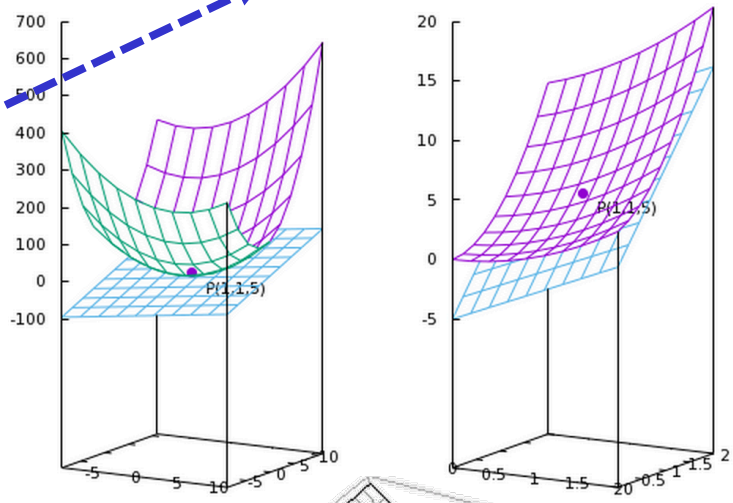
Recap: Subgradient

If f(x) is convex, the tangent plane is always below f:

$$\forall x' \in \mathbb{R}^n f(x') \geq f(x) + \langle \nabla f(x), x' - x \rangle$$

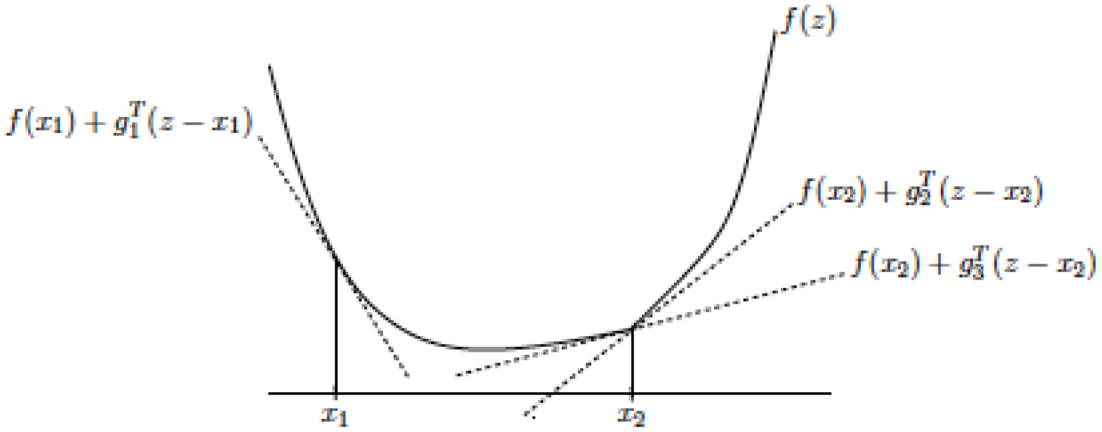
Subgradient of f() at x:
any vector **g** that satisfies:

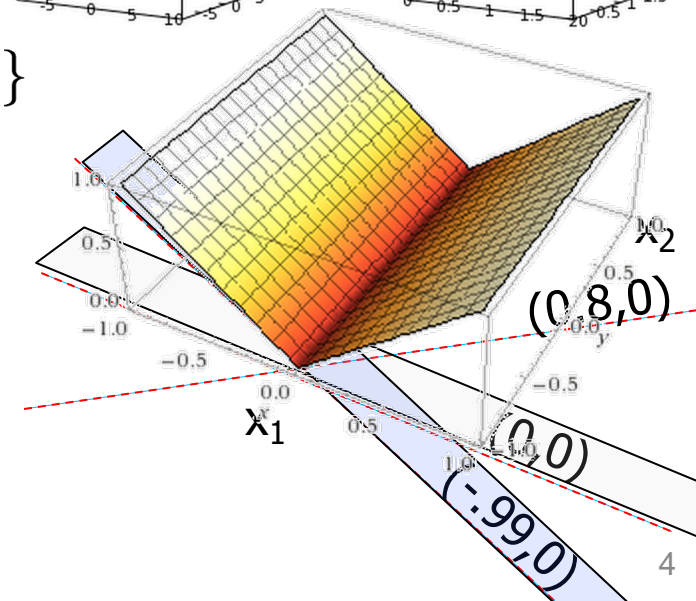
$$\forall x' \in \mathbb{R}^n \quad f(x') \geq f(x) + \langle g, x' - x \rangle$$



Set of subgradients at x is called a **subdifferential** of f() at x

$$\partial f(x) = \{g : \forall x' \in \mathbb{R}^n \quad f(x') \geq f(x) + \langle g, x' - x \rangle\}$$





Condition for Global minimum

- Convex f: differentiable risk + (non-) differentiable penalty
- Necessary and sufficient condition for global minimum for continuous, convex f:

$$f(w) = \hat{R}_{S_m}(w) + \Omega(w)$$

$$\partial f(w) = \left\{ \nabla_w \hat{R}_{S_m}(w) \right\} + \partial\Omega(w)$$

$$w^* = \arg\min_w f(w) \Leftrightarrow 0 \in \partial f(w^*)$$

$$0 \in \left\{ \nabla_w \hat{R}_{S_m}(w^*) \right\} + \partial\Omega(w^*)$$

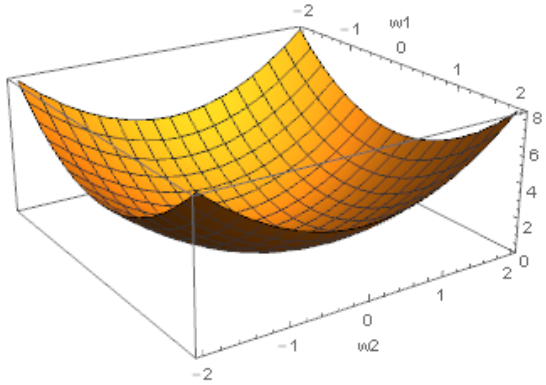
$$-\nabla_w \hat{R}_{S_m}(w^*) \in \partial\Omega(w^*)$$

algebraic (Minkowski) sum of sets :

$$A + B = \{a + b : a \in A, b \in B\} \qquad A + \emptyset = \emptyset$$

Condition for Global minimum

- L_2 norm (differentiable)



$$\Omega(w) = ||w||_2^2 = \sum_{i=1}^{n} w_i^2$$

$$-\nabla_w \hat{R}_{S_m}(w^*) = \nabla_w \Omega(w^*)$$

$$-\frac{1}{2}\nabla_w \hat{R}_{S_m}(w^*) = w^*$$

$$-\frac{1}{2}\frac{\partial \hat{R}_{S_m}(w^*)}{\partial w_i} = w_i^*$$

$$-\frac{1}{2}\sum_{j=1}^{m}\frac{\partial \log(1+\exp(-y_j {w^*}^T x_j))}{\partial w_i} = w_i^*$$

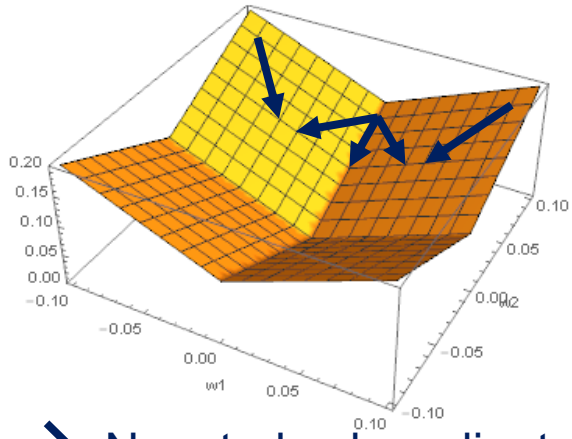
$$w_i^* = 0 \text{ is part of minimum } w^* \text{ if } \frac{\partial \hat{R}_{S_m}(w^*)}{\partial w_i} = 0$$

- Gradient of risk has to exactly cancel gradient of $||w||^2$
  - Unlikely to happen at $w_i=0$
- Built-in feature selection – very unlikely

Subdifferential for L_1

- L_1 norm (2D):

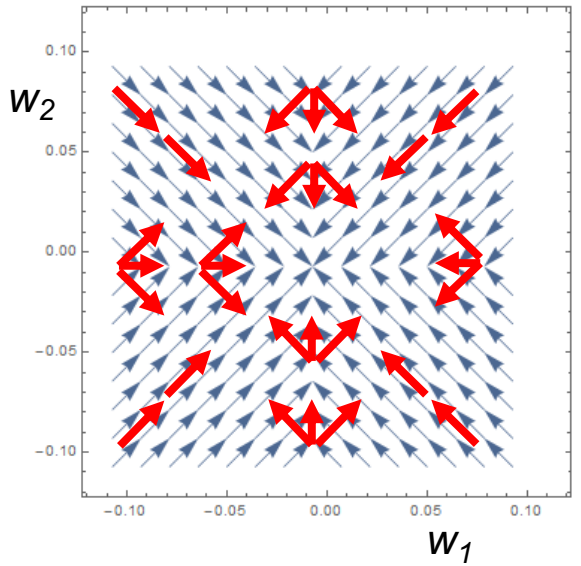
$$f(x) = |x_1| + |x_2|$$



↘ Negated subgradient

$$\partial f(x) = G_1 \times G_2$$

$$G_i(x) = \begin{cases} [-1, 1], & \text{if } x_i = 0 \\ \{\text{sign}(x_i)\}, & \text{if } x_i \neq 0 \end{cases}$$



↘ Negated subgradient

- L_1 norm (n-D):

$$f : \mathbb{R}^n \to \mathbb{R}$$

$$f(x) = \sum_{i=1}^{n} |x_i|$$

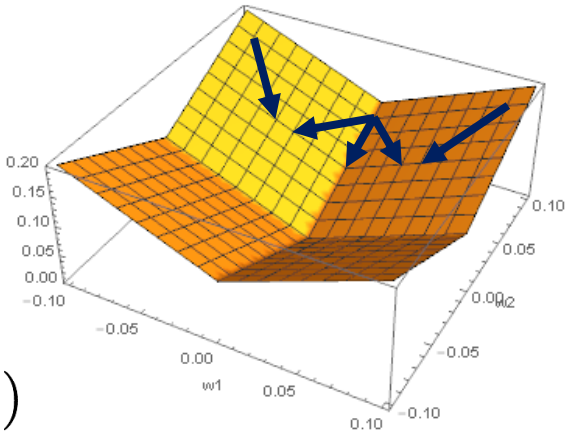
$$\partial f(x) = G_1(x) \times G_2(x) \times \ldots \times G_n(x)$$

Condition for Global minimum

- L_1 penalty (non-differentiable)

$$\Omega(w) = \lambda||w||_1 = \lambda \sum_{i=1}^{n} |w_i|$$

Larger λ = larger penalty for same $||w||_1$



$$-\nabla_w \hat{R}_{S_m}(w^*) \in \partial\Omega(w^*)$$

$$-\nabla_w \hat{R}_{S_m}(w^*) \in \lambda(G_1(w^*) \times G_2(w^*) \times \ldots \times G_n(w^*))$$

$$-\frac{\partial \hat{R}_{S_m}(w^*)}{\partial w_i} \in \lambda G_i(w^*)$$

$$G_i(x) = \begin{cases} [-1,1], & \text{if } x_i = 0 \\ \{\text{sign}(x_i)\}, & \text{if } x_i \neq 0 \end{cases}$$

$$\frac{\partial \hat{R}_{S_m}(w^*)}{\partial w_i} \in \begin{cases} [-\lambda,\lambda], & \text{if } w_i^* = 0 \\ \{-\lambda \text{sign}(w_i^*)\}, & \text{if } w_i^* \neq 0 \end{cases}$$

$$w_i^* = 0 \text{ is part of minimum } w^* \text{ if } \left|\frac{\partial \hat{R}_{S_m}(w^*)}{\partial w_i}\right| \leq \lambda$$

- Feature selection – no longer so unlikely!
  - Gradient of risk has a range to fall into
  - Larger λ=>wider range λG_i => $w_i^*=0$ more likely

Condition for Global minimum

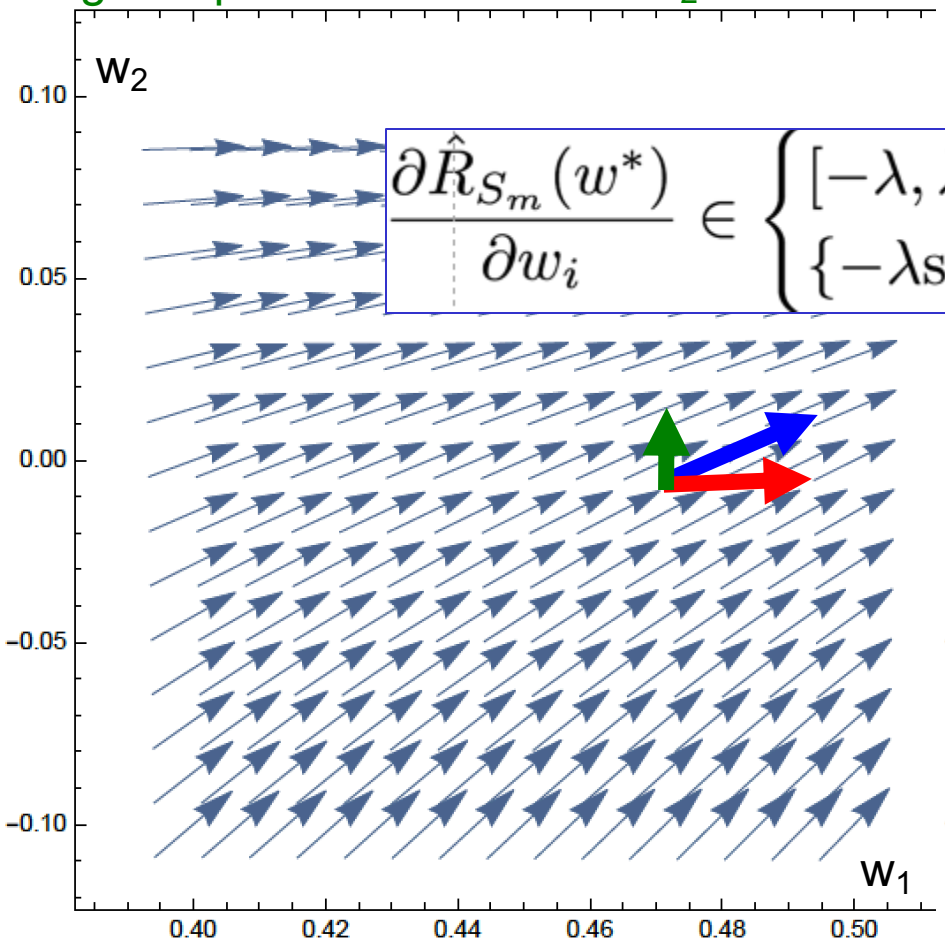
Risk + L_1 penalty: gradient plots

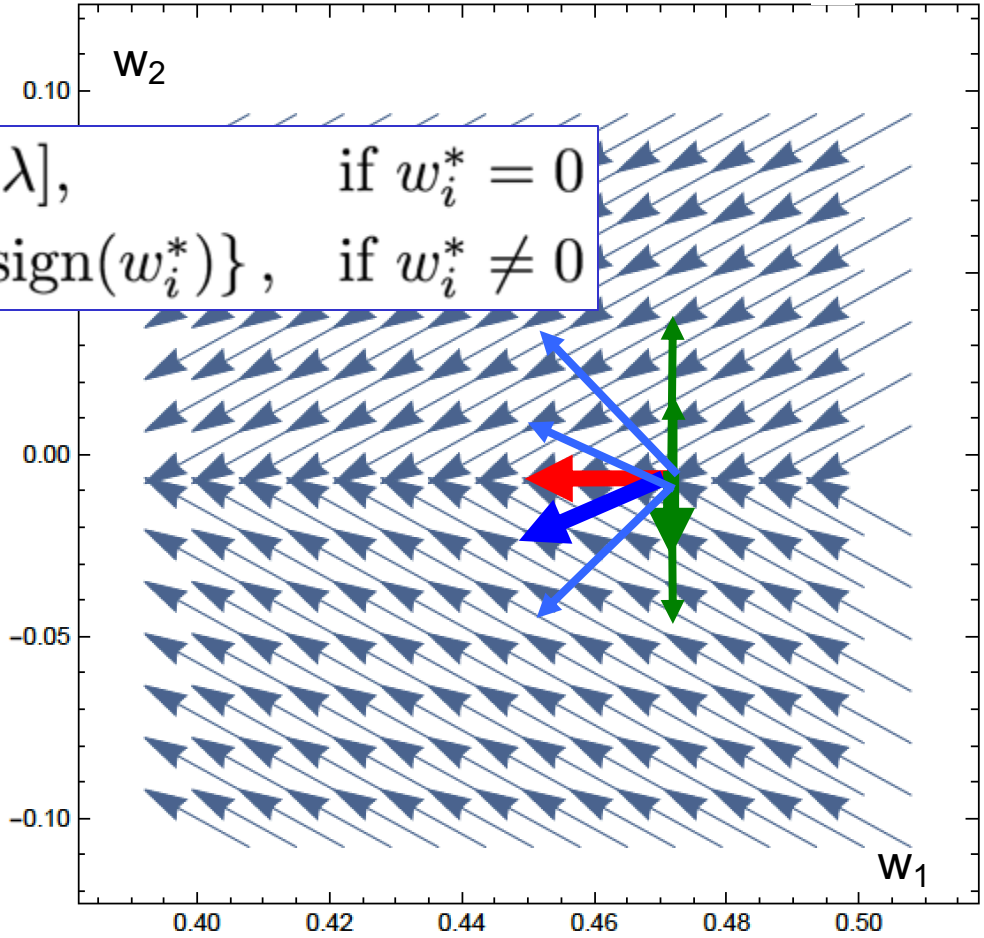
Minimum:
w_1=0.465, w_2=0

Empirical risk at w_1=0.46, w_2=0:
Negated gradient (one)
Negated partial derivative for w_1
Negated partial derivative for w_2

L_1 penalty at w_1=0.46, w_2=0:
Subgradients (more then one)
Negated w_1 coordinate of subgradients = -λ
Negated w_2 coordinate of subgradients ∈ [-λ, λ]

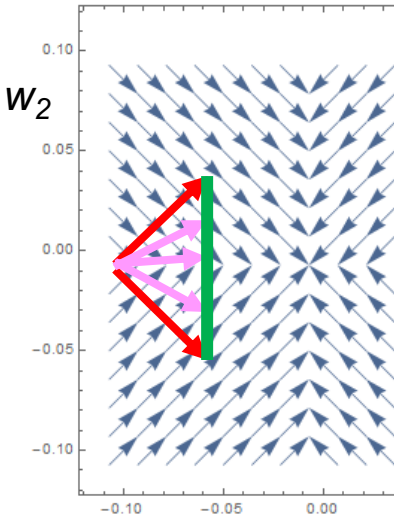
$$\frac{\partial \hat{R}_{S_m}(w^*)}{\partial w_i} \in \begin{cases} [-\lambda, \lambda], & \text{if } w_i^* = 0 \\ \{-\lambda \text{sign}(w_i^*)\}, & \text{if } w_i^* \neq 0 \end{cases}$$





Clarke subdifferential

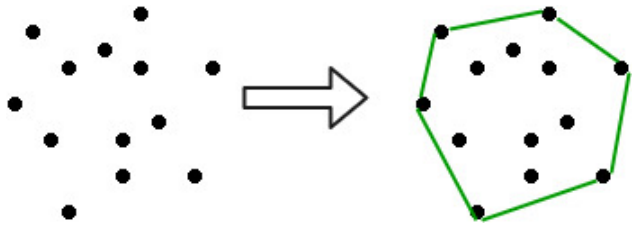
- Subgradient / subdifferential above was defined for convex functions
- What about non-convex functions?
  - E.g. f(x)=-|x|
    - It seems it's not really that different from f(x)=|x|
- Clarke's generalized gradient



Definition 1 (Clarke, 1975 [8]). *The* generalized gradient $\partial f(x)$ *of a locally Lipschitz function* f *at* x *is defined as*

$$\partial f(x) = \text{conv}\left\{\lim_{x_k \to x} \nabla f(x_k)\right\},$$

where the limit is over all convergent sequences of x_k *where gradient exists, and* conv *denotes convex hull.*



Big picture

- loss/risk is differentiable and convex,
  penalty terms may not be differentiable (although still convex)
  - loss function/risk tells us if a classifier matches the training data well
  - penalty terms reflect our prior knowledge about what a good classifier would look like
    - Training data is limited+noisy, prior knowledge may help
- **How to find global minimum of risk + penalty?**
- **Generic (slow) technique: method of subgradients**
- But we know that our risk+penalty optimization problems have a particular structure:
  - **Objective function is a sum of a differentiable term and a non-differentiable term**
  - **Can we have a faster method designed for such objective functions?**

Convex optimization

Function f has **Lipschitz continuous gradient** with constant L (f is in class of functions $F_L^{1,1}$) if

$| \nabla f(x) - \nabla f(y) | <= L || x-y ||$ for all x,y

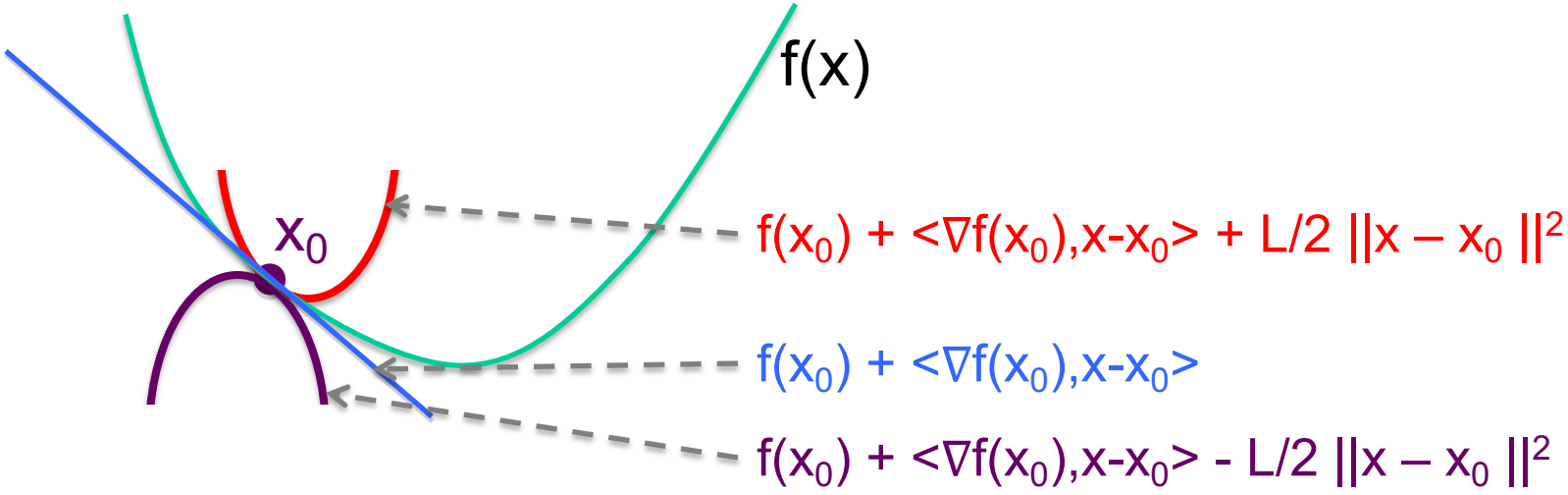
Then (not only for convex functions), for any x,y:

$$| f(y) - f(x) + \langle \nabla f(x), y-x \rangle | <= L/2 \, ||y - x||^2$$

That means:

$$f(x) <= f(x_0) + \langle \nabla f(x_0), x-x_0 \rangle + L/2 \, ||x - x_0||^2$$ ← Descent lemma

$$f(x) >= f(x_0) + \langle \nabla f(x_0), x-x_0 \rangle - L/2 \, ||x - x_0||^2$$



Descent lemma: 1D intuition

For any specific z: based on f(x), we can create function φ(y) by: shifting horizontally by –z, vertically by f(z), and subtracting a linear term $\langle \nabla f(x), z-x \rangle$, and we have: $z=0,\ \varphi(0)=0,\ \nabla\varphi(0)=0$

if: $\| \nabla f(x) - \nabla f(z) \| \le L \| x-z \|$ for all x

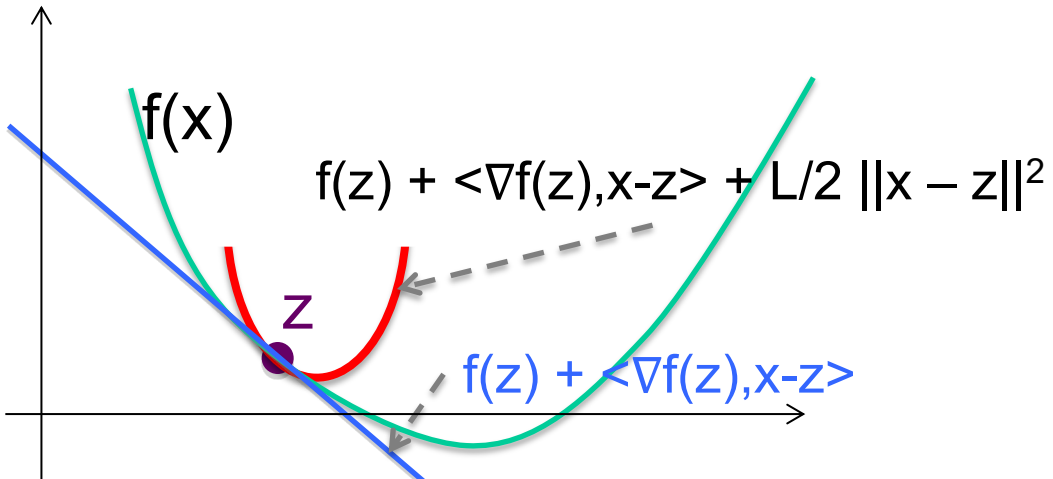
then: $f(x) \le f(z) + \langle \nabla f(z), x-z \rangle + L/2 \|x - z\|^2$

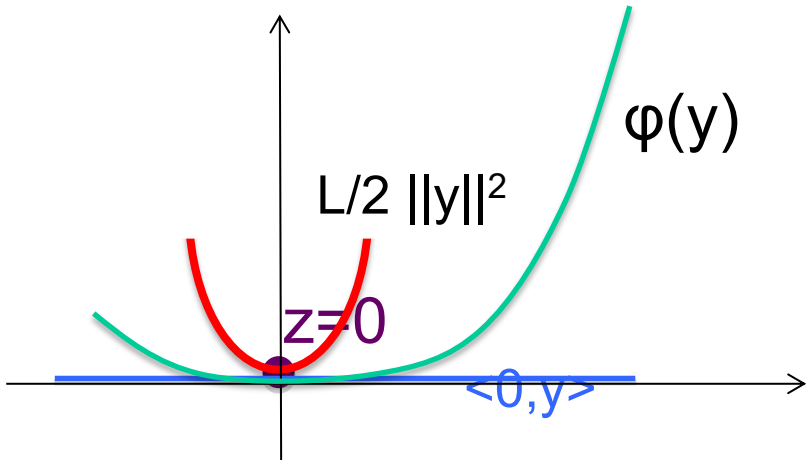
becomes:

if $\| \nabla\varphi(y) \| \le L \| y \|$ for all y then $\varphi(y) \le L/2 \|y\|^2$

In 1D:

$\varphi(y) = \int_0^y (d\varphi(x)/dx)\, dx \le \int_0^y |d\varphi(x)/dx|\, dx \le \int_0^y L|x|dx = L/2\, y^2$





Descent lemma: proof (skip)

$$f(x) \leq f(z) + < \nabla f(z), x - z > + \frac{L}{2}||x-z||^2$$

$$f(z+\delta) - f(z) \leq < \nabla f(z), \delta > + \frac{L}{2}||\delta||^2$$

$$f(z+\delta) - f(z) = g(1) - g(0)$$

$$x = z + \delta$$
$$g(\alpha) = f(z + \alpha\delta)$$

1: $\int h = \int k + \int h - \int k = \int k + \int (h-k) \leq \int k + |\int (h-k)|$

2: $|\langle a,b \rangle| = |a^T b| \leq ||a||\,||b||$ (Cauchy-Schwarz)

3: $|\int h| \leq \int |h|$

4: $||\nabla f(x) - \nabla f(z)|| \leq L\,||x-z||$ for all x,z (L-Lipschitz ∇ assumption)

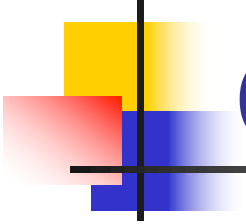
$$g(1) - g(0) = \int_0^1 \frac{\partial g(\alpha)}{\partial \alpha} d\alpha = \int_0^1 \frac{\partial f(z+\alpha\delta)}{\partial \alpha} d\alpha = \int_0^1 \delta^T \nabla f(z + \alpha\delta) d\alpha$$

(from 1) $$\leq \int_0^1 \delta^T \nabla f(z) d\alpha + \left| \int_0^1 \delta^T \left[\nabla f\left(z + \alpha\delta\right) - \nabla f\left(z\right) \right] d\alpha \right|$$

(from 2,3) $$\leq \delta^T \nabla f(z) \int_0^1 d\alpha + \int_0^1 ||\delta|| \cdot ||\nabla f\left(z + \alpha\delta\right) - \nabla f\left(z\right)|| d\alpha$$

(from 4) $$\leq \delta^T \nabla f(z) + \int_0^1 ||\delta|| L\alpha ||\delta|| d\alpha = \delta^T \nabla f(z) + L||\delta||^2 \int_0^1 \alpha d\alpha$$

$$\leq \; < \nabla f(z), \delta > + \frac{L}{2}||\delta||^2 = \; < \nabla f(z), x - z > + \frac{L}{2}||x-z||^2$$



Convex optimization

How can descent lemma help us find the global minimum of a convex function?

Majorization – minimization strategy

Convex optimization

Majorization – minimization strategy

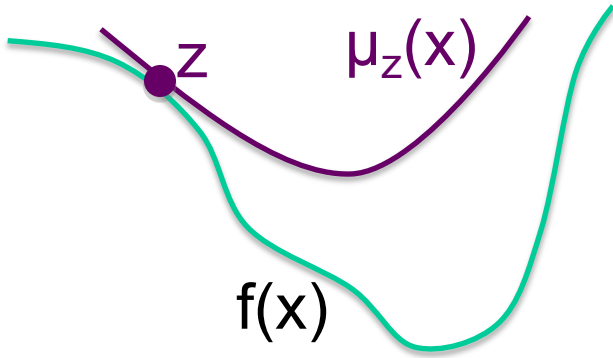
Instead of directly minimizing function f(x)

We design a family of "easier" functions μ_z such that:

$f(x) \leq \mu_z(x)$ for all x

$f(z) = \mu_z(z)$

μ_z is said to majorize function f(x) at z



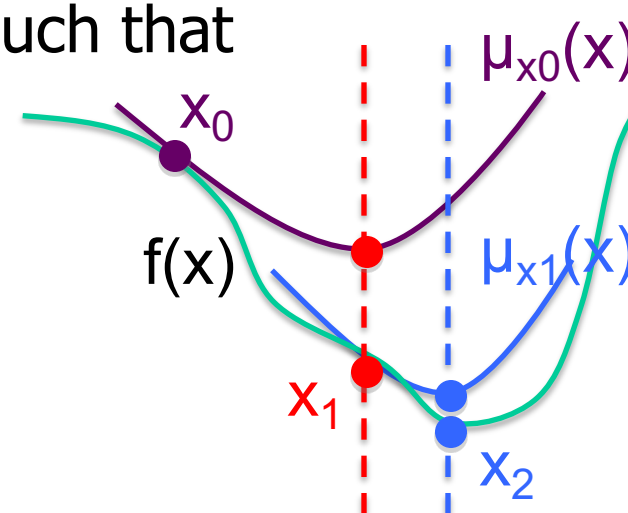
Iterative **majorization-minimization (MM)** procedure constructs a sequence $\{x_n\}$ such that $f(x_{n+1}) \leq f(x_n)$:

We start with arbitrary x_0

We construct $\mu_{x0}(x)$ and find its minimum x_1

We construct $\mu_{xn}(x)$ and find its minimum x_{n+1}

We can show that $f(x_{n+1}) \leq f(x_n)$



Convex optimization

Majorization – minimization strategy

We design a family of functions μ_z such that:

$f(x) \leq \mu_z(x)$ for all x (1)

$f(z) = \mu_z(z)$ (2)

Let:

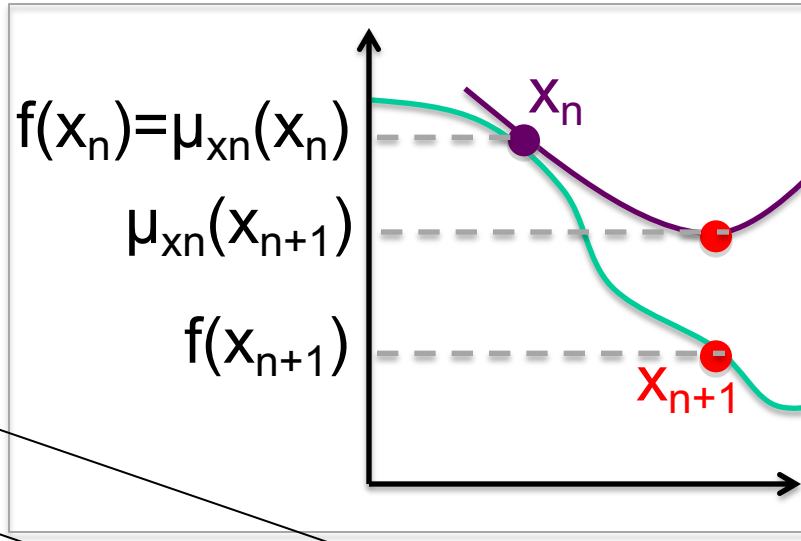
$x_{n+1} = \text{argmin}_x \mu_{xn}(x)$ (3)

Why $f(x_{n+1}) \leq f(x_n)$?

$f(x_{n+1}) \leq \mu_{xn}(x_{n+1})$ by (1)

$\leq \mu_{xn}(x_n)$ by (3)

$= f(x_n)$ by (2)



$\mu_{xn}()$ majorizes f

x_{n+1} was min of $\mu_{xn}()$

$\mu_{xn}()$ equals to f on x_n

Convex optimization

Majorization – minimization strategy

We design a family of functions μ_z such that:

$f(x) \leq \mu_z(x)$ for all x (1)

$f(z) = \mu_z(z)$ (2)

Let:

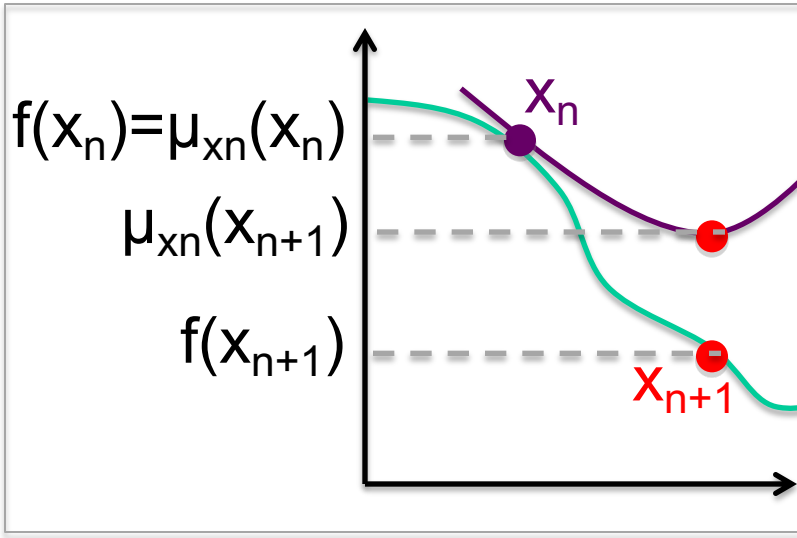
$x_{n+1} = \text{argmin}_x \, \mu_{xn}(x)$ (3)

Why $f(x_{n+1}) \leq f(x_n)$?

$f(x_{n+1}) \leq \mu_{xn}(x_{n+1})$ by (1)

$\leq \mu_{xn}(x_n)$ by (3)

$= f(x_n)$ by (2)



Consequence: we get a sequence of solution x_n going towards local minimum (or global minimum if each minimum is global)

Convex optimization

Majorization – minimization strategy

We design a family of functions μ_z such that:

$f(x) \leq \mu_z(x)$ for all x (1)

$f(z) = \mu_z(z)$ (2)

$x_{n+1} = \text{argmin}_x \, \mu_{xn}(x)$ (3)

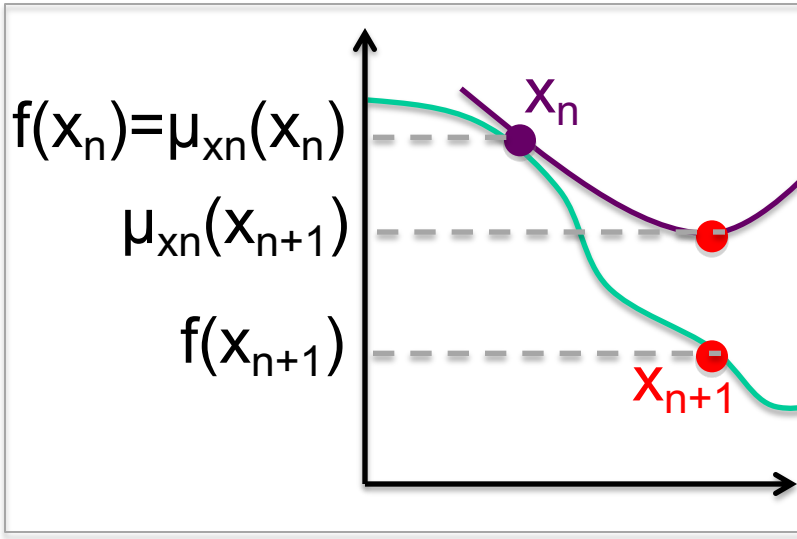
We have $f(x_{n+1}) \leq f(x_n)$

When $f(x_{n+1}) = f(x_n)$?

Only if:

$f(x_{n+1}) = \mu_{xn}(x_{n+1}) = \mu_{xn}(x_n)$

We don't move down only if x_n is already a minimum of $\mu_{xn}(x)$



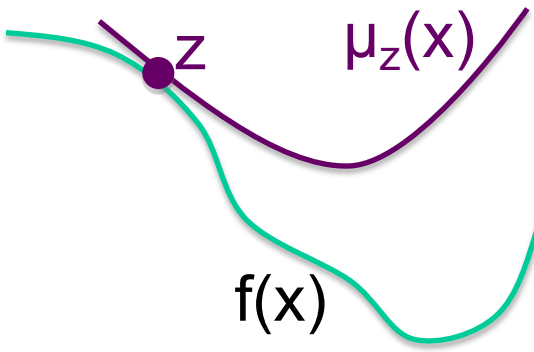
Convex optimization

<x,z+y>=<x,z>+<x,y>
<ax,by>=ab<x,y>
<x,y>=<y,x>
<x , x> = ||x||^2

Majorization – minimization using gradient

For any f with L-Lipschitz gradient, we have:

$f(x) <= f(z) + <\nabla f(z), x-z> + L/2 \, ||x - z||^2$



Let $\mu_z(x) = f(z) + <\nabla f(z), x-z> + L/2 \, ||x - z||^2$

then $f(z) = \mu_z(z)$ and $f(x) \leq \mu_z(x)$ for all x

that is, **μ_z majorizes f(x) at z**

Is minimum of $\mu_z(x)$ easy to find?

$\mu_z(x)$ is: a constant + linear f. of x + quadratic f. of x

μ_z is convex, $\nabla\mu_z(x) = 0$ is sufficient condition for global minimum

$0 = \nabla\mu_z(x) = \nabla f(z) + \nabla<\nabla f(z), x> + \nabla<\nabla f(z), -z> + \nabla L/2 \, ||x - z||^2$

$0 = \nabla<\nabla f(z), x> + L/2 \, \nabla<x-z, x-z>$

$0 = \nabla f(z) + L/2 \, \nabla(<x,x>+<x,-z>+<-z,x>+<z,z>)$

$-\nabla f(z)/L = 1/2 \, (\nabla<x,x> - 2\nabla<x,z>)$

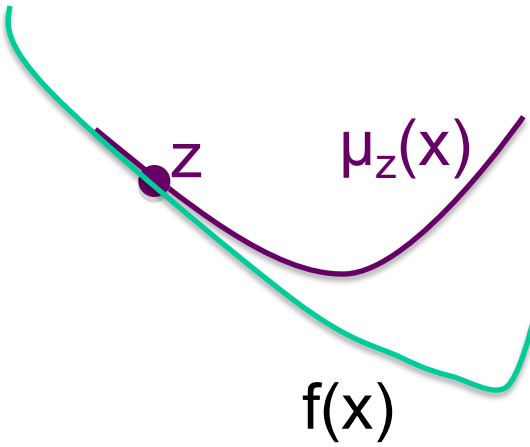
$-\nabla f(z)/L = 1/2 \, (2x - 2z)$

$x = z - \nabla f(z)/L$ - we got x that is the minimum of $\mu_z(x)$

Gradient descent

For any f with L-Lipschitz gradient:

$$f(x) \le f(z) + \langle \nabla f(z), x-z \rangle + L/2 \, \|x - z\|^2$$



Let $\mu_z(x) = f(z) + \langle \nabla f(z), x-z \rangle + L/2 \, \|x - z\|^2$

μ_z majorizes f(x) at z

Minimum of $\mu_z(x)$ is $x=z - \nabla f(z)/L$

Let's apply the MM strategy using $\mu_z(x)$:

We start from x_0

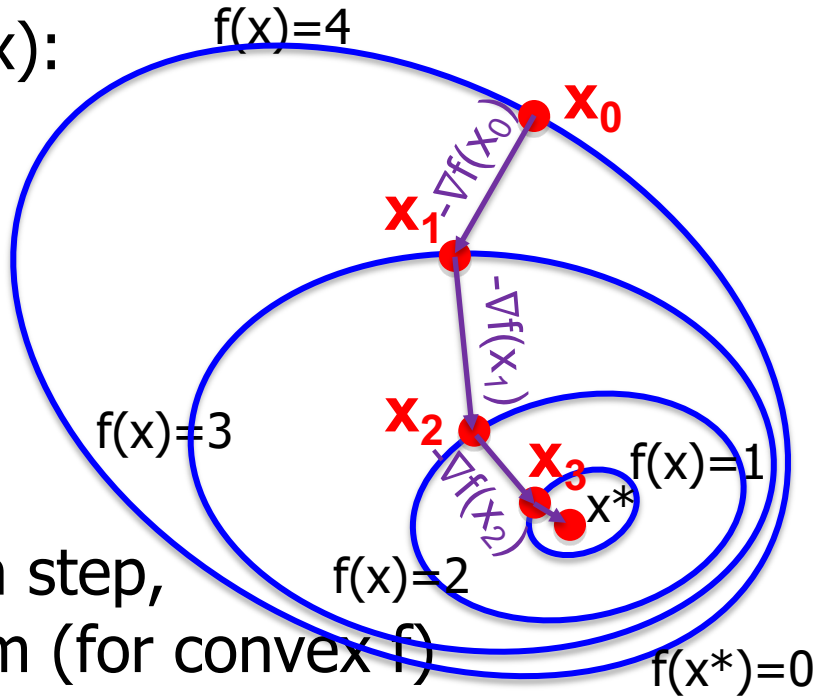
We calculate $x_1=x_0 - \nabla f(x_0)/L$

We calculate $x_2=x_1 - \nabla f(x_1)/L$

$x_{n+1}= x_n - \nabla f(x_n)/L$

We derived **gradient descent**!

We have a proof it goes down in each step, converging towards global minimum (for convex f)



Gradient descent

$\langle x,z+y\rangle=\langle x,z\rangle+\langle x,y\rangle$
$\langle ax,by\rangle=ab\langle x,y\rangle$
$\langle x,y\rangle=\langle y,x\rangle$
$\langle x , x\rangle = ||x||^2$

Let: $\mu_z(x) = f(z) + \langle \nabla f(z), x-z\rangle + L/2\ ||x - z\ ||^2$

Then: $\mu_z(x) = f(z) + L/2\ ||x - [z - \nabla f(z)/L]||^2 - 1/2L\ ||\nabla f(z)||^2$

Red is just another form of blue (let's denote $\nabla_z=\nabla f(z)$):

$\mu_z(x)= f(z) + L/2\ ||\ x - [z - \nabla_z/L]\ \ ||^2 - 1/2L\ ||\ \nabla_z\ ||^2$

$= f(z) + L/2\ ||(x - z) + \nabla_z/L]||^2 - 1/2L\ ||\ \nabla_z\ ||^2$

$= f(z) + L/2\ \langle (x - z) + \nabla_z/L, (x - z) + \nabla_z/L\rangle - 1/2L\ \langle \nabla_z, \nabla_z\rangle$

$= f(z) + L/2\ \{\ \langle x - z, x - z\rangle + 2\langle \nabla_z/L,(x - z)\rangle + \langle \nabla_z/L, \nabla_z/L\rangle\ \} - L/2\ \langle \nabla_z/L,\nabla/L\rangle$

$= f(z) + L/2\ \langle x - z, x - z\rangle + \langle \nabla_z, (x - z)\rangle + L/2\langle \nabla_z/L, \nabla_z/L\rangle - L/2\ \langle \nabla_z/L,\nabla_z/L\rangle$

$= f(z) + L/2\ \langle x - z, x - z\rangle + \langle \nabla_z, (x - z)\rangle$

$= f(z) + L/2\ ||x - z\ ||^2 + \langle\ \nabla_z, (x - z)\rangle$

$= f(z) + \langle \nabla f(z), (x - z)\rangle + L/2\ ||x - z\ ||^2 = \mu_z(x)$

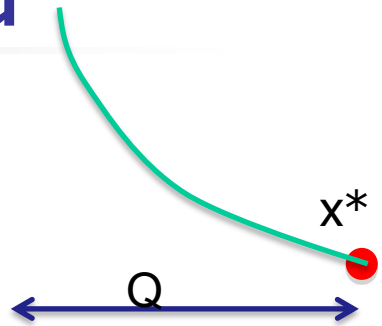
Now it's even easier to see that $x=z - \nabla f(z)/L$ is the minimum of $\mu_z(x)$

$\mu_z(x) = f(z) + L/2\ ||x - [z - \nabla f(z)/L]||^2 - 1/2L\ ||\nabla f(z)||^2$

Only the green part above depends on x, it's always non-negative, and we have
$||\ [z - \nabla f(z)/L] - [z - \nabla f(z)/L]\ ||^2 = 0$

Gradient projection method

Problem: minimize convex f(x)
s.t. $x \in Q$, where Q is a convex set



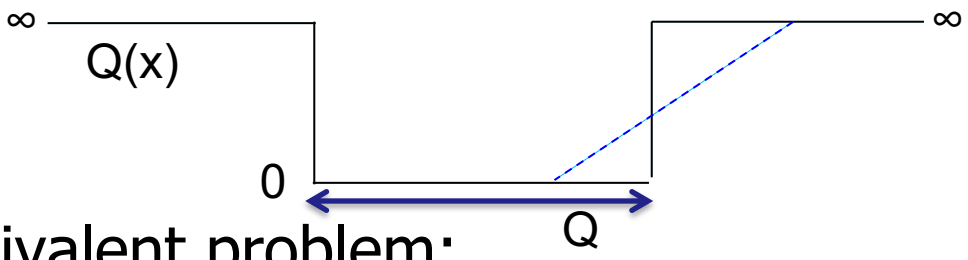
Let Q(x) be an *indicator function* for set Q

$Q(x) = 0$ if $x \in Q$,

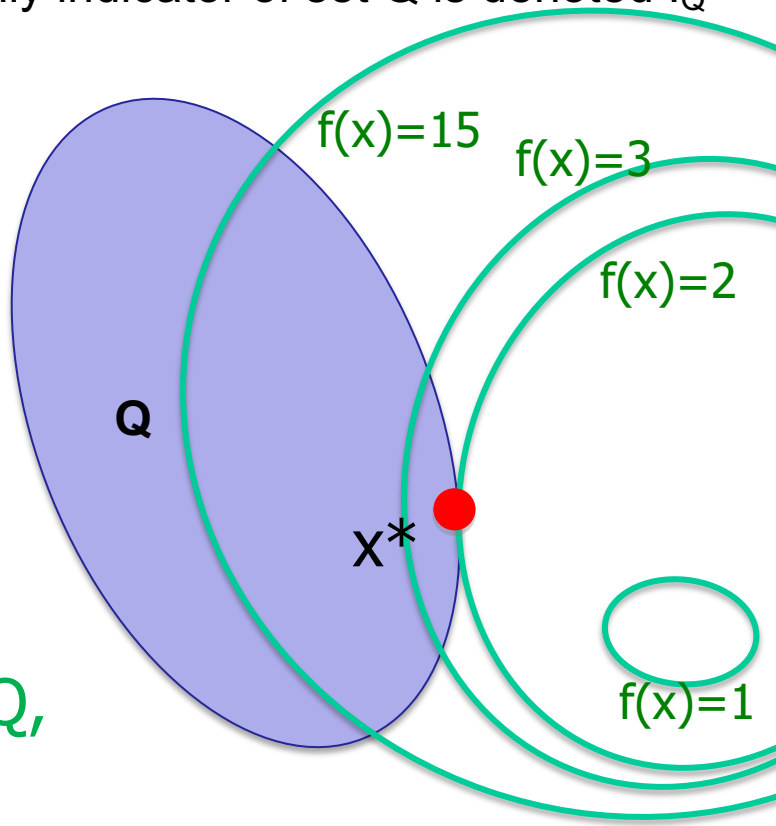
$Q(x) = \infty$ otherwise

(has to be infinity; if finite large const., then Q(x) not convex)

This is not the usual notation in literature; typically indicator of set Q is denoted I_Q



Equivalent problem:
minimize convex f(x)+Q(x)



f+Q has finite values only for $x \in Q$, so minimum is in Q