DUALITY

WHY DUALITY?

No constraints

$$\text{minimize} \quad f(x)$$

Gradient descent
Newton's method
Quasi-newton
Conjugate gradients
etc…

Non-differentiable

$$\text{minimize} \quad f(x)$$

????

Constrained problems?

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g(x) \leq 0 \\ & h(x) = 0 \end{aligned}$$

????

LAGRANGIAN

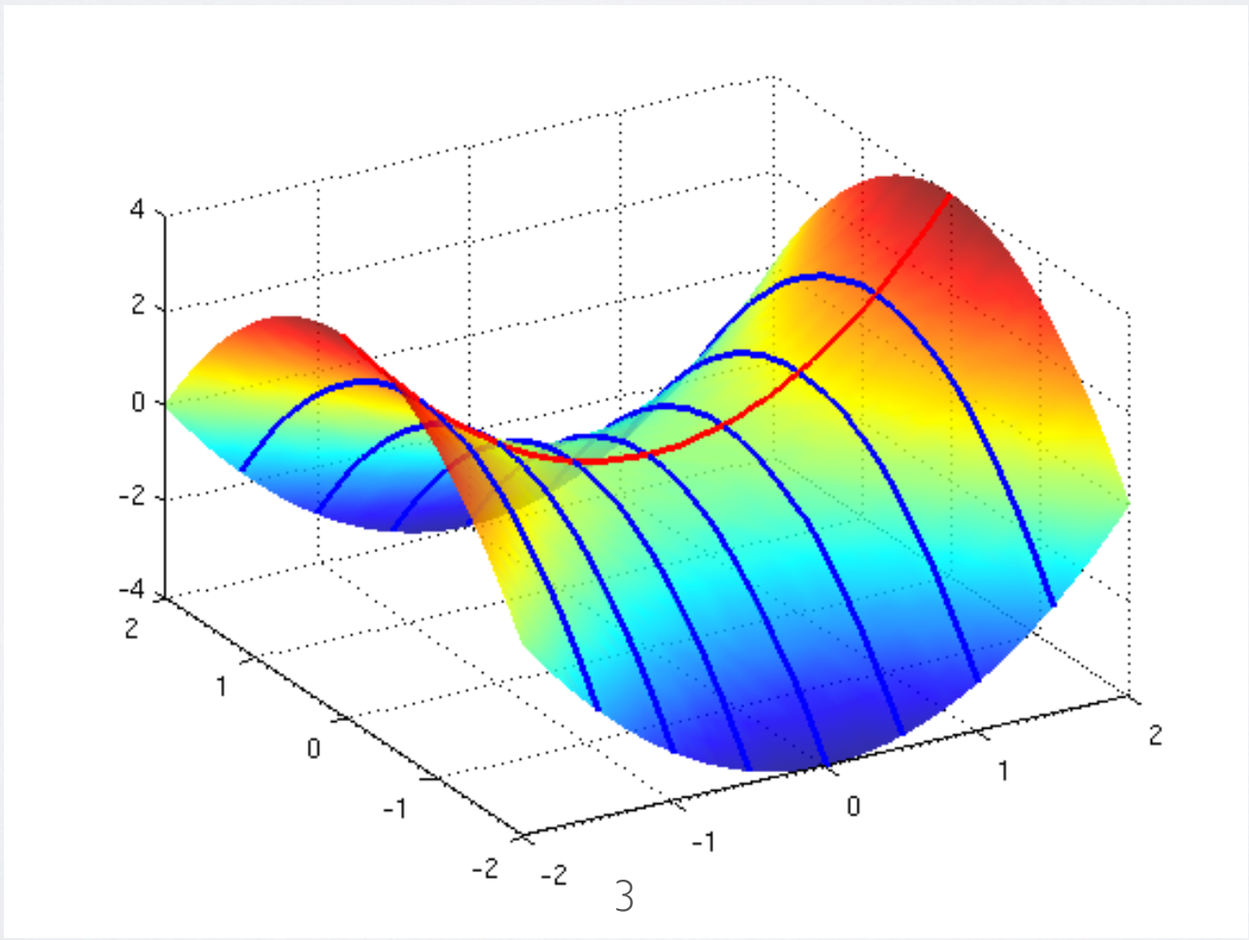
Simple case

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & Ax - b = 0 \end{aligned}$$

"Saddle-point" form

$$\min_x \max_\lambda \underline{f(x) + \langle \lambda, Ax - b \rangle}$$

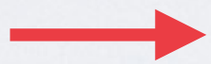
Lagrangian



SADDLE-POINT FORM

Simple case

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & Ax - b = 0 \end{aligned}$$



"Saddle-point" form

$$\min_x \max_\lambda f(x) + \langle \lambda, Ax - b \rangle$$

Lagrangian

$$\max_\lambda f(x) + \langle \lambda, Ax - b \rangle = \begin{cases} f(x), & Ax - b = 0 \\ \infty, & \text{otherwise} \end{cases}$$

$$\min_x \max_\lambda f(x) + \langle \lambda, Ax - b \rangle = \min_x f(x)$$

INEQUALITY CONSTRAINTS

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & Ax - b = 0 \\ & Cx - d \leq 0 \end{aligned}$$

$$\min_{x} \max_{\lambda, \nu \geq 0} f(x) + \langle \lambda, Ax - b \rangle + \langle \nu, Cx - d \rangle$$

non-negative constraint

Why does this work?

$$Cx - d \leq 0 : \quad \nu = 0$$

GENERAL FORM LAGRANGIAN

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g(x) \leq 0 \\ & h(x) = 0 \end{aligned}$$

Lagrangian

$$L(x, \lambda, \nu) = f(x) + \langle \lambda, h(x) \rangle + \langle \nu, g(x) \rangle$$

Saddle-point formulation

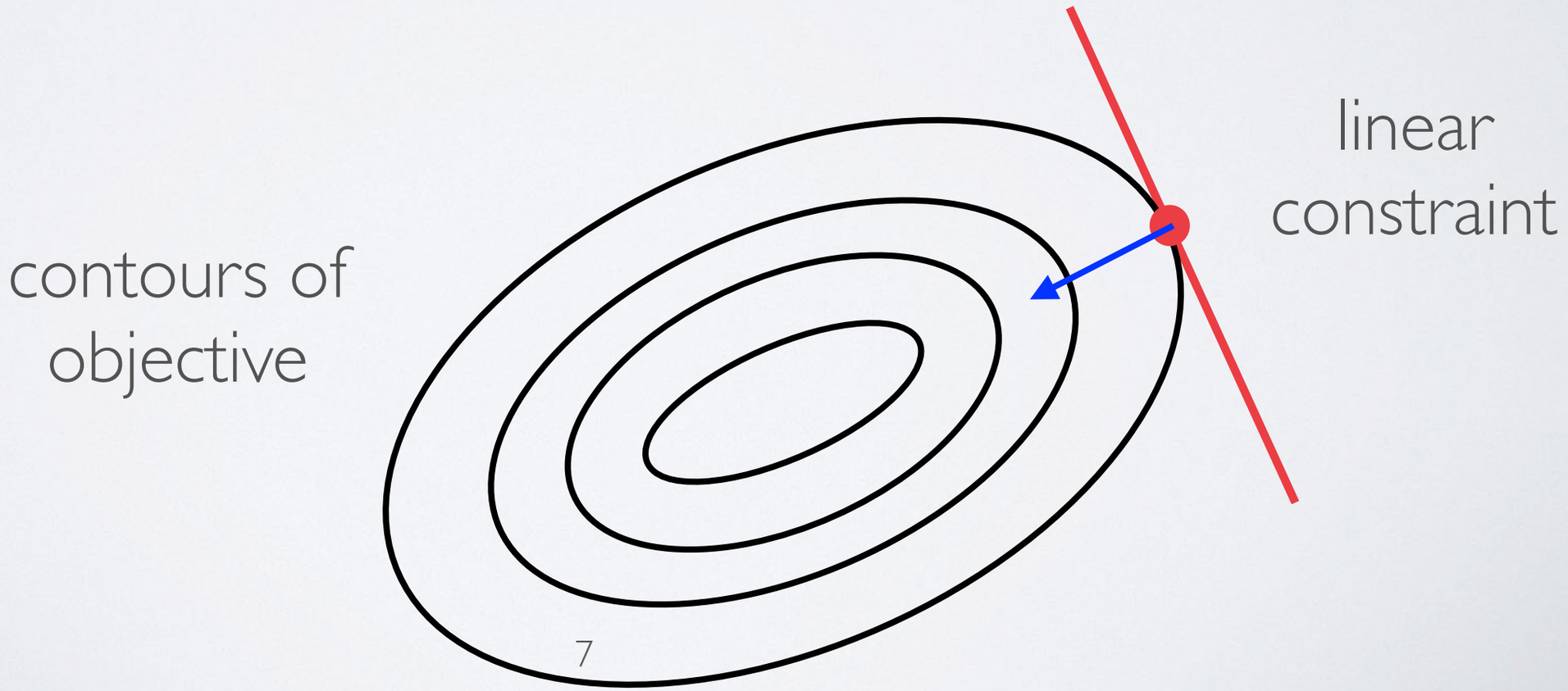
$$f(x^\star) = \min_x \max_{\lambda, \nu \geq 0} f(x) + \langle \lambda, h(x) \rangle + \langle \nu, g(x) \rangle$$

CALCULUS INTERPRETATION

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & h(x) = 0 \end{aligned}$$

gradient of objective parallel to gradient of constraint

$$-\nabla f(x) = \nabla h(x)\lambda \quad \longrightarrow \quad \nabla f(x) + \nabla h(x)\lambda = 0$$



CALCULUS INTERPRETATION

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & h(x) = 0 \end{aligned}$$

gradient of objective parallel to gradient of constraint

$$-\nabla f(x) = \nabla h(x)\lambda \quad \longrightarrow \quad \nabla f(x) + \nabla h(x)\lambda = 0$$

$$L(x, \lambda) = f(x) + \langle \lambda, h(x) \rangle$$

$\min_x$

optimality condition

$$\partial_x L(x, \lambda) = \nabla f(x) + \nabla h(x)\lambda = 0$$

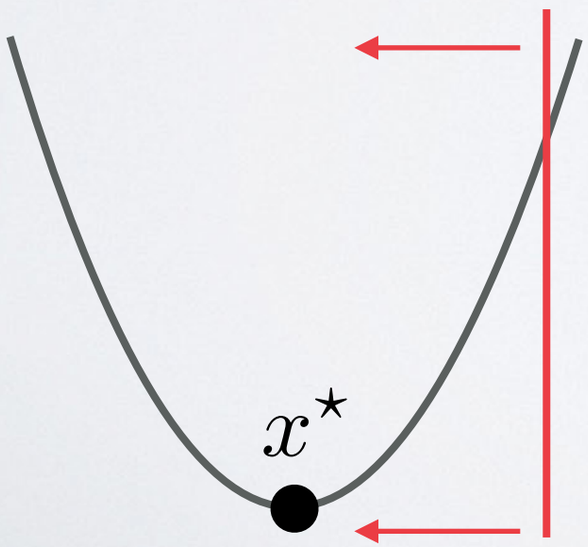
INEQUALITY CONDITIONS

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g(x) \leq 0 \end{aligned}$$

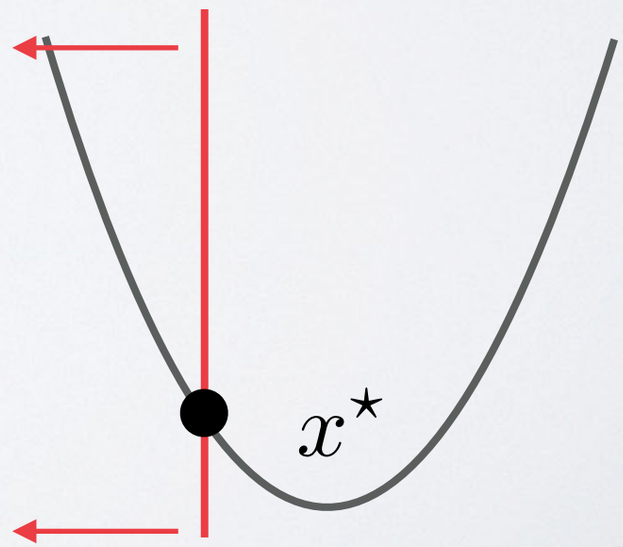
$$L(x, \nu) = f(x) + \langle \nu, g(x) \rangle$$

$$\min_{x} \max_{\nu \geq 0} \; f(x) + \langle \nu, g(x) \rangle$$

Inactive: $\nu = 0$



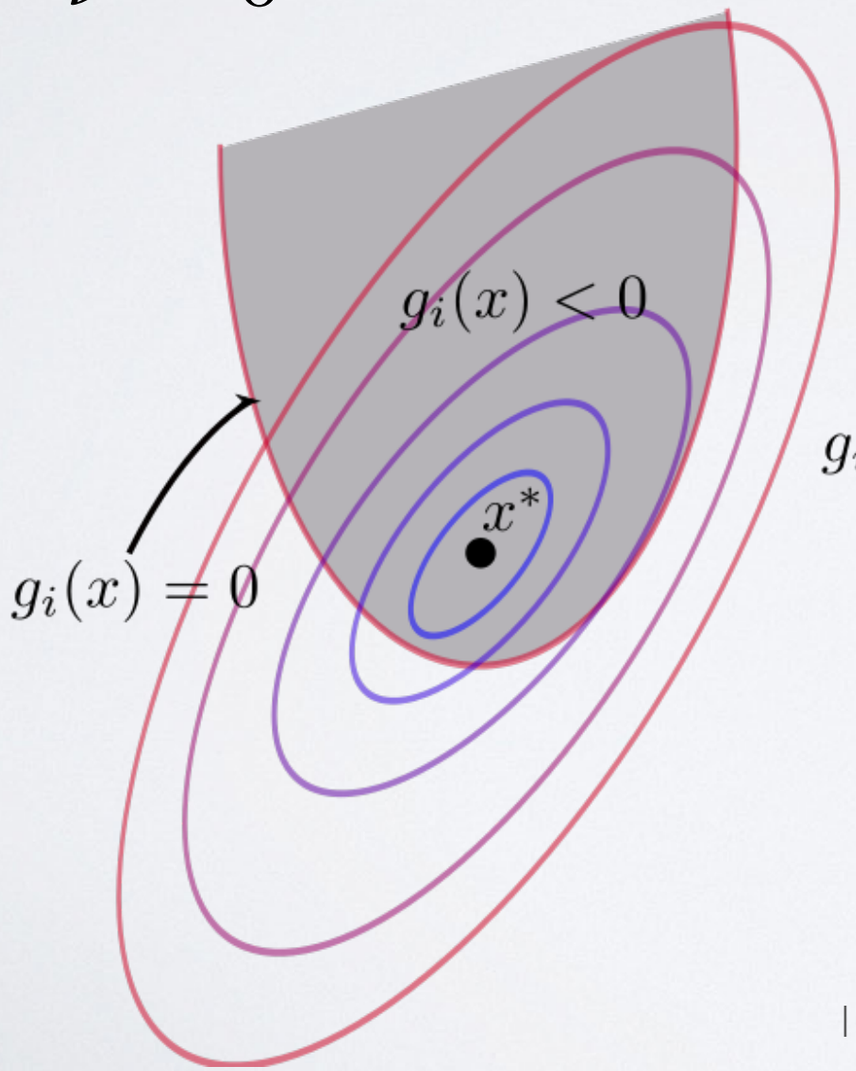
Active: $\nu > 0$



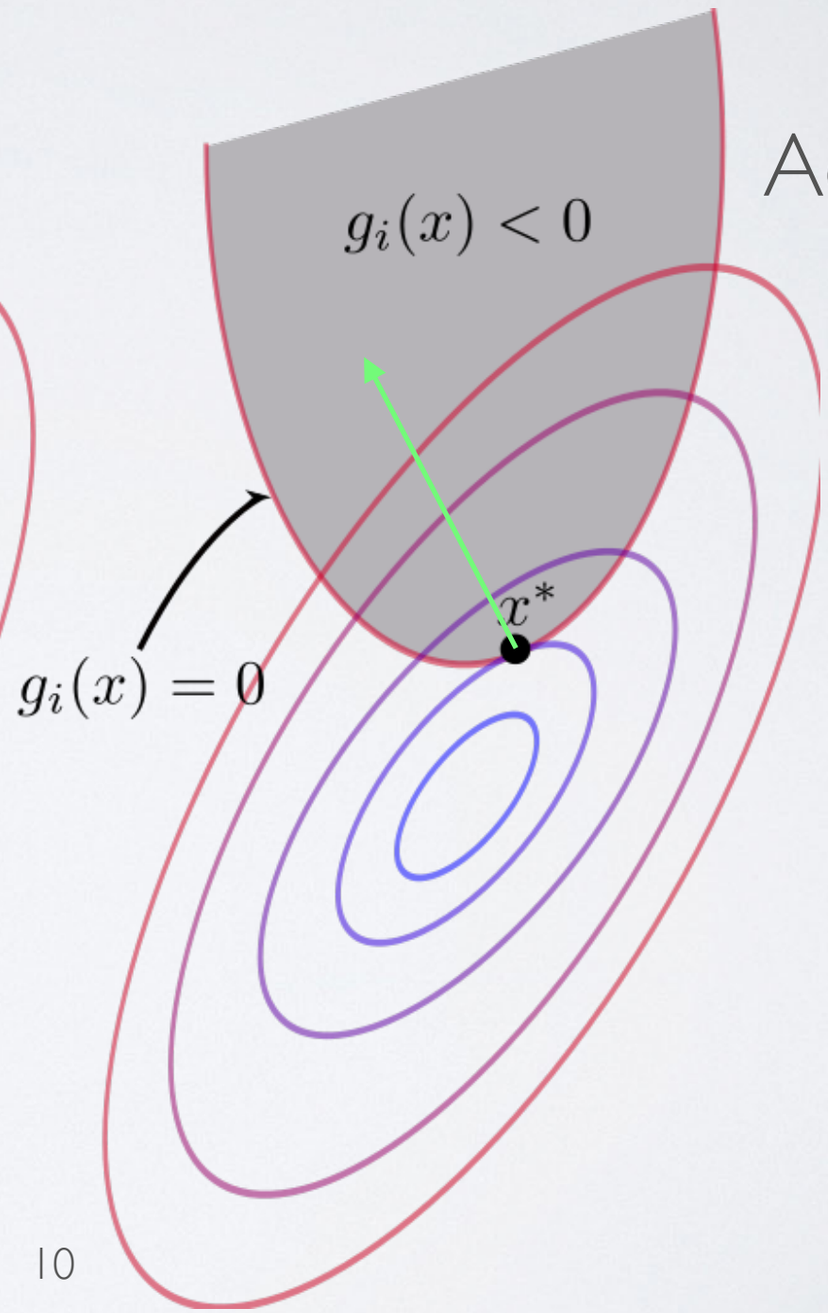
INEQUALITY CONDITIONS

$$\partial_x L(x, \lambda) = \nabla f(x) + \nabla g(x)\nu = 0$$

Inactive: $\nu = 0$



Active: $\nu > 0$



OPTIMALITY CONDITIONS: KKT SYSTEM

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g(x) \leq 0 \\ & h(x) = 0 \end{aligned}$$

$$L(x, \lambda, \nu) = f(x) + \langle \lambda, h(x) \rangle + \langle \nu, g(x) \rangle$$

necessary conditions

Karush-Kuhn-Tucker

$$\nabla f(x^\star) + \nabla h(x^\star)\lambda + \nabla g(x^\star)\nu = 0 \quad \text{Primal/dual optimality}$$

$$\left.\begin{aligned} g(x^\star) &\leq 0 \\ h(x^\star) &= 0 \end{aligned}\right\} \quad \text{Primal feasibility}$$

$$\nu \geq 0 \quad \text{Dual feasibility}$$

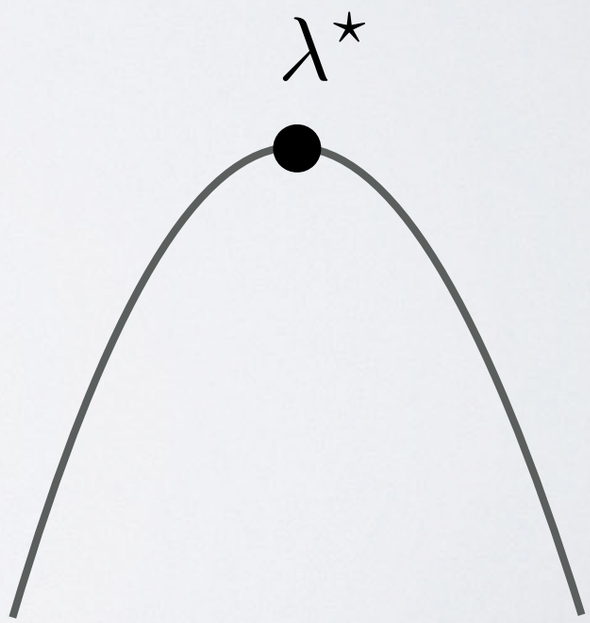
$$\nu_i g_i(x^\star) = 0 \quad \text{Comp slackness}$$

DUAL FUNCTION

$$d(\lambda, \nu) = \min_x L(x, \lambda, \nu)$$
$$= \min_x f(x) + \langle \lambda, h(x) \rangle + \langle \nu, g(x) \rangle$$

Dual is **concave**

Why?
Does this depend on convexity of f ?



DUAL FUNCTION

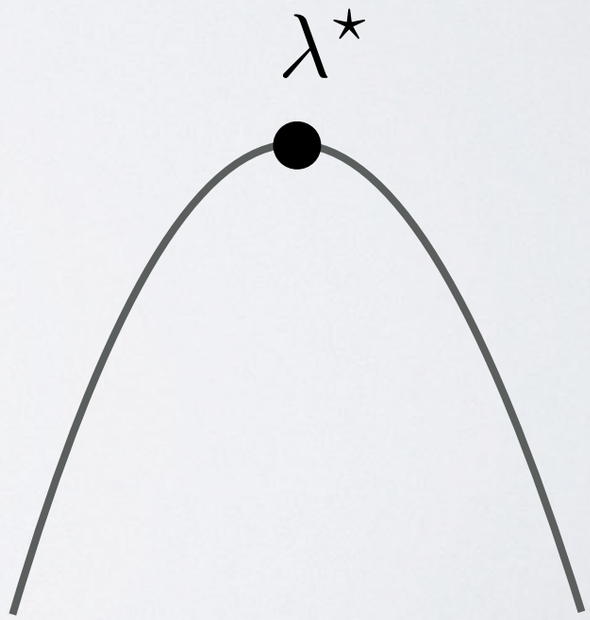
$$d(\lambda, \nu) = \min_x L(x, \lambda, \nu)$$
$$= \min_x f(x) + \langle \lambda, h(x) \rangle + \langle \nu, g(x) \rangle$$

Dual is **lower bound** to optimal objective

$$\min_x f(x) + \langle \lambda, h(x) \rangle + \langle \nu, g(x) \rangle \leq f(x^\star)$$

Why?

(because optimal x satisfies constraints)



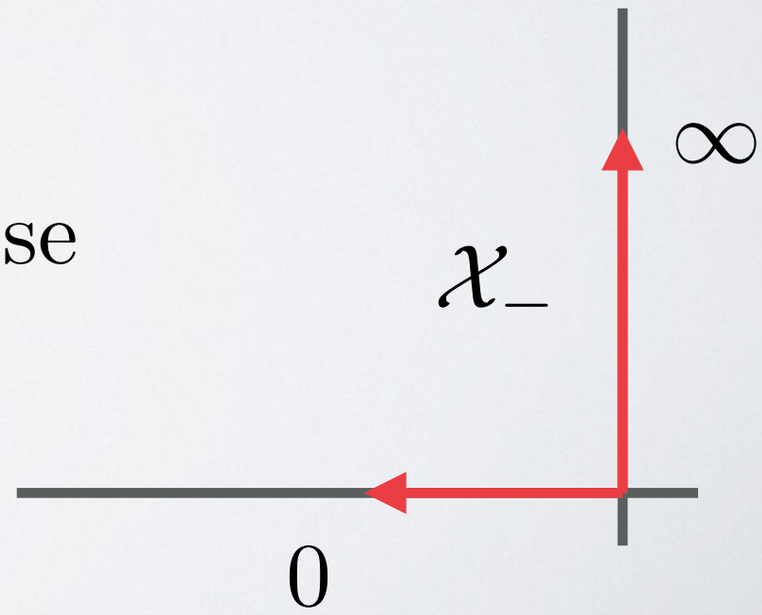
GEOMETRIC INTERPRETATION OF LOWER BOUND

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g(x) \leq 0 \\ & h(x) = 0 \end{aligned}$$

Implicit constraints

$$\text{minimize} \quad f(x) + \mathcal{X}_0(h(x)) + \mathcal{X}_-(g(x))$$

$$\mathcal{X}_-(z) = \begin{cases} 0, & z \leq 0 \\ \infty, & \text{otherwise} \end{cases}$$

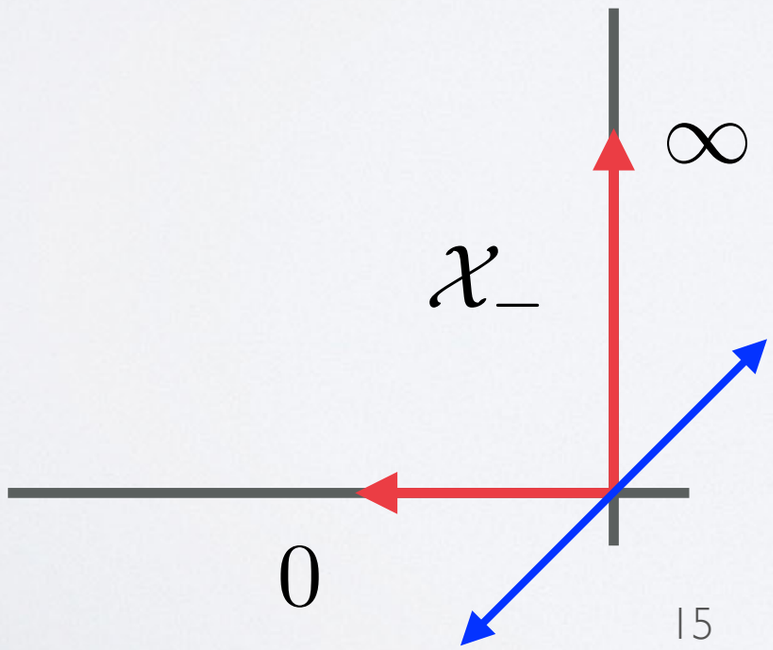


GEOMETRIC INTERPRETATION

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g(x) \leq 0 \\ & h(x) = 0 \end{aligned}$$

Implicit constraints

$$\text{minimize} \quad f(x) + \mathcal{X}_0(h(x)) + \mathcal{X}_-(g(x))$$



linear
lower bound

$$\mathcal{X}_-(z) \geq \langle z, \nu \rangle, \ \nu \geq 0$$

GEOMETRIC INTERPRETATION

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g(x) \leq 0 \\ & h(x) = 0 \end{aligned}$$

Implicit constraints

$$\text{minimize} \quad f(x) + \mathcal{X}_0(h(x)) + \mathcal{X}_-(g(x))$$

Lower bound

$$\text{minimize} \quad f(x) + \langle \lambda, h(x) \rangle + \langle \nu, g(x) \rangle$$

MAX OF DUAL

$$\max_{\lambda,\nu\geq 0} d(\lambda,\nu) = \max_{\lambda,\nu\geq 0} \min_{x} \underbrace{f(x) + \langle \lambda, h(x)\rangle + \langle \nu, g(x)\rangle}_{L(x,\lambda,\nu)}$$

$$\max_{\lambda,\nu\geq 0} d(\lambda,\nu) = \max_{\lambda,\nu\geq 0} \min_{x} L(x,\lambda,\nu) \overset{???}{=} \underbrace{\min_{x} \max_{\lambda,\nu\geq 0} L(x,\lambda,\nu)}_{\text{Solution}}$$

Does maximizing dual = minimizing primal ???

WEAK/STRONG DUALITY

Weak duality

$$\max_{\lambda,\nu \geq 0} d(\lambda, \nu) \leq f(x^\star)$$

Always holds: even for non-convex problems

Strong duality

$$\max_{\lambda,\nu \geq 0} d(\lambda, \nu) = f(x^\star)$$

Holds "most of the time" for convex problems

SLATER'S CONDITION

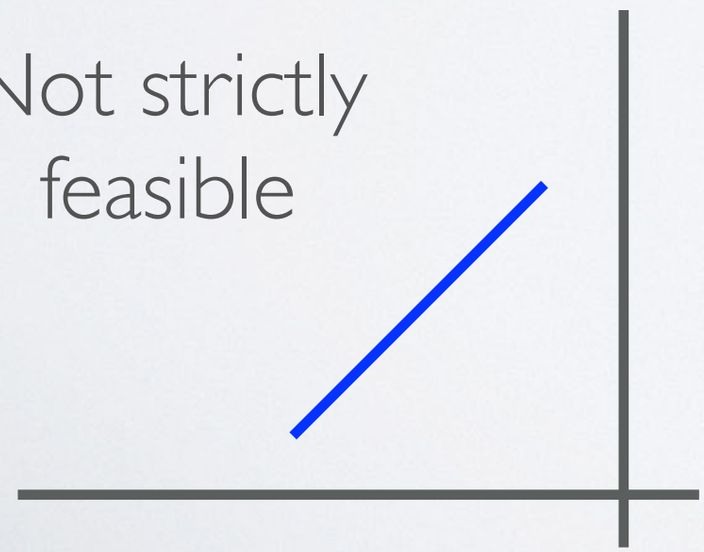
$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g(x) \leq 0 \\ & Ax = b \end{aligned}$$

"Constraint qualification"

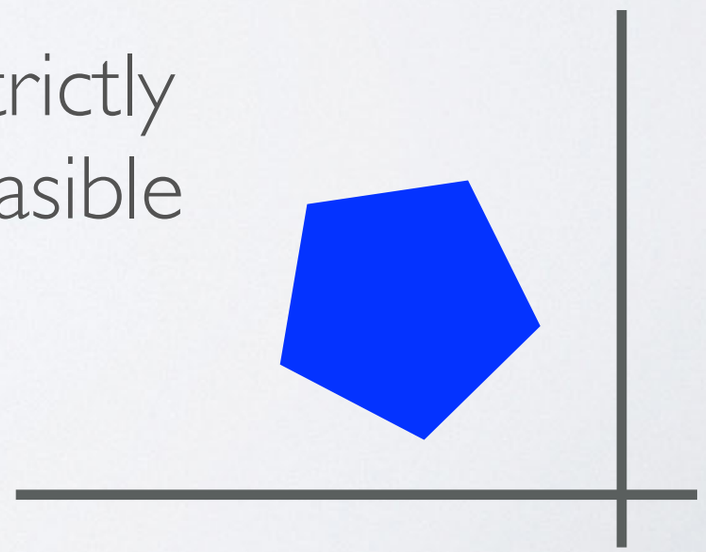
Slater's condition holds if there is a **strictly feasible** point

$$\begin{aligned} f(x) &< \infty \\ g(x) &< 0 \\ Ax &= b \end{aligned}$$

Not strictly feasible



Strictly feasible



SLATER'S CONDITION

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & g(x) \leq 0 \\ & Ax = b \end{aligned}$$

Slater's condition holds if there is a **strictly feasible** point →

$$\begin{aligned} f(x) &< \infty \\ g(x) &< 0 \\ Ax &= b \end{aligned}$$

Theorem

Convex+(linear equalities)+(Slater's condition) = strong duality

$$\max_{\lambda,\nu \geq 0} \min_{x} L(x, \lambda, \nu) = \min_{x} \max_{\lambda,\nu \geq 0} L(x, \lambda, \nu)$$

$$\max_{\lambda,\nu \geq 0} d(\lambda, \nu) = f(x^\star)$$

NON-HOMOGENOUS SVM

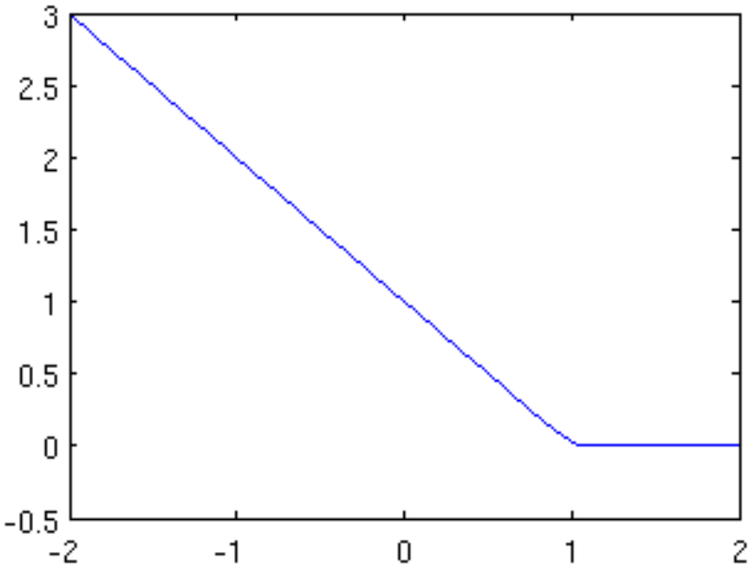
$$\underset{w,b}{\text{minimize}} \quad \frac{1}{2}\|w\|^2 + C\sum h[y_i(x_i^T w - b)]$$

Choose line with largest margin:

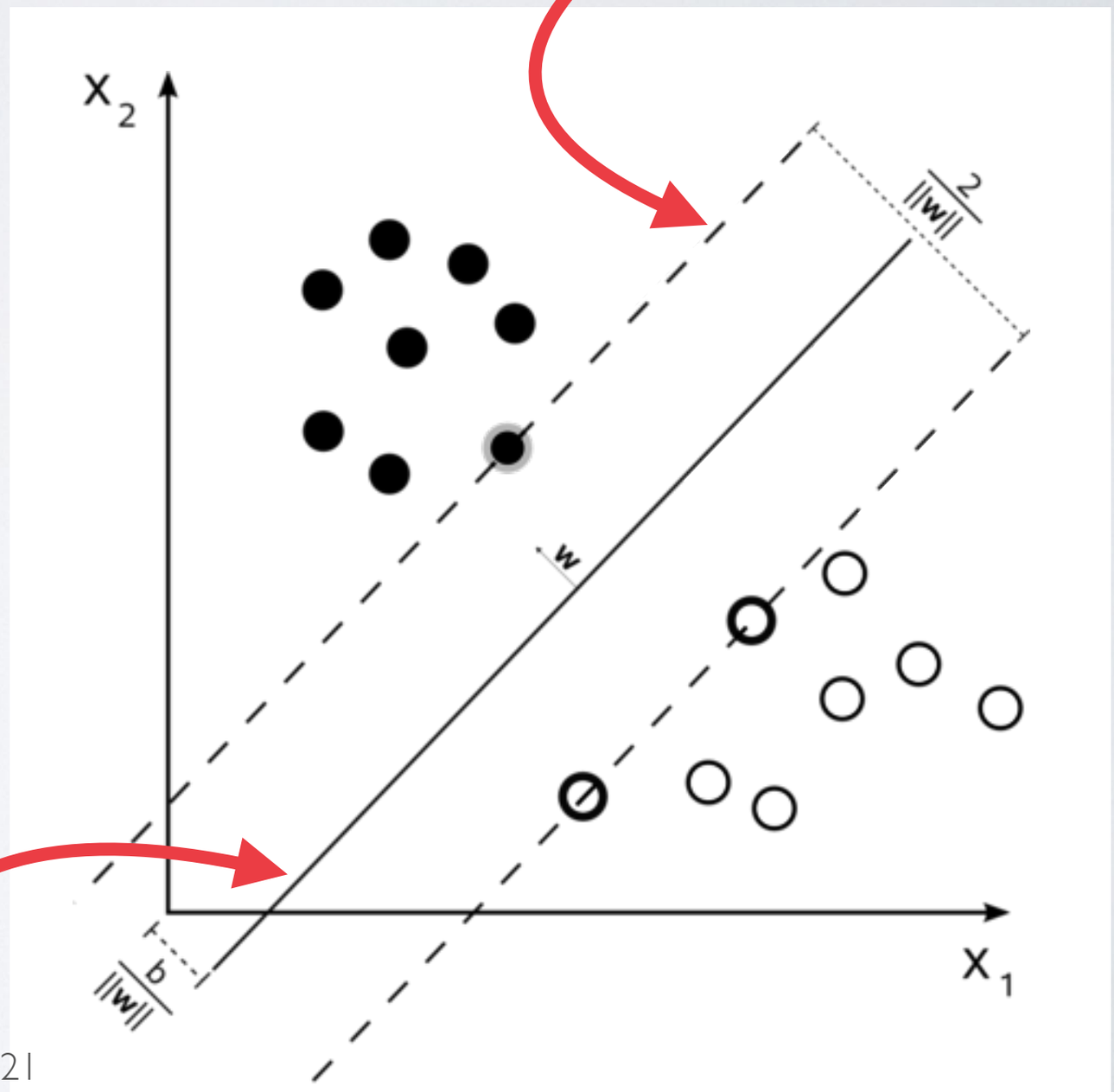
$$\frac{2}{\|w\|}$$

Push data to "right" side of line:

$$h(l_i[x_i^T w - b])$$



$$x^T w - b = 1$$



$$x^T w - b = 0$$

DUAL: THE HARD WAY

$$\underset{w,b}{\text{minimize}} \quad \frac{1}{2}\|w\|^2 + C\sum h[y_i(x_i^T w - b)]$$

$$h(z) = \max\{z - 1, 0\}$$

$$\underset{w,b}{\text{minimize}} \quad \frac{1}{2}\|w\|^2 + C\sum \max\{1 - y_i(x_i^T w - b), 0\}$$

Idea: make this differentiable

$$\underset{w,b}{\text{minimize}} \quad \frac{1}{2}\|w\|^2 + C\sum \max\{\mathbf{1} - YXw + yb, 0\}$$

Standard form

$$\begin{aligned} \underset{w,b,v}{\text{minimize}} \quad & \frac{1}{2}\|w\|^2 + C\langle \mathbf{1}, v\rangle \\ \text{subject to} \quad & v \geq \mathbf{1} - YXw + yb \\ & v \geq 0 \end{aligned}$$

DUAL: THE HARD WAY

$$\begin{aligned} \underset{w,b,v}{\text{minimize}} \quad & \frac{1}{2}\|w\|^2 + C\langle \mathbf{1}, v\rangle \\ \text{subject to} \quad & v \geq \mathbf{1} - YXw + yb \\ & v \geq 0 \end{aligned}$$

Positive multipliers

$$d(\alpha,\gamma) = \min_{w,b,v} \frac{1}{2}\|w\|^2 + C\langle \mathbf{1}, v\rangle + \langle \alpha, \mathbf{1} - YXw + yb - v\rangle + \langle \gamma, -v\rangle$$

Break it up!

$$d(\alpha,\gamma) = \min_{w,b,v} \frac{1}{2}\|w\|^2 + C\langle \mathbf{1}, v\rangle + \langle \alpha, \mathbf{1}\rangle - \langle \alpha, YXw\rangle + \langle \alpha, y\rangle b - \langle \alpha, v\rangle - \langle \gamma, v\rangle$$

DUAL: THE HARD WAY

$$d(\alpha,\gamma) = \min_{w,b,v} \frac{1}{2}\|w\|^2 + C\langle \mathbf{1}, v\rangle + \langle \alpha, \mathbf{1}\rangle - \langle \alpha, YXw\rangle + \langle \alpha, y\rangle b - \langle \alpha, v\rangle - \langle \gamma, v\rangle$$

This is too complicated. Let's reduce it...

Put all the v terms together

$$d(\alpha,\gamma) = \min_{w,b,v} \frac{1}{2}\|w\|^2 + \langle \alpha, \mathbf{1}\rangle - \langle \alpha, YXw\rangle + \langle \alpha, y\rangle b + \underline{\langle C\mathbf{1} - \alpha - \gamma, v\rangle}$$

derivative for v

$$C\mathbf{1} - \alpha - \gamma = 0, \text{ with } \alpha, \gamma \geq 0$$

$$0 \leq \alpha_i \leq C$$

remember this constraint!

DUAL: THE HARD WAY

$$d(\alpha,\gamma) = \min_{w,b,v} \frac{1}{2}\|w\|^2 + C\langle \mathbf{1}, v\rangle + \langle \alpha, \mathbf{1}\rangle - \langle \alpha, YXw\rangle + \langle \alpha, y\rangle b - \langle \alpha, v\rangle - \langle \gamma, v\rangle$$

This is too complicated. Let's reduce it...

$$d(\alpha,\gamma) = \min_{w,b,v} \frac{1}{2}\|w\|^2 + \langle \alpha, \mathbf{1}\rangle - \langle \alpha, YXw\rangle + \langle \alpha, y\rangle b + \langle C\mathbf{1} - \alpha - \gamma, v\rangle$$

$$d(\alpha,\gamma) = \min_{w,b,v} \frac{1}{2}\|w\|^2 + \langle \alpha, \mathbf{1}\rangle - \langle \alpha, YXw\rangle + \underline{\langle \alpha, y\rangle b}$$

derivative for b

$$\langle \alpha, y\rangle = 0$$

remember this constraint!

DUAL: THE HARD WAY

$$d(\alpha,\gamma) = \min_{w,b,v} \frac{1}{2}\|w\|^2 + C\langle \mathbf{1}, v\rangle + \langle \alpha, \mathbf{1}\rangle - \langle \alpha, YXw\rangle + \langle \alpha, y\rangle b - \langle \alpha, v\rangle - \langle \gamma, v\rangle$$

This is too complicated. Let's reduce it...

$$d(\alpha,\gamma) = \min_{w,b,v} \frac{1}{2}\|w\|^2 + \langle \alpha, \mathbf{1}\rangle - \langle \alpha, YXw\rangle + \langle \alpha, y\rangle b + \langle C\mathbf{1} - \alpha - \gamma, v\rangle$$

$$d(\alpha,\gamma) = \min_{w,b,v} \frac{1}{2}\|w\|^2 + \langle \alpha, \mathbf{1}\rangle - \langle \alpha, YXw\rangle + \langle \alpha, y\rangle b$$

$$d(\alpha,\gamma) = \min_{w,b,v} \frac{1}{2}\|w\|^2 + \langle \alpha, \mathbf{1}\rangle - \langle \alpha, YXw\rangle$$

derivative for w

$$w - X^T Y\alpha = 0, \text{ or } w = X^T Y\alpha$$

DUAL: THE HARD WAY

$$d(\alpha, \gamma) = \min_{w,b,v} \frac{1}{2}\|w\|^2 + \langle \alpha, \mathbf{1} \rangle - \langle \alpha, YXw \rangle$$

derivative for w

$$w - X^T Y \alpha = 0, \text{ or } w = X^T Y \alpha$$

$$d(\alpha) = \frac{1}{2}\|X^T Y \alpha\|^2 + \langle \alpha, \mathbf{1} \rangle - \langle \alpha, YXX^T Y \alpha \rangle$$

$$d(\alpha) = \frac{1}{2}\|X^T Y \alpha\|^2 + \langle \alpha, \mathbf{1} \rangle - \|X^T Y \alpha\|^2$$

$$d(\alpha) = -\frac{1}{2}\|X^T Y \alpha\|^2 + \langle \alpha, \mathbf{1} \rangle$$

DUAL: THE HARD WAY

Dual problem…

$$\max \; d(\alpha) = -\frac{1}{2}\|X^T Y \alpha\|^2 + \langle \alpha, \mathbf{1} \rangle$$

$$\text{subject to} \;\; 0 \leq \alpha \leq C$$

$$\langle \alpha, y \rangle = 0$$

note: $w = X^T Y \alpha$

Solvable via "coordinate descent"
LIBSVM

DUAL: THE HARD WAY

Dual problem…

$$\max \; d(\alpha) = -\frac{1}{2}\|X^T Y \alpha\|^2 + \langle \alpha, \mathbf{1} \rangle$$

$$\text{subject to} \;\; 0 \leq \alpha \leq C$$

$$\langle \alpha, y \rangle = 0$$

"kernel form"

$$d(\alpha) = -\frac{1}{2}\alpha^T Y \underline{X X^T} Y \alpha + \langle \alpha, \mathbf{1} \rangle$$

kernel

KERNELS

$$(XX^T)_{ij} = \langle x_i, x_j \rangle = \begin{cases} \text{big, when } x_i \text{ and } x_j \text{ are close} \\ \text{small, when } x_i \text{ and } x_j \text{ are far apart} \end{cases}$$

kernel function

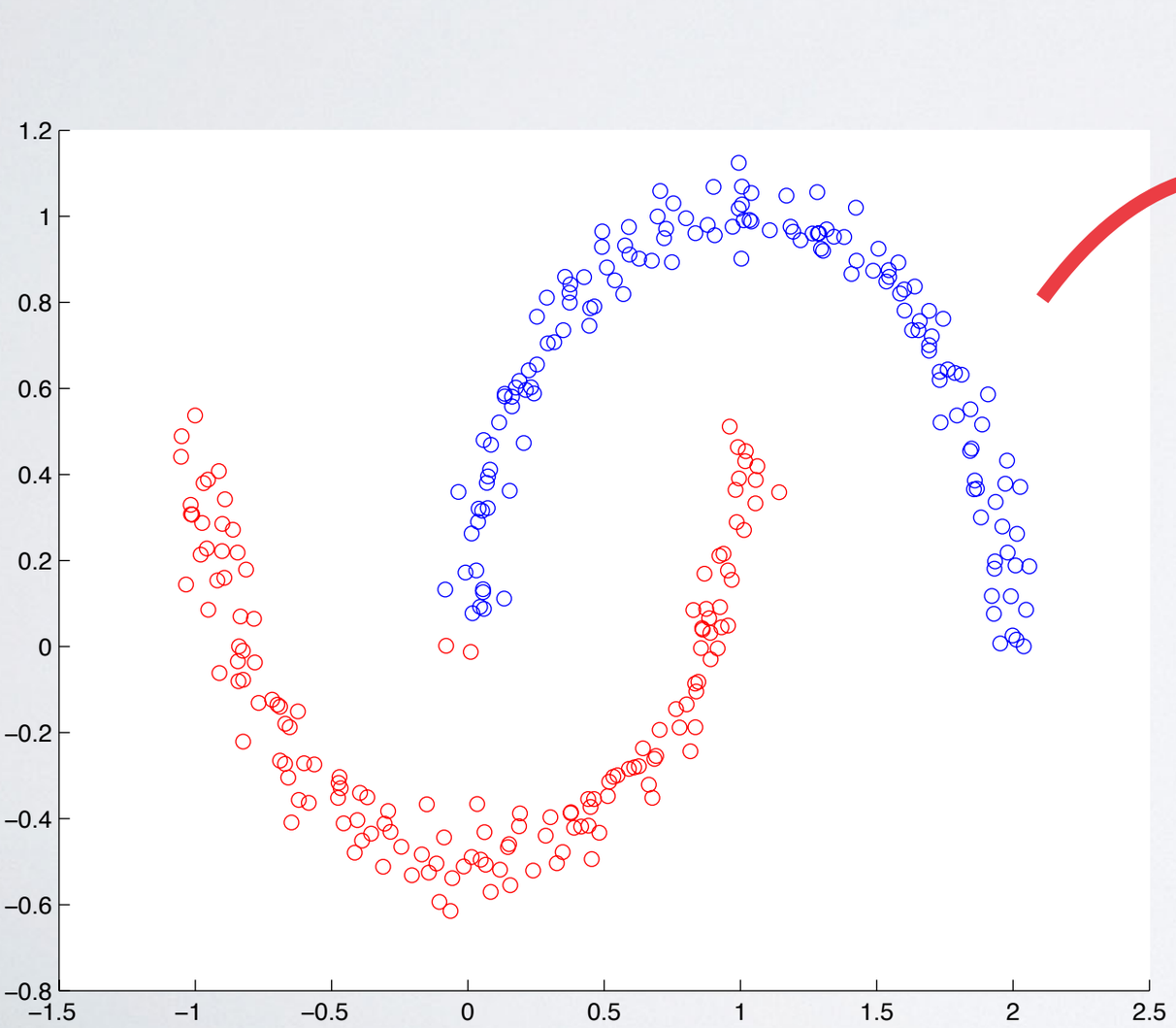
$$k(x_i, x_j) = \begin{cases} \text{big, when } x_i \text{ and } x_j \text{ are “similar”} \\ \text{small, when } x_i \text{ and } x_j \text{ are “different”} \end{cases}$$

Dual SVM: only need to know similarity function

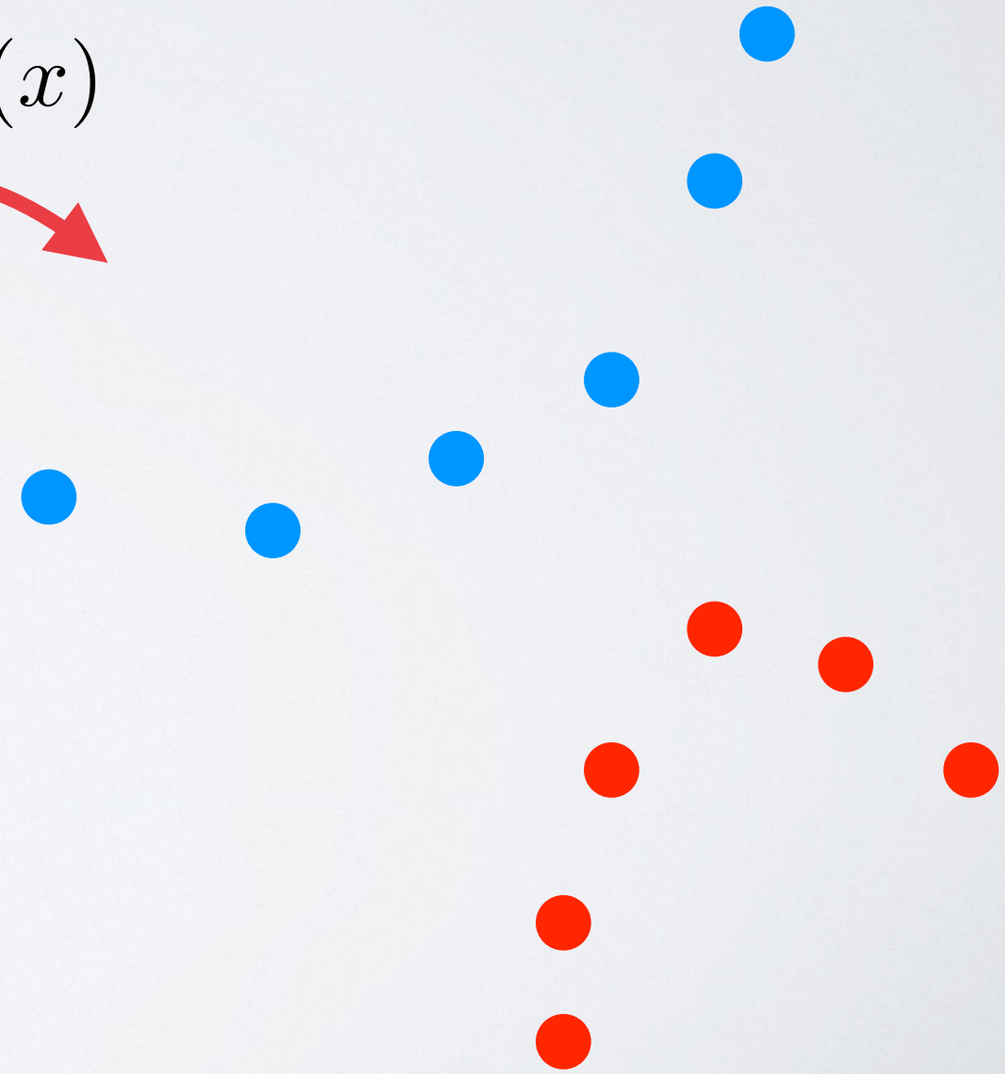
Kernel methods: replace inner product with some other similarity

EXAMPLE: TWO MOONS

I want a mapping that linearizes the problem. I can't write down that mapping, but I can write down a similarity measure for the mapped points!



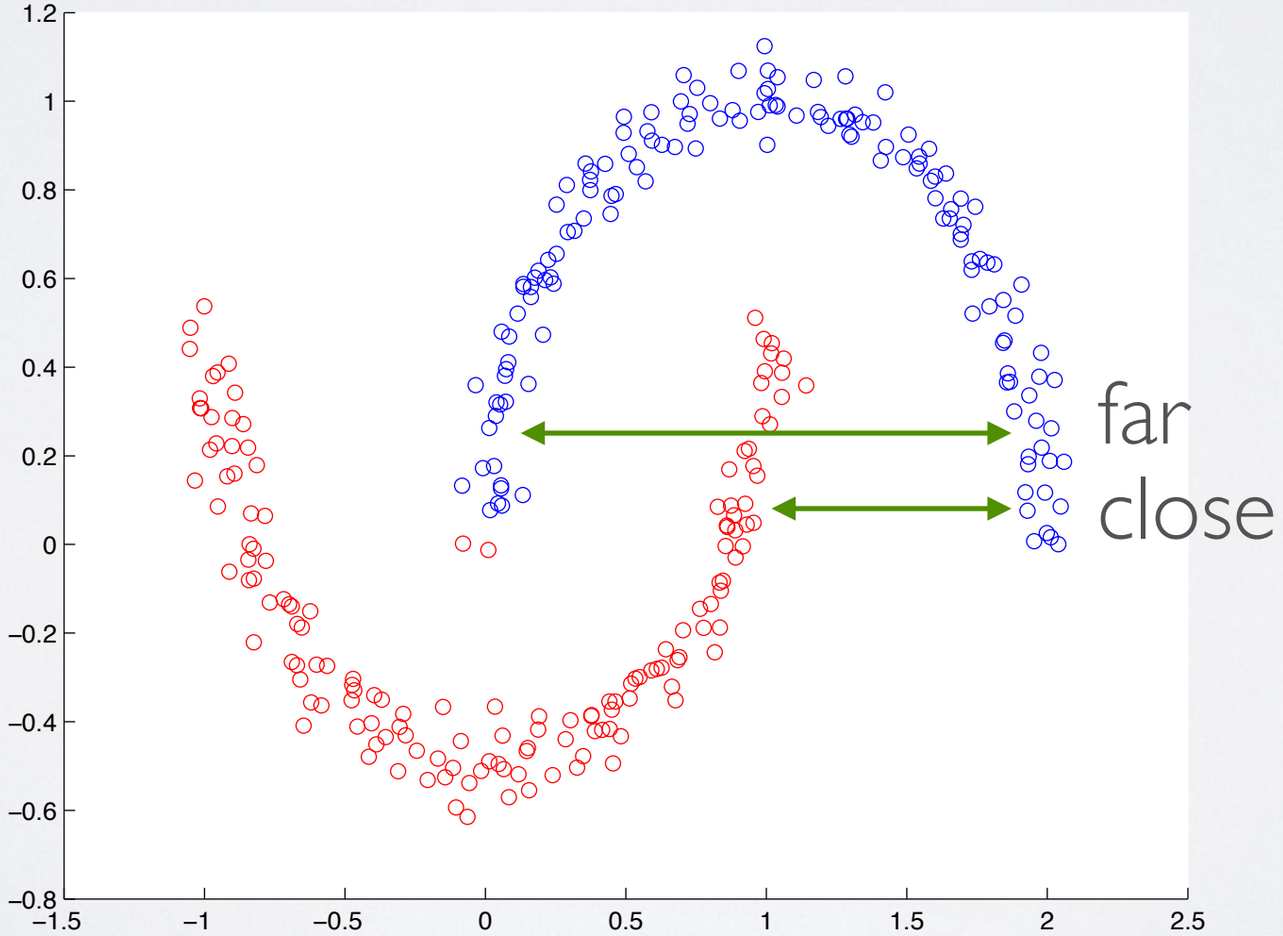
$\phi(x)$



EXAMPLE: TWO MOONS

kernel function

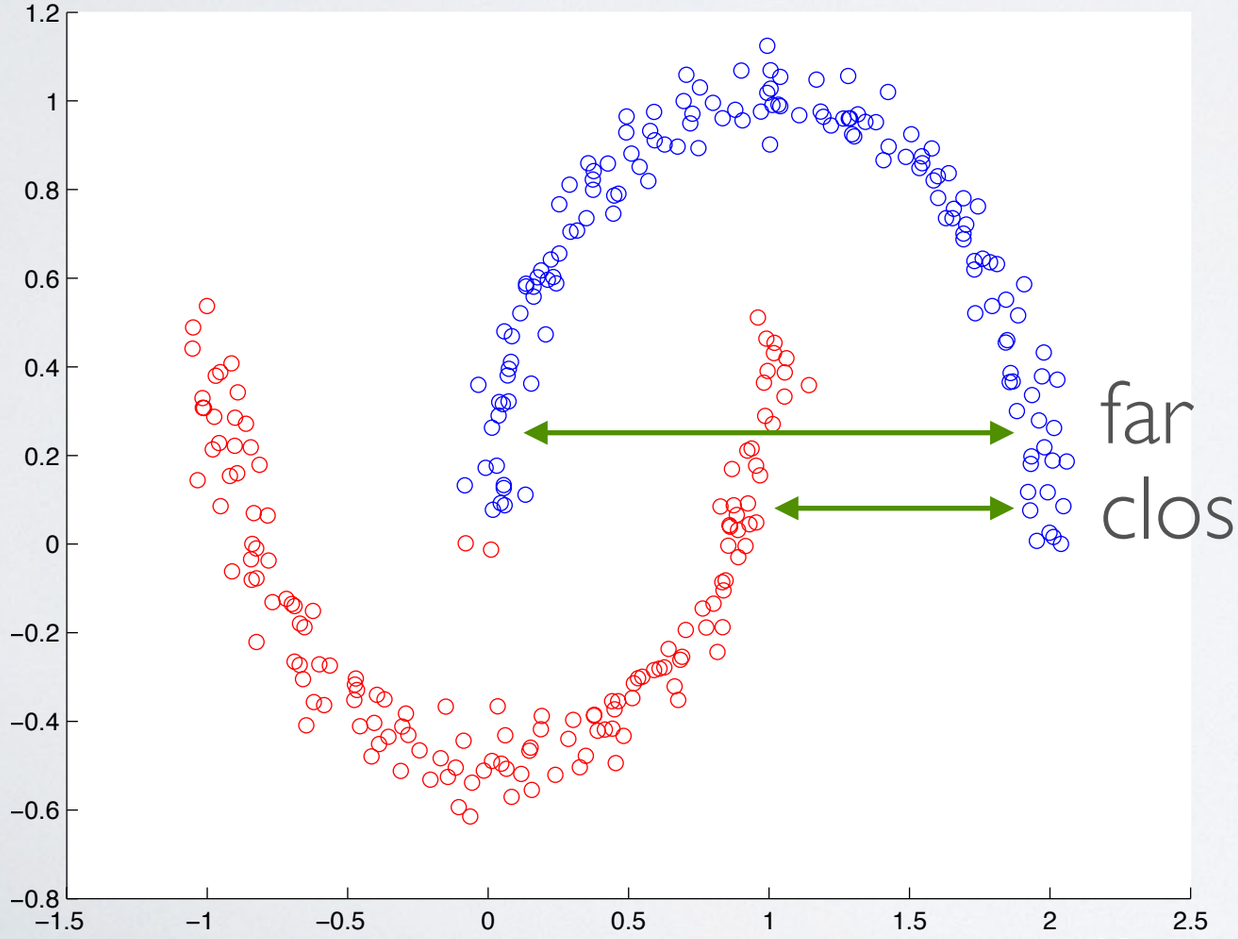
$$k(x_i, x_j) = \begin{cases} \text{big, when } x_i \text{ and } x_j \text{ are “similar”} \\ \text{small, when } x_i \text{ and } x_j \text{ are “different”} \end{cases}$$



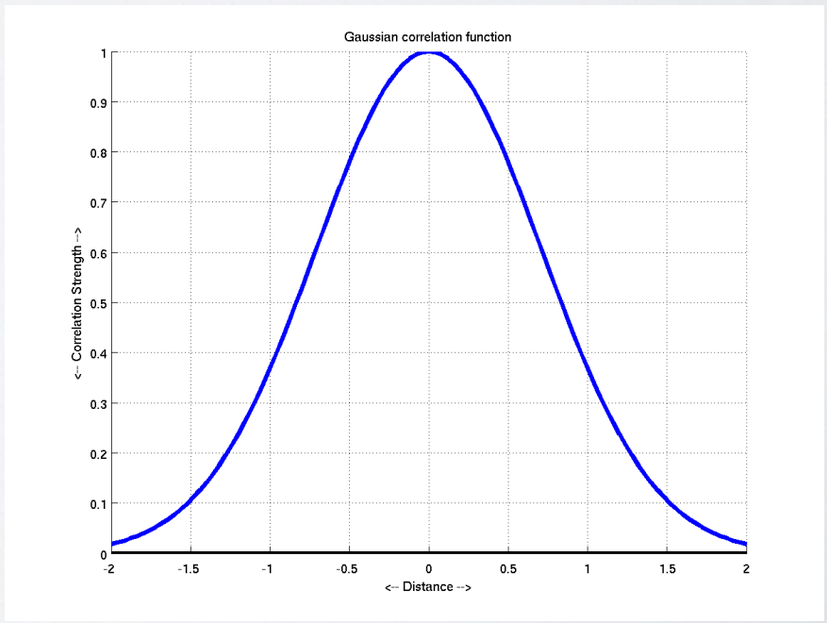
EXAMPLE: TWO MOONS

kernel function

$$k(x_i, x_j) = \begin{cases} \text{big, when } x_i \text{ and } x_j \text{ are “similar”} \\ \text{small, when } x_i \text{ and } x_j \text{ are “different”} \end{cases}$$



$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$



EXAMPLE: TWO MOONS

$$\text{maximize} \quad d(\alpha) = -\frac{1}{2}\alpha^T Y \underline{X X^T} Y \alpha + \langle \alpha, \mathbf{1} \rangle$$

kernel

$$\text{subject to} \quad 0 \leq \alpha \leq C$$

$$\langle \alpha, l \rangle = 0$$

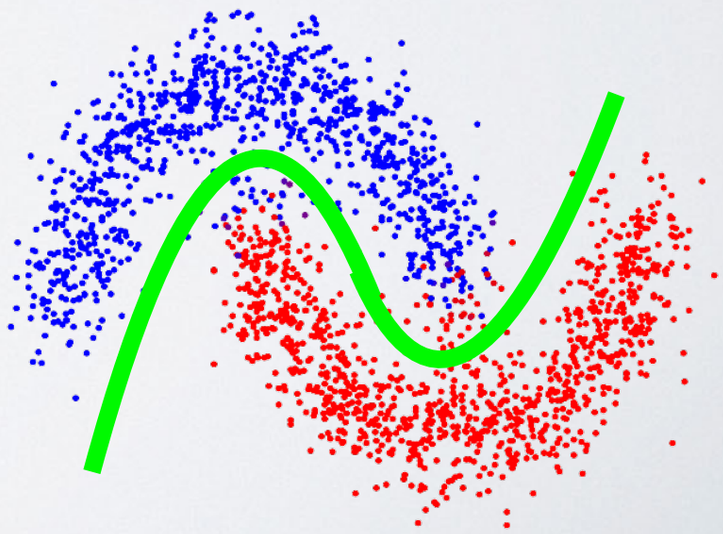
$$K_{ij} = k(x_i, x_j)$$

$$\underset{\alpha}{\text{minimize}} \quad d(\alpha) = \frac{1}{2}\alpha^T Y K Y \alpha - \langle \alpha, \mathbf{1} \rangle$$

$$\text{subject to} \quad 0 \leq \alpha \leq C$$

$$\langle \alpha, l \rangle = 0$$

Return to this later!!



CLASSIFYING TEST DATA

$$\text{maximize} \quad d(\alpha) = -\frac{1}{2}\alpha^T Y X X^T Y \alpha + \langle \alpha, \mathbf{1} \rangle$$

$$\underset{\alpha}{\text{minimize}} \quad d(\alpha) = \frac{1}{2}\alpha^T Y K Y \alpha - \langle \alpha, \mathbf{1} \rangle$$

Recall: $w = X^T Y \alpha$

For a new data point, x

$$w^T x = \langle X^T Y \alpha, x \rangle = \sum y_i \alpha_i \langle x_i, x \rangle = \sum y_i \alpha_i K(x_i, x)$$

POLYNOMIAL KERNEL

$$K(x,y) = (x^T y + 1)^2$$

$$K(x,y) = \left(1 + \sum x_i y_i\right)^2 = 1 + \sum_i x_i y_i + \sum_i x_i^2 y_i^2 + 2\sum_{i \neq j} x_i x_j y_i y_j$$

$$= \langle \phi(x), \phi(y) \rangle$$

$$\phi(x) = (1, x_0, \cdots, x_n, x_0^2, \cdots, x_n^2, \sqrt{2}x_0x_1, \sqrt{2}x_0x_2, \cdots, \sqrt{2}x_{n-1}x_n)$$

Constant

First order

Second order

In general

$$K(x,y) = (x^T y + 1)^p$$

TEXT KERNELS

Based on…

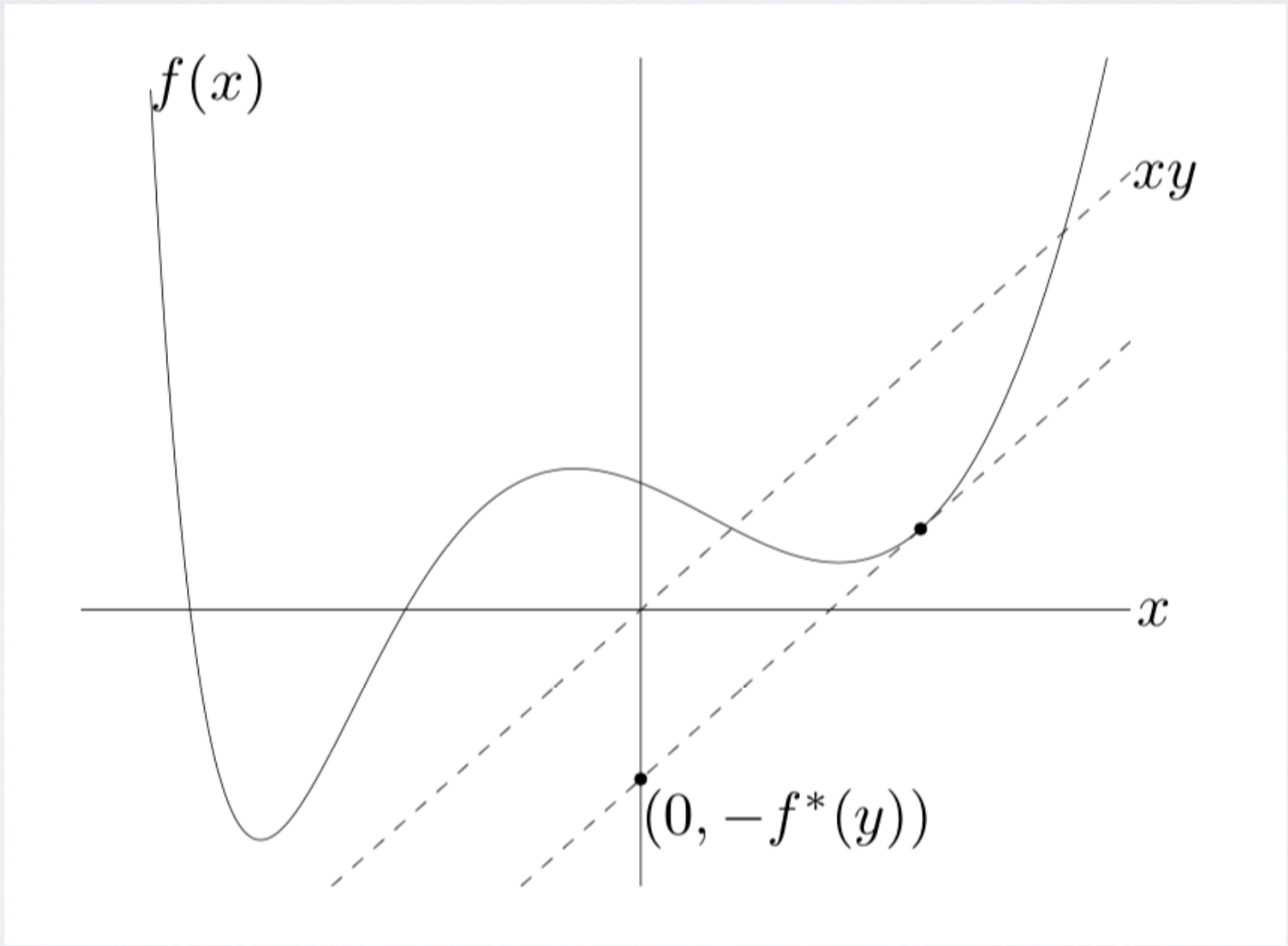
Edit distance between words/paragraphs

Bag of words models

Genome mutations

CONJUGATE FUNCTION

$$f^*(y) = \max_x \; y^T x - f(x)$$



Is it convex?

CONJUGATE OF NORM

$$f(x) = \|x\|$$

$$f^*(y) = \max_x \; y^T x - \|x\|$$

dual norm

$$\|y\|_* \triangleq \max_x y^T x / \|x\|$$

Holder inequality

$$y^T x \leq \|y\|_* \|x\|$$

$$\|y\|_* \leq 1 \longrightarrow f^* = 0$$ why?

$$\|y\|_* > 1 \longrightarrow f^* = \infty$$ why?

$$f^*(y) = \begin{cases} 0, \; \|y\|_* \leq 1 \\ \infty, \; \text{otherwise} \end{cases}$$

EXAMPLES

$$f(x) = \|x\|_2, \qquad f^*(y) = \mathcal{X}_2(y) = \begin{cases} 0, & \|x\|_2 \leq 1 \\ \infty, & \|x\|_2 > 1 \end{cases}$$

$$f(x) = |x|, \qquad f^*(y) = \mathcal{X}_\infty(y) = \begin{cases} 0, & \|x\|_\infty \leq 1 \\ \infty, & \|x\|_\infty > 1 \end{cases}$$

$$f(x) = \|x\|_\infty, \qquad f^*(y) = \mathcal{X}_1(y) = \begin{cases} 0, & |x| \leq 1 \\ \infty, & |x| > 1 \end{cases}$$

HOW TO USE CONJUGATE

$$\text{minimize} \quad g(x) + f(Ax + b)$$

write this as…

$$\begin{aligned} \text{minimize} \quad & g(x) + f(y) \\ \text{subject to} \quad & y = Ax + b \end{aligned}$$

$$\begin{aligned} d(\lambda) &= \min_{x,y} g(x) + f(y) + \langle \lambda, Ax + b - y \rangle \\ &= \min_{x,y} g(x) + \langle \lambda, Ax \rangle + f(y) - \langle \lambda, y \rangle + \langle \lambda, b \rangle \\ &= \langle \lambda, b \rangle - \max_{x,y} -g(x) - \langle A^T \lambda, x \rangle - f(y) + \langle \lambda, y \rangle \end{aligned}$$

$$d(\lambda) = \langle \lambda, b \rangle - g^*(-A^T \lambda) - f^*(\lambda)$$

EXAMPLE: LASSO

$$\text{minimize} \quad \mu|x| + \frac{1}{2}\|Ax - b\|^2$$

how big?

$$\text{minimize} \quad g(x) + f(Ax + b)$$

$$d(\lambda) = \langle \lambda, b \rangle - g^*(-A^T\lambda) - f^*(\lambda)$$

note: $(\mu J)^*(y) = \mu J^*(y/\mu)$

$$\text{maximize} \quad -\langle \lambda, b \rangle - \mathcal{X}_\infty\left(\frac{-1}{\mu}A^T\lambda\right) - \frac{1}{2}\|\lambda\|^2$$

change variables to eliminate negative sign

$$\text{maximize} \quad \langle \lambda, b \rangle - \mathcal{X}_\infty\left(\frac{1}{\mu}A^T\lambda\right) - \frac{1}{2}\|\lambda\|^2$$

EXAMPLE: LASSO

$$\text{minimize} \quad \mu|x| + \frac{1}{2}\|Ax - b\|^2$$

$$\text{maximize} \quad \langle \lambda, b \rangle - \mathcal{X}_\infty \left(\frac{1}{\mu} A^T \lambda \right) - \frac{1}{2}\|\lambda\|^2$$

complete square

$$\text{maximize} \quad -\frac{1}{2}\|\lambda - b\|_2^2 + \frac{1}{2}\|b\|^2$$

$$\text{subject to} \quad \|A^T \lambda\|_\infty \leq \mu$$

EXAMPLE: LASSO

$$\text{minimize} \quad \mu|x| + \frac{1}{2}\|Ax - b\|^2$$

dual problem

$$\text{maximize} \quad -\|\lambda - b\|_2^2 + \frac{1}{2}\|b\|^2$$

$$\text{subject to} \quad \|A^T\lambda\|_\infty \leq \mu$$

Zero is a solution when the optimal objective is

$$\mu|0| + \frac{1}{2}\|A0 - b\|^2 = \frac{1}{2}\|b\|^2$$

This coincides with dual variable $\lambda = b$

$$\mu \geq \|A^T b\|$$

EXAMPLE: LASSO

$$\text{minimize} \quad \mu|x| + \frac{1}{2}\|Ax - b\|^2$$

Solution is zero when $\mu \geq \|A^T b\|$

$$\mu_{guess} = \frac{1}{10}\|A^T b\|$$

EXAMPLE: TV

$$\text{minimize} \quad \frac{1}{2}\|x - f\|^2 + \mu|\nabla x|$$

$$\text{minimize} \quad g(x) + f(Ax)$$

$$f(z) = \mu|z| \quad \longrightarrow \quad f^*(y) = \mathcal{X}_\infty(y/\mu)$$

$$(\mu J)^*(y) = \mu J^*(y/\mu)$$

$$g(z) = \frac{1}{2}\|z - f\|^2 \quad \longrightarrow \quad g^*(y) = \frac{1}{2}\|y + f\|^2 - \frac{1}{2}\|f\|^2$$

form dual problem

$$d(\lambda) = -g^*(-A^T\lambda) - f^*(\lambda)$$

$$\text{maximize} \quad -\frac{1}{2}\|\nabla^T\lambda - f\|^2 - \mathcal{X}_\infty(\lambda/\mu)$$

EXAMPLE: TV

$$\text{maximize} \quad -\frac{1}{2}\|\nabla^T \lambda - f\|^2 - \mathcal{X}_\infty(\lambda/\mu)$$

$$\underset{\lambda}{\text{maximize}} \quad -\frac{1}{2}\|\nabla^T \lambda - f\|^2$$

$$\text{subject to} \quad \|\lambda\|_\infty \leq \mu$$

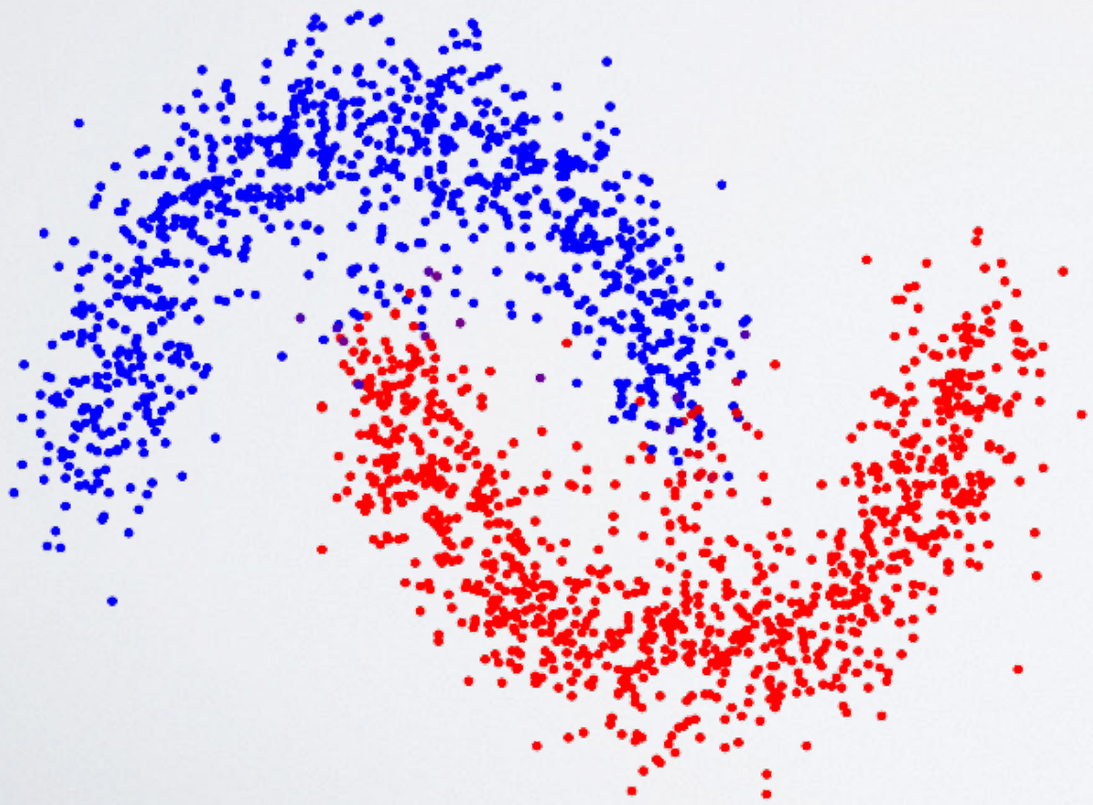
smooth

simple

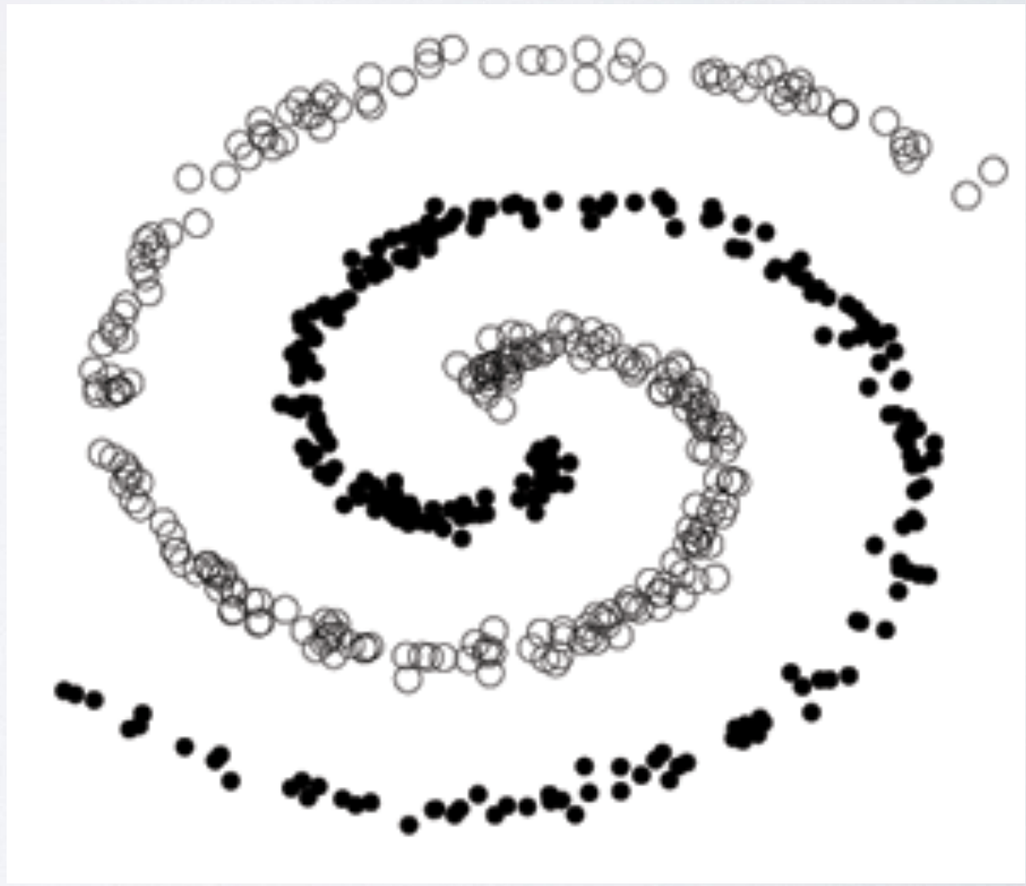
NON-CONVEX PROBLEM: SPECTRAL CLUSTERING

Non-linearly separable classes

Two moons



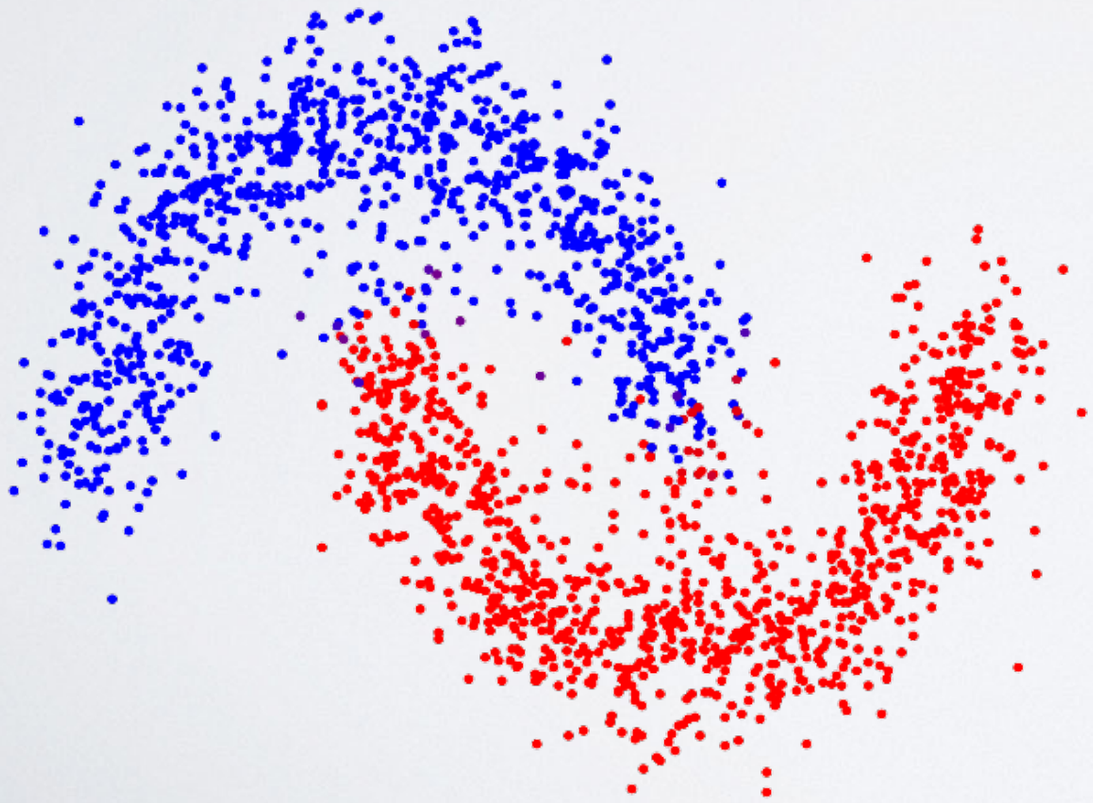
Swiss roll



SPECTRAL CLUSTERING

Non-linearly separable classes

Two moons



(Dis)similarity matrix

$$W_{ij} = \delta - e^{-\|d_i - d_j\|^2/\sigma^2}$$

$$x_i \in \{-1, 1\}$$

Dissimilar = positive

Similar = negative

LABELING PROBLEM

$$\text{minimize} \quad x^T W x = \sum_{i,j} x_i x_j W_{ij}$$

$$\text{subject to} \quad x_i^2 = 1$$

Big W → different labels

Small W → same label

Is this convex? Why? How hard is this problem?

Dual function

$$d(\lambda) = \text{minimize} \quad x^T W x + \langle \lambda, x^2 - 1 \rangle$$

Is this convex? Why? How hard is this problem?

LABELING PROBLEM

Dual function

$$d(\lambda) = \text{minimize} \quad x^T W x + \langle \lambda, x^2 - 1 \rangle$$

$$\sum \lambda_i x_i^2 = x^T \mathtt{diag}(\lambda) x$$

$$d(\lambda) = \text{minimize} \quad x^T (W + \mathtt{diag}(\lambda)) x - \langle \lambda, 1 \rangle$$

$$= \begin{cases} -\langle \lambda, 1 \rangle, \ W + \mathtt{diag}(\lambda) \succeq 0 \\ -\infty, \ \text{otherwise} \end{cases}$$

LABELING PROBLEM

Dual function

$$d(\lambda) = \text{minimize} \quad x^T(W + \texttt{diag}(\lambda))x - \langle \lambda, 1 \rangle$$

$$= \begin{cases} -\langle \lambda, 1 \rangle, & W + \texttt{diag}(\lambda) \succeq 0 \\ -\infty, & \text{otherwise} \end{cases}$$

Pick the dual vector to be smallest constant we can get away with

$$\lambda_i = -\lambda_{\min}(W), \ \forall i$$

$$d(x) = \langle \lambda_{\min}, 1 \rangle$$

smallest eigenvector

$$x = \arg\min \ x^T(W - \lambda_{min} I)x + \langle \lambda_{min}, 1 \rangle = e_{min}$$

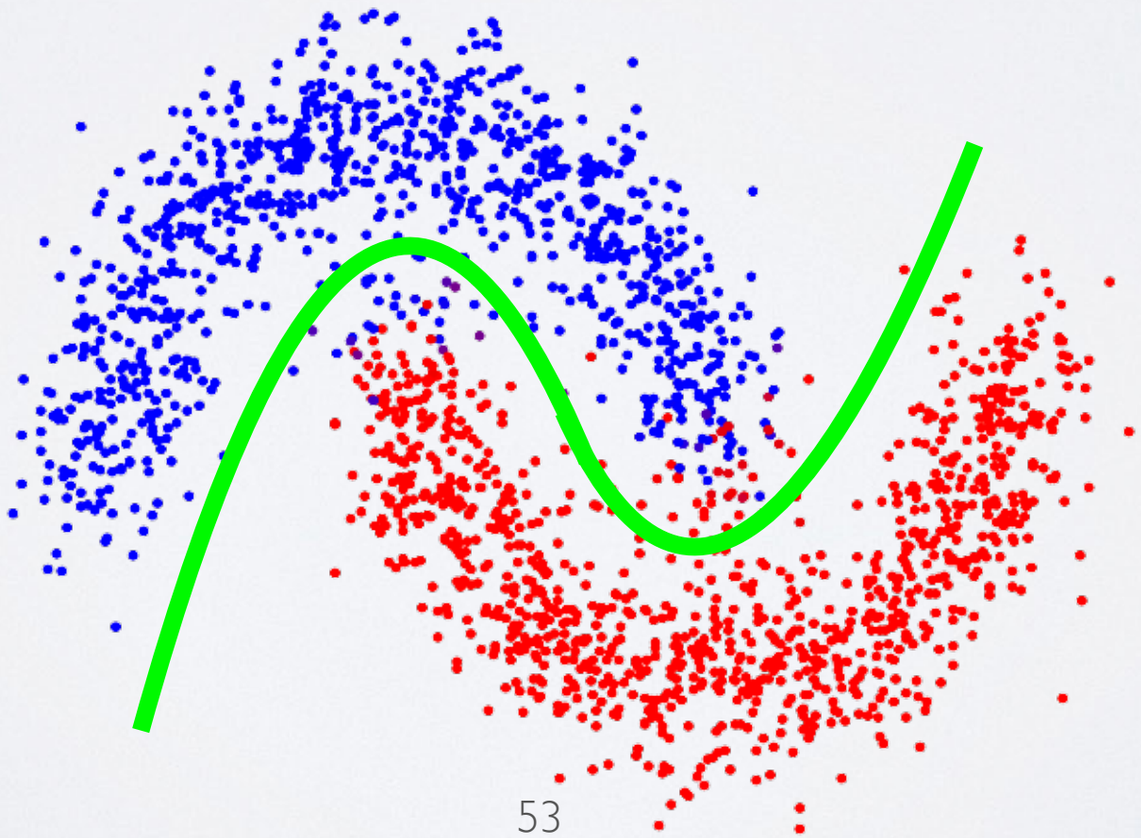
LABELING PROBLEM

Approximate solution

$$x = \arg\min\ x^T(W - \lambda_{min}I)x + \langle \lambda_{min}, 1 \rangle = e_{min}$$

Final step: integer rounding

$$x^{\star}_{approx} = \texttt{round}(e_{min})$$



OVERALL STRATEGY

$$\begin{aligned} \text{minimize} \quad & x^T W x \\ \text{subject to} \quad & x_i^2 = 1 \end{aligned}$$

convex relaxation

Dual function

$$d(\lambda) = \text{minimize} \quad x^T W x + \langle \lambda, x^2 - 1 \rangle$$

Why approximate?

Approximate solution

$$x = \arg\min \; x^T (W - \lambda_{min} I) x + \langle \lambda_{min}, 1 \rangle = e_{min}$$

Final step: integer rounding

$$x^\star_{approx} = \texttt{round}(e_{min})$$

note: we could also definite SIMilarity matrix, and use largest eigenvalue

NEWTON'S METHOD

smooth problem

$$\text{minimize} \quad f(x)$$

quadratic Approximation

$$\text{minimize} \quad \frac{1}{2}(x - x^k)^T H(x - x^k) + \langle x - x^k, g \rangle$$

What about a constrained problem?

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & Ax = b \end{aligned}$$

CONSTRAINED NEWTON

Constrained problem

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & Ax = b \end{aligned}$$

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}(x - x^k)^T H(x - x^k) + \langle x - x^k, g \rangle \\ \text{subject to} \quad & Ax = b \end{aligned}$$

$$L(x, \lambda) = \frac{1}{2}(x - x^k)^T H(x - x^k) + \langle x - x^k, g \rangle + \langle \lambda, Ax - b \rangle$$

KKT Conditions

$$\begin{aligned} H(x - x^k) + g + A^T \lambda &= 0 \\ Ax &= b \end{aligned}$$

CONSTRAINED NEWTON

KKT Conditions

$$H(x - x^k) + g + A^T\lambda = 0$$

$$Ax = b$$

Newton direction $d = x - x^k$

$$\begin{pmatrix} H & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} d \\ \lambda \end{pmatrix} = \begin{pmatrix} -g \\ b - Ax^k \end{pmatrix}$$

Solution is approximate minimizer
Solution satisfies constraints

NEWTON ALGORITHM

- Compute Hessian and gradient
- Solve Newton system

$$\begin{pmatrix} H & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} d \\ \lambda \end{pmatrix} = \begin{pmatrix} -g \\ b - Ax^k \end{pmatrix}$$

- Armijo search

$$f(x^k + \tau d) \leq f(x^k) + \frac{\tau}{10} g^T d$$

- Update iterate

$$x^{k+1} = x^k + \tau d$$

When stepsize=1, constraints are satisfied **exactly**

NEWTON COMPLEXITY

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & Ax = b \end{aligned}$$

Theorem

Suppose the Hessian of f is Lipschitz continuous. Then after a finitely many constrained Newton steps the unit stepsize is an Armijo step, and

$$(\tau = 1)$$

$$\|x^{k+1} - x^\star\| \leq C\|x^k - x^\star\|^2$$

Furthermore, the constraint are exactly satisfied.