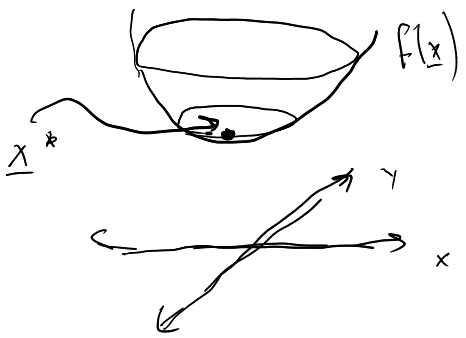
Optimization

Topics: 1) Problem definition, types of problems
2) Convex problems
3) Solution methods
4) SVMs

Boyd & Vandenberghe
Convex Optimization

General problem:

$$\underline{x}^* = \underset{\underline{x} \in D}{\operatorname{argmin}} f(\underline{x}) \quad \text{s.t.} \quad g_j(\underline{x}) \le 0, \; j = 1, 2, \ldots \quad h_k(\underline{x}) = 0, \; k = 1, 2, \ldots$$



- $\underline{x} \in D$: domain (often $\mathbb{R}^n$)
- $f(\underline{x})$: objective fn. (cost, loss, energy, ...)
- $g_j(\underline{x}) \le 0$, $h_k(\underline{x}) = 0$: constraints

Feasible set: Set of $\underline{x} \in D$ that satisfy constraints.

Brute force (grid search): Divide D into grid w/ width δ. # points grows as $(1/\delta)^n$

Type	Domain	Objective	Constraints	Solution
Linear	$\mathbb{R}^n$	$\underline{c}^T \underline{x}$	$A\underline{x} \le \underline{b}$, $\underline{x} \ge 0$	Easy (simplex method)
Integer	$\mathbb{N}^n$	" "	" "	NP-hard in general
Constraint-sat.	$\{0,1\}^n$	Constant	Boolean	NP-hard in general
Convex	$\mathbb{R}^n$	Convex fn.	Convex set	Easy! (interior-point method)
Quadratic	$\mathbb{R}^n$	$\underline{x}^T Q \underline{x} + c^T \underline{x}$	$A\underline{x} \le b$	Easy if Q positive-definite
...				

Ex (least squares) $y = X \cdot \beta$. Given $\{x_i, y_i\}$, $i = 1 \dots P$, find optimal β

$$\beta^* = \underset{\beta}{\operatorname{argmin}} \| X\beta - y \|^2$$

X: $P \times N$, β: $N \times 1$, y: $P \times 1$

$$f(\beta) = (X\beta - y)^T (X\beta - y)$$

$$= \beta^T X^T X \beta - 2 y^T X \beta + \underbrace{y^T y}_{\text{constant, ignore}}$$

$$\beta^* = \underset{\beta}{\operatorname{argmin}} \frac{1}{2} \beta^T Q \beta + c^T \beta, \quad Q = X^T X$$

$$c = -2 y^T X$$

Quadratic problem (convex)

Regularization:

$$f(\beta) = \underbrace{(X\beta - y)^T (X\beta - y)}_{\text{fitting to data}} + \underbrace{\lambda \beta^T \beta}_{\text{penalty on large } \beta_i^2}$$

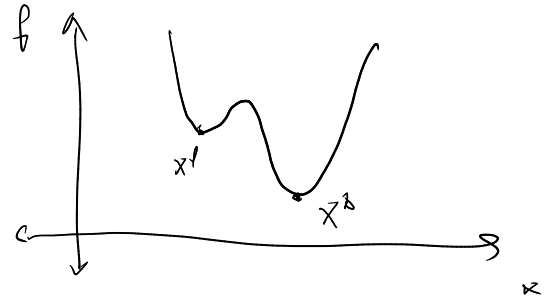
$$= \lambda \beta^T I \beta$$

Same as above with

$$Q \leftarrow Q + \lambda I$$

$\underline{x}^*$ is global optimum. May be local optima $\underline{x}^l$ s.t.

$$f(\underline{x}_l) \le f(\underline{x}) \quad \text{for} \quad \|\underline{x} - \underline{x}^l\| < \varepsilon.$$

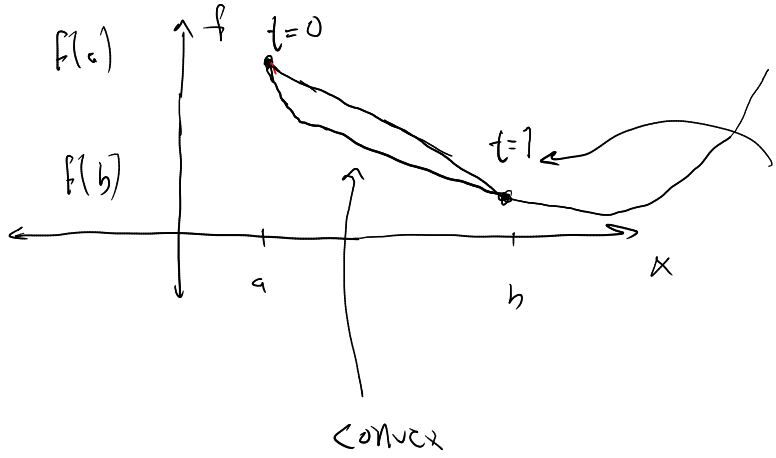


Def: $f(\underline{x})$ is convex if $f(t\underline{a} + (1-t)\underline{b}) \le t f(\underline{a}) + (1-t) f(\underline{b})$

$t \in [0, 1]$

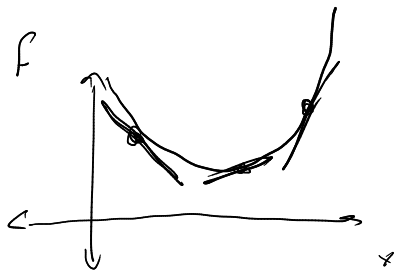
(if $<$, strictly convex)

$t f(a) + (1-t) f(b)$



$\Rightarrow$ If $\nabla f(a)$ is slope at a,

$$f(a) + \nabla f(a)[x - a] < f(x)$$



If f convex, all local min. are global min.

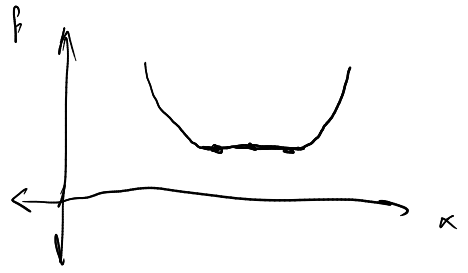
Strictly convex, one global min.

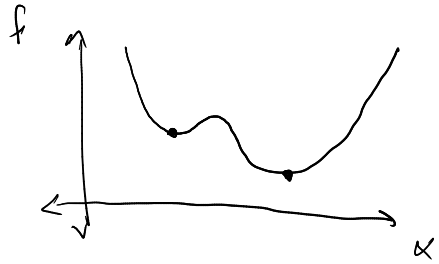
Ex: $f(x) = x^2$. $f(ta + (1-t)b) = t^2a^2 + (1-t)^2 b^2 + 2t(1-t)ab$ (1)

$t f(a) + (1-t) f(b) = ta^2 + (1-t) b^2$ (2)

(1) $-$ (2) $= (t^2 - t) a^2 + ((1-t)^2 - (1-t)) b^2 + 2t(1-t)ab$

$= t(t-1) a^2 + t(t-1) b^2 - 2t(t-1) ab = t(t-1)(a-b)^2 \le 0$





Higher d:

Gradient of $f(\underline{x})$: $\nabla f(\underline{x}) = \begin{pmatrix} \partial f/\partial x_1 \\ \partial f/\partial x_2 \\ \vdots \\ \partial f/\partial x_n \end{pmatrix}$

Hessian: $H(\underline{x}) = \begin{pmatrix} \partial^2 f/\partial x_1^2 & \partial^2 f/\partial x_1 \partial x_2 & \cdots \\ \partial^2 f/\partial x_2 \partial x_2 & \partial^2 f/\partial x_2^2 & \\ \vdots & & \ddots \end{pmatrix}$

1d: $f(x) \approx f(a) + f'(a)(x-a) + \frac{1}{2} f''(a)(x-a)^2$

Minimum: $f' = 0$, $f'' > 0$.

Strictly convex if $f'' > 0$ everywhere.

higher-d: $f(\underline{x}) \approx f(\underline{a}) + \nabla f(\underline{a})(\underline{x}-\underline{a}) + \frac{1}{2}(\underline{x}-\underline{a})^T H(\underline{a})(\underline{x}-\underline{a})$

Minimum: $\nabla f = 0$, H positive definite ($\underline{v}^T H \underline{v} > 0 \ \forall \underline{v}$)

Strictly convex if H positive definite everywhere

Analytical approaches:

Unconstrained problem: look for $\underline{x}^*$ w/ $\nabla f(\underline{x}^*) = 0$

Equality constraints → method of Lagrange multipliers.

$$\min_{\underline{x}} f(\underline{x}) \quad \text{s.t.} \quad g_j(\underline{x}) = 0 \qquad j = 1, 2, \ldots$$

Let $\mathcal{L}(\underline{x}, \underline{\lambda}) = f(\underline{x}) - \sum_j \lambda_j g_j(\underline{x})$

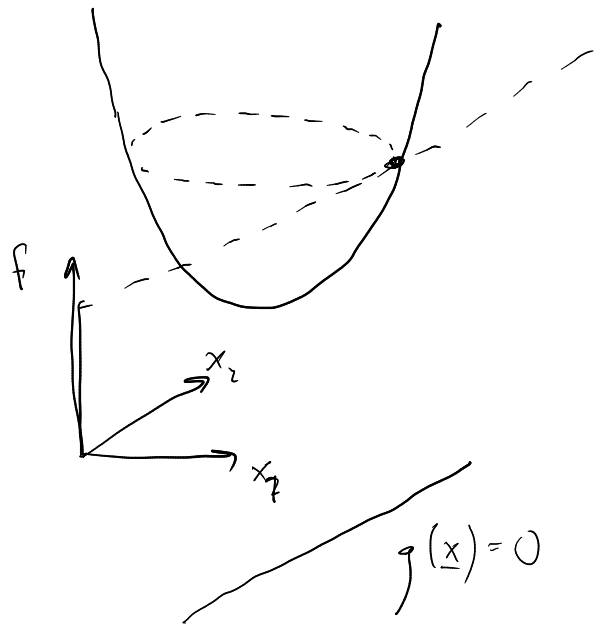
λ_j are Lagrange multipliers.

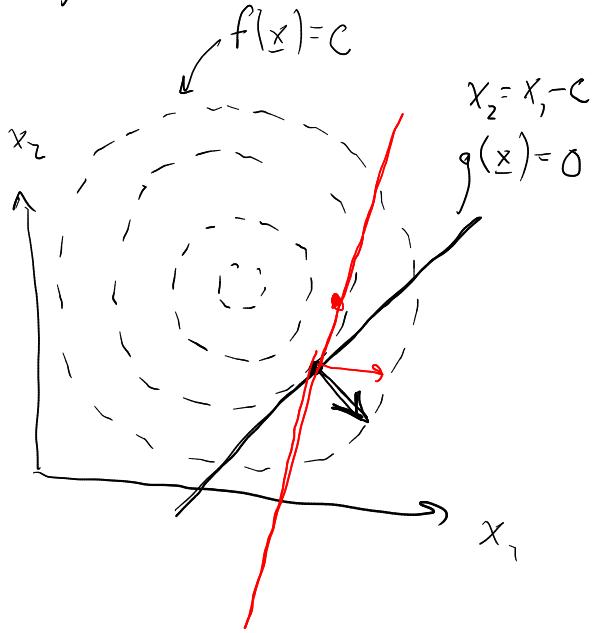
Look for $\underline{x}^*, \underline{\lambda}$ s.t. $\nabla \mathcal{L}(\underline{x}, \underline{\lambda}) = 0$

Note $\frac{\partial}{\partial \lambda_j} \mathcal{L} = -g_j(\underline{x}) = 0$

$$\nabla_{\underline{x}} f(\underline{x}) - \sum_j \lambda_j \nabla_{\underline{x}} g_j(\underline{x}) = 0$$

"level sets"





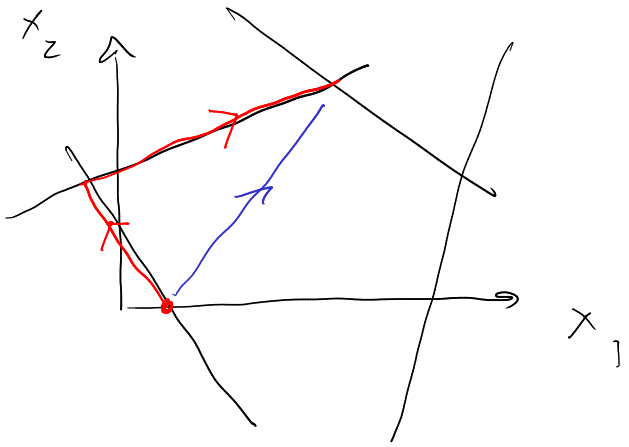
With equality constraints $g_j(\underline{x}) \leq 0$, $h_k(\underline{x}) = 0$.

Karush-Kuhn Tucker (KKT) conditions

Numerical approaches:

Linear problems:

$$\min \underline{c}^T \underline{x} \quad \text{s.t.} \quad A\underline{x} \leq \underline{b}, \quad \underline{x} \geq 0$$



Simplex method.

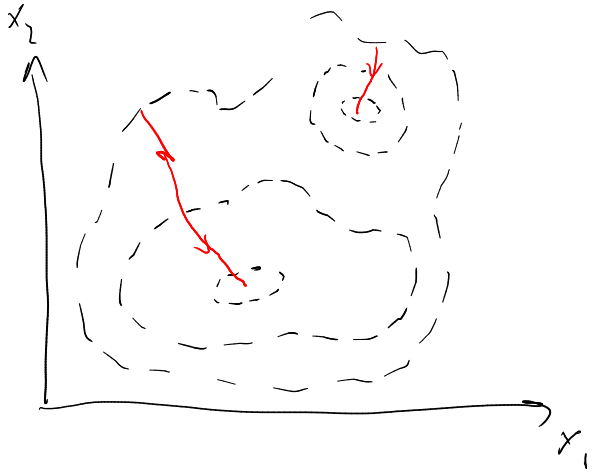
Convex problems: Interior point methods.

Gradient descent:

Given initial value $\underline{x}_0$,

$$\underline{x}_{n+1} = \underline{x}_n - \eta_n \nabla f(\underline{x}_n)$$

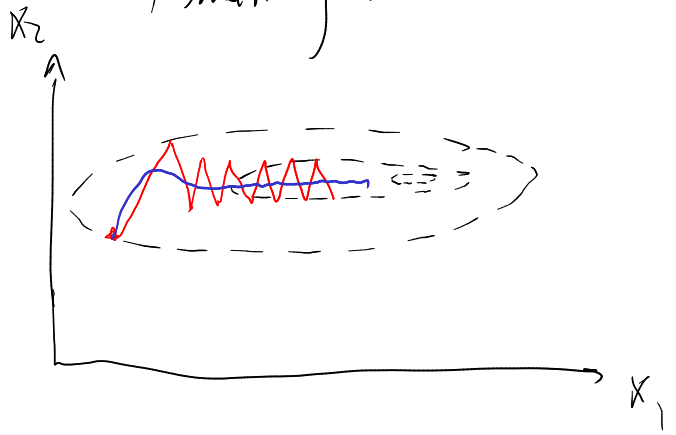
η_n: step size, "learning rate"



Often $\eta_{n+1} < \eta_n$.

"line search": Choose η_n to minimize $f(\underline{x}_{n+1})$.

Problems: 1) multiple minima (multiple initial conditions, noise)
2) small gradients



Momentum: $\underline{z}_{n+1} = \beta \underline{z}_n + \nabla f(\underline{x}_n)$ $\quad \beta = 0.99$

$\underline{x}_{n+1} = \underline{x}_n - \eta_n \underline{z}_{n+1}$

Ex: Last time: Lyapunov fn. for Hopfield net.

$$\mathcal{L}(\underline{x}) = \sum_i \int_0^{x_i} dz\, z F'(z) - F(x_i) I - \frac{1}{2} F(x_i) \sum_j W_{ij} F(x_j)$$

$$\frac{dx_i}{dt} = -\frac{\partial \mathcal{L}}{\partial x_i} F'(x_i) \Rightarrow \frac{d\underline{x}}{dt} = -(\nabla \mathcal{L}) \cdot (\underline{F}'(x_i))$$

If F linear, $F' = 1$, and $\underline{x}(t)$ follows gradient of $\mathcal{L}$.

If $F' > 0$, $\underline{x}(t)$ moves in direction that matches sign of gradient.

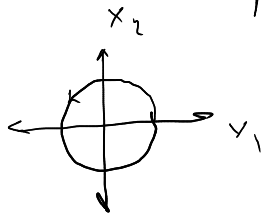
Can we always write ODE as gradient system?

$$\dot{\underline{x}} = -\nabla f(\underline{x})$$

No. Necessary & sufficient:

$$\dot{x}_i = -\frac{\partial}{\partial x_i} f \quad \dot{x}_j = -\frac{\partial}{\partial x_j} f \implies \frac{\partial}{\partial x_j}\dot{x}_i = \frac{\partial}{\partial x_i}\dot{x}_j \quad \forall\, i,j$$

Counterexample: $\dot{\underline{x}} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \underline{x}$ $\quad \dot{x}_1 = x_2 \quad \frac{\partial}{\partial x_2}\dot{x}_1 = 1$ $\quad \dot{x}_2 = -x_1 \quad \frac{\partial}{\partial x_1}\dot{x}_2 = -1$



Can write Lyapunov function $\mathcal{L}(t) = \frac{1}{2}\|\nabla f\|^2 \implies \frac{d\mathcal{L}}{dt} = \sum_i \frac{\partial \mathcal{L}}{\partial x_i}\frac{\partial x_i}{\partial t}$

$$= \frac{1}{2}\sum_i \left(\frac{\partial f}{\partial x_i}\right)^2$$

$$= -\sum_i \left(\frac{\partial f}{\partial x_i}\right)^2 = -\sum_i (\dot{x}_i)^2$$

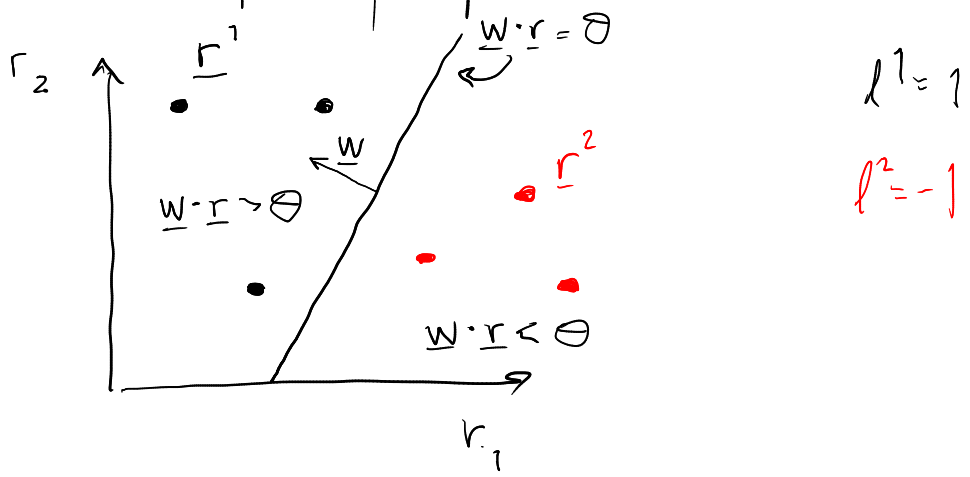
Gradient system cannot have periodic orbits:

If $\underline{x}(0) = \underline{x}(T)$, $\mathcal{L}(0) = \mathcal{L}(T)$. But

$$\mathcal{L}(T) = \int_0^T \frac{d\mathcal{L}}{dt}\,dt = \int_0^T -\sum_i (\dot{x}_i)^2 dt \implies \dot{x}_i = 0$$

Convex optimization example: Support vector machines (SVM)

Problem: Given P patterns $\underline{r}^\mu$, $\mu = 1 \dots P$, and labels $\ell^\mu = \pm 1$, find linear separating hyperplane that optimally separates $\ell = 1$ and $\ell = -1$.



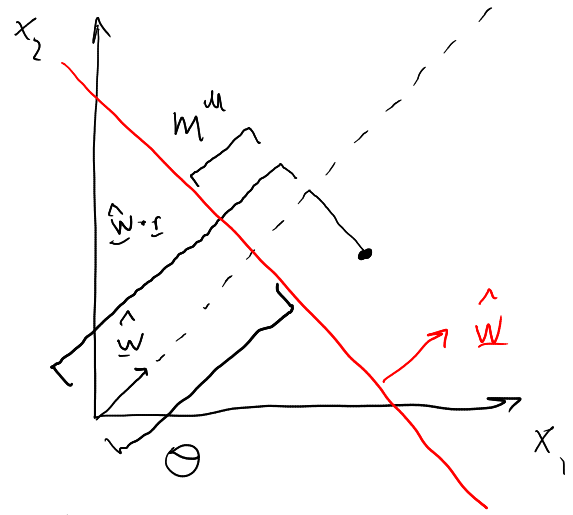
Classifier: $\ell = \text{sign}\left(\underline{w} \cdot \underline{r} - \theta\right)$

($\underline{w}$: weights; θ: threshold)

How to choose optimal $\underline{w}, \theta$? May be multiple valid solutions (draw).

Idea: Maximize margin m^μ (smallest distance from $\underline{r}^\mu$ to boundary). (draw).

If $\|\underline{w}\| = 1$, then $m^\mu = |\hat{\underline{w}} \cdot \underline{r}^\mu - \theta|$



$\underline{w}$ is perpendicular to separating hyperplane.

Optimization problem:

$$\underbrace{m^\mu}$$

$$\underset{\underline{w}}{\text{maximize}} \quad \min_\mu \left| \underline{w} \cdot \underline{r}^\mu - \theta \right| \quad \text{s.t.} \quad \|\underline{w}\| = 1,$$

$$\text{sign}\left(\underline{w} \cdot \underline{r}^\mu - \theta\right) = \ell^\mu$$

Redefine constraints: $\left(\underline{w} \cdot \underline{r}^\mu - \theta\right) \cdot \ell^\mu > 0.$

How to deal with $\|\underline{w}\|$?

If margin $= m$, $\left(\underline{w} \cdot \underline{r}^\mu - \theta\right) \ell^\mu \geq m$. Divide by m:

$$\left(\frac{\underline{w}}{m} \cdot \underline{r}^\mu - \frac{\theta}{m}\right) \ell^\mu \geq 1$$

Reparameterize: $\tilde{\underline{w}} \leftarrow \frac{\underline{w}}{m}$

$$\tilde{\theta} = \frac{\theta}{m}$$

$$\left(\tilde{\underline{w}} \cdot \underline{r}^\mu - \tilde{\theta}\right) \ell^\mu \geq 1.$$

Note: $\|\tilde{\underline{w}}\| = \left\|\frac{\underline{w}}{m}\right\| = \frac{1}{m}$

Maximize margin $\Longleftrightarrow$ minimize $\|\tilde{\underline{w}}\|^2$!

$$\underline{w}^*, \theta^* = \underset{\underline{w}, \theta}{\operatorname{argmin}} \ \underline{w}^T \underline{w} \quad \text{s.t.} \quad \left(\underline{w} \cdot \underline{r}^{(\mu)} - \theta\right) \ell^{(\mu)} \geq 1.$$

$\Rightarrow \|\underline{w}^*\| = \frac{1}{m}$.

Properties: 1) $\underline{w}$ determined only by closest points (those on margin) — support vectors.

2) Binary classification — linear boundary

3) Fully supervised ($\ell^{(\mu)}$ know $\forall \mu$)

4) Soln only exists if data linearly separable

Extensions: 1) Multi-class

2) Kernel SVM (later)

3) "Soft margin"

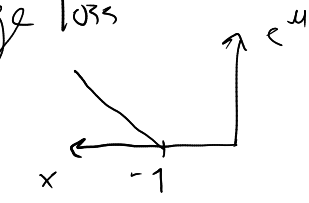
↓

Allow misclassifications:

Penalize w/ $c^{(\mu)} = \max\left(0, 1 - \ell^{(\mu)}\left(\underline{w} \cdot \underline{r}^{(\mu)} - \theta\right)\right)$

$$\min \sum_{\mu=1}^{P} c^{(\mu)} + \lambda \|\underline{w}\|^2$$

"hinge loss"



How to write as convex problem?

Rewrite constraints: $(\underline{w} \cdot \underline{r}^u - \Theta) l^u \geq 1 - c^u$

$$\underline{w}^*, \underline{c}^*, \Theta^* = \underset{\underline{w}, \underline{c}, \Theta}{\operatorname{argmin}} \sum_{u=1}^{p} c^u + \lambda \underline{w}^T \underline{w} \quad \text{s.t.} \quad (\underline{w} \cdot \underline{r}^u - \Theta) l^u \geq 1 - c^u$$

Interpretation of Lagrange mult:

$$\mathcal{L} = f(\underline{x}) + \lambda g(\underline{x}). \qquad \text{Let } g(\underline{x}) = \tilde{g}(\underline{x}) - c = 0$$

$$\Rightarrow \frac{\partial \mathcal{L}}{\partial c} = \lambda. \quad \lambda \text{ is sensitivity of } \mathcal{L} \text{ to change in constraint}$$

For SVMs, $\lambda \neq 0$ only for SVs (on margin).

Duality: "primal problem"

$$\underline{x}^* = \underset{\underline{x}}{\operatorname{argmin}} f(\underline{x}), \qquad \begin{aligned} g_j(\underline{x}) &\leq 0 \\ h_k(\underline{x}) &= 0 \end{aligned}$$

Write Lagrangian

$$\mathcal{L}(x, \underline{\lambda}, \underline{\nu}) = f(\underline{x}) + \sum_j \lambda_j g_j(\underline{x}) + \sum_k \nu_k h_k(\underline{x})$$

Dual function:

$$G(\underline{\lambda}, \underline{\nu}) = \inf_{\underline{x} \in D} \mathcal{L}(\underline{x}, \underline{\lambda}, \underline{\nu})$$

Note $G(\underline{\lambda}, \underline{\nu}) \leq f^*$ if $\lambda_j \geq 0 \ \forall j$.

Why? For any <u>feasible</u> $\underline{x}$, $g_k(\underline{x}) \leq 0$, $h_k(\underline{x}) = 0$

$$\Rightarrow \underbrace{\sum_j \lambda_j g_j(\underline{x})}_{\leq 0} + \underbrace{\sum_k \nu_k h_k(\underline{x})}_{= 0} \leq 0$$

$$\Rightarrow \mathcal{L}(\underline{x}, \underline{\lambda}, \underline{\nu}) \leq f(\underline{x})$$

Can minimize G "dual problem":

$$\underline{\lambda}^*, \underline{\nu}^* = \underset{\underline{\lambda}, \underline{\nu}}{\operatorname{argmax}} \ G(\underline{\lambda}, \underline{\nu}) \quad \text{s.t.} \quad \lambda_j \geq 0$$

If primal problem convex, $G^* = f^*$

For SVM,

$$\mathcal{L} = \frac{1}{2} \underline{w}^T \underline{w} - \sum_\mu \lambda^\mu \left[(\underline{w} \cdot \underline{r}^\mu - \theta) l^\mu - 1 \right]$$

$$G(\underline{\lambda}) = \inf_{\underline{w}, \theta} \mathcal{L} \qquad \frac{\partial \mathcal{L}}{\partial w_i} = 0 \Rightarrow \underline{w} - \sum_\mu \lambda_\mu \underline{r}^\mu l^\mu = 0$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = 0 \Rightarrow \sum_\mu \lambda^\mu l^\mu = 0$$

$$\underline{w} = \sum_\mu \lambda_\mu \underline{r}^\mu l^\mu \leftarrow \text{sum of SVs}$$

Dual problem:

$$\max -\frac{1}{2}\sum_{u}\sum_{v} \lambda_u \lambda_v l_u l_v \left(\underline{r}^u\right)^T \underline{r}^v + \sum_u \lambda^u$$

$$\text{s.t.} \sum_u \lambda_u l_u = 0, \quad \lambda_u \geq 0.$$

Optimization over P variables vs. n.