Optimization

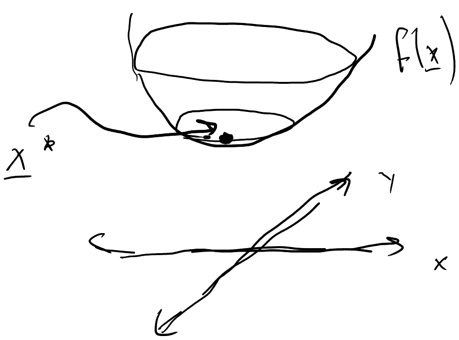
Topics: 1) Problem definition, types of problems
2) Convex problems
3) Solution methods
4) SVMs

Boyd & Vandenberghe
<u>Convex Optimization</u>

Green notes are extra examples and extensions not covered in class.

General problem:

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



domain (often $\mathbb{R}^n$)

objective fn. (cost, loss, energy, ...)

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

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

$$\beta^* = \underset{\beta}{\operatorname{argmin}} \| \underset{P \times N}{X} \underset{N \times 1}{\beta} - \underset{P \times 1}{y} \|^2$$

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

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

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

$$\underline{c} = -2 y^T X$$

Quadratic problem (convex)

Regularization:

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

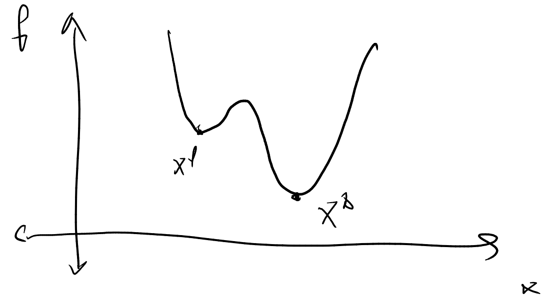
$$= \lambda \beta^T I \beta$$

Same as above with

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

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

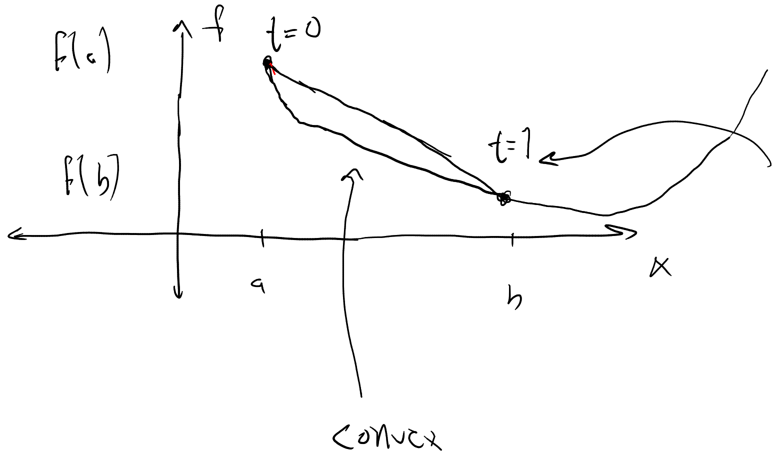
$$f(\underline{x}_l) \leq f(\underline{x}) \text{ for } \|\underline{x} - \underline{x}^l\| < \varepsilon.$$



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

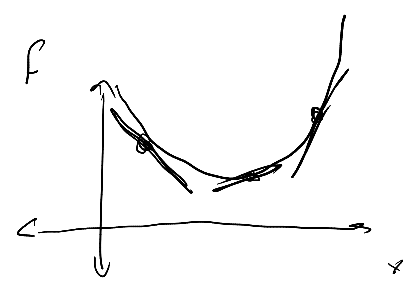
$t \in [0, 1]$

(if $<$, strictly convex)



$\Longrightarrow$ 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.

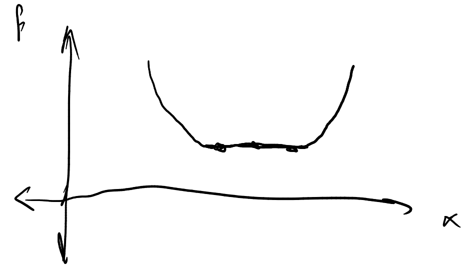
strictly convex, one global min.

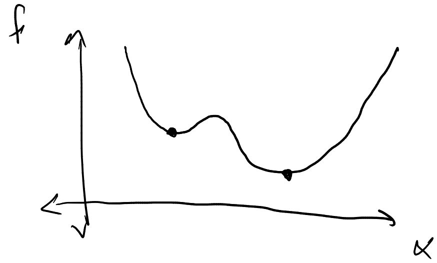
Ex: $f(x) = x^2$. $\quad f(ta + (1-t)b) = t^2 a^2 + (1-t)^2 b^2 + 2t(1-t)ab$ (1)

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

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

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





Higher d:

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

<u>Hessian</u>:

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

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.

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

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

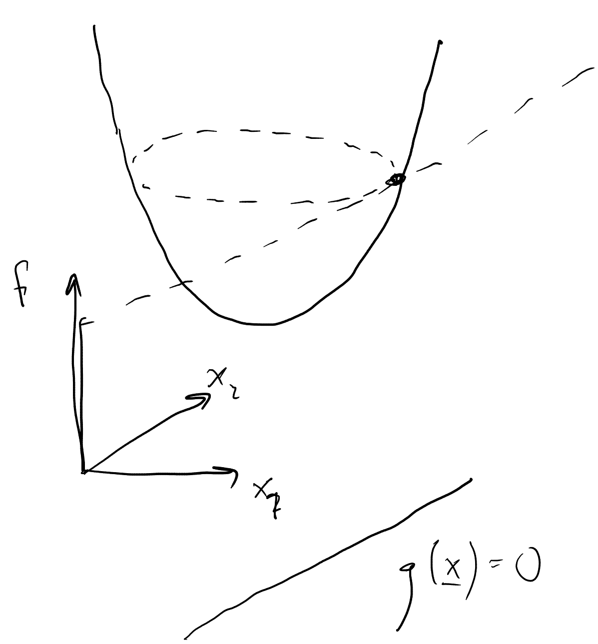
λ_j are Lagrange multipliers.

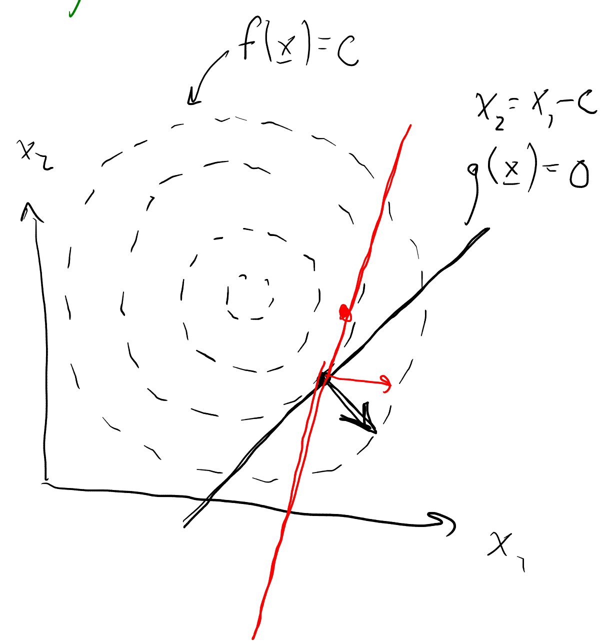
Look for $\underline{x}^*, \underline{\lambda}$ s.t. $\nabla \mathcal{L}(\underline{x}, \underline{\lambda}) = 0$

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"





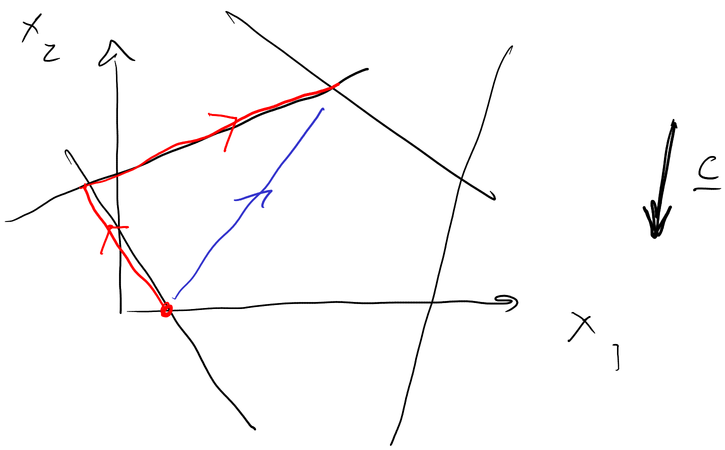
With inequality constraints $g_j(\underline{x}) \le 0$, $h_k(\underline{x}) = 0$.

Karush-Kuhn Tucker (KKT) conditions

Numerical approaches:

Linear problems:

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

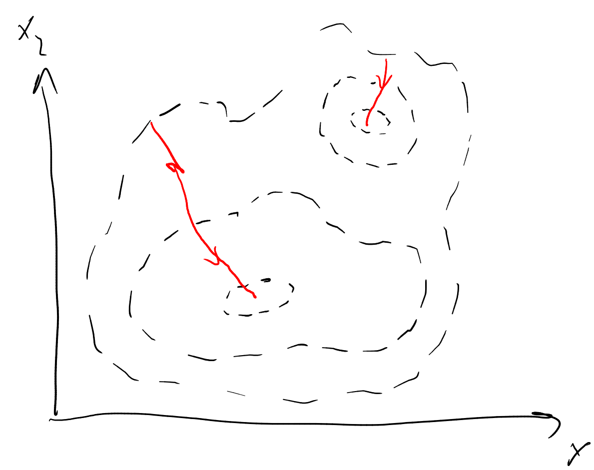


Simplex method.

Convex problems: Interior point methods.

Gradient descent:

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



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

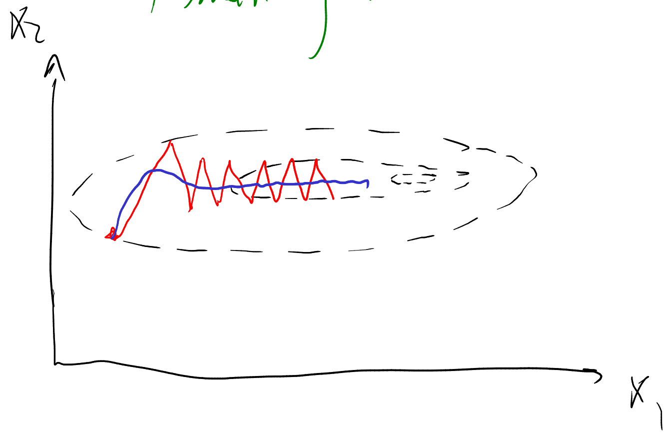
step size "learning rate"

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

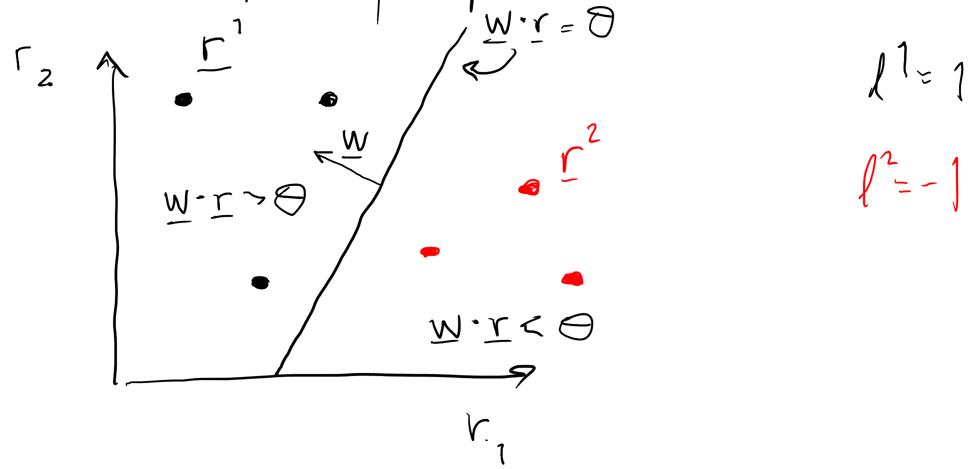


Momentum: $\underline{z}_{n+1} = \beta \underline{z}_n + \nabla f(\underline{x}_n)$ $\qquad \beta = 0.99$

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

Convex optimization example: Support vector machines (SVM)

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



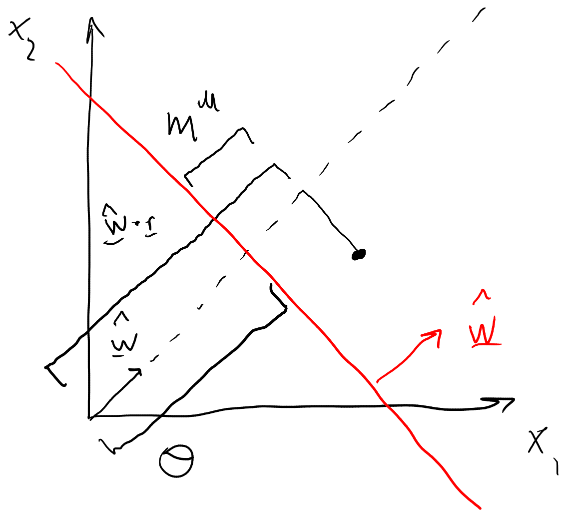
Classifier: $\ell = \text{sign}(\underline{w} \cdot \underline{r} - \theta)$

weights $\uparrow$ ($\underline{w}$)　　threshold $\uparrow$ (θ)

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

Idea: Maximize <u>margin</u> 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:

$$\underset{\underline{w}}{\text{maximize}} \quad \min_{\mu} \overbrace{\left| \underline{w} \cdot \underline{r}^\mu - \theta \right|}^{m^\mu} \quad \text{s.t.} \quad \|\underline{w}\| = 1,$$
$$\operatorname{sign}(\underline{w} \cdot \underline{r}^\mu - \theta) = \ell^\mu$$

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

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

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

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

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

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

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

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

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

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

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

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

2) Binary classification — linear boundary

3) Fully supervized (l^μ know $\forall \mu$)!

4) Sol'n only exists if data linearly seperable

Extensions: 1) Multi-class

2) Kernel SVM (later)

3) "Soft margin"

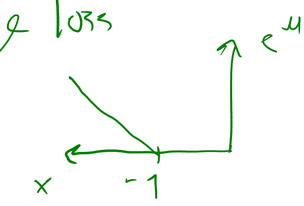
$\downarrow$

Allow misclassifications:

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

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

"hinge loss"



How to write as convex problem?

Rewrite constraints: $(\underline{w} \cdot \underline{r}^\mu - \theta)\, l^\mu \geq 1 - c^\mu$

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

Interpetation of Lagrange mult.:

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

$$\Longrightarrow \frac{\partial \mathcal{L}}{\partial c} = \lambda.$$ λ is sensitivity of $\mathcal{L}$ to change in constraint.

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

Duality: "primal problem"

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

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:

$$G(\underline{\lambda}, \underline{\nu}) = \underset{\underline{x} \in D}{\inf} \mathcal{L}(\underline{x}, \underline{\lambda}, \underline{\nu}) \qquad (\approx \min)$$

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

Why? For any __feasible__ $\underline{x}$, $g_k(\underline{x}) \leq 0$, $h_k(\underline{x}) = 0$

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

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

Can minimize G "dual problem":

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

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:

$$\max -\frac{1}{2}\sum_\mu \sum_\nu \lambda_\mu \lambda_\nu l_\mu l_\nu (\underline{r}^\mu)^T \underline{r}^\nu + \sum_\mu \lambda^\mu$$

$$\text{s.t. } \sum_\mu \lambda_\mu l_\mu = 0 , \quad \lambda_\mu \geq 0.$$

Optimization over P variables vs. n.