Function Approximation Methods-V

CS771: Introduction to Machine Learning

Purushottam Kar

Outline of today's discussion

PART I

- Revisit Lagrangian duality
- Apply duality to SVMs
- Observe a beautiful connection b/w CD and SGD
- See how some state-of-the-art solvers for SVMs are built

PART II

- Learn how to model data (data features that is)
- Look at the naïve Bayes technique for supervised problems
- Get introduced to the GMM technique for unsupervised problems

Duality

How to take the problem you want to solve
… and convert it to a problem you can solve

Fenchel, Lagrange, Wolfe, Pontryagin

Duality

How to take the problem you want to solve
… and convert it to a problem you can solve

Fenchel, Lagrange, Wolfe, Pontryagin

Duality

How to take the problem you want to solve
… and convert it to a problem you can solve

The Lagrangian

The Lagrangian

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$

The Lagrangian

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$

$$\hat{\mathbf{w}} = \arg\min \sum_{i=1}^{n} \left(y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle\right)^2$$

$$s.t. \ \|\mathbf{w}\|_2 \leq r$$

The Lagrangian

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$

The Lagrangian

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$

Want $g(\mathbf{w}) \geq 0$?
$-g(\mathbf{w}) \leq 0$

The Lagrangian

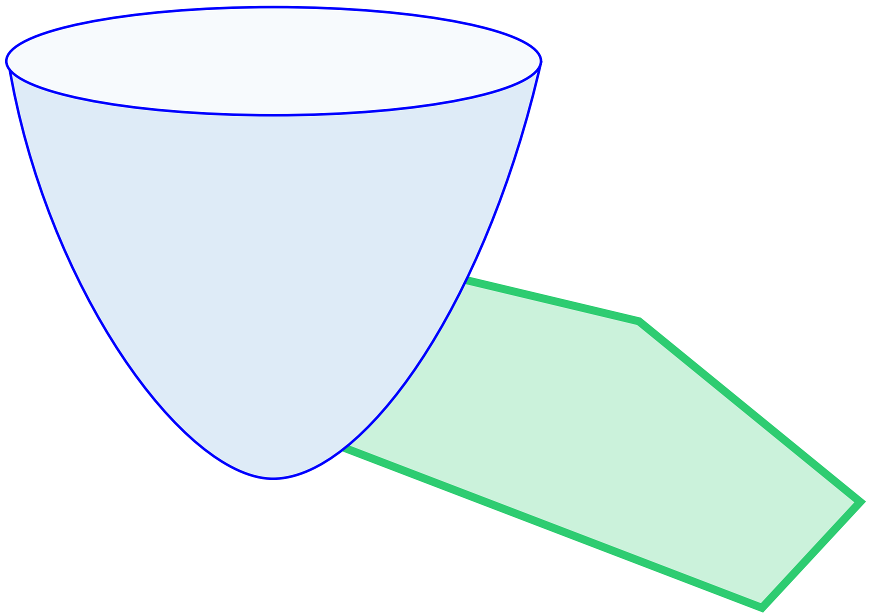
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \; f(\mathbf{w})$$

$$\text{s.t.} \; g(\mathbf{w}) \leq 0$$

The Lagrangian

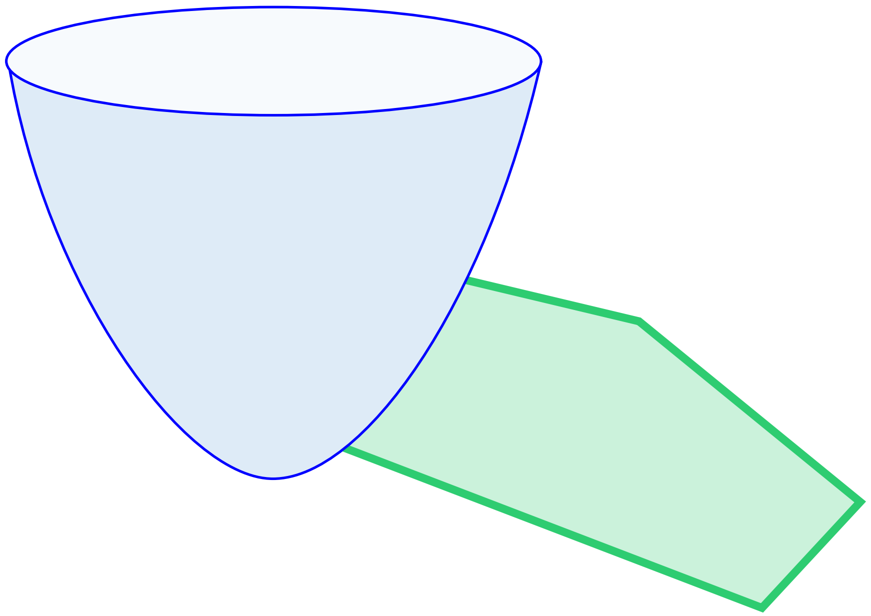
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$



The Lagrangian

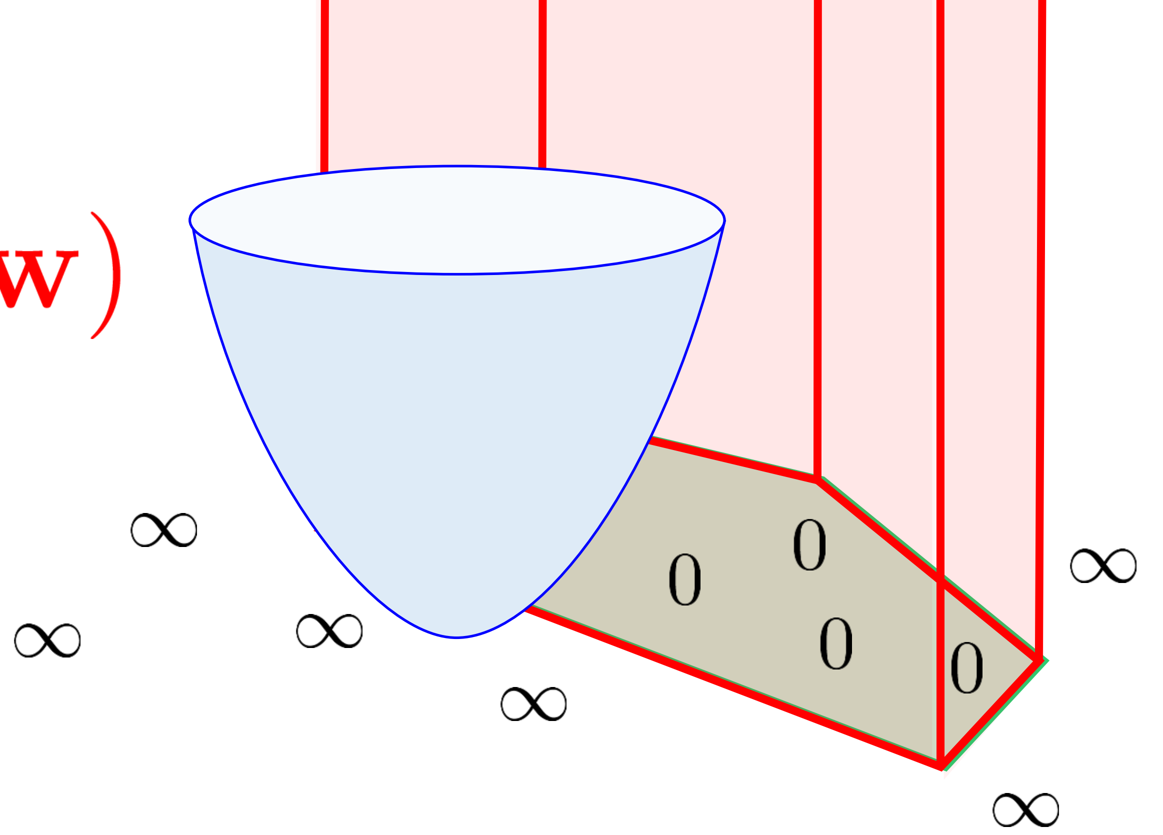
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}) + r(\mathbf{w})$$



The Lagrangian

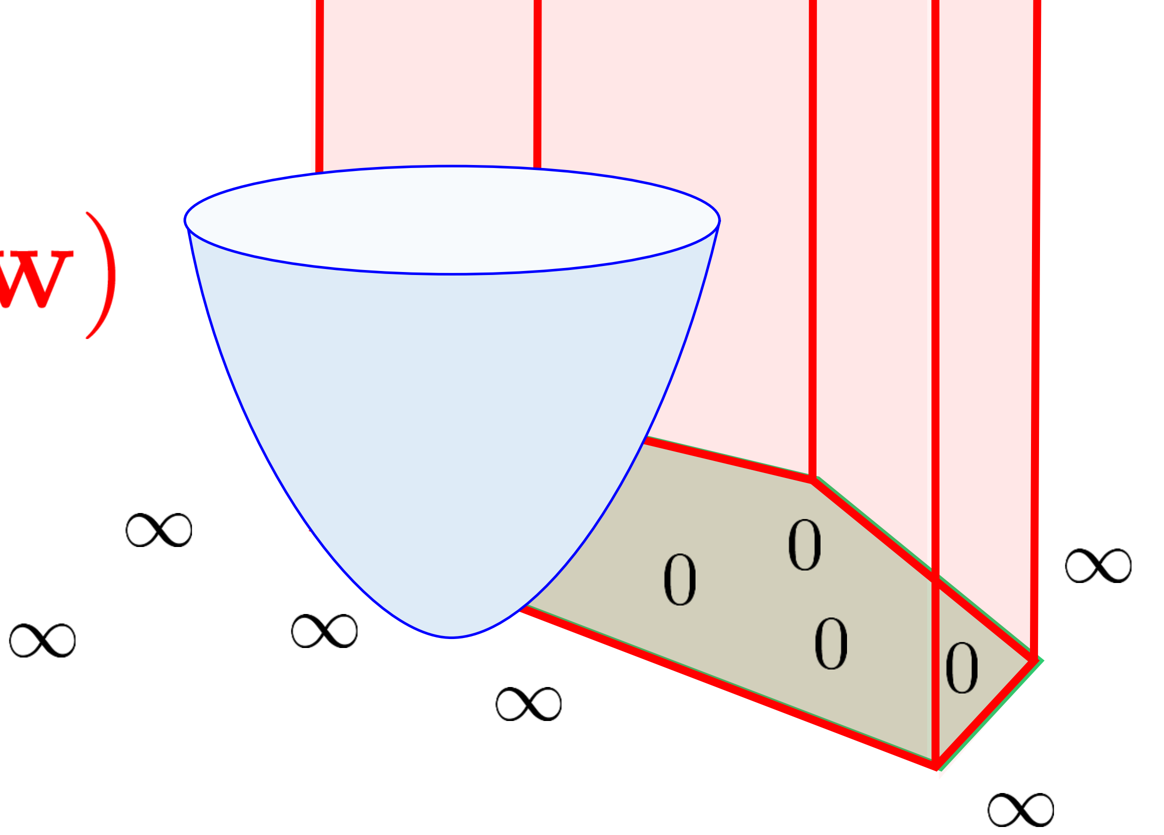
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w}) + r(\mathbf{w})$$

Barrier Function



The Lagrangian

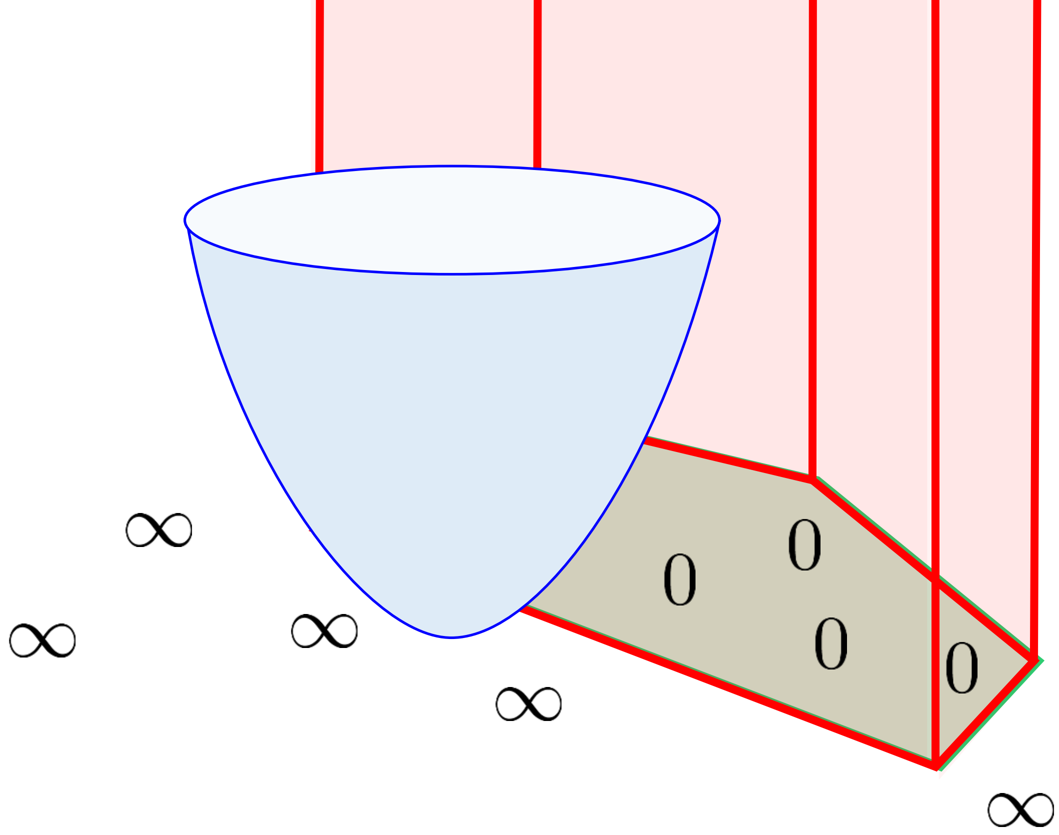
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}) + r(\mathbf{w})$$



The Lagrangian

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

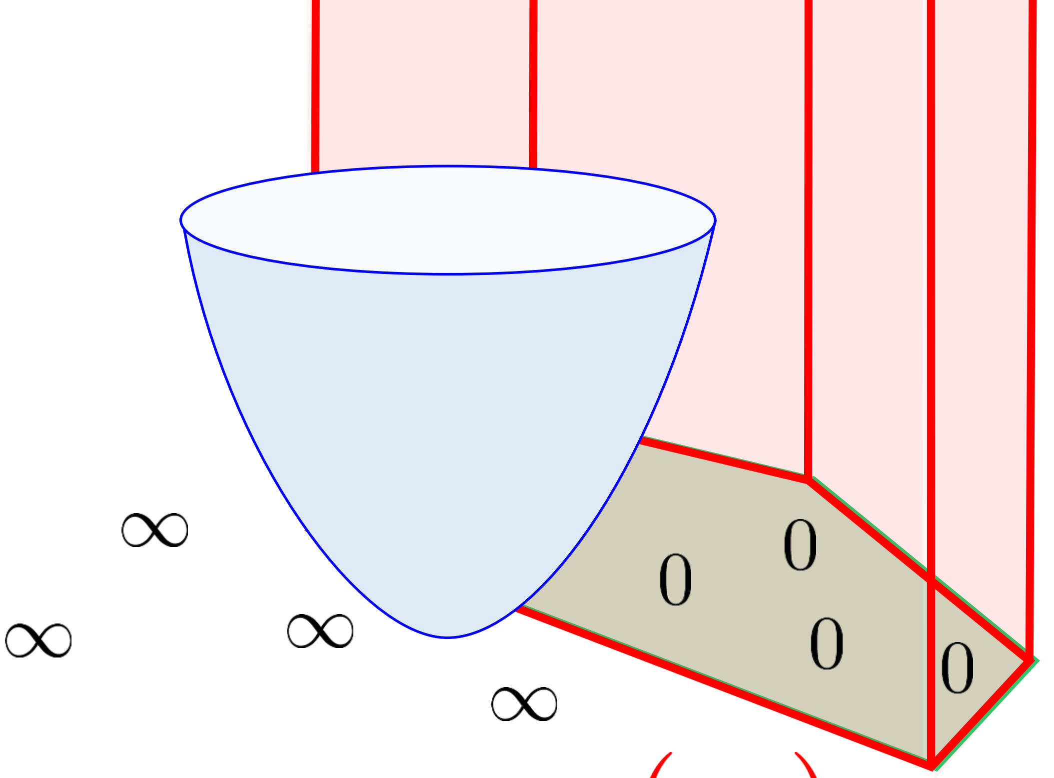
$$\text{s.t. } g(\mathbf{w}) \leq 0$$



The Lagrangian

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$

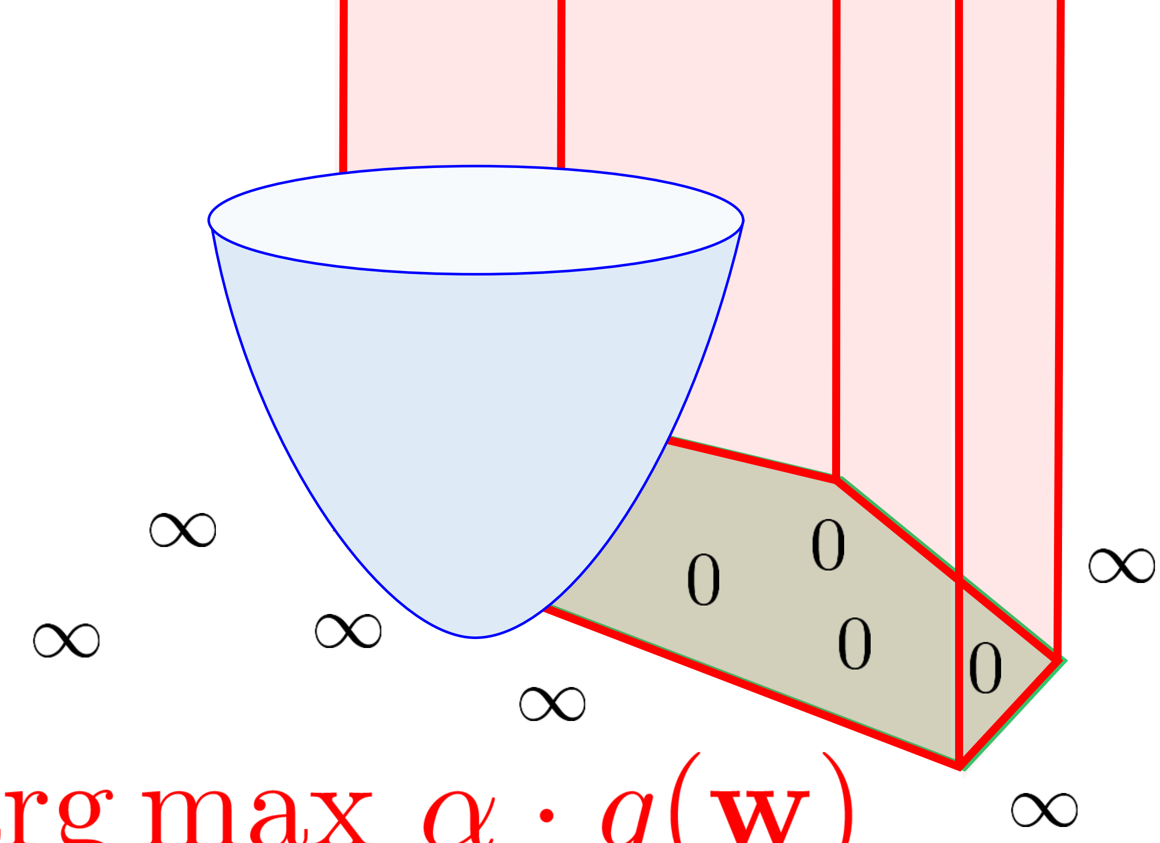


$$\arg\max_{\alpha \geq 0} \alpha \cdot g(\mathbf{w})$$

The Lagrangian

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$



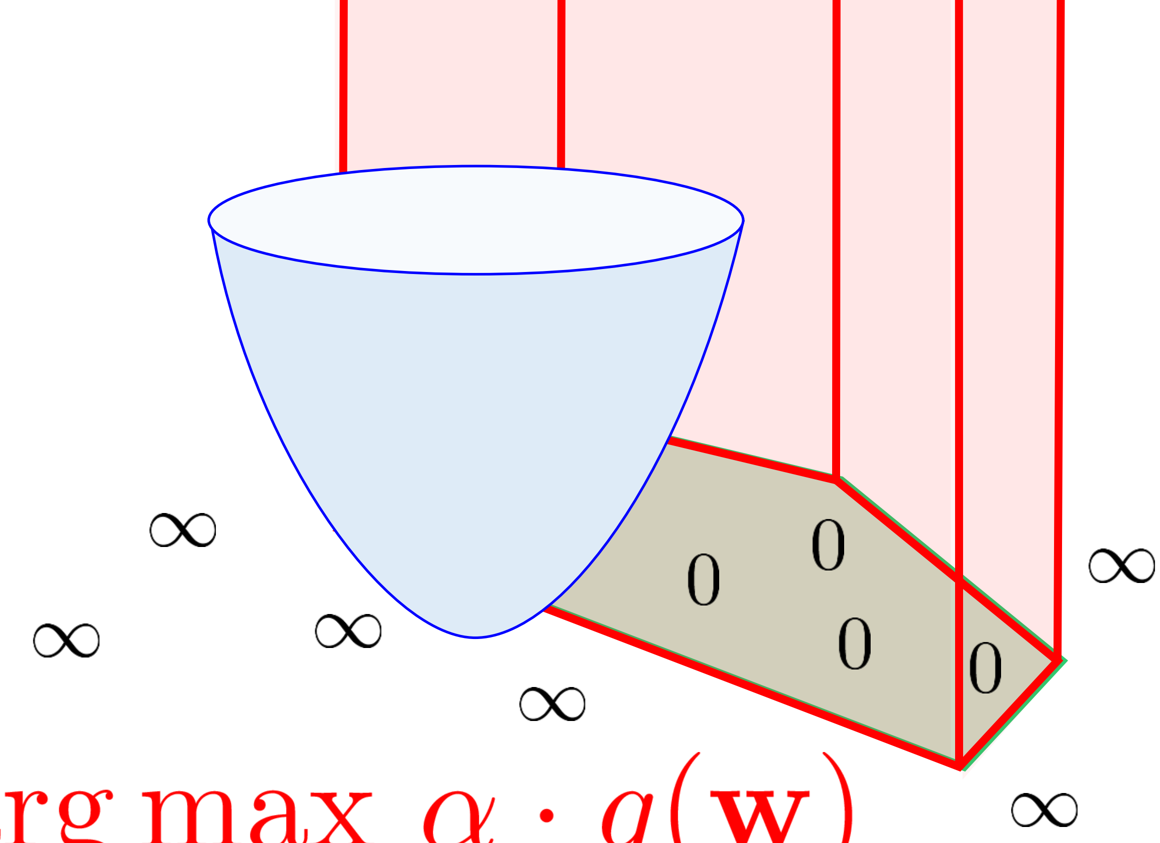
$$\arg\max_{\alpha\geq 0} \alpha \cdot g(\mathbf{w})$$

∞ if $g(\mathbf{w}) > 0$

The Lagrangian

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$



$$\arg\max_{\alpha\geq 0} \alpha \cdot g(\mathbf{w})$$

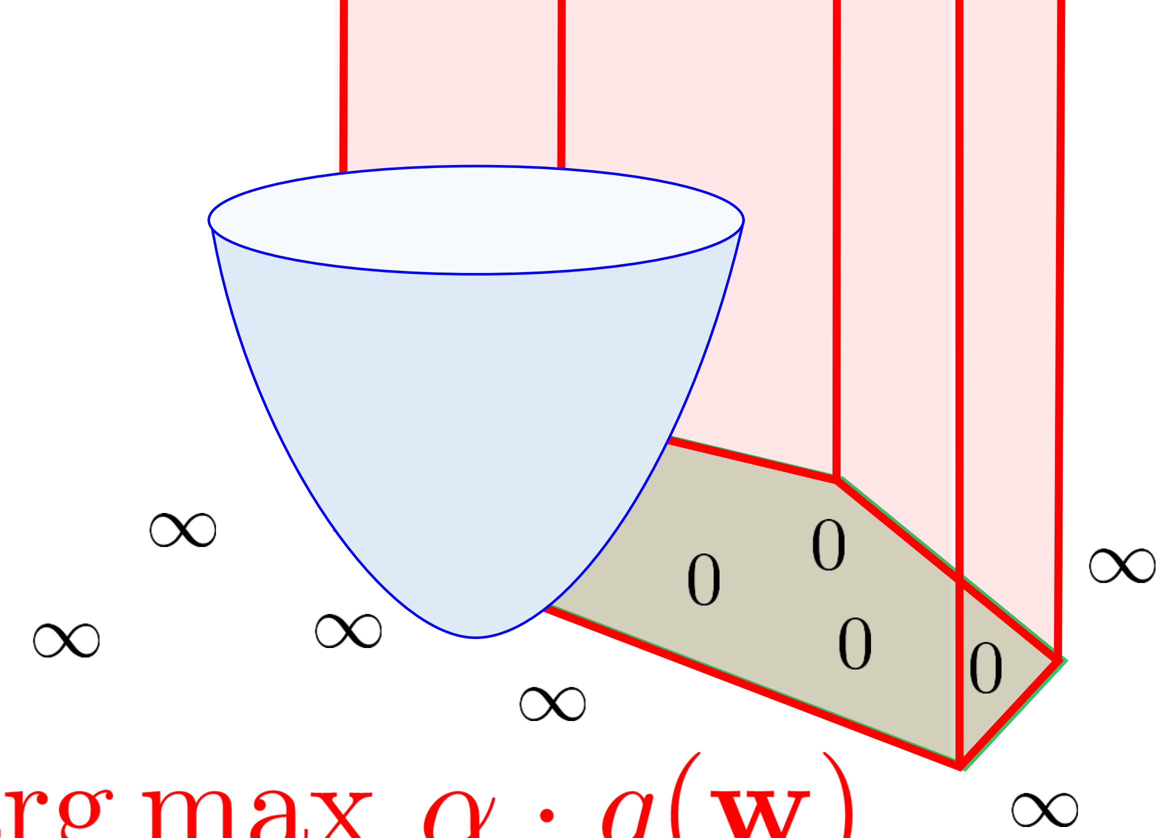
∞ if $g(\mathbf{w}) > 0$

0 if $g(\mathbf{w}) \leq 0$

The Lagrangian

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$



$$f(\mathbf{w}) + \arg\max_{\alpha \geq 0} \alpha \cdot g(\mathbf{w})$$

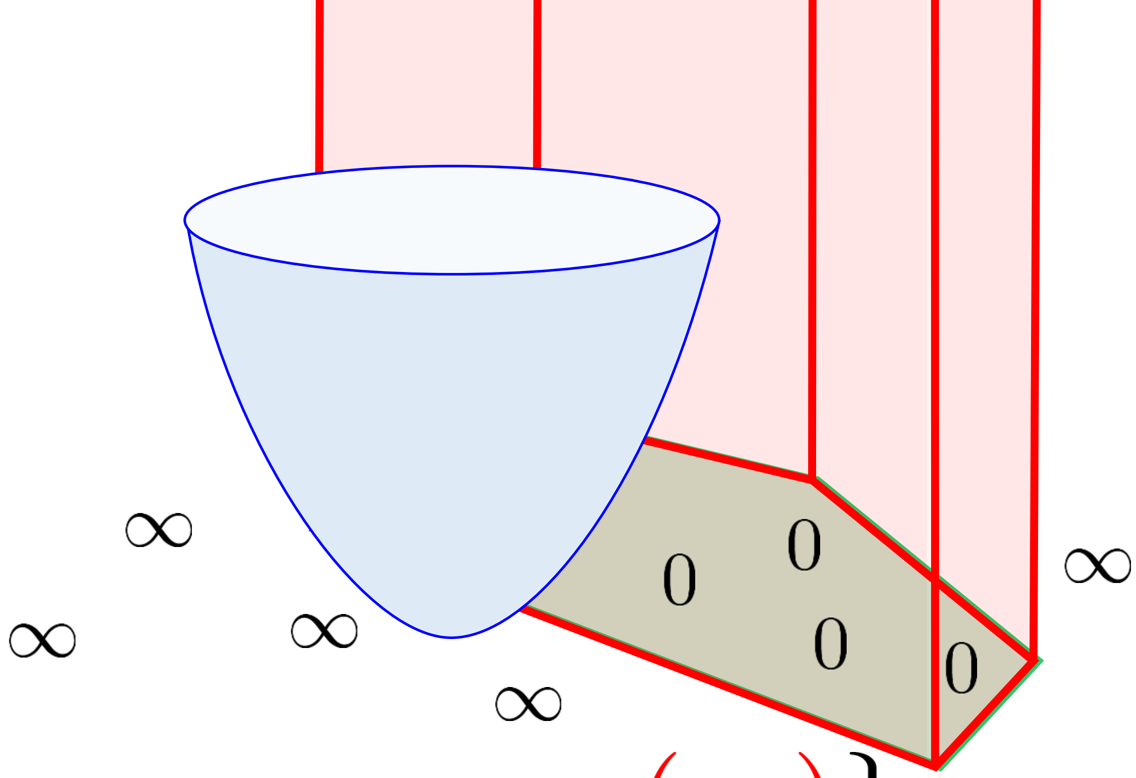
∞ if $g(\mathbf{w}) > 0$

0 if $g(\mathbf{w}) \leq 0$

The Lagrangian

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$



$$\arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{ f(\mathbf{w}) + \arg\max_{\alpha\geq 0} \alpha \cdot g(\mathbf{w}) \right\}$$

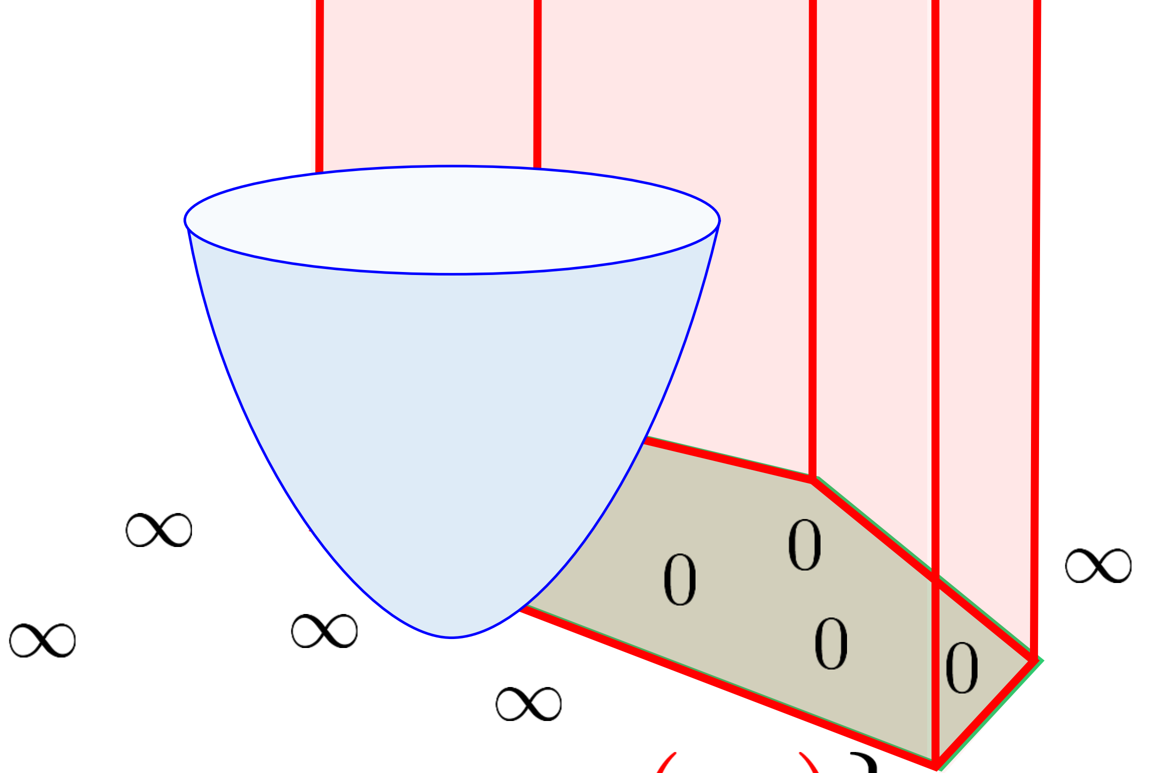
∞ if $g(\mathbf{w}) > 0$

0 if $g(\mathbf{w}) \leq 0$

The Lagrangian

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$



$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{ f(\mathbf{w}) + \arg\max_{\alpha\geq 0} \alpha \cdot g(\mathbf{w}) \right\}$$

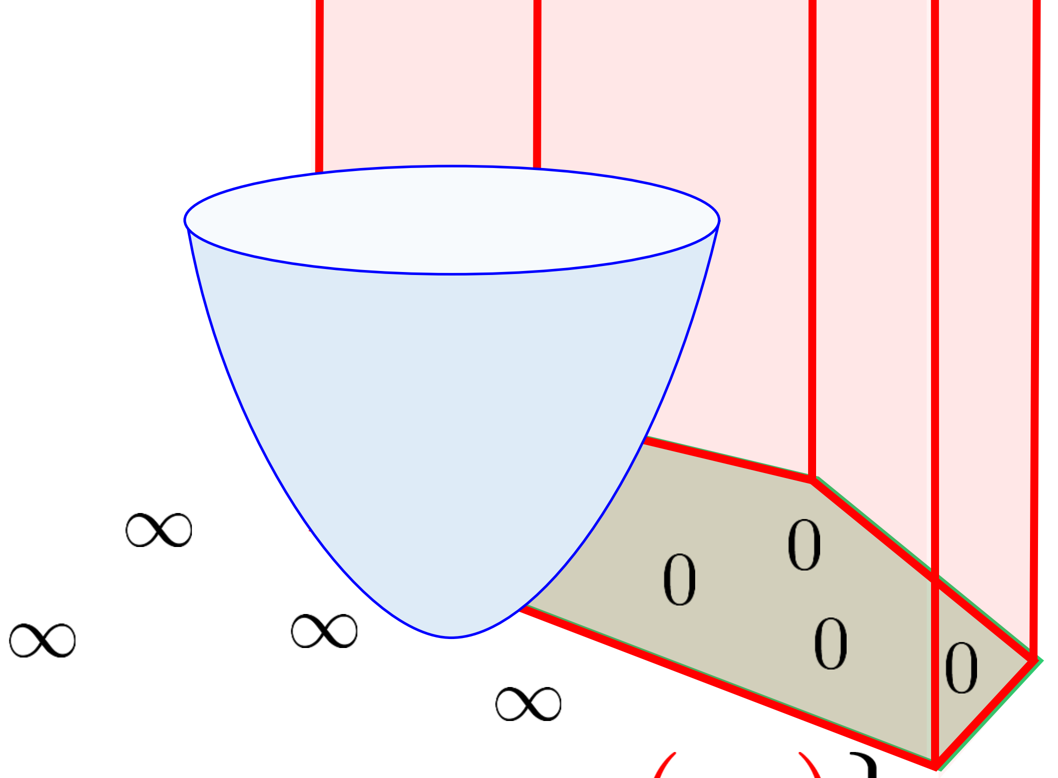
∞ if $g(\mathbf{w}) > 0$

0 if $g(\mathbf{w}) \leq 0$

The Lagrangian

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$



$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{ f(\mathbf{w}) + \arg\max_{\alpha\geq 0} \alpha \cdot g(\mathbf{w}) \right\} \quad \infty$$

∞ if $g(\mathbf{w}) > 0$

0 if $g(\mathbf{w}) \leq 0$

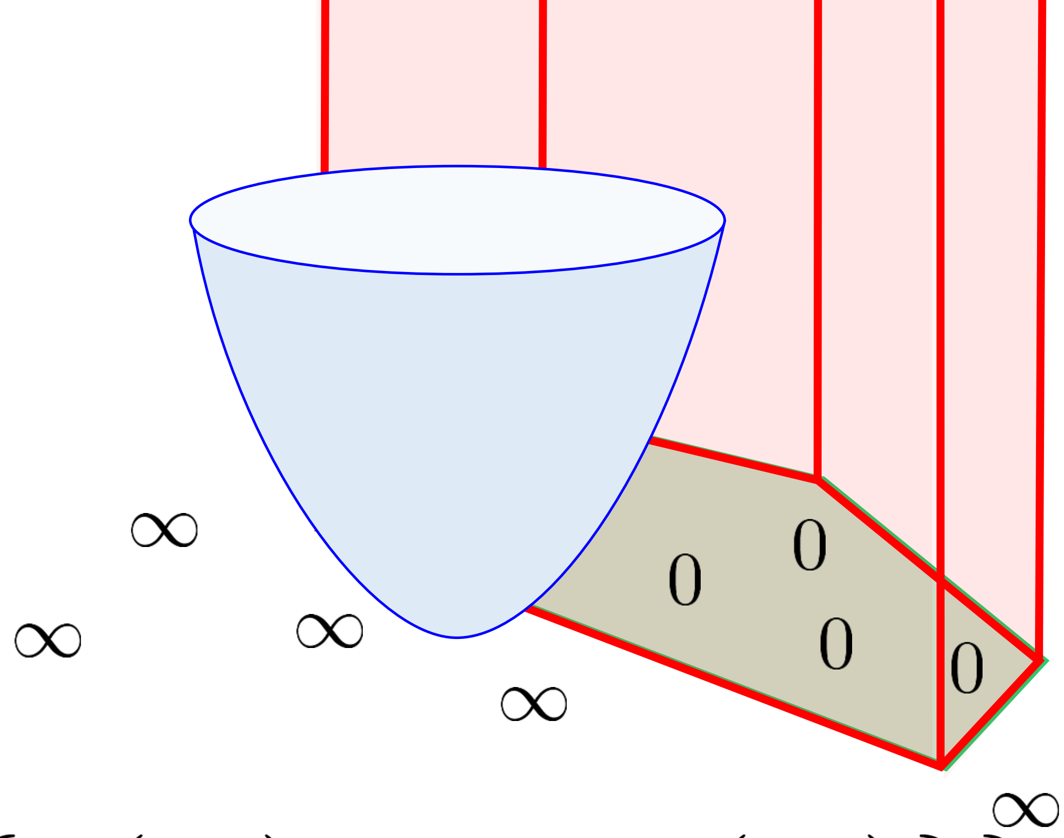
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{ \arg\max_{\alpha\geq 0} \left\{ f(\mathbf{w}) + \alpha \cdot g(\mathbf{w}) \right\} \right\}$$

Exercise

The Lagrangian

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$

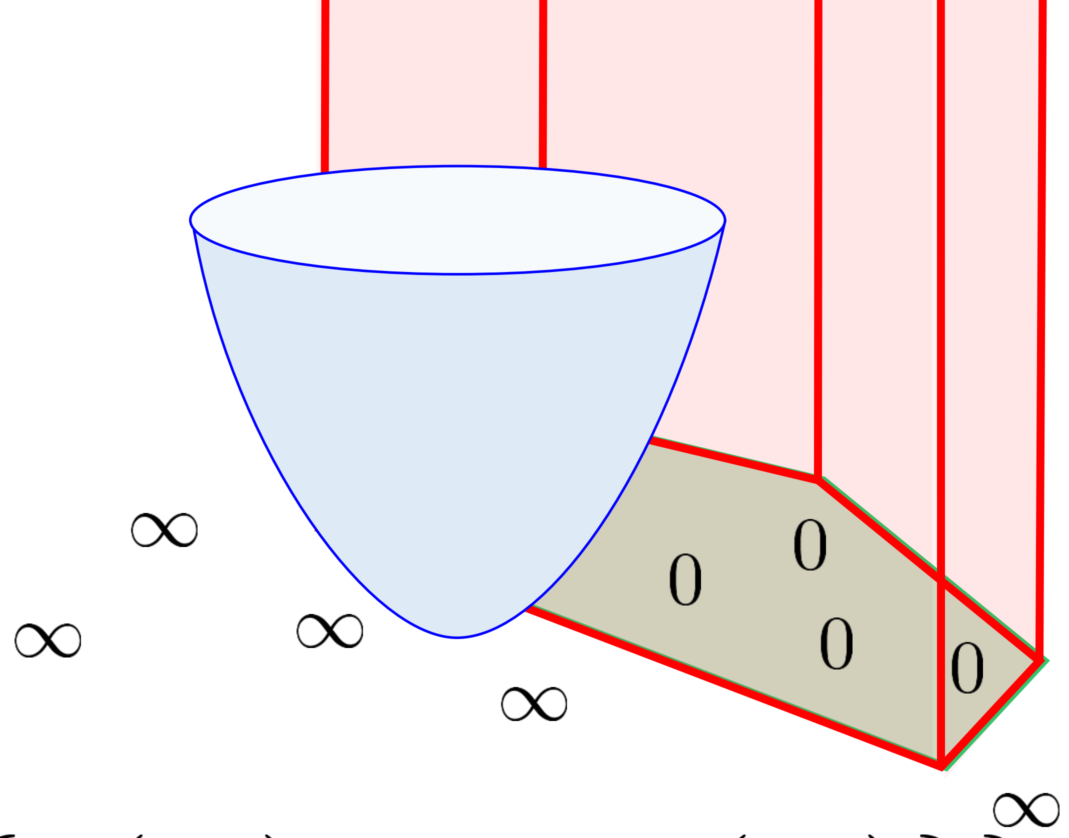


$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{\arg\max_{\alpha\geq 0} \left\{f(\mathbf{w}) + \alpha\cdot g(\mathbf{w})\right\}\right\}$$

The Lagrangian

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$



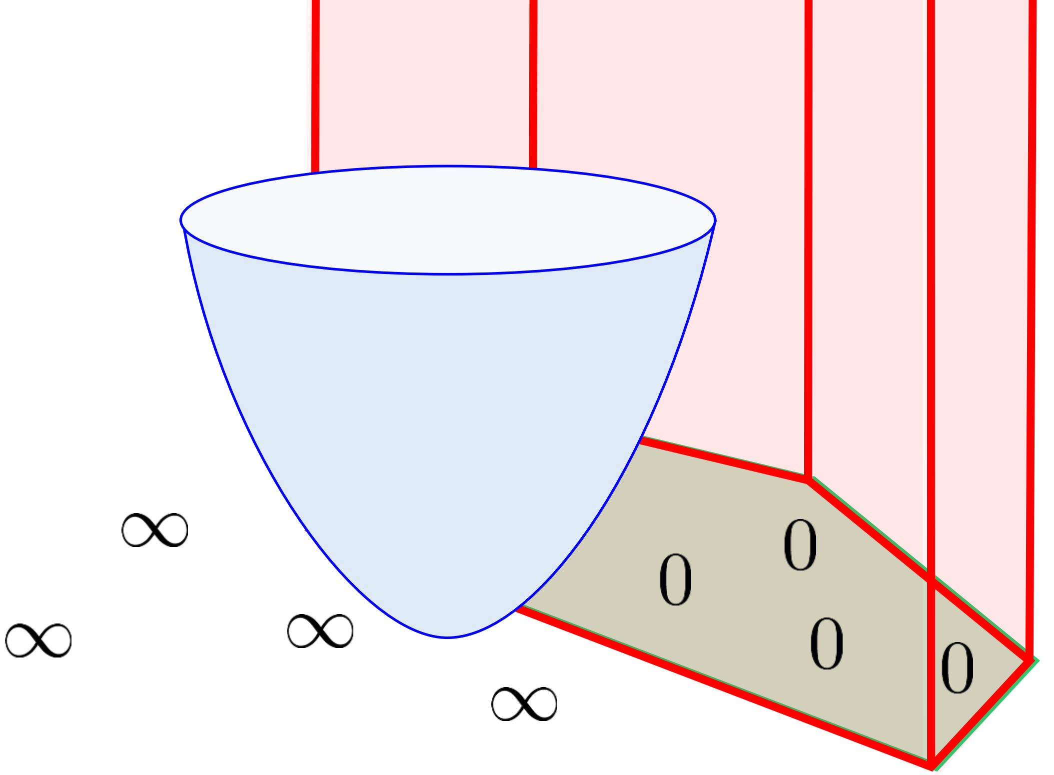
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{ \arg\max_{\alpha\geq 0} \left\{ f(\mathbf{w}) + \alpha \cdot g(\mathbf{w}) \right\} \right\}$$

$$\mathcal{L}(\mathbf{w}, \alpha) = f(\mathbf{w}) + \alpha \cdot g(\mathbf{w})$$

The Lagrangian

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$



$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{\arg\max_{\alpha\geq 0} \left\{f(\mathbf{w}) + \alpha\cdot g(\mathbf{w})\right\}\right\}$$

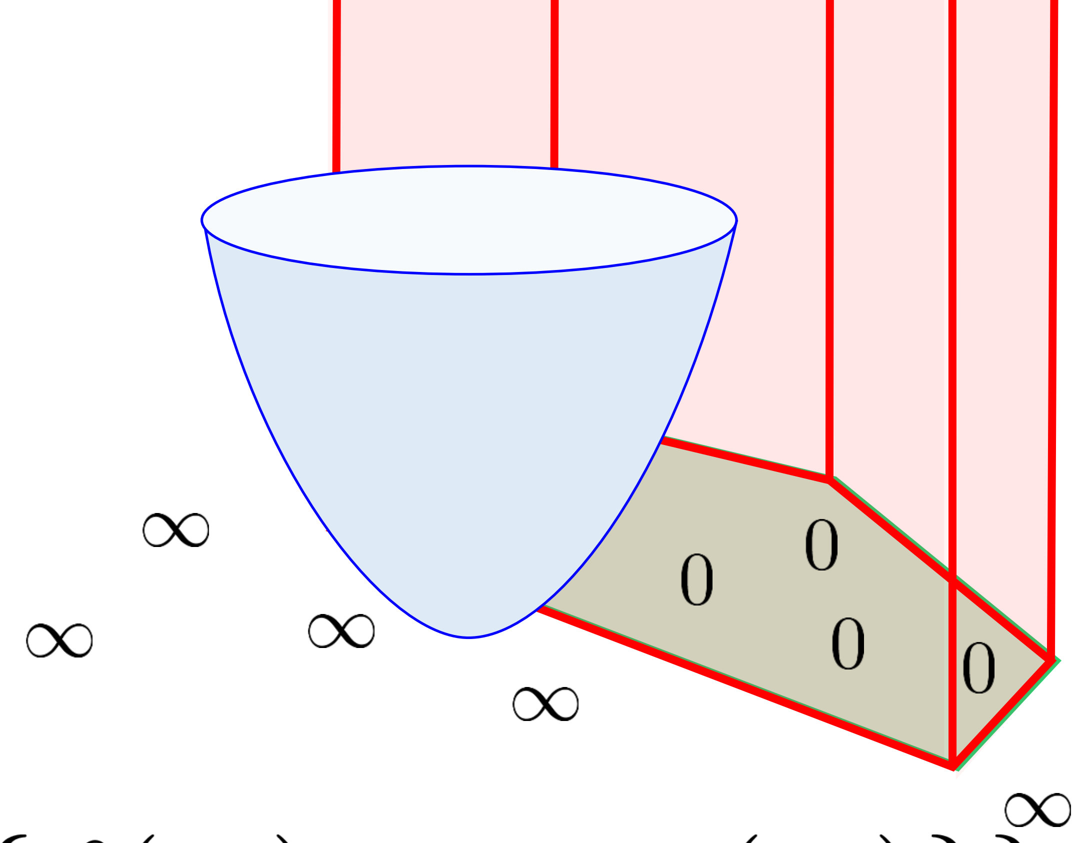
$$\mathcal{L}(\mathbf{w},\alpha) = f(\mathbf{w}) + \alpha\cdot g(\mathbf{w})$$

Lagrangian

The Lagrangian

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$



$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{\arg\max_{\alpha\geq 0} \left\{f(\mathbf{w}) + \alpha\cdot g(\mathbf{w})\right\}\right\}$$

Lagrange multiplier

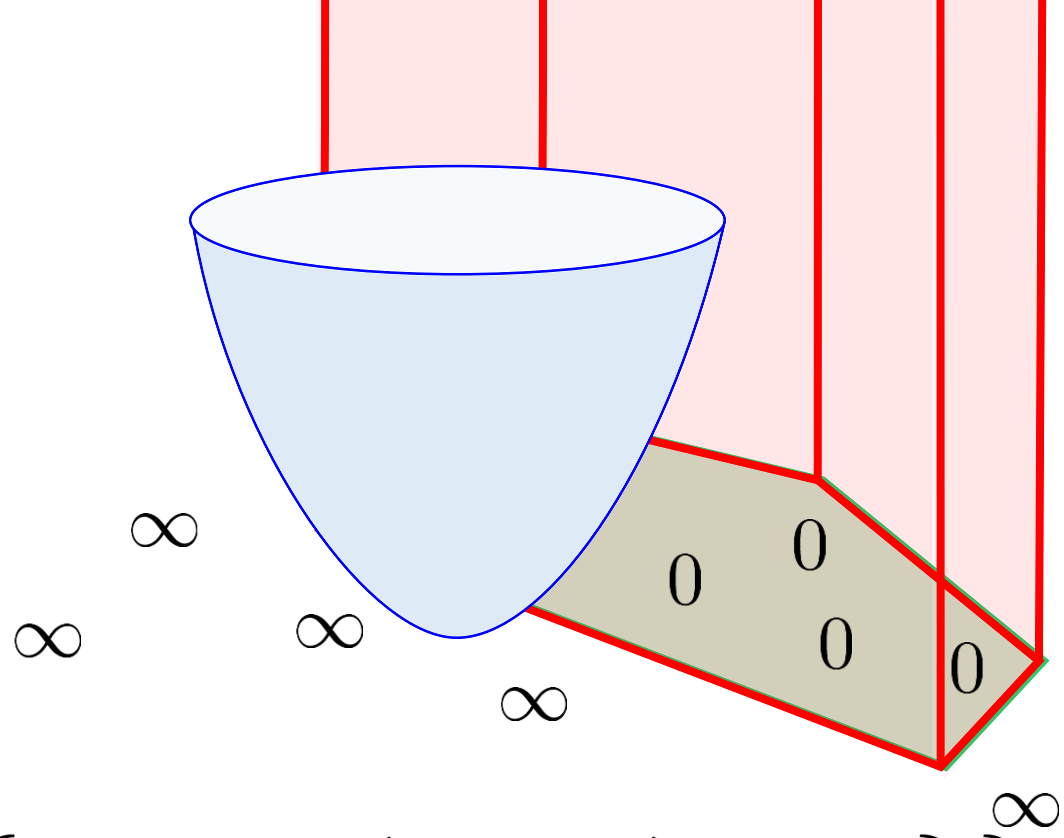
$$\mathcal{L}(\mathbf{w},\alpha) = f(\mathbf{w}) + \alpha\cdot g(\mathbf{w})$$

Lagrangian

The Lagrangian

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$



$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{\arg\max_{\alpha\geq 0} \left\{ \mathcal{L}(\mathbf{w},\alpha) \right\}\right\}$$

Lagrange multiplier

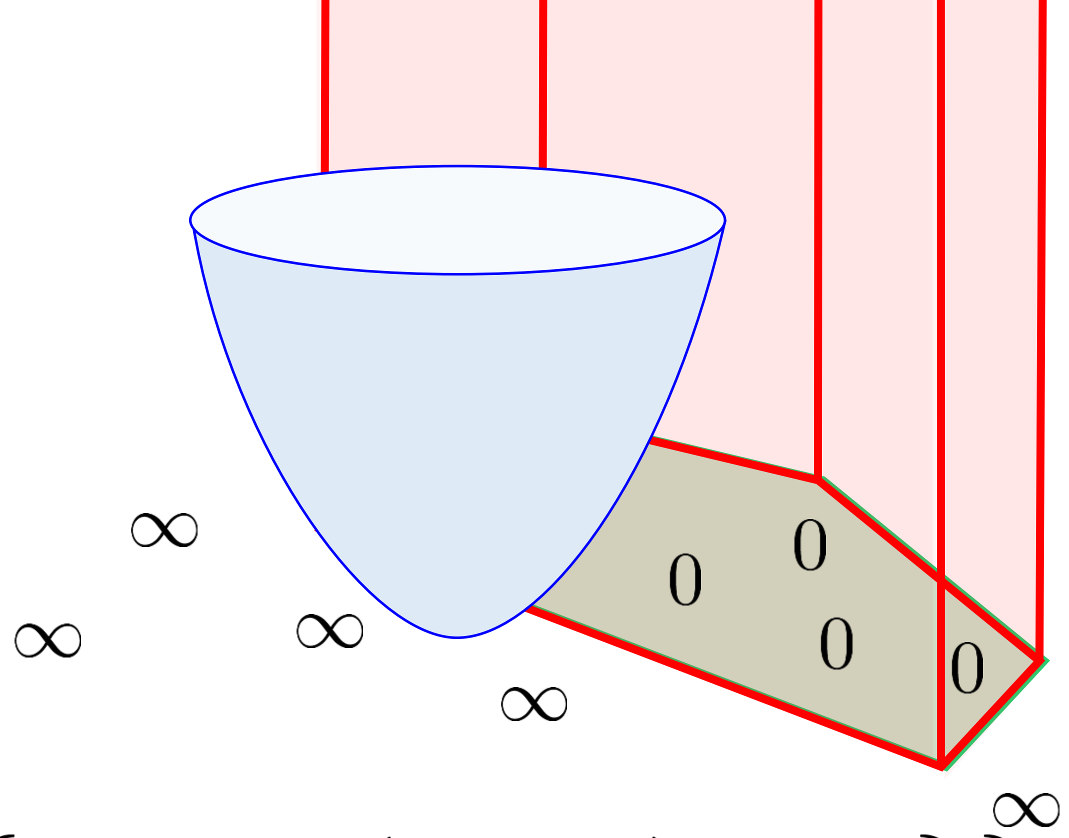
$$\mathcal{L}(\mathbf{w},\alpha) = f(\mathbf{w}) + \alpha \cdot g(\mathbf{w})$$

Lagrangian

The Lagrangian

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \le 0$$



$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{ \arg\max_{\alpha\ge 0} \left\{ \mathcal{L}(\mathbf{w},\alpha) \right\} \right\}$$

Lagrange multiplier

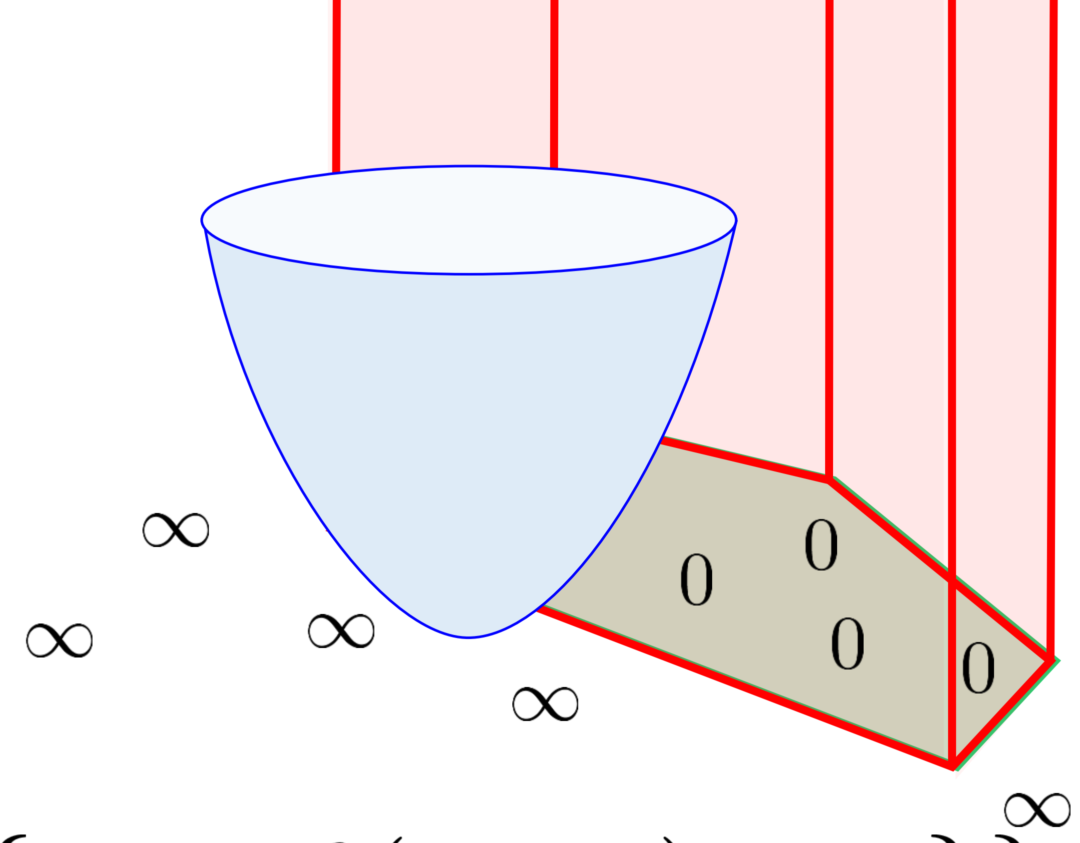
$$\mathcal{L}(\mathbf{w},\alpha) = f(\mathbf{w}) + \alpha \cdot g(\mathbf{w})$$

Lagrangian

The Lagrangian

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$



$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{ \arg\max_{\alpha\geq 0} \left\{ \mathcal{L}(\mathbf{w},\alpha) \right\} \right\}$$

Lagrange multiplier

Primal problem

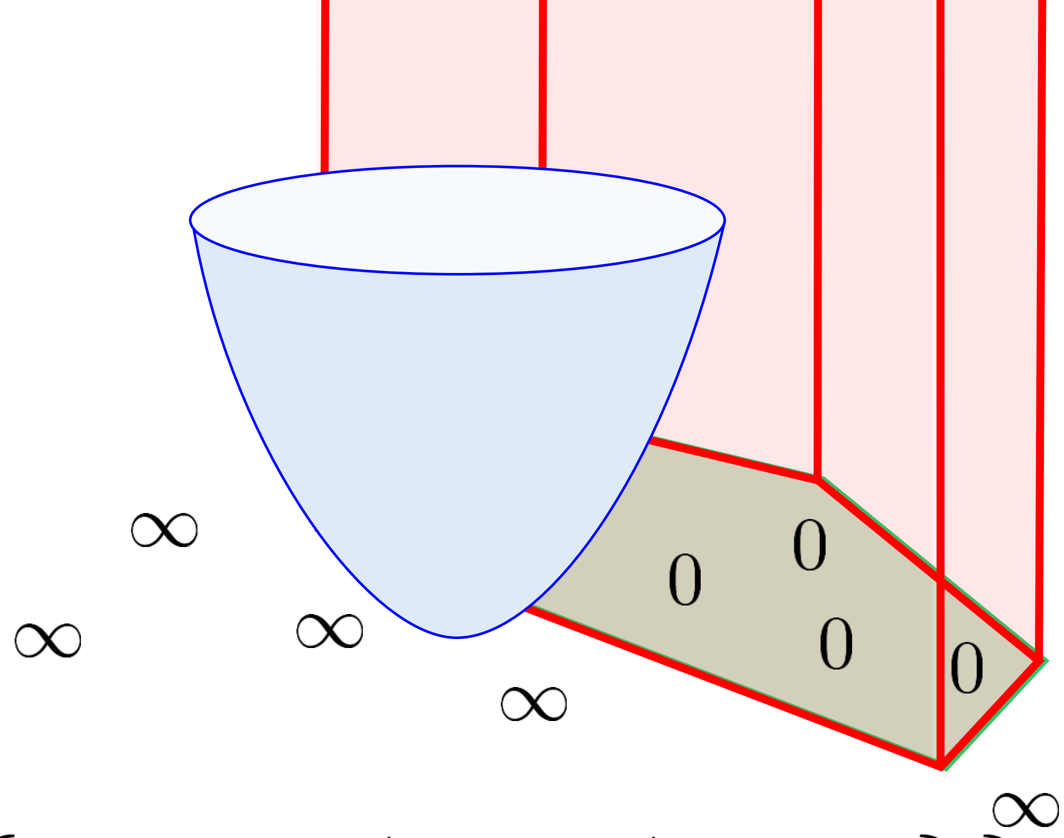
$$\mathcal{L}(\mathbf{w},\alpha) = f(\mathbf{w}) + \alpha \cdot g(\mathbf{w})$$

Lagrangian

The Lagrangian

$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$



$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{\arg\max_{\alpha\geq 0} \left\{ \mathcal{L}(\mathbf{w},\alpha) \right\}\right\}$$

Lagrange multiplier

Primal problem

$$\mathcal{L}(\mathbf{w},\alpha) = f(\mathbf{w}) + \alpha \cdot g(\mathbf{w})$$

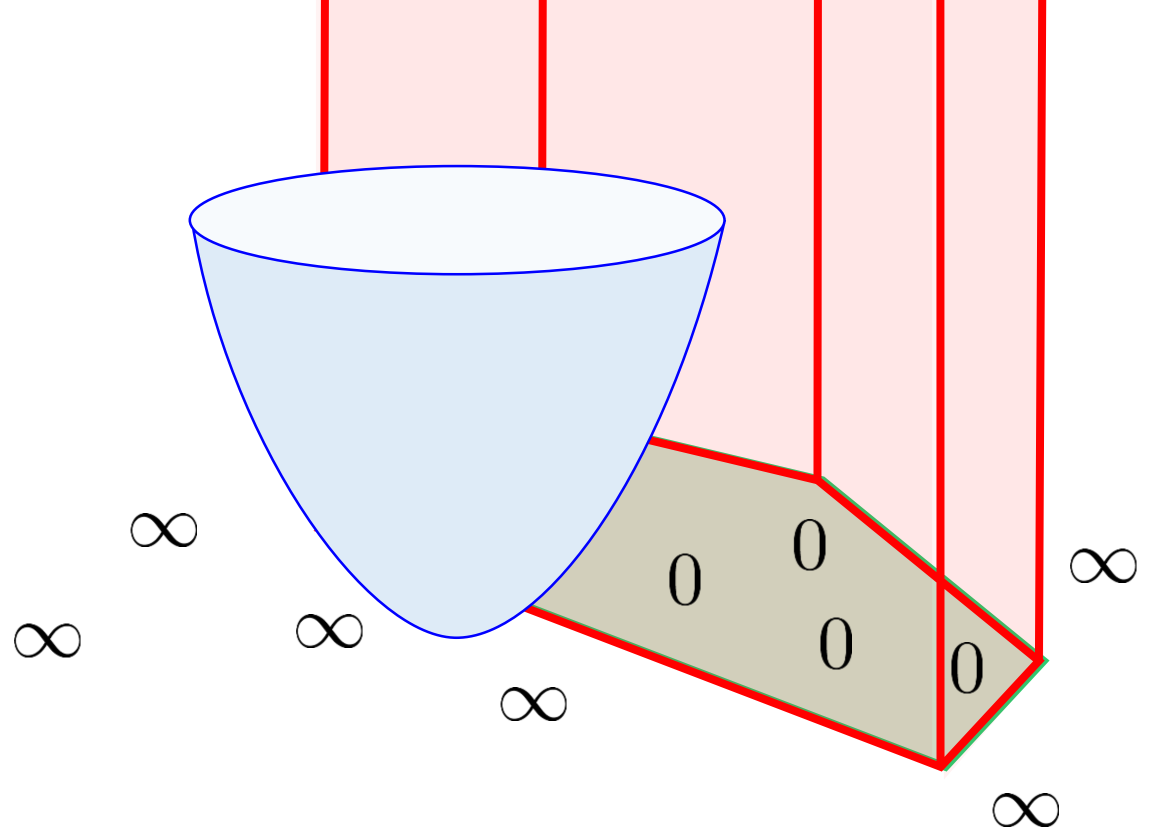
Lagrangian

The Lagrangian

$$\hat{\mathbf{w}}_P = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

Primal problem

$$\text{s.t. } g(\mathbf{w}) \le 0$$



$$\hat{\mathbf{w}}_P = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{\arg\max_{\alpha\ge 0} \left\{ \mathcal{L}(\mathbf{w},\alpha) \right\}\right\}$$

Lagrange multiplier

Primal problem

$$\mathcal{L}(\mathbf{w},\alpha) = f(\mathbf{w}) + \alpha \cdot g(\mathbf{w})$$

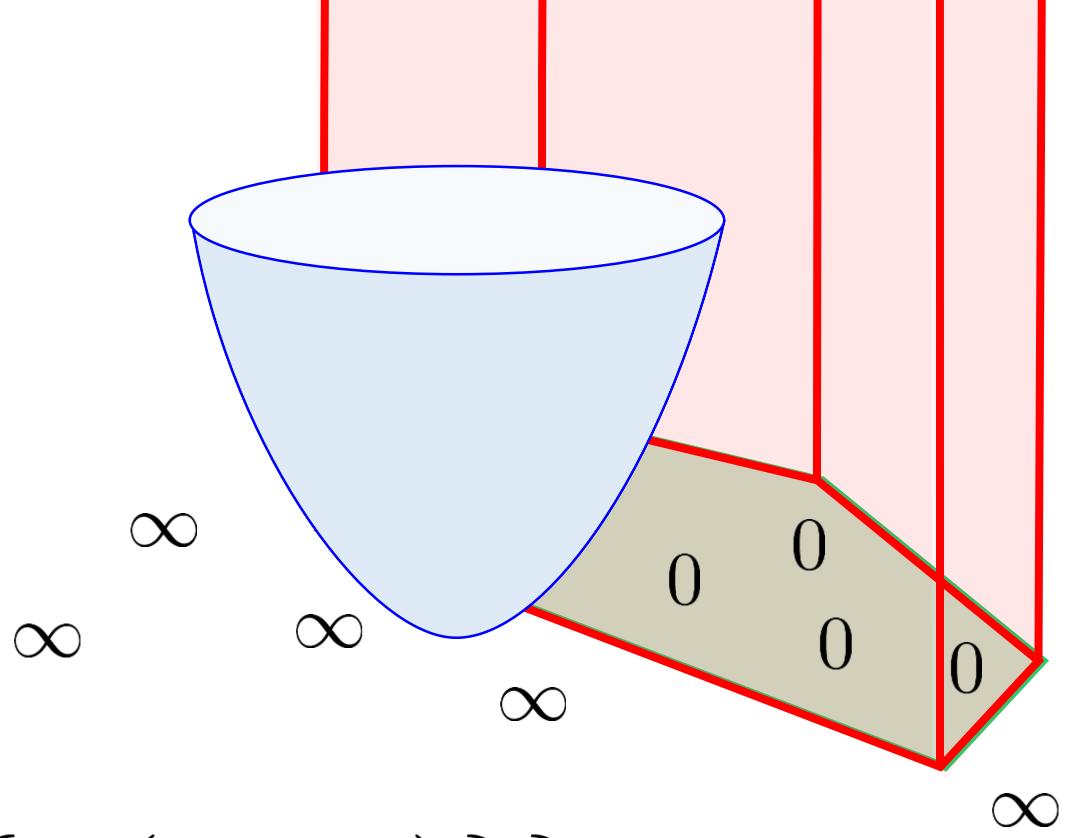
Lagrangian

The Lagrangian

$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$

Primal problem



$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{ \arg\max_{\alpha\geq 0} \left\{ \mathcal{L}(\mathbf{w},\alpha) \right\} \right\}$$

Lagrange multiplier

Primal problem

$$\mathcal{L}(\mathbf{w},\alpha) = f(\mathbf{w}) + \alpha \cdot g(\mathbf{w})$$

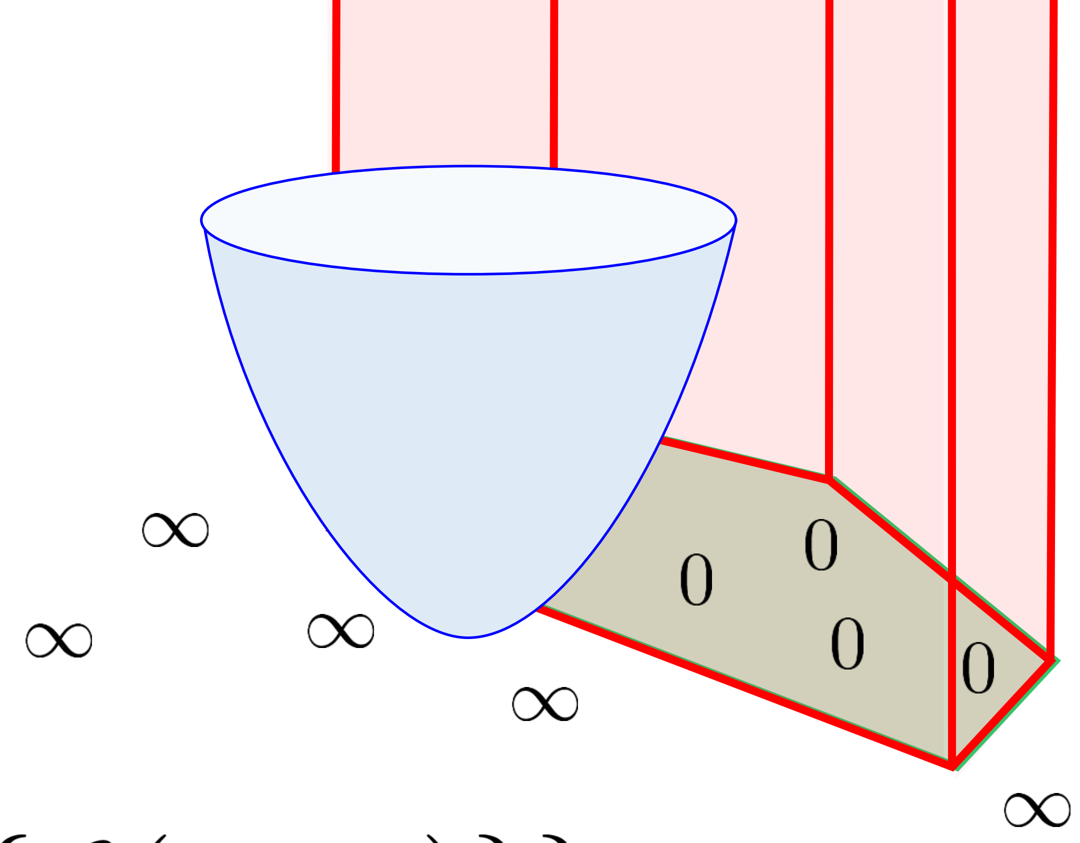
Lagrangian

The Lagrangian

$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \; f(\mathbf{w})$$

$$\text{s.t.} \; g(\mathbf{w}) \leq 0$$

Primal problem



$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{ \arg\max_{\alpha\geq 0} \left\{ \mathcal{L}(\mathbf{w}, \alpha) \right\} \right\}$$

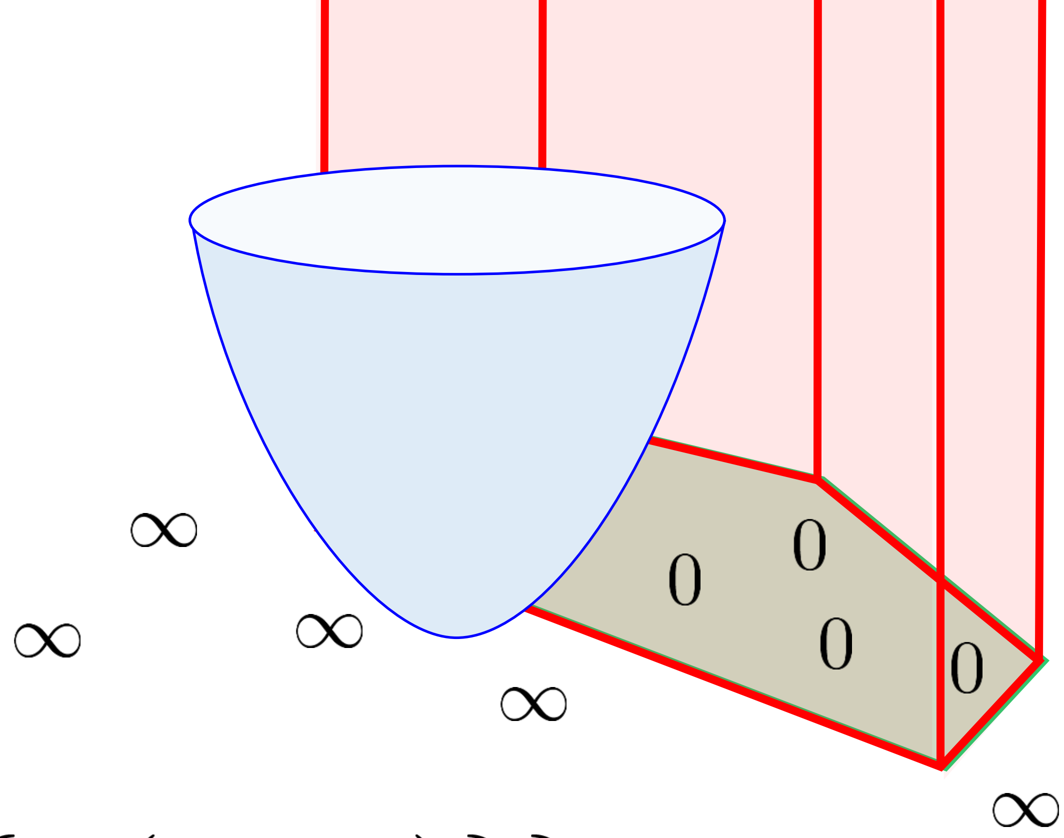
Primal problem

The Lagrangian

$$\hat{\mathbf{w}}_P = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \; f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$

Primal problem



$$\hat{\mathbf{w}}_P = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{\arg\max_{\alpha\geq 0} \left\{\mathcal{L}(\mathbf{w},\alpha)\right\}\right\}$$

Primal problem

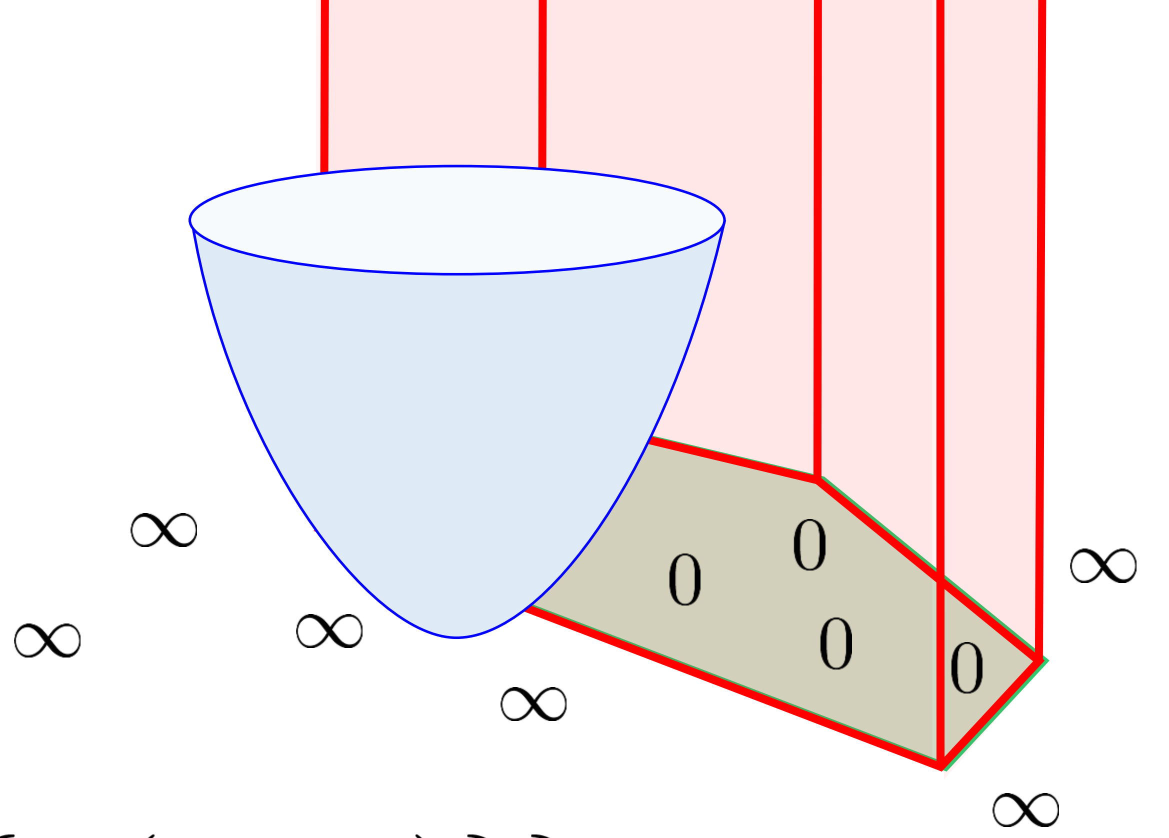
$$\hat{\mathbf{w}}_D = \arg\max_{\alpha\geq 0} \left\{\arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{\mathcal{L}(\mathbf{w},\alpha)\right\}\right\}$$

The Lagrangian

$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$

Primal problem



$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{\arg\max_{\alpha\geq 0} \left\{\mathcal{L}(\mathbf{w},\alpha)\right\}\right\}$$

Primal problem

$$\hat{\mathbf{w}}_{\mathrm{D}} = \arg\max_{\alpha\geq 0} \left\{\arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{\mathcal{L}(\mathbf{w},\alpha)\right\}\right\}$$

Dual problem

The Lagrange Dual Problem

$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$

The Lagrange Dual Problem

$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$

$$\mathcal{L}(\mathbf{w}, \alpha) = f(\mathbf{w}) + \alpha \cdot g(\mathbf{w})$$

The Lagrange Dual Problem

$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$

$$\mathcal{L}(\mathbf{w}, \alpha) = f(\mathbf{w}) + \alpha \cdot g(\mathbf{w})$$

$$\hat{\mathbf{w}}_{\mathrm{D}} = \arg\max_{\alpha \geq 0} \left\{ \arg\min_{\mathbf{w} \in \mathbb{R}^d} \left\{ \mathcal{L}(\mathbf{w}, \alpha) \right\} \right\}$$

The Lagrange Dual Problem

$$\hat{\mathbf{w}}_{\mathrm{P}} = \underset{\mathbf{w}\in\mathbb{R}^d}{\arg\min}\; f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$

$$\hat{\mathbf{w}}_{\mathrm{P}} \overset{?}{=} \hat{\mathbf{w}}_{\mathrm{D}}$$

$$\mathcal{L}(\mathbf{w}, \alpha) = f(\mathbf{w}) + \alpha \cdot g(\mathbf{w})$$

$$\hat{\mathbf{w}}_{\mathrm{D}} = \underset{\alpha\geq 0}{\arg\max}\; \left\{ \underset{\mathbf{w}\in\mathbb{R}^d}{\arg\min}\; \{\mathcal{L}(\mathbf{w}, \alpha)\} \right\}$$

The Lagrange Dual Problem

$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$

$$\hat{\mathbf{w}}_{\mathrm{P}} \stackrel{?}{=} \hat{\mathbf{w}}_{\mathrm{D}}$$

$\hat{\mathbf{w}}_p = \hat{\mathbf{w}}_d$ for nice problems*

$$\mathcal{L}(\mathbf{w}, \alpha) = f(\mathbf{w}) + \alpha \cdot g(\mathbf{w})$$

$$\hat{\mathbf{w}}_{\mathrm{D}} = \arg\max_{\alpha\geq 0} \left\{\arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{\mathcal{L}(\mathbf{w}, \alpha)\right\}\right\}$$

The Lagrange Dual Problem

$$\hat{\mathbf{w}}_\mathrm{P} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$

$$\hat{\mathbf{w}}_\mathrm{P} \overset{?}{=} \hat{\mathbf{w}}_\mathrm{D}$$

$\hat{\mathbf{w}}_p = \hat{\mathbf{w}}_d$ for nice problems*

E.g. if $f(\cdot)$ is convex and $\{\mathbf{w} : g(\mathbf{w}) \leq 0\}$ is convex

$$\mathcal{L}(\mathbf{w}, \alpha) = f(\mathbf{w}) + \alpha \cdot g(\mathbf{w})$$

$$\hat{\mathbf{w}}_\mathrm{D} = \arg\max_{\alpha \geq 0} \left\{\arg\min_{\mathbf{w}\in\mathbb{R}^d} \{\mathcal{L}(\mathbf{w}, \alpha)\}\right\}$$

The Lagrange Dual Problem

$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$

$$\mathcal{L}(\mathbf{w}, \alpha) = f(\mathbf{w}) + \alpha \cdot g(\mathbf{w})$$

$$\hat{\mathbf{w}}_{\mathrm{D}} = \arg\max_{\alpha\geq 0} \left\{\arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{\mathcal{L}(\mathbf{w}, \alpha)\right\}\right\}$$

$$\alpha_{\mathrm{D}} \cdot g(\hat{\mathbf{w}}_{\mathrm{D}}) = 0$$

$$\hat{\mathbf{w}}_{\mathrm{P}} \overset{?}{=} \hat{\mathbf{w}}_{\mathrm{D}}$$

$\hat{\mathbf{w}}_p = \hat{\mathbf{w}}_d$ for nice problems*

E.g. if $f(\cdot)$ is convex and $\{\mathbf{w} : g(\mathbf{w}) \leq 0\}$ is convex

The Lagrange Dual Problem

$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$

$$\hat{\mathbf{w}}_{\mathrm{P}} \stackrel{?}{=} \hat{\mathbf{w}}_{\mathrm{D}}$$

$\hat{\mathbf{w}}_p = \hat{\mathbf{w}}_d$ for nice problems*

E.g. if $f(\cdot)$ is convex and $\{\mathbf{w} : g(\mathbf{w}) \leq 0\}$ is convex

$$\mathcal{L}(\mathbf{w}, \alpha) = f(\mathbf{w}) + \alpha \cdot g(\mathbf{w})$$

$$\hat{\mathbf{w}}_{\mathrm{D}} = \arg\max_{\alpha \geq 0} \left\{ \arg\min_{\mathbf{w} \in \mathbb{R}^d} \left\{ \mathcal{L}(\mathbf{w}, \alpha) \right\} \right\}$$

$$\alpha_{\mathrm{D}} \cdot g(\hat{\mathbf{w}}_{\mathrm{D}}) = 0$$

The Lagrange Dual Problem

$$\hat{\mathbf{w}}_\mathrm{P} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$

$$\hat{\mathbf{w}}_\mathrm{P} \overset{?}{=} \hat{\mathbf{w}}_\mathrm{D}$$

$\hat{\mathbf{w}}_p = \hat{\mathbf{w}}_d$ for nice problems*

E.g. if $f(\cdot)$ is convex and $\{\mathbf{w} : g(\mathbf{w}) \leq 0\}$ is convex

$$\mathcal{L}(\mathbf{w}, \alpha) = f(\mathbf{w}) + \alpha \cdot g(\mathbf{w})$$

$$\hat{\mathbf{w}}_\mathrm{D} = \arg\max_{\alpha\geq 0} \left\{\arg\min_{\mathbf{w}\in\mathbb{R}^d} \{\mathcal{L}(\mathbf{w}, \alpha)\}\right\}$$

The maximizer

$$\alpha_\mathrm{D} \cdot g(\hat{\mathbf{w}}_\mathrm{D}) = 0$$

The Lagrange Dual Problem

$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g(\mathbf{w}) \leq 0$$

$$\hat{\mathbf{w}}_{\mathrm{P}} \stackrel{?}{=} \hat{\mathbf{w}}_{\mathrm{D}}$$

$\hat{\mathbf{w}}_p = \hat{\mathbf{w}}_d$ for nice problems*

$$\mathcal{L}(\mathbf{w}, \alpha) = f(\mathbf{w}) + \alpha \cdot g(\mathbf{w})$$

E.g. if $f(\cdot)$ is convex and $\{\mathbf{w} : g(\mathbf{w}) \leq 0\}$ is convex

$$\hat{\mathbf{w}}_{\mathrm{D}} = \arg\max_{\alpha\geq 0} \left\{\arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{\mathcal{L}(\mathbf{w}, \alpha)\right\}\right\}$$

The maximizer

$$\alpha_{\mathrm{D}} \cdot g(\hat{\mathbf{w}}_{\mathrm{D}}) = 0$$

Complimentary slackness KKT Condition

Multiple Constraints!

Multiple Constraints!

$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g_1(\mathbf{w}) \leq 0, g_2(\mathbf{w}) \leq 0$$

$$h(\mathbf{w}) = 0$$

Multiple Constraints!

$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

$$\text{s.t. } g_1(\mathbf{w}) \leq 0, g_2(\mathbf{w}) \leq 0$$

$$h(\mathbf{w}) = 0$$

$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{\arg\max_{\substack{\alpha,\beta\geq 0\\ \gamma\in\mathbb{R}}} \left\{f(\mathbf{w}) + \alpha\cdot g_1(\mathbf{w}) + \beta\cdot g_2(\mathbf{w}) + \gamma\cdot h(\mathbf{w})\right\}\right\}$$

Multiple Constraints!

$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \; f(\mathbf{w})$$

$$\text{s.t. } g_1(\mathbf{w}) \leq 0, g_2(\mathbf{w}) \leq 0$$

$$h(\mathbf{w}) = 0$$

One Lagrange multiplier for each constraint

$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{ \arg\max_{\substack{\alpha,\beta\geq 0\\ \gamma\in\mathbb{R}}} \left\{ f(\mathbf{w}) + \alpha\cdot g_1(\mathbf{w}) + \beta\cdot g_2(\mathbf{w}) + \gamma\cdot h(\mathbf{w}) \right\} \right\}$$

Multiple Constraints!

$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

One Lagrange multiplier for each constraint

$$\text{s.t. } g_1(\mathbf{w}) \le 0, g_2(\mathbf{w}) \le 0$$

$$h(\mathbf{w}) = 0$$

$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{ \arg\max_{\substack{\alpha,\beta\ge 0\\ \gamma\in\mathbb{R}}} \left\{ f(\mathbf{w}) + \alpha\cdot g_1(\mathbf{w}) + \beta\cdot g_2(\mathbf{w}) + \gamma\cdot h(\mathbf{w}) \right\} \right\}$$

Doesn't allow $g_1(\mathbf{w}) > 0$

Multiple Constraints!

$$\hat{\mathbf{w}}_P = \underset{\mathbf{w} \in \mathbb{R}^d}{\arg\min}\ f(\mathbf{w})$$

$$\text{s.t. } g_1(\mathbf{w}) \leq 0, g_2(\mathbf{w}) \leq 0$$

$$h(\mathbf{w}) = 0$$

One Lagrange multiplier for each constraint

$$\hat{\mathbf{w}}_P = \underset{\mathbf{w} \in \mathbb{R}^d}{\arg\min}\ \left\{ \underset{\substack{\alpha,\beta \geq 0 \\ \gamma \in \mathbb{R}}}{\arg\max}\ \{ f(\mathbf{w}) + \alpha \cdot g_1(\mathbf{w}) + \beta \cdot g_2(\mathbf{w}) + \gamma \cdot h(\mathbf{w}) \} \right\}$$

Doesn't allow $g_1(\mathbf{w}) > 0$

Doesn't allow $g_2(\mathbf{w}) > 0$

Multiple Constraints!

$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

One Lagrange multiplier for each constraint

$$\text{s.t. } g_1(\mathbf{w}) \leq 0, g_2(\mathbf{w}) \leq 0$$

$$h(\mathbf{w}) = 0$$

$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{ \arg\max_{\substack{\alpha,\beta\geq 0\\ \gamma\in\mathbb{R}}} \left\{ f(\mathbf{w}) + \alpha\cdot g_1(\mathbf{w}) + \beta\cdot g_2(\mathbf{w}) + \gamma\cdot h(\mathbf{w}) \right\} \right\}$$

Doesn't allow $g_1(\mathbf{w}) > 0$

Doesn't allow $g_2(\mathbf{w}) > 0$

Doesn't allow $h(\mathbf{w}) \neq 0$

Multiple Constraints!

$$\hat{\mathbf{w}}_{\mathrm{P}} = \underset{\mathbf{w}\in\mathbb{R}^d}{\arg\min}\ f(\mathbf{w})$$

$$\text{s.t. } g_1(\mathbf{w}) \leq 0, g_2(\mathbf{w}) \leq 0$$

$$h(\mathbf{w}) = 0$$

One Lagrange multiplier for each constraint

$$\hat{\mathbf{w}}_{\mathrm{P}} = \underset{\mathbf{w}\in\mathbb{R}^d}{\arg\min} \left\{ \underset{\substack{\alpha,\beta\geq 0\\ \gamma\in\mathbb{R}}}{\arg\max} \left\{ f(\mathbf{w}) + \alpha\cdot g_1(\mathbf{w}) + \beta\cdot g_2(\mathbf{w}) + \gamma\cdot h(\mathbf{w}) \right\} \right\}$$

Multiple Constraints!

$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \; f(\mathbf{w})$$

$$\text{s.t. } g_1(\mathbf{w}) \leq 0, g_2(\mathbf{w}) \leq 0$$
$$h(\mathbf{w}) = 0$$

One Lagrange multiplier for each constraint

$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \; \left\{ \arg\max_{\substack{\alpha,\beta\geq 0 \\ \gamma\in\mathbb{R}}} \; \{ f(\mathbf{w}) + \alpha\cdot g_1(\mathbf{w}) + \beta\cdot g_2(\mathbf{w}) + \gamma\cdot h(\mathbf{w}) \} \right\}$$

$$\hat{\mathbf{w}}_{\mathrm{D}} = \arg\max_{\substack{\alpha,\beta\geq 0 \\ \gamma\in\mathbb{R}}} \; \left\{ \arg\min_{\mathbf{w}\in\mathbb{R}^d} \; \{ f(\mathbf{w}) + \alpha\cdot g_1(\mathbf{w}) + \beta\cdot g_2(\mathbf{w}) + \gamma\cdot h(\mathbf{w}) \} \right\}$$

Multiple Constraints!

$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$$

Complimentary Slackness

$$\alpha_D \cdot g_1(\hat{\mathbf{w}}_D) = 0$$
$$\beta_D \cdot g_2(\hat{\mathbf{w}}_D) = 0$$

$$\text{s.t. } g_1(\mathbf{w}) \leq 0, g_2(\mathbf{w}) \leq 0$$
$$h(\mathbf{w}) = 0$$

One Lagrange multiplier for each constraint

$$\hat{\mathbf{w}}_{\mathrm{P}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{\arg\max_{\substack{\alpha,\beta\geq 0\\ \gamma\in\mathbb{R}}} \left\{f(\mathbf{w}) + \alpha\cdot g_1(\mathbf{w}) + \beta\cdot g_2(\mathbf{w}) + \gamma\cdot h(\mathbf{w})\right\}\right\}$$

$$\hat{\mathbf{w}}_{\mathrm{D}} = \arg\max_{\substack{\alpha,\beta\geq 0\\ \gamma\in\mathbb{R}}} \left\{\arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{f(\mathbf{w}) + \alpha\cdot g_1(\mathbf{w}) + \beta\cdot g_2(\mathbf{w}) + \gamma\cdot h(\mathbf{w})\right\}\right\}$$

Multiple Constraints!

Complimentary Slackness

$$\alpha_D \cdot g_1(\hat{\mathbf{w}}_D) = 0$$
$$\beta_D \cdot g_2(\hat{\mathbf{w}}_D) = 0$$

$$\hat{\mathbf{w}}_{\mathrm{P}} = \underset{\mathbf{w} \in \mathbb{R}^d}{\arg\min} \; f(\mathbf{w})$$

One Lagrange multiplier for each constraint

$$\text{s.t. } g_1(\mathbf{w}) \leq 0, g_2(\mathbf{w}) \leq 0$$
$$h(\mathbf{w}) = 0$$

Only for inequality constraints

$$\hat{\mathbf{w}}_{\mathrm{P}} = \underset{\mathbf{w} \in \mathbb{R}^d}{\arg\min} \left\{ \underset{\substack{\alpha, \beta \geq 0 \\ \gamma \in \mathbb{R}}}{\arg\max} \left\{ f(\mathbf{w}) + \alpha \cdot g_1(\mathbf{w}) + \beta \cdot g_2(\mathbf{w}) + \gamma \cdot h(\mathbf{w}) \right\} \right\}$$

$$\hat{\mathbf{w}}_{\mathrm{D}} = \underset{\substack{\alpha, \beta \geq 0 \\ \gamma \in \mathbb{R}}}{\arg\max} \left\{ \underset{\mathbf{w} \in \mathbb{R}^d}{\arg\min} \left\{ f(\mathbf{w}) + \alpha \cdot g_1(\mathbf{w}) + \beta \cdot g_2(\mathbf{w}) + \gamma \cdot h(\mathbf{w}) \right\} \right\}$$

SVM Revisited

SVM Revisited

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\left\|\mathbf{w}\right\|_2^2$$
$$\text{s.t. } y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \geq 1$$

SVM Revisited

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \leq 0$$

SVM Revisited

n constraints, so n Lagrange multipliers

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i\left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \leq 0$$

SVM Revisited

n constraints, so n Lagrange multipliers

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i\langle \mathbf{w}, \mathbf{x}^i\rangle \leq 0$$

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \boldsymbol{\alpha}_i \left(1 - y^i\langle \mathbf{w}, \mathbf{x}^i\rangle\right)$$

SVM Revisited

n constraints, so n Lagrange multipliers

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \leq 0$$

$\boldsymbol{\alpha} \in \mathbb{R}^n$

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \boldsymbol{\alpha}_i \left(1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle\right)$$

SVM Revisited

n constraints, so n Lagrange multipliers

$\boldsymbol{\alpha} \in \mathbb{R}^n$

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle \leq 0$$

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \boldsymbol{\alpha}_i \left(1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle\right)$$

$$\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{D}} = \max_{\boldsymbol{\alpha} \geq 0} \left\{ \min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) \right\}$$

SVM Revisited

n constraints, so n Lagrange multipliers

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle \leq 0$$

$\boldsymbol{\alpha} \in \mathbb{R}^n$

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \boldsymbol{\alpha}_i \left(1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle\right)$$

$$\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{D}} = \max_{\boldsymbol{\alpha} \geq 0} \left\{ \min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) \right\}$$

$\boldsymbol{\alpha} \geq 0$ is notation for $\boldsymbol{\alpha}_i \geq 0$ for all i

SVM Revisited

n constraints, so n Lagrange multipliers

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle \leq 0$$

$\boldsymbol{\alpha} \in \mathbb{R}^n$

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^n \boldsymbol{\alpha}_i \left(1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle\right)$$

Inner problem is an Unconstrained problem so Gradients must vanish at optimum

$$\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{D}} = \max_{\boldsymbol{\alpha} \geq 0} \left\{ \min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) \right\}$$

$\boldsymbol{\alpha} \geq 0$ is notation for $\boldsymbol{\alpha}_i \geq 0$ for all i

SVM Revisited

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \leq 0$$

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \boldsymbol{\alpha}_i \left(1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle\right)$$

$$\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{D}} = \max_{\boldsymbol{\alpha} \geq 0} \left\{ \min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) \right\}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{0}$$

n constraints, so n Lagrange multipliers

$\boldsymbol{\alpha} \in \mathbb{R}^n$

Inner problem is an Unconstrained problem so Gradients must vanish at optimum

$\boldsymbol{\alpha} \geq 0$ is notation for $\boldsymbol{\alpha}_i \geq 0$ for all i

SVM Revisited

n constraints, so n Lagrange multipliers

$\boldsymbol{\alpha} \in \mathbb{R}^n$

Inner problem is an Unconstrained problem so Gradients must vanish at optimum

$\boldsymbol{\alpha} \geq 0$ is notation for $\boldsymbol{\alpha}_i \geq 0$ for all i

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \leq 0$$

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) = \frac{1}{2} \|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \boldsymbol{\alpha}_i \left(1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle\right)$$

$$\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{D}} = \max_{\boldsymbol{\alpha} \geq 0} \left\{ \min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) \right\}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{0} \Rightarrow \hat{\mathbf{w}}_{\mathrm{D}} = \sum_{i=1}^{n} \boldsymbol{\alpha}_i y^i \mathbf{x}^i$$

SVM Revisited

n constraints, so n Lagrange multipliers

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$
$$\text{s.t. } 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \leq 0$$

$\boldsymbol{\alpha} \in \mathbb{R}^n$

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \boldsymbol{\alpha}_i \left(1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle\right)$$

Inner problem is an Unconstrained problem so Gradients must vanish at optimum

$$\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{D}} = \max_{\boldsymbol{\alpha} \geq 0} \left\{ \min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) \right\}$$

$\boldsymbol{\alpha} \geq 0$ is notation for $\boldsymbol{\alpha}_i \geq 0$ for all i

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{0} \Rightarrow \hat{\mathbf{w}}_{\mathrm{D}} = \sum_{i=1}^{n} \boldsymbol{\alpha}_i y^i \mathbf{x}^i$$

Optimal model $\hat{\mathbf{w}}$ is a weighted sum of training points!

SVM Revisited

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle \leq 0$$

n constraints, so n Lagrange multipliers

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \boldsymbol{\alpha}_i \left(1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle\right)$$

$\boldsymbol{\alpha} \in \mathbb{R}^n$

$$\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{D}} = \max_{\boldsymbol{\alpha} \geq 0} \left\{ \min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) \right\}$$

Inner problem is an Unconstrained problem so Gradients must vanish at optimum

$\boldsymbol{\alpha} \geq 0$ is notation for $\boldsymbol{\alpha}_i \geq 0$ for all i

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{0} \Rightarrow \hat{\mathbf{w}}_{\mathrm{D}} = \sum_{i=1}^{n} \boldsymbol{\alpha}_i y^i \mathbf{x}^i$$

Optimal model $\hat{\mathbf{w}}$ is a weighted sum of training points!

Points with $\boldsymbol{\alpha}_i \neq 0$ support vectors

SVM Revisited

n constraints, so n Lagrange multipliers

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$
$$\text{s.t. } 1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle \leq 0$$

$\boldsymbol{\alpha} \in \mathbb{R}^n$

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^n \boldsymbol{\alpha}_i \left(1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle\right)$$

Inner problem is an Unconstrained problem so Gradients must vanish at optimum

$$\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{D}} = \max_{\boldsymbol{\alpha} \geq 0} \left\{ \min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) \right\}$$

$\boldsymbol{\alpha} \geq 0$ is notation for $\boldsymbol{\alpha}_i \geq 0$ for all i

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{0} \Rightarrow \hat{\mathbf{w}}_{\mathrm{D}} = \sum_{i=1}^n \boldsymbol{\alpha}_i y^i \mathbf{x}^i$$

Lets substitute $\mathbf{w}$ in $\mathcal{L}$ and eliminate it!

Optimal model $\widehat{\mathbf{w}}$ is a weighted sum of training points!

Points with $\boldsymbol{\alpha}_i \neq 0$ support vectors

SVM Revisited

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \leq 0$$

$$\mathcal{L}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \boldsymbol{\alpha}_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_j y^i y^j \left\langle \mathbf{x}^i, \mathbf{x}^j \right\rangle$$

$$\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{D}} = \max_{\boldsymbol{\alpha} \geq 0} \left\{ \min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \boldsymbol{\alpha}) \right\}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{0} \Rightarrow \hat{\mathbf{w}}_{\mathrm{D}} = \sum_{i=1}^{n} \boldsymbol{\alpha}_i y^i \mathbf{x}^i$$

n constraints, so n Lagrange multipliers

$\boldsymbol{\alpha} \in \mathbb{R}^n$

Inner problem is an Unconstrained problem so Gradients must vanish at optimum

$\boldsymbol{\alpha} \geq 0$ is notation for $\boldsymbol{\alpha}_i \geq 0$ for all i

Lets substitute $\mathbf{w}$ in $\mathcal{L}$ and eliminate it!

Optimal model $\hat{\mathbf{w}}$ is a weighted sum of training points!

Points with $\boldsymbol{\alpha}_i \neq 0$ support vectors

SVM Revisited

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\left\|\mathbf{w}\right\|_2^2$$

$$\text{s.t. } 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \leq 0$$

SVM Revisited

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \leq 0$$

$$\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{D}} = \sum_{i=1}^{n} \hat{\boldsymbol{\alpha}}_i y^i \mathbf{x}^i$$

SVM Revisited

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \leq 0$$

$$\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{D}} = \sum_{i=1}^{n} \hat{\boldsymbol{\alpha}}_i y^i \mathbf{x}^i$$

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha} \in \mathbb{R}^n} \sum_{i=1}^{n} \boldsymbol{\alpha}_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_j y^i y^j \left\langle \mathbf{x}^i, \mathbf{x}^j \right\rangle$$

$$\text{s.t. } \boldsymbol{\alpha}_i \geq 0$$

SVM Revisited

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \leq 0$$

$$\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\text{D}} = \sum_{i=1}^{n} \hat{\boldsymbol{\alpha}}_i y^i \mathbf{x}^i$$

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha} \in \mathbb{R}^n} \sum_{i=1}^{n} \boldsymbol{\alpha}_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_j y^i y^j \left\langle \mathbf{x}^i, \mathbf{x}^j \right\rangle$$

$$\text{s.t. } \boldsymbol{\alpha}_i \geq 0$$

$$\hat{\boldsymbol{\alpha}}_i (1 - y^i \left\langle \hat{\mathbf{w}}_{\text{D}}, \mathbf{x}^i \right\rangle) = 0$$

SVM Revisited

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle \leq 0$$

$$\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{D}} = \sum_{i=1}^{n} \hat{\boldsymbol{\alpha}}_i y^i \mathbf{x}^i$$

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \sum_{i=1}^{n} \boldsymbol{\alpha}_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_j y^i y^j \langle \mathbf{x}^i, \mathbf{x}^j \rangle$$

$$\text{s.t. } \boldsymbol{\alpha}_i \geq 0$$

$$\hat{\boldsymbol{\alpha}}_i (1 - y^i \langle \hat{\mathbf{w}}_{\mathrm{D}}, \mathbf{x}^i \rangle) = 0$$

Complimentary slackness

SVM Revisited

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \leq 0$$

$$\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{D}} = \sum_{i=1}^{n} \hat{\boldsymbol{\alpha}}_i y^i \mathbf{x}^i$$

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \sum_{i=1}^{n} \boldsymbol{\alpha}_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_j y^i y^j \left\langle \mathbf{x}^i, \mathbf{x}^j \right\rangle$$

$$\text{s.t. } \boldsymbol{\alpha}_i \geq 0$$

$$\hat{\boldsymbol{\alpha}}_i (1 - y^i \left\langle \hat{\mathbf{w}}_{\mathrm{D}}, \mathbf{x}^i \right\rangle) = 0$$

Complimentary slackness

If $\widehat{\boldsymbol{\alpha}}_i \neq 0$, then $y^i\langle \widehat{\mathbf{w}}_{\mathrm{D}}, \mathbf{x}^i \rangle = 1$

SVM Revisited

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$

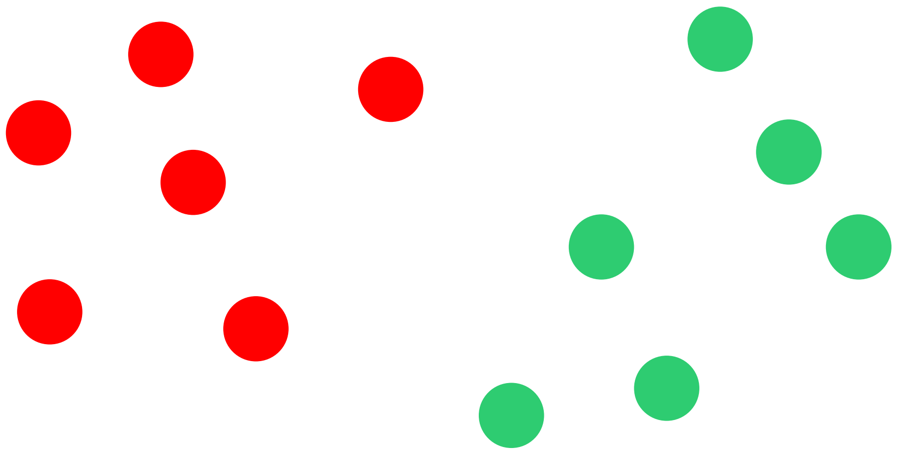
$$\text{s.t. } 1 - y^i\left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \leq 0$$

$$\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{D}} = \sum_{i=1}^{n} \hat{\boldsymbol{\alpha}}_i y^i \mathbf{x}^i$$

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \sum_{i=1}^{n} \boldsymbol{\alpha}_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\alpha}_i\boldsymbol{\alpha}_j y^i y^j \left\langle \mathbf{x}^i, \mathbf{x}^j \right\rangle$$

$$\text{s.t. } \boldsymbol{\alpha}_i \geq 0$$

$$\hat{\boldsymbol{\alpha}}_i(1 - y^i\left\langle \hat{\mathbf{w}}_{\mathrm{D}}, \mathbf{x}^i \right\rangle) = 0$$



Complimentary slackness

If $\widehat{\boldsymbol{\alpha}}_i \neq 0$, then $y^i\langle \widehat{\mathbf{w}}_{\mathrm{D}}, \mathbf{x}^i \rangle = 1$

SVM Revisited

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle \leq 0$$

$$\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{D}} = \sum_{i=1}^{n} \hat{\boldsymbol{\alpha}}_i y^i \mathbf{x}^i$$

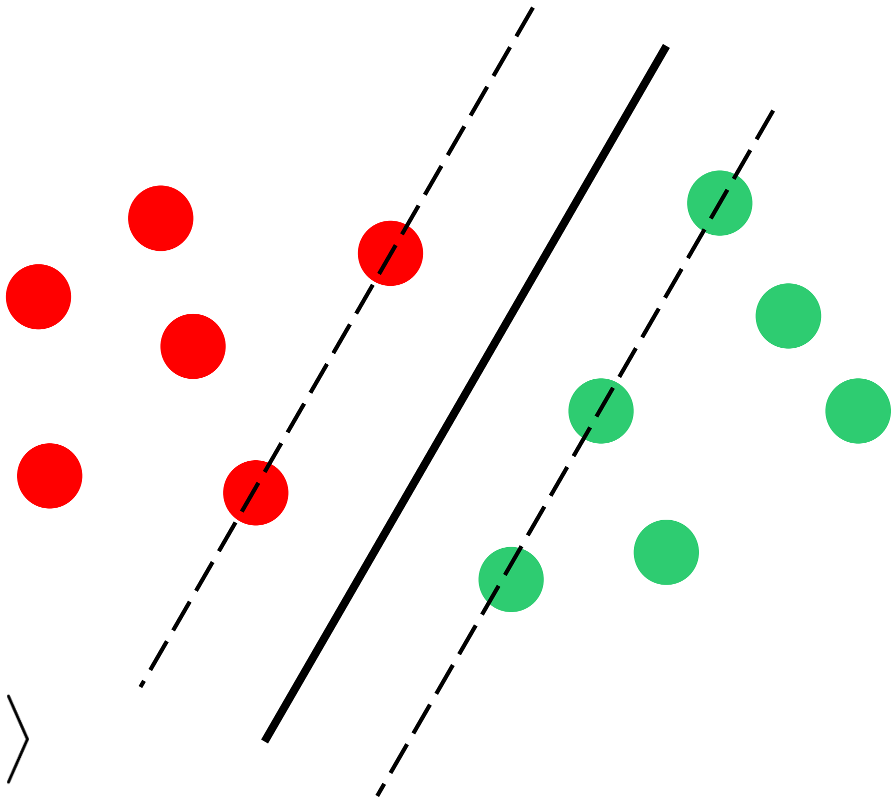
$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha} \in \mathbb{R}^n} \sum_{i=1}^{n} \boldsymbol{\alpha}_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_j y^i y^j \langle \mathbf{x}^i, \mathbf{x}^j \rangle$$

$$\text{s.t. } \boldsymbol{\alpha}_i \geq 0$$

$$\hat{\boldsymbol{\alpha}}_i (1 - y^i \langle \hat{\mathbf{w}}_{\mathrm{D}}, \mathbf{x}^i \rangle) = 0$$

Complimentary slackness

If $\widehat{\boldsymbol{\alpha}}_i \neq 0$, then $y^i \langle \widehat{\mathbf{w}}_{\mathrm{D}}, \mathbf{x}^i \rangle = 1$



SVM Revisited

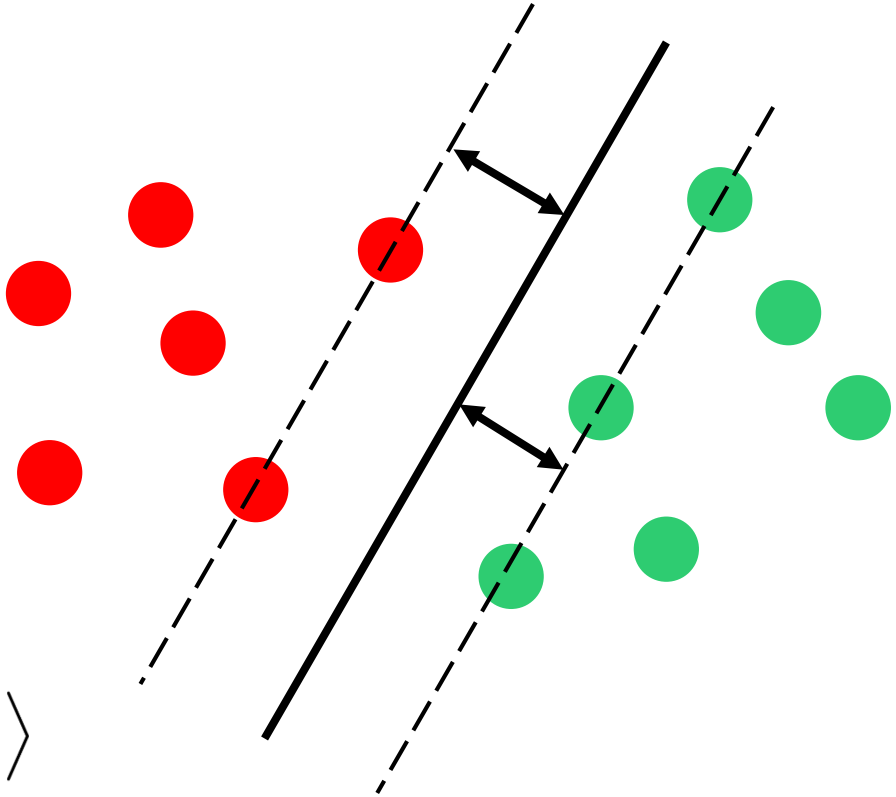
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \leq 0$$

$$\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{D}} = \sum_{i=1}^n \hat{\boldsymbol{\alpha}}_i y^i \mathbf{x}^i$$

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \sum_{i=1}^n \boldsymbol{\alpha}_i - \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \boldsymbol{\alpha}_i\boldsymbol{\alpha}_j y^i y^j \left\langle \mathbf{x}^i, \mathbf{x}^j \right\rangle$$

$$\text{s.t. } \boldsymbol{\alpha}_i \geq 0$$



$$\hat{\boldsymbol{\alpha}}_i(1 - y^i \left\langle \hat{\mathbf{w}}_{\mathrm{D}}, \mathbf{x}^i \right\rangle) = 0$$

Complimentary slackness

If $\hat{\boldsymbol{\alpha}}_i \neq 0$, then $y^i\langle \hat{\mathbf{w}}_{\mathrm{D}}, \mathbf{x}^i\rangle = 1$

SVM Revisited

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i \left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \leq 0$$

$$\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{D}} = \sum_{i=1}^{n} \hat{\boldsymbol{\alpha}}_i y^i \mathbf{x}^i$$

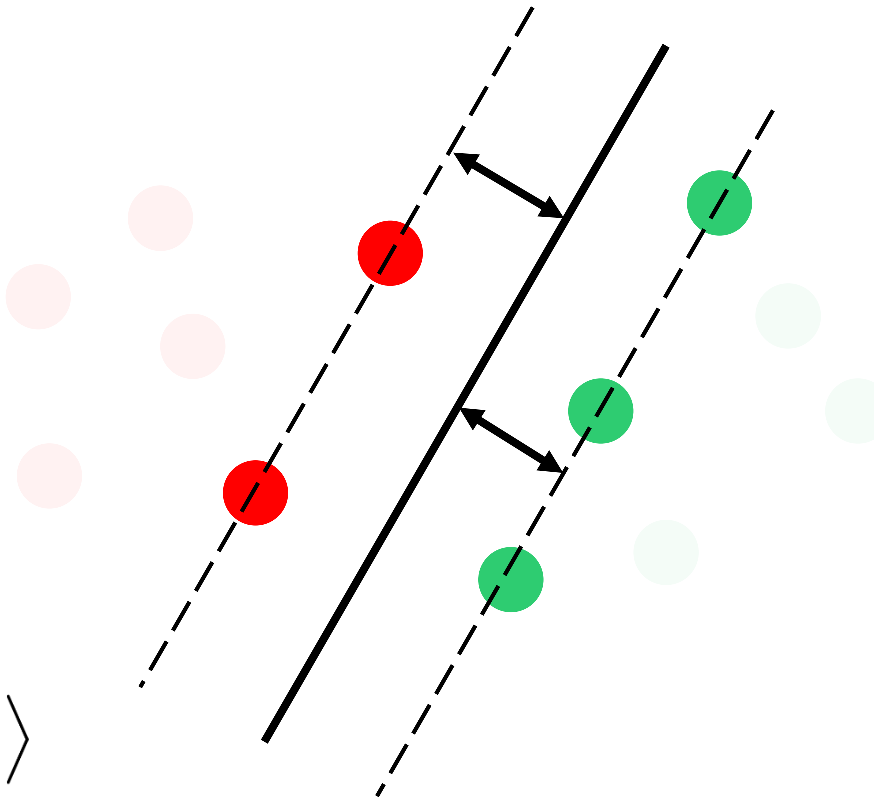
$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \sum_{i=1}^{n} \boldsymbol{\alpha}_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_j y^i y^j \left\langle \mathbf{x}^i, \mathbf{x}^j \right\rangle$$

$$\text{s.t. } \boldsymbol{\alpha}_i \geq 0$$

$$\hat{\boldsymbol{\alpha}}_i(1 - y^i \left\langle \hat{\mathbf{w}}_{\mathrm{D}}, \mathbf{x}^i \right\rangle) = 0$$

Complimentary slackness

If $\widehat{\boldsymbol{\alpha}}_i \neq 0$, then $y^i \langle \widehat{\mathbf{w}}_{\mathrm{D}}, \mathbf{x}^i \rangle = 1$



SVM Revisited

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i\left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \leq 0$$

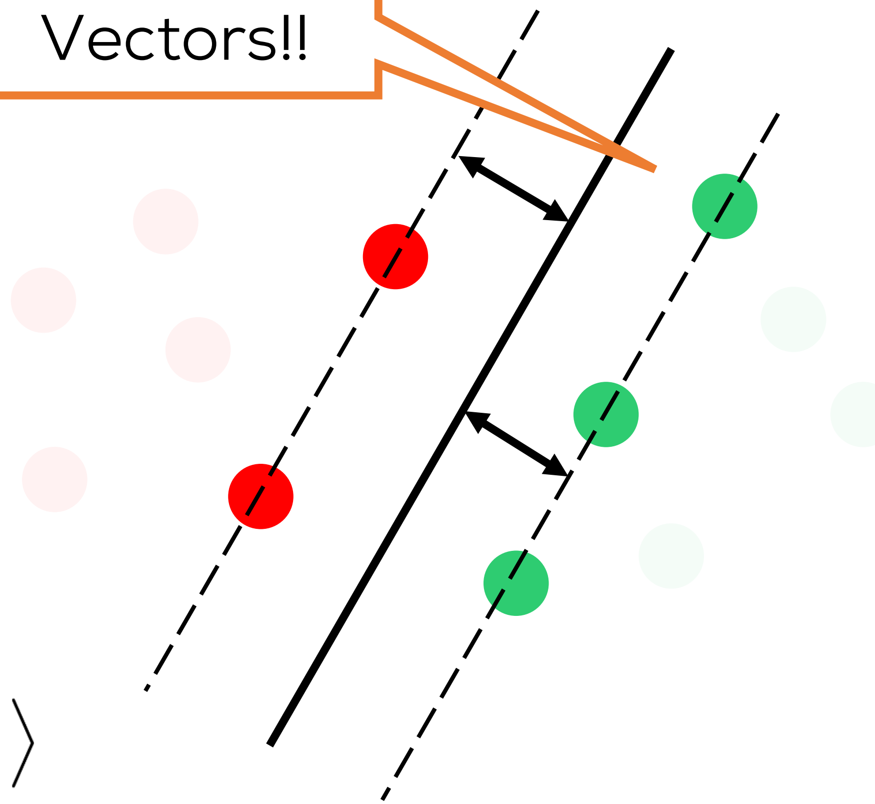
$$\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{D}} = \sum_{i=1}^{n} \hat{\boldsymbol{\alpha}}_i y^i \mathbf{x}^i$$

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \sum_{i=1}^{n}\boldsymbol{\alpha}_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j y^i y^j\left\langle \mathbf{x}^i, \mathbf{x}^j\right\rangle$$

$$\text{s.t. } \boldsymbol{\alpha}_i \geq 0$$

$$\hat{\boldsymbol{\alpha}}_i\left(1 - y^i\left\langle \hat{\mathbf{w}}_{\mathrm{D}}, \mathbf{x}^i\right\rangle\right) = 0$$

Support Vectors!!



Complimentary slackness

If $\hat{\boldsymbol{\alpha}}_i \neq 0$, then $y^i\langle \hat{\mathbf{w}}_{\mathrm{D}}, \mathbf{x}^i\rangle = 1$

SVM Revisited

$$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle \leq 0$$

$$\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{D}} = \sum_{i=1}^{n} \hat{\boldsymbol{\alpha}}_i y^i \mathbf{x}^i$$

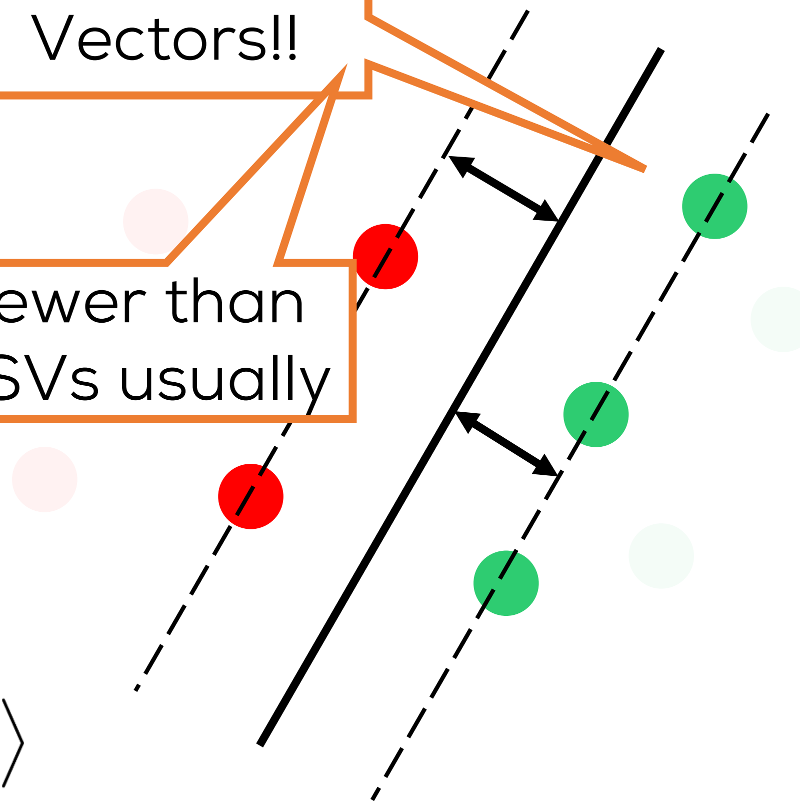
$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha} \in \mathbb{R}^n} \sum_{i=1}^{n} \boldsymbol{\alpha}_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_j y^i y^j \langle \mathbf{x}^i, \mathbf{x}^j \rangle$$

$$\text{s.t. } \boldsymbol{\alpha}_i \geq 0$$

$$\hat{\boldsymbol{\alpha}}_i(1 - y^i \langle \hat{\mathbf{w}}_{\mathrm{D}}, \mathbf{x}^i \rangle) = 0$$

Complimentary slackness

If $\widehat{\boldsymbol{\alpha}}_i \neq 0$, then $y^i \langle \widehat{\mathbf{w}}_{\mathrm{D}}, \mathbf{x}^i \rangle = 1$



SVM Revisited

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle \leq 0$$

$$\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{D}} = \sum_{i=1}^{n} \hat{\boldsymbol{\alpha}}_i y^i \mathbf{x}^i$$

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha} \in \mathbb{R}^n} \sum_{i=1}^{n} \boldsymbol{\alpha}_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_j y^i y^j \langle \mathbf{x}^i, \mathbf{x}^j \rangle$$

$$\text{s.t. } \boldsymbol{\alpha}_i \geq 0$$

$$\hat{\boldsymbol{\alpha}}_i (1 - y^i \langle \hat{\mathbf{w}}_{\mathrm{D}}, \mathbf{x}^i \rangle) = 0$$

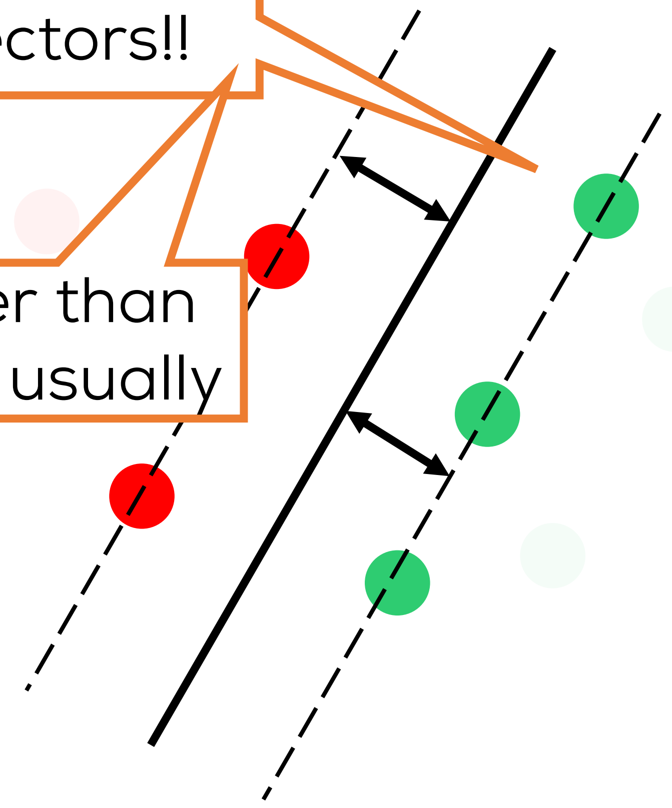
Support Vectors!!

Fewer than n SVs usually

$\|\boldsymbol{\alpha}\|_1$ is a sparsity promoting reg.!!

Complimentary slackness

If $\hat{\boldsymbol{\alpha}}_i \neq 0$, then $y^i \langle \hat{\mathbf{w}}_{\mathrm{D}}, \mathbf{x}^i \rangle = 1$



SVM Revisited

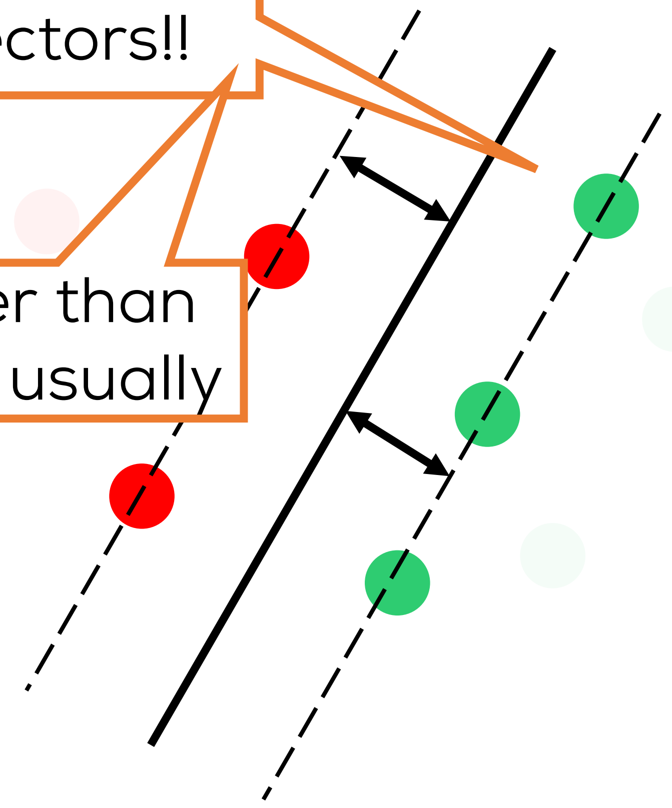
What about the version with slack and all the notational agony?

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle \le 0$$

Support Vectors!!

Fewer than n SVs usually



$$\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{D}} = \sum_{i=1}^{n} \hat{\boldsymbol{\alpha}}_i y^i \mathbf{x}^i$$

$\|\boldsymbol{\alpha}\|_1$ is a sparsity promoting reg.!!

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha} \in \mathbb{R}^n} \sum_{i=1}^{n} \boldsymbol{\alpha}_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_j y^i y^j \langle \mathbf{x}^i, \mathbf{x}^j \rangle$$

$$\text{s.t. } \boldsymbol{\alpha}_i \ge 0$$

$$\hat{\boldsymbol{\alpha}}_i (1 - y^i \langle \hat{\mathbf{w}}_{\mathrm{D}}, \mathbf{x}^i \rangle) = 0$$

Complimentary slackness

If $\hat{\boldsymbol{\alpha}}_i \neq 0$, then $y^i \langle \hat{\mathbf{w}}_{\mathrm{D}}, \mathbf{x}^i \rangle = 1$

SVM Revisited

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i \langle \mathbf{w}, \mathbf{x}^i \rangle \leq 0$$

What about the version with slack and all the notational agony?

Slightly more tedious

$$\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{D}} = \sum_{i=1}^n \hat{\boldsymbol{\alpha}}_i y^i \mathbf{x}^i$$

$\|\boldsymbol{\alpha}\|_1$ is a sparsity promoting reg.!!

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \sum_{i=1}^n \boldsymbol{\alpha}_i - \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \boldsymbol{\alpha}_i\boldsymbol{\alpha}_j y^i y^j \langle \mathbf{x}^i, \mathbf{x}^j \rangle$$

$$\text{s.t. } \boldsymbol{\alpha}_i \geq 0$$

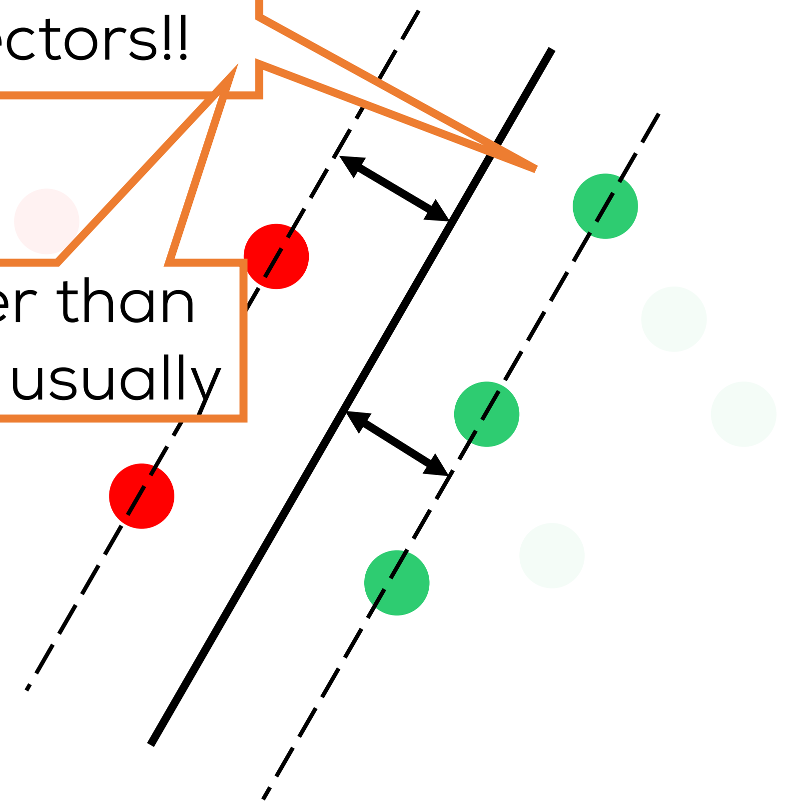
$$\hat{\boldsymbol{\alpha}}_i(1 - y^i \langle \hat{\mathbf{w}}_{\mathrm{D}}, \mathbf{x}^i \rangle) = 0$$

Complimentary slackness

If $\hat{\boldsymbol{\alpha}}_i \neq 0$, then $y^i\langle \hat{\mathbf{w}}_{\mathrm{D}}, \mathbf{x}^i \rangle = 1$

Support Vectors!!

Fewer than n SVs usually



SVM Revisited

What about the version with slack and all the notational agony?

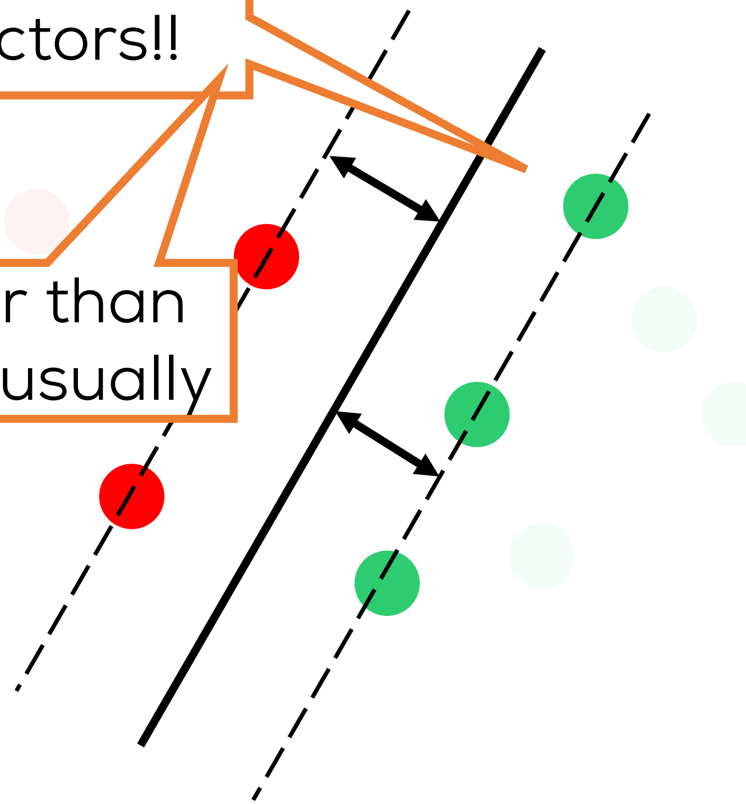
Slightly more tedious

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t. } 1 - y^i\left\langle \mathbf{w}, \mathbf{x}^i \right\rangle \leq 0$$

Support Vectors!!

Fewer than n SVs usually



$$\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{D}} = \sum_{i=1}^{n} \hat{\boldsymbol{\alpha}}_i y^i \mathbf{x}^i$$

$\|\boldsymbol{\alpha}\|_1$ is a sparsity promoting reg.!!

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \sum_{i=1}^{n} \boldsymbol{\alpha}_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\alpha}_i\boldsymbol{\alpha}_j y^i y^j \left\langle \mathbf{x}^i, \mathbf{x}^j \right\rangle$$

$$\text{s.t. } \boldsymbol{\alpha}_i \geq 0$$

$$\hat{\boldsymbol{\alpha}}_i(1 - y^i\left\langle \hat{\mathbf{w}}_{\mathrm{D}}, \mathbf{x}^i \right\rangle) = 0$$

Complimentary slackness

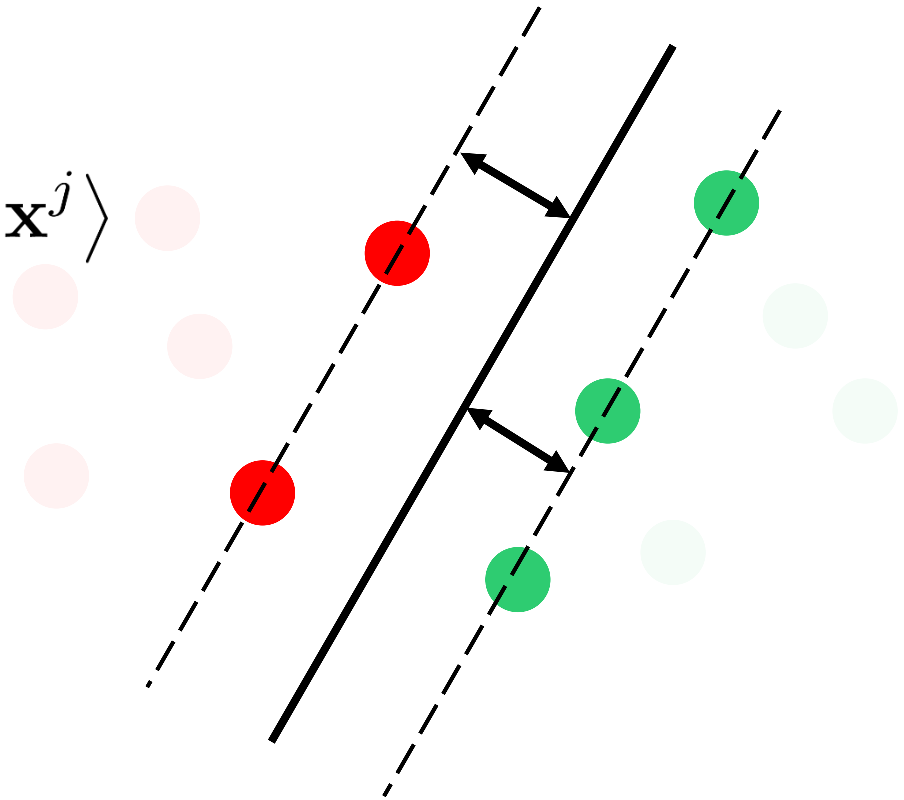
If $\hat{\boldsymbol{\alpha}}_i \neq 0$, then $y^i\langle \hat{\mathbf{w}}_{\mathrm{D}}, \mathbf{x}^i \rangle = 1$

Exercise

App: SVMs via CD

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \sum_{i=1}^{n} \boldsymbol{\alpha}_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_j y^i y^j \left\langle \mathbf{x}^i, \mathbf{x}^j \right\rangle$$

$$\text{s.t. } \boldsymbol{\alpha}_i \geq 0$$



Coordinate Descent

- Similar to gradient descent except update one variable at a time
- Notation: $\nabla_i f(\mathbf{w}) = [\nabla f(\mathbf{w})]_{\mathrm{i}}$ i.e. the i-th directional derivative

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}) + r(\mathbf{w})$$

COORDINATE DESCENT

1. Initialize $\mathbf{w}^0 \in \mathbb{R}^d$
2. Select a coordinate $j_t \in [d]$
3. Let $\mathbf{g}_{j_t}^t \leftarrow \nabla_{\mathrm{j_t}} f(\mathbf{w}^t) + \nabla_{\mathrm{j_t}} r(\mathbf{w}^t) \in \mathbb{R}$
4. Update $\mathbf{w}_{j_t}^{t+1} \leftarrow \mathbf{w}_{j_t}^t - \eta_t \cdot \mathbf{g}_{j_t}^t$
5. Preserve other coord. $\mathbf{w}_k^{t+1} \leftarrow \mathbf{w}_k^t,\ k \neq j_t$
6. Repeat until convergence

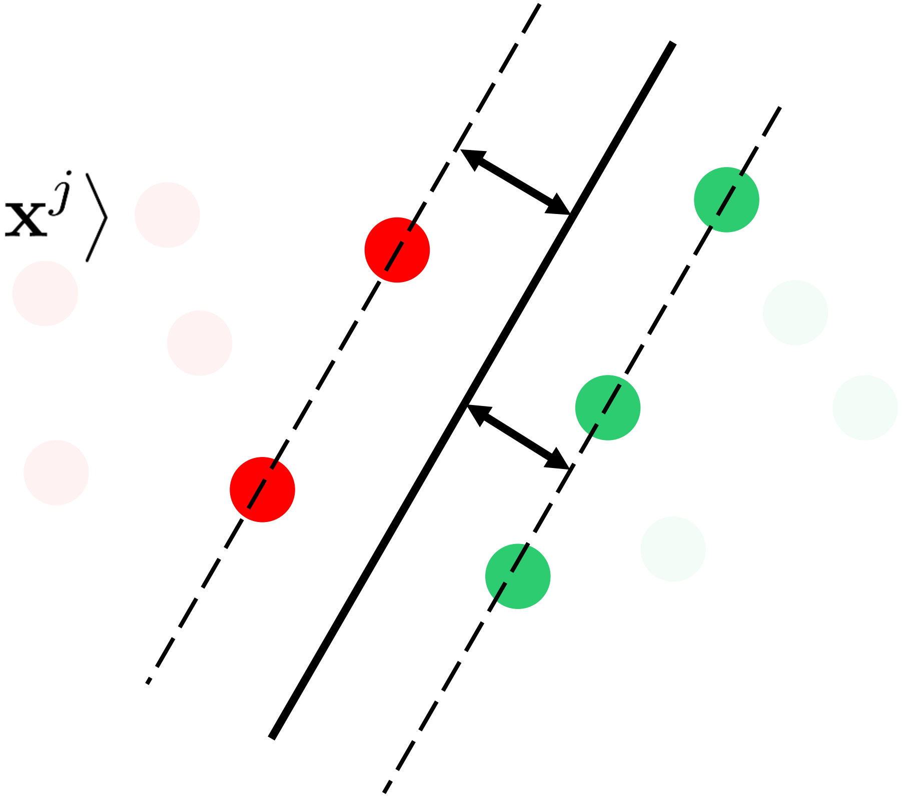
Random (Stochastic CD), Cyclic $1,2,\ldots,d,1,2,3\ldots d,\ldots$

Only $O(n)$ time per iter!

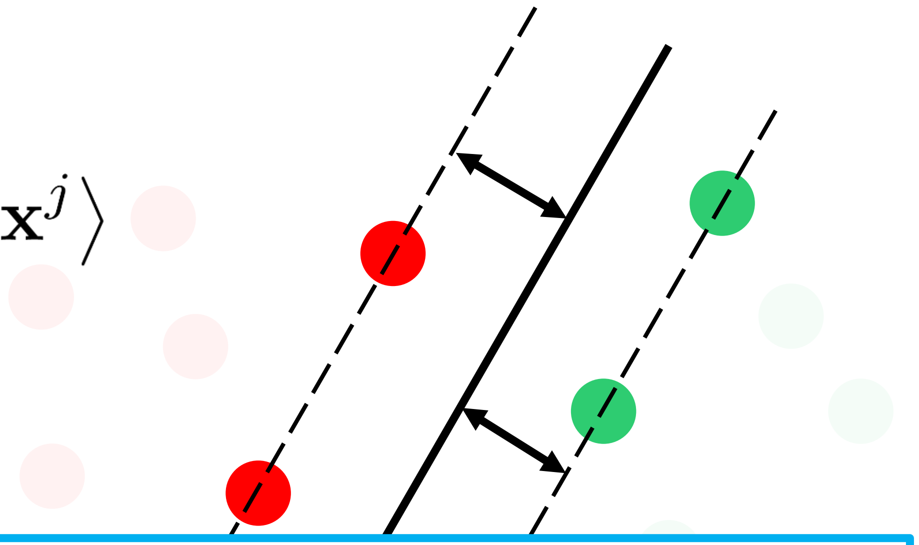
App: SVMs via CD

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha} \in \mathbb{R}^n} \sum_{i=1}^{n} \boldsymbol{\alpha}_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_j y^i y^j \left\langle \mathbf{x}^i, \mathbf{x}^j \right\rangle$$

$$\text{s.t. } \boldsymbol{\alpha}_i \geq 0$$



App: SVMs via CD



$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha} \in \mathbb{R}^n} \sum_{i=1}^{n} \boldsymbol{\alpha}_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_j y^i y^j \left\langle \mathbf{x}^i, \mathbf{x}^j \right\rangle$$

$$\text{s.t. } \boldsymbol{\alpha}_i \geq 0$$

- Choose $i_t \in [n]$ at each time step (random, cyclic etc)
- Update only $\boldsymbol{\alpha}_{i_t}$, leave all other $\boldsymbol{\alpha}_i, \mathrm{i} \neq i_t$ untouched
  - Use projected gradient descent –constraint is pretty simple
- Note: $\boldsymbol{\alpha}_{i_t}$ corresponds to the data point $\left(\mathbf{x}^{i_t}, y^{i_t}\right)$
  - Only the i_t-th data point required to perform this update - exercise
  - Stochastic CD in dual looks like stochastic GD in primal!
- Current state of the art for SVMs – Liblinear, Scikit-learn

On to Data Modelling!

Hope you had fun with FA

Will revisit this method soon

For now, back to PML …

Please give your Feedback

http://tinyurl.com/ml17-18afb