UVA CS : Machine Learning

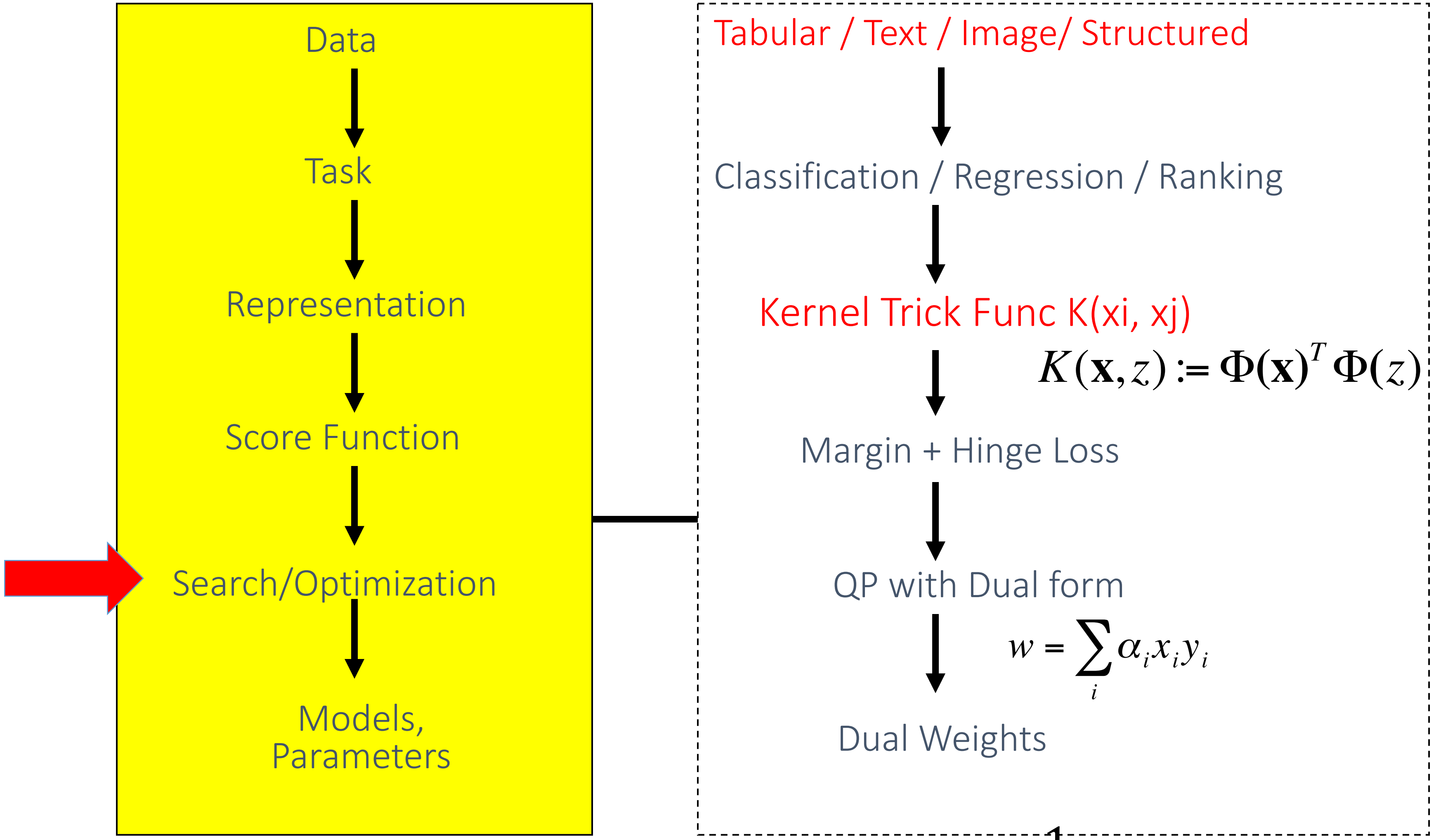
S4 Lecture 21 Extra Extra: Optimization with Dual Form and for SVM

Dr. Yanjun Qi
University of Virginia
Department of Computer Science

Today Extra

- Optimization of SVM
  - ✓ SVM as QP
  - ✓ A simple example of constrained optimization
  - ✓ SVM Optimization with dual form
  - ✓ KKT condition
  - ✓ SMO algorithm for fast SVM dual optimization

This: Kernel Support Vector Machine



$$\underset{\mathbf{w},b}{\operatorname{argmin}} \sum_{i=1}^{p} w_i^2 + C\sum_{i=1}^{n} \varepsilon_i$$

$$\text{subject to } \forall \mathbf{x}_i \in Dtrain : y_i\left(\mathbf{x}_i \cdot \mathbf{w} + b\right) \geq 1 - \varepsilon_i$$

$$\max_{\alpha} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\sum_i \alpha_i y_i = 0, \qquad \alpha_i \geq 0 \qquad \forall i$$

Optimization with Quadratic programming (QP)

Quadratic programming solves optimization problems of the following form:

$$\min_U \frac{u^T R u}{2} + d^T u + c$$

subject to n inequality constraints:

$$a_{11}u_1 + a_{12}u_2 + \ldots \le b_1$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$a_{n1}u_1 + a_{n2}u_2 + \ldots \le b_n$$

and k equivalency constraints:

$$a_{n+1,1}u_1 + a_{n+1,2}u_2 + \ldots = b_{n+1}$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$a_{n+k,1}u_1 + a_{n+k,2}u_2 + \ldots = b_{n+k}$$

$u \rightarrow$ variable

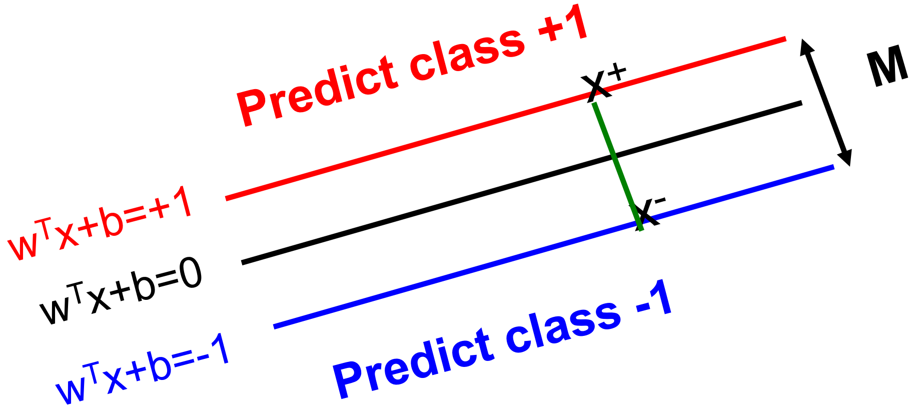
$f(u) \rightarrow$ object

Quadratic term

$g_i(u) \rightarrow$ constraints

When a problem can be specified as a QP problem we can use solvers that are better than gradient descent or simulated annealing

SVM as a QP problem



$$M = \frac{2}{\sqrt{w^T w}}$$

R as I matrix, d as zero vector, c as 0 value

$$\min_U \frac{u^T R u}{2} + d^T u + c$$

subject to n inequality constraints:

$$a_{11}u_1 + a_{12}u_2 + \ldots \le b_1$$
$$\vdots \qquad \vdots \qquad \vdots$$
$$a_{n1}u_1 + a_{n2}u_2 + \ldots \le b_n$$

and k equivalency constraints:

$$a_{n+1,1}u_1 + a_{n+1,2}u_2 + \ldots = b_{n+1}$$
$$\vdots \qquad \vdots \qquad \vdots$$
$$a_{n+k,1}u_1 + a_{n+k,2}u_2 + \ldots = b_{n+k}$$

Min $(w^T w)/2$

subject to the following inequality constraints:

For all x in class + 1

$w^T x + b >= 1$

For all x in class - 1

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

A total of n constraints if we have n input samples

Optimization Review:
Ingredients

- Objective function
- Variables
- Constraints

Find values of the variables
that minimize or maximize the objective function
while satisfying the constraints

Today Extra

- Optimization of SVM
  - ✓ SVM as QP
  - ✓ A simple example of constrained optimization and dual
  - ✓ Optimization with dual form
  - ✓ KKT condition
  - ✓ SMO algorithm for fast SVM dual optimization

Optimization Review:
Lagrangian Duality

- The Primal Problem

Primal:

$$\min_w \quad f_0(w)$$
$$\text{s.t.} \quad f_i(w) \le 0, \quad i = 1, \ldots, k$$

"Method of Lagrange multipliers"
convert to a higher-dimensional problem

The generalized Lagrangian:

$$\mathcal{L}(w,\alpha) = f_0(w) + \sum_{i=1}^{k} \alpha_i f_i(w)$$

the α's ($\alpha_i \ge 0$) are called the Lagarangian multipliers

Lemma:

$$\max_{\alpha,\alpha_i \ge 0} \mathcal{L}(w,\alpha) = \begin{cases} f_0(w) & \text{if } w \text{ satisfies primal constraints} \\ \infty & \text{o/w} \end{cases}$$

A re-written Primal:

$$\min_w \max_{\alpha,\alpha_i \ge 0} \mathcal{L}(w,\alpha)$$

Optimization Review:
Lagrangian Duality, cont.

- Recall the Primal Problem:

$$\min_w \max_{\alpha, \alpha_i \geq 0} L(w,\alpha)$$

- The Dual Problem:

$$\max_{\alpha, \alpha_i \geq 0} \min_w L(w,\alpha)$$

- **Theorem (weak duality):**

$$d^* = \max_{\alpha, \alpha_i \geq 0} \min_w L(w,\alpha) \quad \leq \quad \min_w \max_{\alpha, \alpha_i \geq 0} L(w,\alpha) = p^*$$

- **Theorem (strong duality):**

Iff there exist a saddle point of $L(w,\alpha)$

we have $d^* = p^*$

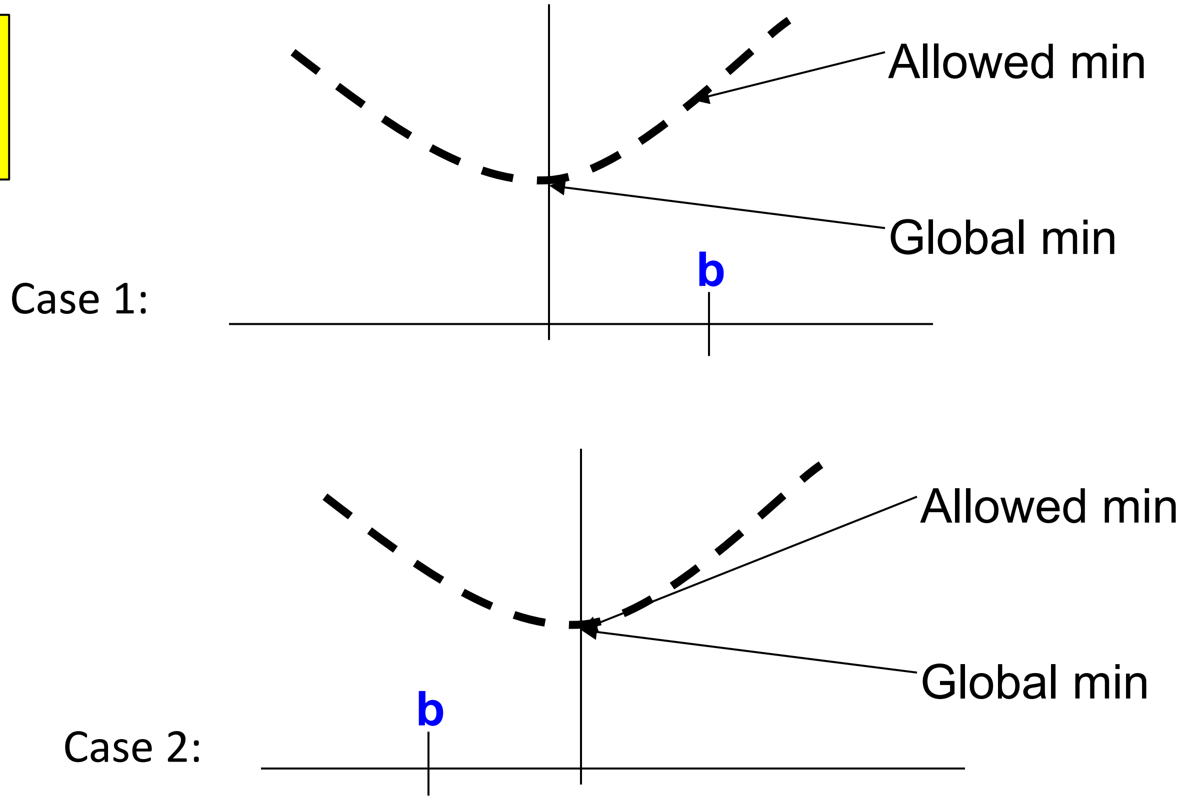
$\min_u u^2$

s.t. $u >= b$

Optimization Review:
Constrained Optimization

$\min_u u^2$

s.t. $u >= b$

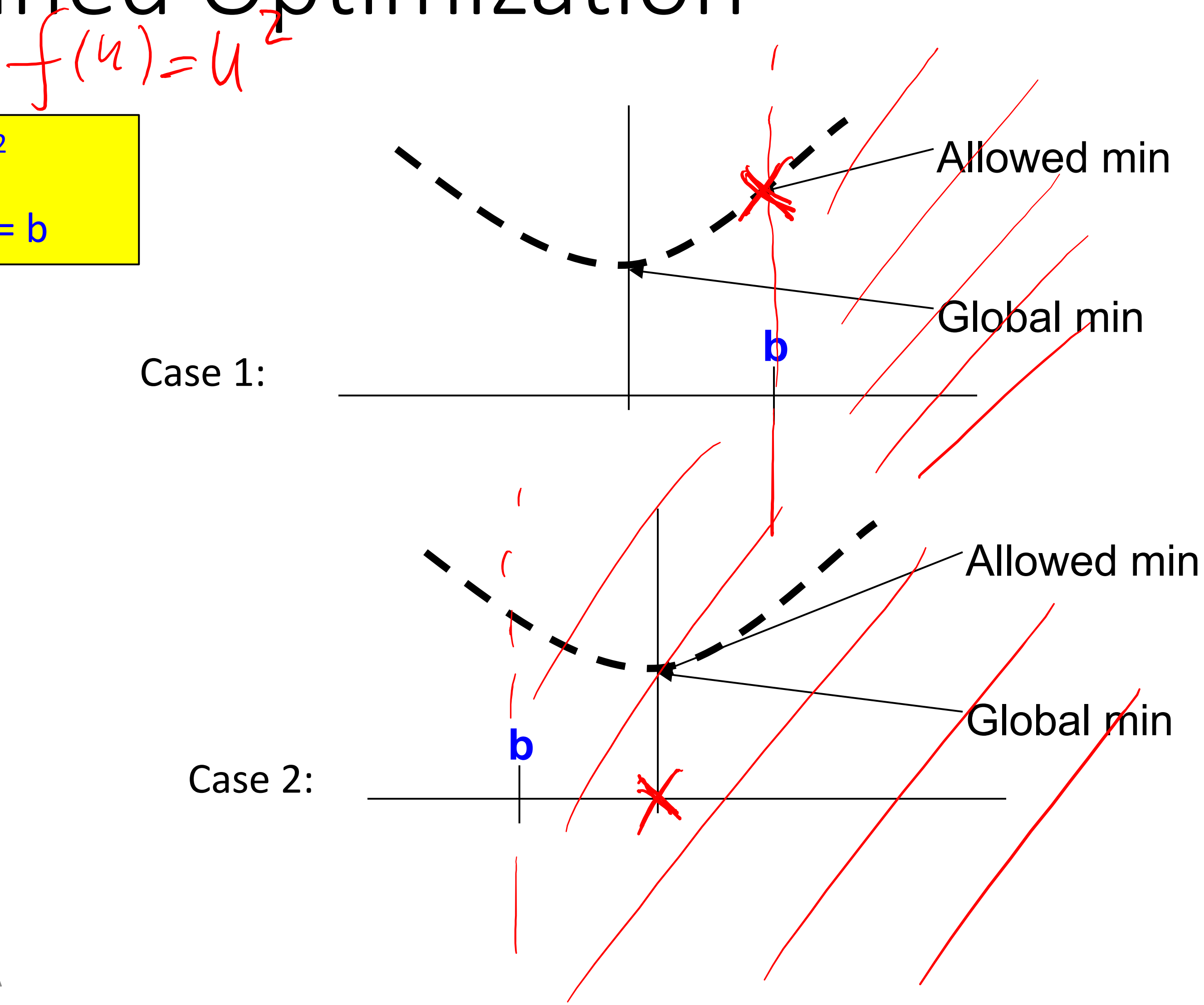


Optimization Review:

Constrained Optimization

$\min_u u^2$

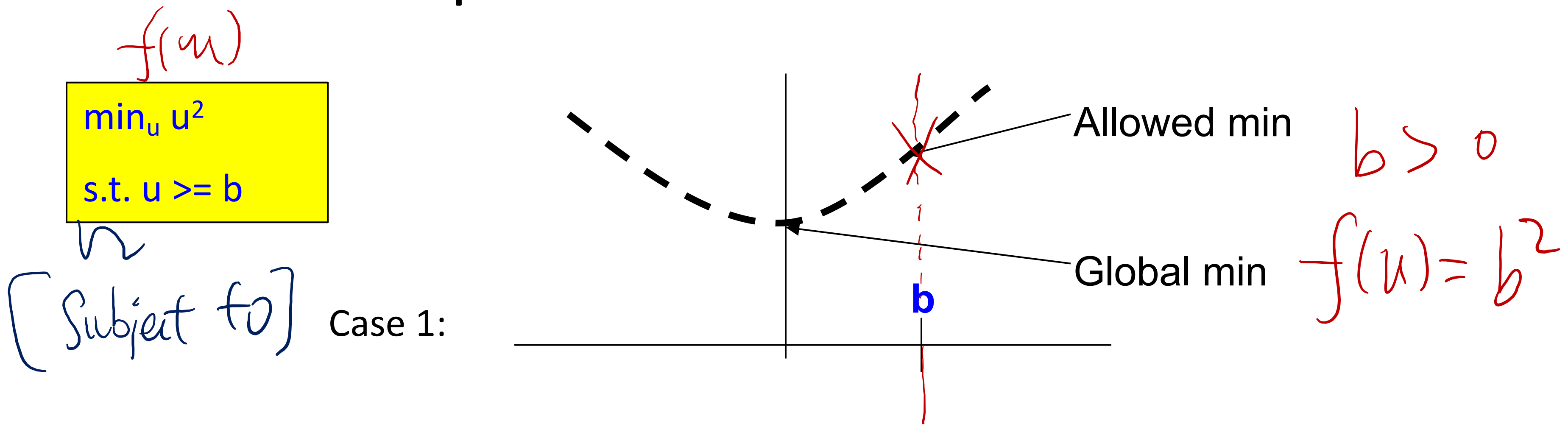
s.t. $u >= b$

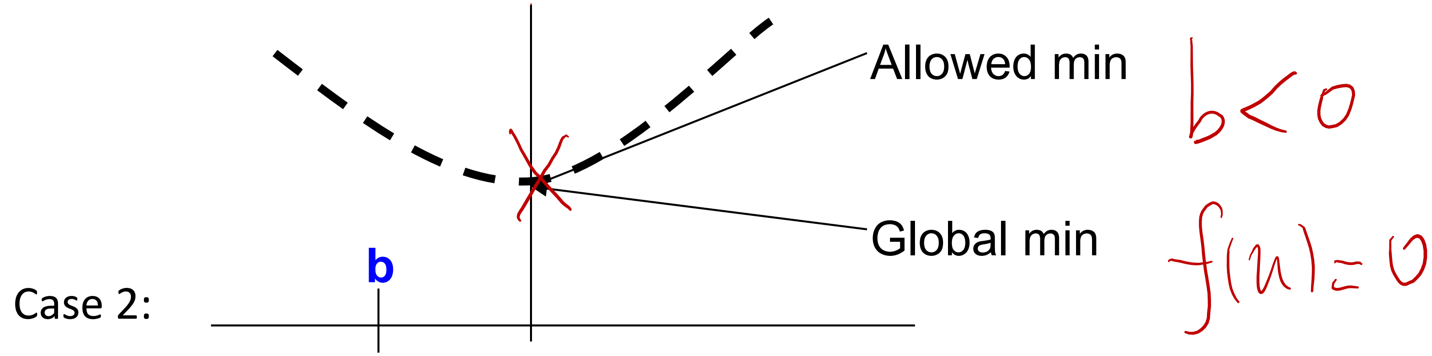


Optimization Review:
Constrained Optimization

$\min_u u^2$

s.t. $u >= b$





$\min_u u^2$

s.t. $u >= b$

$$① \begin{cases} \min_u f_0(u) = u^2 \\ \text{s.t.} \quad b - u \le 0 \end{cases}$$

primal problem

$\min_u u^2$

s.t. $u >= b$

① $\begin{cases} \min_u f_0(u) = u^2 \\ \text{s.t.} \quad b - u \le 0 \end{cases}$

⇒ multiplier variable

② $L(u, \alpha) = u^2 + \underbrace{\alpha}_{\ge 0} \underbrace{(b-u)}_{\le 0}$

u: 1×1, α: 1×1

$\min_u u^2$

s.t. $u >= b$

① $\begin{cases} \min_u f_0(u) = u^2 \\ \text{s.t.} \quad b - u \le 0 \end{cases}$

② $L(\underset{1\times1}{u}, \underset{1\times1}{\alpha}) = u^2 + \underbrace{\alpha}_{\ge 0}\underbrace{(b-u)}_{\le 0}$

③ $\frac{\partial L(u,\alpha)}{\partial u} = 2u - \alpha = 0$

$u = \frac{\alpha}{2}$

$\rightarrow \operatorname{argmin}_u L(u,\alpha)$

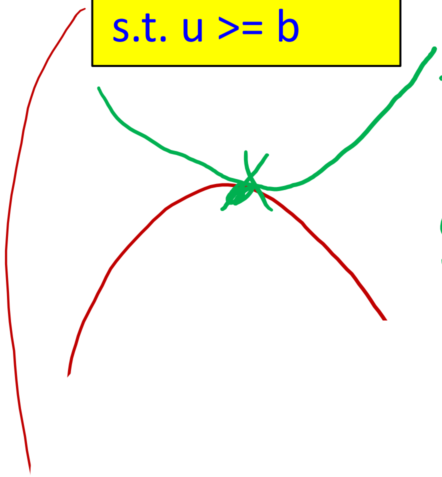
$\min_u u^2$

s.t. $u >= b$

$$g(\alpha) = L(u, \alpha) = \frac{\alpha^2}{4} + \alpha\left(b - \frac{\alpha}{2}\right)$$

$\min_u u^2$

s.t. $u >= b$



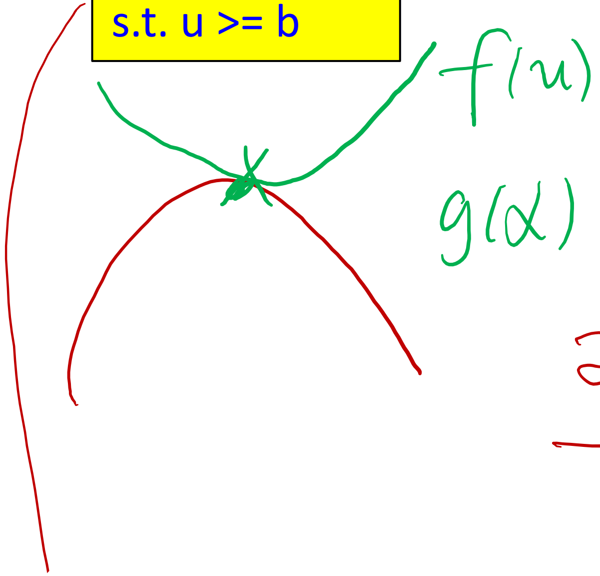
$$g(\alpha) = L(u, \alpha) = \frac{\alpha^2}{4} + \alpha\left(b - \frac{\alpha}{2}\right)$$

$$u = \alpha/2$$

$$g(\alpha) = -\frac{\alpha^2}{4} + b\alpha$$

$\min_u u^2$

s.t. $u >= b$



$$g(\alpha) = L(u, \alpha) = \frac{\alpha^2}{4} + \alpha\left(b - \frac{\alpha}{2}\right)$$

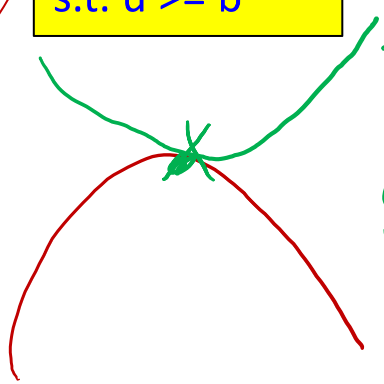
$$u = \alpha/2$$

$$g(\alpha) = -\frac{\alpha^2}{4} + b\alpha$$

$$\frac{\partial g(\alpha)}{\partial \alpha} = -\frac{\alpha}{2} + b = 0, \quad \alpha \geq 0$$

$\min_u u^2$

s.t. $u \ge b$



$$g(\alpha) = L(u,\alpha) = \frac{\alpha^2}{4} + \alpha\left(b - \frac{\alpha}{2}\right)$$

$u = \alpha/2$

$$g(\alpha) = -\frac{\alpha^2}{4} + b\alpha$$

$$\frac{\partial g(\alpha)}{\partial \alpha} = -\frac{\alpha}{2} + b = 0, \quad \alpha \ge 0$$

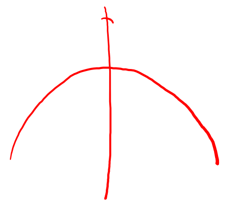
Dual $\Rightarrow$
- $b > 0$, $\alpha = 2b$, $g(\alpha) = b^2$
- $b < 0$, $\alpha = 0$, $g(\alpha) = 0$

Primal $\Rightarrow$
- $b > 0$, $f(u) = b^2$, $u = b$
- $b < 0$, $f(u) = 0$, $u = 0$

Primal : $\min_{w} \max_{\alpha} L(w, \alpha)$

Dual : $\max_{\alpha} \min_{w} L(w, \alpha)$

$\Rightarrow \max_{\alpha} g(\alpha)$



$$f(u): \begin{cases} \min u^2 \\ s.t. \quad u \geq b \end{cases}$$

$$g(\alpha): \begin{cases} \max -\frac{\alpha^2}{4} + b\alpha = \max\left\{-\left(\frac{\alpha}{2}-b\right)^2 + b^2\right\} \\ s.t. \quad \alpha \geq 0 \end{cases}$$

$$\begin{cases} \text{if}_{\ b \geq 0} \ \ b = \alpha/2, \quad u = b, \quad g = b^2 \\ \text{if}_{\ b<0} \ \ b \neq \alpha/2, \quad \alpha = 0, \quad g = 0 \end{cases}$$

$$\Rightarrow \quad \alpha(b-u) = 0$$

KKT condition

Today Extra

- Optimization of SVM
  - ✓ SVM as QP
  - ✓ A simple example of constrained optimization
  - ✓ SVM Optimization with dual form
  - ✓ KKT condition
  - ✓ SMO algorithm for fast SVM dual optimization

$$\min_{w,b} \max_{\alpha} \frac{w^T w}{2} - \sum_i \alpha_i [(w^T x_i + b) y_i - 1]$$

$$\alpha_i \geq 0 \qquad \forall i$$

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w - \sum_i \alpha_i x_i y_i = 0$$

train

$$\min_{w,b} \max_{\alpha} \left\{ \frac{w^T w}{2} - \sum_i \alpha_i [(w^T x_i + b) y_i - 1] \right\} \Rightarrow \max_{\alpha} \min_{w,b} L(w,b,\alpha)$$

$$\alpha_i \geq 0 \qquad \forall i$$

train

$$\begin{cases} \frac{\partial L}{\partial w} = 0 \Rightarrow w - \sum_i \alpha_i x_i y_i = 0 \\ \frac{\partial L}{\partial b} = 0 \Rightarrow \sum \alpha_i y_i = 0 \end{cases}$$

$$L_{primal} = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i \left(y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1 \right)$$

$$L_{dual} = \frac{1}{2} \left(\sum \alpha_i x_i y_i\right)^T \left(\sum_j \alpha_j x_j y_j\right) - \sum_i \alpha_i y_i \left(\sum_j \alpha_j x_j y_j\right)^T x_i$$

$$- \underbrace{\sum \alpha_i y_i}_{0} b + \sum \alpha_i$$

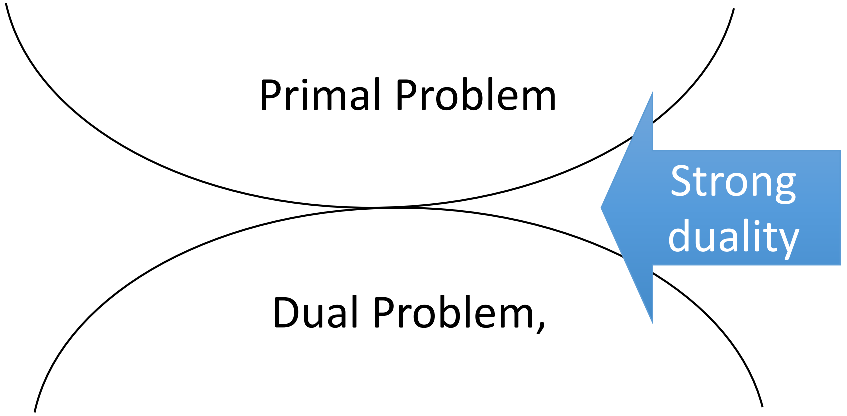
$$= \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \left(x_i^T x_j \right)$$

Optimization Review: Dual Problem

- Solving dual problem if the dual form is easier than primal form

- Need to change primal minimization to dual maximization (OR → Need to change primal maximization to dual minimization)

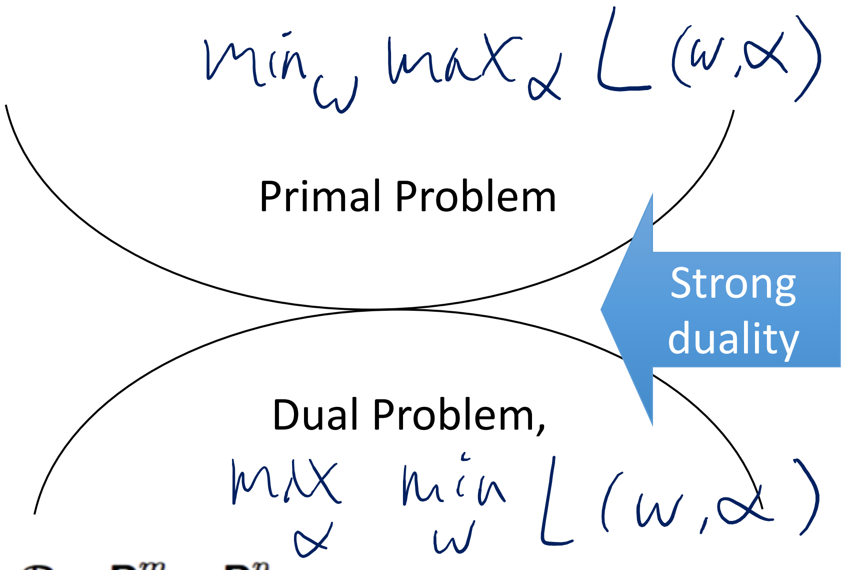
- Only valid when the original optimization problem is convex/concave (strong duality)



Today Extra

- Optimization of SVM
  - ✓ SVM as QP
  - ✓ A simple example of constrained optimization
  - ✓ SVM Optimization with dual form
  - ✓ KKT condition
  - ✓ SMO algorithm for fast SVM dual optimization

KKT Condition for Strong Duality



$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p \end{aligned}$$

Lagrangian: $L : \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}$, with $\mathbf{dom}\, L = \mathcal{D} \times \mathbf{R}^m \times \mathbf{R}^p$,

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

complementary slackness: $\lambda_i f_i(x) = 0,\ i = 1, \ldots, m$

Key for SVM Dual

Optimization Review: Lagrangian (even more general standard form)

standard form problem (not necessarily convex)

$$
\begin{array}{lll}
\text{minimize} & f_0(x) & \\
\text{subject to} & f_i(x) \le 0, & i = 1, \ldots, m \\
 & h_i(x) = 0, & i = 1, \ldots, p
\end{array}
$$

variable $x \in \mathbf{R}^n$, domain $\mathcal{D}$, optimal value $p^\star$

Lagrangian: $L : \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}$, with $\mathbf{dom}\, L = \mathcal{D} \times \mathbf{R}^m \times \mathbf{R}^p$,

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

- weighted sum of objective and constraint functions
- λ_i is Lagrange multiplier associated with $f_i(x) \le 0$
- ν_i is Lagrange multiplier associated with $h_i(x) = 0$

From Stanford "Convex Optimization 5. Duality — Boyd & Vandenberghe

Optimization Review: Lagrange dual function

Lagrange dual function: $g : \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}$,

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu)$$
$$= \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right)$$

g is concave, can be $-\infty$ for some λ, ν

lower bound property: if $\lambda \succeq 0$, then $g(\lambda, \nu) \leq p^\star$

Inf(.): greatest lower bound

proof: if $\tilde{x}$ is feasible and $\lambda \succeq 0$, then

$$f_0(\tilde{x}) \geq L(\tilde{x}, \lambda, \nu) \geq \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) = g(\lambda, \nu)$$

minimizing over all feasible $\tilde{x}$ gives $p^\star \geq g(\lambda, \nu)$

From Stanford "Convex Optimization 5. Duality — Boyd & Vandenberghe

Optimization Review:

Complementary slackness

assume strong duality holds, $x^\star$ is primal optimal, $(\lambda^\star, \nu^\star)$ is dual optimal

inf (.): greatest lower bound

$$
\begin{aligned}
f_0(x^\star) = g(\lambda^\star, \nu^\star) \quad &= \quad \inf_x \left(f_0(x) + \sum_{i=1}^{m} \lambda_i^\star f_i(x) + \sum_{i=1}^{p} \nu_i^\star h_i(x) \right) \\
&\leq \quad f_0(x^\star) + \sum_{i=1}^{m} \lambda_i^\star f_i(x^\star) + \sum_{i=1}^{p} \nu_i^\star h_i(x^\star) \\
&\leq \quad f_0(x^\star)
\end{aligned}
$$

hence, the two inequalities hold with equality

- $x^\star$ minimizes $L(x, \lambda^\star, \nu^\star)$
- $\lambda_i^\star f_i(x^\star) = 0$ for $i = 1, \ldots, m$ (known as complementary slackness):

$$
\lambda_i^\star > 0 \Longrightarrow f_i(x^\star) = 0, \qquad f_i(x^\star) < 0 \Longrightarrow \lambda_i^\star = 0
$$

From Stanford “Convex Optimization 5. Duality — Boyd & Vandenberghe

Optimization Review:

Complementary slackness

assume strong duality holds, $x^\star$ is primal optimal, $(\lambda^\star, \nu^\star)$ is dual optimal

inf (.): greatest lower bound

$$
\begin{aligned}
f_0(x^\star) = g(\lambda^\star, \nu^\star) &= \inf_x \left(f_0(x) + \sum_{i=1}^{m} \lambda_i^\star f_i(x) + \sum_{i=1}^{p} \nu_i^\star h_i(x) \right) \\
&\le f_0(x^\star) + \sum_{i=1}^{m} \lambda_i^\star f_i(x^\star) + \sum_{i=1}^{p} \nu_i^\star h_i(x^\star) \\
&\le f_0(x^\star)
\end{aligned}
$$

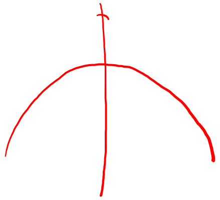
hence, the two inequalities hold with equality

- $x^\star$ minimizes $L(x, \lambda^\star, \nu^\star)$
- $\lambda_i^\star f_i(x^\star) = 0$ for $i = 1, \ldots, m$ (known as complementary slackness):

$$
\lambda_i^\star > 0 \Longrightarrow f_i(x^\star) = 0, \qquad f_i(x^\star) < 0 \Longrightarrow \lambda_i^\star = 0
$$

From Stanford “Convex Optimization 5. Duality — Boyd & Vandenberghe

$$f(u): \begin{cases} \min u^2 \\ \text{s.t. } u \geq b \end{cases}$$



$$g(\alpha): \begin{cases} \max -\frac{\alpha^2}{4} + b\alpha = \max\left\{-\left(\frac{\alpha}{2}-b\right)^2 + b^2\right\} \\ \text{s.t. } \alpha \geq 0 \end{cases}$$

$$\begin{cases} \text{if}_{\,b \geq 0} \; b = \alpha/2, & u = b, \; g = b^2 \\ \text{if}_{\,b < 0} \; b \neq \alpha/2, & \alpha = 0, \; g = 0 \end{cases}$$

$$\Rightarrow \alpha(b - u) = 0$$

KKT condition

Optimization Review:

Karush-Kuhn-Tucker (KKT) conditions

the following four conditions are called KKT conditions (for a problem with differentiable f_i, h_i):

1. primal constraints: $f_i(x) \leq 0$, $i = 1, \ldots, m$, $h_i(x) = 0$, $i = 1, \ldots, p$
2. dual constraints: $\lambda \succeq 0$
3. complementary slackness: $\lambda_i f_i(x) = 0$, $i = 1, \ldots, m$
4. gradient of Lagrangian with respect to x vanishes:

Key for SVM Dual

$$\nabla f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) + \sum_{i=1}^{p} \nu_i \nabla h_i(x) = 0$$

from page 5–17: if strong duality holds and x, λ, ν are optimal, then they must satisfy the KKT conditions

From Stanford “Convex Optimization 5. Duality — Boyd & Vandenberghe

Dual formulation for linearly non separable case (soft SVM)

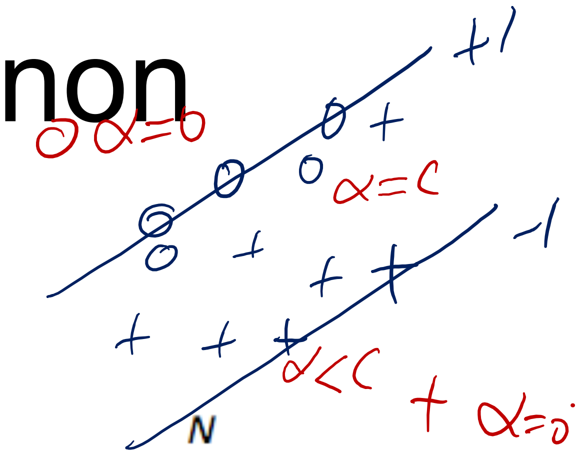
Substituting (1), (2), and (3) into the Lagrange, we have:

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N} \alpha_i \alpha_k y_i y_k x_i^T x_k, \text{ with } 0 \le \alpha_i \le C \text{ and } \sum_{i=1}^{N} \alpha_i y_i = 0. \quad (4)$$

- $\hat{\alpha}_i > 0$: which implies $y_i(x_i^T \hat{\mathbf{w}} + \hat{b}) - 1 + \hat{\xi}_i = 0$ according to (5). These points are the *support vectors*.
  - $\hat{\xi}_i = 0$: which implies $\hat{\mu}_i > 0$ from (6) and so $\hat{\alpha}_i < C$ from (3). There are the support points which lie on the edge of the margin.
  - $\hat{\xi}_i > 0$: which implies $\hat{\mu}_i = 0$ from (6) and so $\hat{\alpha}_i = C$ from (3). There are the support points which violate the margin.
- $\hat{\alpha}_i = 0$: These points are not support vectors, which play no role in determining the hyperplane.

From Stanford "Convex Optimization 5. Duality — Boyd & Vandenberghe

Dual formulation for linearly non separable case



Substituting (1), (2), and (3) into the Lagrange, we have:

$$L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \alpha_i \alpha_k y_i y_k x_i^T x_k, \text{ with } 0 \leq \alpha_i \leq C \text{ and } \sum_{i=1}^{N} \alpha_i y_i = 0. \quad (4)$$

- $\hat{\alpha}_i > 0$: which implies $y_i(x_i^T \hat{\mathbf{w}} + \hat{b}) - 1 + \hat{\xi}_i = 0$ according to (5). These points are the *support vectors*.
  - $\hat{\xi}_i = 0$: which implies $\hat{\mu}_i > 0$ from (6) and so $\hat{\alpha}_i < C$ from (3). There are the support points which lie on the edge of the margin.
  - $\hat{\xi}_i > 0$: which implies $\hat{\mu}_i = 0$ from (6) and so $\hat{\alpha}_i = C$ from (3). There are the support points which violate the margin.
- $\hat{\alpha}_i = 0$: These points are not support vectors, which play no role in determining the hyperplane.

From Stanford “Convex Optimization 5. Duality — Boyd & Vandenberghe

Support Vectors for the Soft-case

- Support vectors are
  - Samples on the margin: $y_i(\mathbf{x}_i \cdot \mathbf{w} + b) = 1,$
    $0 < \alpha_i < C$
  - Sample violate (mostly inside the margin area):

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) < 1,$$
$$\alpha_i = C$$

Today Extra

- Optimization of SVM
  - ✓ SVM as QP
  - ✓ A simple example of constrained optimization
  - ✓ SVM Optimization with dual form
  - ✓ KKT condition
  - ✓ SMO algorithm for fast SVM dual optimization

Fast SVM Implementations

- SMO: Sequential Minimal Optimization
- SVM-Light
- LibSVM
- BSVM
- ……

J. Platt (1999),
Fast Training of Support Vector Machines Using Sequential Minimal Optimization
https://pdfs.semanticscholar.org/d1fa/8485ad749d51e7470d801bc1931706597601.pdf

SMO: Sequential Minimal Optimization

- Key idea
  - Divide the large QP problem of SVM into a series of smallest possible QP problems, which can be solved analytically and thus avoids using a time-consuming numerical QP in the loop (a kind of SQP method).
  - Space complexity: O(n).
  - Since QP is greatly simplified, most time-consuming part of SMO is the evaluation of decision function, therefore it is very fast for linear SVM and sparse data.

SMO

- At each step, SMO chooses 2 Lagrange multipliers to jointly optimize, find the optimal values for these multipliers and updates the SVM to reflect the new optimal values.
- Three components
  - An analytic method to solve for the two Lagrange multipliers
  - A heuristic for choosing which (next) two multipliers to optimize
  - A method for computing b at each step, so that the KTT conditions are fulfilled for both the two examples (corresponding to the two multipliers)

Choosing Which Multipliers to Optimize

- First multiplier
  - Iterate over the entire training set, and find an example that violates the KTT condition.
- Second multiplier
  - Maximize the size of step taken during joint optimization.
  - $|E_1 - E_2|$, where E_i is the error on the *i*-th example.

References

- Big thanks to Prof. Ziv Bar-Joseph and Prof. Eric Xing @ CMU for allowing me to reuse some of his slides
- Elements of Statistical Learning, by Hastie, Tibshirani and Friedman
- Prof. Andrew Moore @ CMU's slides
- Tutorial slides from Dr. Tie-Yan Liu, MSR Asia
- A Practical Guide to Support Vector Classification Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, 2003-2010
- Tutorial slides from Stanford "Convex Optimization I — Boyd & Vandenberghe