Optimizing SVM Multiclass Dual Solvers

Using GPU to power fast SVM solvers

Advisor: **Prof. Yutong Wang**

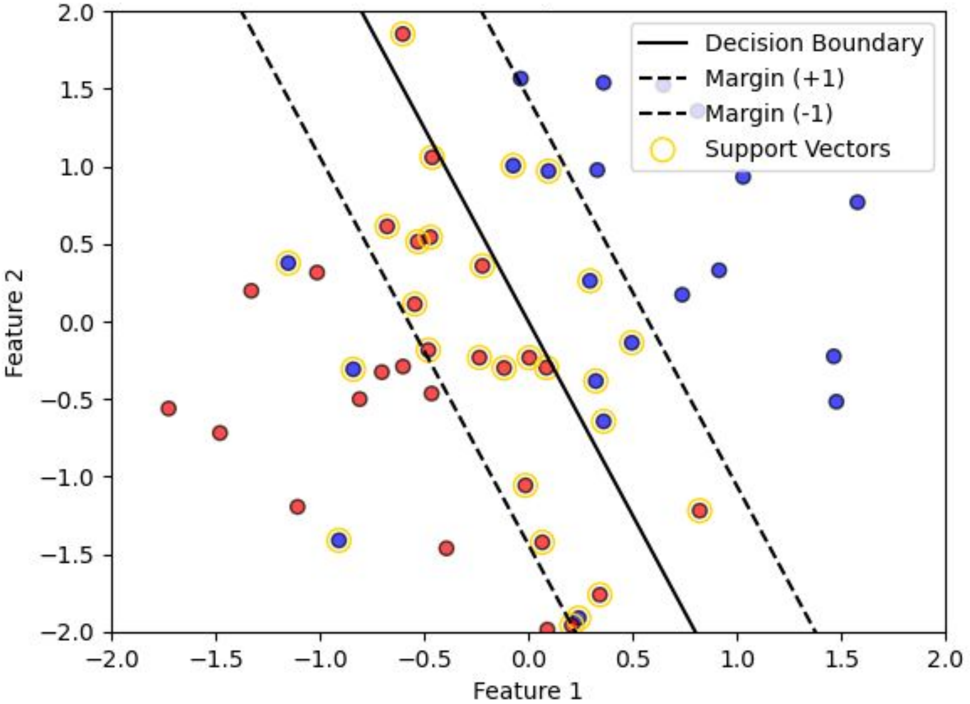
Student: **Vinay Prabhakar**

EMC² Research Lab

Binary Classification - Soft Margin SVM

Setup

$$h(x) = \text{sign}(w^\top x)$$



Symbols

$n :$ Number of training samples

$d :$ Dimension of feature space

$x_i \in \mathbb{R}^d :$ Feature vector for the i^th sample

$y_i \in \{-1, 1\} :$ Class label for the ith sample

$w \in \mathbb{R}^d :$ Weight vector

$\epsilon_i \geq 0 :$ Slack variable for the ith sample

$C > 0 :$ Regularization parameter (penalty for misclassification)

Binary Classification Primal Solver

Objective Function

$$\min_{w,\epsilon_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\epsilon_i$$

subject to,

Classification Constraint:

$$y_i(w \cdot x_i) \geq 1 - \epsilon_i, \quad \forall i \qquad \text{—(1)}$$

Slack Constraint:

$$\epsilon_i \geq 0, \quad \forall i \qquad \text{—(2)}$$

Hinge Loss

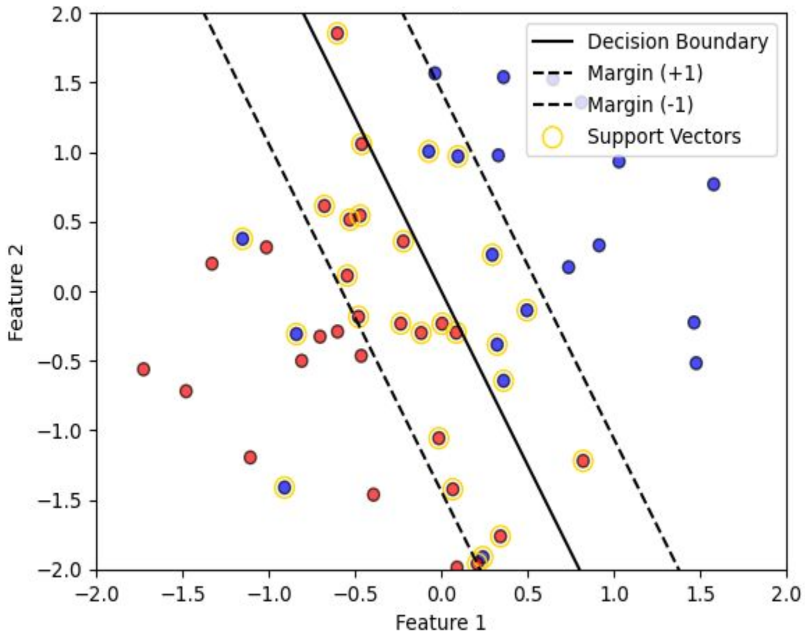
$$y_i(w^\top x_i) \geq 1 - \epsilon_i$$
$$\epsilon_i \geq 1 - y_i(w^\top x_i)$$
$$\epsilon_i \geq \max(0, 1 - y_i(w^\top x_i))$$
$$\epsilon_i = \max(0, 1 - y_i(w^\top x_i))$$
$$\ell_{\text{hinge}}(x_i, y_i) = \max(0, 1 - y_i(w^\top x_i))$$

Slack Interpretation

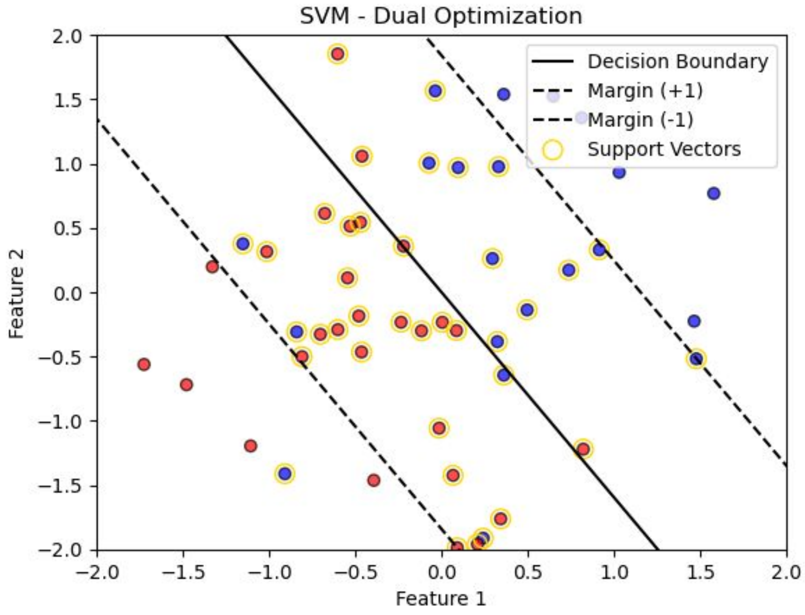
Condition on ϵ_i	**Interpretation w.r.t. Margin**
$\epsilon_i = 0$	Outside or on the correct margin
$0 < \epsilon_i < 1$	Inside the margin
$\epsilon_i = 1$	On the decision boundary
$\epsilon_i > 1$	On the wrong side of the hyperplane

Binary Classification Primal Solver - Training with different ‘C’

C=1



C=0.1



Binary Classification Primal Solver - Algorithm

Algorithm 1 Stochastic Gradient Descent for Primal SVM $(X, y, C, \eta, num\_epochs)$

Initialize $w \leftarrow \mathbf{0} \in \mathbb{R}^d$
for $t = 1$ to T **do**
    Compute margin: $m_i \leftarrow y_i(w^\top x_i)$ for all i
    Initialize gradient: $g \leftarrow \mathbf{0}$
    for $i = 1$ to n **do**
        if $m_i < 1$ **then**
            $g \leftarrow g - C \cdot y_i \cdot x_i$
        end if
    end for
    $g \leftarrow g + w$
    $w \leftarrow w - \eta \cdot g$
end for
return w

Dual Solver - Formulation

Step 1: Lagrangian Form

Introducing multipliers $\alpha_i \geq 0$, $\beta_i \geq 0$, we have:

$$\mathcal{L}(w, \epsilon_i, \alpha_i, \beta_i) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\epsilon_i - \sum_{i=1}^{n}\alpha_i[y_i(w \cdot x_i) - 1 + \epsilon_i] - \sum_{i=1}^{n}\beta_i\epsilon_i$$

Step 2: KKT Conditions - Stationarity

Set derivatives w.r.t w, ϵ_i to zero:

$$\frac{\partial \mathcal{L}}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{n}\alpha_i y_i x_i \qquad \text{—(3)}$$

$$\frac{\partial \mathcal{L}}{\partial \epsilon_i} = 0 \Rightarrow C - \alpha_i - \beta_i = 0 \Rightarrow \alpha_i + \beta_i = C, \quad \forall i \qquad \text{—(4)}$$

Step 3: KKT Conditions - Complementary Slackness

$$\alpha_i[y_i(w \cdot x_i) - 1 + \epsilon_i] = 0 \qquad \text{—(5)}$$

$$\beta_i\epsilon_i = 0 \qquad \text{—(6)}$$

Dual Solver - Formulation

Step 4: Substitute Stationarity into Lagrangian
Replace w using (3), β using (4) and simplify:

$$\mathcal{L}(\alpha_i) = \frac{1}{2}\left\|\sum_{i=1}^{n} \alpha_i y_i x_i\right\|^2 + \sum_{i=1}^{n}(C - \alpha_i - \beta_i)\epsilon_i - \sum_{i=1}^{n} \alpha_i[y_i(\sum_{j=1}^{n} \alpha_j y_j x_j \cdot x_i) - 1]$$

Since $C - \alpha_i - \beta_i = 0$, ϵ_i terms vanish:

$$\mathcal{L}(\alpha_i) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

Step 5: Dual Form
Thus, the dual optimization explicitly is:

$$\max_{\alpha_i} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

subject to *(from step 1 & 2)*:

$$0 \leq \alpha_i \leq C, \quad \forall i$$

Dual Solver - Algorithm

Algorithm 2 Coordinate Ascent for Dual SVM (X, y, C, num_epochs)

Initialize $\alpha_i \sim \mathcal{U}(0, C)$ for $i = 1, \ldots, n$

Initialize gradient: $g \leftarrow \mathbf{1} - \left(y_i y_j (x_i^\top x_j)\right) \alpha$

for $t = 1$ to num_epochs **do**

    for each i in a random permutation of $\{1, \ldots, n\}$ **do**

        Compute coefficients a, b using $\boxed{\text{cached gradient}}$

        $\alpha_i^{\text{new}} \leftarrow \arg\max_{x \in [0, C]} ax^2 + bx$

        $\Delta \leftarrow \alpha_i^{\text{new}} - \alpha_i$

        $\alpha_i \leftarrow \alpha_i^{\text{new}}$

        Update g using Δ

    end for

end for

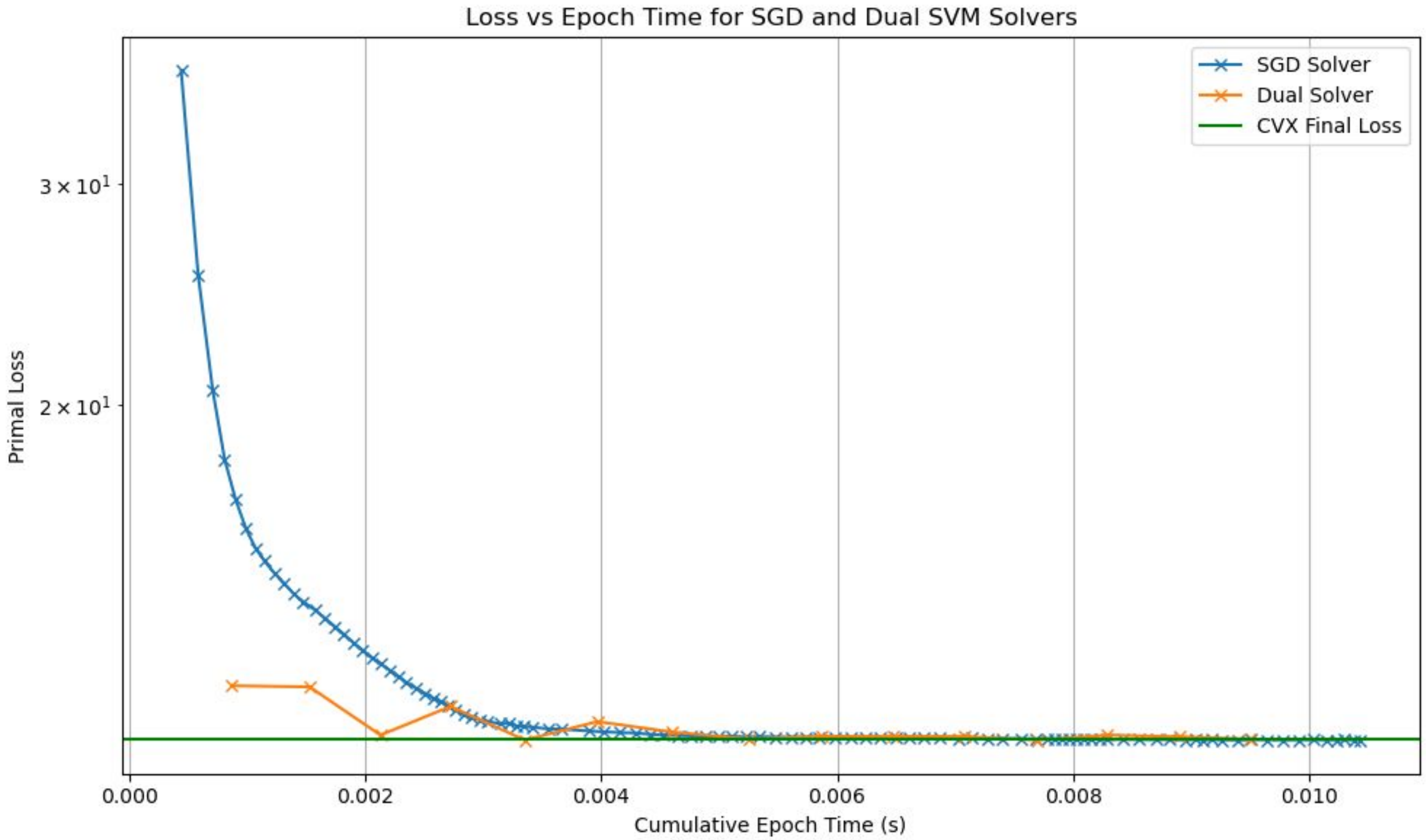
return α, w

Dual Solver - Algorithm

COEFFICIENTS VISUALIZATION

Dual Solver - Algorithm

Observations



Why Dual Solver over Primal ?

- Constraints are easier to handle
- Efficient in higher dimensions when d>>>n
- More stable in large scale data
- Converges faster

Reference: Abe, Shigeo. (2009). Is Primal Better Than Dual. 854-863. 10.1007/978-3-642-04274-4_88.

Multiclass SVM Approaches - Binary Reduction vs. Single Loss Function

Binary Reduction

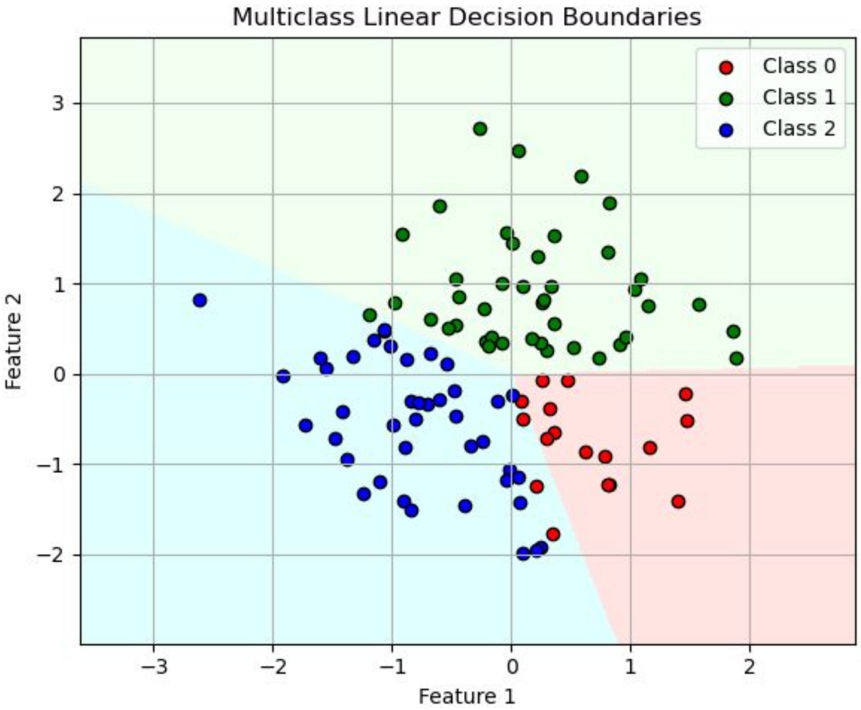
- Reduce into subproblems
- One vs. One - ${}^{k}C_{2}$ classifiers
- One vs. All - k classifiers

Unified Optimization

- Single function
- Consider all classes at once
- Better interclass relationships

Multi Classification - Soft Margin SVM

Setup $$h(x) = \arg\max_{m \in [k]} w_m^\top x$$



Symbols

n : Number of training samples

d : Dimension of feature space

k : Number of classes

$x_i \in \mathbb{R}^d$: Feature vector for the i^{th} sample

$y_i \in \{0, 1, \ldots, k-1\}$: Class label for the i^{th} sample

$w_m \in \mathbb{R}^d$: Weight vector for the m^{th} class

$\xi_i \geq 0$: Slack variable for the i^{th} sample

$C > 0$: Regularization parameter

Crammer Singer SVM

Objective Function

$$\min_{\{w_m\},\{\xi_i\}} \frac{1}{2}\sum_{m=1}^{k} \|w_m\|^2 + C\sum_{i=1}^{n}\xi_i$$

subject to,

Classification Constraint:

$$w_{y_i}^\top x_i - w_m^\top x_i \geq 1 - \xi_i,\ \forall m \neq y_i \quad \text{—(1)}$$

Slack Constraint:

$$\xi_i \geq 0,\ \forall i \quad \text{—(2)}$$

Multiclass Hinge Loss

$$w_{y_i}^\top x_i - w_m^\top x_i \geq 1 - \xi_i,\ \forall m \neq y_i$$

Penalizing only the largest violator class
(Highest scoring class where, $m \neq y_i$),

$$\max_{m\neq y_i} w_{y_i}^\top x_i - w_m^\top x_i \geq 1 - \xi_i,\ \forall m \neq y_i$$

$$\xi_i \geq 1 + \max_{m\neq y_i} w_m^\top x_i - w_{y_i}^\top x_i$$

$$\xi_i \geq \max\left(0, 1 + \max_{m\neq y_i} w_m^\top x_i - w_{y_i}^\top x_i\right)$$

$$\xi_i = \max\left(0, 1 + \max_{m\neq y_i} w_m^\top x_i - w_{y_i}^\top x_i\right)$$

$$\ell_{\text{hinge}}^{\text{multi}}(x_i, y_i) = \max\left(0, 1 + \max_{m\neq y_i} w_m^\top x_i - w_{y_i}^\top x_i\right)$$

Dual Formulation

Primal Form

$$\min_{\{w_m\}_{m=1}^{k}} \frac{1}{2}\sum_{m=1}^{k}\|w_m\|^2 + C\sum_{i=1}^{n}\left[1+\max_{m\neq y_i} w_m^\top x_i - w_{y_i}^\top x_i\right]_+$$

$$\text{s.t. } w_{y_i}^\top x_i - w_m^\top x_i \geq 1-\xi_i,\ \forall m \neq y_i$$

$$\xi_i \geq 0,\ \forall i$$

Dual Form

$$\min_{\alpha} f(\alpha) = \frac{1}{2}\sum_{m=1}^{k}\left\|\sum_{i=1}^{n}\alpha_i^m x_i\right\|^2 + \sum_{i=1}^{n}\sum_{m=1}^{k}\Delta_i^m \alpha_i^m$$

$$\text{subject to: } \sum_{m=1}^{k}\alpha_i^m = 0, \quad \alpha_i^m \leq C_i^m,\ \forall i$$

$$\Delta_i^m = 0 \text{ if } m = y_i \text{ else1}, \quad C_i^m = C \text{ when } i = y_i \text{ else } 0$$

Blondel SVM Algorithm

Algorithm 3 Blondel Multiclass SVM Dual Solver $(X, y, C, num\_epochs, \epsilon)$

Initialize $\alpha \leftarrow \mathbf{0} \in \mathbb{R}^{n \times K}$, $W \leftarrow \mathbf{0} \in \mathbb{R}^{K \times d}$
Precompute x_i norms: $||x_i||^2$ for all i
for $t = 1$ to num_epochs **do**
$v_{\max} \leftarrow 0$
for $i = 1$ to n **do**
$g_i \leftarrow Wx_i + 1$; $g_i[y_i] \leftarrow g_i[y_i] - 2$
Compute C_i: C at index y_i, 0 elsewhere
Compute violation $v_i \leftarrow \max(g_i) - \min\{g_m \mid \alpha_i[m] < C_i[m]\}$
$v_{\max} \leftarrow \max(v_{\max}, v_i)$
if $v_i \leq \epsilon$ **then**
continue ▷ Skip if no violation
end if
$\hat{\beta} \leftarrow \|x_i\|(C_i - \alpha_i) + \frac{g_i}{\|x_i\|}$
$z \leftarrow C \cdot \|x_i\|$
$\beta \leftarrow \texttt{SimplexProjection}(\hat{\beta}, z)$
$\delta_i \leftarrow C_i - \alpha_i - \frac{\beta}{\|x_i\|}$
$\alpha_i \leftarrow \alpha_i + \delta_i$
for $m = 1$ to K **do**
$W_m \leftarrow W_m + \delta_i[m] \cdot x_i$
end for
end for
Record α, W, loss, and time
end for
return α, W

Filtering samples for sub-problem

Calculate gradient

$$g_i^m = w_m^\top x_i + \Delta_i^m$$

Check for violation

KTT Conditions satisfaction

$$\begin{aligned} \alpha_i^m = 0 \quad &\Rightarrow \quad g_i^m \leq 0 \\ 0 < \alpha_i^m < C \quad &\Rightarrow \quad g_i^m = 0 \\ \alpha_i^m = C \quad &\Rightarrow \quad g_i^m \geq 0 \end{aligned}$$

Violation

$$v_i = \max_m g_i^m - \min_{m:\alpha_i^m < C} g_i^m$$

Restricted Sub Problem

Solve introducing a small change

$$\delta_i := \alpha_i^{\text{new}} - \alpha_i^{\text{current}}$$

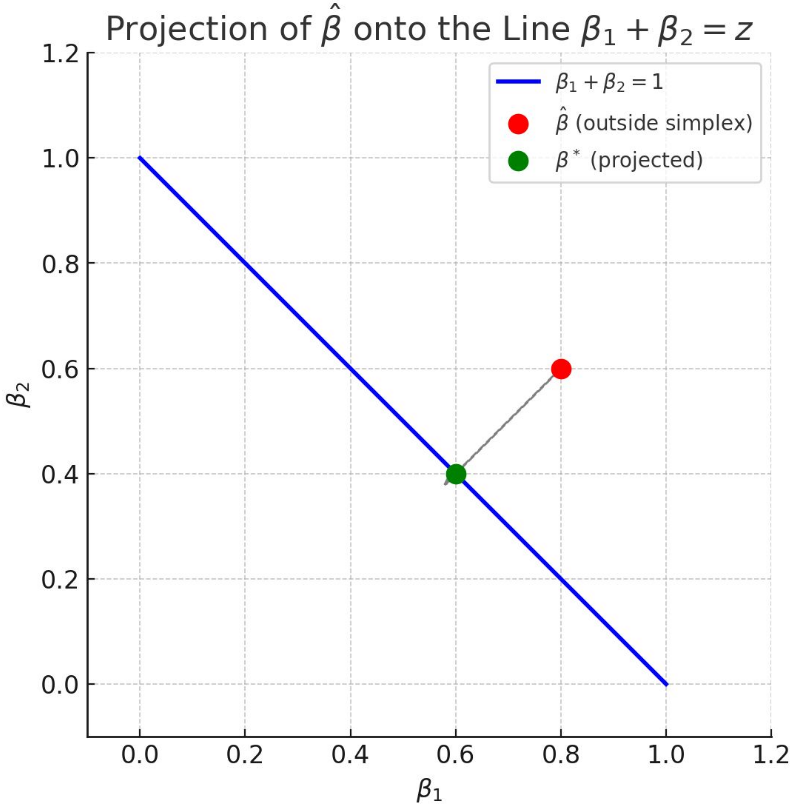
Taylor Expansion

$$f(\alpha + \delta_i) \approx f(\alpha) + g_i^\top \delta_i + \frac{1}{2}\delta_i^\top H_i \delta_i$$

$$f(\alpha + \delta_i) \approx f(\alpha) + g_i^\top \delta_i + \frac{1}{2}\|\delta_i\|^2$$

$$\boxed{\min_{\delta_i \in \mathbb{R}^k} \frac{1}{2}\|\delta_i\|^2 + g_i^\top \delta_i \quad \text{subject to: } \delta_i^\top \mathbf{1} = 0,\ \delta_i \le C_i - \alpha_i}$$

Simplex Projection



Change in variables

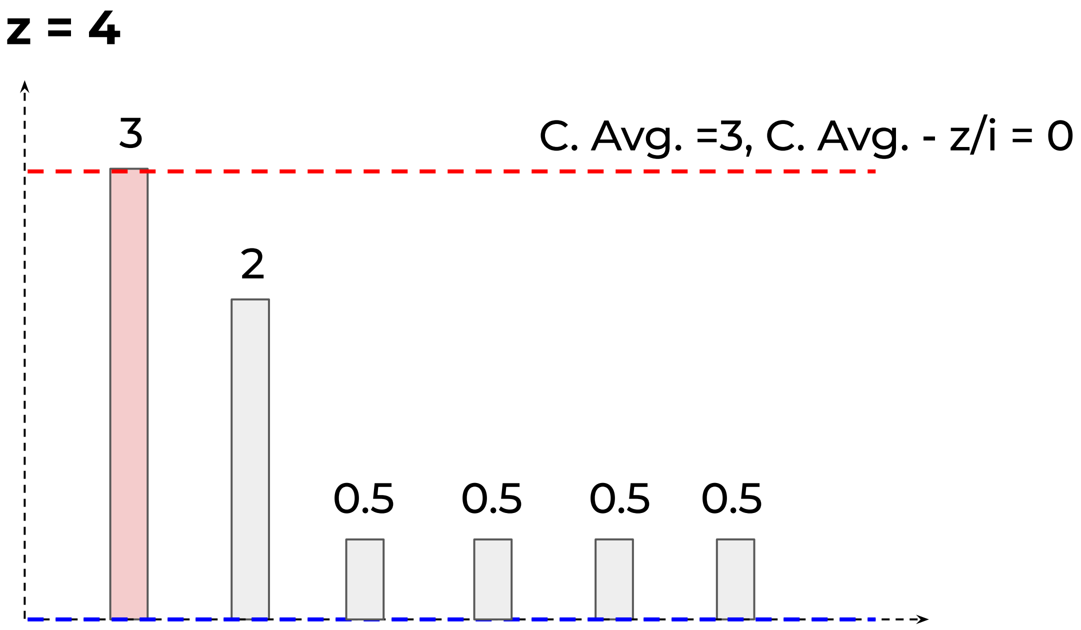
$$\delta_i = C_i - \alpha_i - \frac{\beta}{\|x_i\|} \quad \Rightarrow \quad \beta = \|x_i\|(C_i - \alpha_i - \delta_i)$$

Simplex

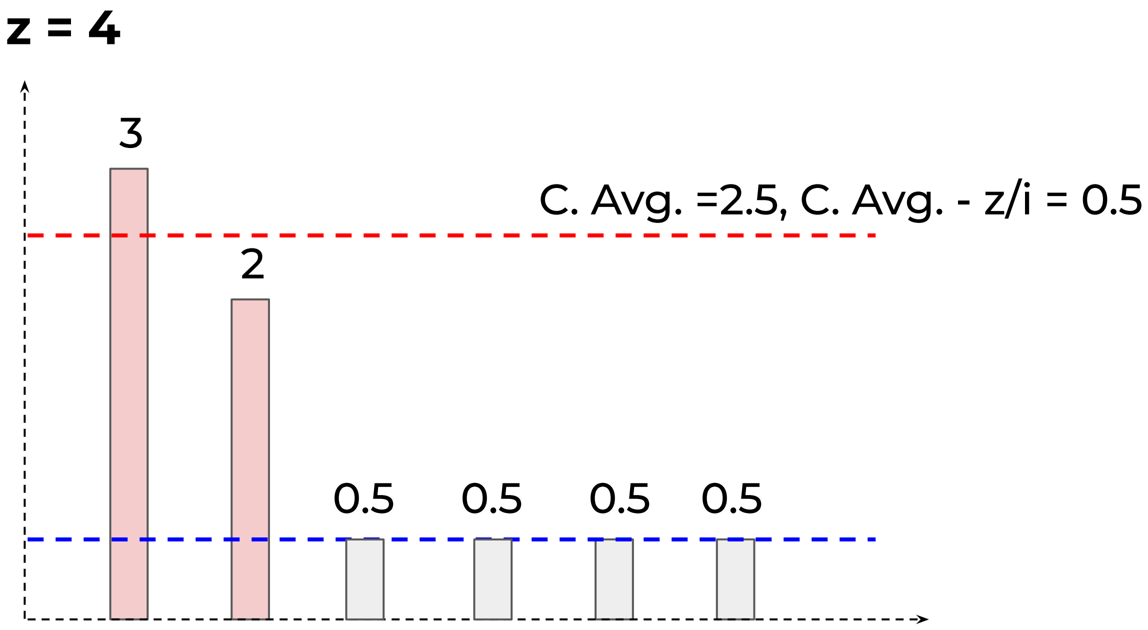
$$\min_{\beta \in \mathbb{R}^k} \frac{1}{2}\|\beta - \hat{\beta}\|^2 \quad \text{subject to: } \beta \geq 0, \quad \sum_{m=1}^{k} \beta_m = z$$

$$\hat{\beta} = \|x_i\|(C_i - \alpha_i) + \frac{g_i}{\|x_i\|}, \quad z = C \cdot \|x_i\|$$

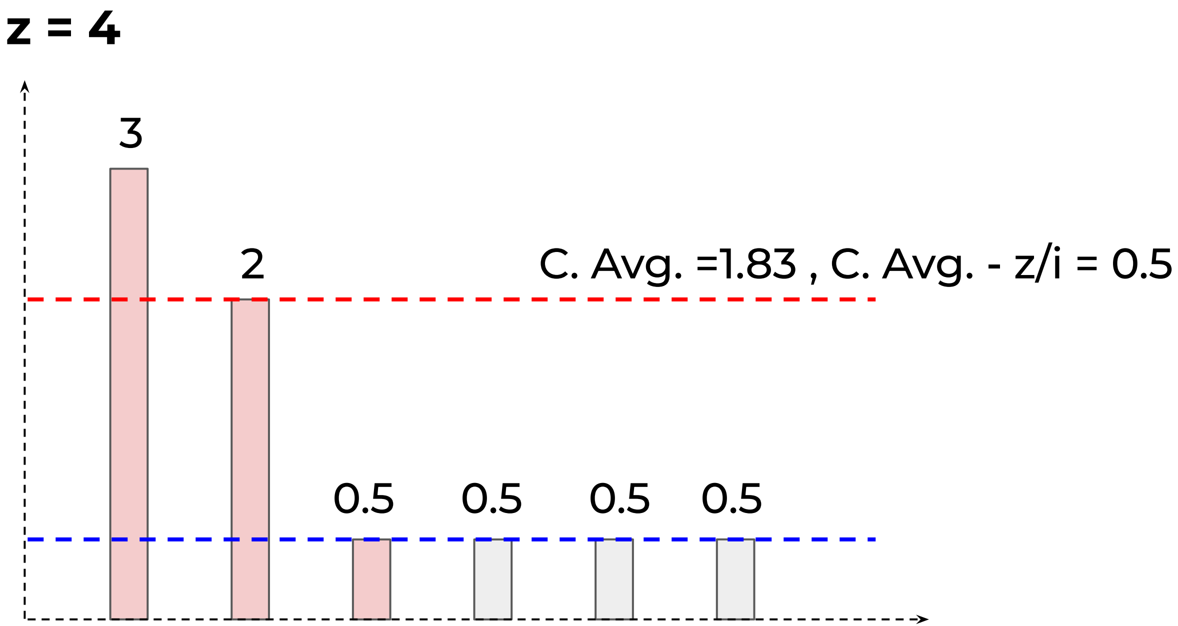
Sorting



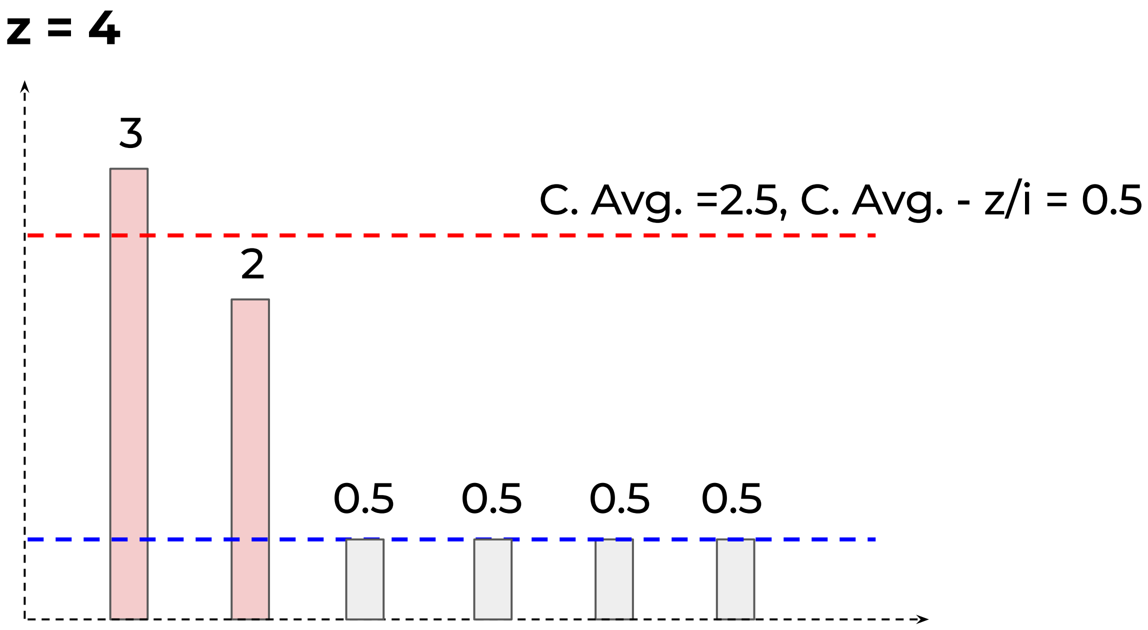
Sorting



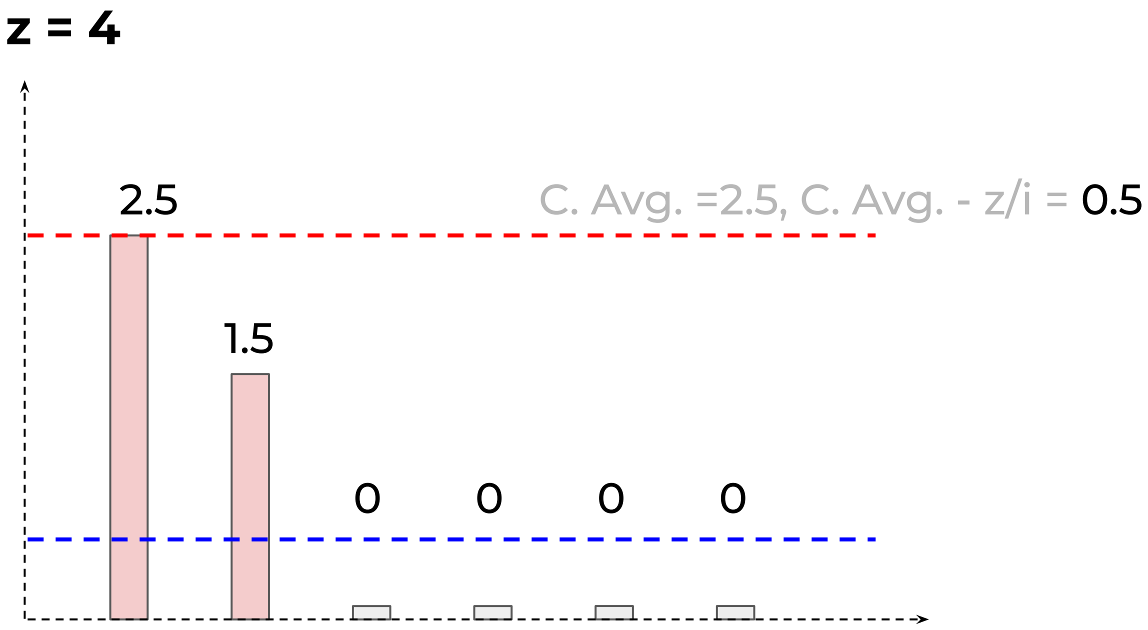
Sorting



Sorting



Sorting



Weston Watkins SVM

Objective Function

$$\min_{\{w_m\},\{\xi_i\}} \frac{1}{2}\sum_{m=1}^{k}\|w_m\|^2 + C\sum_{i=1}^{n}\sum_{m=1}^{k}\xi_{im}$$

subject to,

Classification Constraint:

$$w_{y_i}^\top x_i - w_m^\top x_i \geq 1 - \xi_{im},\ \forall m \neq y_i \quad \text{—(1)}$$

Slack Constraint:

$$\xi_{im} \geq 0,\ \forall i\ ,\ \forall m \quad \text{—(2)}$$

Multiclass Hinge Loss

$$w_{y_i}^\top x_i - w_m^\top x_i \geq 1 - \xi_{im}, \quad \forall m \neq y_i$$

Penalizing all largest violator classes

$$\xi_{im} \geq 1 - (w_{y_i}^\top x_i - w_m^\top x_i)$$

$$\xi_{im} \geq \max(0,\ 1 - (w_{y_i}^\top x_i - w_m^\top x_i))$$

$$\xi_{im} = \max(0,\ 1 - (w_{y_i}^\top x_i - w_m^\top x_i))$$

$$\xi_{im} = \max(0,\ 1 - (w_{y_i}^\top x_i - w_m^\top x_i))$$

$$\ell^{\mathrm{WW}}(x_i, y_i) = \sum_{m \neq y_i} \xi_{im}$$

$$\ell^{\mathrm{WW}}(x_i, y_i) = \sum_{m \neq y_i} \max(0,\ 1 - (w_{y_i}^\top x_i - w_m^\top x_i))$$

Research Goal

Implement **Weston Watkins SVM on GPU**

Weston Watkins Loss is one of the best performing loss function for SVMs

Reference: Doğan, Ürün., Glasmachers, Tobias., & Igel, Christian. (2016). A Unified View on Multi-class Support Vector Classification. Journal of Machine Learning Research, 17(1), 1–32.


PM
29
Wednesday - December 04
Do not open
Start time 5pm
End time 7pm
If your answer is on scratch sheet, staple it to the exam and mark it clearly.
Do not turn in cheatsheet
GAP
1969