Optimization in Machine Learning

Lecture 20: Project Gradient Descent, Karush Kuhn Tucker Conditions and Duality, Further results in Discrete Optimization

Ganesh Ramakrishnan

Department of Computer Science
Dept of CSE, IIT Bombay
https://www.cse.iitb.ac.in/~ganesh

April, 2025

Outline of next few topics

- Gradient Descent Analysis for Lipschitz Smoothness/Continuity/(Strong) Convexity [Done]
  - Our running Colab Notebook:
- Nestorov's and Polyak's accelerated gradient descent [Done]
- Sub-gradient Descent and Analysis [Done]
- Stochastic Gradient Descent [Done]
- Generalized Gradient Descent [Next]
- Proximal Gradient Descent [Dpme]
- Projected Gradient Descent for Constrained Optimization [Partly done]
- Discrete/Combinatorial Optimization for Subset Selection
- Constrained Optimization, Duality & KKT Conditions
- And so on...more such algorithms

Recap: Generalized Gradient Descent and its Special Cases

$$prox_{\gamma}(\mathbf{z}) = \text{argmin}_{\mathbf{x}} \frac{1}{2\gamma}||\mathbf{x} - \mathbf{z}||^2 + c(\mathbf{x})$$

It's special cases are:

1. Gradient Descent: $c(\mathbf{x}) = 0$
2. Projected Gradient Descent: $c(\mathbf{x}) = I_{\mathcal{C}}(\mathbf{x})$ (Example: $= \sum_i I_{g_i}(\mathbf{x})$)
3. Alternating Projection/Proximal Minimization: $f(\mathbf{x}) = 0$
4. Alternating Direction Method of Multipliers
5. Special Cases for Specific Objectives
   - LASSO: (Fast) Iterative Shrinkage Thresholding Algorithm (ISTA/FISTA)

How would you solve for the prox operator in general for projected gradient descent?

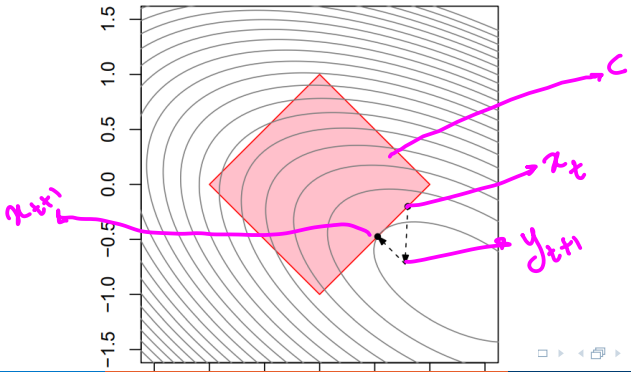
(Hint: See following two slides)

Recap: Summary of Proximal Gradient Descent

- Proximal Gradient Descent becomes gradient descent if $c = 0$
- If f is 0 (i.e. no smooth function), one can minimize a non-differentiable function c as long as the prox operator is easy to compute: **Proximal Point Algorithm**.
- Key to Proximal GD: Being able to compute Proximal operator.
- What if Prox can be efficiently computed approximately?
- There are papers (Schmidt et al 2011, Inexact Proximal Gradient Methods) where Prox is computed approximately but one can still derive the convergence rates if the errors due to approximation can be controlled!
- Accelerated Proximal GD: Similar to GD, one can accelerate GD to get optimal convergence rates! (Beck and Teboulle 2008)

Recap: An important special case: Projected Gradient Descent

- Consider the Problem of Constrained Convex Minimization: $\min_{x \in \mathcal{C}} f(x)$
- A simple modification of the gradient descent procedure is:
  1. At every iteration t: (Gradient Step): Compute $y_{t+1} = x_t - \alpha \nabla f(x_t)$
  2. (Projection step) $x_{t+1} = P_{\mathcal{C}}(y_{t+1})$
- Key here is the Projection step. Define $P_{\mathcal{C}}(x) = \text{argmin}_{y \in \mathcal{C}} \frac{1}{2}||x - y||^2$



Recap: Projected Gradient Descent and Proximal Gradient Descent

- There is a close connection between Proximal and Projected Gradient Descent.
- Define $c(x) = I(x \in \mathcal{C})$ where $I(.)$ is the Indicator function.
- Its easy to see that the $\text{prox}_c(x) = P_{\mathcal{C}}(x)$, i.e. the Prox operator is exactly the same as a projection operator.
- As a result, projected gradient descent becomes a special case of proximal gradient descent.
- Theoretical results of Proj. GD: All results for standard Gradient descent carry over to the projected case as long as the projection operator is easy to compute!

KKT conditions, Lagrange Dual etc

Algorithm: Projected Gradient Descent (We use $\mathbf{x}_u^t$ instead of $\mathbf{z}^t$)

Find a starting point $\mathbf{x}_p^0 \in \mathcal{C}$.
Set $t = 1$
repeat
1. Choose a step size $\gamma_t \propto 1/\sqrt{t}$.
2. Set $\mathbf{x}_u^t = \mathbf{x}_p^{t-1} - \gamma_t \nabla f(\mathbf{x}_p^{t-1})$.
3. Set $\mathbf{x}_p^t = \text{argmin}_{\mathbf{z} \in \mathcal{C}} ||\mathbf{x}_u^t - \mathbf{z}||_2^2$.
4. Set $t = t + 1$.

until stopping criterion (such as $||\mathbf{x}_p^t - \mathbf{x}_p^{t-1}|| \leq \epsilon$ or $f(\mathbf{x}_p^t) > f(\mathbf{x}_p^{t-1})$) is satisfied[a]

[a] Better criteria can be found using Lagrange duality theory, such as in the form of duality gap, etc.

Figure 1: The projected gradient descent algorithm.

Recap: Computing the Projection Operator

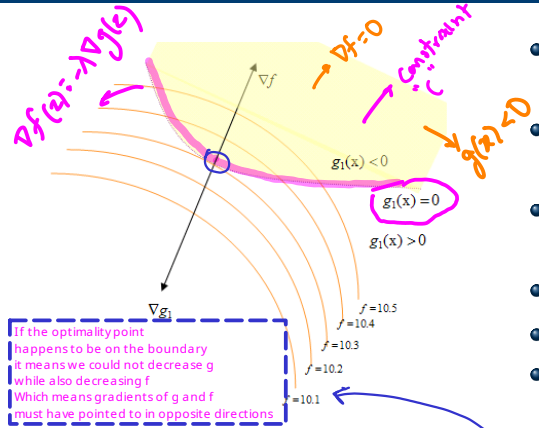


Figure 2: In this Figure $c = 0$ and f is the function being optimized. See optional course material referred to at the end for understanding this figure completely

From projection perspective this is simply one iteration of projection

- Lets assume for simplicity that $\mathcal{C} = \{x | g(x) \leq c\}$ and $f = \frac{1}{2}||z - x||^2$ is the function being minimized
- Computing the projection step involves solving: $\min_z \left\{\frac{1}{2}||z - x||^2, \text{ such that } g(z) \leq c\right\}$.
- Use the idea of Lagrange multipliers and the KKT conditions (see subsequent slides)!
- Define $L(z, \lambda) = \frac{1}{2}||z - x||^2 + \lambda(g(z) - c)$.
- Optimality conditions are: $\nabla_z L = 0$ and $\nabla_\lambda L = 0$!
- There are two options.
  1. Either $x \in \mathcal{C}$, and the constraints are not active ($g(z) < c$), in which case we need $\nabla_z L = 0$ and $\lambda = 0$ (that is $z = x$). if the optimal point is NOt in the boundary we have the regular condition of unconstrained optimization
  2. Or x is on the boundary of $\mathcal{C}$, in which case $g(z) = c$ and $z - x + \lambda \nabla g(z) = 0$. if the optimal point happens to be on the bondary
- If both can be solved in closed form, we are done!

Computing the Projection Operator

- Optimality conditions imply: $g(z) = c$ and $z - x + \lambda \nabla g(z) = 0$. If both these can be solved in closed form, we are done!
- How to compute the Projection operators for constraints $\mathcal{C}_g = \{x \mid g(x) \leq c\}$ when
  - $g(x) = a^T x$
  - $g(x) = ||x||_2^2$
  - $g(x) = ||x||_1$
  - $g(x) = ||x - x_0||^2$
  - $g(x) = x^T A x + bx + c$
  - $g(x) = ||x||_\infty$

$a^T x = c$ & $z - x + \lambda a = 0$ if x is optimal pt on bndy

else $x = z$

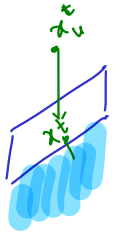
Try out these very simple cases

Easy to Project Sets $\mathcal{C}$ (with closed form solutions)

- Solution set of a linear system $\mathcal{C} = \{\mathbf{x} \in \Re^n : A^T\mathbf{x} = \mathbf{b}\}$
- Affine images $\mathcal{C} = \{A\mathbf{x} + \mathbf{b} : \mathbf{x} \in \Re^n\}$
- Nonnegative orthant $\mathcal{C} = \{\mathbf{x} \in \Re^n : \mathbf{x} \succeq 0\}$. It may be hard to project on arbitrary polyhedron.
- Norm balls $\mathcal{C} = \{\mathbf{x} \in \Re^n : \|\mathbf{x}\|_p \leq 1\}$, for $p = 1, 2, \infty$

Homework: For each of the example constrained g's stated here, apply the necessary condition for constrained optimality stated (and motivated) on the previous slide and derive the projection operation.

Easy to Project Sets $\mathcal{C}$ (with closed form solutions)



- Solution set of a linear system $\mathcal{C} = \{\mathbf{x} \in \Re^n : A^T\mathbf{x} = \mathbf{b}\}$
- Affine images $\mathcal{C} = \{A\mathbf{x} + \mathbf{b} : \mathbf{x} \in \Re^n\}$
- Nonnegative orthant $\mathcal{C} = \{\mathbf{x} \in \Re^n : \mathbf{x} \succeq 0\}$. It may be hard to project on arbitrary polyhedron.
- Norm balls $\mathcal{C} = \{\mathbf{x} \in \Re^n : \|\mathbf{x}\|_p \leq 1\}$, for $p = 1, 2, \infty$

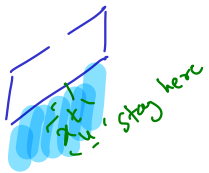


Table of Orthogonal Projections: See https://archive.siam.org/books/mo25/mo25_ch6.pdf

$$P_C(\mathbf{z}) = prox_{I_C}(\mathbf{z}) = \text{argmin}_{\mathbf{x}} \frac{1}{2\gamma}||\mathbf{x}-\mathbf{z}||^2 + I_C(\mathbf{x}) = \text{argmin}_{\mathbf{x}\in C} \frac{1}{2\gamma}||\mathbf{x}-\mathbf{z}||^2$$

Set $C =$	For $\gamma = 1$, $P_C(\mathbf{z}) =$	Assumptions
$\Re^n_+$	$[\mathbf{z}]_+$	
$\text{Box}[\mathbf{l},\mathbf{u}]$	$P_C(\mathbf{z})_i = min\{max\{z_i, l_i\}, u_i\}$	$l_i \leq u_i$
$\text{Ball}[\mathbf{c}, r]$	$\mathbf{c} + \frac{r}{max\{\|\mathbf{z}-\mathbf{c}\|_2, r\}}(\mathbf{z}-\mathbf{c})$	$\|.\|_2$ ball, centre $\mathbf{c} \in \Re^n$ & radius $r > 0$
$\{\mathbf{x} \| A\mathbf{x} = \mathbf{b}\}$	$\mathbf{z} - A^T(AA^T)^{-1}(A\mathbf{z}-\mathbf{b})$	$A \in \Re^{m\times n}$, $\mathbf{b} \in \Re^m$, A is full row rank
$\{\mathbf{x} \| \mathbf{a}^T\mathbf{x} \leq b\}$	$\mathbf{z} - \frac{[\mathbf{a}^T\mathbf{z}-b]_+}{\|\mathbf{a}\|^2}\mathbf{a}$	$0 \neq \mathbf{a} \in \Re^n$ $b \in \Re$
Δ_n (n-simplex)	$[\mathbf{z} - \mu^*\mathbf{e}]_+$ where $\mu^* \in \Re$ satisfies $\mathbf{e}^T[\mathbf{z}-\mu^*\mathbf{e}]_+ = 1$	
$H_{\mathbf{a},b} \cap \text{Box}[\mathbf{l},\mathbf{u}]$	$P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z}-\mu^*\mathbf{a})$ where $\mu^* \in \Re$ satisfies $\mathbf{a}^T P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z}-\mu^*\mathbf{a}) = b$	$0 \neq \mathbf{a} \in \Re^n$ $b \in \Re$
$H^-_{\mathbf{a},b} \cap \text{Box}[\mathbf{l},\mathbf{u}]$	$P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z})$ $\quad \mathbf{a}^T P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z}) \leq b$ $P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z}-\lambda^*\mathbf{a})$ $\quad \mathbf{a}^T P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z}) > b$ where $\lambda^* \in \Re$ satisfies $\mathbf{a}^T P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z}-\lambda^*\mathbf{a}) = b$ & $\lambda^* > 0$	$0 \neq \mathbf{a} \in \Re^n$ $b \in \Re$
$B_{\|.\|_1}[0,\alpha]$	$\mathbf{z}$ $\quad \|z\|_1 \leq \alpha$ $[\mathbf{z}-\lambda^*\mathbf{e}]_+ \odot sign(\mathbf{z})$ $\quad \|z\|_1 > \alpha$ where $\lambda^* > 0$, & $[\mathbf{z}-\lambda^*\mathbf{e}]_+ \odot sign(\mathbf{z}) = \alpha$	$\alpha > 0$

Table of Orthogonal Projections: See https://archive.siam.org/books/mo25/mo25_ch6.pdf

$$P_C(\mathbf{z}) = prox_{I_C}(\mathbf{z}) = \text{argmin}_{\mathbf{x}} \frac{1}{2\gamma}||\mathbf{x}-\mathbf{z}||^2 + I_C(\mathbf{x}) = \text{argmin}_{\mathbf{x}\in C} \frac{1}{2\gamma}||\mathbf{x}-\mathbf{z}||^2$$

Set $C =$	For $\gamma = 1$, $P_C(\mathbf{z}) =$	Assumptions
$\Re^n_+$	$[\mathbf{z}]_+$	
$\text{Box}[\mathbf{l},\mathbf{u}]$	$P_C(\mathbf{z})_i = min\{max\{z_i, l_i\}, u_i\}$	$l_i \leq u_i$
$\text{Ball}[\mathbf{c}, r]$	$\mathbf{c} + \frac{r}{max\{\|\mathbf{z}-\mathbf{c}\|_2, r\}}(\mathbf{z}-\mathbf{c})$	$\|.\|_2$ ball, centre $\mathbf{c} \in \Re^n$ & radius $r > 0$
$\{\mathbf{x} \vert A\mathbf{x} = \mathbf{b}\}$	$\mathbf{z} - A^T(AA^T)^{-1}(A\mathbf{z}-\mathbf{b})$	$A \in \Re^{m\times n}$, $\mathbf{b} \in \Re^m$, A is full row rank
$\{\mathbf{x} \vert \mathbf{a}^T\mathbf{x} \leq b\}$	$\mathbf{z} - \frac{[\mathbf{a}^T\mathbf{z}-b]_+}{\|\mathbf{a}\|^2}\mathbf{a}$	$0 \neq \mathbf{a} \in \Re^n$ $b \in \Re$
Δ_n (n-simplex)	$[\mathbf{z} - \mu^*\mathbf{e}]_+$ where $\mu^* \in \Re$ satisfies $\mathbf{e}^T[\mathbf{z}-\mu^*\mathbf{e}]_+ = 1$	
$H_{\mathbf{a},b} \cap \text{Box}[\mathbf{l},\mathbf{u}]$	$P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z}-\mu^*\mathbf{a})$ where $\mu^* \in \Re$ satisfies $\mathbf{a}^T P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z}-\mu^*\mathbf{a}) = b$	$0 \neq \mathbf{a} \in \Re^n$ $b \in \Re$
$H^-{}_{\mathbf{a},b} \cap \text{Box}[\mathbf{l},\mathbf{u}]$	$P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z})$ if $\mathbf{a}^T P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z}) \leq b$ $P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z}-\lambda^*\mathbf{a})$ if $\mathbf{a}^T P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z}) > b$ where $\lambda^* \in \Re$ satisfies $\mathbf{a}^T P_{\text{Box}[\mathbf{l},\mathbf{u}]}(\mathbf{z}-\lambda^*\mathbf{a}) = b$ & $\lambda^* > 0$	$0 \neq \mathbf{a} \in \Re^n$ $b \in \Re$
$B_{\|.\|_1}[0,\alpha]$	$\mathbf{z}$ if $\|z\|_1 \leq \alpha$ $[\mathbf{z}-\lambda^*\mathbf{e}]_+ \odot sign(\mathbf{z})$ if $\|z\|_1 > \alpha$ where $\lambda^* > 0$, & $[\mathbf{z}-\lambda^*\mathbf{e}]_+ \odot sign(\mathbf{z}) = \alpha$	$\alpha > 0$

Calculus of projection with multiple constraints

Radial direction for Projection

Geometric intuitions are clear. Question is how these analytical expressions are derived?
Lagrange function $(f + \lambda g)$ and KKT conditions

Convergence Results for Projected Gradient Descent (PGD)

- Lipschitz continuous function f (C) using PGD: R^2B^2/ϵ^2 iterations
- Lipschitz continuous functions + Strongly Convex f (CS) using PGD: $2B^2/\epsilon - 1$ iterations
- Smooth Function f using PGD: $\frac{R^2L}{\epsilon}$ iterations.
- Smooth Functions using Nestorov's PGD: $\sqrt{\frac{2R^2L}{\epsilon}}$ iterations
- Smooth + Strongly Convex f (SS) using PGD: With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{L}{\mu}\log(\frac{R^2L}{2\epsilon})$ iterations.
- Smooth + Strongly Convex f (SS) using Nestorov's PGD: With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\sqrt{\frac{L}{\mu}}\log(\frac{R^2L}{2\epsilon})$ iterations.
- All results for standard Gradient descent carry over to the projected case **as long as the projection operator is easy to compute!**
- Proofs in monograph by Sebastian Bubeck at `https://moodle.iitb.ac.in/mod/resource/view.php?id=36947`

How to solve the Projected Gradient Descent in closed form?

Lagrange Function, KKT Conditions & Duality For Solving Constrained Optimizaton

Karush Kuhn Tucker (KKT) Conditions for Optimality, Conditions for Necessity and Sufficiency *etc.* for..

Calculus of projection with multiple constraints

$$\begin{aligned} & \min_{\mathbf{x} \in \mathcal{D}} && f(\mathbf{x}) \\ & \text{subject to} && g_i(\mathbf{x}) \leq 0 \quad i = 1, \ldots, m \\ & && h_j(\mathbf{x}) = 0 \quad j = 1, \ldots, p \\ & \text{variable } \mathbf{x} = (x_1, \ldots, x_n) \end{aligned} \tag{1}$$

Karush Kuhn Tucker (KKT) Conditions for Optimality, Conditions for Necessity and Sufficiency *etc.* for..

option 1

For one particular iteration of projected gradient descent, f(x) = $\frac{1}{2}\|z - x\|^2$

$$\min_{\mathbf{x} \in \mathcal{D}} \quad f(\mathbf{x})$$

subject to

$$g_i(\mathbf{x}) \leq 0 \quad i = 1, \ldots, m$$
$$h_j(\mathbf{x}) = 0 \quad j = 1, \ldots, p \qquad (1)$$

variable $\mathbf{x} = (x_1, \ldots, x_n)$

n dimensional space with
m inequality constraints and
p equality constraints

option 2

Instead of solving the constrained optimization problem iteratively by gradient descent on f and then projection with constraints, if the f is itself simple, one could solve this constrained optimization problem directly

Briefly: The Dual Theory for Constrained Optimization

Consider the general constrained minimization problem

$$\begin{array}{ll} \min\limits_{\mathbf{x}\in\mathcal{D}} & f(\mathbf{x}) \\ \text{subject to} & g_i(\mathbf{x}) \leq 0,\ i = 1, 2, \ldots, m \\ \text{subject to} & h_j(\mathbf{x}) = 0,\ j = 1, 2, \ldots, p \end{array} \tag{2}$$

- Consider forming the lagrange function by associating prices (called lagrange multipliers) λ_i and μ_j , with constraints involving g_i and h_j respectively.

$$L(\mathbf{x}, \lambda, \mu) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{j=1}^{p} \mu_j h_j(\mathbf{x}) = f(\mathbf{x}) + \lambda^T \mathbf{g}(\mathbf{x}) + \mu^T \mathbf{h}(\mathbf{x})$$

- At each feasible $\mathbf{x}$, for fixed $\lambda_i \geq 0 \, \forall i \in \{1..m\}$,

$$f(\mathbf{x}) \geq \quad L(\mathbf{x}, \lambda, \mu) \quad \text{if } g_i(\mathbf{x}) \leq 0 \ \&\ h_j(\mathbf{x}) = 0 \tag{3}$$

Briefly: The Dual Theory for Constrained Optimization

Consider the general constrained minimization problem

$$\begin{aligned} \min_{\mathbf{x}\in\mathcal{D}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \leq 0,\ i = 1, 2, \ldots, m \\ \text{subject to} \quad & h_j(\mathbf{x}) = 0,\ j = 1, 2, \ldots, p \end{aligned} \tag{2}$$

Feasible region

+ve λ's * -ve g's

h's are 0's

- Consider forming the lagrange function by associating prices (called lagrange multipliers) λ_i and μ_j , with constraints involving g_i and h_j respectively.

$$L(\mathbf{x}, \lambda, \mu) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{j=1}^{p} \mu_j h_j(\mathbf{x}) = f(\mathbf{x}) + \lambda^T \mathbf{g}(\mathbf{x}) + \mu^T \mathbf{h}(\mathbf{x})$$

- At each feasible $\mathbf{x}$, for fixed $\lambda_i \geq 0 \, \forall i \in \{1..m\}$,

Rationale behind lagrange function:
At the boundary, the gradient of f must lie in the space spanned by the gradients of g's and h's with some sign constraints

$$f(\mathbf{x}) \geq \quad L(\mathbf{x}, \lambda, \mu) \quad \text{if } g_i(\mathbf{x}) \leq 0 \ \&\ h_j(\mathbf{x}) = 0 \tag{3}$$

Briefly: The Dual Theory for Constrained Optimization

- For $\lambda_i \geq 0 \, \forall i \in \{1..m\}$ and any μ_j, minimizing the right hand side of (3) over all feasible $\mathbf{x}$

$$f(\mathbf{x}) \geq \min_{\mathbf{x} \; s.t \; g_i(\mathbf{x}) \leq 0, h_j(\mathbf{x}) = 0} L(\mathbf{x}, \lambda, \mu) \quad \triangleq L^*(\lambda, \mu) \tag{4}$$

- $L^*(\lambda, \mu)$ is a pointwise (w.r.t $\mathbf{x} \in g_i(\mathbf{x}) \leq 0, h_j(\mathbf{x}) = 0$) minimum of linear functions ($L(\mathbf{x}, \lambda, \mu)$) and is therefore always a concave function.
- Since $f(\mathbf{x}) \geq L^*(\lambda, \mu)$ for all primal feasible $\mathbf{x}$ and dual feasible *i.e.*, $\lambda_i \geq 0$ and μ_j, , we can maximize the lower bound $L^*(\lambda, \mu)$ to give the following **Dual Problem**

$$\begin{aligned} \max_{\lambda \in \Re^m, \mu \in \Re^p} \quad & L^*(\lambda, \mu) \\ \text{subject to} \quad & \lambda \geq \mathbf{0} \end{aligned}$$

Briefly: The Dual Theory for Constrained Optimization

- For $\lambda_i \geq 0 \, \forall i \in \{1..m\}$ and any μ_j, minimizing the right hand side of (3) over all feasible $\mathbf{x}$

$$f(\mathbf{x}) \geq \min_{\mathbf{x} \; s.t \; g_i(\mathbf{x}) \leq 0, h_j(\mathbf{x}) = 0} L(\mathbf{x}, \lambda, \mu) \triangleq L^*(\lambda, \mu) \tag{4}$$

Inequality will hold even if I minimize the RHS with respect to all feasible x

- $L^*(\lambda, \mu)$ is a pointwise (w.r.t $\mathbf{x} \in g_i(\mathbf{x}) \leq 0, h_j(\mathbf{x}) = 0$) minimum of linear functions ($L(\mathbf{x}, \lambda, \mu)$) and is therefore always a concave function.
- Since $f(\mathbf{x}) \geq L^*(\lambda, \mu)$ for all primal feasible $\mathbf{x}$ and dual feasible *i.e.*, $\lambda_i \geq 0$ and μ_j, , we can maximize the lower bound $L^*(\lambda, \mu)$ to give the following **Dual Problem**

$$\begin{aligned} \max_{\lambda \in \Re^m, \mu \in \Re^p} \quad & L^*(\lambda, \mu) \\ \text{subject to} \quad & \lambda \geq \mathbf{0} \end{aligned}$$

Briefly: The Dual Theory for Constrained Optimization (contd.)

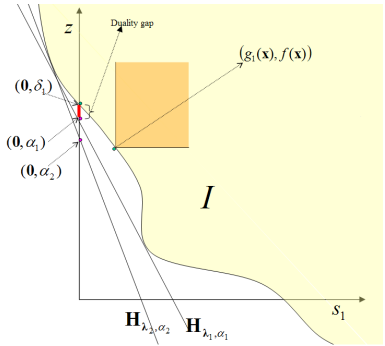


Figure 3: See optional course material referred to at the end for understanding this figure completely

The **Dual Problem** restated in two ways:

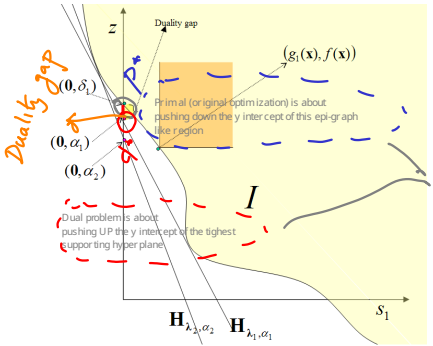
$$\begin{aligned} \max_{\lambda \in \Re^m, \mu \in \Re^p} \quad & L^*(\lambda, \mu) \\ \text{subject to} \quad & \lambda \geq \mathbf{0} \end{aligned} \tag{5}$$

$$\max_{\lambda \in \Re^m, \mu \in \Re^p} \quad \min_{\mathbf{x} \; s.t \; g_i(\mathbf{x}) \leq 0, h_j(\mathbf{x}) = 0} \quad L(\mathbf{x}, \lambda, \mu)$$
$$\text{subject to} \quad \lambda \geq \mathbf{0}$$

Theorem

(i) The dual function $L^*(\lambda, \mu)$ is always concave. (ii) If p^* is solution of (2) and d^* of (5) then $p^* \geq d^*$ and the gap $p^* - d^*$ is called the **duality gap**

Briefly: The Dual Theory for Constrained Optimization (contd.)



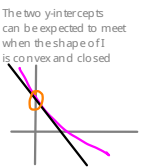


Figure 3: See optional course material referred to at the end for understanding this figure completely

The **Dual Problem** restated in two ways:

x axis is space of values of g

v axis is space of value of f

$$\max_{\lambda \in \Re^m, \mu \in \Re^p} L^*(\lambda, \mu) \quad \text{subject to} \quad \lambda \geq \mathbf{0} \tag{5}$$

duhg L

$$\max_{\lambda \in \Re^m, \mu \in \Re^p} \min_{\mathbf{x} \; s.t \; g_i(\mathbf{x}) \leq 0, h_j(\mathbf{x}) = 0} L(\mathbf{x}, \lambda, \mu) \quad \text{subject to} \quad \lambda \geq \mathbf{0}$$

Theorem

(i) The dual function $L^*(\lambda, \mu)$ is always concave. (ii) If p^* is solution of (2) and d^* of (5) then $p^* \geq d^*$ and the gap $p^* - d^*$ is called the **duality gap**

Zero Duality Gap ⇒ KKT Conditions

Theorem

Assuming requisite differentiability, we can state the following Karush-Kuhn-Tucker (KKT) necessary conditions conditions in (6) for zero duality gap:

$$
\begin{array}{lrcll}
(1) & \nabla f(\widehat{\mathbf{x}}) + \sum_{i=1}^{m} \widehat{\lambda}_i \nabla g_i(\widehat{\mathbf{x}}) + \sum_{j=1}^{p} \widehat{\mu}_j \nabla h_j(\widehat{\mathbf{x}}) & = & \mathbf{0} & \\
(2) & g_i(\widehat{\mathbf{x}}) & \leq & 0 & i = 1, 2, \ldots, m \\
(3) & \widehat{\lambda}_i & \geq & 0 & i = 1, 2, \ldots, m \\
(4) & \widehat{\lambda}_i g_i(\widehat{\mathbf{x}}) & = & 0 & i = 1, 2, \ldots, m \\
(5) & h_j(\widehat{\mathbf{x}}) & = & 0 & j = 1, 2, \ldots, p
\end{array}
\tag{6}
$$

Zero Duality Gap ⇒ KKT Conditions

CONVEXITY IS NOT PART OF NECESSARY CONDITIONS...

Theorem

Assuming requisite differentiability, we can state the following Karush-Kuhn-Tucker (KKT) necessary conditions conditions in (6) for zero duality gap:

If $\hat{x} = \arg\min_x f(x)$ s.t. $g_i(x) \le 0$, $h_j(x) = 0$ } Then there must exist $\hat{\lambda}$ & $\hat{\mu}$ & following conditions with $(\hat{x}, \hat{\lambda}, \hat{\mu})$

$$
\begin{aligned}
&(1) \quad \nabla f(\hat{\mathbf{x}}) + \textstyle\sum_{i=1}^{m} \hat{\lambda}_i \nabla g_i(\hat{\mathbf{x}}) + \sum_{j=1}^{p} \hat{\mu}_j \nabla h_j(\hat{\mathbf{x}}) = \mathbf{0} \\
&(2) \quad g_i(\hat{\mathbf{x}}) \le 0 \quad i = 1, 2, \ldots, m \\
&(3) \quad \hat{\lambda}_i \ge 0 \quad i = 1, 2, \ldots, m \\
&(4) \quad \hat{\lambda}_i g_i(\hat{\mathbf{x}}) = 0 \quad i = 1, 2, \ldots, m \\
&(5) \quad h_j(\hat{\mathbf{x}}) = 0 \quad j = 1, 2, \ldots, p
\end{aligned}
\tag{6}
$$

Primal feasibility conditions

COMPLEMENTARY SLACKNESS

Either you are on boundary $g_i(\hat{x}) = 0$ & therefore $\hat{\lambda}_i$ plays some role
OR you are inside $g_i(\hat{x}) < 0$ & therefore $\hat{\lambda}_i$ plays no role

Zero Duality Gap $\Leftarrow$ KKT Conditions+Convexity

Theorem

If the function f is convex, g_i are convex and h_j are affine, then KKT conditions in (6) are necessary and sufficient conditions for zero duality gap.

$$
\begin{array}{lrcll}
(1) & \nabla f(\widehat{\mathbf{x}}) + \sum_{i=1}^{m} \widehat{\lambda}_i \nabla g_i(\widehat{\mathbf{x}}) + \sum_{j=1}^{p} \widehat{\mu}_j \nabla h_j(\widehat{\mathbf{x}}) & = & \mathbf{0} & \\
(2) & g_i(\widehat{\mathbf{x}}) & \leq & 0 & i = 1, 2, \ldots, m \\
(3) & \widehat{\lambda}_i & \geq & 0 & i = 1, 2, \ldots, m \\
(4) & \widehat{\lambda}_i g_i(\widehat{\mathbf{x}}) & = & 0 & i = 1, 2, \ldots, m \\
(5) & h_j(\widehat{\mathbf{x}}) & = & 0 & j = 1, 2, \ldots, p
\end{array}
\tag{6}
$$

Zero Duality Gap $\Leftarrow$ KKT Conditions+Convexity

sufficiency owing to convexity

Theorem

If the function f is convex, g_i are convex and h_j are affine, then KKT conditions in (6) are necessary and sufficient conditions for zero duality gap.

ALL THE ALGEBRAIC EXPRESSIONS OF PROJECTION OPERATIONS THAT WE SHOWED IN THE EARLIER TABLE FOR PROJECTION CAN BE DERIVED USING THESE KKT CONDITIONS! H/W TRY THESE ON SOME!

$$
\begin{array}{lrcll}
(1) & \nabla f(\widehat{\mathbf{x}}) + \sum_{i=1}^{m} \widehat{\lambda}_i \nabla g_i(\widehat{\mathbf{x}}) + \sum_{j=1}^{p} \widehat{\mu}_j \nabla h_j(\widehat{\mathbf{x}}) & = & \mathbf{0} & \\
(2) & g_i(\widehat{\mathbf{x}}) & \leq & 0 & i = 1, 2, \ldots, m \\
(3) & \widehat{\lambda}_i & \geq & 0 & i = 1, 2, \ldots, m \\
(4) & \widehat{\lambda}_i g_i(\widehat{\mathbf{x}}) & = & 0 & i = 1, 2, \ldots, m \\
(5) & h_j(\widehat{\mathbf{x}}) & = & 0 & j = 1, 2, \ldots, p
\end{array}
\tag{6}
$$

Convexity, KKT Conditions, Slater's Qualification, Zero Duality Gap

Problem type	Objective Function	Constraints	$L^*(\lambda)$	Dual constraints	Strong duality
Linear Program	$\mathbf{c}^T\mathbf{x}$	$A\mathbf{x} \leq \mathbf{b}$	$-\mathbf{b}^T\lambda$	$A^T\lambda + \mathbf{c} = \mathbf{0}$ $\lambda \geq \mathbf{0}$	Feasible primal and dual
Quadratic Program	$\frac{1}{2}\mathbf{x}^T Q\mathbf{x} + \mathbf{c}^T\mathbf{x}$ for $Q \in \mathcal{S}^n_{++}$	$A\mathbf{x} \leq \mathbf{b}$	$-\frac{1}{2}\left(\mathbf{c} - A^T\lambda\right)^T Q^{-1}\left(\mathbf{c} - A^T\lambda\right) + \mathbf{b}^T\lambda$	$\lambda \geq \mathbf{0}$	Always
Entropy maximization	$x_i \sum_{i=1}^n \ln x_i$	$A\mathbf{x} \leq \mathbf{b}$ $\mathbf{x}^T\mathbf{1} = 1$	$-\mathbf{b}^T\lambda - \mu - e^{-\mu-1}\sum_{i=1}^n e^{-\mathbf{a}_i^T\lambda}$ $\mathbf{a}_i$ is the i^{th} column of A	$\lambda \geq \mathbf{0}$	Primal constraints are satisfied.

Table 1: Some optimization problems, their duals and conditions for strong duality.

HOMEWORK: CAN THE FORM OF THE DUAL BECOME SIGNIFICANTLY SIMPLER THAN THE PRIMAL
IS THAT WHAT HAPPENED IN SVMs?

Topics covered in Extra optional material from previous offering

- Geometric Interpretation of Necessary conditions of optimality[1]
- Geometric Interpretation of Zero duality gap and connection with convexity[2]
- Weak and Strong Duality Theorem and their proofs, Slater's Constraint Qualification[3]
- Conditional Gradient Descent (Frank Wolfe Method), Convex Conjugate, Dual Ascent, ADMM[4]

[1] https://www.youtube.com/watch?v=O3vfurE_rsO&list=PLyo3HAXSZD3xRZ3mLOW6ERtQ3huVVmyoO&index=27
[2] https://www.youtube.com/watch?v=yG6HxDqHj1Y&list=PLyo3HAXSZD3xRZ3mLOW6ERtQ3huVVmyoO&index=29
[3] https://www.youtube.com/watch?v=r4PcHOek-YI&list=PLyo3HAXSZD3xRZ3mLOW6ERtQ3huVVmyoO&index=31
[4] https://www.youtube.com/watch?v=Jlmsr3uqz3Q&list=PLyo3HAXSZD3xRZ3mLOW6ERtQ3huVVmyoO&index=34

Support Vector Regression and its Dual

Self Reading Material for practical illustration of Duality, KKT Conditions, Necessity, Sufficiency, *etc.*

KKT and Dual for SVR

- $\min_{\mathbf{w},b,\xi_i,\xi_i^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*)$
  s.t. $\forall i$,
  $y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i$,
  $b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i \leq \epsilon + \xi_i^*$,
  $\xi_i, \xi_i^* \geq 0$
- Let's consider the lagrange multipliers α_i, α_i^*, μ_i and μ_i^* corresponding to the above-mentioned constraints.
- The Lagrange Function is

KKT and Dual for SVR

- $\min_{\mathbf{w},b,\xi_i,\xi_i^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*)$
  s.t. $\forall i$,
  $y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i$,
  $b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i \leq \epsilon + \xi_i^*$,
  $\xi_i, \xi_i^* \geq 0$
- Let's consider the lagrange multipliers α_i, α_i^*, μ_i and μ_i^* corresponding to the above-mentioned constraints.
- The Lagrange Function is

$$L(\mathbf{w},\alpha,\alpha^*,\mu,\mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i+\xi_i^*) + \sum_{i=1}^m \alpha_i\left(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right) + \sum_{i=1}^m \alpha_i^*\left(b + \mathbf{w}^\top\phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^m \mu_i\xi_i - \sum_{i=1}^m \mu_i^*\xi_i^*$$

KKT conditions for SVR

$$L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*) + \sum_{i=1}^m \alpha_i \left(y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right) + \sum_{i=1}^m \alpha_i^* \left(b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^m \mu_i \xi_i - \sum_{i=1}^m \mu_i^* \xi_i^*$$

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$,

KKT conditions for SVR

$$L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*) + \sum_{i=1}^m \alpha_i \left(y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right) + \sum_{i=1}^m \alpha_i^* \left(b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^m \mu_i \xi_i - \sum_{i=1}^m \mu_i^* \xi_i^*$$

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$,

  $\mathbf{w} - \alpha_i \phi(\mathbf{x}_i) + \alpha_i^* \phi(\mathbf{x}_i) = 0$ *i.e.*, $\mathbf{w} = \sum_{i=1}^m (\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$
- Differentiating the Lagrangian w.r.t. ξ_i,

KKT conditions for SVR

$$L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*) + \sum_{i=1}^m \alpha_i \left(y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right) + \sum_{i=1}^m \alpha_i^* \left(b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^m \mu_i \xi_i - \sum_{i=1}^m \mu_i^* \xi_i^*$$

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$,
  $\mathbf{w} - \alpha_i \phi(\mathbf{x}_i) + \alpha_i^* \phi(\mathbf{x}_i) = 0$ *i.e.*, $\mathbf{w} = \sum_{i=1}^m (\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$
- Differentiating the Lagrangian w.r.t. ξ_i,
  $C - \alpha_i - \mu_i = 0$ *i.e.*, $\alpha_i + \mu_i = C$
- Differentiating the Lagrangian w.r.t ξ_i^*,

KKT conditions for SVR

$$L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*) + \sum_{i=1}^m \alpha_i \left(y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right) + \sum_{i=1}^m \alpha_i^* \left(b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^m \mu_i \xi_i - \sum_{i=1}^m \mu_i^* \xi_i^*$$

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$,
  $\mathbf{w} - \alpha_i \phi(\mathbf{x}_i) + \alpha_i^* \phi(\mathbf{x}_i) = 0$ *i.e.*, $\mathbf{w} = \sum_{i=1}^m (\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$
- Differentiating the Lagrangian w.r.t. ξ_i,
  $C - \alpha_i - \mu_i = 0$ *i.e.*, $\alpha_i + \mu_i = C$
- Differentiating the Lagrangian w.r.t ξ_i^*,
  $\alpha_i^* + \mu_i^* = C$
- Differentiating the Lagrangian w.r.t b,

KKT conditions for SVR

$$L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*) + \sum_{i=1}^m \alpha_i \left(y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right) + \sum_{i=1}^m \alpha_i^* \left(b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^m \mu_i \xi_i - \sum_{i=1}^m \mu_i^* \xi_i^*$$

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$,
  $\mathbf{w} - \alpha_i \phi(\mathbf{x}_i) + \alpha_i^* \phi(\mathbf{x}_i) = 0$ *i.e.*, $\mathbf{w} = \sum_{i=1}^m (\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$
- Differentiating the Lagrangian w.r.t. ξ_i,
  $C - \alpha_i - \mu_i = 0$ *i.e.*, $\alpha_i + \mu_i = C$
- Differentiating the Lagrangian w.r.t ξ_i^*,
  $\alpha_i^* + \mu_i^* = C$
- Differentiating the Lagrangian w.r.t b,
  $\sum_i (\alpha_i^* - \alpha_i) = 0$
- Complimentary slackness:

KKT conditions for SVR

$$L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*) + \sum_{i=1}^m \alpha_i \left(y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i\right) + \sum_{i=1}^m \alpha_i^* \left(b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*\right) - \sum_{i=1}^m \mu_i \xi_i - \sum_{i=1}^m \mu_i^* \xi_i^*$$

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$,
  $\mathbf{w} - \alpha_i \phi(\mathbf{x}_i) + \alpha_i^* \phi(\mathbf{x}_i) = 0$ *i.e.*, $\mathbf{w} = \sum_{i=1}^m (\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$
- Differentiating the Lagrangian w.r.t. ξ_i,
  $C - \alpha_i - \mu_i = 0$ *i.e.*, $\alpha_i + \mu_i = C$
- Differentiating the Lagrangian w.r.t ξ_i^*,
  $\alpha_i^* + \mu_i^* = C$
- Differentiating the Lagrangian w.r.t b,
  $\sum_i (\alpha_i^* - \alpha_i) = 0$
- Complimentary slackness:
  $\alpha_i(y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0$ AND $\mu_i \xi_i = 0$ AND
  $\alpha_i^*(b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*) = 0$ AND $\mu_i^* \xi_i^* = 0$

Conclusions from the KKT conditions:

$\alpha_i \in (0, C) \Rightarrow ?$

$\alpha_i^* \in (0, C) \Rightarrow ?$

KKT conditions clearly outline support vectors!

KKT conditions

- Differentiating the Lagrangian w.r.t. $\mathbf{w}$,
  $w - \alpha_i \phi(\mathbf{x}_i) + \alpha_i^* \phi(\mathbf{x}_i) = 0$
  i.e. $\mathbf{w} = \sum_{i=1}^m (\alpha_i - \alpha_i^*) \phi(\mathbf{x}_i)$
- Differentiating the Lagrangian w.r.t. ξ_i,
  $C - \alpha_i - \mu_i = 0$
  i.e. $\alpha_i + \mu_i = C$
- Differentiating the Lagrangian w.r.t ξ_i^*,
  $\alpha_i^* + \mu_i^* = C$
- Differentiating the Lagrangian w.r.t b,
  $\sum_i^m (\alpha_i^* - \alpha_i) = 0$
- Complimentary slackness:
  $\alpha_i (y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0$
  $\mu_i \xi_i = 0$
  $\alpha_i^* (b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*) = 0$
  $\mu_i^* \xi_i^* = 0$

Conclusions from the KKT conditions:

$$\alpha_i(y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0$$

and

$$\alpha_i^*(b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*) = 0$$

$\Rightarrow$?

Conclusions from the KKT conditions:

$$\alpha_i \in (0, C) \Rightarrow ?$$

$$(C - \alpha_i)\xi_i = 0 \Rightarrow ?$$

$$\alpha_i^* \in (0, C) \Rightarrow ?$$

$$(C - \alpha_i^*)\xi_i^* = 0 \Rightarrow ?$$

For Support Vector Regression, since the original objective and the constraints are convex, any $(\mathbf{w}, b, \alpha, \alpha^*, \mu, \mu^*, \xi, \xi^*)$ that satisfy the necessary KKT conditions gives optimality (conditions are also sufficient)

Some observations

- $\alpha_i, \alpha_i^* \geq 0$, $\mu_i, \mu_i^* \geq 0$, $\alpha_i + \mu_i = C$ and $\alpha_i^* + \mu_i^* = C$
  Thus, $\alpha_i, \mu_i, \alpha_i^*, \mu_i^* \in [0, C]$, $\forall i$
- If $0 < \alpha_i < C$, then $0 < \mu_i < C$
  (as $\alpha_i + \mu_i = C$)
- $\mu_i \xi_i = 0$ and $\alpha_i(y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0$ are complementary slackness conditions
  So $0 < \alpha_i < C \Rightarrow \xi_i = 0$ and $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b = \epsilon + \xi_i = \epsilon$
  - All such points lie on the boundary of the ϵ band
  - Using any point $\mathbf{x}_j$ (that is with $\alpha_j \in (0, C)$) on margin, we can recover b as:
    $b = y_j - \mathbf{w}^\top \phi(\mathbf{x}_j) - \epsilon$

Support Vector Regression
Dual Objective

Weak Duality

- $L^*(\alpha, \alpha^*, \mu, \mu^*) = \min_{\mathbf{w}, b, \xi, \xi^*} L(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*)$
- By weak duality theorem, we have:
  $\min_{\mathbf{w}, b, \xi, \xi^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^m (\xi_i + \xi_i^*) \geq L^*(\alpha, \alpha^*, \mu, \mu^*)$
  s.t. $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b \leq \epsilon - \xi_i$, and
  $\mathbf{w}^\top \phi(\mathbf{x}_i) + b - y_i \leq \epsilon - \xi_i^*$, and
  $\xi_i, \xi^* \geq 0$, $\forall i = 1, \ldots, n$
- The above is true for any $\alpha_i, \alpha_i^* \geq 0$ and $\mu_i, \mu_i^* \geq 0$
- Thus,

Weak Duality

- $L^*(\alpha, \alpha^*, \mu, \mu^*) = \min_{\mathbf{w}, b, \xi, \xi^*} L(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*)$
- By weak duality theorem, we have:
  $\min_{\mathbf{w}, b, \xi, \xi^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^m (\xi_i + \xi_i^*) \geq L^*(\alpha, \alpha^*, \mu, \mu^*)$
  s.t. $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b \leq \epsilon - \xi_i$, and
  $\mathbf{w}^\top \phi(\mathbf{x}_i) + b - y_i \leq \epsilon - \xi_i^*$, and
  $\xi_i, \xi^* \geq 0$, $\forall i = 1, \ldots, n$
- The above is true for any $\alpha_i, \alpha_i^* \geq 0$ and $\mu_i, \mu_i^* \geq 0$
- Thus,

$$\min_{\mathbf{w}, b, \xi, \xi^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^m (\xi_i + \xi_i^*) \geq \max_{\alpha, \alpha^*, \mu, \mu^*} L^*(\alpha, \alpha^*, \mu, \mu^*)$$

s.t. $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b \leq \epsilon - \xi_i$, and
$\mathbf{w}^\top \phi(\mathbf{x}_i) + b - y_i \leq \epsilon - \xi_i^*$, and
$\xi_i, \xi^* \geq 0$, $\forall i = 1, \ldots, n$

Dual objective

- $L^*(\alpha, \alpha^*, \mu, \mu^*) = \min_{\mathbf{w}, b, \xi, \xi^*} L(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*)$
- Assume: In case of SVR, we have a strictly convex objective and linear constraints $\Rightarrow$ KKT conditions are necessary and sufficient and strong duality holds:

$$\min_{\mathbf{w}, b, \xi, \xi^*} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) = \max_{\alpha, \alpha^*, \mu, \mu^*} L^*(\alpha, \alpha^*, \mu, \mu^*)$$

s.t. $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b \leq \epsilon - \xi_i$, and
$w^\top \phi(\mathbf{x}_i) + b - y_i \leq \epsilon - \xi_i^*$, and
$\xi_i, \xi^* \geq 0, \forall i = 1, \ldots, n$

- This value is precisely obtained at the $(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*)$ that satisfies the necessary (and sufficient) KKT optimality conditions
- Given strong duality, we can equivalently solve

$$\max_{\alpha, \alpha^*, \mu, \mu^*} L^*(\alpha, \alpha^*, \mu, \mu^*)$$

- $L(\alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) +$
$\sum_{i=1}^{m}\left(\alpha_i(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i) + \alpha_i^*(\mathbf{w}^\top\phi(\mathbf{x}_i) + b - y_i - \epsilon - \xi_i^*)\right) - \sum_{i=1}^{m}(\mu_i\xi_i + \mu_i^*\xi_i^*)$
- We obtain $\mathbf{w}$, b, ξ_i, ξ_i^* in terms of α, α^*, μ and μ^* by using the KKT conditions derived earlier as $\mathbf{w} = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$ and $\sum_{i=1}^{m}(\alpha_i - \alpha_i^*) = 0$ and $\alpha_i + \mu_i = C$ and $\alpha_i^* + \mu_i^* = C$
- Thus, we get:

- $L(\alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) +$
$\sum_{i=1}^{m}\left(\alpha_i(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i) + \alpha_i^*(\mathbf{w}^\top\phi(\mathbf{x}_i) + b - y_i - \epsilon - \xi_i^*)\right) - \sum_{i=1}^{m}(\mu_i\xi_i + \mu_i^*\xi_i^*)$
- We obtain $\mathbf{w}$, b, ξ_i, ξ_i^* in terms of α, α^*, μ and μ^* by using the KKT conditions derived earlier as $\mathbf{w} = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$ and $\sum_{i=1}^{m}(\alpha_i - \alpha_i^*) = 0$ and $\alpha_i + \mu_i = C$ and $\alpha_i^* + \mu_i^* = C$
- Thus, we get:
$L(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*)$
$= \frac{1}{2}\sum_i\sum_j(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\phi^\top(\mathbf{x}_i)\phi(\mathbf{x}_j) + \sum_i(\xi_i(C - \alpha_i - \mu_i) + \xi_i^*(C - \alpha_i^* - \mu_i^*)) -$
$b\sum_i(\alpha_i - \alpha_i^*) - \epsilon\sum_i(\alpha_i + \alpha_i^*) + \sum_i y_i(\alpha_i - \alpha_i^*) - \sum_i\sum_j(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\phi^\top(\mathbf{x}_i)\phi(\mathbf{x}_j)$

- $L(\alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) +$
$\sum_{i=1}^{m}\left(\alpha_i(y_i - \mathbf{w}^\top\phi(\mathbf{x}_i) - b - \epsilon - \xi_i) + \alpha_i^*(\mathbf{w}^\top\phi(\mathbf{x}_i) + b - y_i - \epsilon - \xi_i^*)\right) - \sum_{i=1}^{m}(\mu_i\xi_i + \mu_i^*\xi_i^*)$
- We obtain $\mathbf{w}$, b, ξ_i, ξ_i^* in terms of α, α^*, μ and μ^* by using the KKT conditions derived earlier as $\mathbf{w} = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi(\mathbf{x}_i)$ and $\sum_{i=1}^{m}(\alpha_i - \alpha_i^*) = 0$ and $\alpha_i + \mu_i = C$ and $\alpha_i^* + \mu_i^* = C$
- Thus, we get:
$L(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*)$
$= \frac{1}{2}\sum_i\sum_j(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\phi^\top(\mathbf{x}_i)\phi(\mathbf{x}_j) + \sum_i(\xi_i(C - \alpha_i - \mu_i) + \xi_i^*(C - \alpha_i^* - \mu_i^*)) -$
$b\sum_i(\alpha_i - \alpha_i^*) - \epsilon\sum_i(\alpha_i + \alpha_i^*) + \sum_i y_i(\alpha_i - \alpha_i^*) - \sum_i\sum_j(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\phi^\top(\mathbf{x}_i)\phi(\mathbf{x}_j)$
$= -\frac{1}{2}\sum_i\sum_j(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\phi^\top(\mathbf{x}_i)\phi(\mathbf{x}_j) - \epsilon\sum_i(\alpha_i + \alpha_i^*) + \sum_i y_i(\alpha_i - \alpha_i^*)$

Kernel function: $K(\mathbf{x}_i, \mathbf{x}_j) = \phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j)$

- $\mathbf{w} = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi(x_i) \Rightarrow$ the final decision function $f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi^T(\mathbf{x}_i)\phi(\mathbf{x}) + y_j - \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j) - \epsilon$ $\mathbf{x}_j$ is any point with $\alpha_j \in (0, C)$. Recall similarity with

Kernel function: $K(\mathbf{x}_i, \mathbf{x}_j) = \phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j)$

- $\mathbf{w} = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi(x_i) \Rightarrow$ the final decision function
  $f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi^T(\mathbf{x}_i)\phi(\mathbf{x}) + y_j - \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j) - \epsilon$
  $\mathbf{x}_j$ is any point with $\alpha_j \in (0, C)$. Recall similarity with kernelized expression for Ridge Regression
- The dual optimization problem to compute the α's for SVR is:

Kernel function: $K(\mathbf{x}_i, \mathbf{x}_j) = \phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j)$

- $\mathbf{w} = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi(x_i) \Rightarrow$ the final decision function
  $f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi^T(\mathbf{x}_i)\phi(\mathbf{x}) + y_j - \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j) - \epsilon$
  $\mathbf{x}_j$ is any point with $\alpha_j \in (0, C)$. Recall similarity with kernelized expression for Ridge Regression
- The dual optimization problem to compute the α's for SVR is:

$$max_{\alpha_i, \alpha_i^*} -\frac{1}{2}\sum_i\sum_j(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\phi^\top(\mathbf{x}_i)\phi(\mathbf{x}_j)$$

$$-\epsilon\sum_i(\alpha_i + \alpha_i^*) + \sum_i y_i(\alpha_i - \alpha_i^*)$$

  s.t.
  - $\sum_i(\alpha_i - \alpha_i^*) = 0$
  - $\alpha_i, \alpha_i^* \in [0, C]$
- **We notice that the only way these three expressions involve ϕ is through $\phi^\top(\mathbf{x}_i)\phi(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j)$, for some i, j**

An Example Kernel

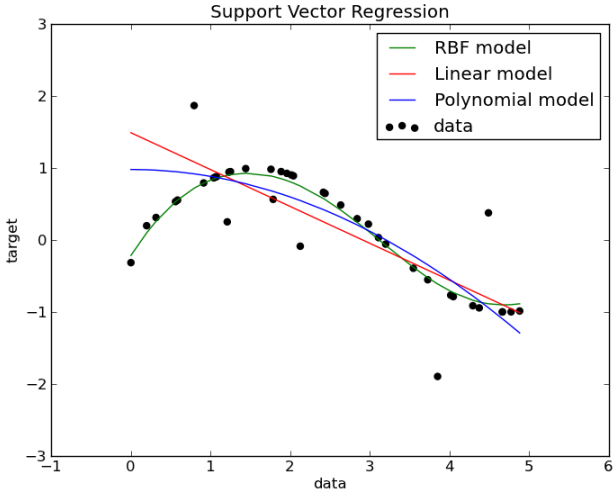
- Let $K(\mathbf{x}_1, \mathbf{x}_2) = (1 + \mathbf{x}_1^\top \mathbf{x}_2)^2$
- What $\phi(\mathbf{x})$ will give $\phi^\top(\mathbf{x}_1)\phi(\mathbf{x}_2) = K(\mathbf{x}_1, \mathbf{x}_2) = (1 + \mathbf{x}_1^\top \mathbf{x}_2)^2$
- Is such a ϕ guaranteed to exist?
- Is there a unique ϕ for given K?

An Example Kernel

- We can prove that such a ϕ exists
- For example, for a 2-dimensional $\mathbf{x}_i$:

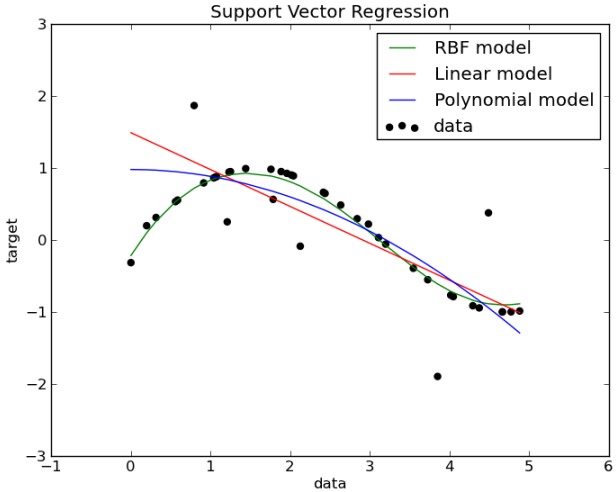
$$\phi(\mathbf{x}_i) = \begin{bmatrix} 1 \\ x_{i1}\sqrt{2} \\ x_{i2}\sqrt{2} \\ x_{i1}x_{i2}\sqrt{2} \\ x_{i1}^2 \\ x_{i2}^2 \end{bmatrix}$$

- $\phi(\mathbf{x}_i)$ exists in a 5-dimensional space
- But, to compute $K(\mathbf{x}_1, \mathbf{x}_2)$, all we need is $x_1^\top x_2$ without having to enumerate $\phi(\mathbf{x}_i)$


Support Vector Regression
RBF model
Linear model
Polynomial model
data
target
data
3
2
1
0
−1
−2
−3
−1
0
1
2
3
4
5
6

Kernels in SVR

- Recall:
  $max_{\alpha_i,\alpha_i^*} -\frac{1}{2}\sum_i\sum_j(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(\mathbf{x}_i,\mathbf{x}_j) - \epsilon\sum_i(\alpha_i+\alpha_i^*) + \sum_i y_i(\alpha_i-\alpha_i^*)$
  and the decision function:
  $f(x) = \sum_i(\alpha_i-\alpha_i^*)K(\mathbf{x}_i,\mathbf{x}) + b$
  are all in terms of the kernel $K(\mathbf{x}_i,\mathbf{x}_j)$ only
- *One can now employ any mercer kernel in SVR or Ridge Regression to implicitly perform linear regression in higher dimensional spaces*


Support Vector Regression
RBF model
Linear model
Polynomial model
data
target
data
3
2
1
0
−1
−2
−3
−1
0
1
2
3
4
5
6

Equivalent Forms of Ridge Regression

- Consider the formulation in which we limit the weights of the coefficients by putting a constraint on size of the L2 norm of the weight vector:

$$\text{argmin}_{\mathbf{w}}(\mathbf{\Phi w} - \mathbf{y})^T(\mathbf{\Phi w} - \mathbf{y})$$
$$\|\mathbf{w}\|_2^2 \leq \xi$$

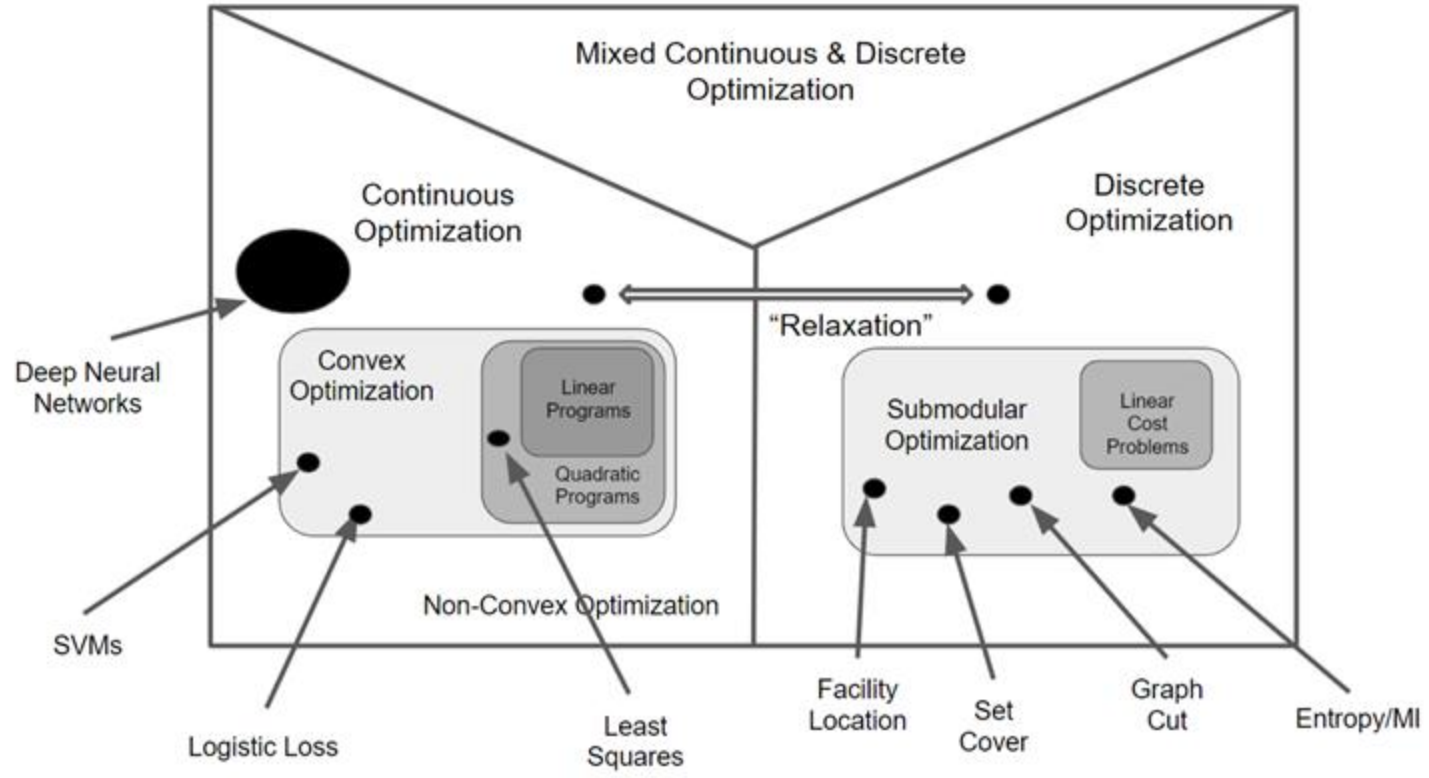
- The objective function, namely $f(\mathbf{w}) = (\mathbf{\Phi w} - \mathbf{y})^\mathbf{T}(\mathbf{\Phi w} - \mathbf{y})$ is strictly convex. The constraint function, $g(\mathbf{w}) = \|\mathbf{w}\|_2^2 - \xi$, is also convex.
- For convex $g(\mathbf{w})$, the set $\{\mathbf{w}|\mathbf{g}(\mathbf{w}) \leq \mathbf{0}\}$, is also convex. (Why?)

Algorithms for (Submodular function) optimization, with (Matroid) constraints, strategies for improving efficiency

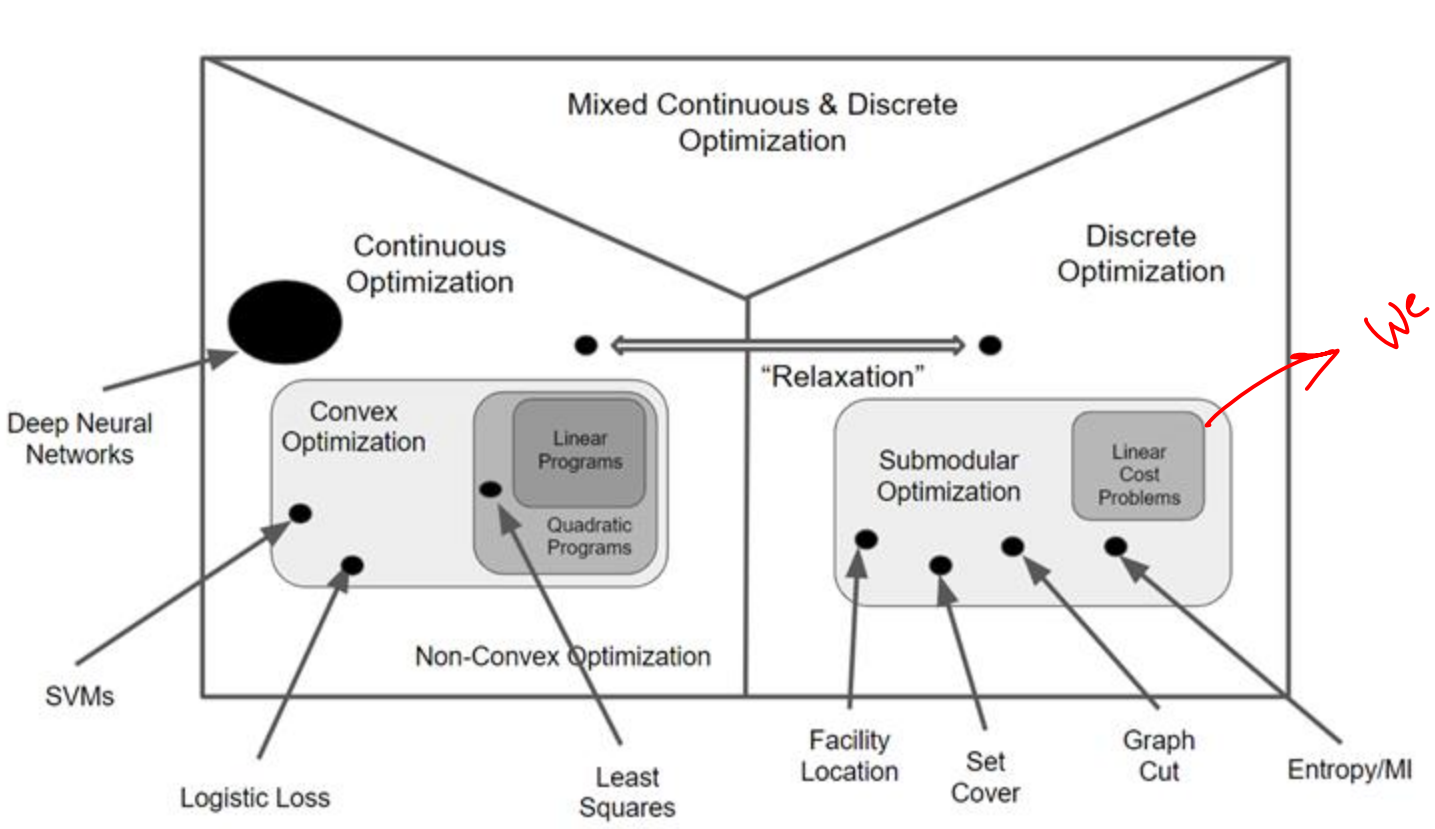
Optimization in Machine Learning

Lecture 20

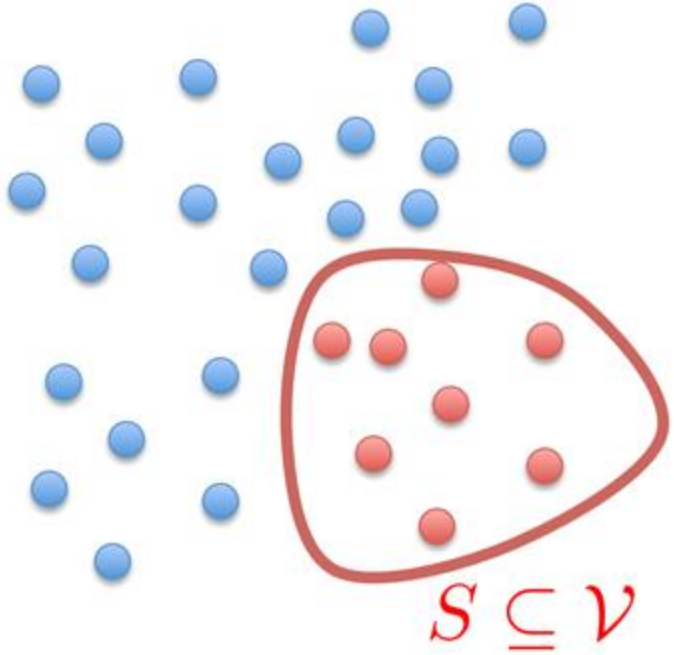
Big Picture: Continuous and Discrete Optimization



Big Picture: Continuous and Discrete Optimization



Discrete Optimization

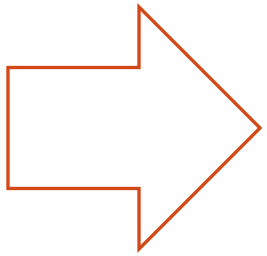


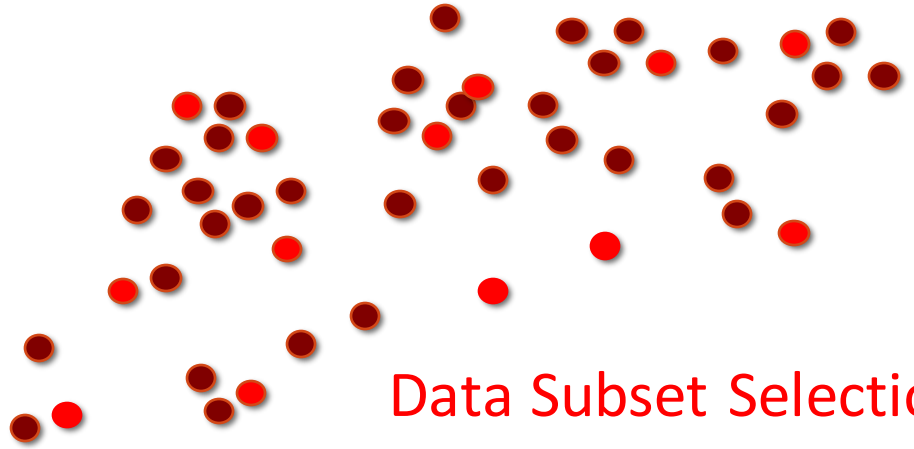
$S \subseteq \mathcal{V}$

ground set $\mathcal{V}$

(scoring) function

$F : 2^{\mathcal{V}} \rightarrow \mathbb{R}_+$

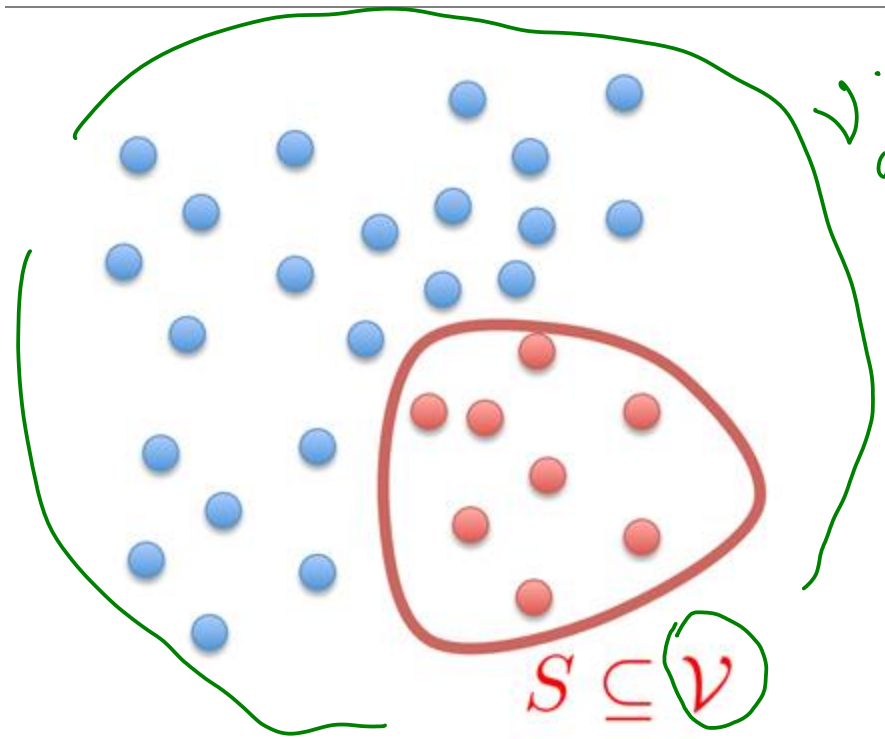




Data Subset Selection

Summarization

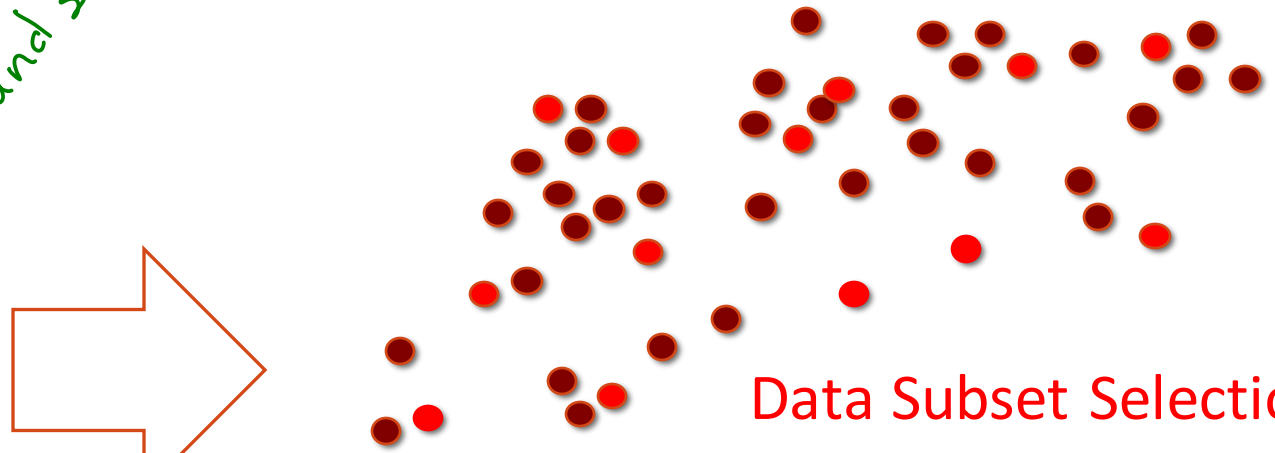
Discrete Optimization



ground set $\mathcal{V}$

(scoring) function

$F : 2^{\mathcal{V}} \rightarrow \mathbb{R}_+$



Data Subset Selection

Summarization

Acknowledgements

Slides borrowed from several sources:

1. Submodular Optimization course at UW from Jeff Bilmes
2. Tutorial on Submodular Optimization by Stefanie Jegelka, Andreas Krause and Jeff Bilmes at ICML and NIPS

Useful Material

- Fujishige, “Submodular Functions and Optimization”, 2005
- Narayanan, “Submodular Functions and Electrical Networks”, 1997
- Welsh, “Matroid Theory”, 1975
- Oxley, “Matroid Theory”, 1992 (and 2011).
- Lawler, “Combinatorial Optimization: Networks and Matroids”, 1976.
- Schrijver, “Combinatorial Optimization”, 2003
- Gruenbaum, “Convex Polytopes, 2nd Ed”, 2003.
- Additional readings that will be announced here.

Acknowledgements

Slides borrowed from several sources:

1. Submodular Optimization course at UW from Jeff Bilmes
2. Tutorial on Submodular Optimization by Stefanie Jegelka, Andreas Krause and Jeff Bilmes at ICML and NIPS

Useful material

- Some custom slides. Also slides from our tutorials at
  - IJCAI 2020: https://sites.google.com/view/ijcaitutorial2020summarization/home
  - ECAI 2020: https://sites.google.com/view/ecaitutorial2020summ/home
- Jeff's Class: https://people.ece.uw.edu/bilmes/classes/ee563_spring_2018/
- Stefanie Jegelka & Andreas Krause's 2013 ICML tutorial: http://techtalks.tv/talks/submodularity-in-machine-learning-new-directions-part-i/58125/
- Jeff's NIPS, 2013 tutorial on submodularity: http://melodi.ee.washington.edu/~bilmes/pgs/b2hd-bilmes2013-nips-tutorial.html and http://youtu.be/c4rBof38nKQ
- Andreas Krause's web page: http://submodularity.org
- Francis Bach's updated 2013 text: http://hal.archives-ouvertes.fr/docs/00/87/06/09/PDF/submodular_fot_revisd_hal.pdf
- Tom McCormick's overview paper on submodular minimization: http://people.commerce.ubc.ca/faculty/mccormick/sfmchap8a.pdf

Discrete Optimization in Machine Learning

- MAP inference in Probabilistic Models: Ising Models, DPPs
- Feature Subset Selection
- Data Partitioning
- Data Subset Selection
- Data Summarization: Text, Images, Video Summarization
- Social networks, Influence Maximization
- Natural Language Processing: words, phrases, n-grams, syntax trees, semantic structures
- Computer Vision: Image Segmentation, Image Correspondence
- Genomics and Computational Biology: cell types or assay selection, selecting peptides and proteins

Recap: Examples from our earlier lectures

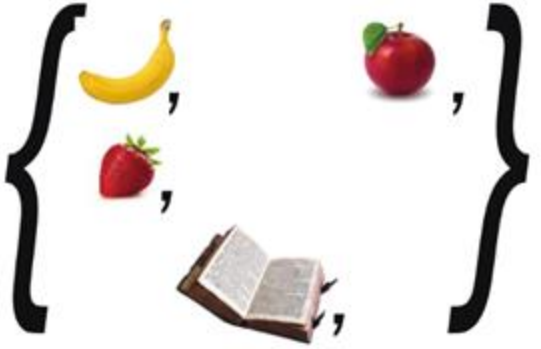
Outline

- Discrete Optimization in Machine Learning
- Definition and Intuition of Submodularity
  - Modeling Power of Submodular/Set Functions
  - Examples of Submodular Functions
  - Examples of Submodular Optimization
- Optimization Algorithms for Different Function Classes and Constraints
  - Practical Applications

Combinatorial Subset Selection Problems

$V =$ 

$f : 2^V \rightarrow \mathbb{R}$

$A =$ 

Choose Subset $A \subseteq V$

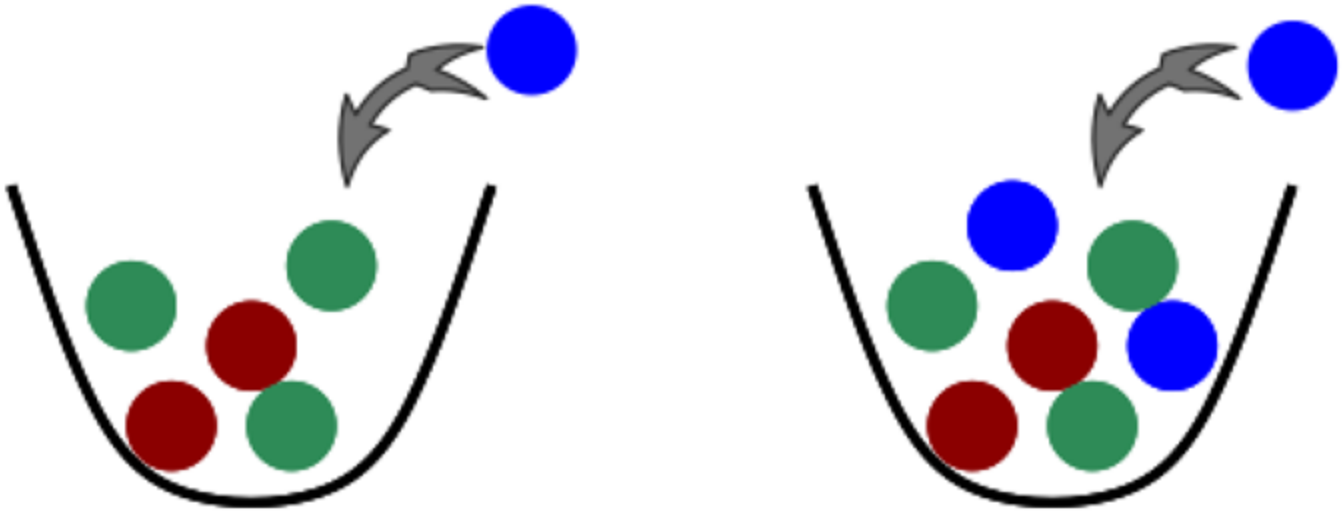
$f(A) = 22$

General Set function Optimization: very hard!

What if there is some special structure?

Submodular Functions

$$f(A \cup v) - f(A) \geq f(B \cup v) - f(B), \text{ if } A \subseteq B$$



$f = \#$ of distinct colors of balls in the urn.

Negative of a Submodular Function is a Super-modular Function!

Q: Is it necessary for a submodular fn f that $f(A) \leq f(B)$ if $A \subseteq B$?

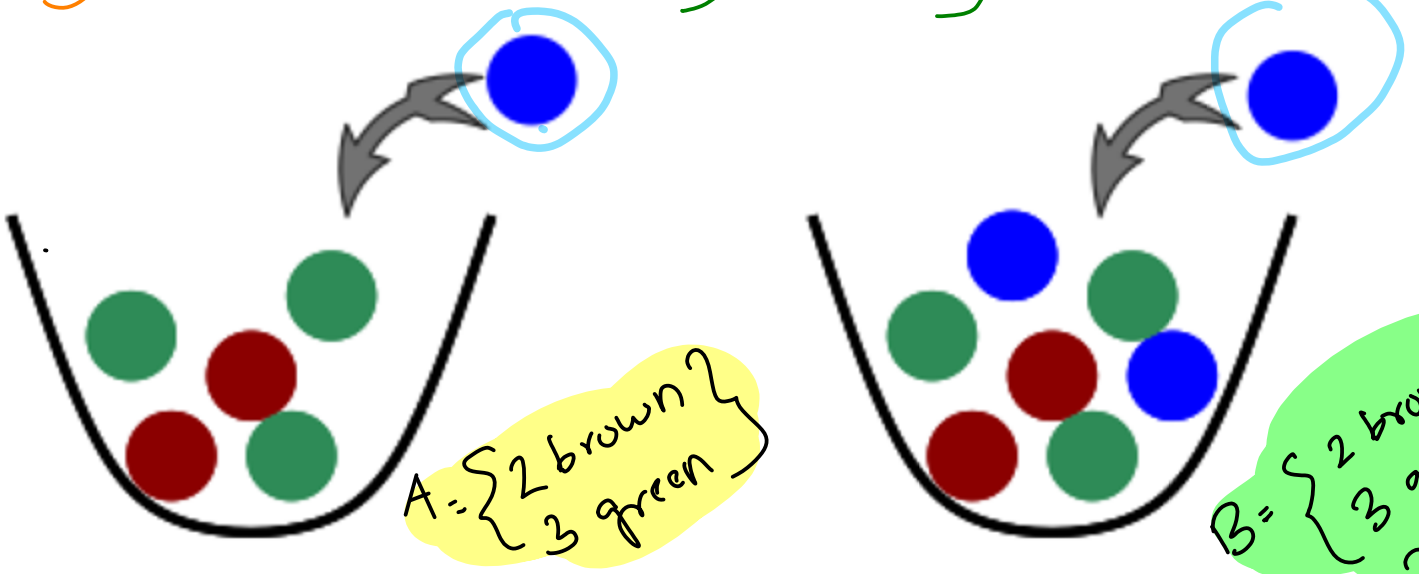
Ans: No

Prove or disprove using a counter example

Submodular Functions

$$1 = f(A \cup v) - f(A) \geq f(B \cup v) - f(B), \text{ if } A \subseteq B$$

3 2 3 3



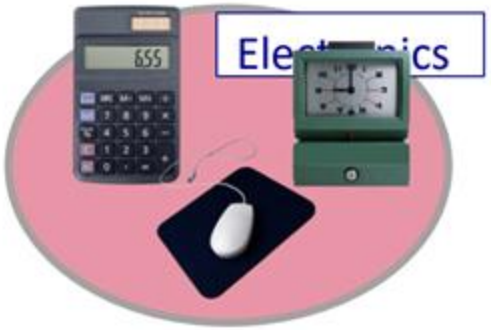
Negative of a Submodular Function is a Super-modular Function!

f = # of distinct colors of balls in the urn.

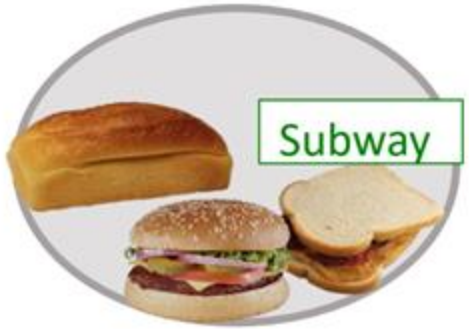
Submodular Functions

$f(\text{fries, drink}) - f(\text{fries}) \geq f(\text{burger, drink}) - f(\text{burger, fries})$

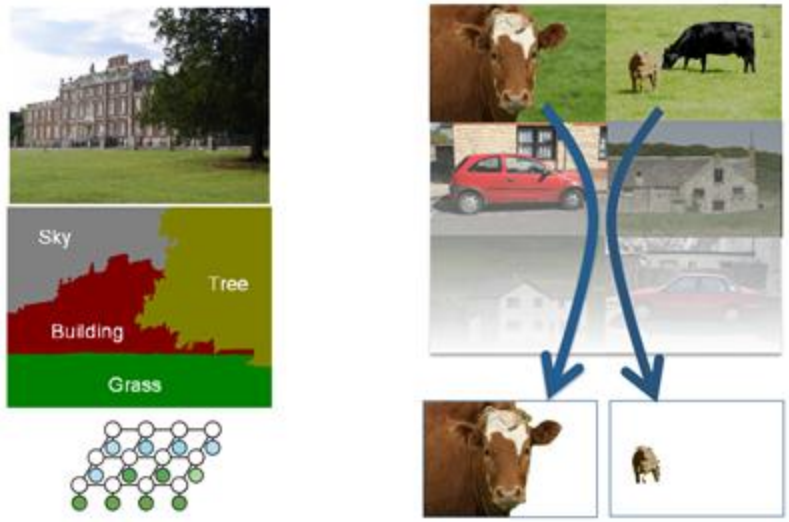
Diminishing Returns!



The more items you buy, the more the discount!



Submodular Optimization in Machine Learning

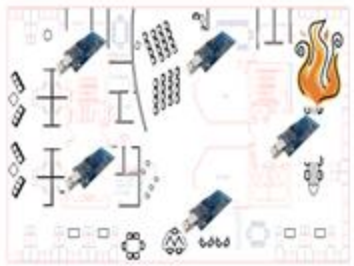


$F(S) = \text{coherence} + \text{likelihood}$

Discrete Labeling

$F(S) = \text{relevance} + \text{diversity or coverage}$

Summarization



- where put sensors?
- which experiments?
- summarization

$F(S) = \text{“information”}$

Sensor Placement

Modular Functions

- each element e has a weight $w(e)$

$$F(S) = \sum_{e \in S} w(e)$$

$A \subset B$

$$F(A \cup e) - F(A) = w(e) \quad = \quad F(B \cup e) - F(B) = w(e)$$

Modular Functions are both submodular and super-modular!

Modular Functions

- each element e has a weight $w(e)$

$$F(S) = \sum_{e \in S} w(e)$$

$A \subset B$

$$F(A \cup e) - F(A) = w(e) \quad = \quad F(B \cup e) - F(B) = w(e)$$

Modular Functions are both submodular and super-modular!

Monotone Submodular Functions

- A set function is called monotonic if
  $$A \subseteq B \subseteq V \Rightarrow F(A) \leq F(B)$$

- Examples:
  - Influence in social networks [Kempe et al KDD '03]
  - For discrete RVs, entropy $F(A) = H(X_A)$ is monotonic:
    Suppose $B = A \cup C$. Then
    $$F(B) = H(X_A, X_C) = H(X_A) + H(X_C \mid X_A) \geq H(X_A) = F(A)$$
  - Information gain: $F(A) = H(Y) - H(Y \mid X_A)$

Q: Is it necessary for a submodular fn f that $f(A) \leq f(B)$ if $A \subseteq B$?

Ans: !!

Prove or disprove using a counter example

Monotone Submodular Functions

- A set function is called monotonic if

  $A \subseteq B \subseteq V \Rightarrow F(A) \leq F(B)$

If a modular function has non-negative weights it will be monotone

Function value should not decrease as you grow the set

Eg: Cardinality function which is also modular

- Examples:
  - Influence in social networks [Kempe et al KDD '03]
  - For discrete RVs, entropy $F(A) = H(X_A)$ is monotonic:

    Suppose $B = A \cup C$. Then

    $F(B) = H(X_A, X_C) = H(X_A) + H(X_C \mid X_A) \geq H(X_A) = F(A)$

    H/w: Verify
  - Information gain: $F(A) = H(Y) - H(Y \mid X_A)$

Always monotone. But only under some (Naive Bayes like conditional independence) conditions is it submodular

Suppose entropy were both monotone and submodular, it would still not mean that there is no function that is submodular but not monotone (or vice versa)

Not every monotone function is submodular

- **Eg:**

	$F_1(A)$
{}	0
{a}	1
{b}	1
{a,b}	3

- this is a form of attractive potentials...Monotone but not submodular

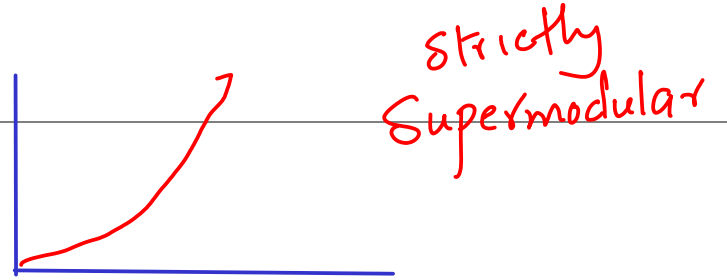
- In fact it is (strictly) supermodular.

Its negative is (strictly) submodular.

- **Eg: Conventional Information gain based on entropy – it is a difference of two submodular functions**

- Not every submodular function is montone
  - We will see examples of diversity based functions such as logDet and disparity-min-sum

Not every monotone function is submodular



- **Eg:**

	$F_1(A)$
{}	0
{a}	1
{b}	1
{a,b}	3

- this is a form of attractive potentials...Monotone but not submodular

- In fact it is (strictly) supermodular.

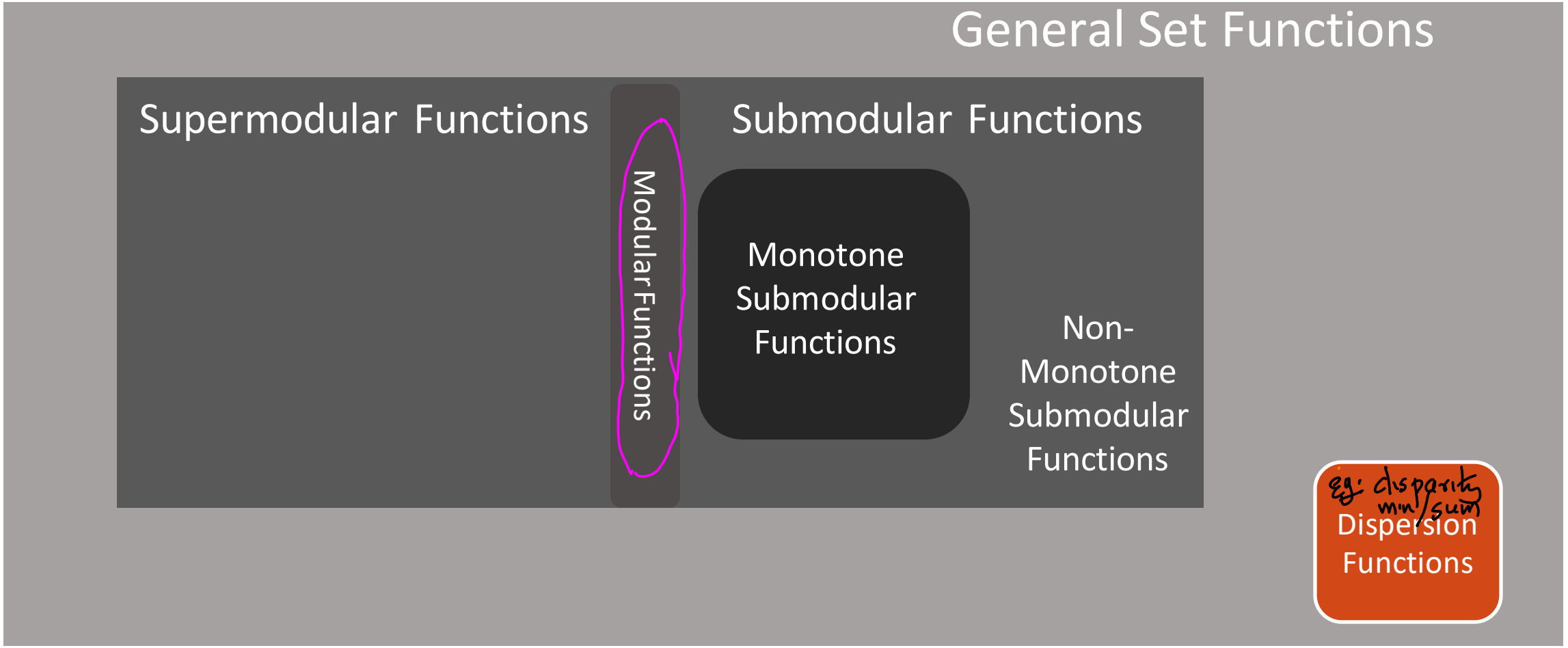
Its negative is (strictly) submodular.

$V = \{a, b\}$

- **Eg: Conventional Information gain based on entropy – it is a difference of two submodular functions**

- Not every submodular function is montone
  - We will see examples of diversity based functions such as logDet and disparity-min-sum

Summary of Function Classes



Properties of Submodular Functions

- Convex Combinations of Submodular Functions are Submodular
- Intersections with Fixed Sets is Submodular (Restrictions)
- Unions with Fixed Sets is Submodular (Conditioning)
- Complement Functions are Submodular (Reflection)
- Two equivalent definitions of submodular functions
- Minimum and Maximum of Submodular Functions
- Information functions
- Submodularity and Convexity
- Submodularity and Concavity

Properties of Submodular Functions

- Convex Combinations of Submodular Functions are Submodular
- Intersections with Fixed Sets is Submodular (Restrictions)
- Unions with Fixed Sets is Submodular (Conditioning)
- Complement Functions are Submodular (Reflection)
- Two equivalent definitions of submodular functions
- Minimum and Maximum of Submodular Functions ?
- Information functions
- Submodularity and Convexity
- Submodularity and Concavity

[Do you recall any of these from our discussions during the course? Any projects in this space?]

Convex Combinations of Submodular Functions

$F_1,\ldots,F_m$ submodular functions on V and $\lambda_1,\ldots,\lambda_m > 0$

Then: $F(A) = \sum_i \lambda_i F_i(A)$ is submodular

Submodularity closed under nonnegative linear combinations!

Extremely useful fact:

- $F_\theta(A)$ submodular ➔ $\sum_\theta P(\theta) F_\theta(A)$ submodular!
- Multicriterion optimization
- A basic proof technique! ☺

Convex Combinations of Submodular Functions

$F_1,\dots,F_m$ submodular functions on V and $\lambda_1,\dots,\lambda_m > 0$

Then: $F(A) = \sum_i \lambda_i F_i(A)$ is submodular

$$\sum_i \lambda_i [F_i(A \cup \{v\}) - F_i(A)] \geq \sum_i \lambda_i [F_i(B \cup \{v\}) - F_i(B)] \quad \forall\ A \subseteq B$$

Submodularity closed under nonnegative linear combinations!

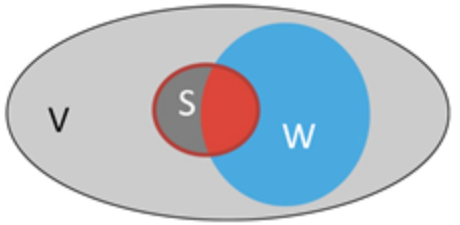
Extremely useful fact:

- $F_\theta(A)$ submodular ➔ $\sum_\theta P(\theta) F_\theta(A)$ submodular!
- Multicriterion optimization
- A basic proof technique! ☺

More Properties

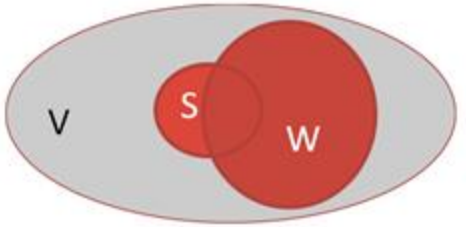
- Restriction: F(S) submodular on V, W subset of V

  Then $F'(S) = F(S \cap W)$ is submodular



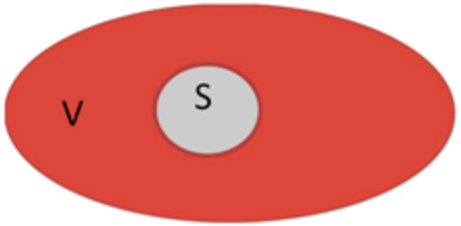
- Conditioning: F(S) submodular on V, W subset of V

  Then $F'(S) = F(S \cup W)$ is submodular



- Reflection: F(S) submodular on V

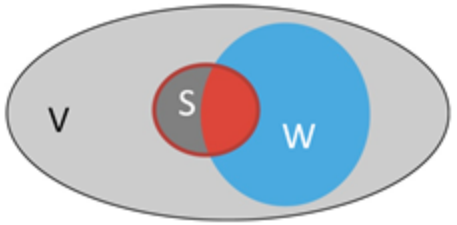
  Then $F'(S) = F(V \setminus S)$ is submodular



More Properties

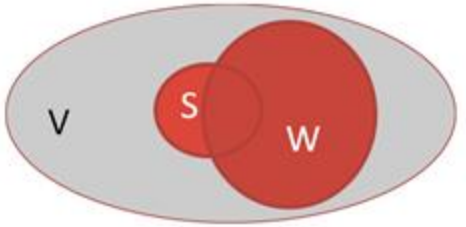
- Restriction: F(S) submodular on V, W subset of V

  Then $F'(S) = F(S \cap W)$ is submodular



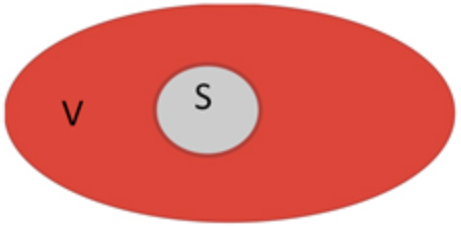
- Conditioning: F(S) submodular on V, W subset of V

  Then $F'(S) = F(S \cup W)$ is submodular



- Reflection: F(S) submodular on V

  Then $F'(S) = F(V \setminus S)$ is submodular

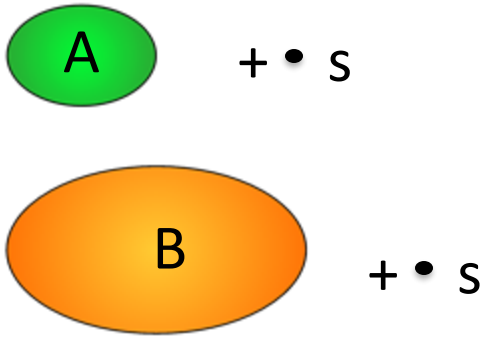


F'(S) is the submodular function F applied to the complement of S

Two Equivalent Definitions of Submodularity

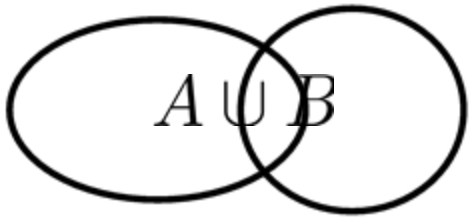
- Diminishing gains: for all $A \subseteq B \subseteq V$

$$F(A \cup s) - F(A) \quad \geq \quad F(B \cup s) - F(B)$$



- Union-Intersection: for all $A, B \subseteq V$

$$F(A) + F(B) \quad \geq \quad F(A \cup B) + F(A \cap B)$$

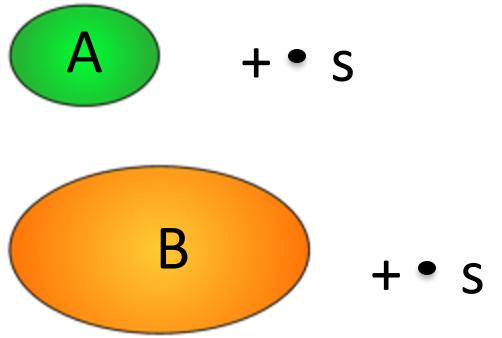


Can you try and prove that the two above definitions of submodularity are equivalent?

Two Equivalent Definitions of Submodularity

- Diminishing gains: for all $A \subseteq B \subseteq V$ Already seen

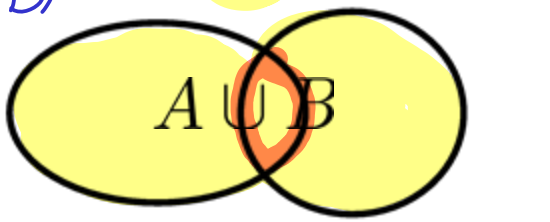
$$F(A \cup s) - F(A) \quad \geq \quad F(B \cup s) - F(B)$$



- Union-Intersection: for all $A, B \subseteq V$

$$F(A) + F(B) \quad \geq \quad F(A \cup B) + F(A \cap B)$$

If F were cardinality $F(A) + F(B) = F(A \cup B) + F(A \cap B)$



Can you try and prove that the two above definitions of submodularity are equivalent?

Step 1: Define sets

Let:

- $X = A \cap B$
- $Y = A \setminus B$
- $Z = B \setminus A$

Then:

- $A = X \cup Y$
- $B = X \cup Z$
- $A \cup B = X \cup Y \cup Z$

Step 2: Expand $F(A)$ and $F(B)$ from $F(X)$

We write $F(A)$ as:

$$F(A) = F(X) + \sum_{y \in Y} [F(X \cup \{y_1, \ldots, y_i\}) - F(X \cup \{y_1, \ldots, y_{i-1}\})]$$

Similarly,

$$F(B) = F(X) + \sum_{z \in Z} [F(X \cup \{z_1, \ldots, z_i\}) - F(X \cup \{z_1, \ldots, z_{i-1}\})]$$

Let's denote:

- $F(A) = F(X) + \Delta_Y$
- $F(B) = F(X) + \Delta_Z$

HOMEWORK: COMPLETE THE ABOVE STEPS OF DERIVATION