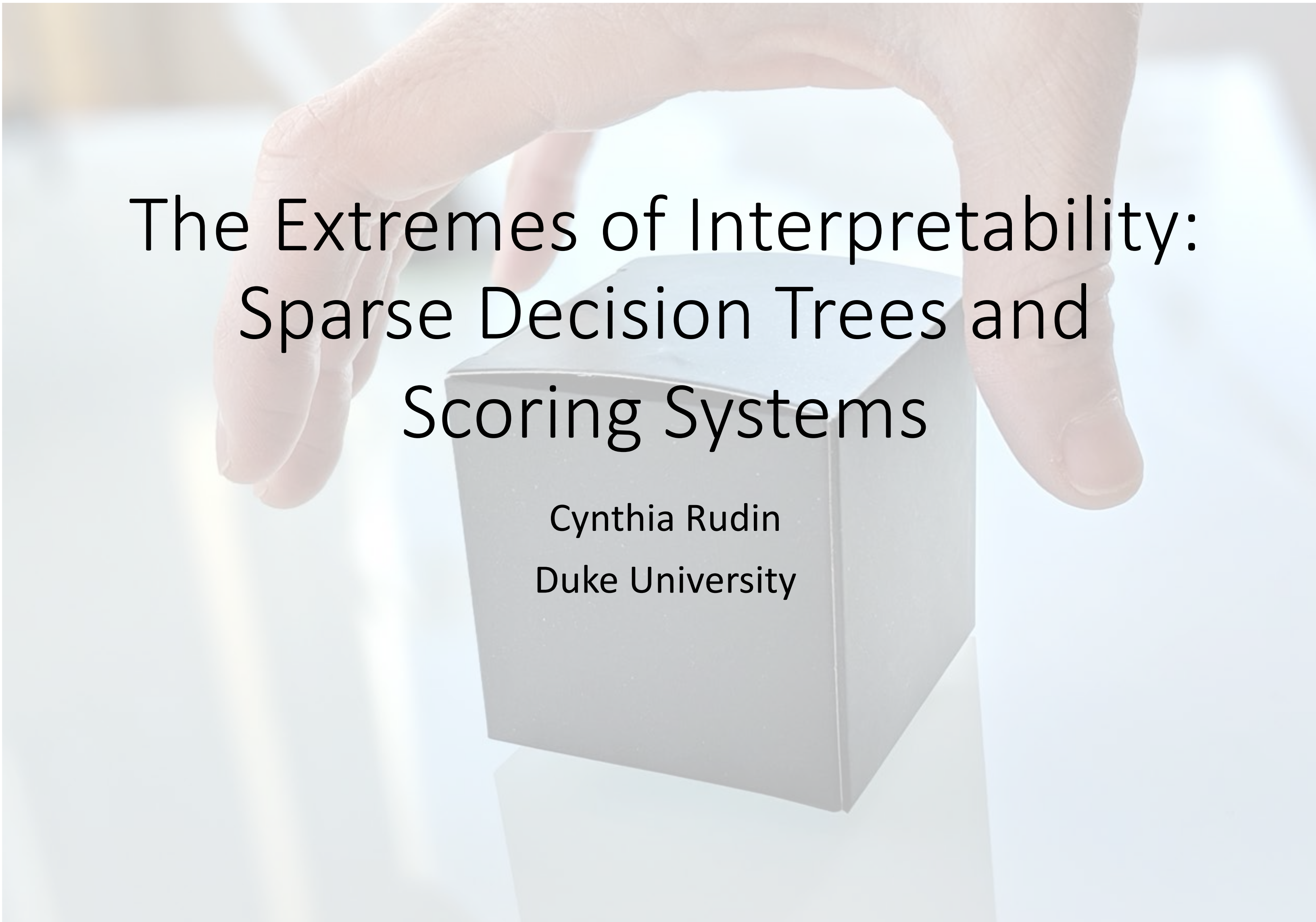
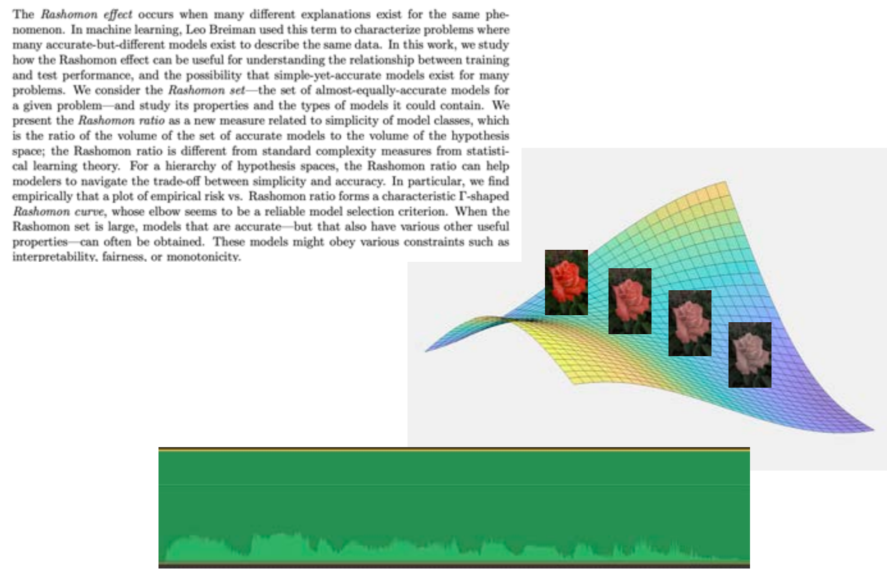


The Extremes of Interpretability: Sparse Decision Trees and Scoring Systems

Cynthia Rudin
Duke University

Problem spectrum

age 45
congestive heart failure? yes
takes aspirin
smoking? no
gender M
exercise? yes
allergies? no
number of past strokes 2
diabetes? yes



The *Rashomon effect* occurs when many different explanations exist for the same phenomenon. In machine learning, Leo Breiman used this term to characterize problems where many accurate-but-different models exist to describe the same data. In this work, we study how the Rashomon effect can be useful for understanding the relationship between training and test performance, and the possibility that simple-yet-accurate models exist for many problems. We consider the *Rashomon set*—the set of almost-equally-accurate models for a given problem—and study its properties and the types of models it could contain. We present the *Rashomon ratio* as a new measure related to simplicity of model classes, which is the ratio of the volume of the set of accurate models to the volume of the hypothesis space; the Rashomon ratio is different from standard complexity measures from statistical learning theory. For a hierarchy of hypothesis spaces, the Rashomon ratio can help modelers to navigate the trade-off between simplicity and accuracy. In particular, we find empirically that a plot of empirical risk vs. Rashomon ratio forms a characteristic Γ-shaped *Rashomon curve*, whose elbow seems to be a reliable model selection criterion. When the Rashomon set is large, models that are accurate—but that also have various other useful properties—can often be obtained. These models might obey various constraints such as interpretability, fairness, or monotonicity.

Tabular: All features are interpretable
- many problems in criminal justice, healthcare, social sciences, equipment reliability & maintenance, etc.
- features include counts, categorical data

Raw: Features are individually uninterpretable
- pixels/voxels, words, a bit of a sound wave

Problem spectrum

Very sparse models (trees, scoring systems)

With minor pre-processing, all methods have similar performance

Neural networks

Tabular: All features are interpretable

- many problems in criminal justice, healthcare, social sciences, equipment reliability & maintenance, etc.
- features include counts, categorical data

Raw: Features are individually uninterpretable

- pixels/voxels, words, a bit of a sound wave

- …But don’t they lose accuracy?

  - Explainable Machine Challenge (credit scoring data from FICO)

  - Florida COMPAS data (criminal recidivism)

The New York Times

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

232

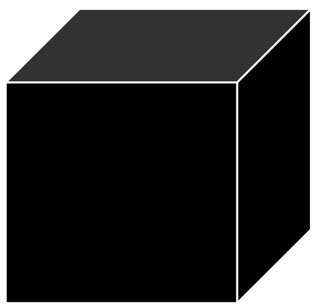
Glenn Rodriguez was denied parole because of a miscalculated “COMPAS” score.

A typographical error in a COMPAS score can lead to years of extra prison time.

How accurate is COMPAS?

COMPAS vs. CORELS

COMPAS: (Correctional Offender Management Profiling for Alternative Sanctions)

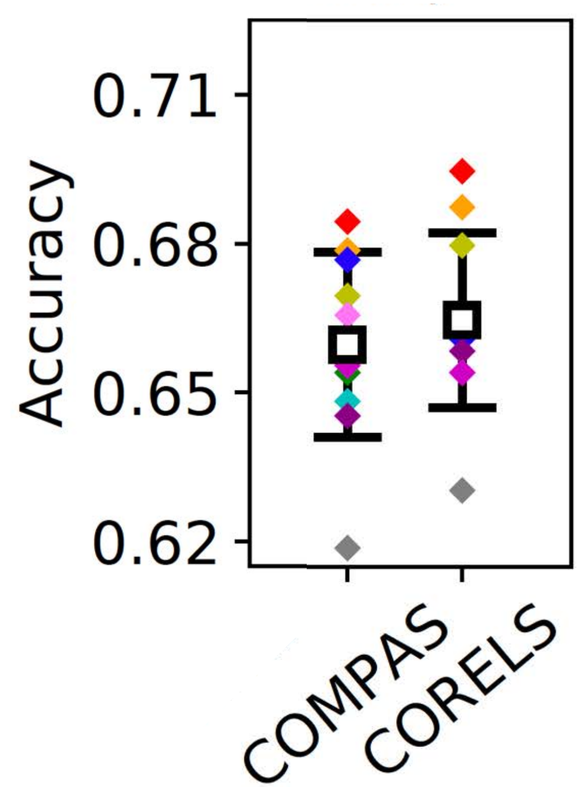


CORELS: (Certifiably Optimal RulE ListS, with Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, and Margo Seltzer, KDD 2017 & JMLR 2018)

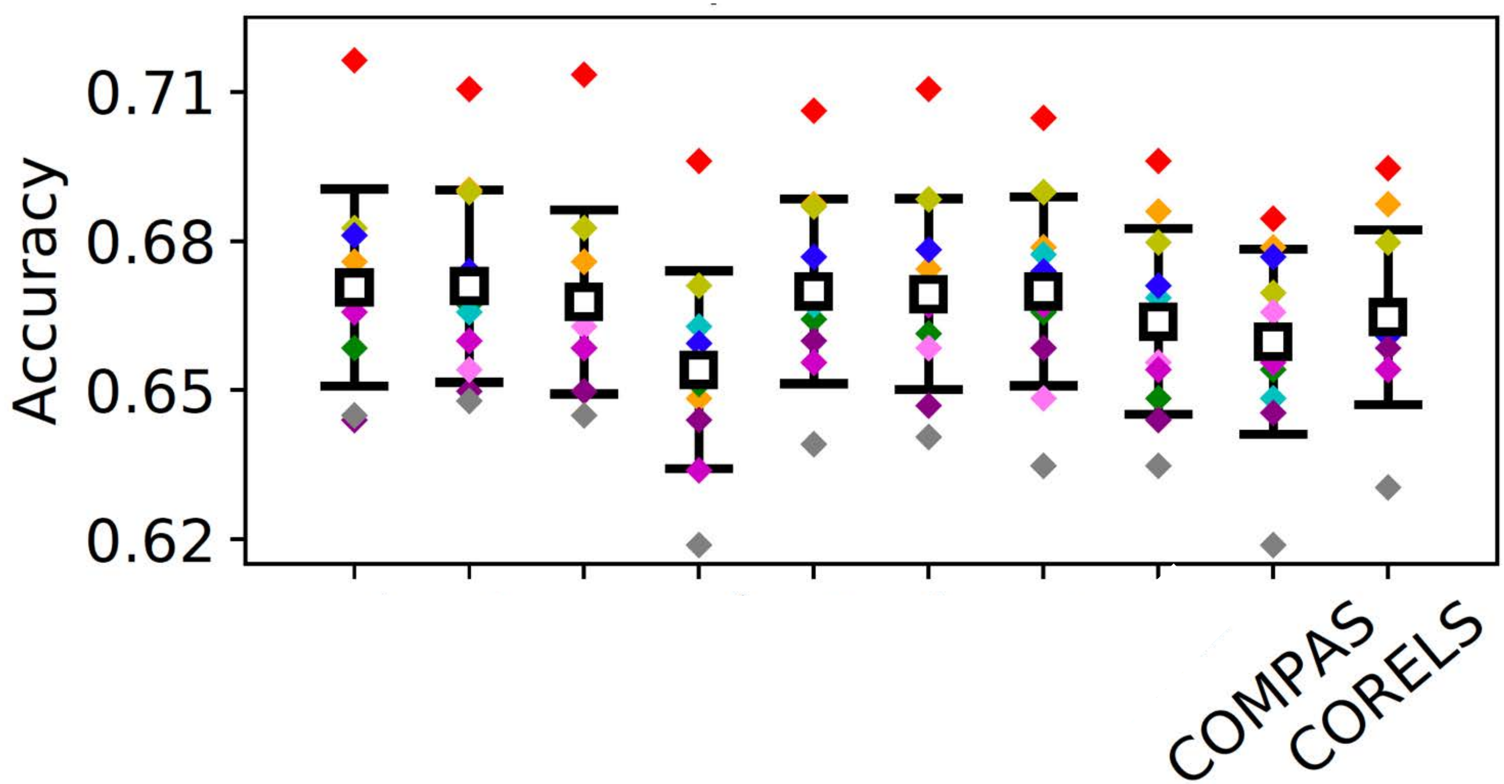
Here is the machine learning model:

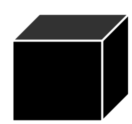
If age=19-20 and sex=male, then predict arrest
else if age=21-22 and priors=2-3 then predict arrest
else if priors >3 then predict arrest
else predict no arrest

Prediction of re-arrest within 2 years



Prediction of re-arrest within 2 years





If age=19-20 and sex=male, then predict arrest
else if age=21-22 and priors=2-3 then predict arrest
else if priors >3 then predict arrest
else predict no arrest

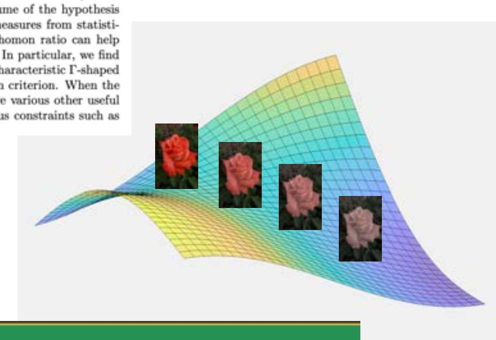
Problem spectrum

age 45
congestive heart failure? yes
takes aspirin
smoking? no
gender M
exercise? yes
allergies? no
number of past strokes 2
diabetes? yes

Tabular: All features are interpretable

- many problems in criminal justice, healthcare, social sciences, equipment reliability & maintenance, etc.
- features include counts, categorical data

The *Rashomon effect* occurs when many different explanations exist for the same phenomenon. In machine learning, Leo Breiman used this term to characterize problems where many accurate-but-different models exist to describe the same data. In this work, we study how the Rashomon effect can be useful for understanding the relationship between training and test performance, and the possibility that simple-yet-accurate models exist for many problems. We consider the *Rashomon set*—the set of almost-equally-accurate models for a given problem—and study its properties and the types of models it could contain. We present the *Rashomon ratio* as a new measure related to simplicity of model classes, which is the ratio of the volume of the set of accurate models to the volume of the hypothesis space; the Rashomon ratio is different from standard complexity measures from statistical learning theory. For a hierarchy of hypothesis spaces, the Rashomon ratio can help modelers to navigate the trade-off between simplicity and accuracy. In particular, we find empirically that a plot of empirical risk vs. Rashomon ratio forms a characteristic Γ-shaped *Rashomon curve*, whose elbow seems to be a reliable model selection criterion. When the Rashomon set is large, models that are accurate—but that also have various other useful properties—can often be obtained. These models might obey various constraints such as interpretability, fairness, or monotonicity.

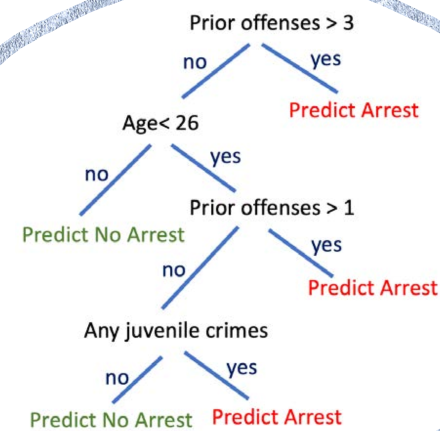


Raw: Features are individually uninterpretable

- pixels/voxels, words, a bit of a sound wave

The Extremes of Interpretability:

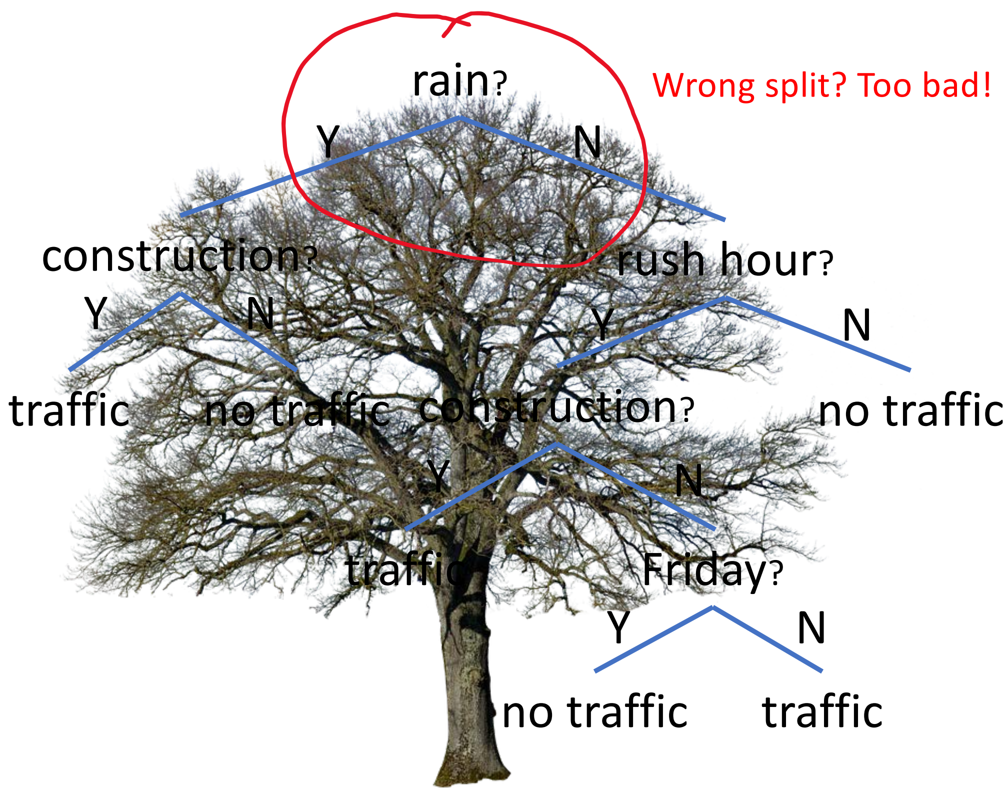
- Optimal decision trees
- Scoring systems

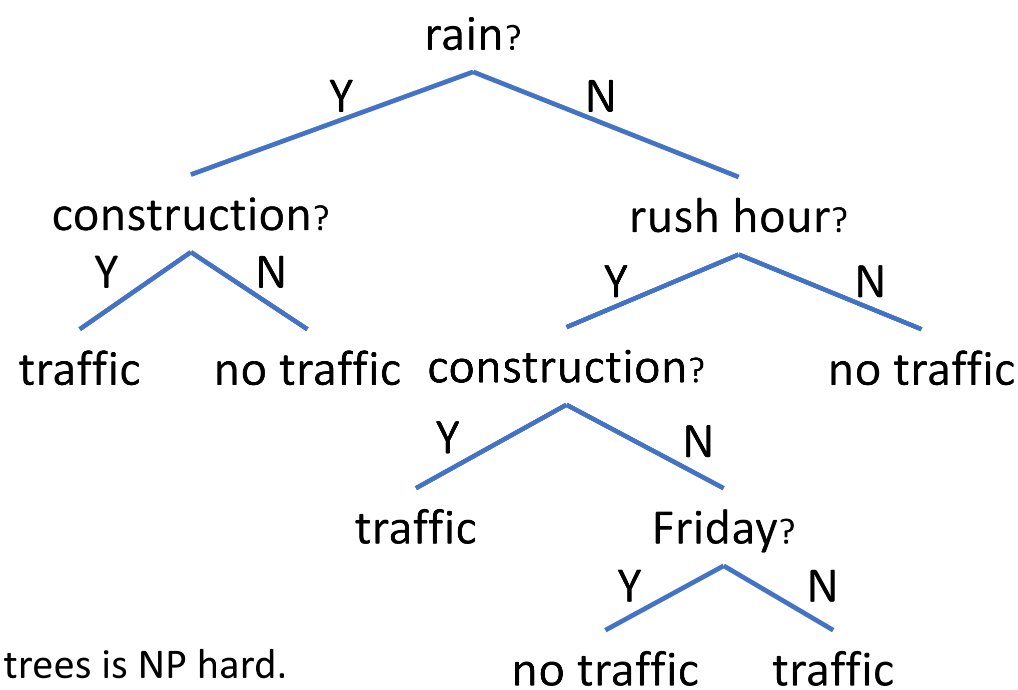


1.	Any cEEG Pattern with Frequency **2 H**z	1 point		···
2.	**E**pileptiform Discharges	1 point	+	···
3.	Patterns include [**L**PD, LRDA, BIPD]	1 point	+	···
4.	**P**atterns Superimposed with Fast or Sharp Activity	1 point	+	···
5.	Prior **S**eizure	1 point	+	···
6.	**B**rief Rhythmic Discharges	**2** points	+	···
		SCORE	=	···

SCORE	0	1	2	3	4	5	6+
RISK	<5%	11.9%	26.9%	50.0%	73.1%	88.1%	95.3%

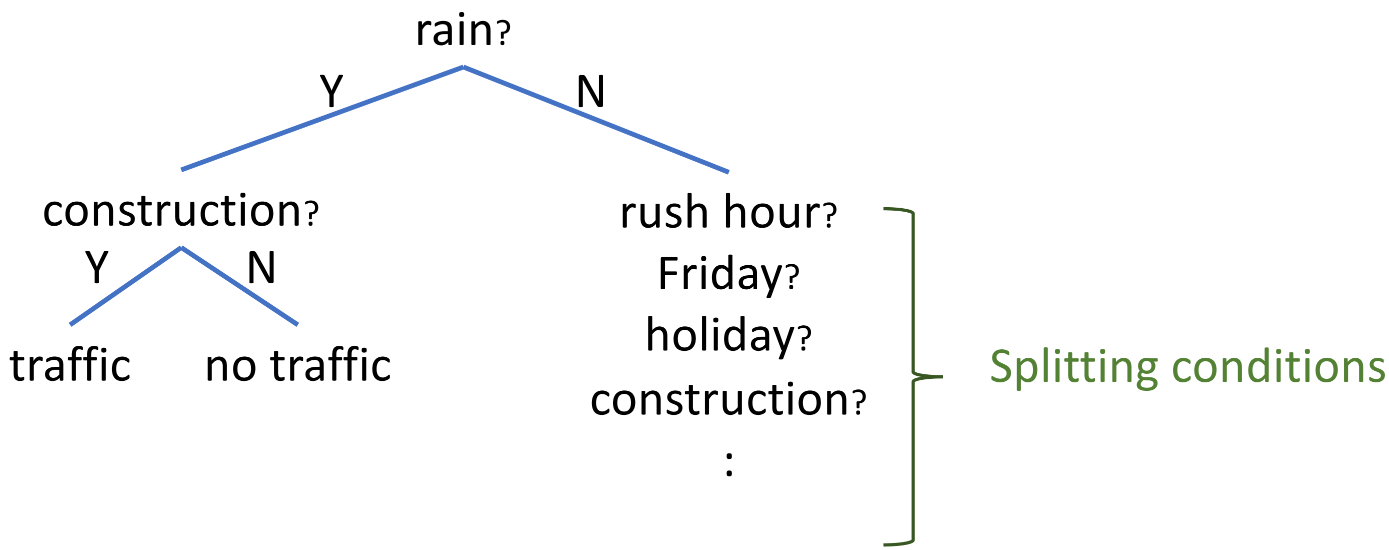
Optimal Sparse Decision Trees





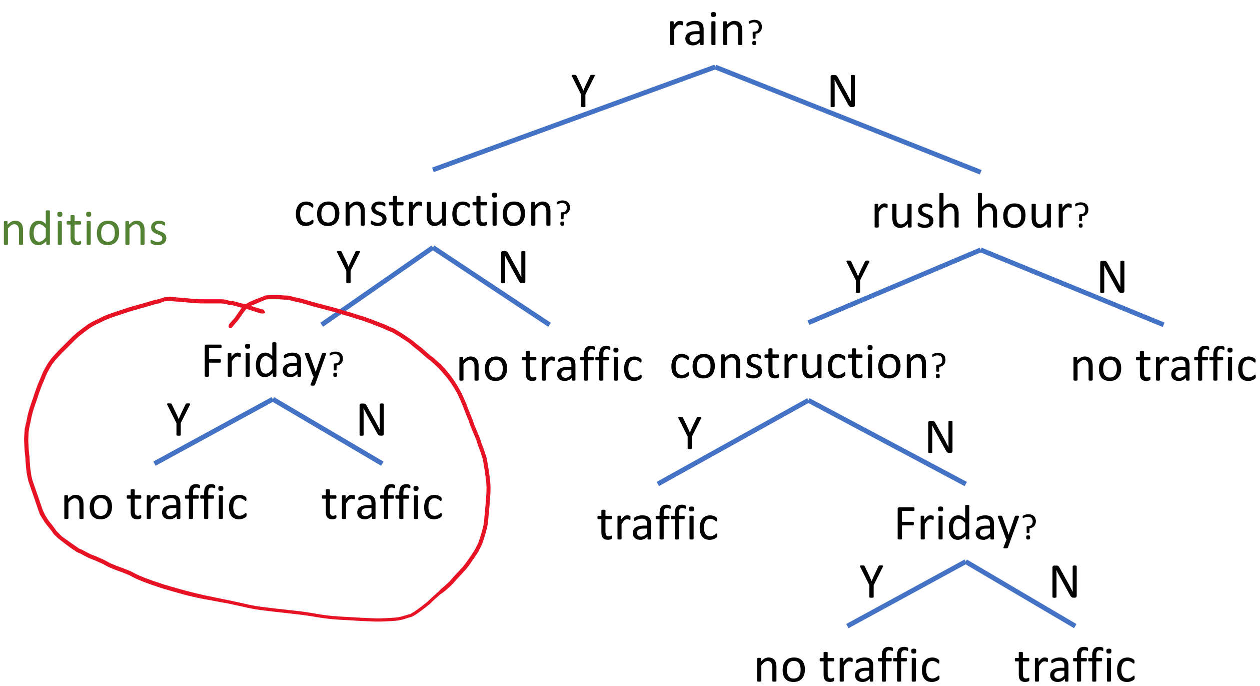
Optimal sparse decision trees is NP hard.
Factorial in the number of variables.

Greedy construction: both the splitting and pruning conditions are based on statistical testing.



Greedy construction: both the splitting and pruning conditions are based on statistical testing.

Pruning conditions



Automatic Interaction Detection (AID) (Morgan & Sonquist, 1963) regression trees

↓

THeta Automatic Interaction Detection (THAID) (Messenger & Mandell, 1972), classification trees

↓

CHi-squared Automatic Interaction Detector (CHAID) (Kass, 1980)

↓

Classification And Regression Trees (CART) (Breiman *et al.*, 1984

↓

ID3 (Quinlan, 1986), C4.5 (Quinlan, 1993)

Ensemble methods: Random Forest,
Boosted Decision Trees, BART

Other problems:
longitudinal data, survival curves:
Segal (1992), Simonoff (several papers)

Global tree optimization, mid-1990's
Bennett, Street, Mangasarian

Improvements in splitting criteria for classification and regression
Hypothesis tests, de-biasing (Strobl), missing variables

Global Tree Optimization:
A Non-greedy Decision Tree Algorithm

1994

Kristin P. Bennett
Email bennek@rpi.edu
Department of Mathematical Sciences
Rensselaer Polytechnic Institute
Troy, NY 12180 *

Tutorials (Murthy 1998, Loh 2014, L. Rokach & O. Maimon 2004 - beware)

Ensemble methods: Random Forest,
Boosted Decision Trees, BART

Global tree optimization, mid-1990's
Bennett, Street, Mangasarian

Improvements in splitting criteria for classification and regression
Hypothesis tests, de-biasing (Strobl), missing variables

Other problems:
longitudinal data, survival curves:
Segal (1992), Simonoff (several papers)

Ensemble methods: Random Forest,
Boosted Decision Trees, BART

Global tree optimization, mid-1990's
Bennett, Street, Mangasarian

Improvements in splitting criteria for classification and regression
Hypothesis tests, de-biasing (Strobl), missing variables

Other problems:
longitudinal data, survival curves:
Segal (1992), Simonoff (several papers)

"Trees sometimes choose irrelevant variables."

"Trees are sometimes 10% worse than ensembles."

"We can't tell how close to optimality our trees are."

"We need new splitting criteria for each objective."

What I hope:

Fully optimal decision trees. User picks objective:

classification accuracy, weighted accuracy, F-score, AUC, partial AUC, precision, recall, etc.

Regularize with sparsity for interpretability.

Adapt to handle missing data / biases, etc.

Adapt to other problems

Fully optimal decision trees. User picks objective:

classification accuracy, weighted accuracy, F-score, AUC, partial AUC, precision, recall, etc.

Regularize with sparsity for interpretability.

Fully optimal decision trees. User picks objective:

classification accuracy, weighted accuracy, F-score, AUC, partial AUC, precision, recall, etc.

Regularize with sparsity for interpretability.

Approaches:

- Genetic programming (e.g., Fan & Gray, 2005, Janikow & Malatkar, 2011), or neural networks
  - no optimality gap
- For classification data that is able to be perfectly separated: SAT solvers (Narodytska et al., 2018, Janota 2020)

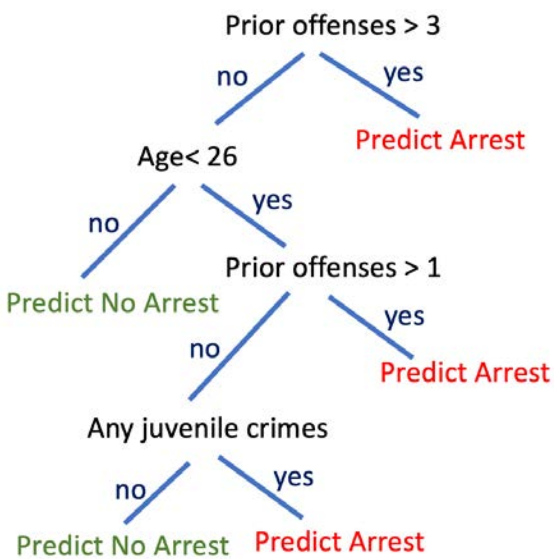
- Mathematical programming solvers (Bennett mid-1990's, Blanquero et al., 2018, Menickelly et al., 2018; Vilas Boas et al., 2019, Verwer & Zhang, BinOCT 2019)

- Dynamic programming / Branch and Bound
  - Garofalakis et al., DTC, 2003 (less relevant since it just finds subtrees of greedy-grown trees)
  - Nijssen & Fromont, DL8, 2007, Nijssen et al., DL8.5, 2020
  - Angelino et al, CORELS, 2018, Hu et al., OSDT 2019, Lin et al., GOSDT, 2020

with Jimmy Lin, Chudi Zhong, Diane Hu, Margo Seltzer

$$\min_{\text{tree}} \hat{L}(\text{tree},\{(x_i,y_i)\}_i) \text{ where}$$

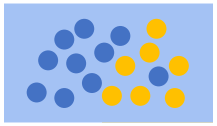
$$\hat{L}(\text{tree},\{(x_i,y_i)\}_i) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} 1_{[\text{tree}(x_i)\neq y_i]}}_{\text{Misclassification error}} + C\underbrace{(\#\,\text{leaves in tree})}_{\text{Sparsity}}$$



An example of an optimal tree on the Broward County Florida re-arrest data

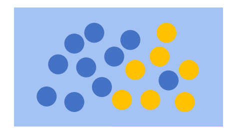
Dynamic programming / Branch and Bound

Start with the full dataset and a naive label

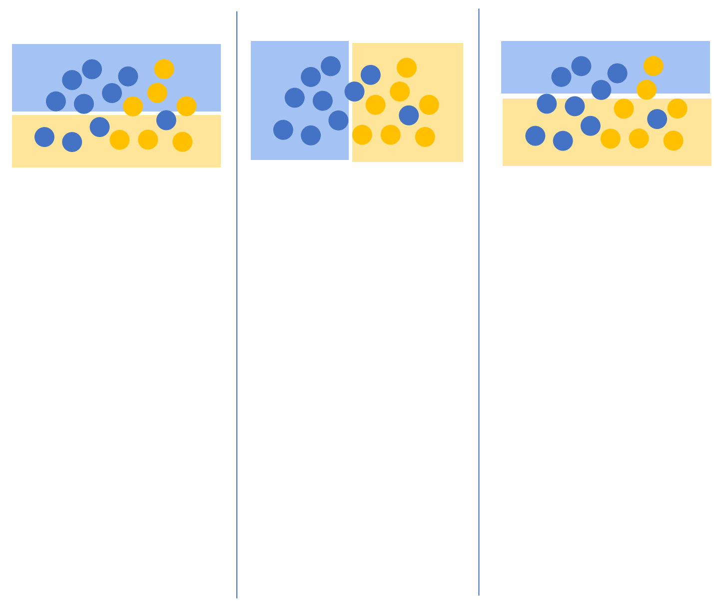


Dynamic programming / Branch and Bound

Start with the full dataset and a naive label

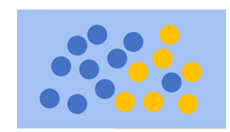


Split it into subsets using each feature

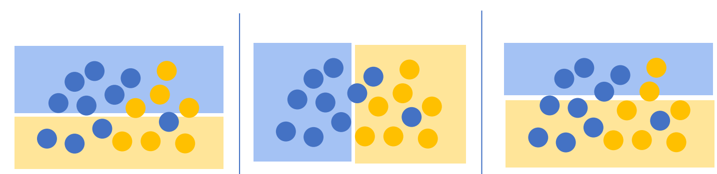


Dynamic programming / Branch and Bound

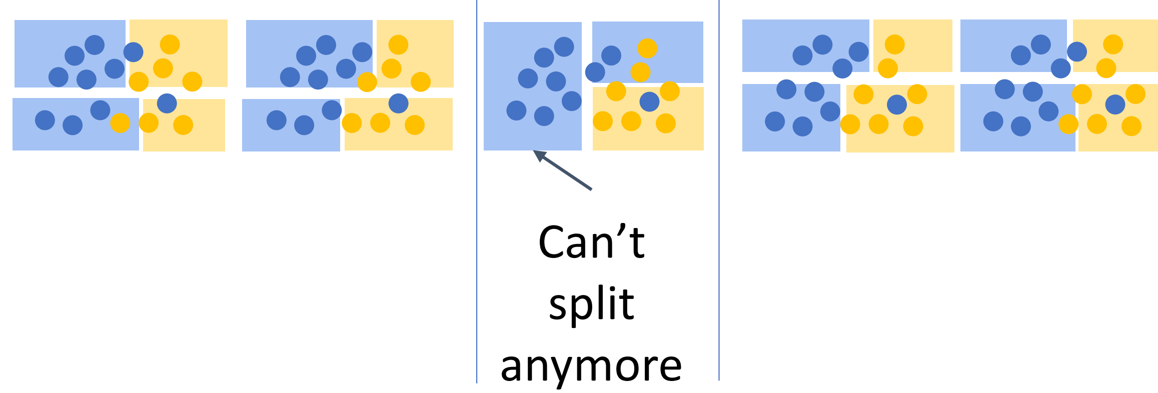
Start with the full dataset and a naive label



Split it into subsets using each feature



Keep splitting (if permitted)



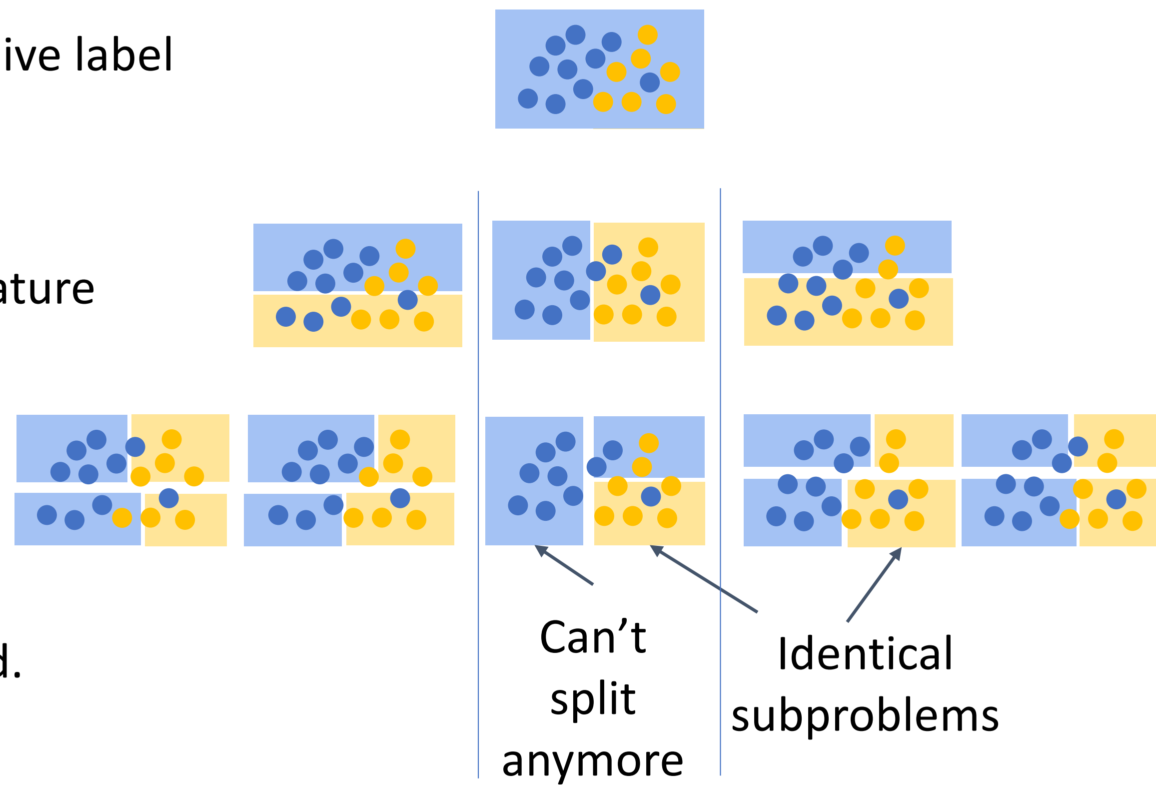
Dynamic programming / Branch and Bound

Start with the full dataset and a naive label

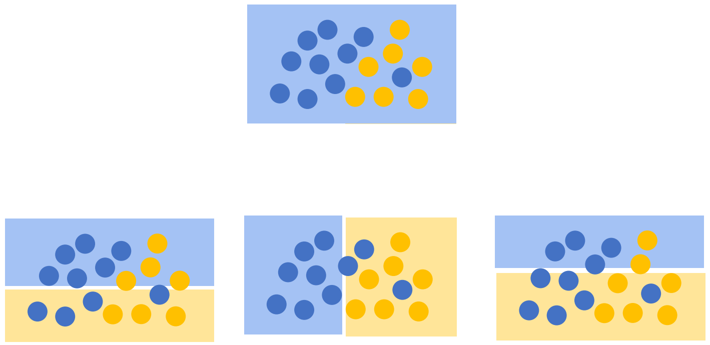
Split it into subsets using each feature

Keep splitting (if permitted)

Consolidate any duplication found.

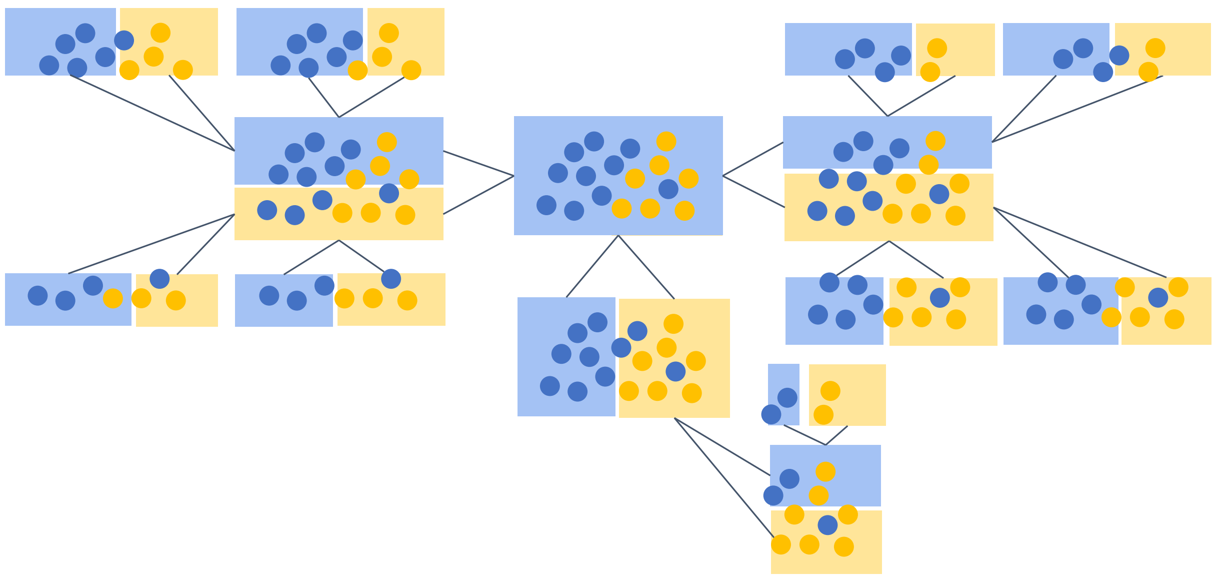


Dynamic programming / Branch and Bound



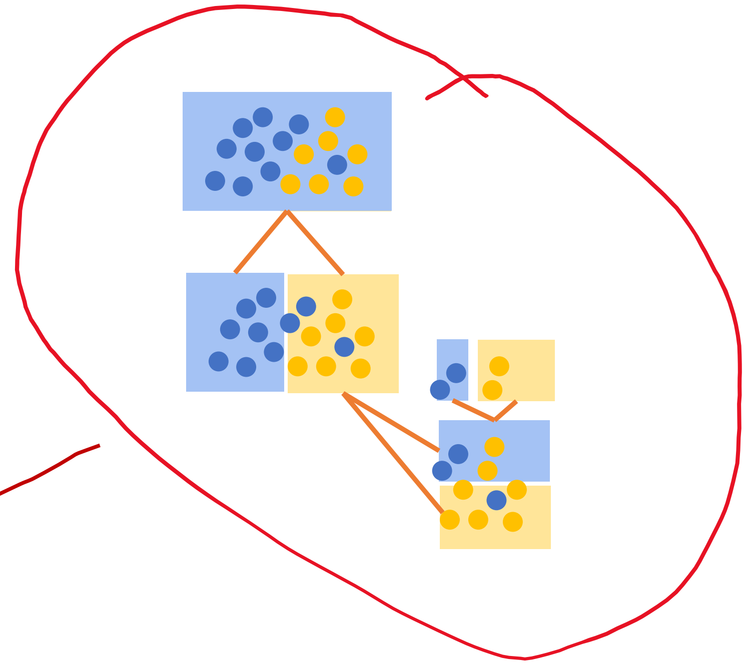
Dynamic programming / Branch and Bound

The solution to each subproblem yields the best feature to split on.



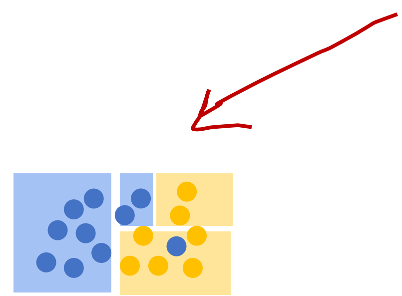
Dynamic programming / Branch and Bound

The solution to each subproblem yields the best feature to split on.



The optimal solution is found after all subproblems are "completed"

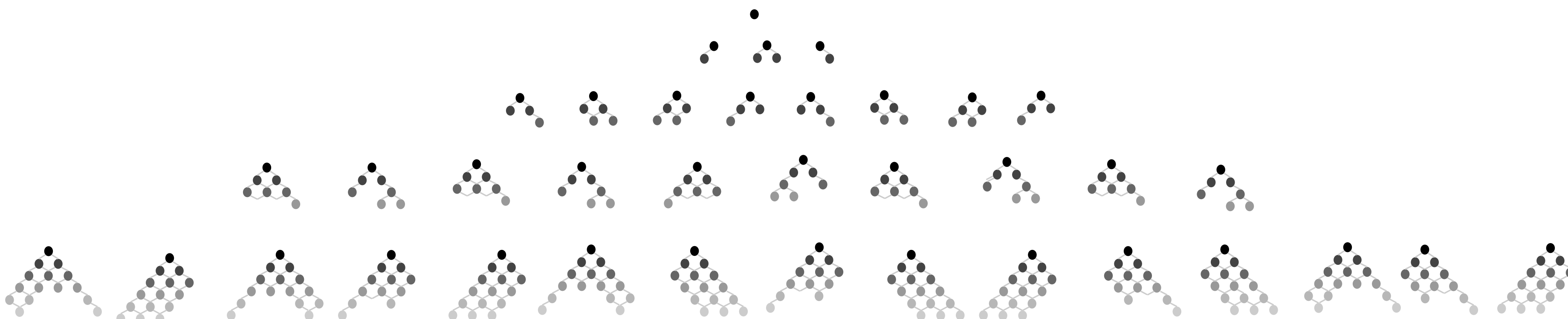
Some subproblems can be proven to yield non-optimal solutions



Dynamic programming / Branch and Bound

Analytical Bounds Reduce the Search Space

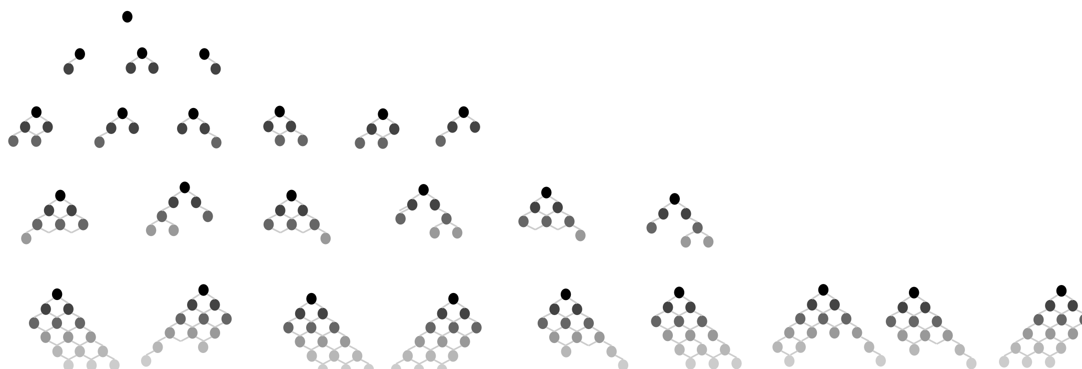
Theorems show that some partial trees can never be extended to form optimal trees.



Dynamic programming / Branch and Bound

Analytical Bounds Reduce the Search Space

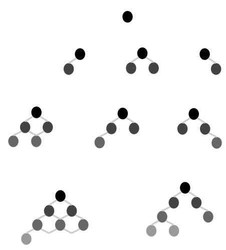
Theorems show that some partial trees can never be extended to form optimal trees.



Dynamic programming / Branch and Bound

Analytical Bounds Reduce the Search Space

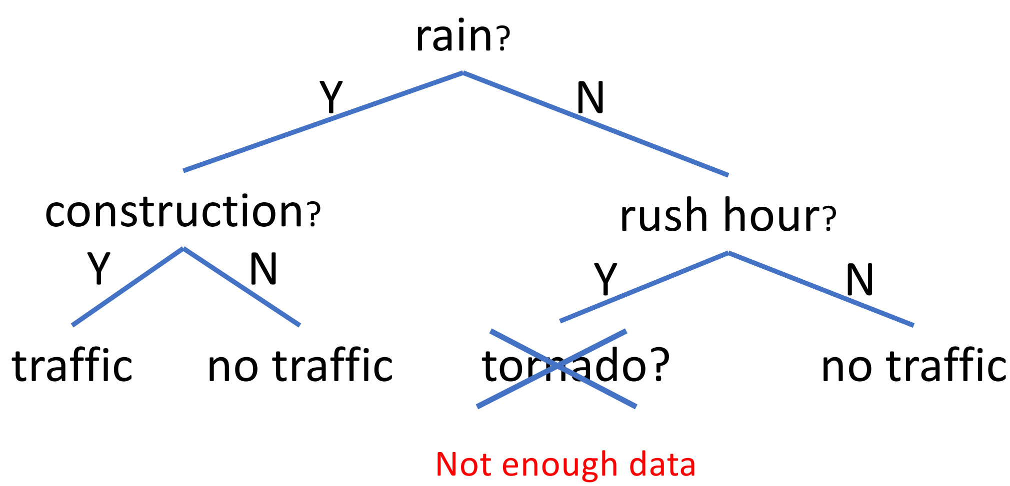
Theorems show that some partial trees can never be extended to form optimal trees.



GOSDT - Generalized and Scalable Optimal Sparse Decision Trees
(Lin et al., ICML 2020)

Analytical Bounds Reduce the Search Space

Theorems show that some partial trees can never be extended to form optimal trees.



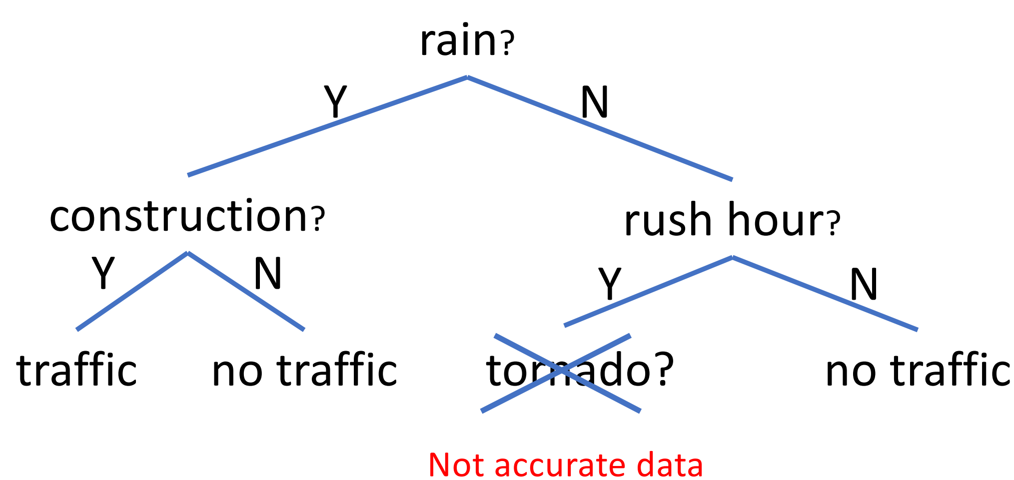
"Theorem":

If the amount of data traveling through an internal node is < 2C (where C is the regularization parameter), the tree cannot achieve the minimum of the objective.

GOSDT - Generalized and Scalable Optimal Sparse Decision Trees
(Lin et al., ICML 2020)

Analytical Bounds Reduce the Search Space

Theorems show that some partial trees can never be extended to form optimal trees.



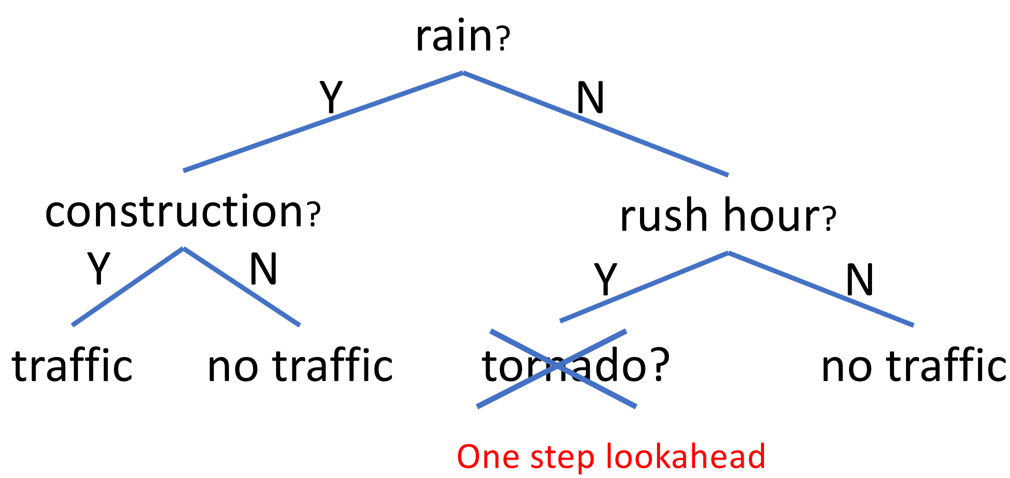
"Theorem":

If a proposed split leads to < C correctly classified data going to either side of the split, then this split can be excluded, and we can exclude that feature anywhere further down the tree extending that leaf.

GOSDT - Generalized and Scalable Optimal Sparse Decision Trees
(Lin et al., ICML 2020)

Analytical Bounds Reduce the Search Space

Theorems show that some partial trees can never be extended to form optimal trees.



"Theorem": Consider a tree with lower bound $b \leq R_{\text{current best}}$.
If $b + C \geq R_{\text{current best}}$, we can prune all of its child trees.

GOSDT - Generalized and Scalable Optimal Sparse Decision Trees
(Lin et al., ICML 2020)

Represent a tree by its leaves

rain & construction & traffic

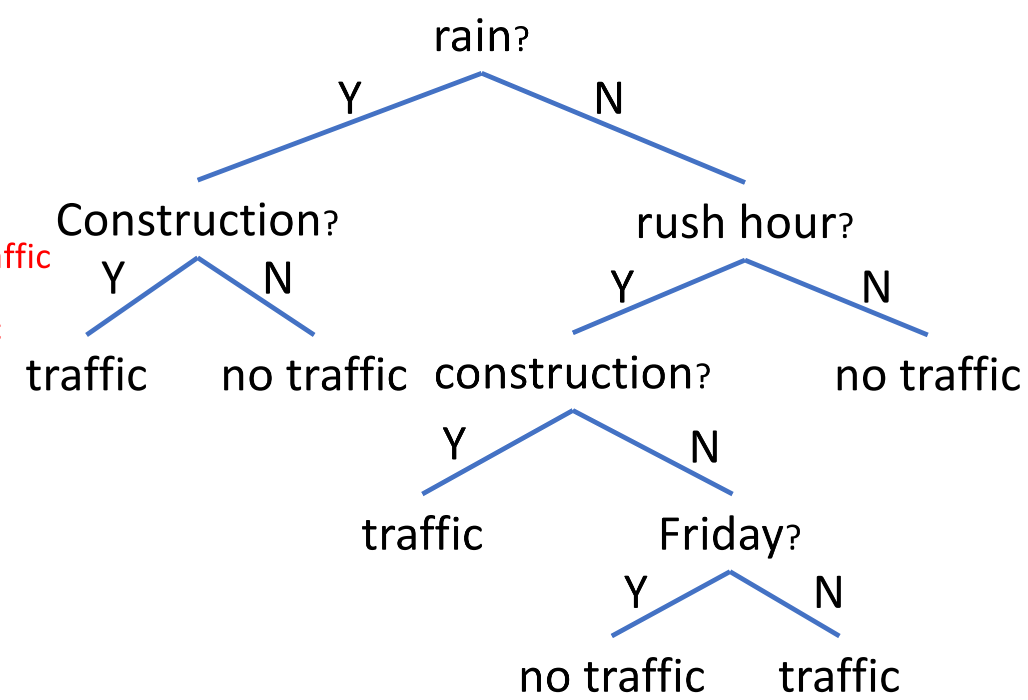
rain & no construction & no traffic

no rain & rush hour & construction & traffic

no rain & rush hour & no construction & Friday and no traffic

no rain & rush hour & no construction & Friday and traffic

no rain & no rush hour & no traffic



GOSDT - Generalized and Scalable Optimal Sparse Decision Trees
(Lin et al., ICML 2020)

Permutation map: Discover identical trees already evaluated

rain & construction & traffic

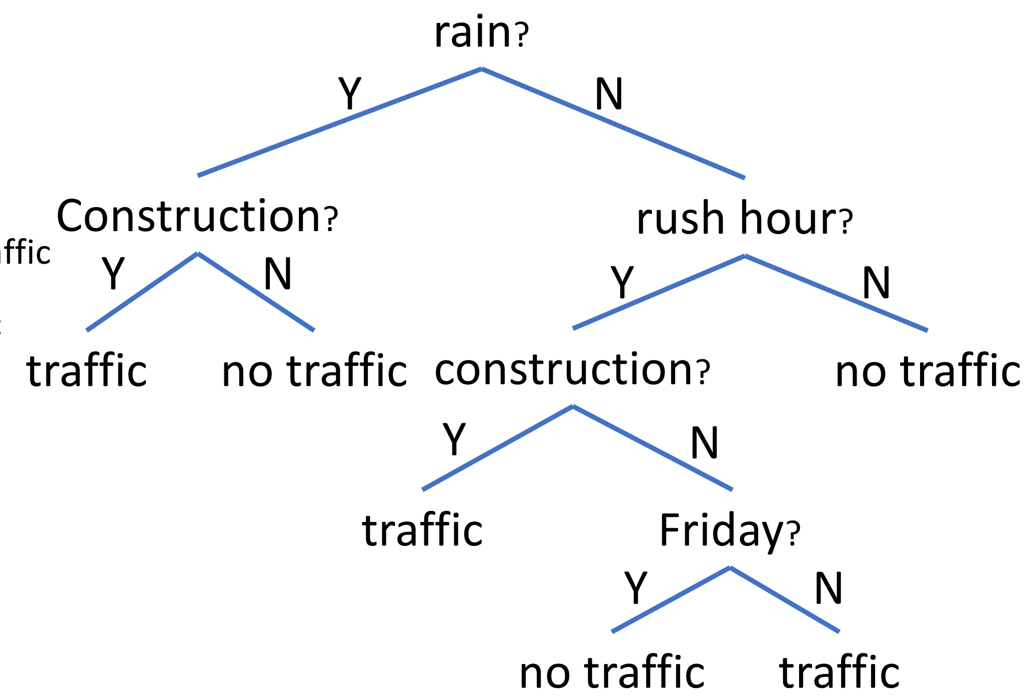
rain & no construction & no traffic

no rain & rush hour & construction & traffic

no rain & rush hour & no construction & Friday and no traffic

no rain & rush hour & no construction & Friday and traffic

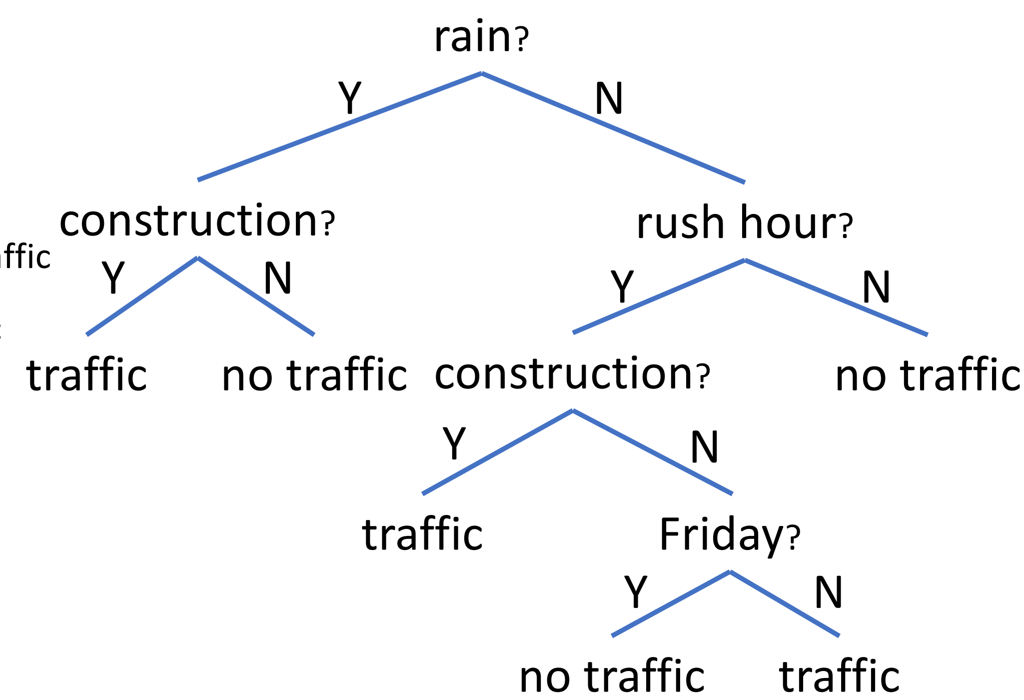
no rain & no rush hour & no traffic



GOSDT - Generalized and Scalable Optimal Sparse Decision Trees
(Lin et al., ICML 2020)

Bit-vectors describe data represented by each leaf

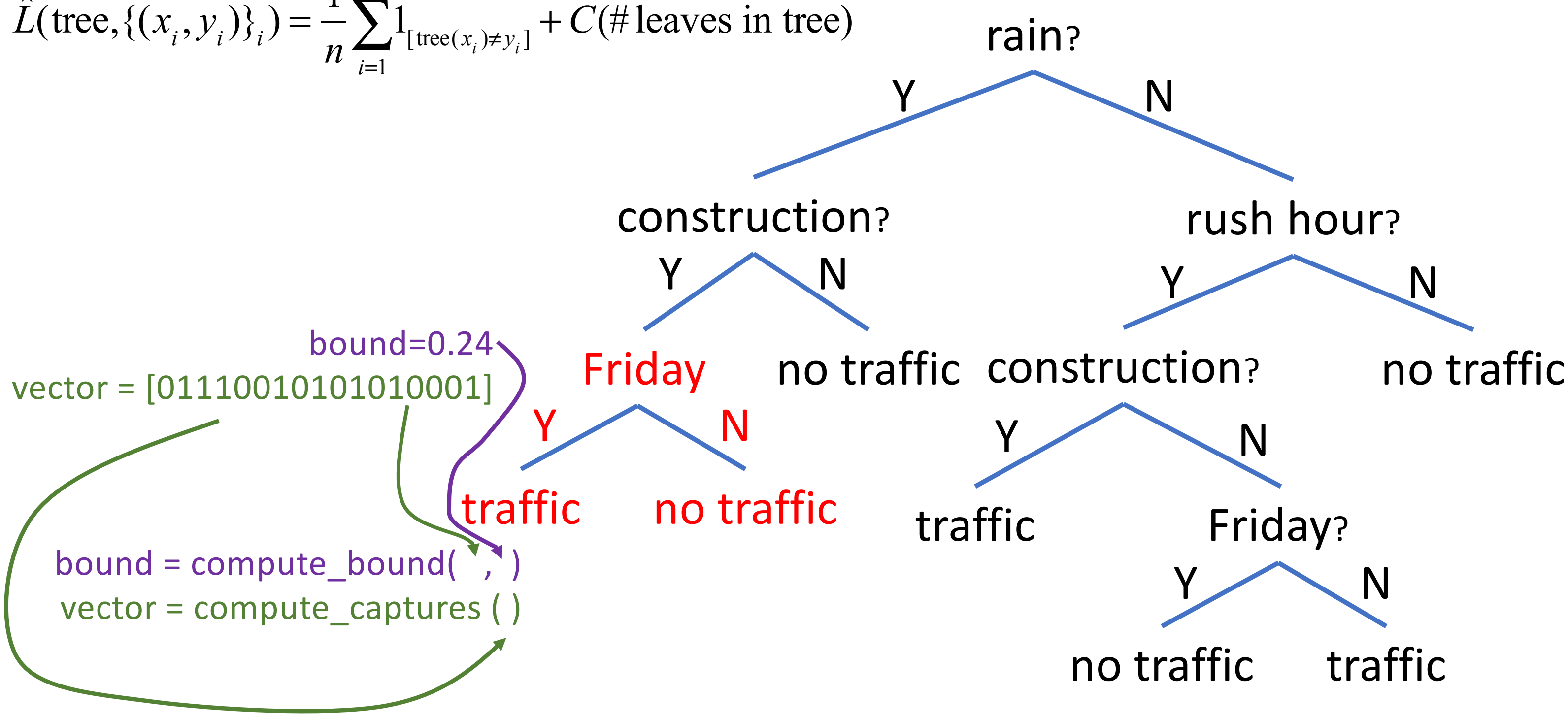
rain & construction & traffic
[10000100010011100000..........................0]
rain & no construction & no traffic
[01100010000000000110..........................1]
no rain & rush hour & construction & traffic
[00010001000000001000..........................0]
no rain & rush hour & no construction & Friday and no traffic
[00001000000000000001..........................0]
no rain & rush hour & no construction & Friday and traffic
[00000000100000000000..........................0]
no rain & no rush hour & no traffic
[00000000000110000000..........................0]



GOSDT - Generalized and Scalable Optimal Sparse Decision Trees
(Lin et al., ICML 2020)

Incremental computation of objective and bounds

$$\hat{L}(\text{tree},\{(x_i,y_i)\}_i) = \frac{1}{n}\sum_{i=1}^{n} 1_{[\text{tree}(x_i)\neq y_i]} + C(\#\,\text{leaves in tree})$$



GOSDT - Generalized and Scalable Optimal Sparse Decision Trees
(Lin et al., ICML 2020)

Strong analytical bounds

Leaf-based representation

Permutation map = Fast Implementation

Caching of intermediate results

Incremental computation

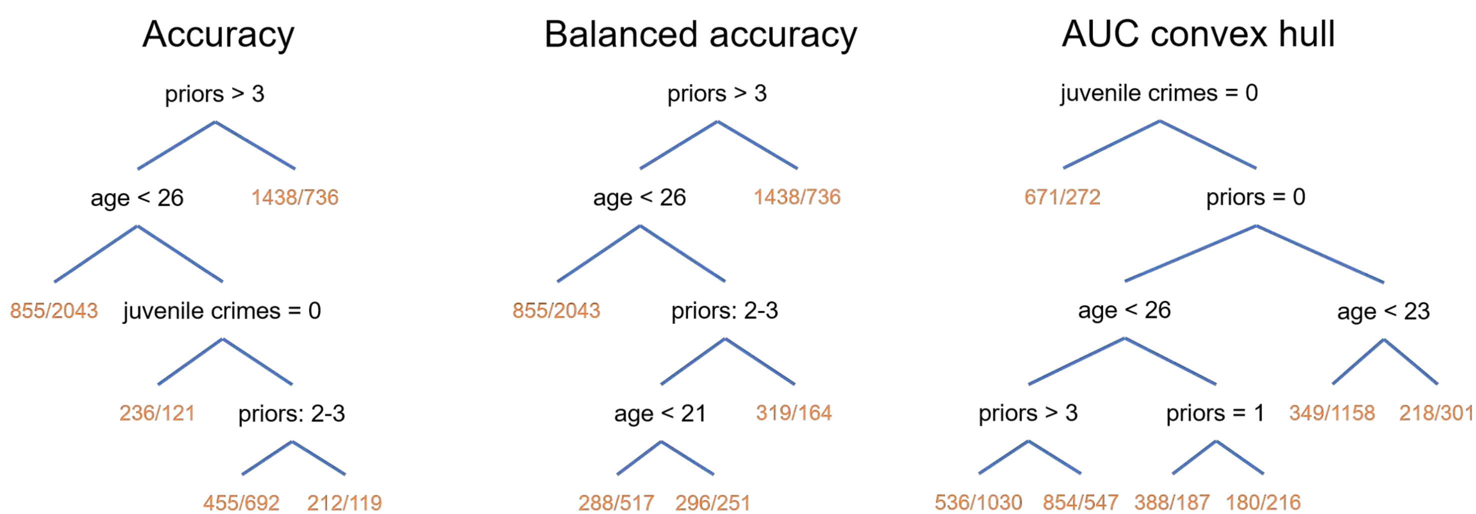
Consolidation of repeated subproblems

GOSDT - Generalized and Scalable Optimal Sparse Decision Trees
(Lin et al., ICML 2020)

$$\min_{\text{tree}} \hat{L}(\text{tree}, \{(x_i, y_i)\}_i) \text{ where}$$

$$\hat{L}(\text{tree}, \{(x_i, y_i)\}_i) = \ell(\text{tree}, \{(x_i, y_i)\}_i) + \lambda(\#\text{ leaves in tree})$$

- Can optimize any loss function monotonically increasing in FP and FN (Balanced accuracy, weighted accuracy, F-1, precision, …)
- Can optimize rank statistics (AUC and partial AUC under the ROC convex hull)



GOSDT - Generalized and Scalable Optimal Sparse Decision Trees
(Lin et al., ICML 2020)

$$\min_{\text{tree}} \hat{L}(\text{tree}, \{(x_i, y_i)\}_i) \text{ where}$$
$$\hat{L}(\text{tree}, \{(x_i, y_i)\}_i) = \ell(\text{tree}, \{(x_i, y_i)\}_i) + \lambda(\text{\# leaves in tree})$$

- Can optimize any loss function monotonically increasing in FP and FN (Balanced accuracy, weighted accuracy, F-1, precision, …)
- Can optimize rank statistics (AUC and partial AUC under the ROC convex hull)

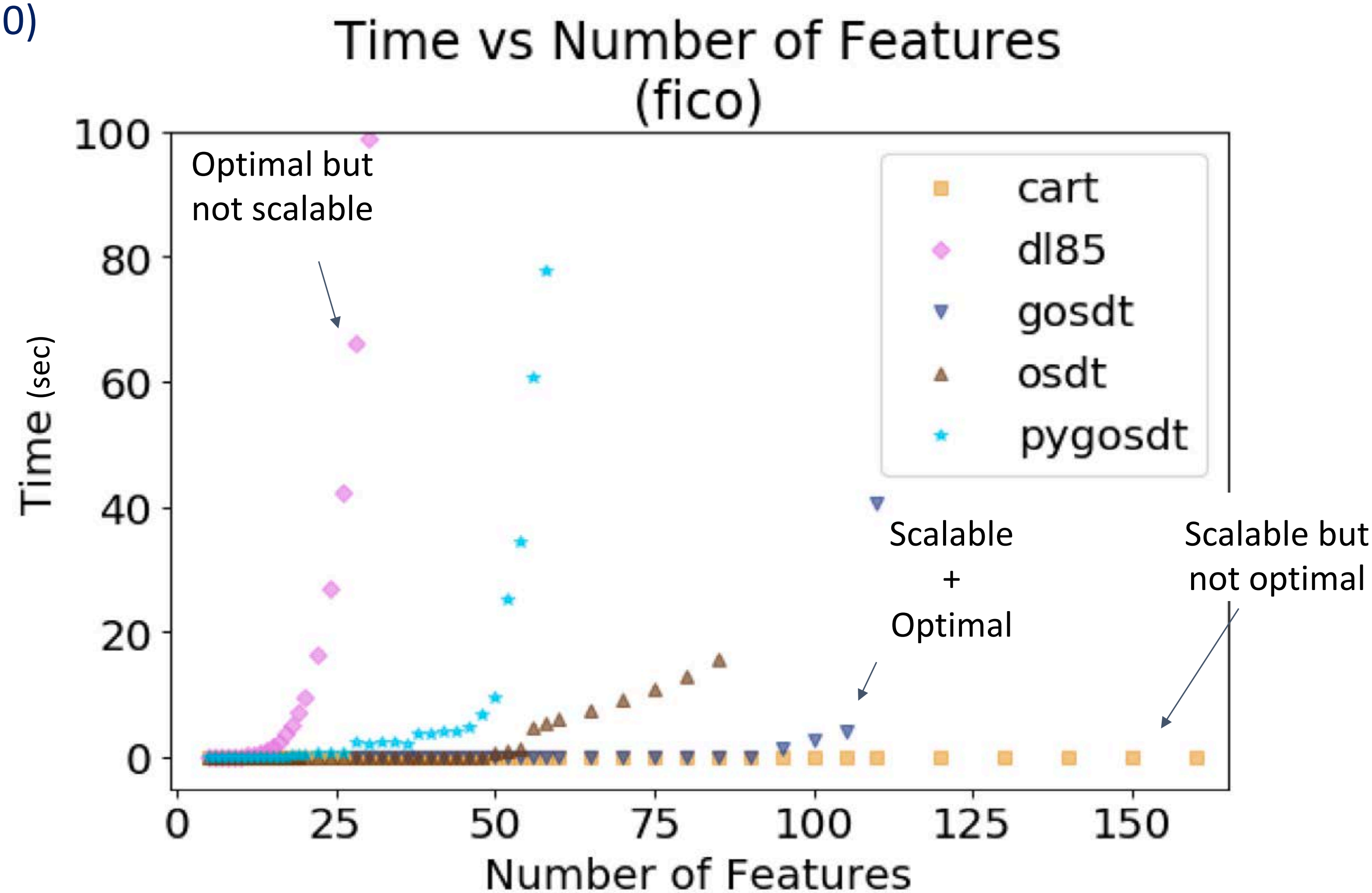
Main experimental results:

- Similar classification error to black box methods.
- For custom losses, much better loss values than greedy decision trees.
- Sparser than all heuristic methods
- Orders of magnitude faster than the next best method.

GOSDT - Generalized and Scalable Optimal Sparse Decision Trees
(Lin et al., ICML 2020)

Scalability

Improvements in orders of magnitude

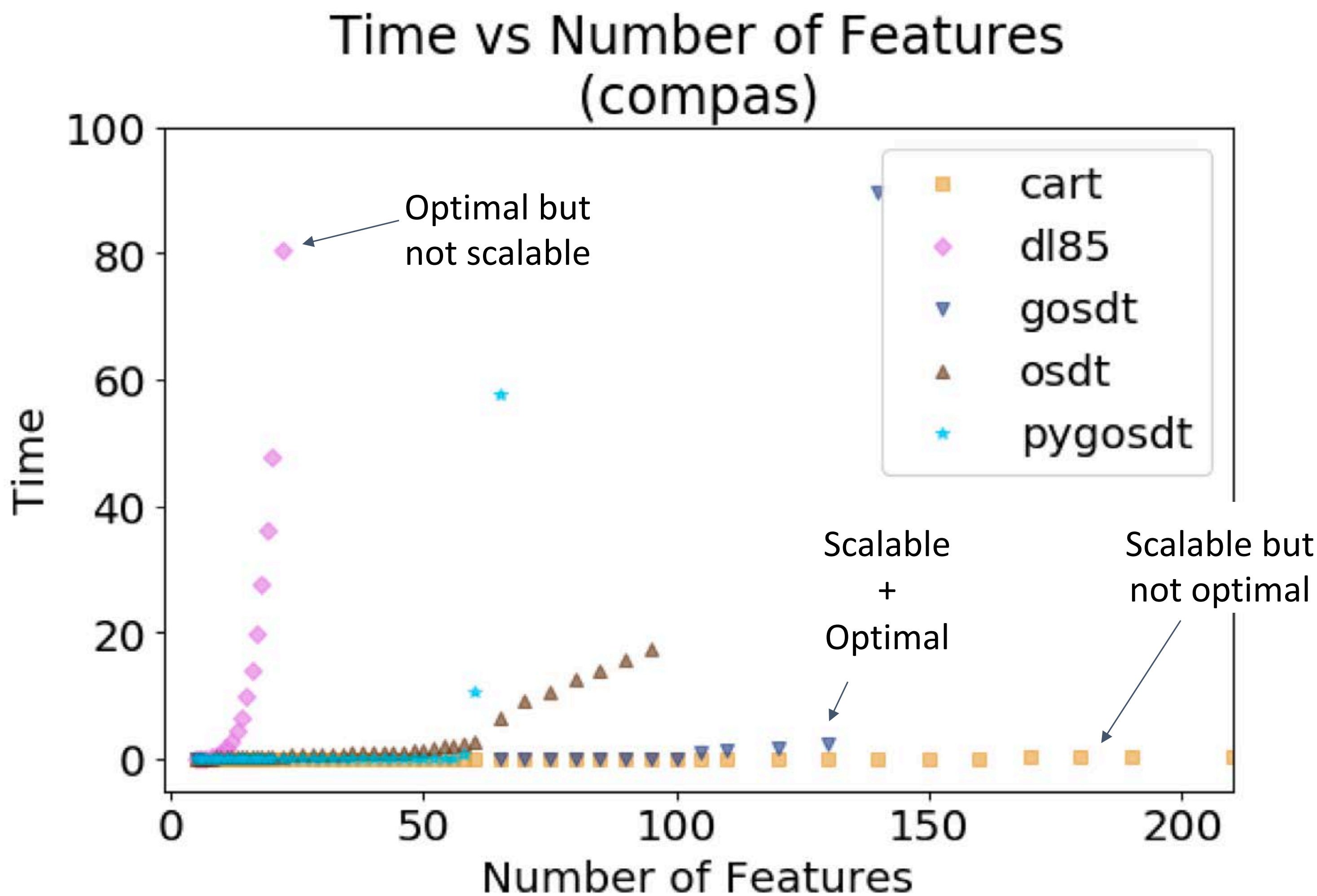


Note: BinOCT too slow to include.

GOSDT - Generalized and Scalable Optimal Sparse Decision Trees
(Lin et al., ICML 2020)

Scalability

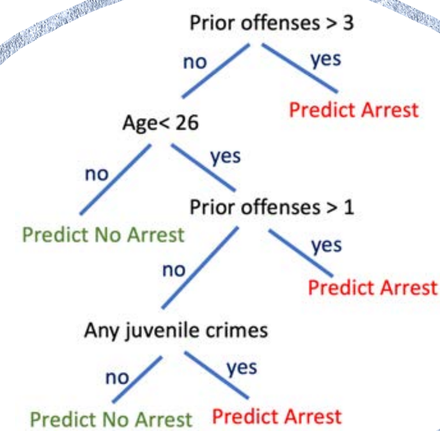
Improvements in orders of magnitude



Note: BinOCT too slow to include.

In this talk

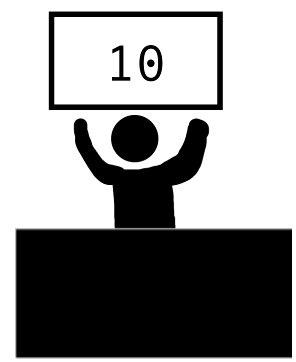
- Optimal decision trees
- Scoring systems



1.	Any cEEG Pattern with Frequency **2 H**z	1 point	···
2.	**E**pileptiform Discharges	1 point	+ ···
3.	Patterns include [**L**PD, LRDA, BIPD]	1 point	+ ···
4.	**P**atterns Superimposed with Fast or Sharp Activity	1 point	+ ···
5.	Prior **S**eizure	1 point	+ ···
6.	**B**rief Rhythmic Discharges	**2** points	+ ···
		SCORE	= ···

SCORE	0	1	2	3	4	5	6+
RISK	<5%	11.9%	26.9%	50.0%	73.1%	88.1%	95.3%

Scoring systems



1 point if person has **social type** with **below average** parole violation rate

Social Type	Violation Rate
All persons	26.5%
Ne'er-do-well	25.6
Mean citizen	30.0
Drunkard	38.9
Gangster	23.2
Recent immigrant	16.7
Farm boy	10.2
Drug addict	66.7

total score over all 21 significant factors predicts **success at parole**

Points for Number of Factors	Per Cent Non-violators of Parole
16-21	98.5
14-15	97.8
13	91.2
12	84.9
11	77.3
10	65.9
7-9	56.1
5-6	32.9
2-4	24.0

Burgess. Factors determining success or failure on parole. 1928

FACTOR	Score *
Gender	
Female	0
Male	1
Age	
Less than 24	3
24-29	2
30-49	1
50+	0
County	
Rural counties	0
Smaller, urban count	1
Allegheny and Philadelphia Counties	2
Total number of prior arrests	
0	0
1	1
2 to 4	2
5 to 12	3
13+	4
Prior property arrests	
No	0
Yes	1
Prior drug arrests	
No	0
Yes	1
Property offender	
No	0
Yes	1
Offense gravity score (OGS)	
4+	0

Risk score	N	% Arrested
0	3	0.0
1	47	17.0
2	181	9.9
3	436	23.6
4	737	24.8
5	1,036	32.4
6	1,067	40.7
7	1,434	47.2
8	1,934	55.5
9	2,103	62.3
10	1,829	69.9
11	1,098	72.2
12	278	79.1
13	25	80.0
14	3	66.7

Pennsylvania Commission on Sentencing, 2013

1. Lived with both biological parents to age 16 (except for death of parent):
Yes .... -2
No .... +3
Evidence:

2. Elementary School Maladjustment:
No Problems.. -1
Slight (Minor discipline or attendance)
or Moderate Problems............................ +2
Severe Problems (Frequent disruptive behavior and/or attendance or behavior resulting in expulsion or serious suspensions) .... +5
(Same as CATS Item)

3. History of alcohol problems *(Check if present):*
˜ Parental Alcoholism ˜ Teenage Alcohol Problem
˜ Adult Alcohol Problem ˜ Alcohol involved in prior offense
˜ Alcohol involved in index offense
No boxes checked................................... -1
1 or 2 boxes checked 0
3 boxes checked .... +1
4 or 5 boxes checked .... +2
Evidence:

4. Marital status (at the time of or prior to index offense):
Ever married (or lived common law in the same home for at least six months) .... -2
Never married... +1
Evidence:

5. Criminal history score for nonviolent offenses prior to the index offense
Score 0 .... -2
Score 1 or 2.. 0
Score 3 or above .... +3
(from the Cormier-Lang system, see below)

6. Failure on prior conditional release (includes parole or probation violation or revocation, failure to comply, bail violation, and any new arrest while on conditional release):
No...0
Yes .... +3
Evidence:

7. Age at index offense
Enter Date of Index Offense: ___/___/_____
Enter Date of Birth: ___/___/_____
Subtract to get Age:
39 or over .... -5
34 - 38 .... -2
28 - 33 .... -1
27 0
26 or less.. +2

8. Victim Injury (for index offense; the most serious is scored):
Death... -2
Hospitalized...0
Treated and released.............................. +1
None or slight (includes no victim)........... +2
Note: admission for the gathering of forensic evidence only is NOT considered as either treated or hospitalized; ratings should be made based on the degree of injury.
Evidence:

9. Any female victim (for index offense)
Yes .... -1
No (includes no victim)............................ +1
Evidence:

10. Meets DSM criteria for any personality disorder (must be made by appropriately licensed or certified professional)
No... -2
Yes .... +3
Evidence:

11. Meets DSM criteria for schizophrenia (must be made by appropriately licensed or certified professional)
Yes .... -3
No .... +1
Evidence:

12. a. Psychopathy Checklist score (if available, otherwise use item 12.b. CATS score)........
4 or under .... -3
5 – 9... -3
10-14 .... -1
15-24 0
25-34 .... +4
35 or higher .... +12
Note: If there are two or more PCL scores, average the scores.
Evidence:

12. b. CATS score (from the CATS worksheet)
0 or 1 .... -3
2 or 3 0
4 ....+2
5 or higher .... +3

12. WEIGHT (Use the highest circled weight from 12 a. or 12 b.) .... _____

TOTAL VRAG SCORE (SUM CIRCLED SCORES FOR ITEMS 1 – 11 PLUS THE WEIGHT FOR ITEM 12): _________

VRAG Score	Category of Risk
-24	Low
-23	Low
-22	Low
-20	Low
-19	Low
-18	Low
-17	Low
-16	Low
-15	Low
-14	Low
-13	Low
-12	Low
-11	Low
-10	Low
-9	Low
-8	Low
-7	Medium
-6	Medium
-5	Medium
-4	Medium
-3	Medium
-2	Medium
-1	Medium
0	Medium
1	Medium
2	Medium
3	Medium
4	Medium
5	Medium
6	Medium
7	Medium
8	Medium
9	Medium
10	Medium
11	Medium
2	Medium
13	Medium
14	High
15	High
16	High
17	High
18	High
19	High
20	High
21	High
22	High
23	High
24	High
25	High
26	High
28	High
32	High

Violence Risk Appraisal Guide (Quinsey et al, 2006)

Calculators

By Category Alphabetically

Addiction Medicine

Anesthesiology

Cardiac Surgery

Cardiology

COVID-19

Critical Care

Emergency

- Intracerebral Hemorrhage
- Ischemic Stroke
- Movement Disorder
- Multiple Sclerosis & Demyelinating Disease
- Neurophysiology
- Seizure
  - 2HELPS2B Score
  - Phenytoin Adjustment in Renal Failure
  - Seizure vs Syncope
- Subarachnoid Hemorrhage

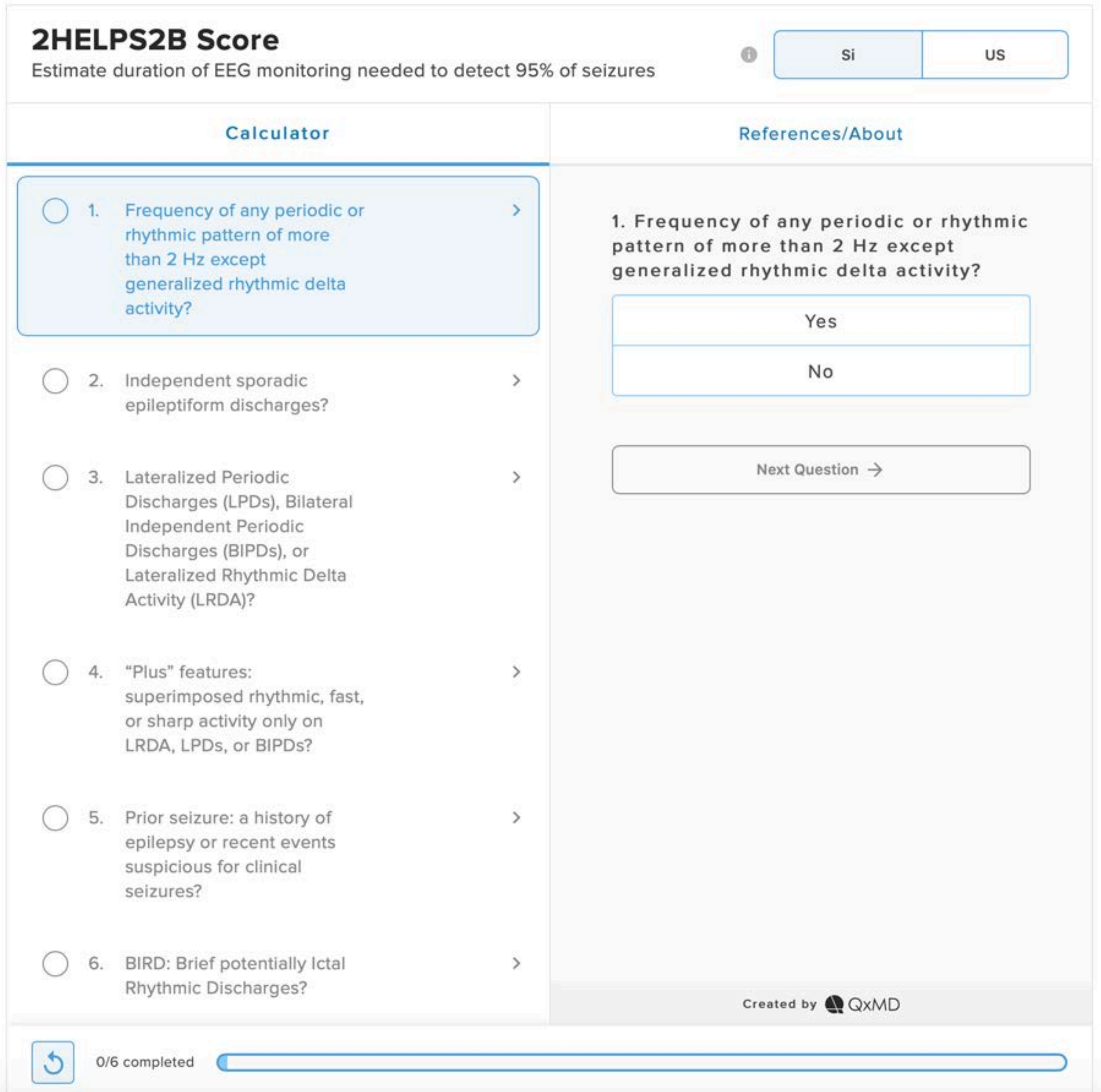
Obstetrics & Gynecology

Oncology

Orthopedics

Otolaryngology (ENT)

Pathology & Lab Medicine


2HELPS2B Score
Estimate duration of EEG monitoring needed to detect 95% of seizures
SI
US
Calculator
References/About
1. Frequency of any periodic or rhythmic pattern of more than 2 Hz except generalized rhythmic delta activity?
2. Independent sporadic epileptiform discharges?
3. Lateralized Periodic Discharges (LPDs), Bilateral Independent Periodic Discharges (BIPDs), or Lateralized Rhythmic Delta Activity (LRDA)?
4. "Plus" features: superimposed rhythmic, fast, or sharp activity only on LRDA, LPDs, or BIPDs?
5. Prior seizure: a history of epilepsy or recent events suspicious for clinical seizures?
6. BIRD: Brief potentially Ictal Rhythmic Discharges?
1. Frequency of any periodic or rhythmic pattern of more than 2 Hz except generalized rhythmic delta activity?
Yes
No
Next Question
Created by QxMD
0/6 completed

Key challenges:

- Constraints (e.g., FP<20%, fairness, etc.)
- Integrality

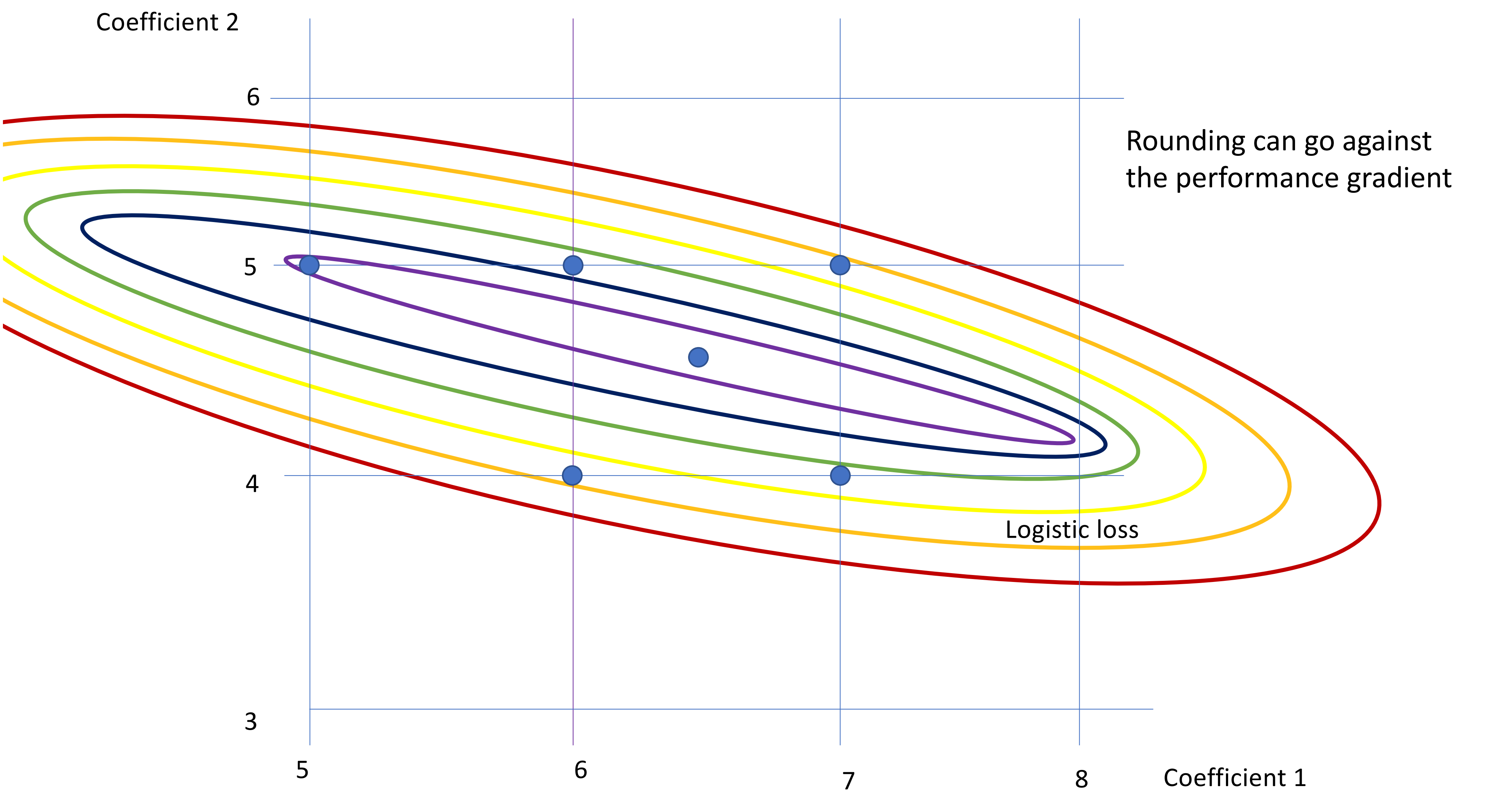
1. *Congestive Heart Failure*	1 point		$\cdots$
2. *Hypertension*	1 point	+	$\cdots$
3. *Age* ≥ 75	1 point	+	$\cdots$
4. *Diabetes Mellitus*	1 point	+	$\cdots$
5. *Prior Stroke or Transient Ischemic Attack*	2 points	+	$\cdots$
ADD POINTS FROM ROWS 1–5	**SCORE**	=	$\cdots$

SCORE	0	1	2	3	4	5	6
STROKE RISK	1.9%	2.8%	4.0%	5.9%	8.5%	12.5%	18.2%

Typical approach:

(Gage et al., 2001), CHADS2 score for stroke prediction: panel of experts

(Antman et al., 2000), TIMI risk score for unstable angina/non-ST elevation MI: preliminary feature selection, followed by logistic regression with the chosen features, scaling, and rounding



Coefficient 2
6
5
4
3
5
6
7
8
Coefficient 1
Rounding can go against
the performance gradient
Logistic loss

Elastic Net

$$\textbf{SCORE} = 1.42 \text{ Rhythmic Patterns Include [BiPD, LRDA, LPD]} + 0.31 \text{ Prior Seizure} + 0.21 \text{ Epileptiform Discharges} + 0.26 \text{ Patterns Superimposed with Fast or Sharp Activity} + 0.25 \text{ Brief Rhythmic Discharges} - 2.54$$

Elastic Net + Rounding

SCORE	=	1	Rhythmic Patterns Include [BiPD, LRDA, LPD]
		+ 0	~~Prior Seizure~~
		+ 0	~~Epileptiform Discharges~~
		+ 0	~~Patterns Superimposed with Fast or Sharp Activity~~
		+ 0	~~Brief Rhythmic Discharges~~
		− 3	

Elastic Net

$$
\begin{aligned}
\textbf{SCORE} = \; & 1.42 \quad \text{Rhythmic Patterns Include [BiPD, LRDA, LPD]} \\
& + 0.31 \quad \text{Prior Seizure} \\
& + 0.21 \quad \text{Epileptiform Discharges} \\
& + 0.26 \quad \text{Patterns Superimposed with Fast or Sharp Activity} \\
& + 0.25 \quad \text{Brief Rhythmic Discharges} \\
& - 2.54
\end{aligned}
$$

Elastic Net + Scaling + Rounding

SCORE	=	6	Rhythmic Patterns Include [BiPD, LRDA, LPD]
		+ 1	Prior Seizure
		+ 1	Epileptiform Discharges
		+ 1	Patterns Superimposed with Fast or Sharp Activity
		+ 1	Brief Rhythmic Discharges
		− 10	

Elastic Net

SCORE	=	1.42	Rhythmic Patterns Include [BiPD, LRDA, LPD]
		+ 0.31	Prior Seizure
		+ 0.21	Epileptiform Discharges
		+ 0.26	Patterns Superimposed with Fast or Sharp Activity
		+ 0.25	Brief Rhythmic Discharges
		− 2.54	

RiskSLIM model (optimized)

1.	BriefRhythmicDischarge	2 points		...
2.	PatternsInclude LPD	2 points	+	...
3.	PriorSeizure	1 point	+	...
4.	EpiletiformDischarge	1 point	+	
		SCORE	=	

SCORE	0	1	2	3	4	5	6
RISK	4.7%	11.9%	26.9%	50.0%	73.1%	88.1%	95.3%

(This one is better calibrated and has large AUC.)

Ustun & R, Optimized Risk Scores, JMLR 2019

Risk-Calibrated Supersparse Linear Integer Models (Risk-SLIM)

(Ustun, R, 2019)

$$\min_{\lambda \in L} \overbrace{\sum_{i=1}^{n} \log\left(1+e^{-y_i \boldsymbol{x}_i^\top \lambda}\right)}^{\text{Logistic Loss}} + \overbrace{C\|\lambda\|_0}^{\text{Model Size}}$$

MINLP – really hard...

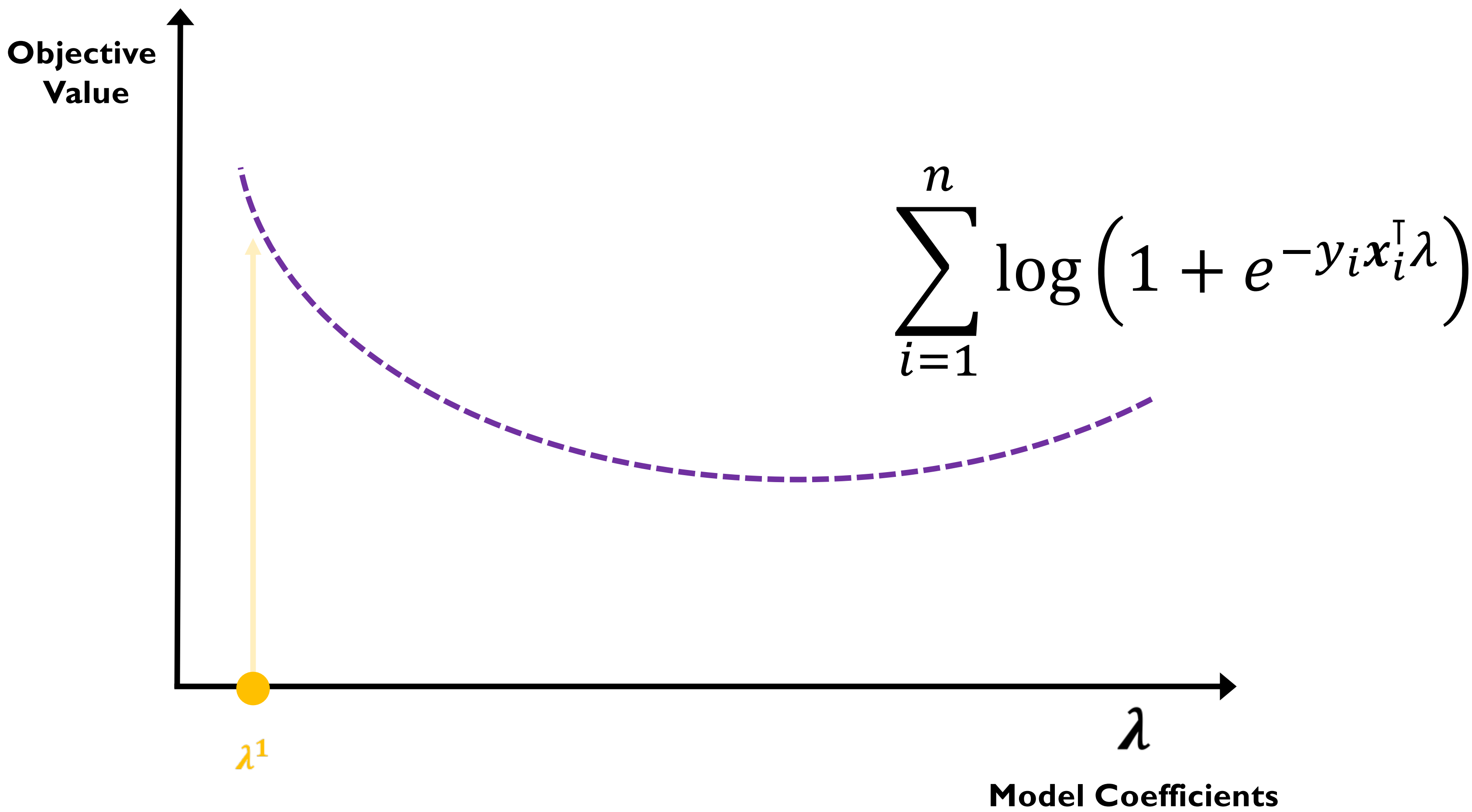
$\lambda \in L$ means that $\forall j,\ \lambda_j \in \underbrace{\{-10,-9,\ldots,0,\ldots,9,10\}}_{\text{Small Integer Coefficients}}$

(optional: additional constraints)

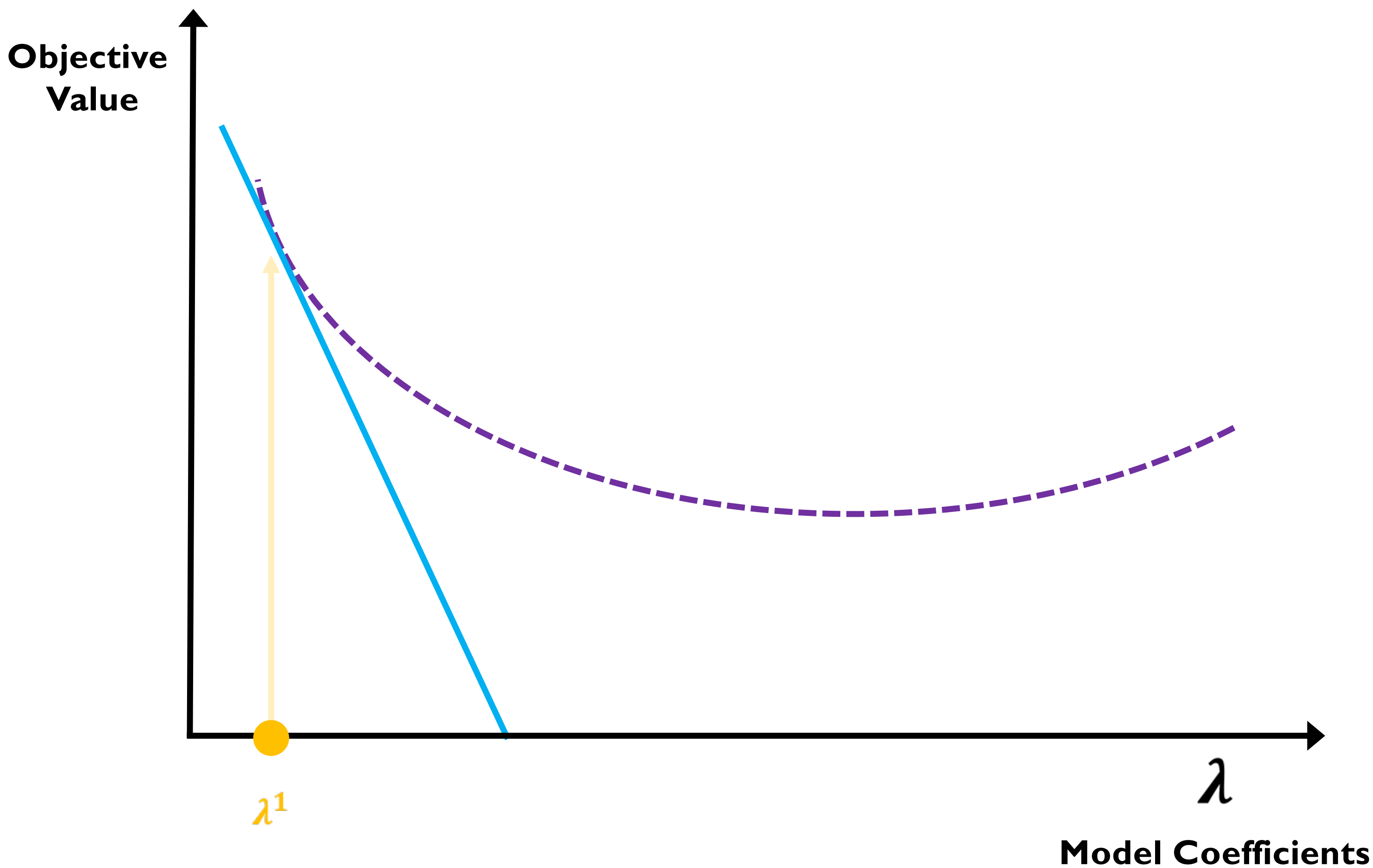
Cutting Planes (Traditional)

$$\min_{\lambda} \sum_{i=1}^{n} \log\left(1 + e^{-y_i \boldsymbol{x}_i^\intercal \lambda}\right)$$

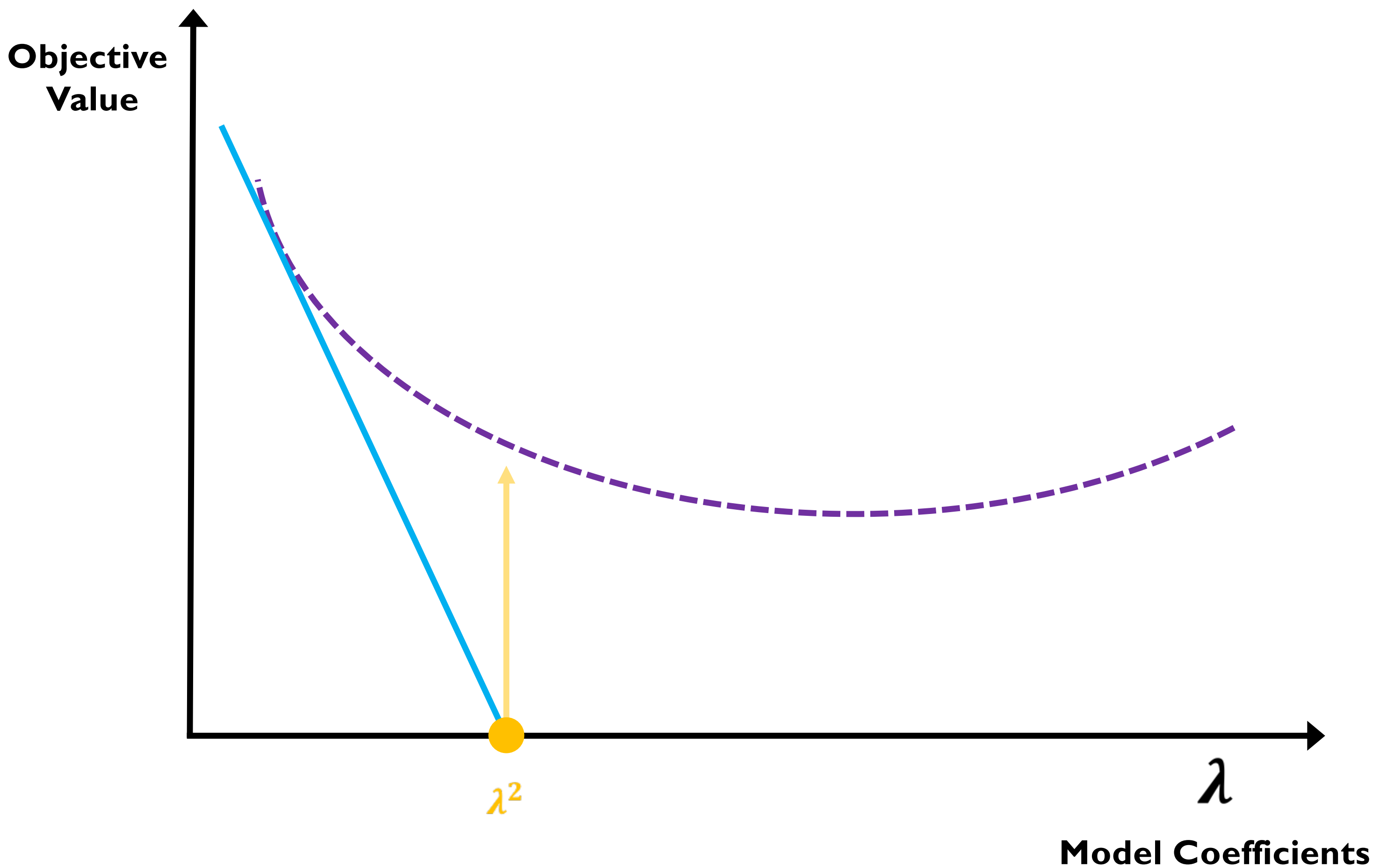
Traditional cutting planes



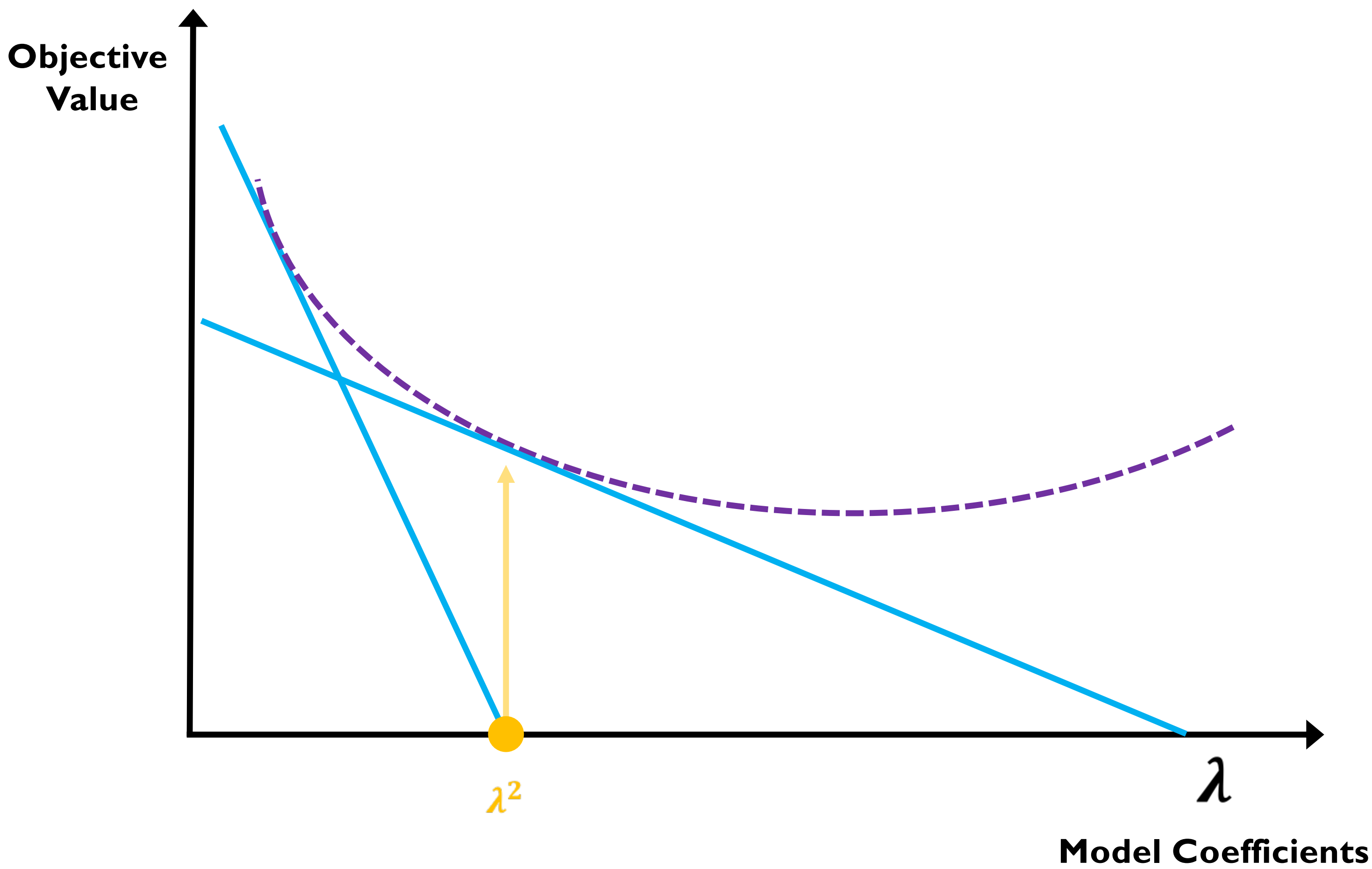
Traditional cutting planes



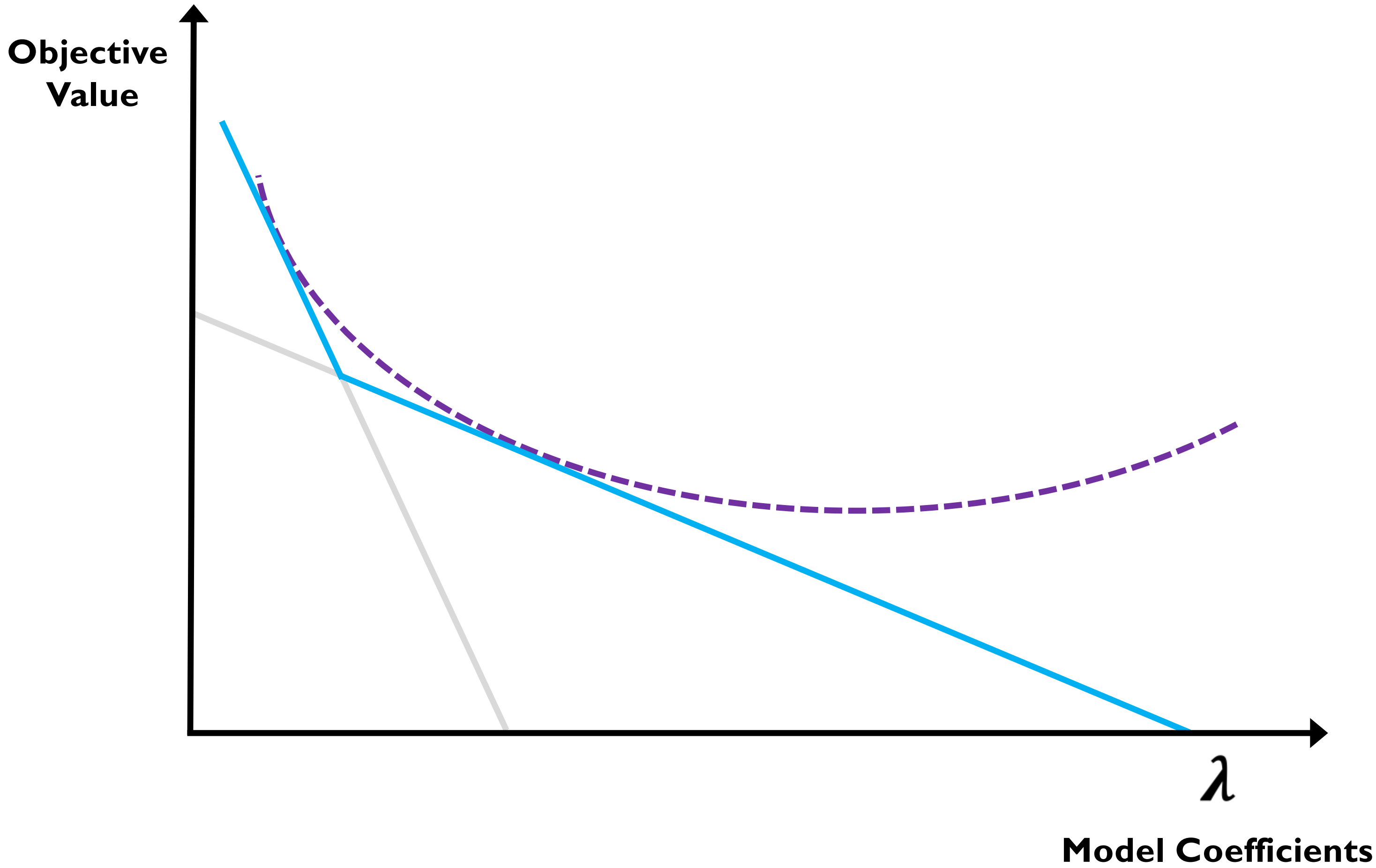
Traditional cutting planes



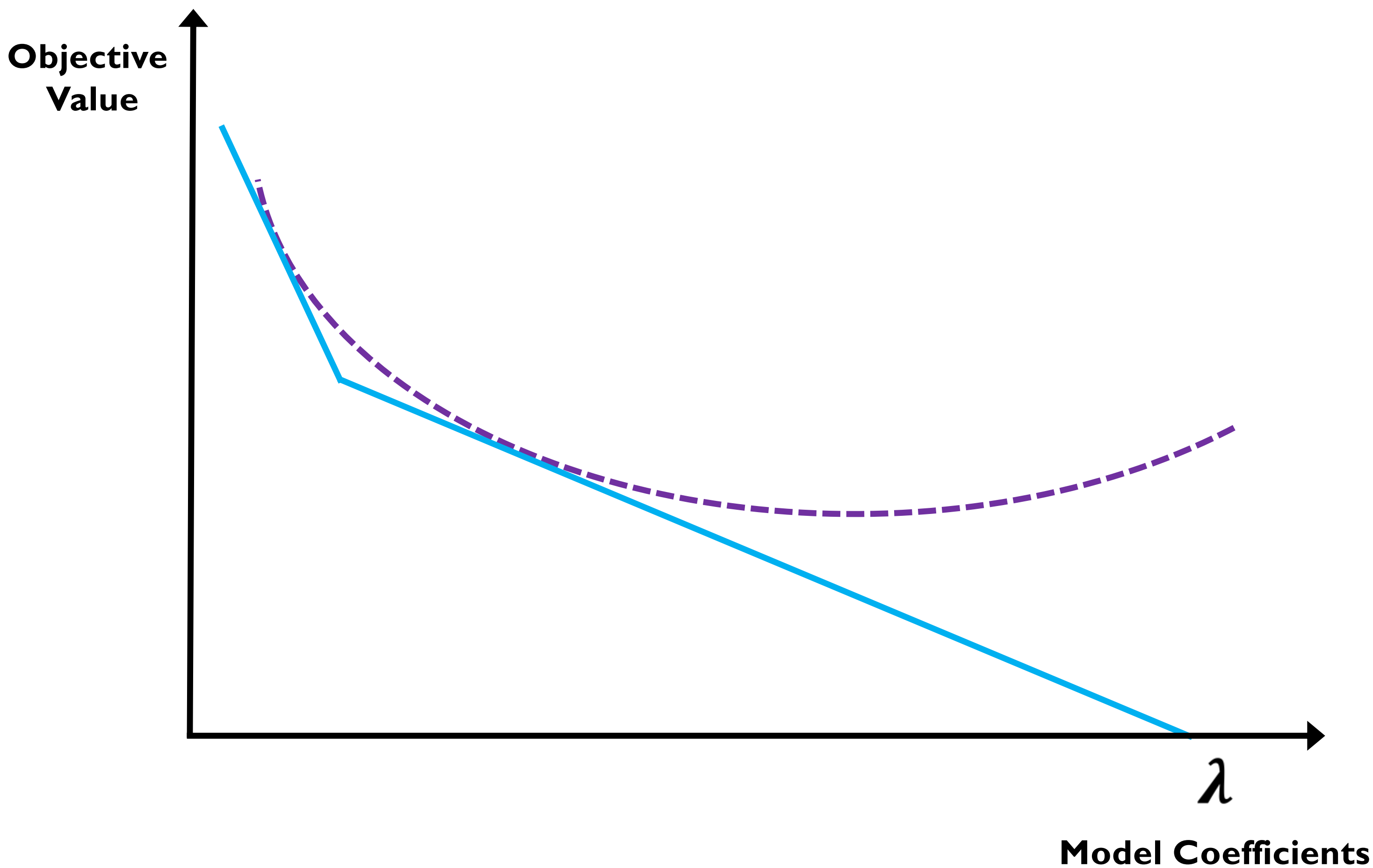
Traditional cutting planes



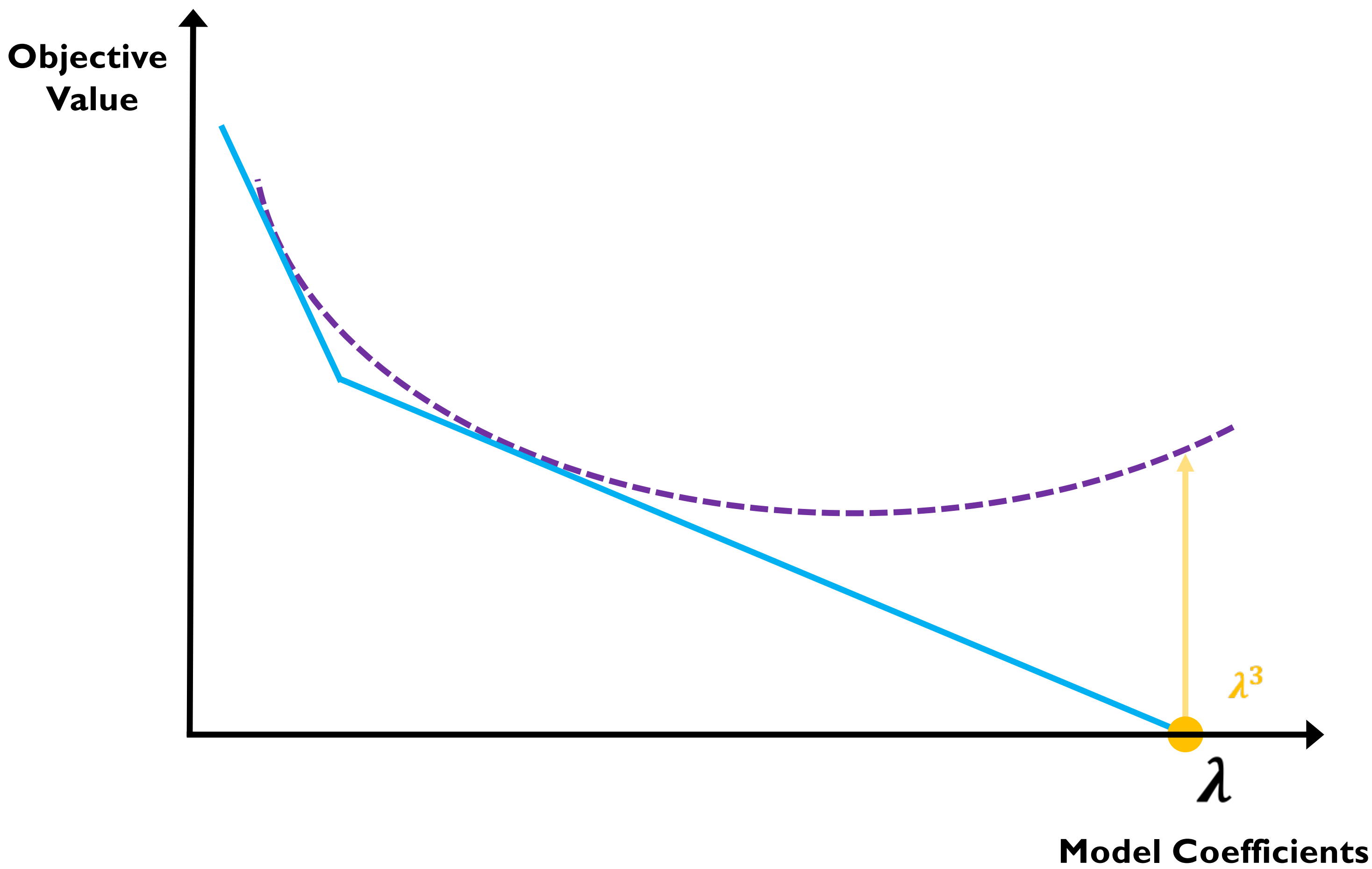
Traditional cutting planes



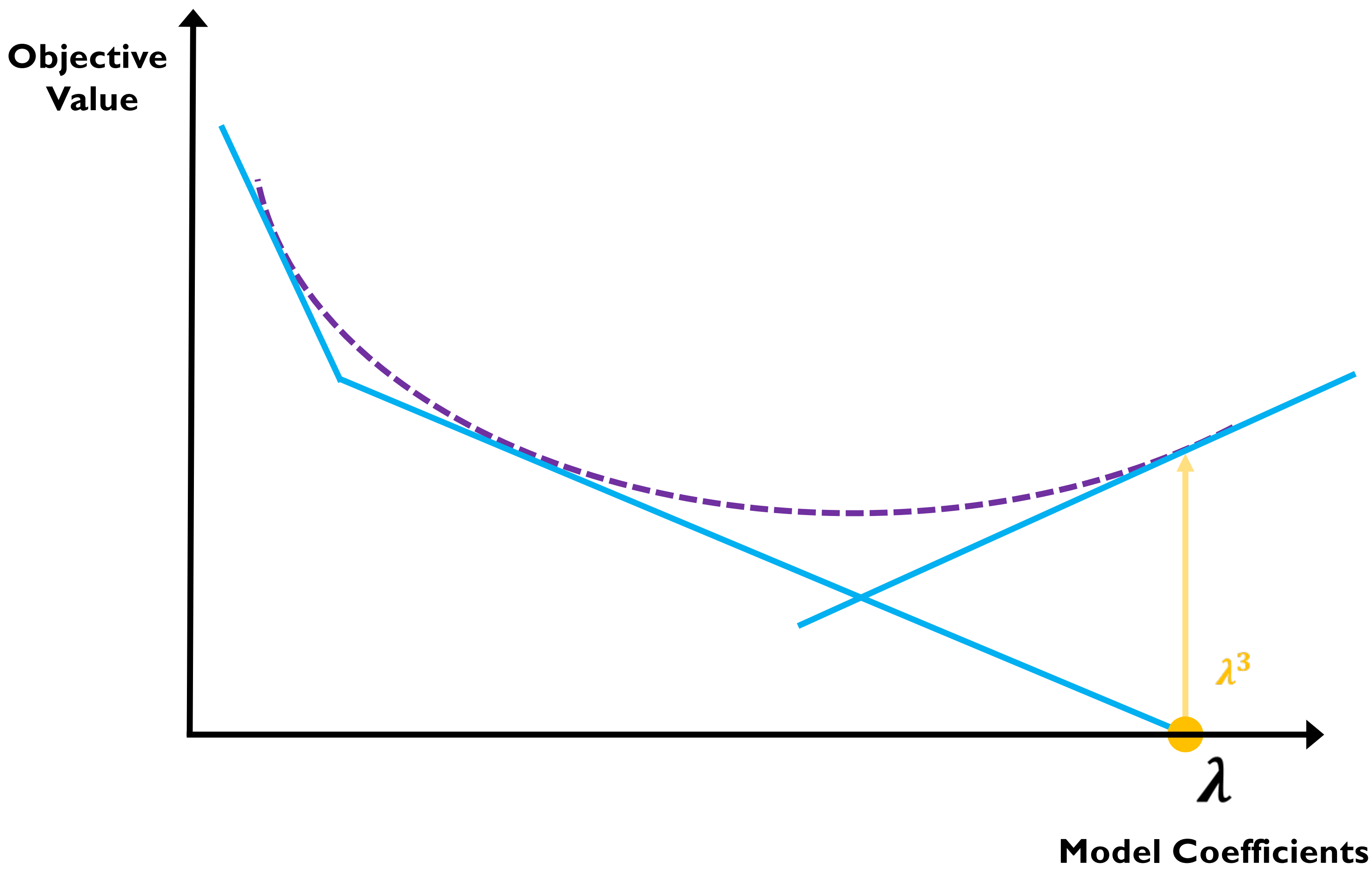
Traditional cutting planes



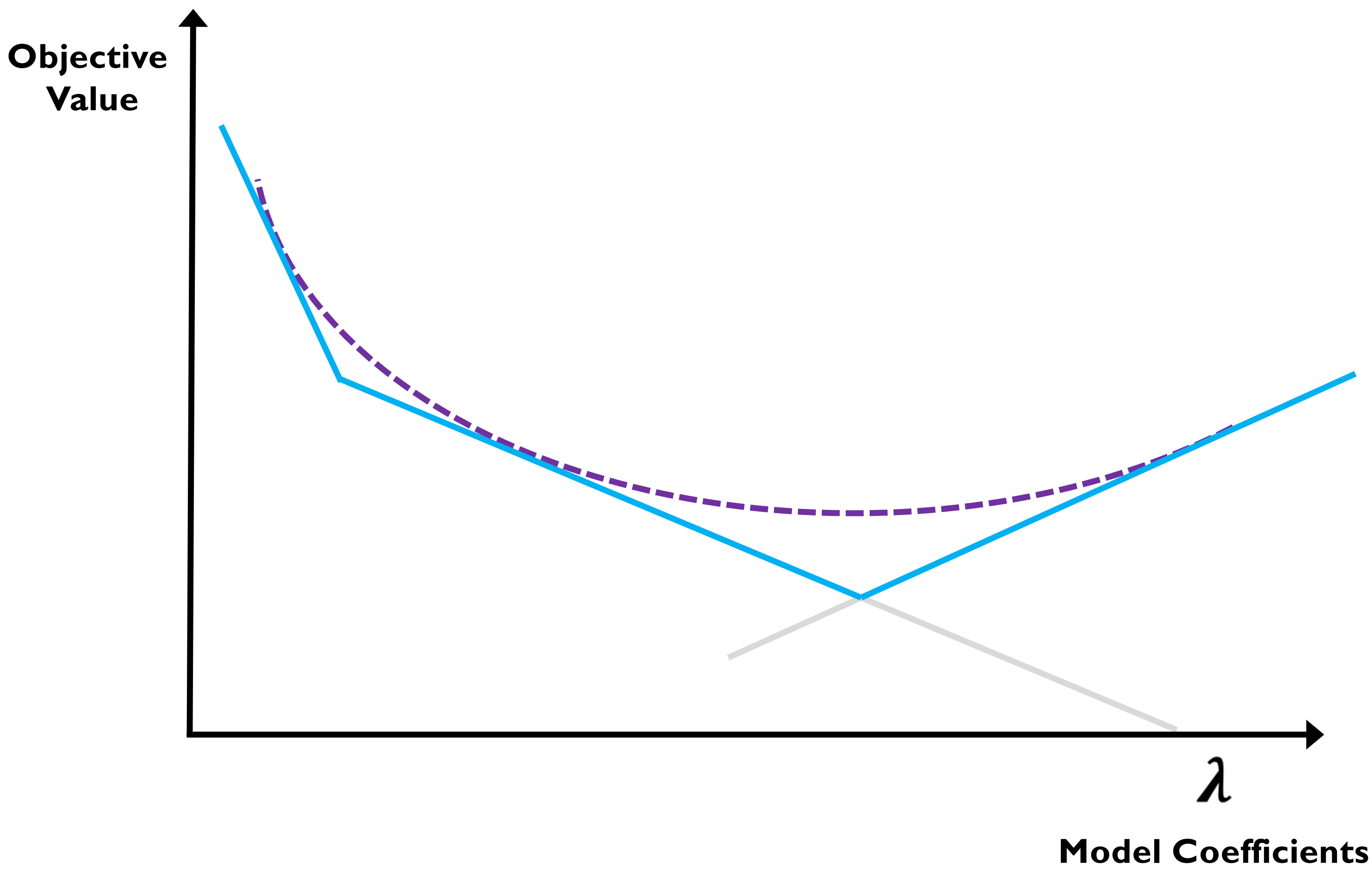
Traditional cutting planes



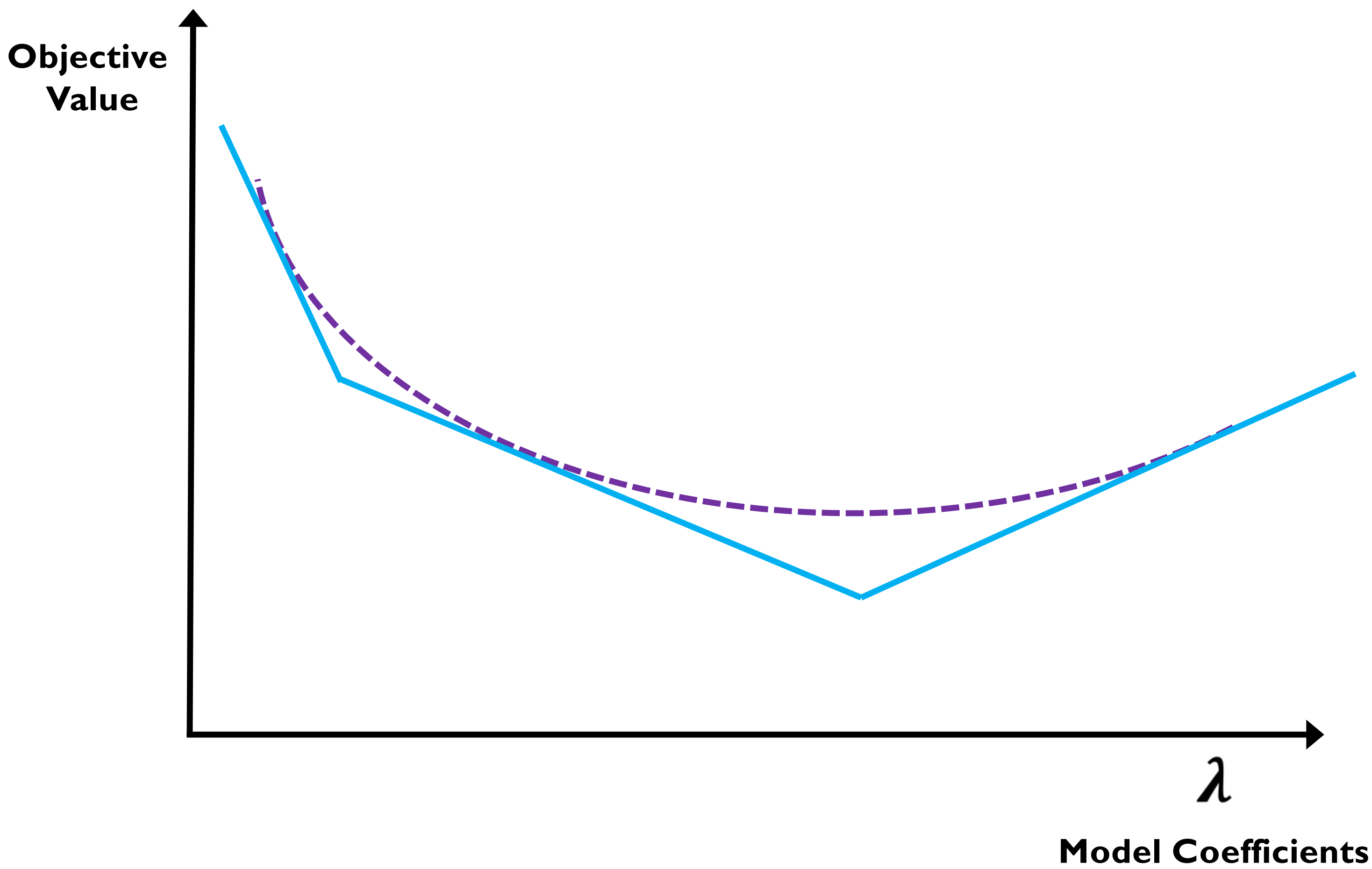
Traditional cutting planes



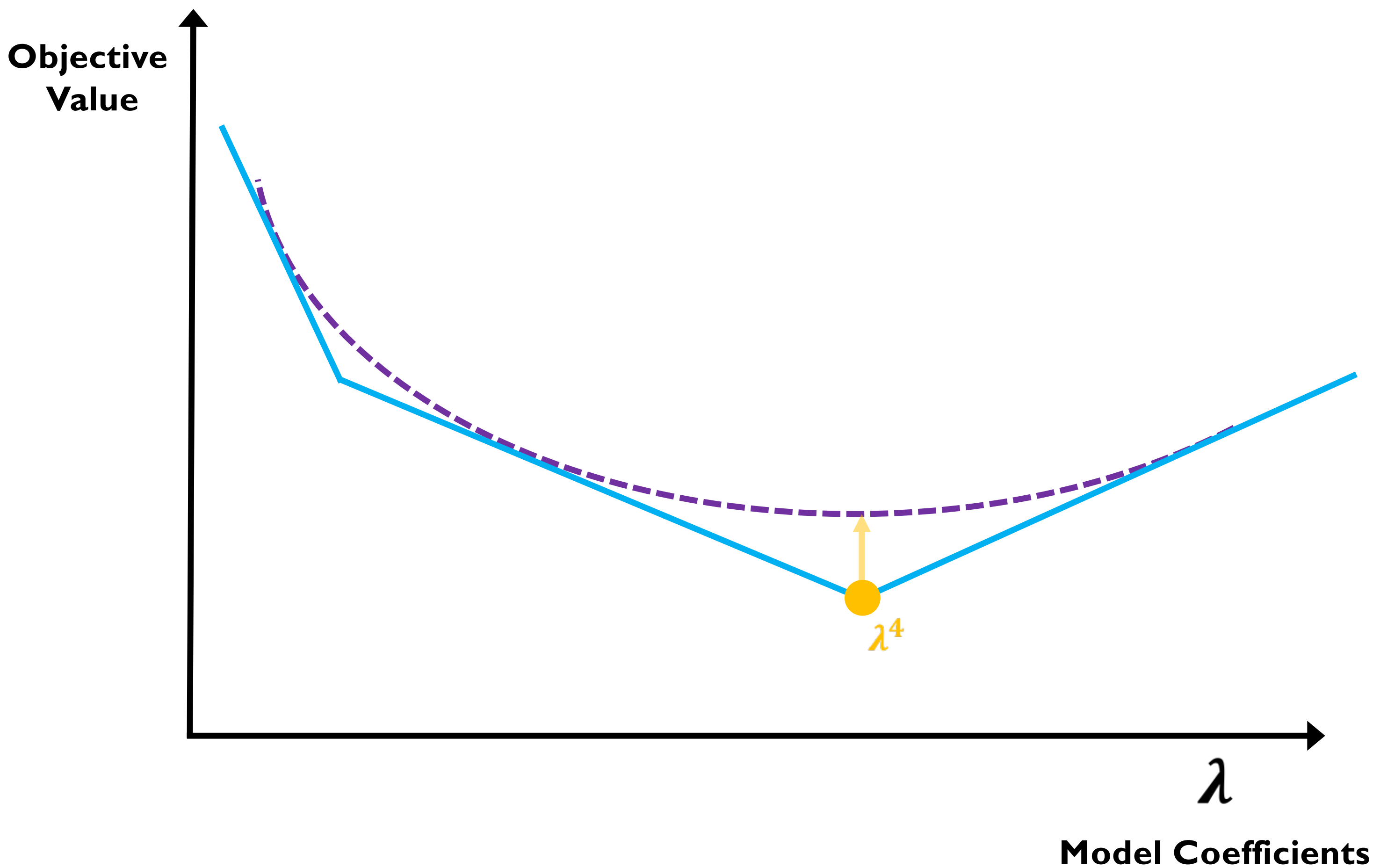
Traditional cutting planes



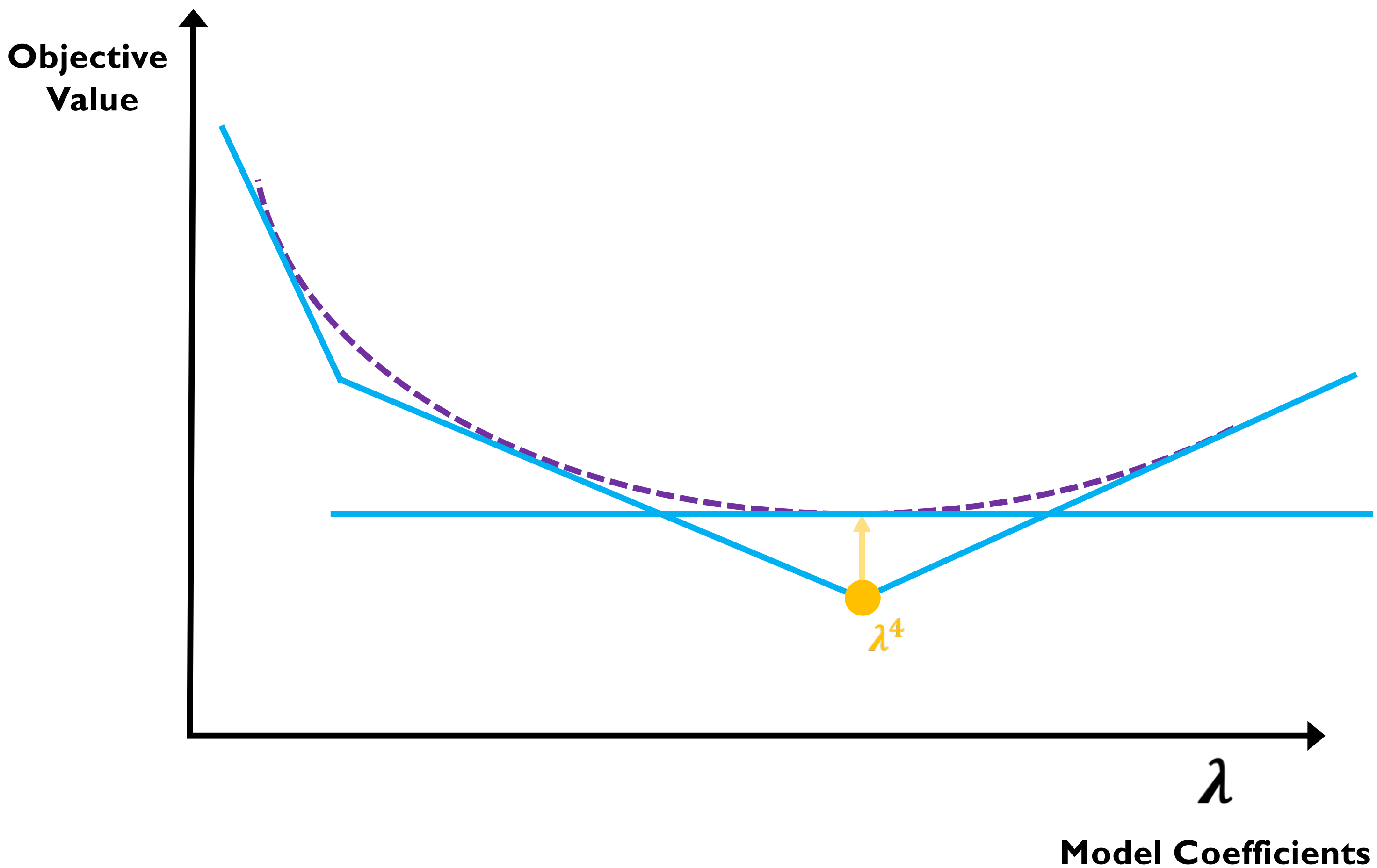
Traditional cutting planes



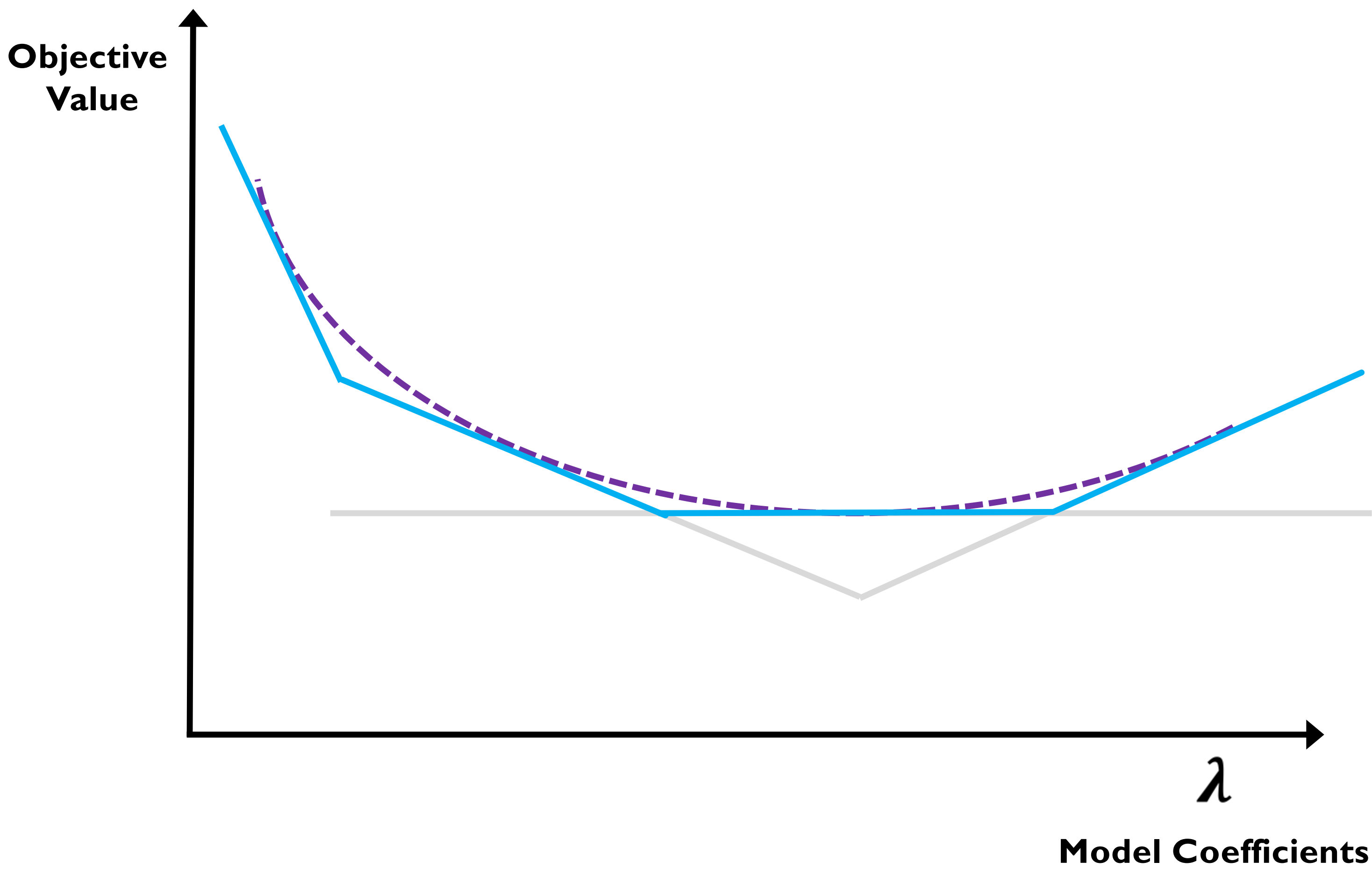
Traditional cutting planes



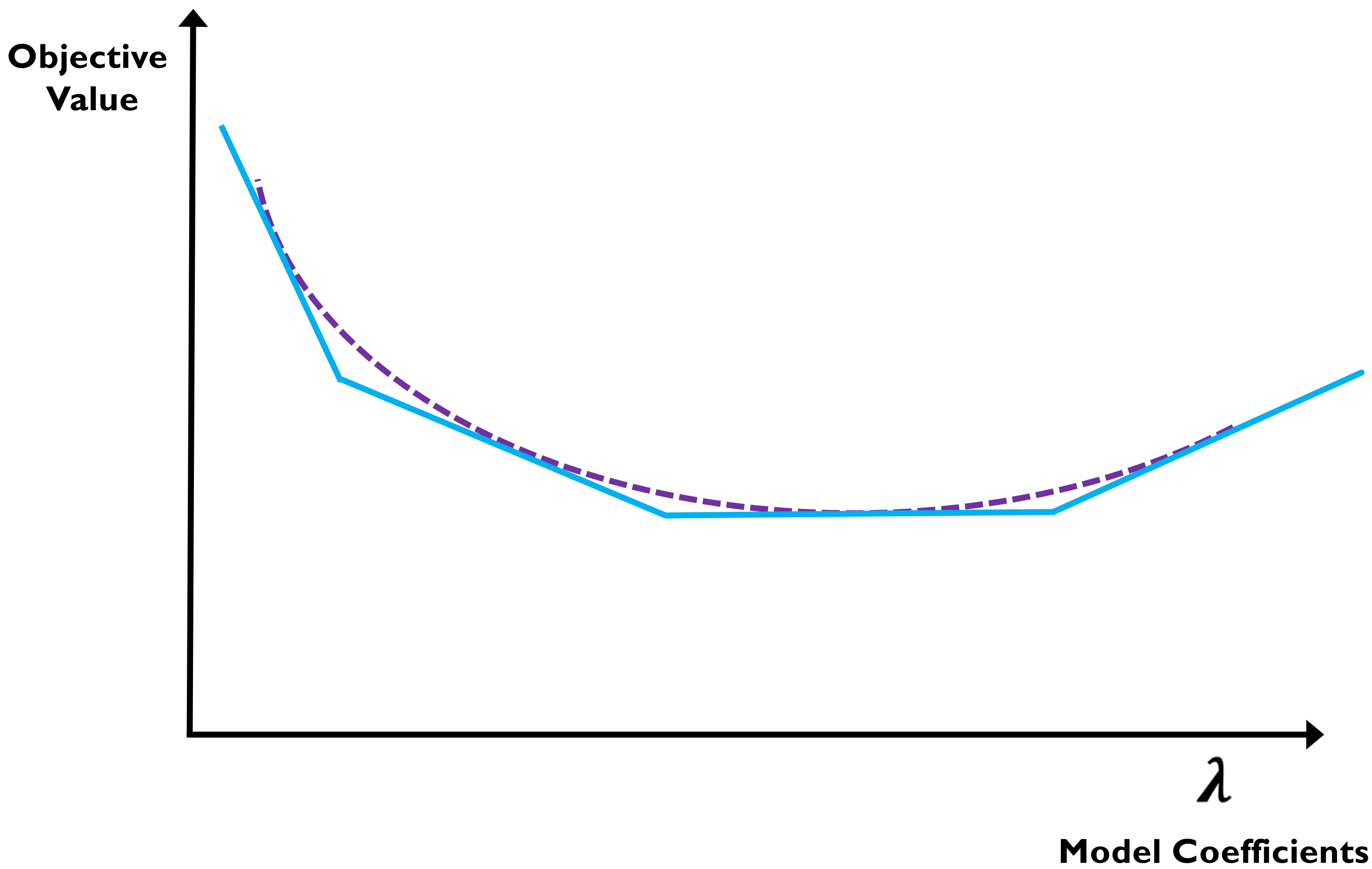
Traditional cutting planes



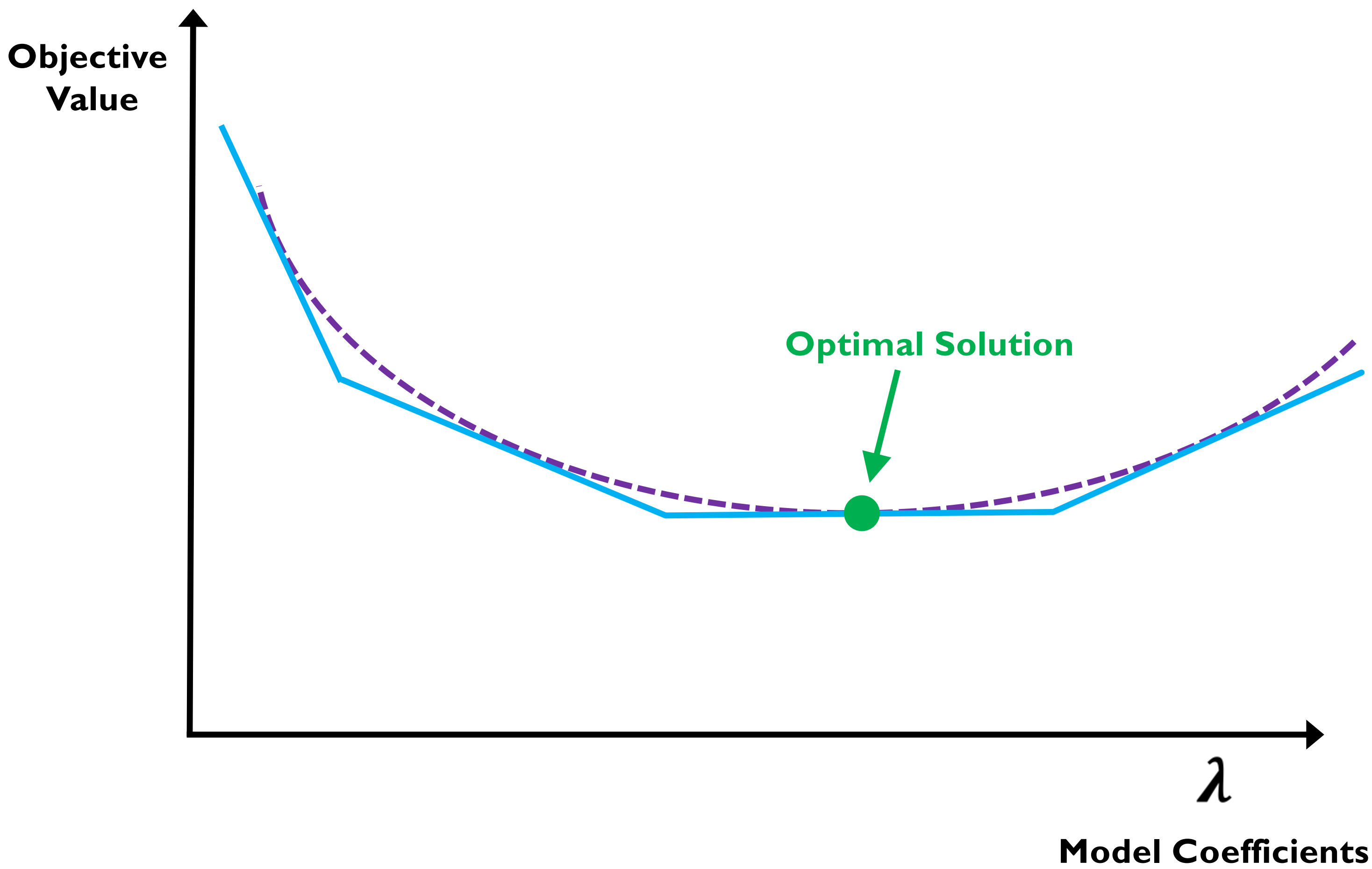
Traditional cutting planes



Traditional cutting planes

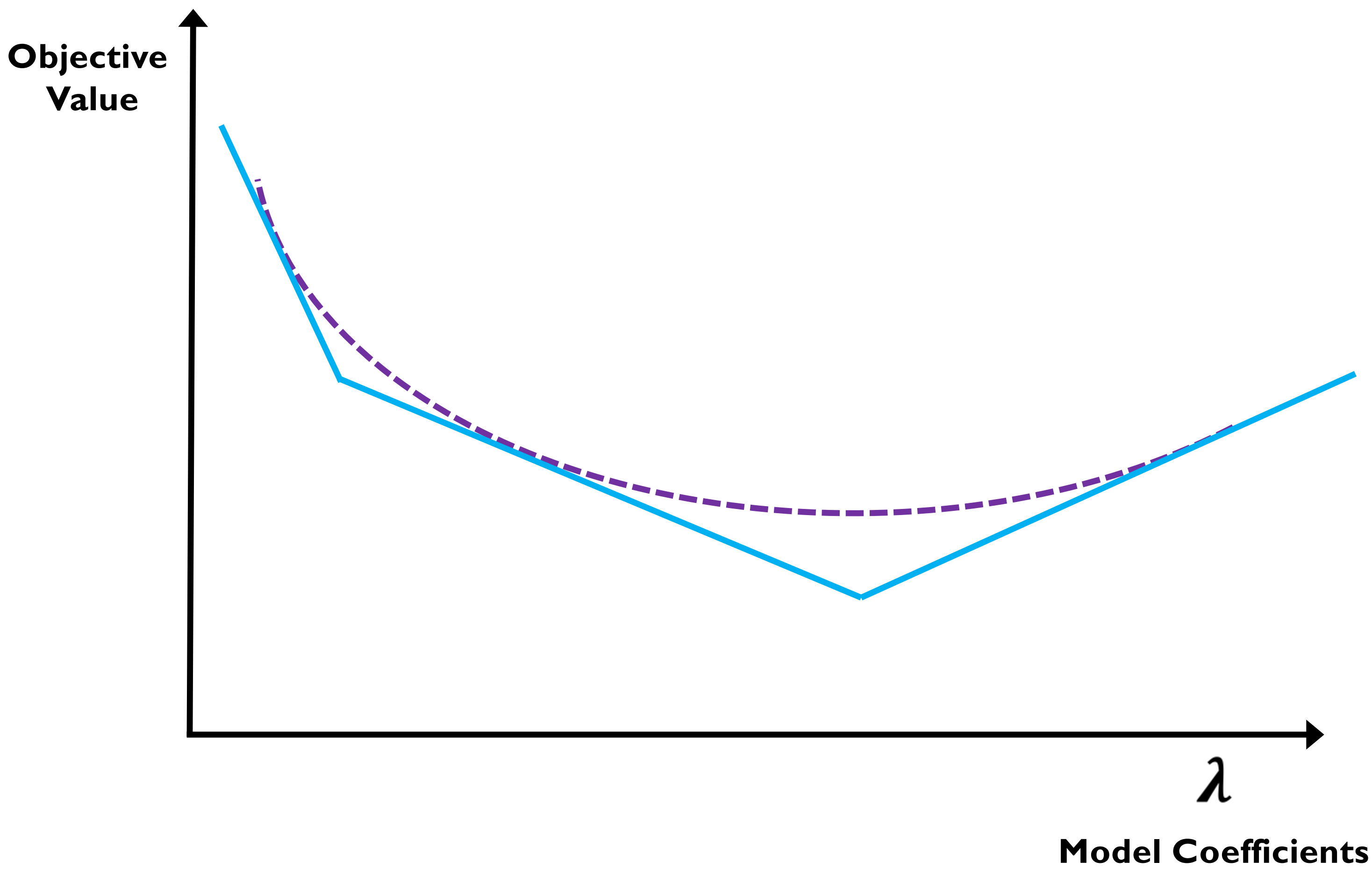


Traditional cutting planes

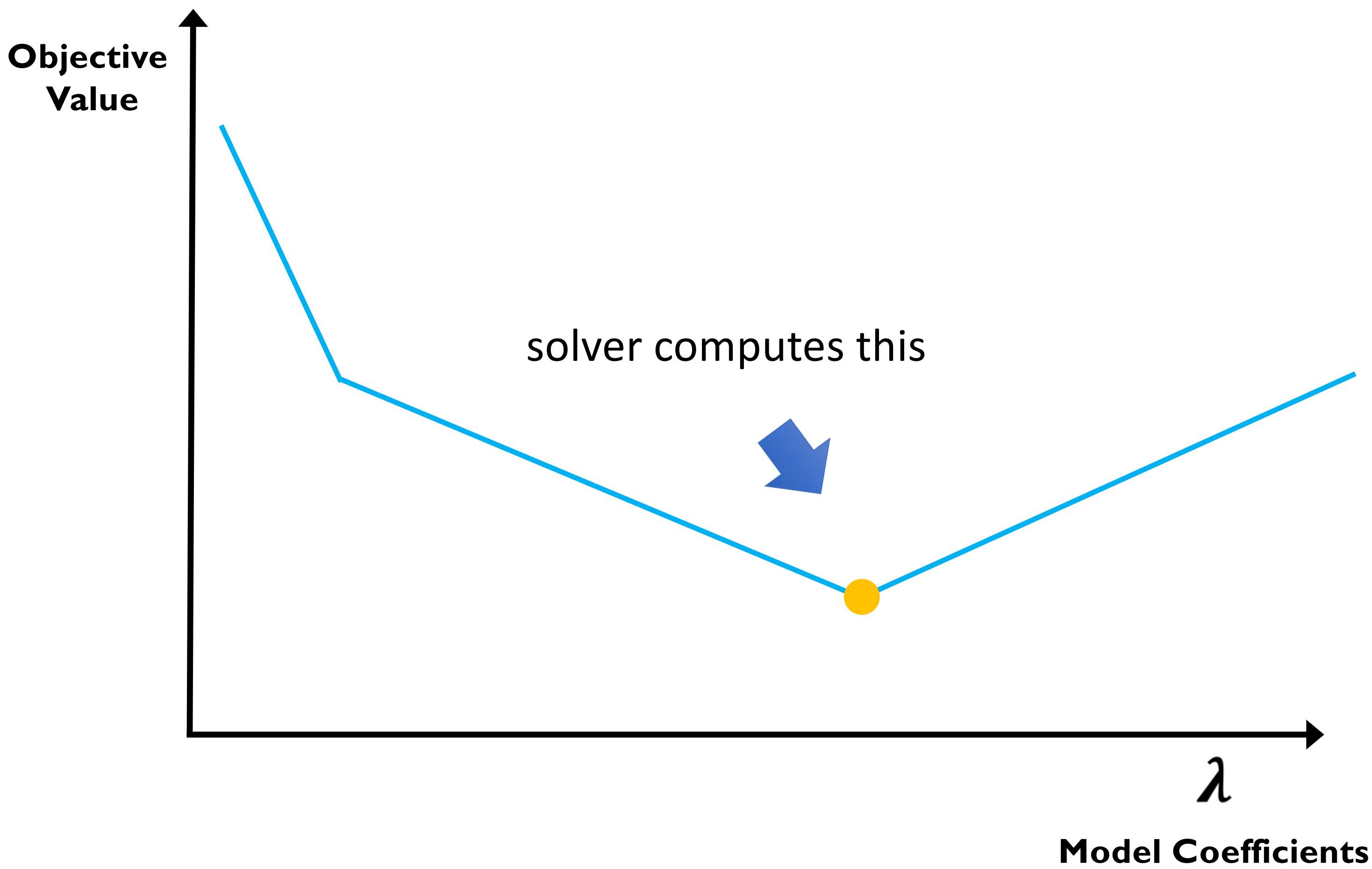


- Something goes wrong when creating models with integer coefficients.

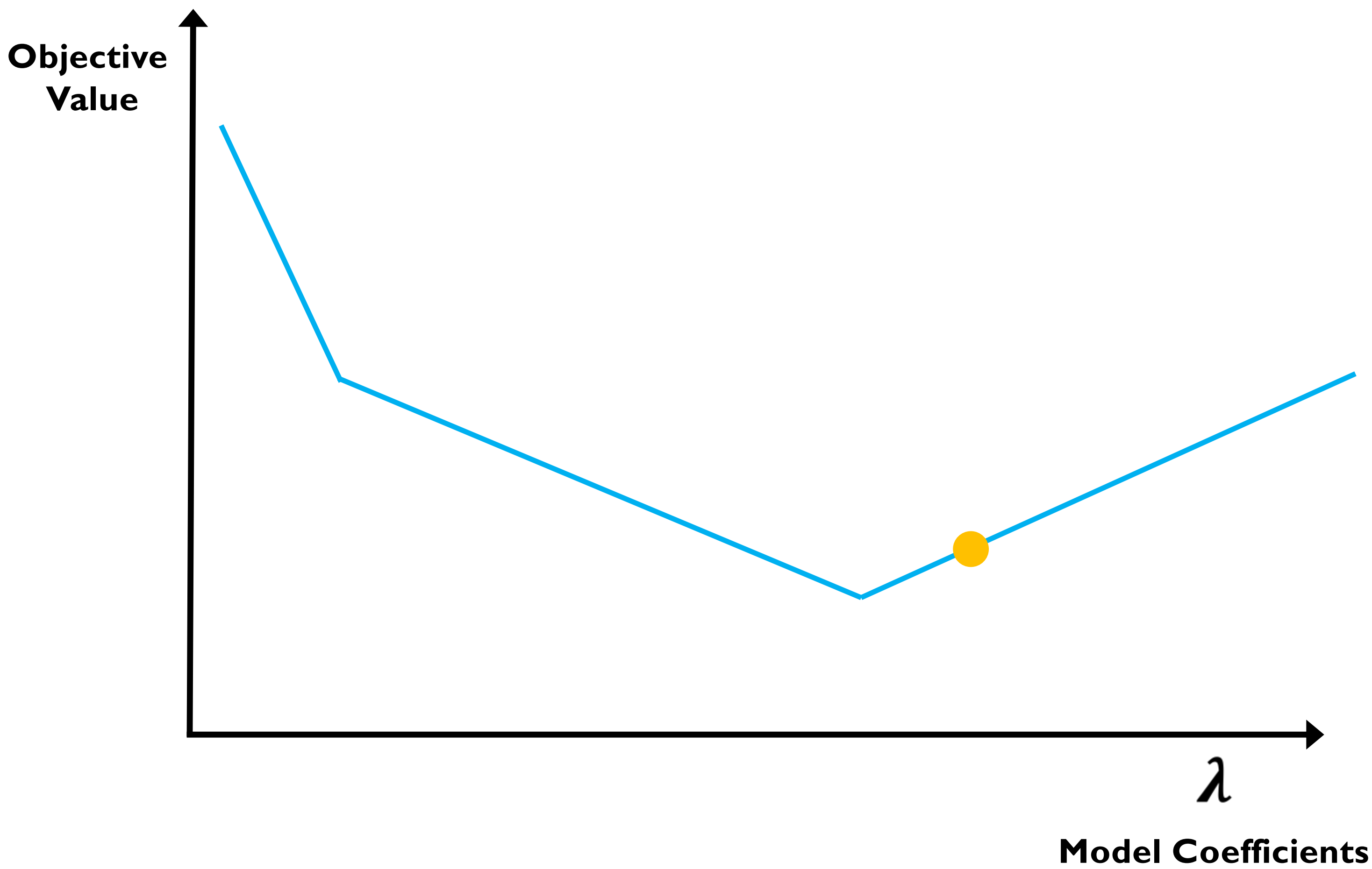
Traditional cutting planes



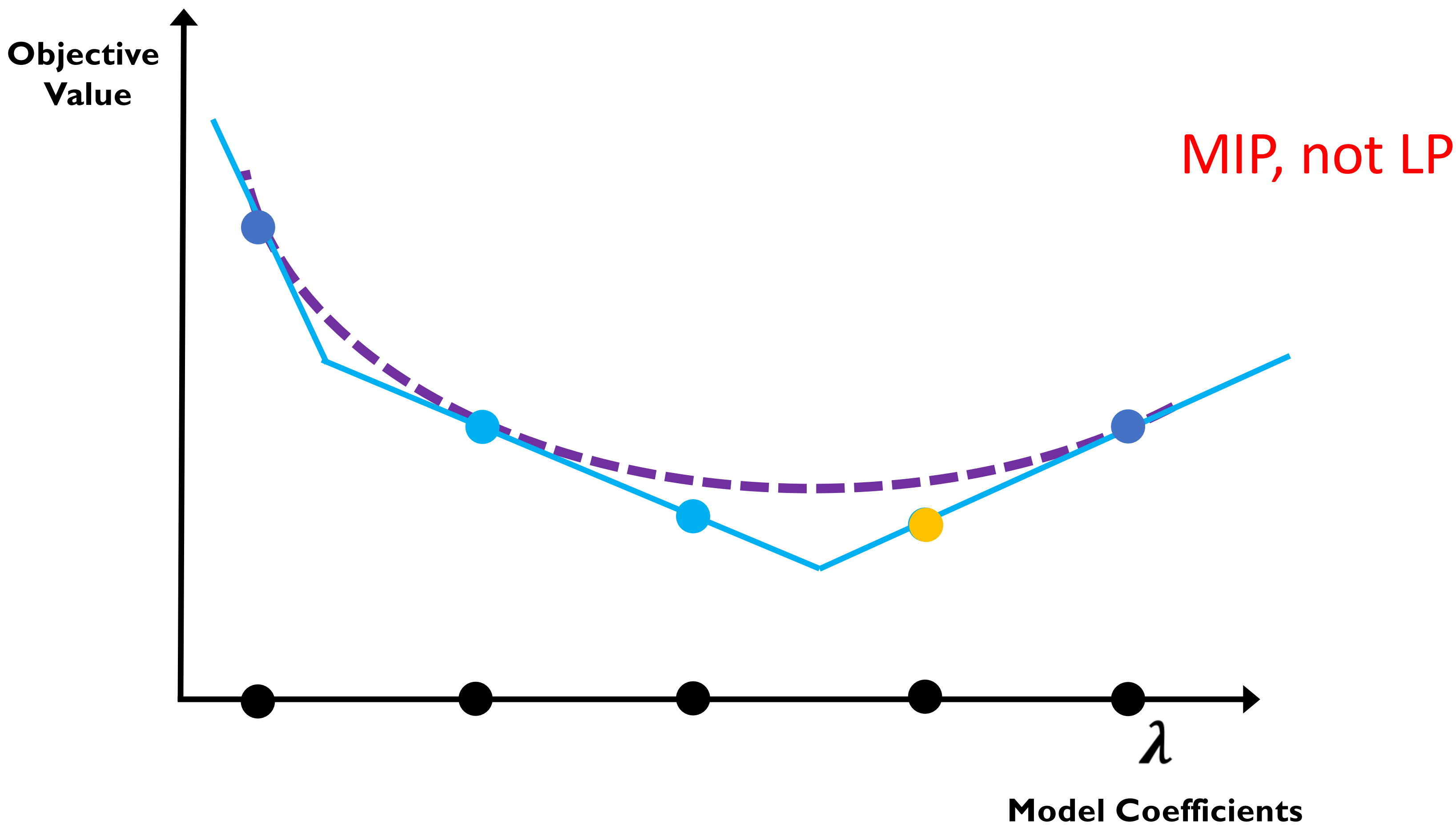
Traditional cutting planes



Traditional cutting planes



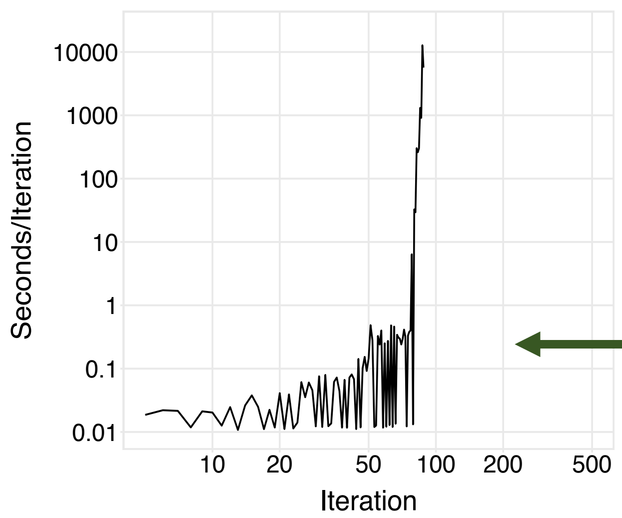
Traditional cutting planes



Stalling

$d = 20$

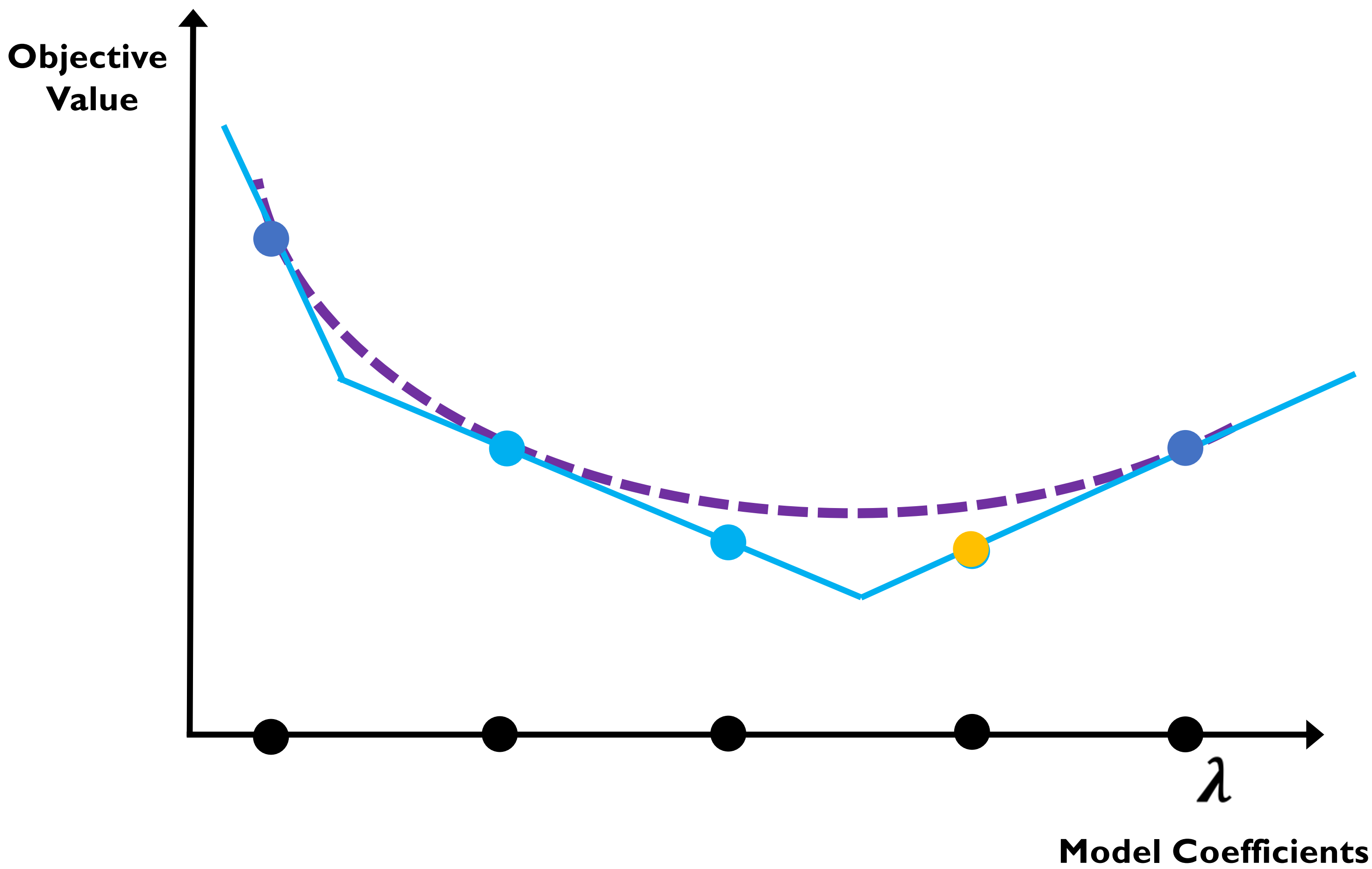
Seconds per iteration



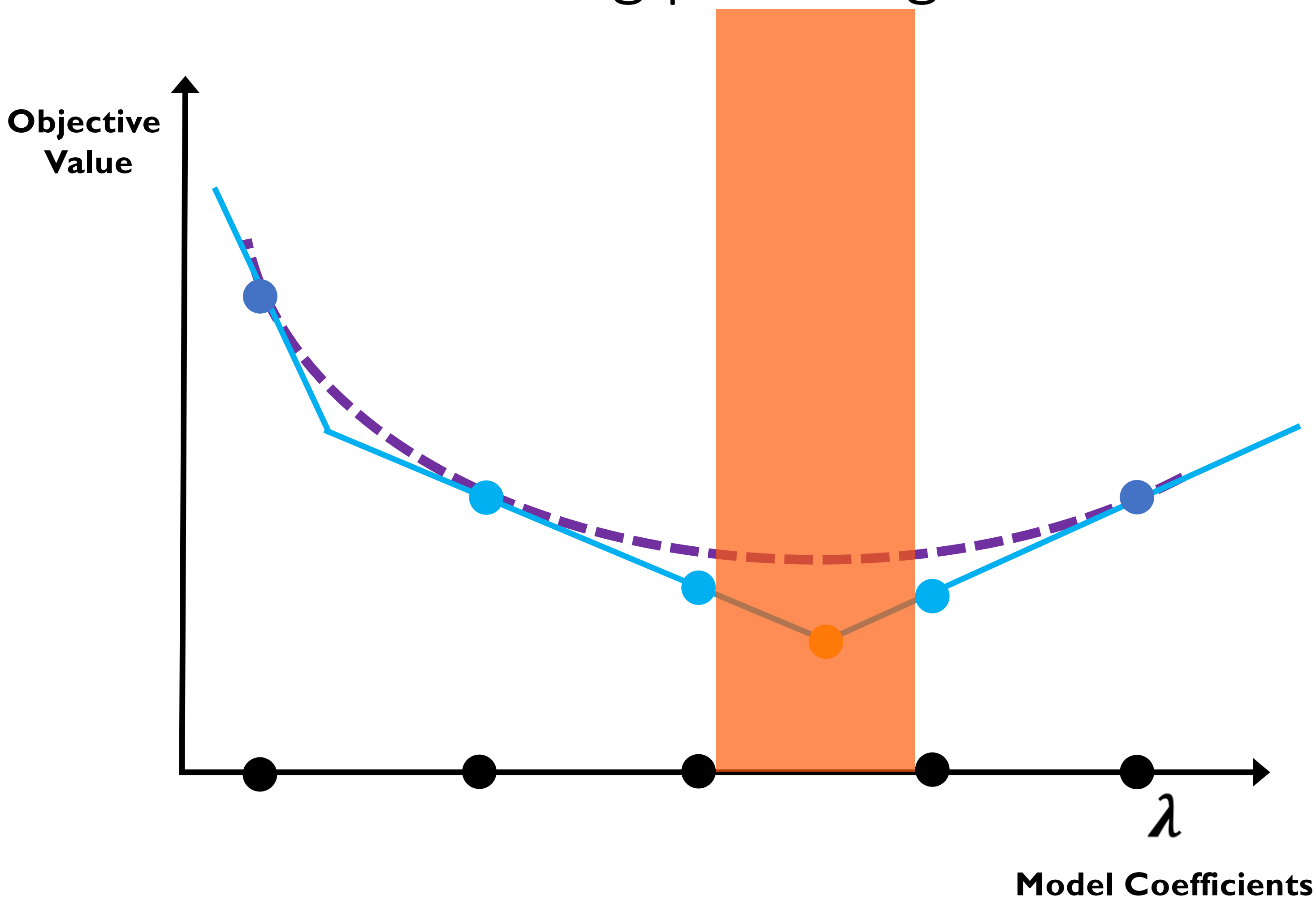
Stalling in traditional cutting planes

RiskSLIM's *Lattice Cutting Plane Algorithm*
(Ustun & Rudin, KDD 17)

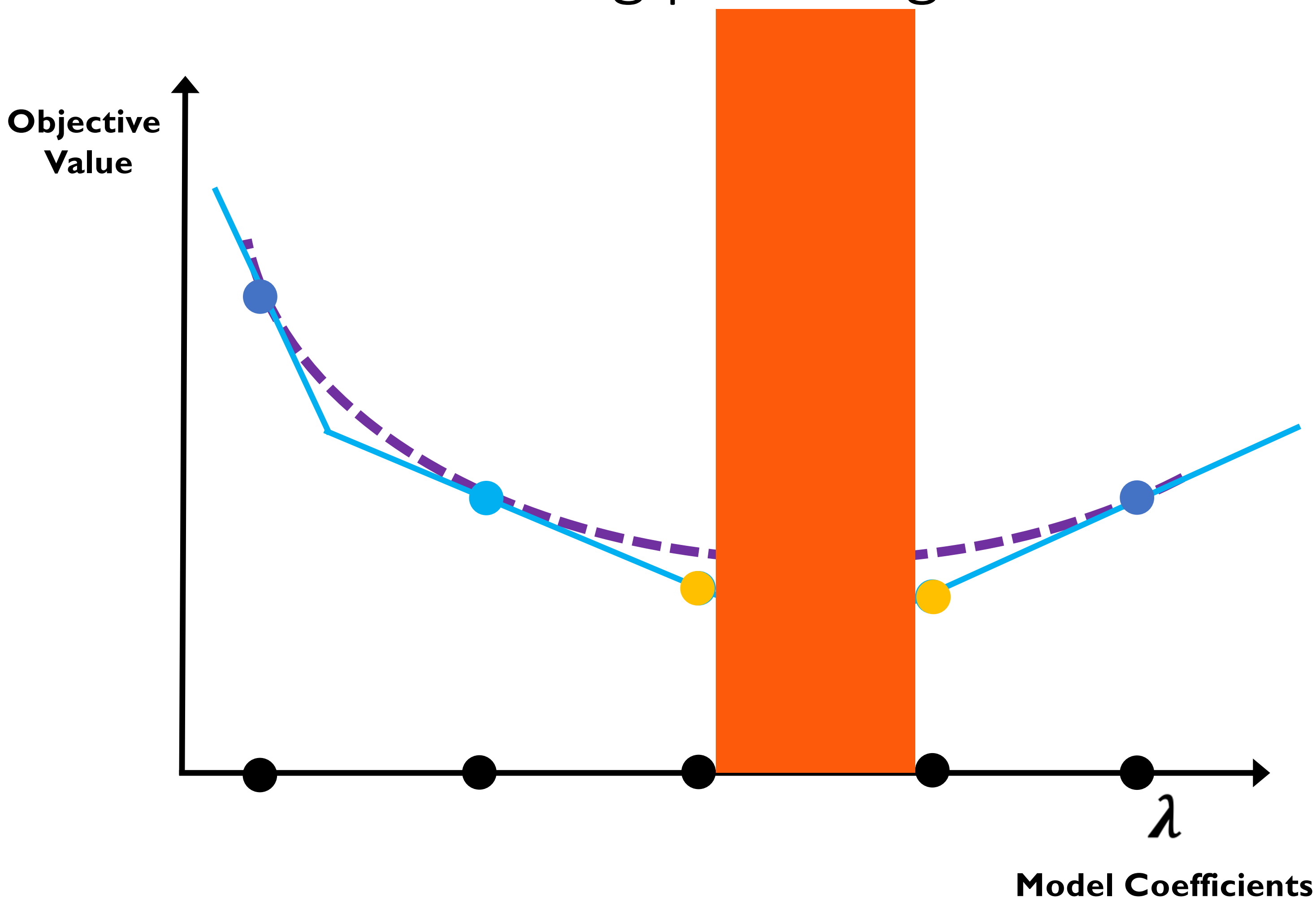
Lattice cutting plane algorithm



Lattice cutting plane algorithm

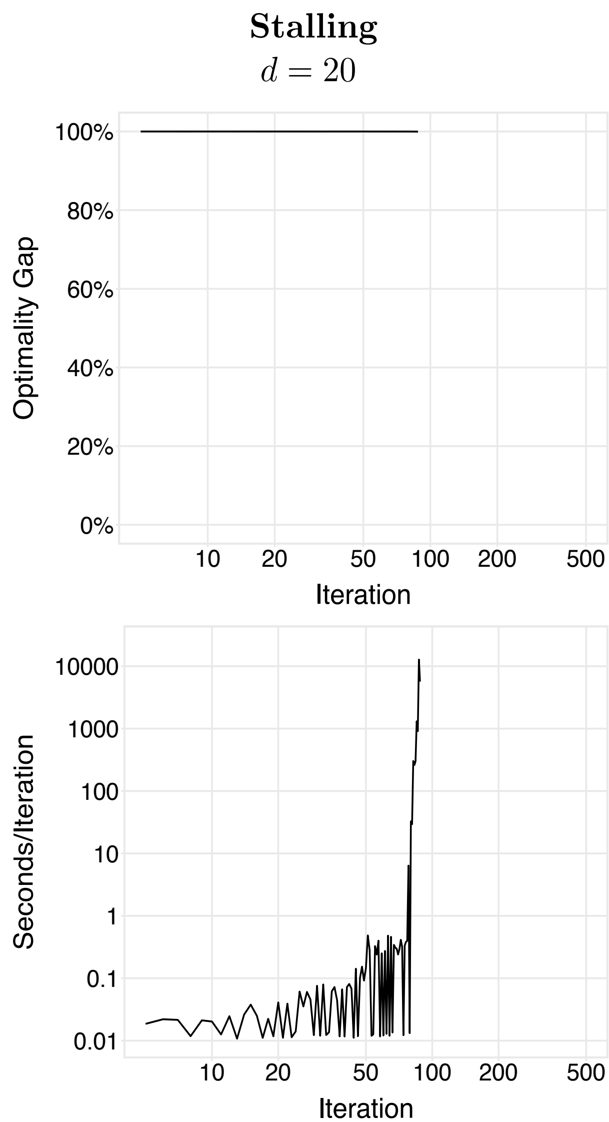


Lattice cutting plane algorithm



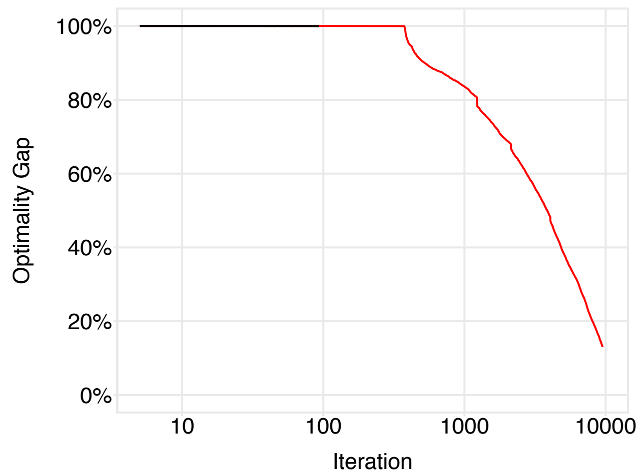
Optimality Gap

Seconds per iteration

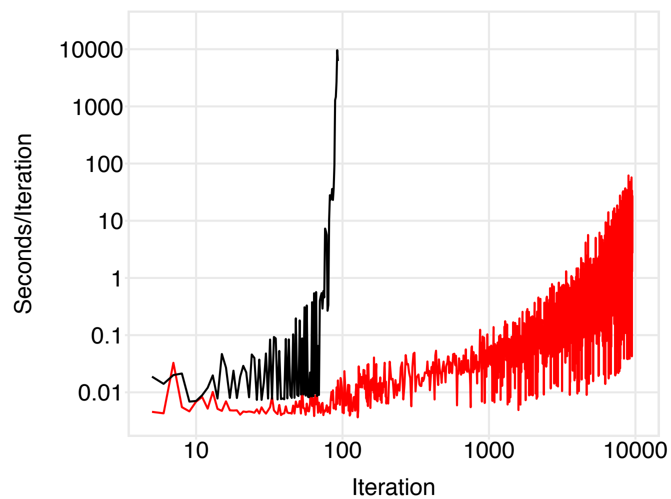


No Stalling

Optimality Gap



Seconds per iteration

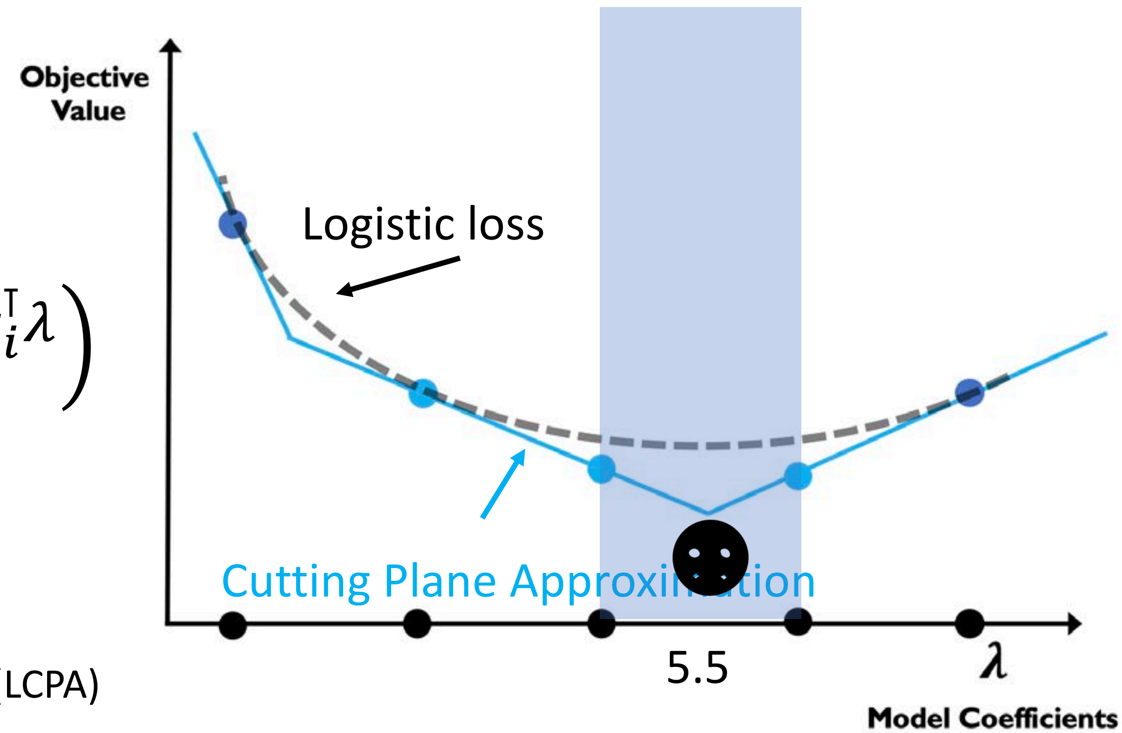


No Stalling for Lattice Cutting Plane Algorithm

Risk-SLIM

(Ustun, R, JMLR 2019)

$$\min_{\lambda \in L} \sum_{i=1}^{n} \log\left(1 + e^{-y_i \boldsymbol{x}_i^{\mathsf{T}} \lambda}\right)$$



RiskSLIM's Lattice Cutting Plane Algorithm (LCPA)

1 2 1 5.5 6.3 3.8 1 0 9 7

≤ 5 ≥ 6

2 subproblems

If a subproblem leads to a feasible integer solution, add a cutting plane.

Otherwise split into 2 subproblems (linear programs).

If min cutting planes = objective, solved!

Risk-SLIM

(Ustun, R, JMLR 2019)

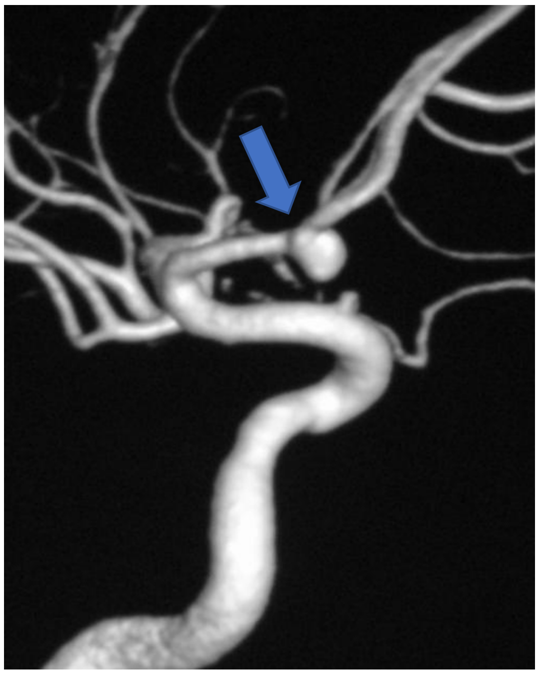
- LCPA is the only method that generates solutions within a reasonable time.
  - MINLP solvers don't work
  - standard cutting planes require solving larger and larger MIPs.

Polishing with SequentialRounding and Discrete Coordinate Descent (DCD)

(Ustun, R, 2019)

1	2	1	5.5	6.3	3.8	1	0	9.8	7	SequentialRounding
1	2	1	5.5	6.3	4	1	0	9.8	7	
1	2	1	5	6.3	4	1	0	9.8	7	
1	2	1	5	7	4	1	0	9.8	7	
1	2	1	5	7	4	1	0	10	7	DCD
1	2	1	5	7	4	2	0	10	7	
1	2	4	5	7	4	1	0	10	7	
1	1	4	5	7	4	1	0	10	7	"1-opt solution"

Preventing Brain Damage in Critically Ill Patients



CT-angiography, Anterior Communicating Saccular Aneurysm

Head CT without contrast showing Subarachnoid Hemorrhage

- Seizure are common (20%)
- Seizure→ Brain Damage
- Need EEG to detect seizures

Need to use EEG data to predict seizures, determine EEG duration

EEG is expensive and limited: 24hrs of monitoring is $1600-$4000

- 2HELPS2B was not created by doctors
- It is a ML model
- It is just as accurate as black box models.
- Doctors can decide themselves whether to trust it
- Doctors can calibrate the score with information not in the database
- Score can be explained to non-physicians

2HELPS2B

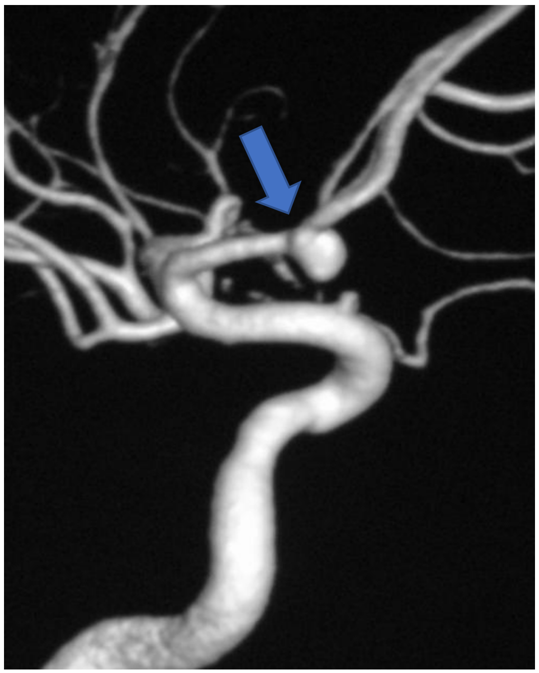
1.	Any cEEG Pattern with Frequency **2 H**z	1 point	···
2.	**E**pileptiform Discharges	1 point	+ ···
3.	Patterns include [**L**PD, LRDA, BIPD]	1 point	+ ···
4.	**P**atterns Superimposed with Fast or Sharp Activity	1 point	+ ···
5.	Prior **S**eizure	1 point	+ ···
6.	**B**rief Rhythmic Discharges	**2** points	+ ···
		SCORE	= ···

SCORE	0	1	2	3	4	5	6+
RISK	<5%	11.9%	26.9%	50.0%	73.1%	88.1%	95.3%

There are many variables to choose from.

Variable
PDR
BRDs
Unreactive background
Prior Sz
GRDA
LRDA
GPDs
LPDs
BIPDs
Infection
Inflammation
Neoplasm
ICH
Metabolic encephalopathy
Stroke
SAH
SDH
TBI
Hypoxic/ischemic
IVH
Hydrocephalus
Discharges
Frequency (>2Hz)[c]

Preventing Brain Damage in Critically Ill Patients



CT-angiography, Anterior Communicating Saccular Aneurysm

Head CT without contrast showing Subarachnoid Hemorrhage

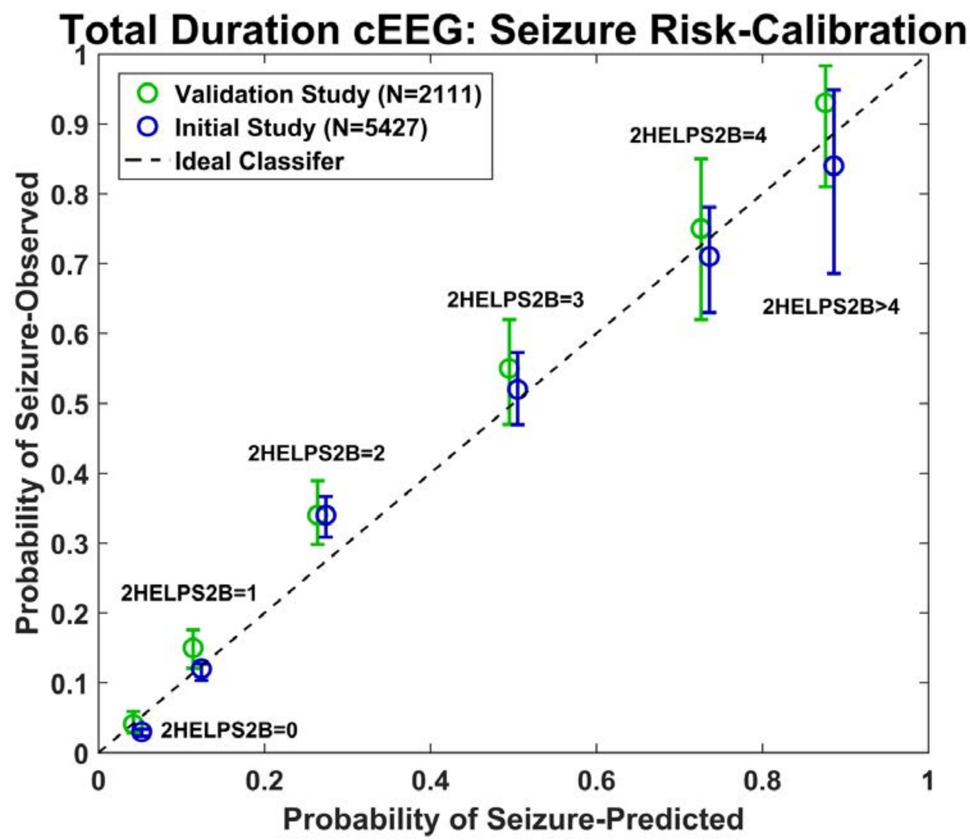
1-Hour Screening EEG

2HELPS2B=3 (high-risk)

- Placed on Continuous EEG for >72H
- Start on preventative medications

So far…

- 2HELPS2B validated on independent multicenter cohort (N=2111)
- Implemented: University of Wisconsin, Massachusetts General Hospital/Harvard Medical School
- Ongoing implementation: Emory University, Duke University, Medical University of South Carolina, Free University of Brussels (Belgium)
- Resulted in **63.6%** reduction in duration of EEG monitoring per patient
  - $1,134.831 saving per patient[1]
- **2.82 X** More Patients Monitored
- **$6.1M** estimated savings in FY 2018 at MGH,UW



[1]2016 Medicare Reimbursement Most Common Professional Code

Problem spectrum

Very sparse models (trees, scoring systems)

With minor pre-processing, all methods have similar performance

Neural networks

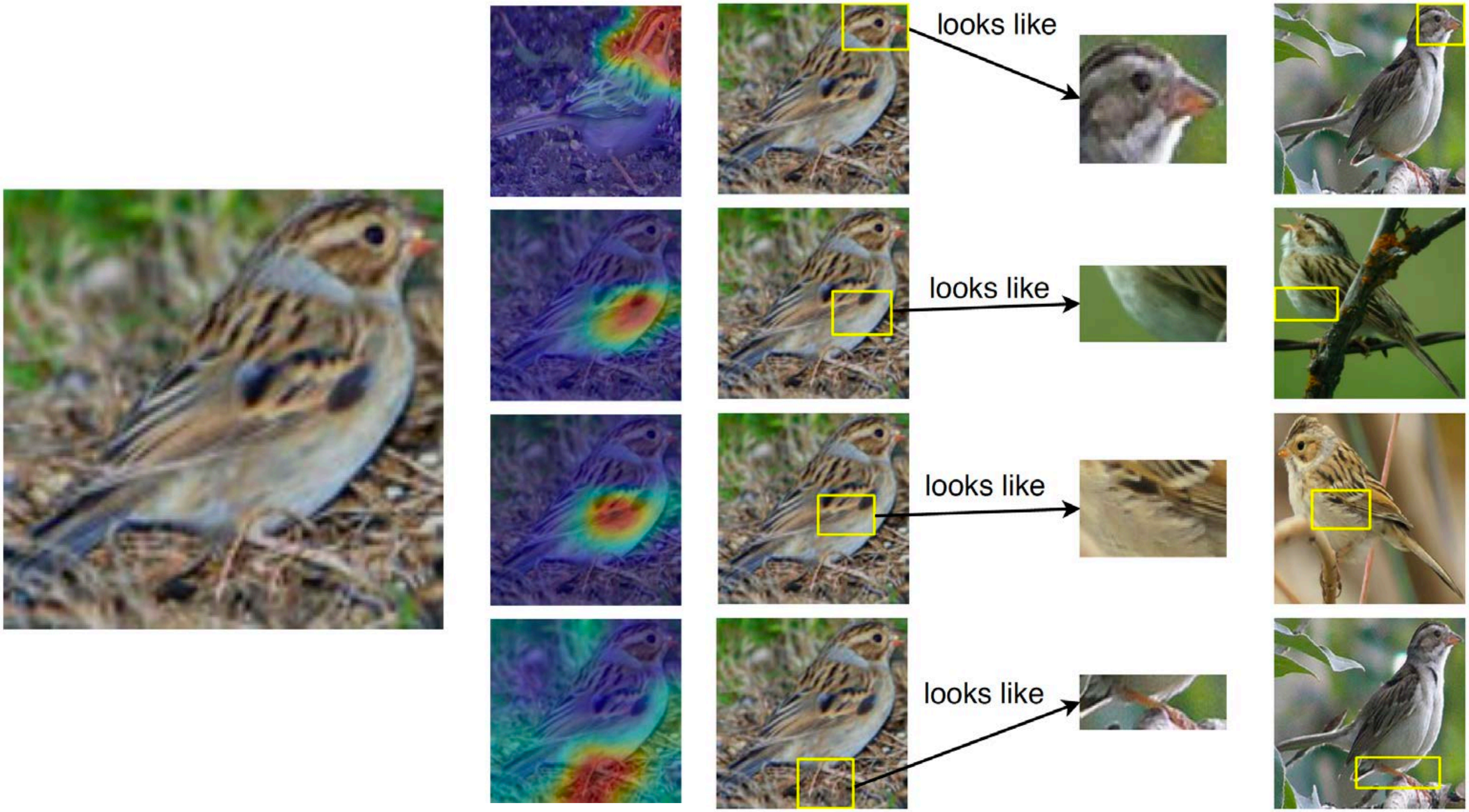
`Tabular`: All features are interpretable

- many problems in criminal justice, healthcare, social sciences, equipment reliability & maintenance, etc.
- features include counts, categorical data

Raw: Features are individually uninterpretable

- pixels/voxels, words, a bit of a sound wave

Interpretable neural networks?


looks like
looks like
looks like
looks like

NeurIPS 2019 (spotlight)

arXiv.org > cs > arXiv:1806.10574

Search or Articl

(Help | Advanced se

Computer Science > Machine Learning

This looks like that: deep learning for interpretable image recognition

Chaofan Chen, Oscar Li, Alina Barnett, Jonathan Su, Cynthia Rudin

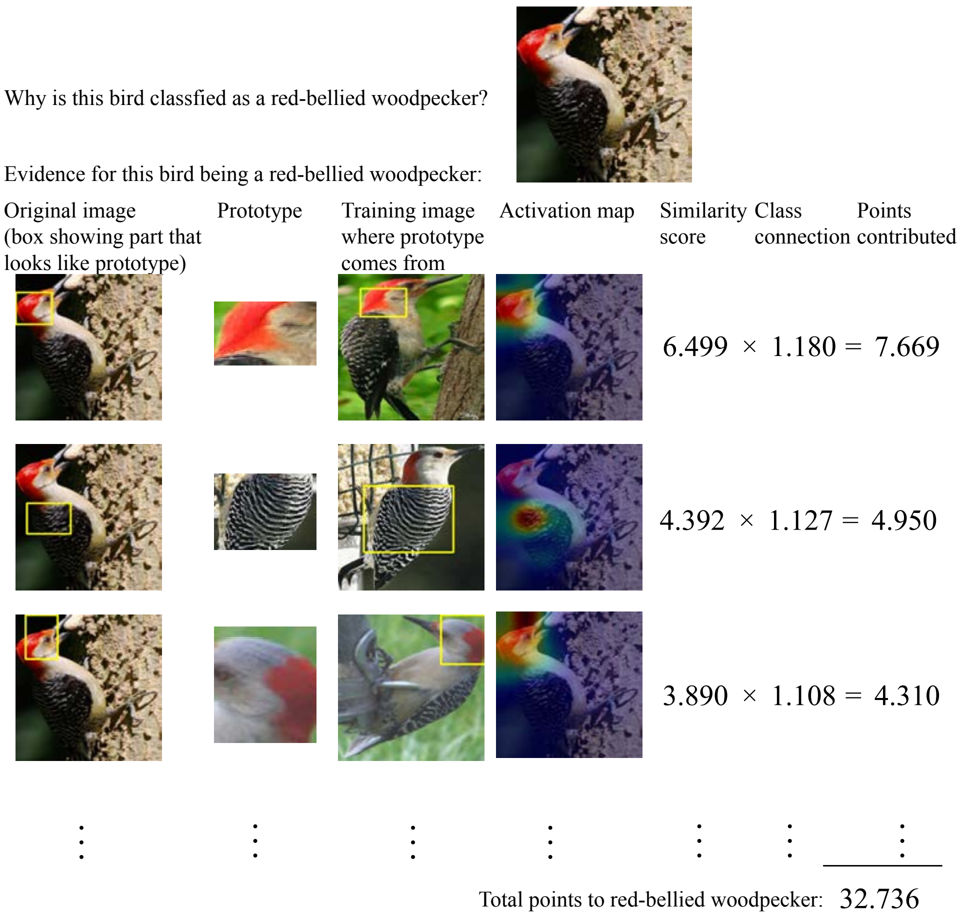
(Submitted on 27 Jun 2018)

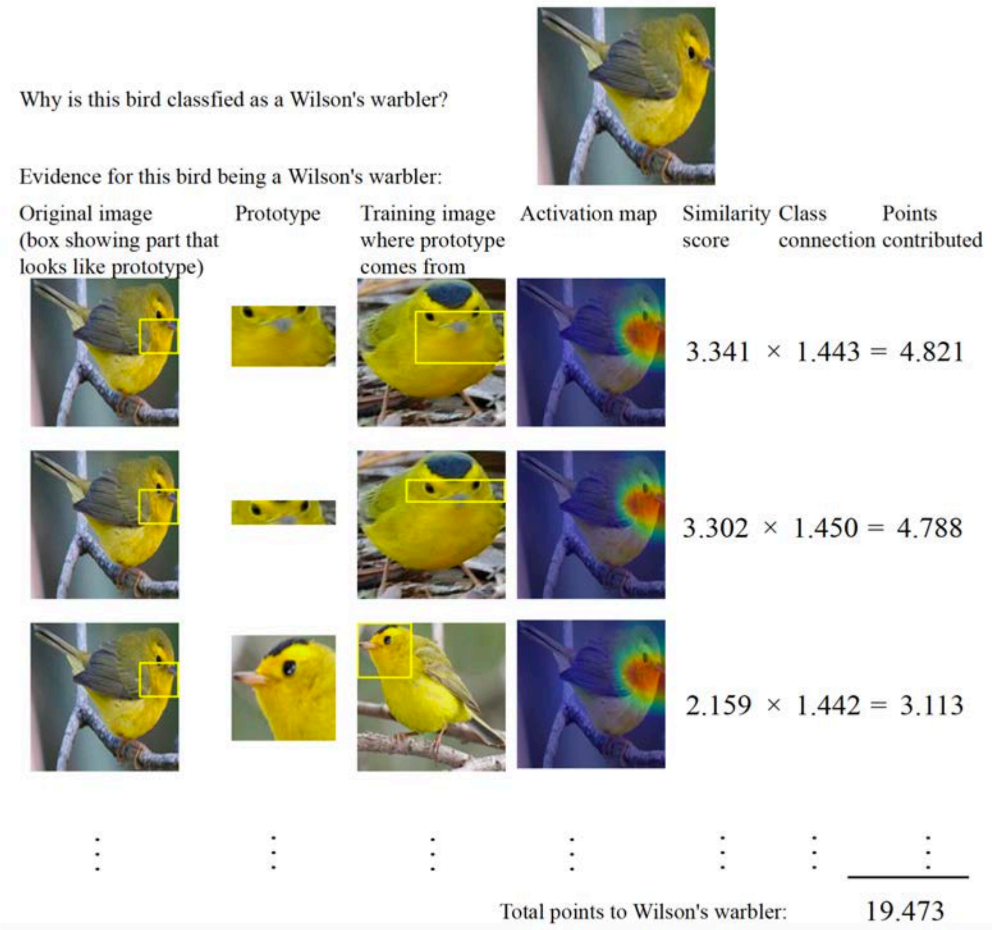
When we are faced with challenging image classification tasks, we often explain our reasoning by dissecting the image, and pointing out prototypical aspects of one class or another. The mounting evidence for each of the classes helps us make our final decision. In this work, we introduce a deep network architecture that reasons in a similar way: the network dissects the image by finding prototypical parts, and combines evidence from the prototypes to make a final classification. The algorithm thus reasons in a way that is qualitatively similar to the way ornithologists, physicians, geologists, architects, and others would explain to people on how to solve challenging image classification tasks. The network uses only image-level labels for training, meaning that there are no labels for parts of images. We demonstrate the method on the CIFAR-10 dataset and 10 classes from the CUB-200-2011 dataset.

Chaofan

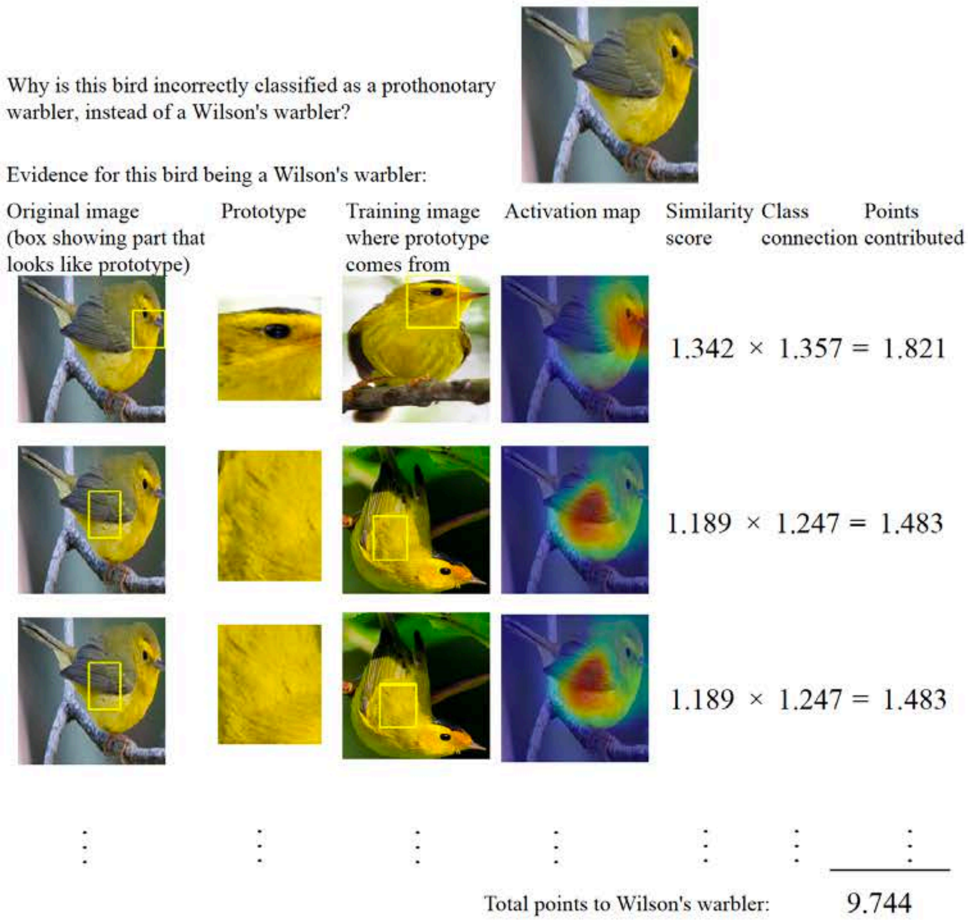
Oscar

Accuracy ≈ black box baselines


Why is this bird classfied as a red-bellied woodpecker?
Evidence for this bird being a red-bellied woodpecker:
Original image (box showing part that looks like prototype)
Prototype
Training image where prototype comes from
Activation map
Similarity score
Class connection
Points contributed
6.499 × 1.180 = 7.669
4.392 × 1.127 = 4.950
3.890 × 1.108 = 4.310
Total points to red-bellied woodpecker: 32.736

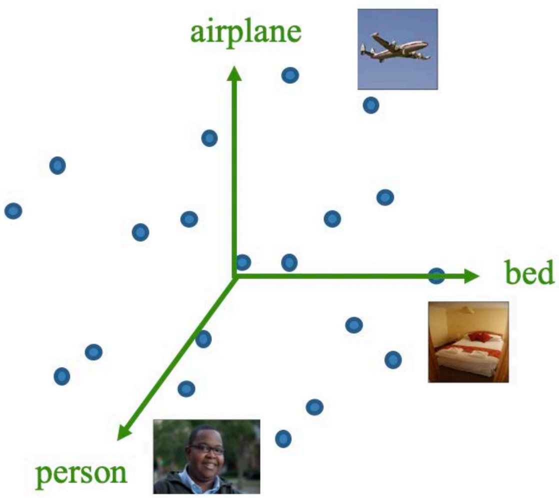


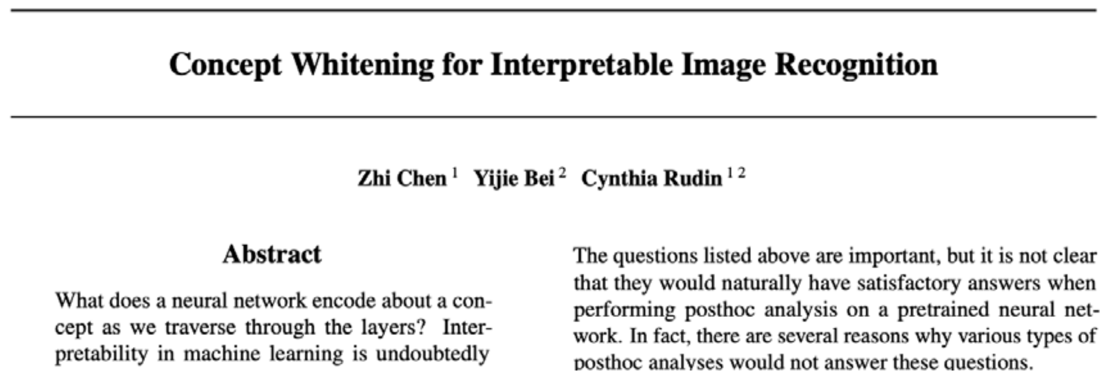
Base model: VGG-16



Base model: DenseNet161

- Even for computer vision, we can still have an interpretable model of the same accuracy as a black box.





Concept Whitening for Interpretable Image Recognition

Zhi Chen [1] Yijie Bei [2] Cynthia Rudin [1 2]

Abstract

What does a neural network encode about a concept as we traverse through the layers? Interpretability in machine learning is undoubtedly

The questions listed above are important, but it is not clear that they would naturally have satisfactory answers when performing posthoc analysis on a pretrained neural network. In fact, there are several reasons why various types of posthoc analyses would not answer these questions.

Nature Machine Intelligence, accepted, Oct 2020

The Idea

- Create a latent space that tells us *how* it is disentangling concepts
- Form the latent space so that its axes represent known concepts
- It's easy to do: Just replace a batch normalization step with a "Concept Whitening" step.
- Instead of normalizing, whiten and rotate.

Summary

- Trees: Modern decision tree methods are not your old CART.
- Scoring systems: Rounding linear model coefficients can go against the performance gradient. LCPA helps.
- Interpretable neural networks for computer vision: yes, they exist.

Jimmy Lin, Chudi Zhong, Diane Hu, Cynthia Rudin, Margo Seltzer
Generalized and Scalable Optimal Sparse Decision Trees. ICML, 2020.

Berk Ustun and Cynthia Rudin
Learning Optimized Risk Scores. JMLR, 2019. Shorter version at KDD 2017.

Aaron F. Struck, Berk Ustun, ….., Cynthia Rudin, M Brandon Westover.
Association of an Electroencephalography-Based Risk Score With Seizure Probability in Hospitalized Patients. JAMA Neurology, 2017

Chaofan Chen, Oscar Li, Chaofan Tao, Alina Barnett, Jonathan Su, Cynthia Rudin
This Looks Like That: Deep Learning for Interpretable Image Recognition. NeurIPS, 2019.

Zhi Chen, Yijie Bei, Cynthia Rudin
Concept Whitening for Interpretable Image Recognition. Nature Machine Intelligence, accepted 2020.