Online Advertising

SECTIONS SEARCH

U.S. INTERNATIONAL 中文

SHOP MARCJACOBS.COM WATCHES

The New York Times

Monday, January 12, 2015 | Today's Paper | Video | 38°F | Hang Seng +0.51% ↑

World U.S. Politics New York Business Opinion Technology Science Health Sports Arts Style Food Home Travel Magazine Real Estate ALL

VOLVO EXPAND

Times Reporter Will Not Be Called to Testify in Leak Case

By MATT APUZZO 9:00 PM ET

The decision ends a seven-year legal fight over whether James Risen could be forced to name the sources of his reports on a botched C.I.A. operation.

39 Comments

The Jiaozhou Bay Bridge, which cost $2.3 billion, is the world's longest sea-crossing bridge.

1 of 7

The Opinion Pages

Choke First, Ask Questions Later

By THE EDITORIAL BOARD

A new report suggests that this disavowed tactic has never gone away and sometimes officers use it as a first, not last, resort.

- Editorial: United in Outrage
- Sheryl Sandberg and Adam Grant: Speaking While Female
- Taking Note: The Sony Hack and the Gender Pay Gap
- The Stone: Why Life Is Absurd

MENAGERIE

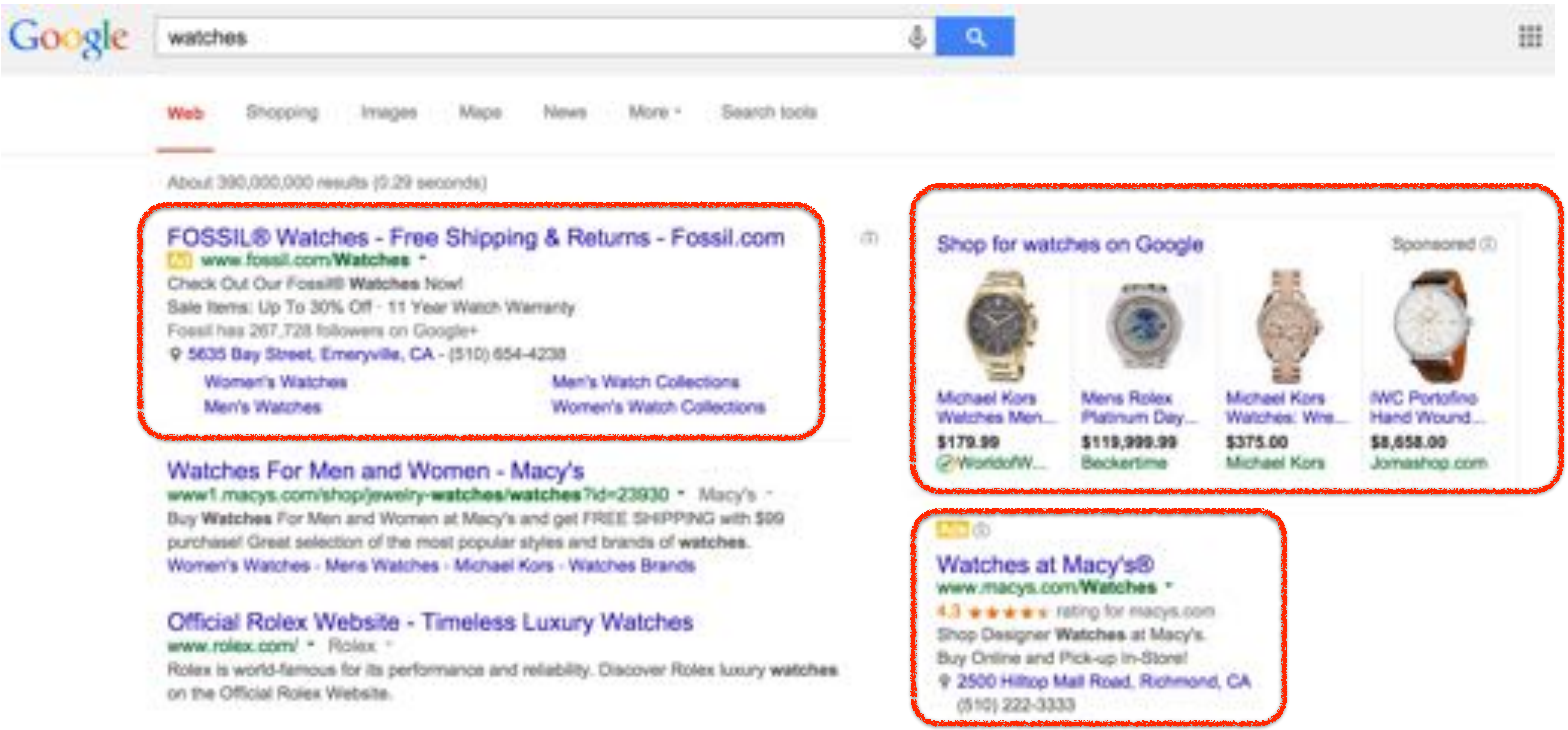
A Swarm in 'Dead City'

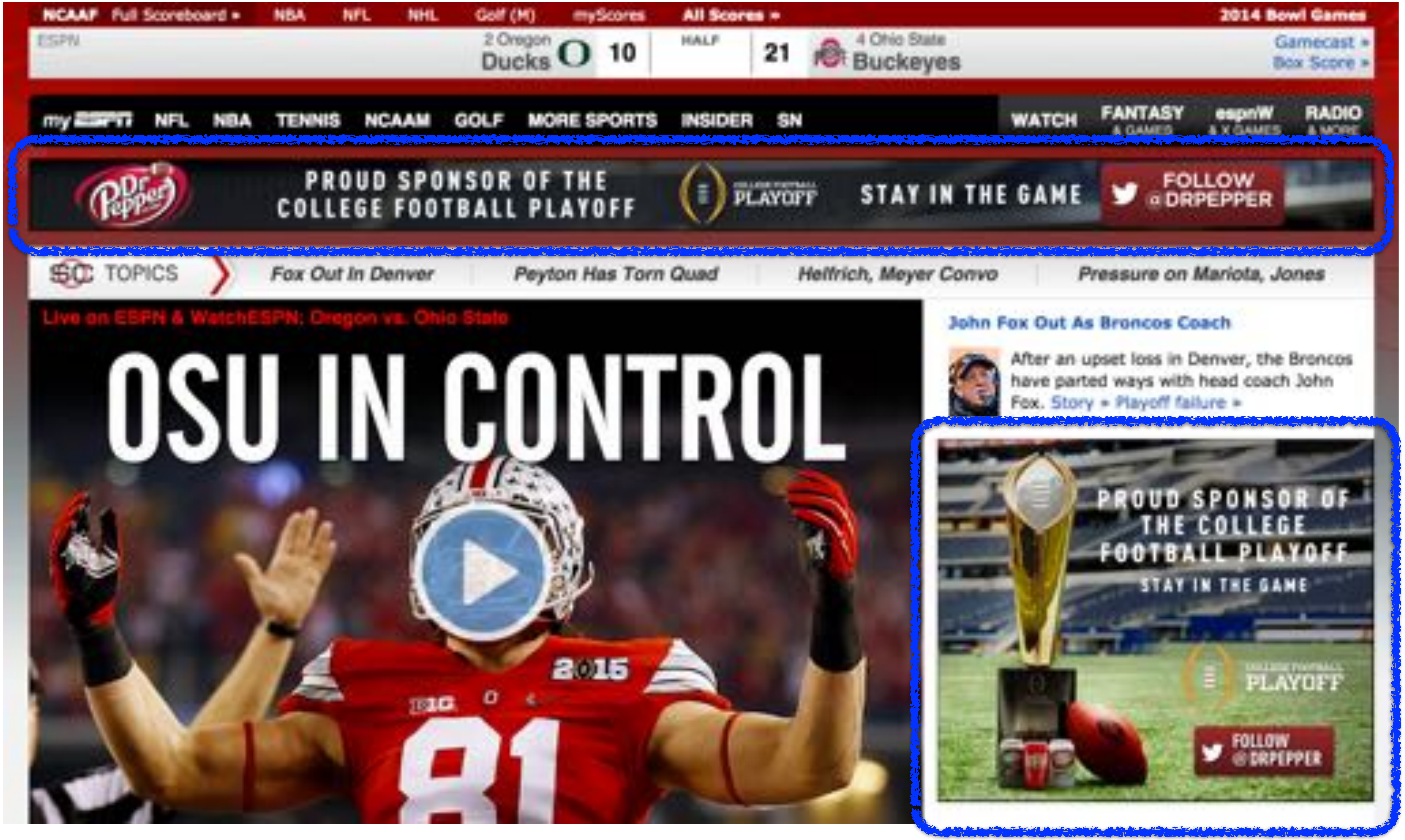
By GABRIELLE SELZ

At 14, I tried to run away. But millions of molting cicadas came between me and my freedom.

- Blow: Tamir Rice and the Value of Life
- Krugman: For the Love of Carbon
- Room for Debate: When Satire Cuts Both Ways
- Bruni, Douthat: Movies and Our Still-Wrenching History

France Orders


Google
watches
Web
Shopping
Images
Maps
News
More
Search tools
About 390,000,000 results (0.29 seconds)
FOSSIL® Watches - Free Shipping & Returns - Fossil.com
Ad www.fossil.com/Watches
Check Out Our Fossil® Watches Now!
Sale Items: Up To 30% Off · 11 Year Watch Warranty
Fossil has 267,728 followers on Google+
5635 Bay Street, Emeryville, CA - (510) 654-4238
Women's Watches
Men's Watch Collections
Men's Watches
Women's Watch Collections
Watches For Men and Women - Macy's
www1.macys.com/shop/jewelry-watches/watches?id=23930 · Macy's
Buy Watches For Men and Women at Macy's and get FREE SHIPPING with $99 purchase! Great selection of the most popular styles and brands of watches.
Women's Watches · Mens Watches · Michael Kors · Watches Brands
Official Rolex Website - Timeless Luxury Watches
www.rolex.com/ · Rolex
Rolex is world-famous for its performance and reliability. Discover Rolex luxury watches on the Official Rolex Website.
Shop for watches on Google
Sponsored
Michael Kors Watches Men...
$179.99
WorldofW...
Mens Rolex Platinum Day...
$119,999.99
Beckertime
Michael Kors Watches: Wre...
$375.00
Michael Kors
IWC Portofino Hand Wound...
$8,658.00
Jomashop.com
Ad
Watches at Macy's®
www.macys.com/Watches
4.3 ★★★★★ rating for macys.com
Shop Designer Watches at Macy's.
Buy Online and Pick-up In-Store!
2500 Hilltop Mall Road, Richmond, CA
(510) 222-3333


NCAAF Full Scoreboard NBA NFL NHL Golf (M) myScores All Scores 2014 Bowl Games
ESPN
2 Oregon Ducks 10
HALF
21
4 Ohio State Buckeyes
Gamecast
Box Score
NFL NBA TENNIS NCAAM GOLF MORE SPORTS INSIDER SN WATCH FANTASY RADIO
PROUD SPONSOR OF THE COLLEGE FOOTBALL PLAYOFF
PLAYOFF
STAY IN THE GAME
FOLLOW @DRPEPPER
TOPICS
Fox Out In Denver
Peyton Has Torn Quad
Helfrich, Meyer Convo
Pressure on Mariota, Jones
Live on ESPN & WatchESPN: Oregon vs. Ohio State
OSU IN CONTROL
John Fox Out As Broncos Coach
After an upset loss in Denver, the Broncos have parted ways with head coach John Fox. Story Playoff failure
PROUD SPONSOR OF THE COLLEGE FOOTBALL PLAYOFF
STAY IN THE GAME
PLAYOFF
FOLLOW @DRPEPPER

Online Advertising is Big Business

Multiple billion dollar industry

$43B in 2013 in USA, 17% increase over 2012
[PWC, Internet Advertising Bureau, April 2013]

Higher revenue in USA than cable TV and nearly the same as broadcast TV
[PWC, Internet Advertising Bureau, Oct 2013]

Large source of revenue for Google and other search engines

Canonical Scalable ML Problem

Problem is hard; we need all the data we can get!

- Success varies by type of online ad (banner, sponsor search, email, etc.) and by ad campaign, but can be less than 1% [Andrew Stern, iMedia Connection, 2010]

Lots of Data

- Lots of people use the internet
- Easy to gathered labeled data

A great success story for scalable ML

The Players

Publishers: NYTimes, Google, ESPN

- Make money displaying ads on their sites

Advertisers: Marc Jacobs, Fossil, Macy’s, Dr. Pepper

- Pay for their ads to be displayed on publisher sites
- They want to attract business

Matchmakers: Google, Microsoft, Yahoo

- Match publishers with advertisers
- In real-time (i.e., as a specific user visits a website)

Why Advertisers Pay?

Impressions

- Get message to target audience
- e.g., brand awareness campaign

Performance

- Get users to do something
- e.g., click on ad (pay-per-click) ← Most common
- e.g., buy something or join a mailing list

Efficient Matchmaking

Idea: Predict probability that user will click each ad and choose ads to maximize probability

- Estimate $\mathbb{P}$(click | predictive features)
- Conditional probability: probability **given** predictive features

Predictive features

- Ad's historical performance
- Advertiser and ad content info
- Publisher info
- User info (e.g., search / click history)

Publishers Get Billions of Impressions Per Day

But, data is high-dimensional, sparse, and skewed

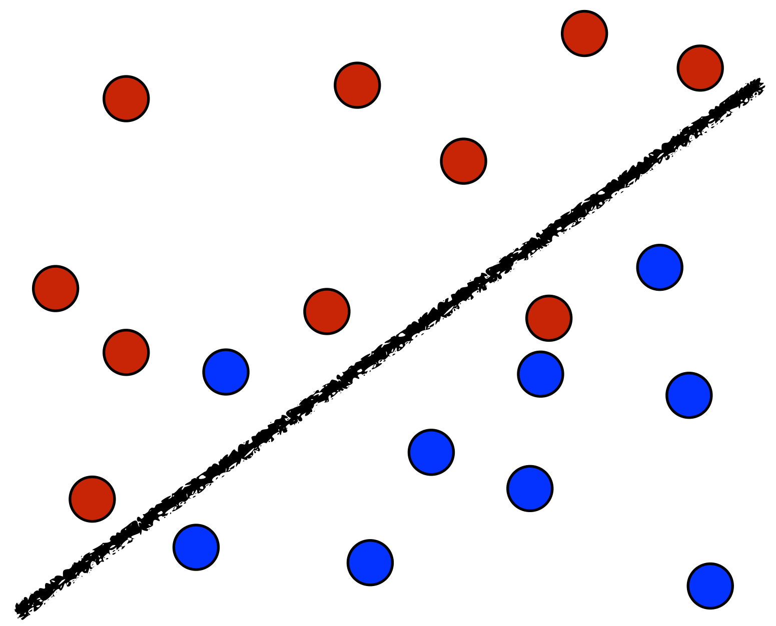
- Hundreds of millions of online users
- Millions of unique publisher pages to display ads
- Millions of unique ads to display
- Very few ads get clicked by users

Massive datasets are crucial to tease out signal

Goal: Estimate $\mathbb{P}(\text{click} \mid \text{user, ad, publisher info})$
Given: Massive amounts of labeled data

Linear Classification and Logistic Regresssion

Classification

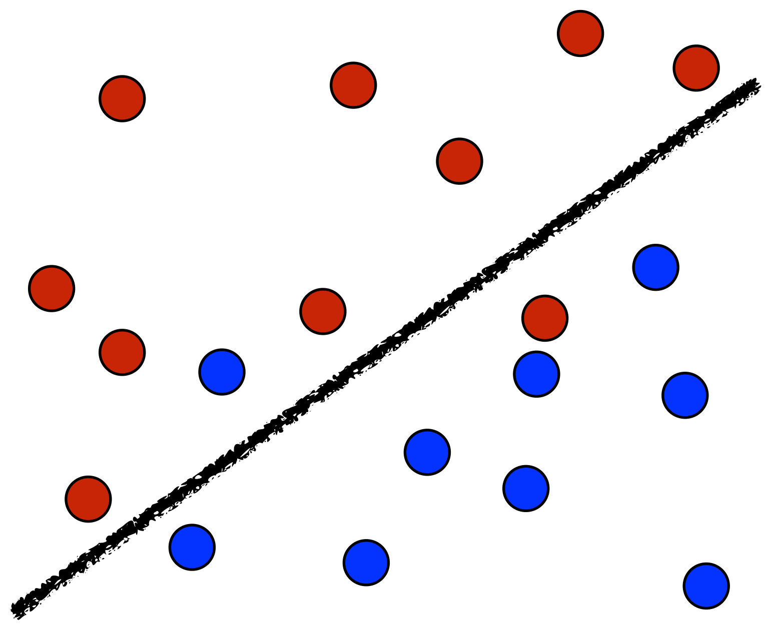


Goal: Learn a mapping from observations to discrete labels given a set of training examples (supervised learning)

Example: Spam Classification

- Observations are emails
- Labels are {spam, not-spam} (Binary Classification)
- Given a set of labeled emails, we want to predict whether a new email is spam or not-spam

Classification

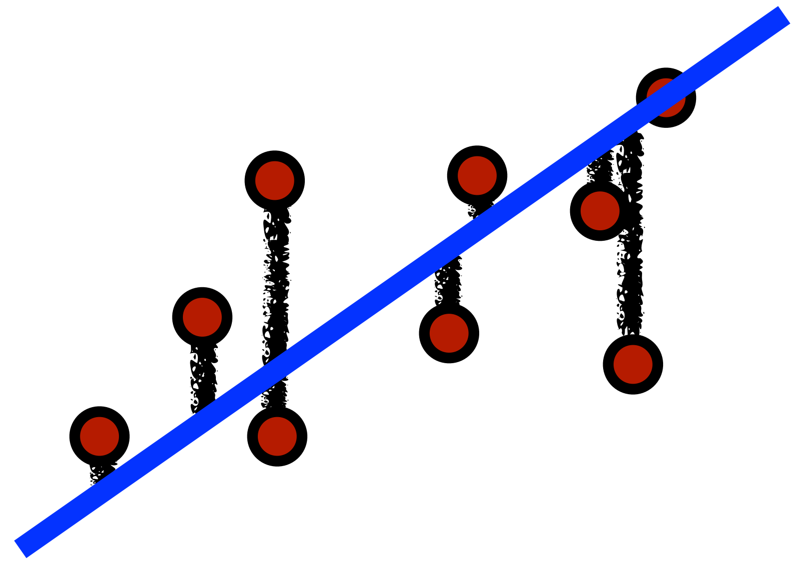


Goal: Learn a mapping from observations to discrete labels given a set of training examples (supervised learning)

Example: Click-through Rate Prediction

- Observations are user-ad-publisher triples
- Labels are {not-click, click} (Binary Classification)
- Given a set of labeled observations, we want to predict whether a new user-ad-publisher triple will result in a click

Reminder: Linear Regression



Example: Predicting shoe size from height, gender, and weight

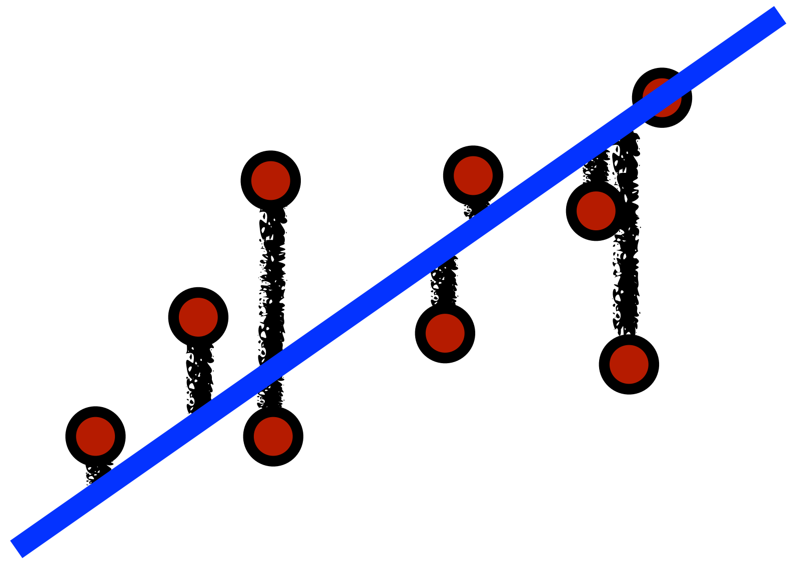
For each observation we have a feature vector, $\mathbf{x}$, and label, y

$$\mathbf{x}^\top = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}$$

We assume a *linear* mapping between features and label:

$$y \approx w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3$$

Reminder: Linear Regression



Example: Predicting shoe size from height, gender, and weight

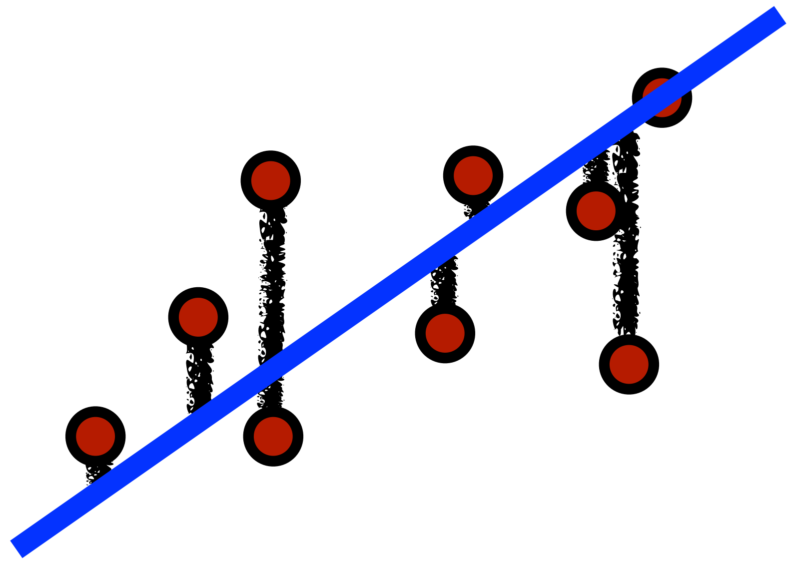
We can augment the feature vector to incorporate offset:

$$\mathbf{x}^\top = \begin{bmatrix} 1 & x_1 & x_2 & x_3 \end{bmatrix}$$

We can then rewrite this linear mapping as scalar product:

$$y \approx \hat{y} = \sum_{i=0}^{3} w_i x_i = \mathbf{w}^\top \mathbf{x}$$

Why a Linear Mapping?



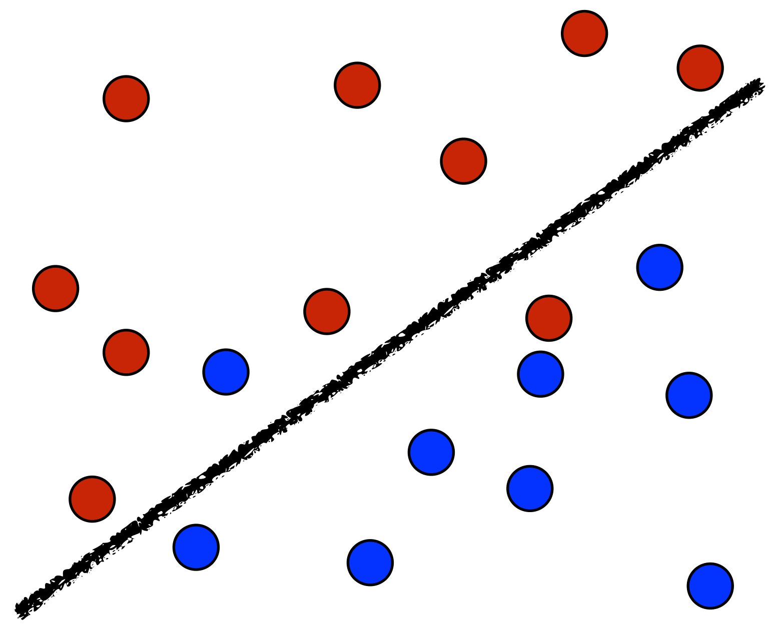
Simple

Often works well in practice

Can introduce complexity via feature extraction

Can we do something similar for classification?

Linear Regression ⇒ Linear Classifier



Example: Predicting rain from temperature, cloudiness, and humidity

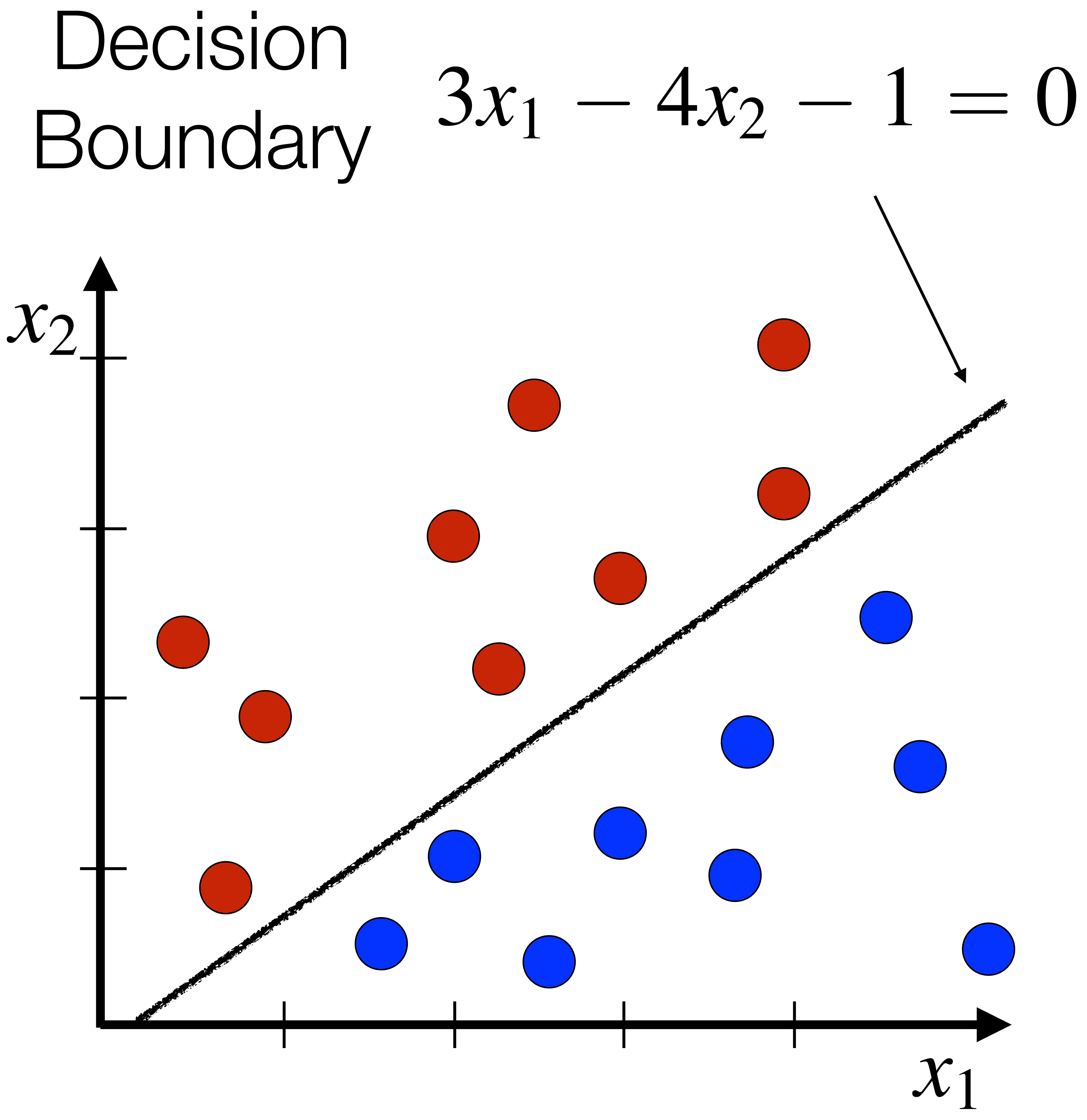
Use the same feature representation: $\mathbf{x}^\top = \begin{bmatrix} 1 & x_1 & x_2 & x_3 \end{bmatrix}$

How can we make class predictions?

- {not-rain, rain}, {not-spam, spam}, {not-click, click}
- We can threshold by sign

$$\hat{y} = \sum_{i=0}^{3} w_i x_i = \mathbf{w}^\top \mathbf{x} \implies \hat{y} = \mathrm{sign}(\mathbf{w}^\top \mathbf{x})$$

Linear Classifier Decision Boundary



Imagine $\mathbf{w}^\top = \begin{bmatrix} -1 & 3 & -4 \end{bmatrix}$

$\mathbf{x}^\top = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} : \mathbf{w}^\top \mathbf{x} = -7$

$\mathbf{x}^\top = \begin{bmatrix} 1 & 2 & 1 \end{bmatrix} : \mathbf{w}^\top \mathbf{x} = 1$

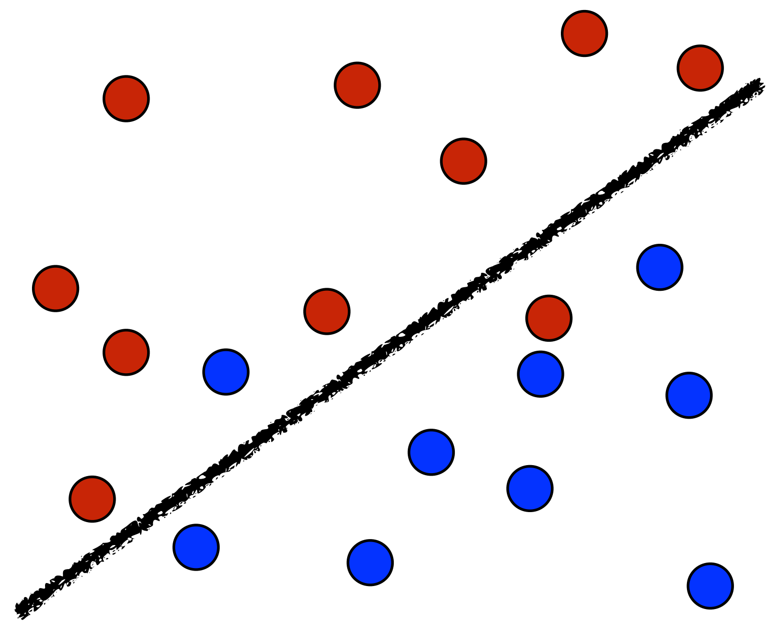
$\mathbf{x}^\top = \begin{bmatrix} 1 & 5 & .5 \end{bmatrix} : \mathbf{w}^\top \mathbf{x} = 12$

$\mathbf{x}^\top = \begin{bmatrix} 1 & 3 & 2.5 \end{bmatrix} : \mathbf{w}^\top \mathbf{x} = -2$

Let's interpret this rule: $\hat{y} = \mathbf{sign}(\mathbf{w}^\top \mathbf{x})$

- $\hat{y} = 1 : \mathbf{w}^\top \mathbf{x} > 0$
- $\hat{y} = -1 : \mathbf{w}^\top \mathbf{x} < 0$
- Decision boundary: $\mathbf{w}^\top \mathbf{x} = 0$

Evaluating Predictions



Regression: can measure ‘closeness’ between label and prediction

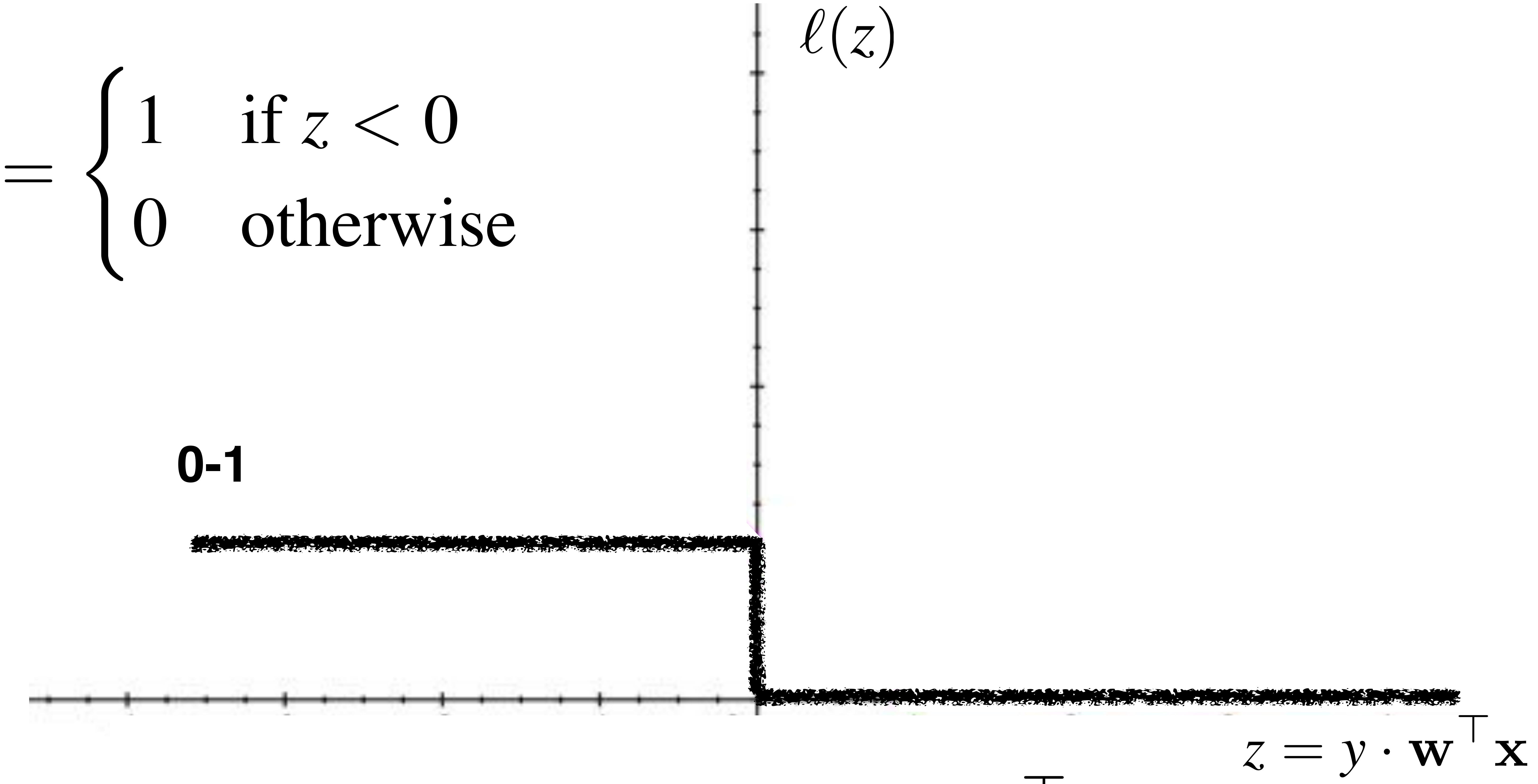
- Song year prediction: better to be off by a year than by 20 years
- Squared loss: $(y - \hat{y})^2$

Classification: Class predictions are discrete

- 0-1 loss: Penalty is 0 for correct prediction, and 1 otherwise

0/1 Loss Minimization

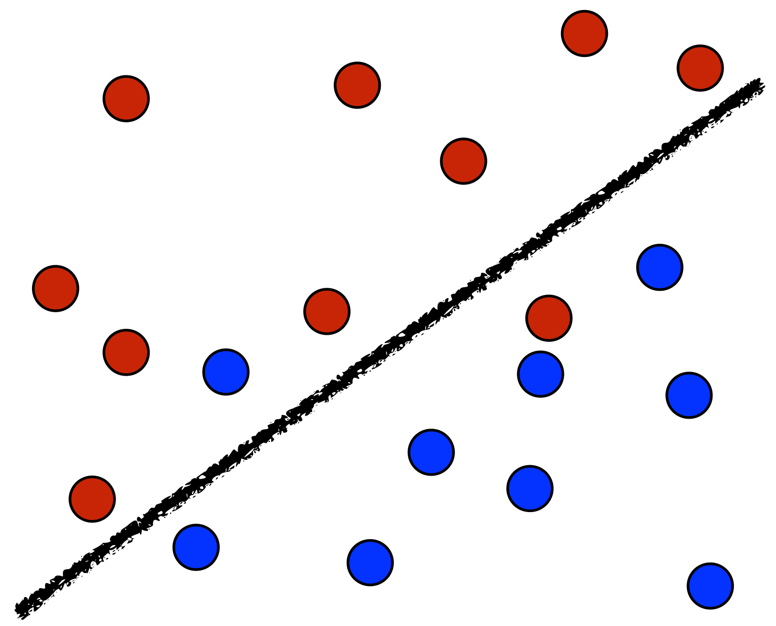
$$\ell_{0/1}(z) = \begin{cases} 1 & \text{if } z < 0 \\ 0 & \text{otherwise} \end{cases}$$



Let $y \in \{-1, 1\}$ and define $z = y \cdot \mathbf{w}^\top \mathbf{x}$

z is positive if y and $\mathbf{w}^\top \mathbf{x}$ have same sign, negative otherwise

How Can We Learn Model ($\mathbf{w}$)?



Assume we have n training points, where $\mathbf{x}^{(i)}$ denotes the ith point

Recall two earlier points:

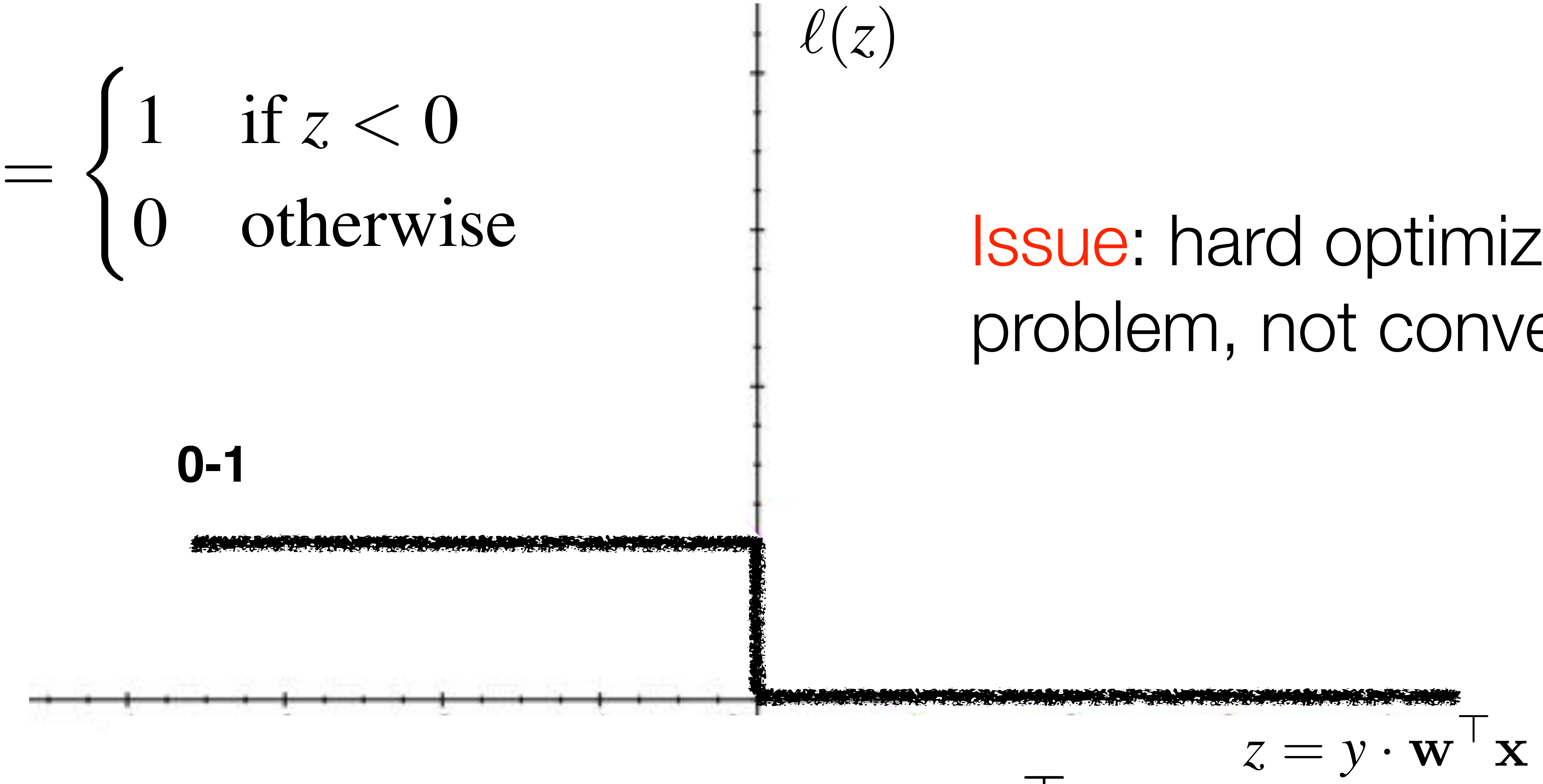
- *Linear* assumption: $\hat{y} = \mathbf{sign}(\mathbf{w}^\top \mathbf{x})$
- We use 0-1 *loss*: $\ell_{0/1}(z)$

Idea: Find $\mathbf{w}$ that minimizes average 0-1 loss over training points:

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \ell_{0/1}\left(y^{(i)} \cdot \mathbf{w}^\top \mathbf{x}^{(i)}\right)$$

0/1 Loss Minimization

$$\ell_{0/1}(z) = \begin{cases} 1 & \text{if } z < 0 \\ 0 & \text{otherwise} \end{cases}$$

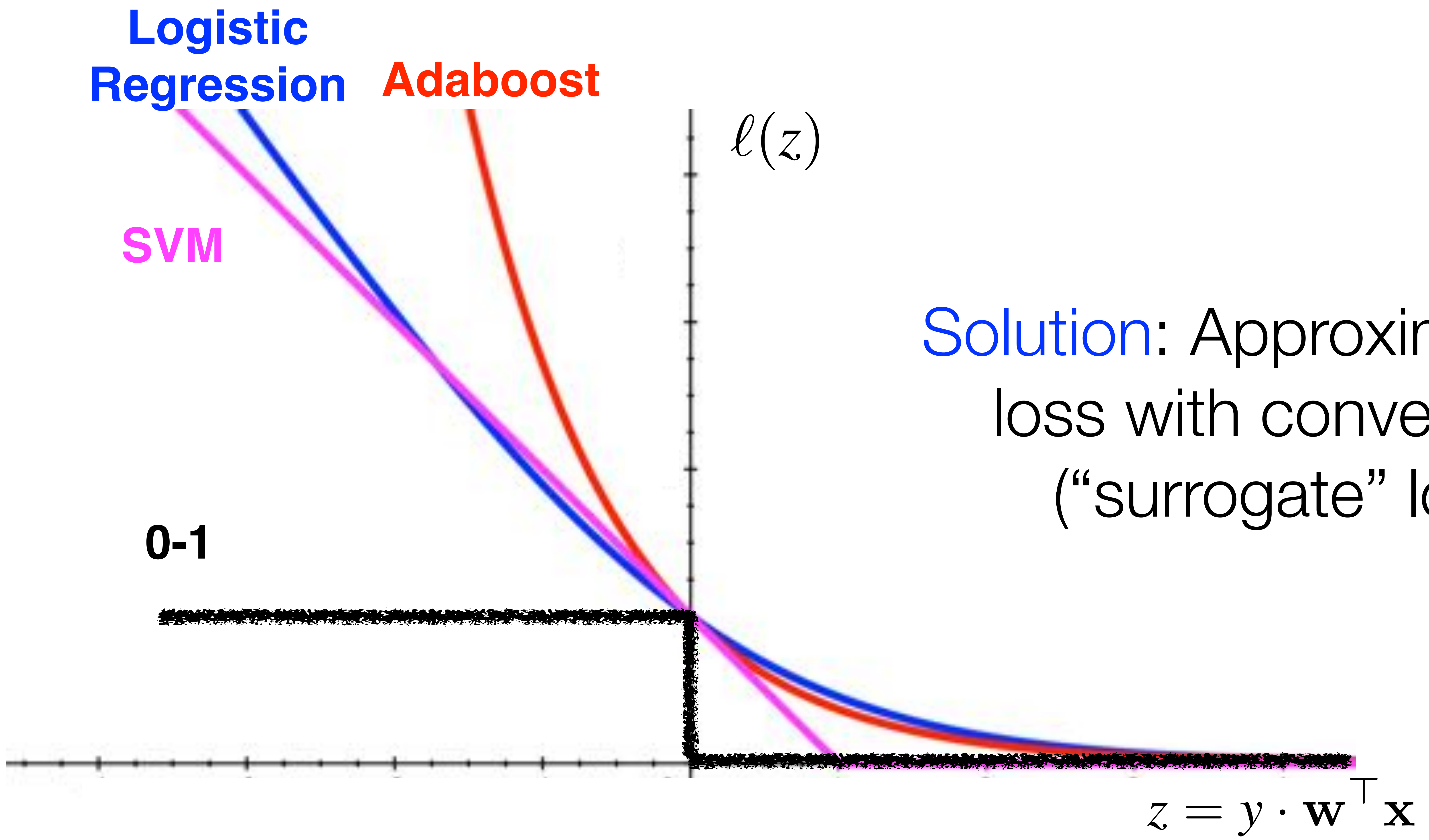


Issue: hard optimization problem, not convex!

Let $y \in \{-1, 1\}$ and define $z = y \cdot \mathbf{w}^\top \mathbf{x}$

z is positive if y and $\mathbf{w}^\top \mathbf{x}$ have same sign, negative otherwise

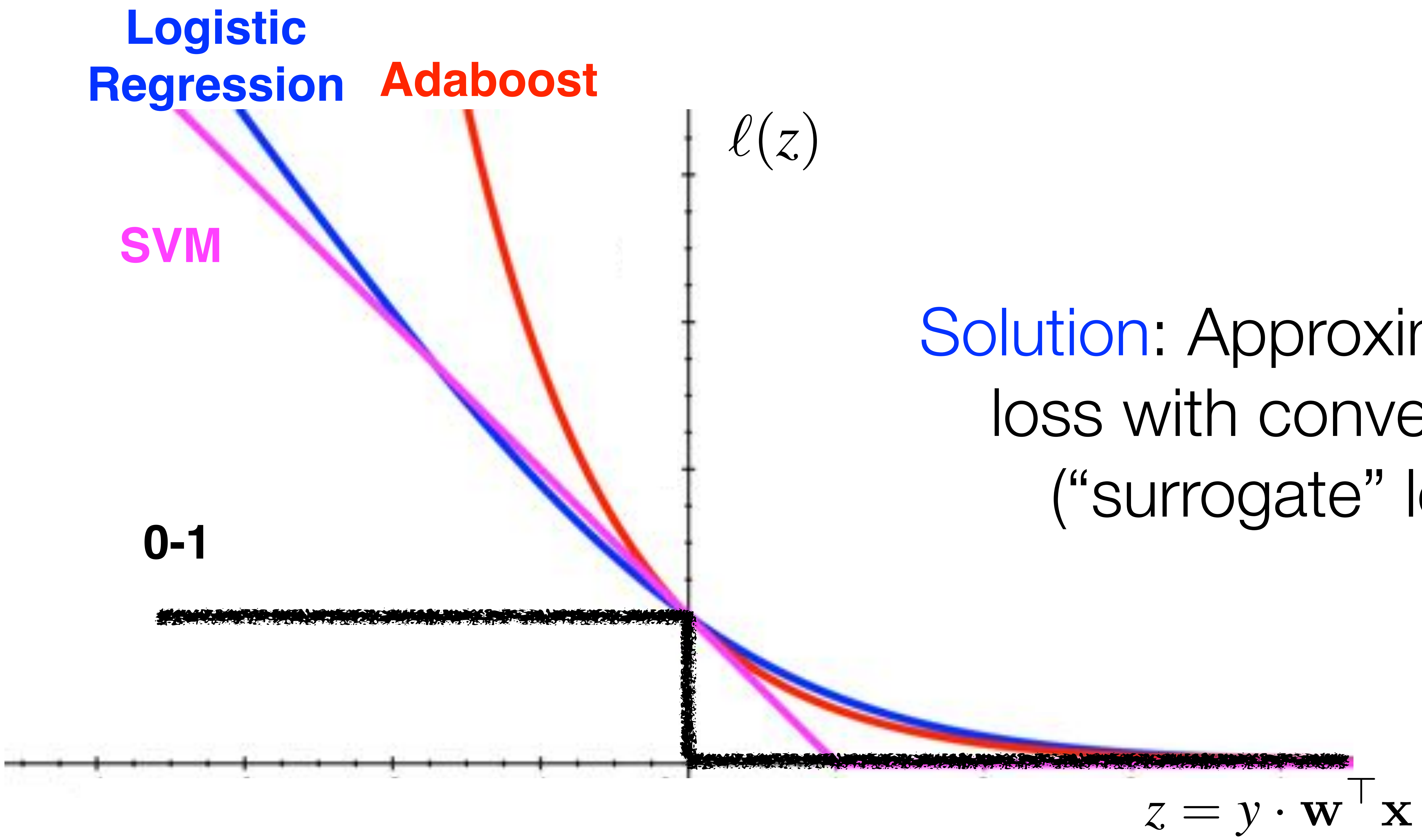
Approximate 0/1 Loss



Solution: Approximate 0/1 loss with convex loss ("surrogate" loss)

SVM (hinge), Logistic regression (logistic), Adaboost (exponential)

Approximate 0/1 Loss



Solution: Approximate 0/1 loss with convex loss ("surrogate" loss)

Logistic loss (logloss): $\ell_{log}(z) = \log(1 + e^{-z})$

Logistic Regression Optimization

Logistic Regression: Learn mapping ($\mathbf{w}$) that minimizes logistic loss on training data

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \ell_{log}\left(y^{(i)} \cdot \mathbf{w}^\top \mathbf{x}^{(i)}\right)$$

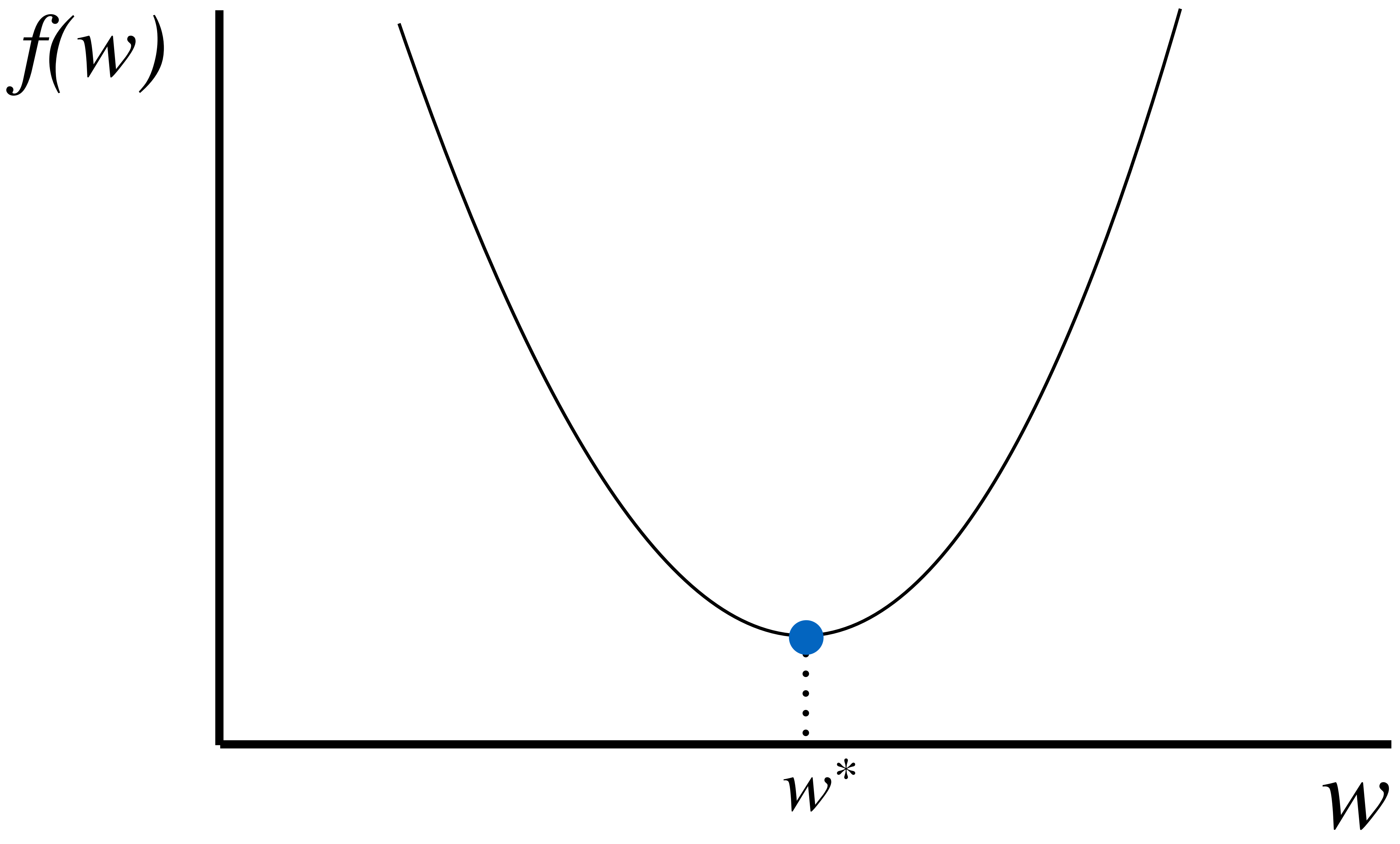
- Convex
- Closed form solution doesn’t exist

Logistic Regression Optimization

Goal: Find $\mathbf{w}^*$ that minimizes

$$f(\mathbf{w}) = \sum_{i=1}^{n} \ell_{log}\left(y^{(i)} \cdot \mathbf{w}^\top \mathbf{x}^{(i)}\right)$$

- Can solve via Gradient Descent



Step Size

Update Rule: $\mathbf{w}_{i+1} = \mathbf{w}_i - \alpha \nabla f(\mathbf{w})$

$$\sum_{j=1}^{n} \left[1 - \frac{1}{1 + \exp(-y^{(j)} \mathbf{w}_i^\top \mathbf{x}^{(j)})}\right]\left(-y^{(j)} \mathbf{x}^{(j)}\right)$$

Gradient

Logistic Regression Optimization

Regularized

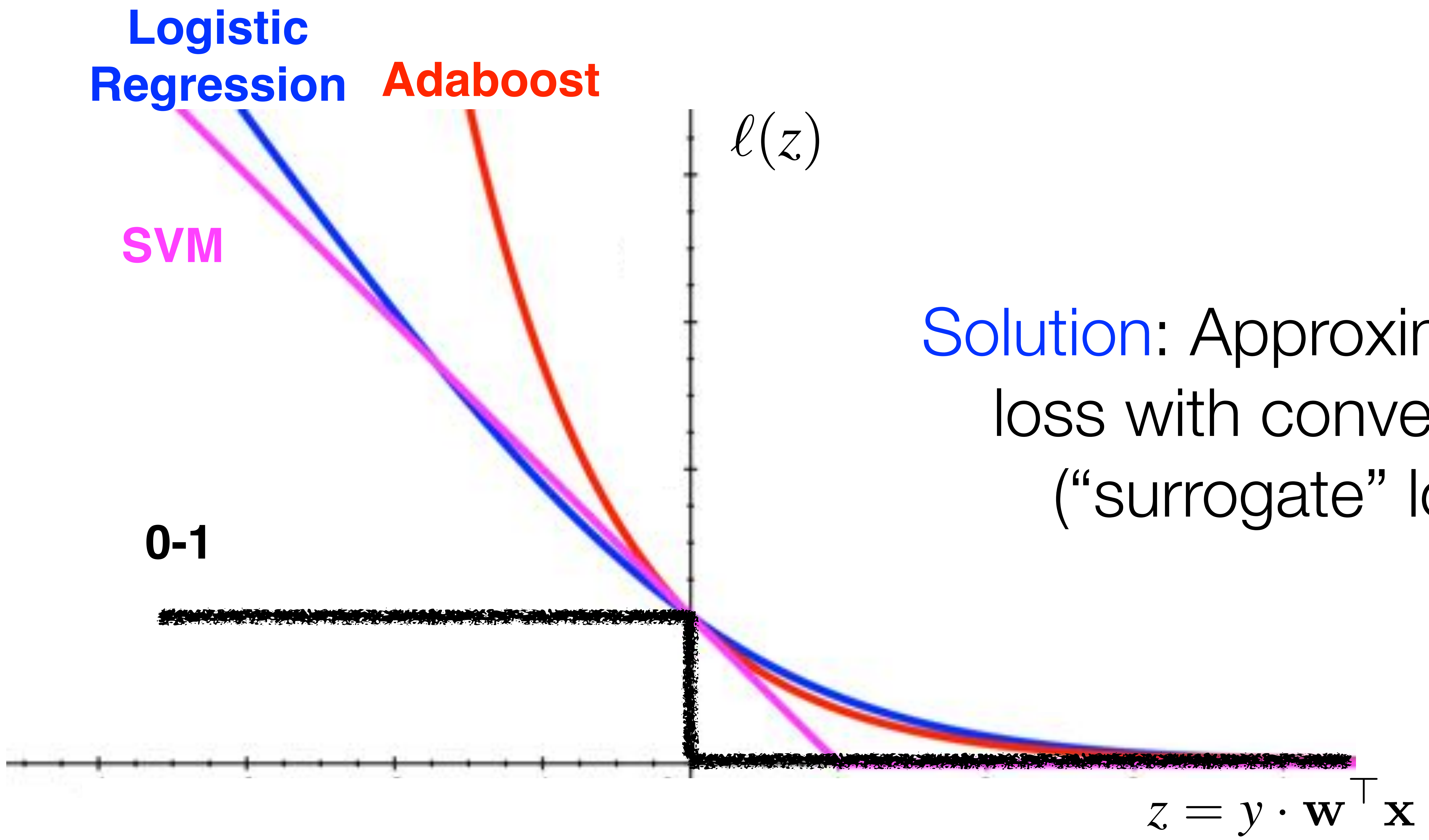
Logistic Regression: Learn mapping ($\mathbf{w}$) that minimizes logistic loss on training data with a regularization term

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \overbrace{\ell_{log}\left(y^{(i)} \cdot \mathbf{w}^\top \mathbf{x}^{(i)}\right)}^{\text{Training LogLoss}} + \overbrace{\lambda ||\mathbf{w}||_2^2}^{\text{Model Complexity}}$$

- Convex
- Closed form solution doesn't exist
- Can add regularization term (as in ridge regression)

Logistic Regression: Probabilistic Interpretation

Approximate 0/1 Loss



Solution: Approximate 0/1 loss with convex loss ("surrogate" loss)

SVM (hinge), Logistic regression (logistic), Adaboost (exponential)

Probabilistic Interpretation

Goal: Model conditional probability: $\mathbb{P}[\, y = 1 \mid \mathbf{x} \,]$

Example: Predict rain from **t**emperature, **c**loudiness, **h**umidity

- $\mathbb{P}[\, y = \text{rain} \mid \text{t} = 14°\text{F},\ \text{c} = \text{LOW},\ \text{h} = 2\% \,] = .05$
- $\mathbb{P}[\, y = \text{rain} \mid \text{t} = 70°\text{F},\ \text{c} = \text{HIGH},\ \text{h} = 95\% \,] = .9$

Example: Predict click from ad's **h**istorical performance, user's click **f**requency, and publisher page's **r**elevance

- $\mathbb{P}[\, y = \text{click} \mid \text{h} = \text{GOOD},\ \text{f} = \text{HIGH},\ \text{r} = \text{HIGH} \,] = .1$
- $\mathbb{P}[\, y = \text{click} \mid \text{h} = \text{BAD},\ \text{f} = \text{LOW},\ \text{r} = \text{LOW} \,] = .001$

Probabilistic Interpretation

Goal: Model conditional probability: $\mathbb{P}[\, y = 1 \,|\, \mathbf{x} \,]$

First thought: $\mathbb{P}[\, y = 1 \,|\, \mathbf{x} \,] \cancel{=} \mathbf{w}^\top \mathbf{x}$

- Linear regression returns any real number, but probabilities range from 0 to 1!

How can we transform or 'squash' its output?

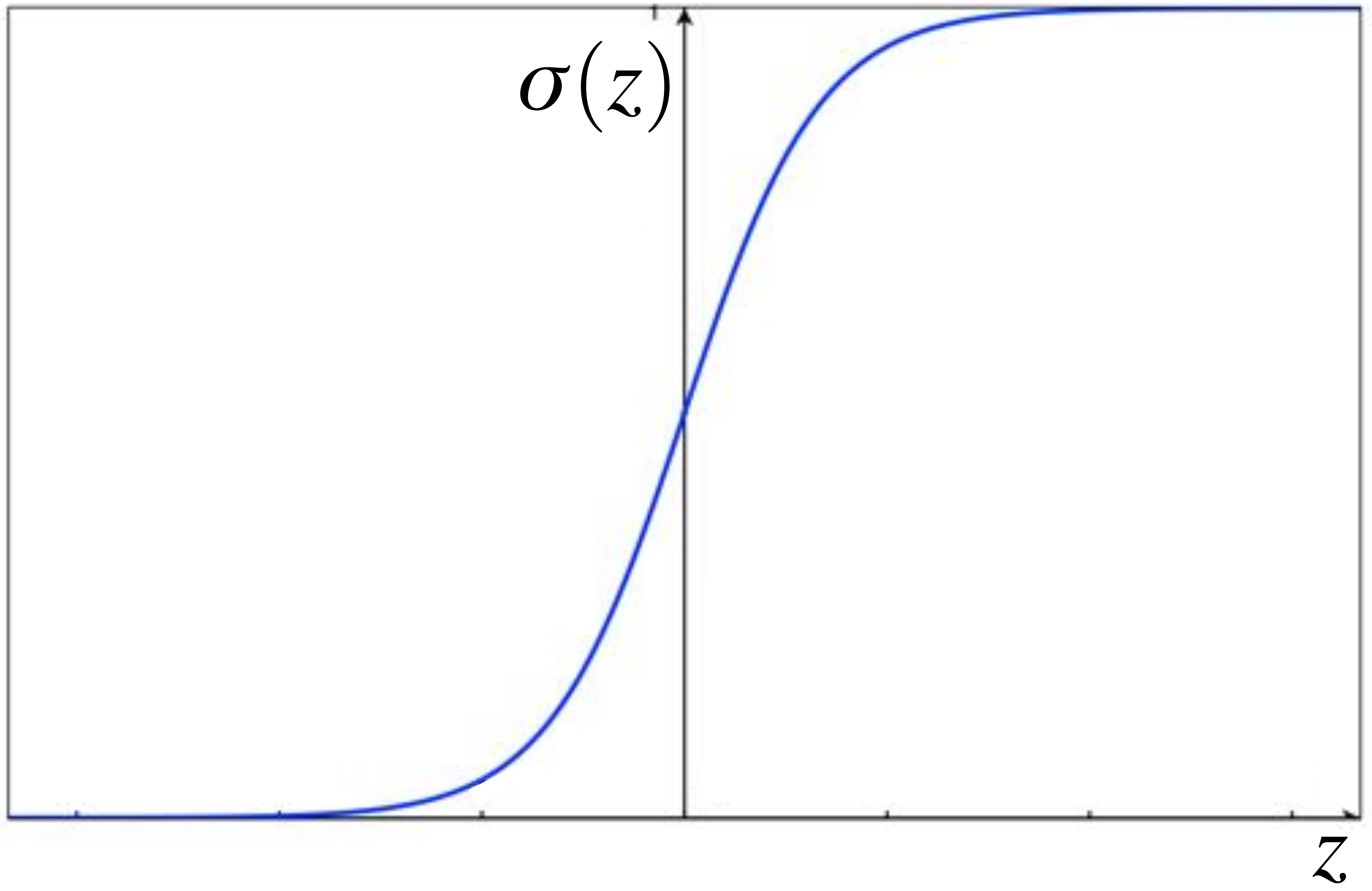
- Use logistic (or sigmoid) function:

$$\mathbb{P}[\, y = 1 \,|\, \mathbf{x} \,] = \sigma(\mathbf{w}^\top \mathbf{x})$$

Logistic Function

Maps real numbers to $[0, 1]$

- Large positive inputs $\Rightarrow 1$
- Large negative inputs $\Rightarrow 0$



$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

Probabilistic Interpretation

Goal: Model conditional probability: $\mathbb{P}[\, y = 1 \,|\, \mathbf{x} \,]$

Logistic regression uses logistic function to model this conditional probability

- $\mathbb{P}[\, y = 1 \,|\, \mathbf{x} \,] = \sigma(\mathbf{w}^\top \mathbf{x})$
- $\mathbb{P}[\, y = 0 \,|\, \mathbf{x} \,] = 1 - \sigma(\mathbf{w}^\top \mathbf{x})$

For notational convenience we now define $y \in \{0, 1\}$

How Do We Use Probabilities?

To make class predictions, we need to convert probabilities to values in $\{0, 1\}$

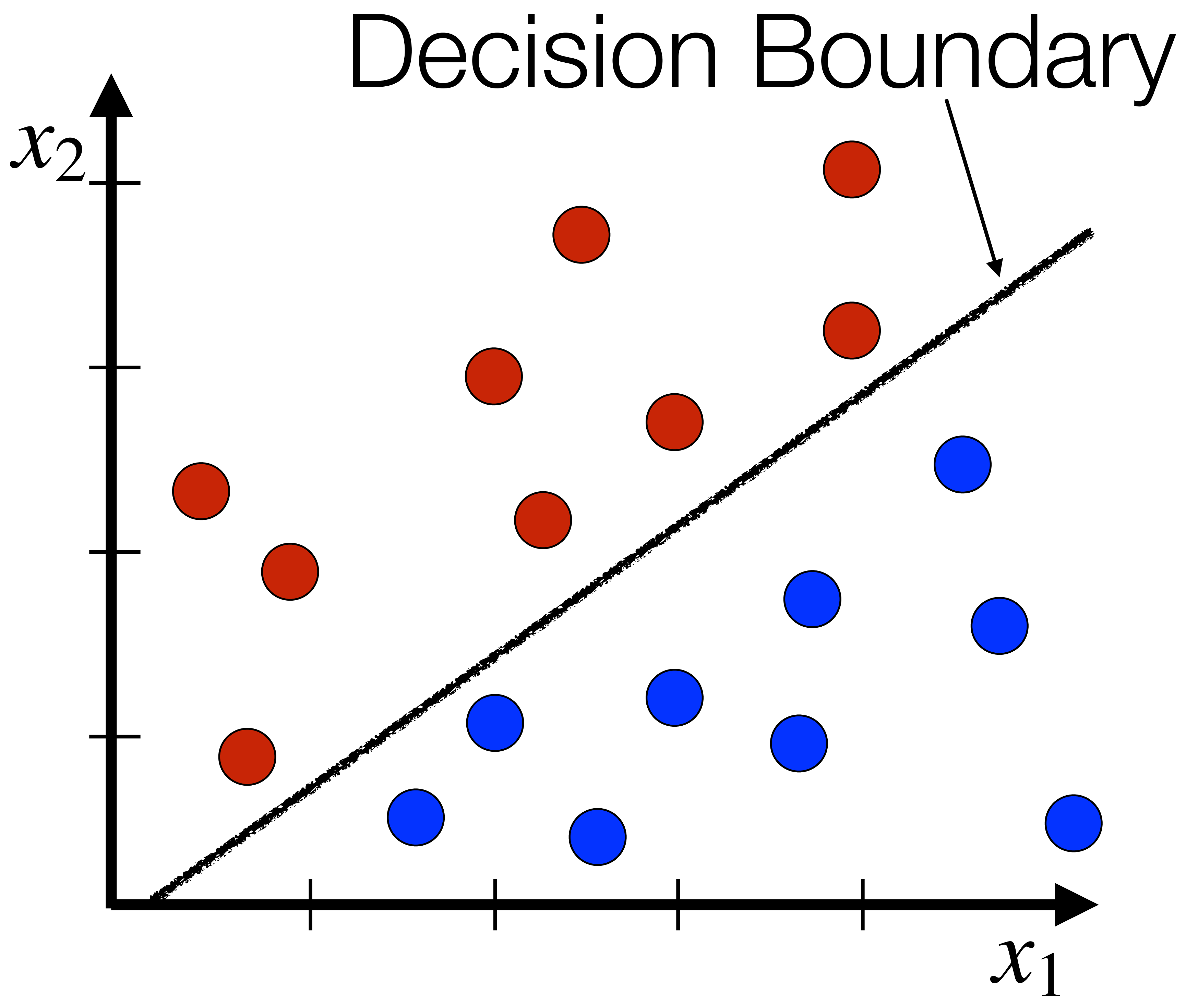
We can do this by setting a threshold on the probabilities

- Default threshold is 0.5
- $\mathbb{P}[\,y = 1 \,|\, \mathbf{x}\,] > 0.5 \implies \hat{y} = 1$

Example: Predict rain from **t**emperature, **c**loudiness, **h**umidity

- $\mathbb{P}[\,y = \text{rain} \,|\, \text{t} = 14°\text{F},\ \text{c} = \text{LOW},\ \text{h} = 2\%\,] = .05$ $\quad\hat{y} = 0$
- $\mathbb{P}[\,y = \text{rain} \,|\, \text{t} = 70°\text{F},\ \text{c} = \text{HIGH},\ \text{h} = 95\%\,] = .9$ $\quad\hat{y} = 1$

Connection with Decision Boundary?



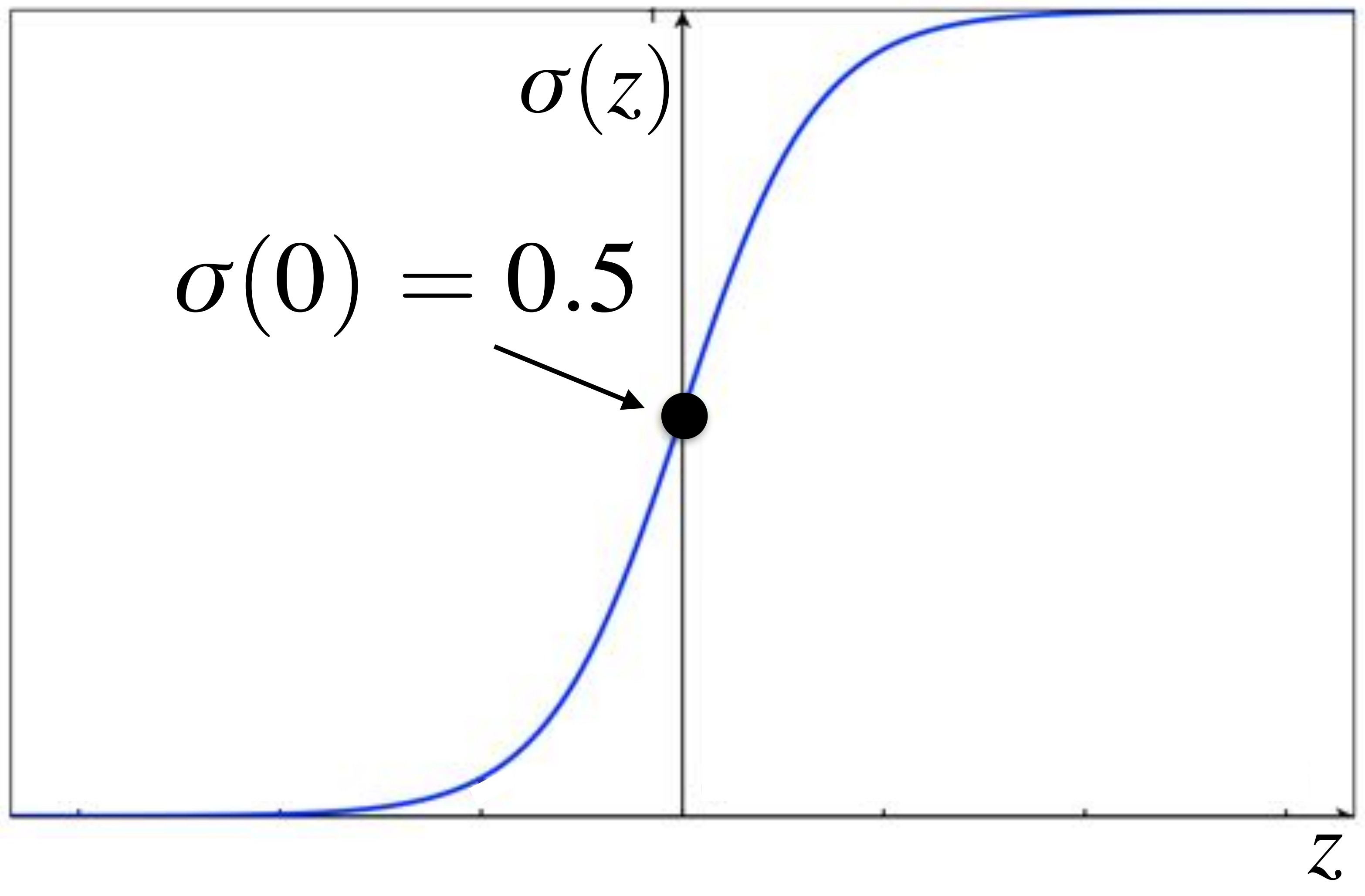
Threshold by sign to make class predictions: $\hat{y} = \mathbf{sign}(\mathbf{w}^\top \mathbf{x})$

- $\hat{y} = 1 : \mathbf{w}^\top \mathbf{x} > 0$
- $\hat{y} = 0 : \mathbf{w}^\top \mathbf{x} < 0$
- decision boundary: $\mathbf{w}^\top \mathbf{x} = 0$

How does this compare with thresholding probability?

- $\mathbb{P}[\, y = 1 \,|\, \mathbf{x} \,] = \sigma(\mathbf{w}^\top \mathbf{x}) > 0.5 \Longrightarrow \hat{y} = 1$

Connection with Decision Boundary?



$$\mathbf{w}^\top \mathbf{x} = 0 \iff \sigma(\mathbf{w}^\top \mathbf{x}) = 0.5$$

Threshold by sign to make class predictions: $\hat{y} = \mathbf{sign}(\mathbf{w}^\top \mathbf{x})$

- $\hat{y} = 1 : \mathbf{w}^\top \mathbf{x} > 0$
- $\hat{y} = 0 : \mathbf{w}^\top \mathbf{x} < 0$
- decision boundary: $\mathbf{w}^\top \mathbf{x} = 0$

How does this compare with thresholding probability?

- $\mathbb{P}[y = 1 \,|\, \mathbf{x}] = \sigma(\mathbf{w}^\top \mathbf{x}) > 0.5 \implies \hat{y} = 1$
- With threshold of 0.5, the decision boundaries are identical!

Using Probabilistic Predictions

How Do We Use Probabilities?

To make class predictions, we need to convert probabilities to values in $\{0, 1\}$

We can do this by setting a threshold on the probabilities

- Default threshold is 0.5
- $\mathbb{P}[\, y = 1 \,|\, \mathbf{x} \,] > 0.5 \implies \hat{y} = 1$

Example: Predict rain from **t**emperature, **c**loudiness, **h**umidity

- $\mathbb{P}[\, y = \text{rain} \,|\, \text{t} = 14^\circ\text{F, c} = \text{LOW, h} = 2\% \,] = .05$ $\quad \hat{y} = 0$
- $\mathbb{P}[\, y = \text{rain} \,|\, \text{t} = 70^\circ\text{F, c} = \text{HIGH, h} = 95\% \,] = .9$ $\quad \hat{y} = 1$

Setting different thresholds

In spam detection application, we model $\mathbb{P}[\, y = \mathbf{spam} \mid \mathbf{x} \,]$

Two types of error

- Classify a not-spam email as spam (*false positive, FP*)
- Classify a spam email as not-spam (*false negative, FN*)

Can argue that false positives are more harmful than false negatives

- Worse to miss an important email than to have to delete spam

We can use a threshold greater than 0.5 to be more ‘conservative’

ROC Plots: Measuring Varying Thresholds

ROC plot displays FPR vs TPR

- Top left is perfect
- Dotted Line is random prediction (i.e., biased coin flips)

Can classify at various thresholds (T)

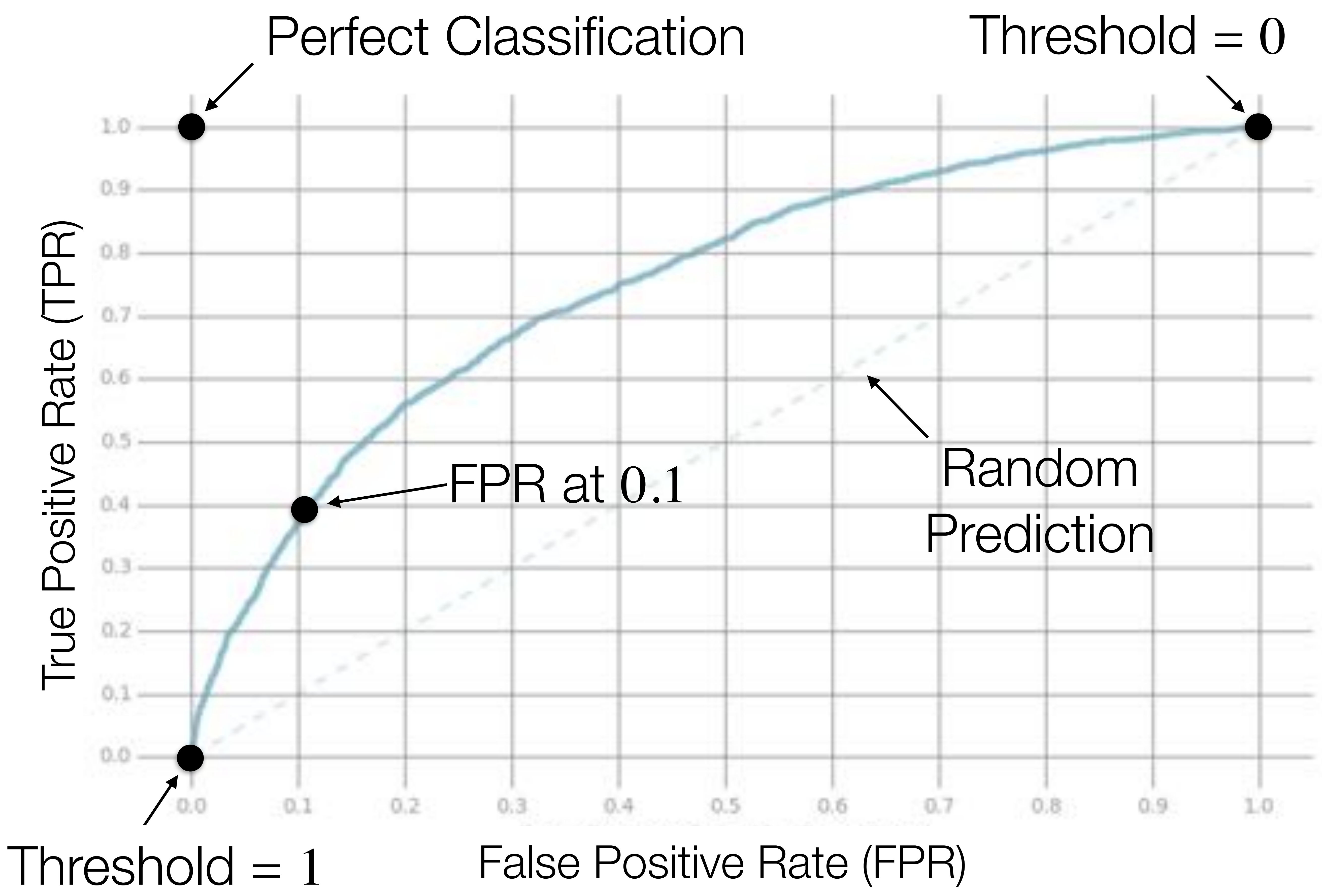
T = 0: Everything is spam

- TPR = 1, but FPR = 1

T = 1: Nothing is spam

- FPR = 0, but TPR = 0

We can tradeoff between TPR/FPR



FPR: % not-spam predicted as spam
TPR: % spam predicted as spam

Working Directly with Probabilities

Example: Predict click from ad's **h**istorical performance, user's click **f**requency, and publisher page's **r**elevance

- $\mathbb{P}[\, y = \text{click} \,|\, \text{h} = \text{GOOD},\ \text{f} = \text{HIGH},\ \text{r} = \text{HIGH} \,] = .1 \quad \hat{y} = 0$
- $\mathbb{P}[\, y = \text{click} \,|\, \text{h} = \text{BAD},\ \text{f} = \text{LOW},\ \text{r} = \text{LOW} \,] = .001 \quad \hat{y} = 0$

Success can be less than 1% [Andrew Stern, iMedia Connection, 2010]

Probabilities provide more granular information

- Confidence of prediction
- Useful when combining predictions with other information

In such cases, we want to evaluate probabilities directly

- Logistic loss makes sense for evaluation!

Logistic Loss

$$\ell_{log}(p, y) = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1-p) & \text{if } y = 0 \end{cases}$$

When $y = 1$, we want $p = 1$

- No penalty at 1
- Increasing penalty away from 1

Similar logic when $y = 0$