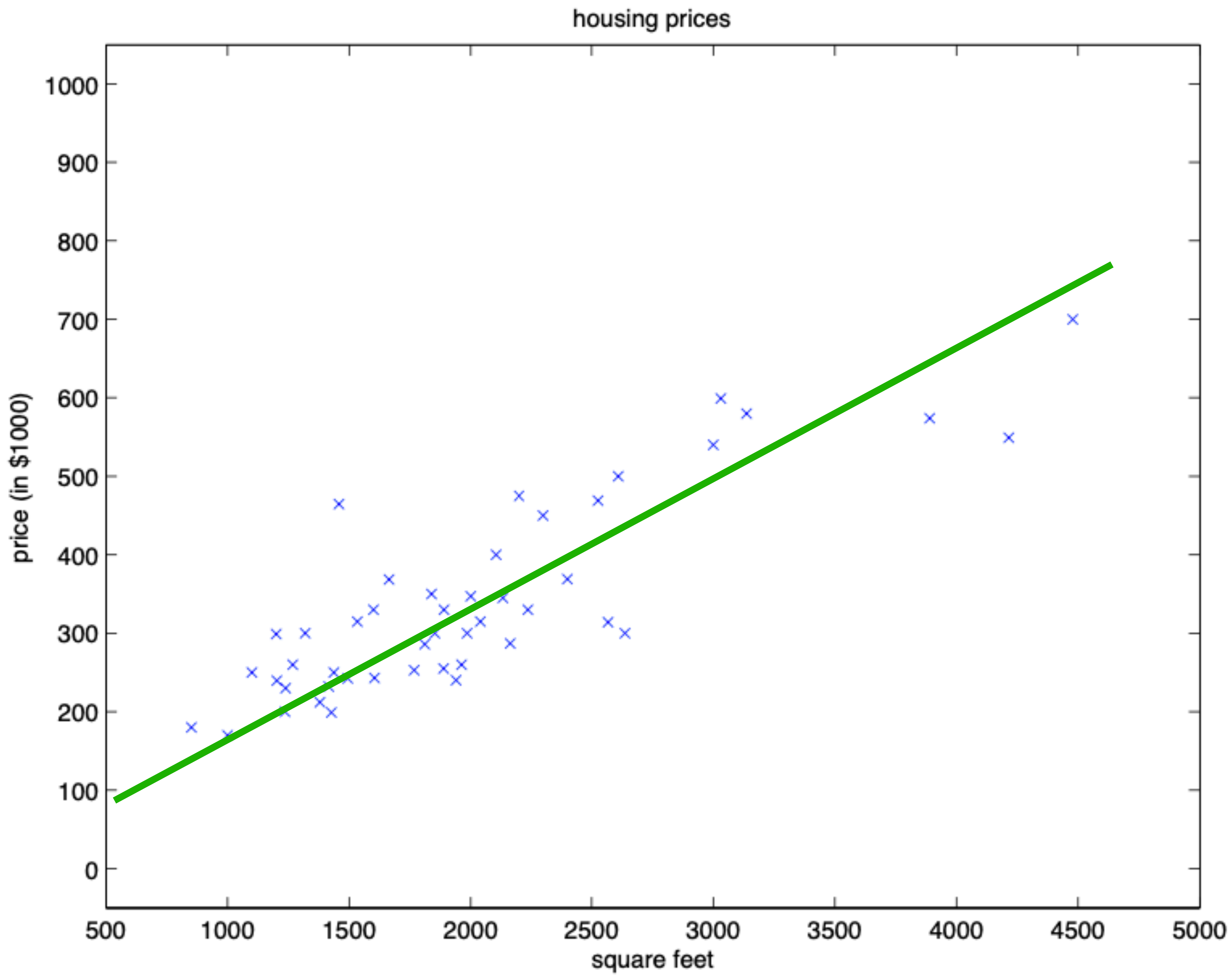
Empirical Risk Minimization

Announcements

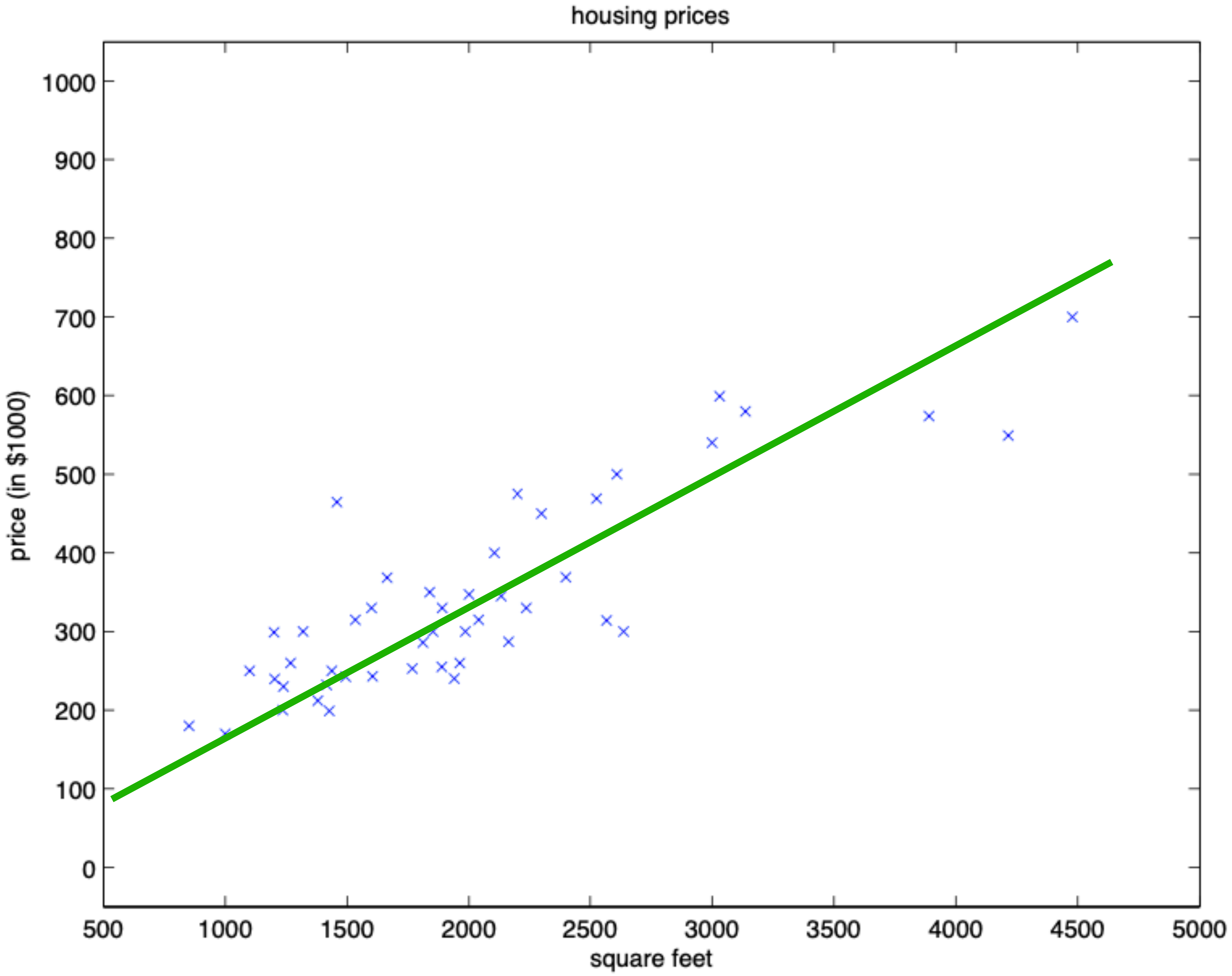
Recap on Linear Regression

Given dataset $\mathscr{D} = \{x_i, y_i\}, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$



Recap on Linear Regression

Given dataset $\mathscr{D} = \{x_i, y_i\}, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$

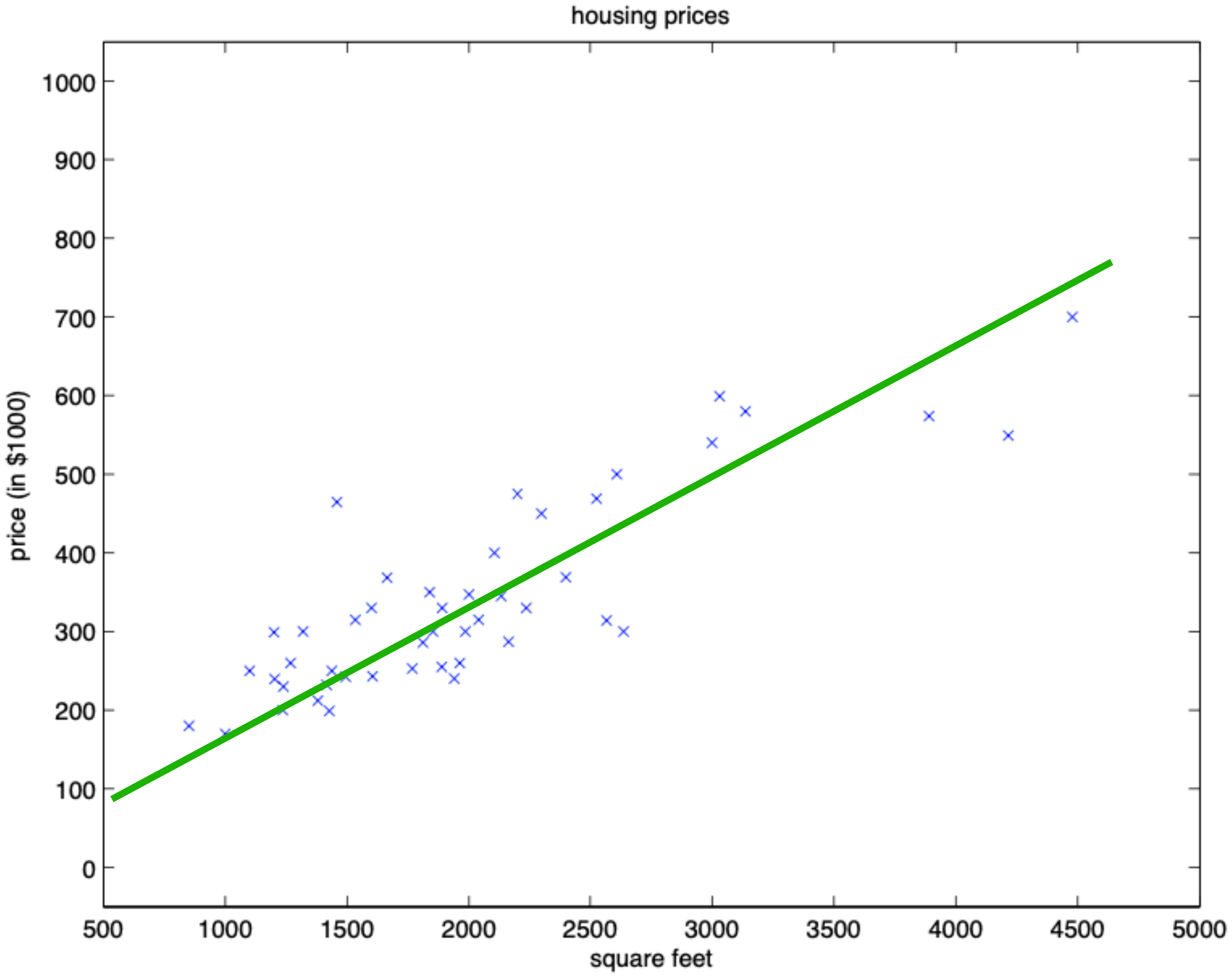


Least Regression with squared loss:

$$\arg\min_w \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

Recap on Linear Regression

Given dataset $\mathscr{D} = \{x_i, y_i\}, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$



Derivation of Normal equation:

$$L(w) := \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

$$\nabla_w L(w) =$$

Recap on SVM

Given dataset $\mathscr{D} = \{x_i, y_i\}, x_i \in \mathbb{R}^d, y_i \in \{+1, -1\}$

Hard margin SVM:

$$\min_{w,b} \|w\|_2^2$$

$$\forall i: \ y_i(w^\top x_i + b) \geq 1$$

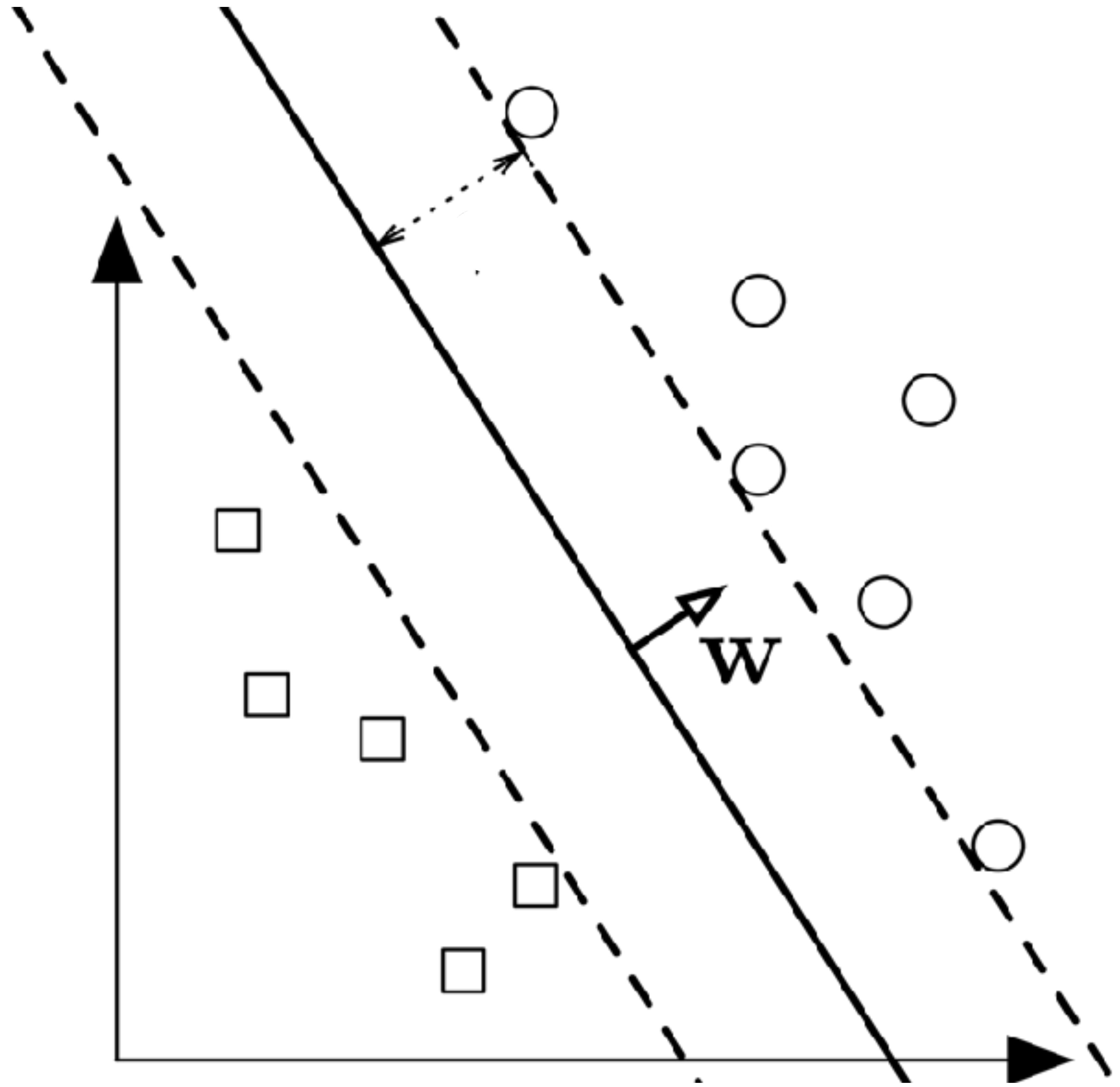
Recap on SVM

Given dataset $\mathscr{D} = \{x_i, y_i\}, x_i \in \mathbb{R}^d, y_i \in \{+1, -1\}$

Hard margin SVM:

$$\min_{w,b} \|w\|_2^2$$

$$\forall i: \; y_i(w^\top x_i + b) \geq 1$$



Width of the "street":

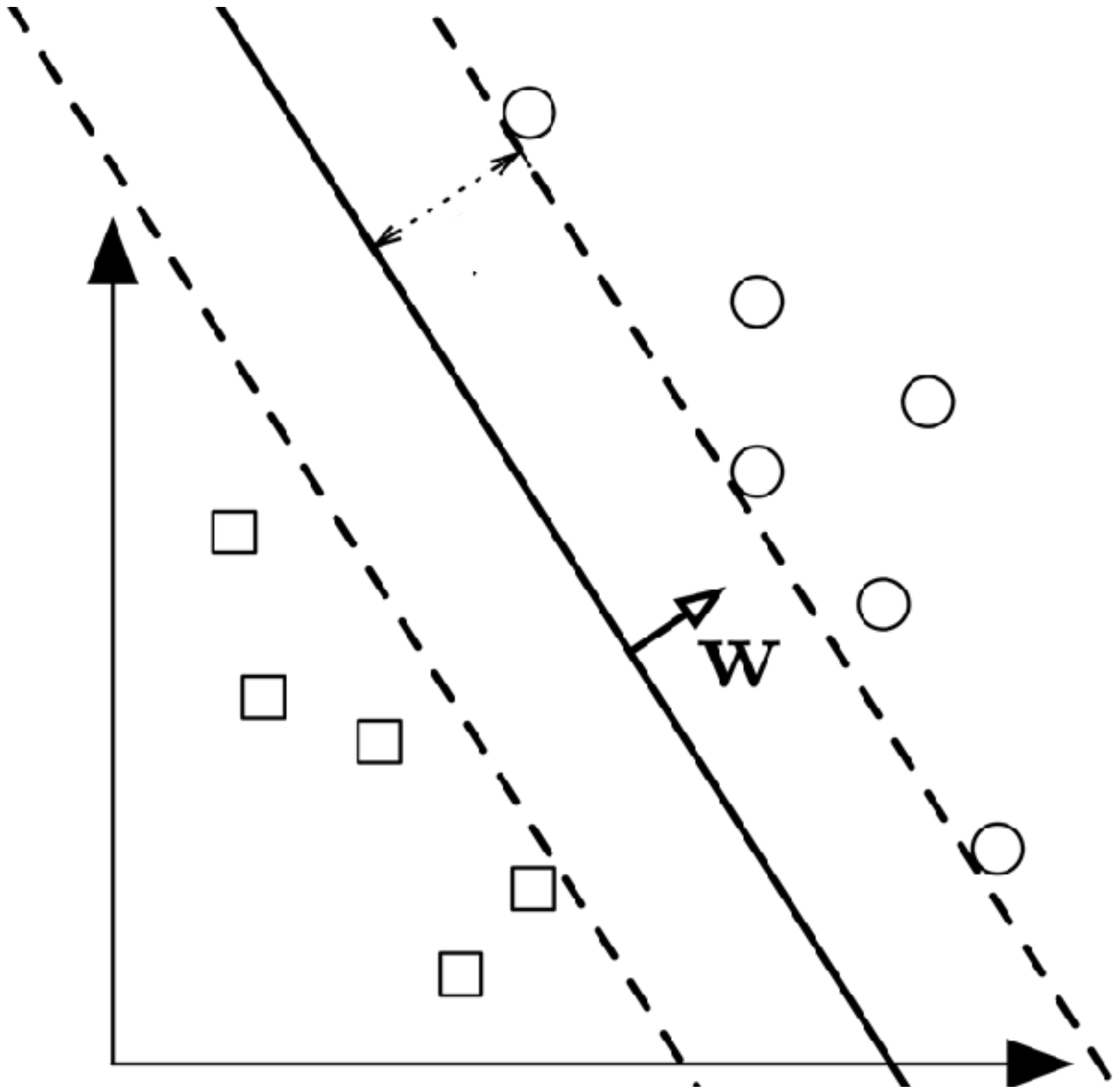
Recap on SVM

Given dataset $\mathscr{D} = \{x_i, y_i\}, x_i \in \mathbb{R}^d, y_i \in \{+1, -1\}$

Hard margin SVM:

$$\min_{w,b} \|w\|_2^2$$

$$\forall i: \; y_i(w^\top x_i + b) \geq 1$$



Width of the "street":

$$2/\|w\|_2$$

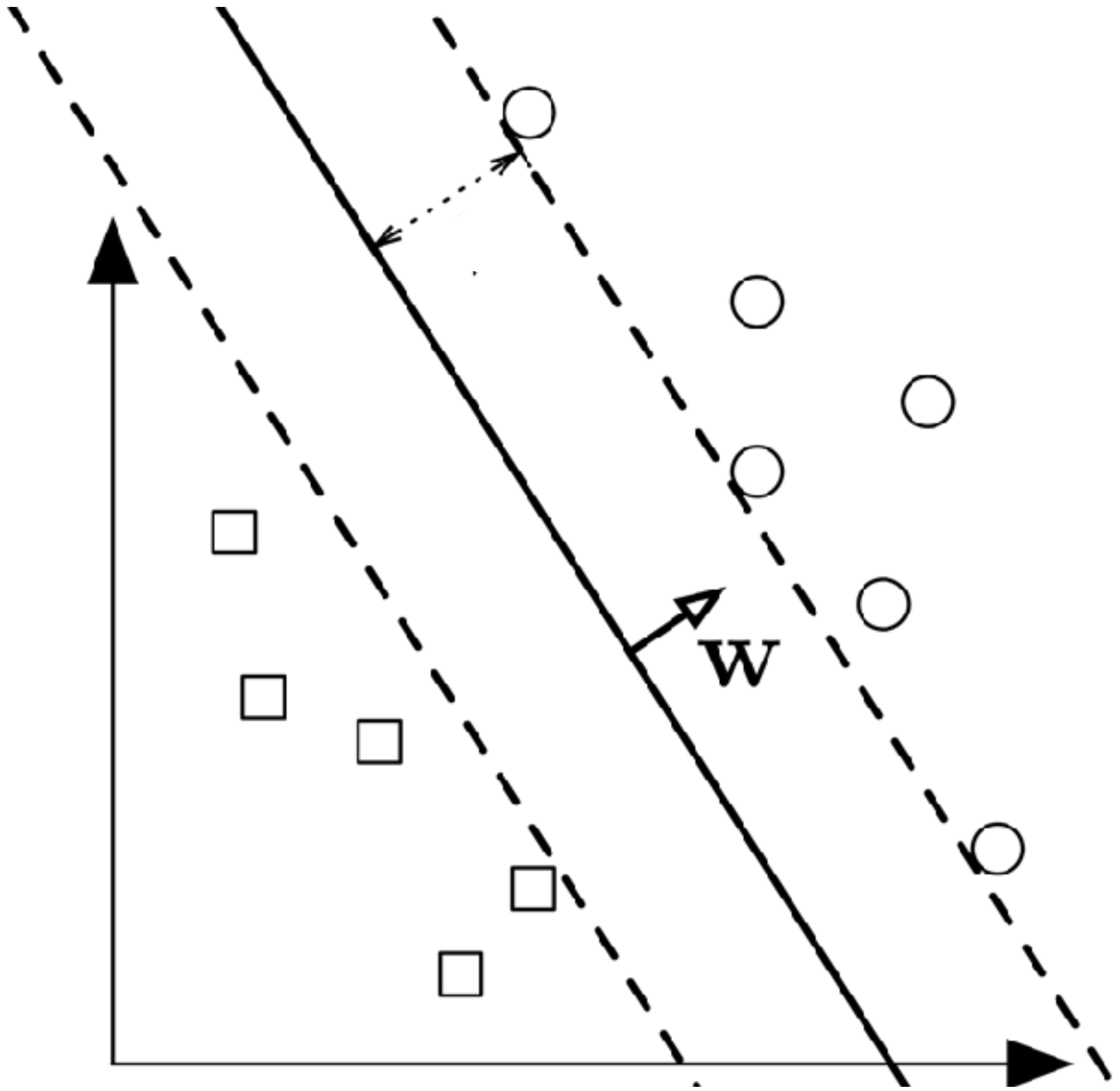
Recap on SVM

Given dataset $\mathscr{D} = \{x_i, y_i\}, x_i \in \mathbb{R}^d, y_i \in \{+1, -1\}$

Hard margin SVM:

$$\min_{w,b} \|w\|_2^2$$

$$\forall i: \; y_i(w^\top x_i + b) \geq 1$$



Width of the "street":

$$2/\|w\|_2$$

Find a "street" that has largest width, while keep all the points outside of the street

Outline for Today

1. Empirical Risk Minimization

2. Examples on loss & hypothesis classes

3. Regularization

ERM

Recall the general supervised learning setting:

ERM

Recall the general supervised learning setting:

We have some distribution P, dataset $\mathscr{D} = \{x_i, y_i\}_{i=1}^{n}$

ERM

Recall the general supervised learning setting:

We have some distribution P, dataset $\mathscr{D} = \{x_i, y_i\}_{i=1}^n$

Each data point is i.i.d sampled from P, i.e., $x_i, y_i \sim P$

ERM

Recall the general supervised learning setting:

We have some distribution P, dataset $\mathscr{D} = \{x_i, y_i\}_{i=1}^n$

Each data point is i.i.d sampled from P, i.e., $x_i, y_i \sim P$

Hypothesis $h : \mathscr{X} \to \mathbb{R}$, & hypothesis class $\mathscr{H} := \{h\} \subset \mathscr{X} \mapsto \mathbb{R}$

ERM

Recall the general supervised learning setting:

We have some distribution P, dataset $\mathscr{D} = \{x_i, y_i\}_{i=1}^n$

Each data point is i.i.d sampled from P, i.e., $x_i, y_i \sim P$

Hypothesis $h : \mathscr{X} \to \mathbb{R}$, & hypothesis class $\mathscr{H} := \{h\} \subset \mathscr{X} \mapsto \mathbb{R}$

Loss function: $\ell(h(x), y)$

ERM

The ultimate objective function:

$$\arg\min_{h \in \mathcal{H}} \mathbb{E}_{x,y \sim P} \left[\ell(h(x), y) \right]$$

ERM

The ultimate objective function:

$$\arg\min_{h \in \mathcal{H}} \mathbb{E}_{x,y \sim P} [\ell(h(x), y)]$$

Unknown

ERM

The ultimate objective function:

$$\arg\min_{h \in \mathscr{H}} \mathbb{E}_{x,y \sim P}\left[\ell(h(x), y)\right]$$

Unknown

Instead we have its **empirical** version

ERM

The ultimate objective function:

$$\arg\min_{h \in \mathcal{H}} \mathbb{E}_{x,y \sim P} [\ell(h(x), y)]$$

Unknown

Instead we have its **empirical** version

$$\arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} [\ell(h(x_i), y_i)]$$

ERM

The ultimate objective function:

$$\arg\min_{h \in \mathcal{H}} \mathbb{E}_{x,y \sim P}\left[\ell(h(x), y)\right]$$

Unknown

Instead we have its **empirical** version

$$\arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \left[\ell(h(x_i), y_i)\right]$$

Empirical risk / Empirical error

The generalization error of ERM solution

$$\hat{h}_{ERM} := \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \left[\ell(h(x_i), y_i) \right]$$

The generalization error of ERM solution

$$\hat{h}_{ERM} := \arg\min_{h \in \mathscr{H}} \frac{1}{n} \sum_{i=1}^{n} \left[\ell(h(x_i), y_i) \right]$$

We often are interested in the true performance of $\hat{h}_{ERM}$:

The generalization error of ERM solution

$$\hat{h}_{ERM} := \arg\min_{h \in \mathscr{H}} \frac{1}{n} \sum_{i=1}^{n} \left[\ell(h(x_i), y_i) \right]$$

We often are interested in the true performance of $\hat{h}_{ERM}$:

$$\mathbb{E}_{\mathscr{D}} \left[\mathbb{E}_{x,y \sim P} \ell(\hat{h}_{ERM}(x), y) \right]$$

The generalization error of ERM solution

$$\hat{h}_{ERM} := \arg\min_{h \in \mathscr{H}} \frac{1}{n} \sum_{i=1}^{n} \left[\ell(h(x_i), y_i) \right]$$

We often are interested in the true performance of $\hat{h}_{ERM}$:

$$\mathbb{E}_{\mathscr{D}} \left[\mathbb{E}_{x,y \sim P} \ell(\hat{h}_{ERM}(x), y) \right]$$

Note $\hat{h}_{ERM}$ is a random quantity as
it depends on data $\mathscr{D}$

The generalization error of ERM solution

$$\hat{h}_{ERM} := \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \left[\ell(h(x_i), y_i) \right]$$

We often are interested in the true performance of $\hat{h}_{ERM}$:

$$\mathbb{E}_{\mathcal{D}} \left[\mathbb{E}_{x,y \sim P} \ell(\hat{h}_{ERM}(x), y) \right]$$

Note $\hat{h}_{ERM}$ is a random quantity as
it depends on data $\mathcal{D}$

e.g., In LR: $\hat{w} = (XX^\top)^{-1}XY$.

The generalization error of ERM solution

Ideally, we want the true loss of ERM to be small:

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{x,y\sim P}\ell(\hat{h}_{ERM}(x), y)\right] \approx \min_{h\in\mathcal{H}} \mathbb{E}_{x,y\sim P}\ell(h(x), y)$$

The generalization error of ERM solution

Ideally, we want the true loss of ERM to be small:

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{x,y\sim P}\ell(\hat{h}_{ERM}(x), y)\right] \approx \min_{h\in\mathcal{H}} \mathbb{E}_{x,y\sim P}\ell(h(x), y)$$

The Minimum expected loss we could get if we knew P

The generalization error of ERM solution

Ideally, we want the true loss of ERM to be small:

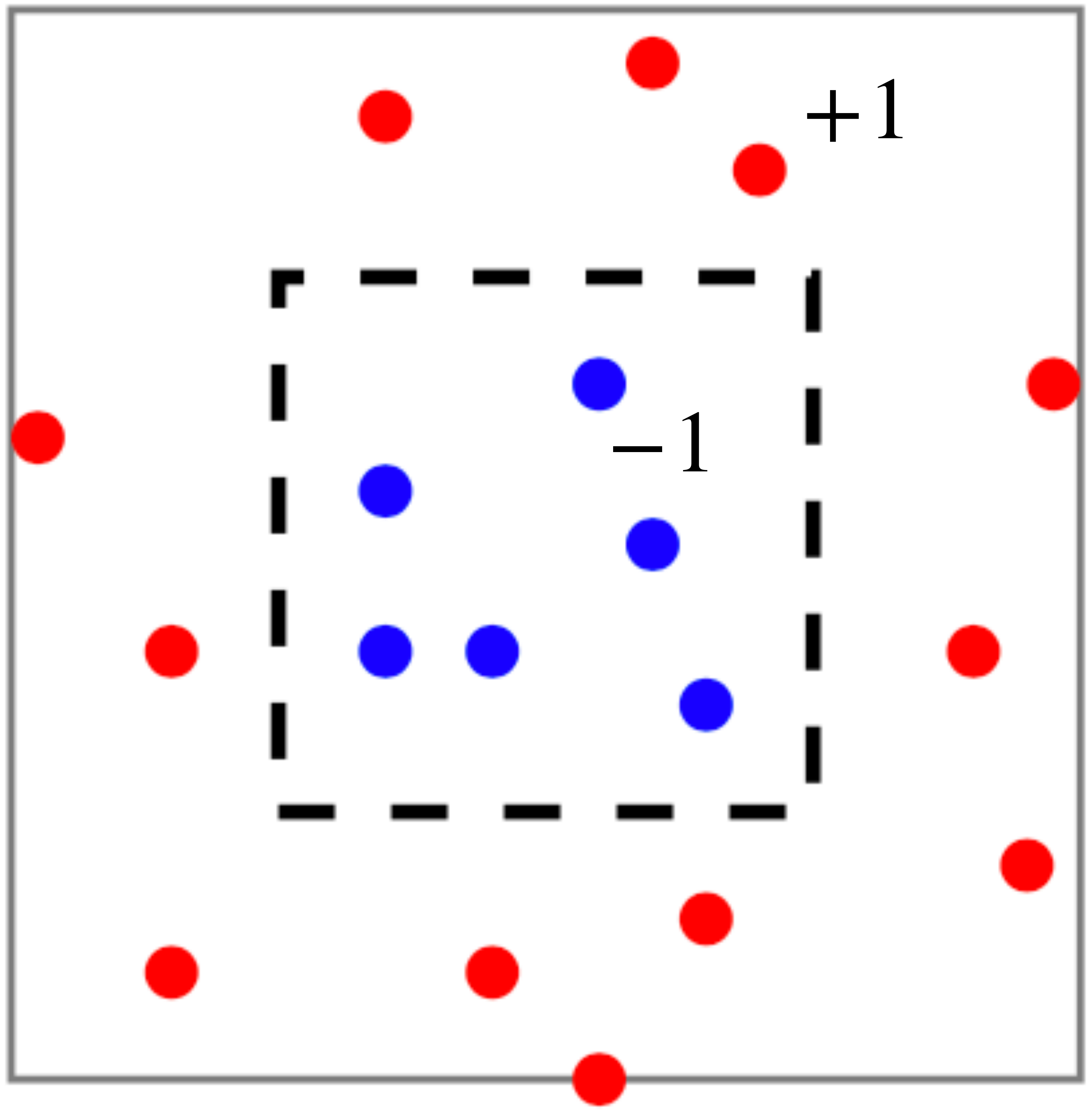
$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{x,y\sim P}\ell(\hat{h}_{ERM}(x), y)\right] \approx \min_{h\in\mathcal{H}} \mathbb{E}_{x,y\sim P}\ell(h(x), y)$$

The Minimum expected loss we could get if we knew P

However, this may not hold if we are not careful about designing $\mathcal{H}$

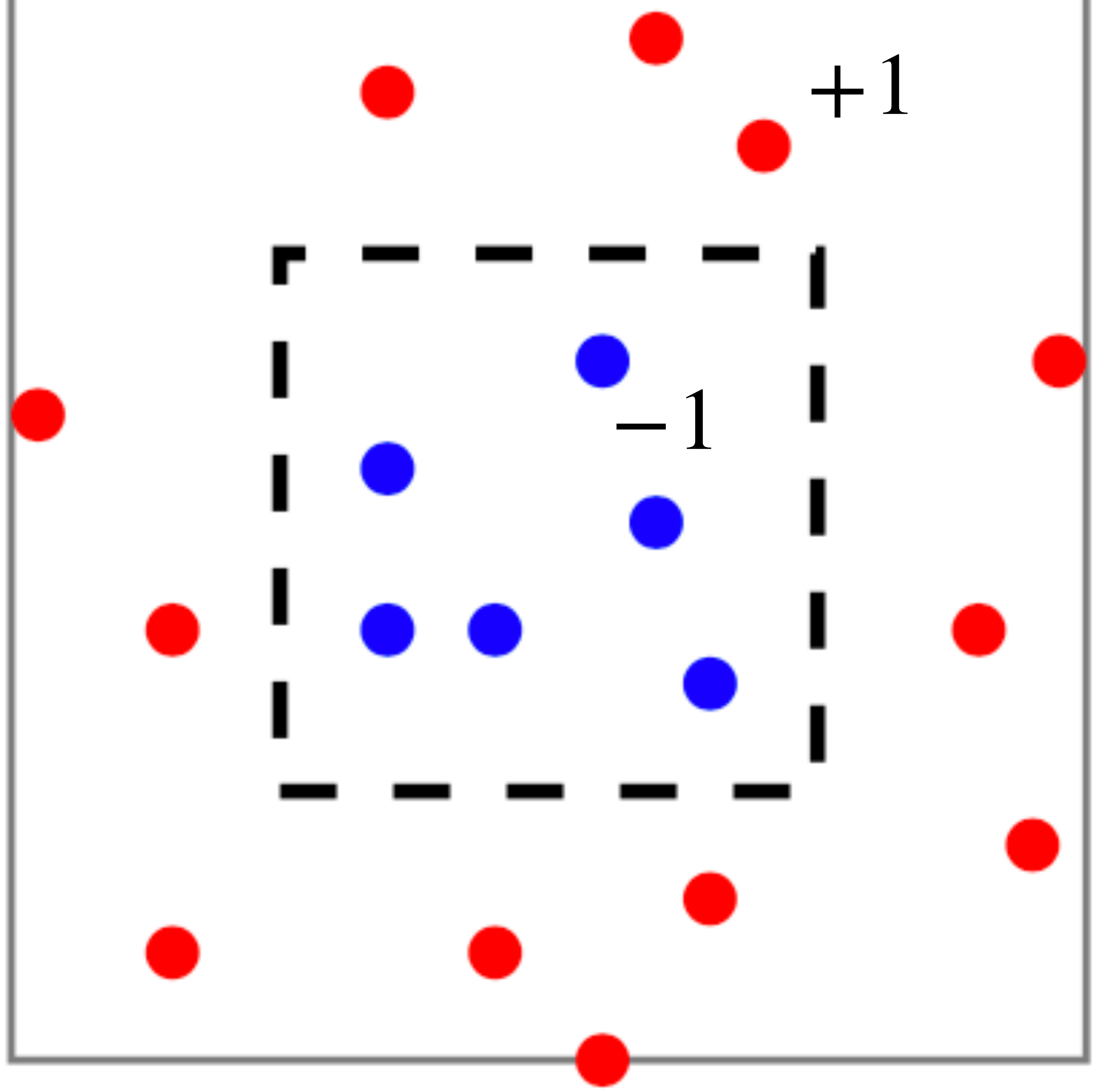
Example:

P: x uniformly distribution over the square;
Label: blue if inside the smaller square, else red



Example:

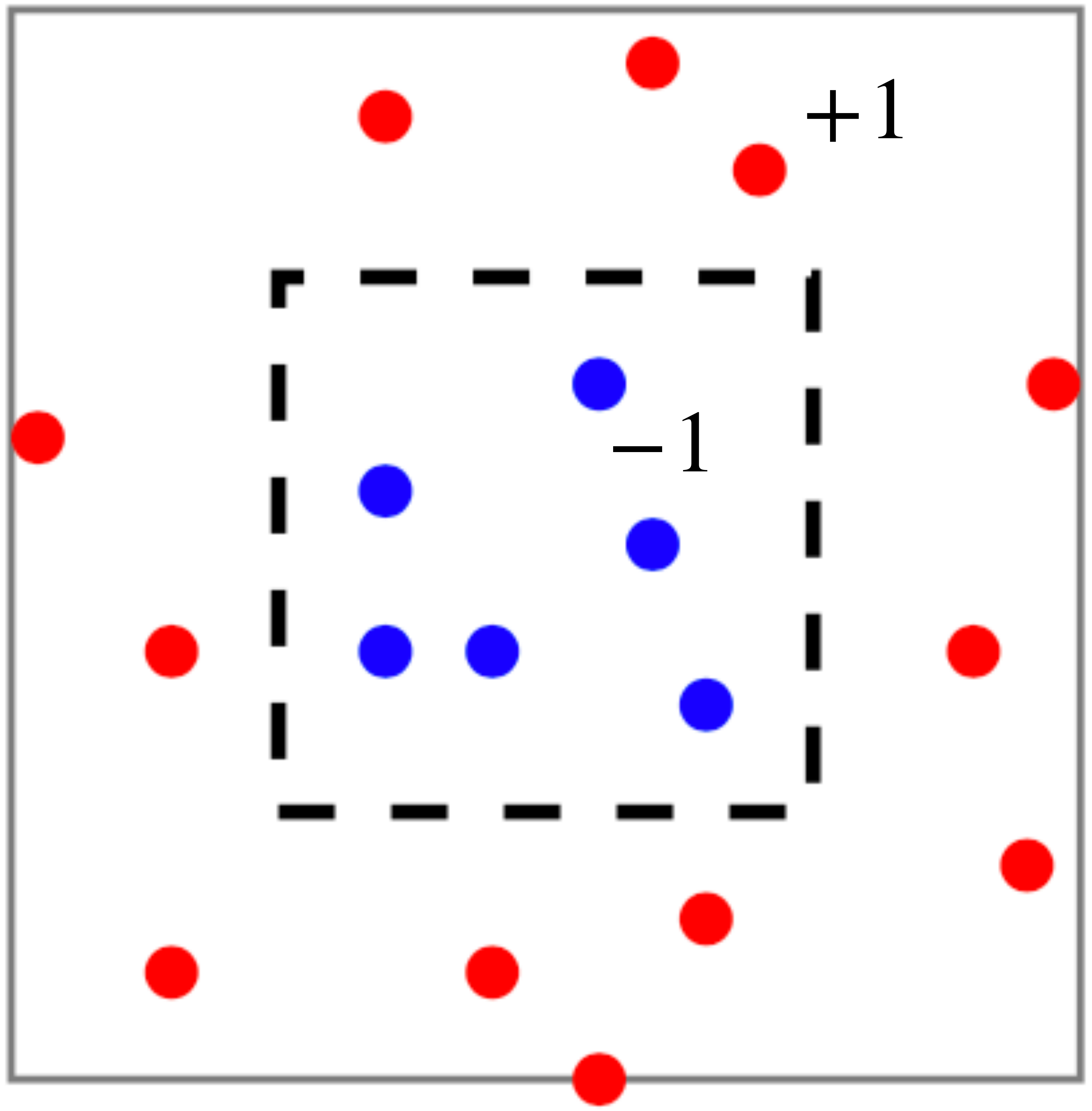
P: x uniformly distribution over the square;
Label: blue if inside the smaller square, else red



Consider a hypothesis class $\mathscr{H}$ contains ALL mappings from $x \to y$

Example:

P: x uniformly distribution over the square;
Label: blue if inside the smaller square, else red

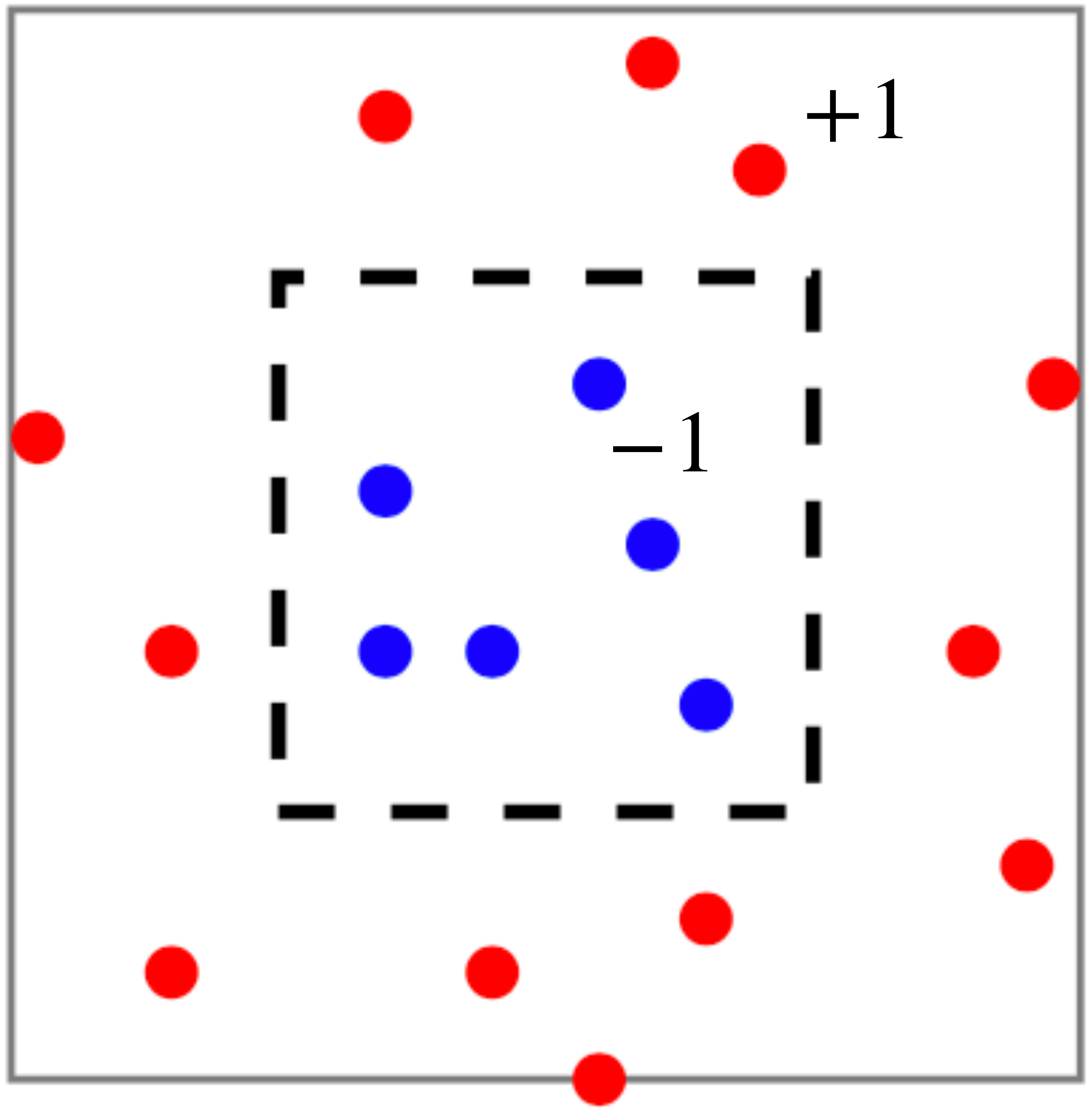


Consider a hypothesis class $\mathscr{H}$ contains ALL mappings from $x \to y$

Zero one loss $\ell(h(x), y) = \mathbf{1}(h(x) \neq y)$

Example:

P: x uniformly distribution over the square;
Label: blue if inside the smaller square, else red



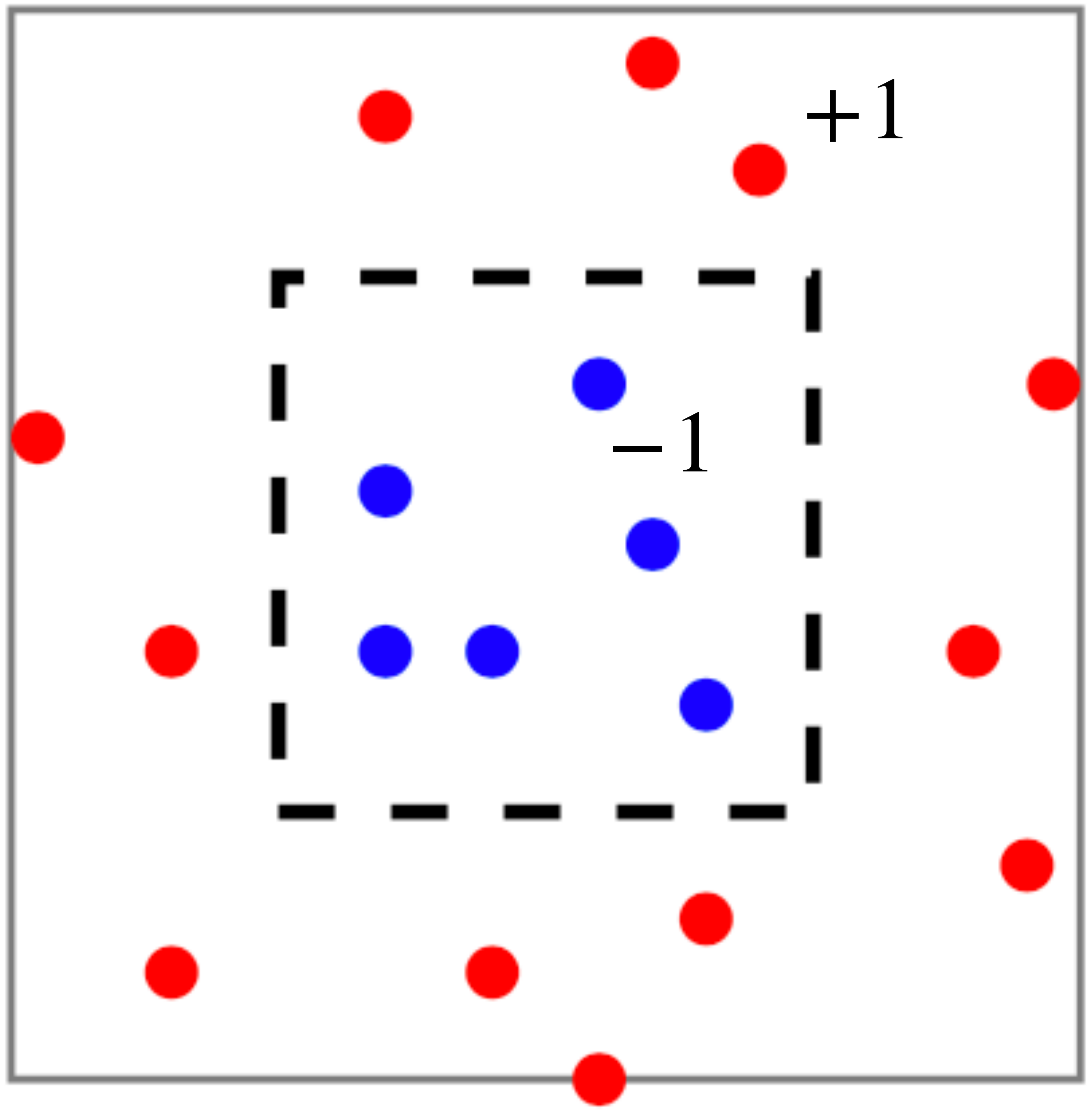
Consider a hypothesis class $\mathscr{H}$ contains ALL mappings from $x \to y$

Zero one loss $\ell(h(x), y) = \mathbf{1}(h(x) \neq y)$

Let us consider this solution that memorizes data:

Example:

P: x uniformly distribution over the square;
Label: blue if inside the smaller square, else red



Consider a hypothesis class $\mathscr{H}$ contains ALL mappings from $x \to y$

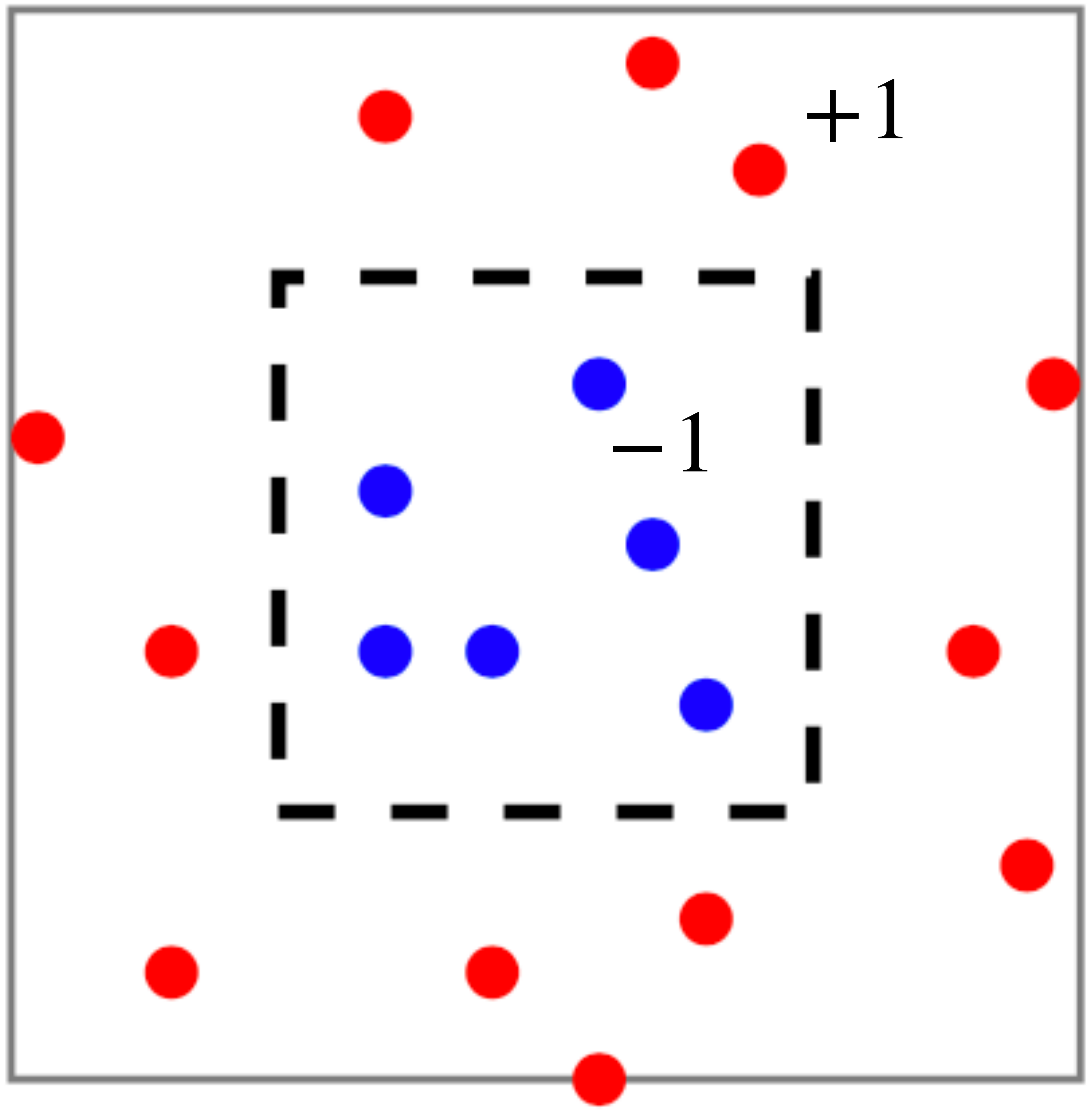
Zero one loss $\ell(h(x), y) = \mathbf{1}(h(x) \neq y)$

Let us consider this solution that memorizes data:

$$\hat{h}(x) = \begin{cases} y_i & \text{if } \exists i, x_i = x \\ +1 & \text{else} \end{cases}$$

Example:

P: x uniformly distribution over the square;
Label: blue if inside the dashed square, else red

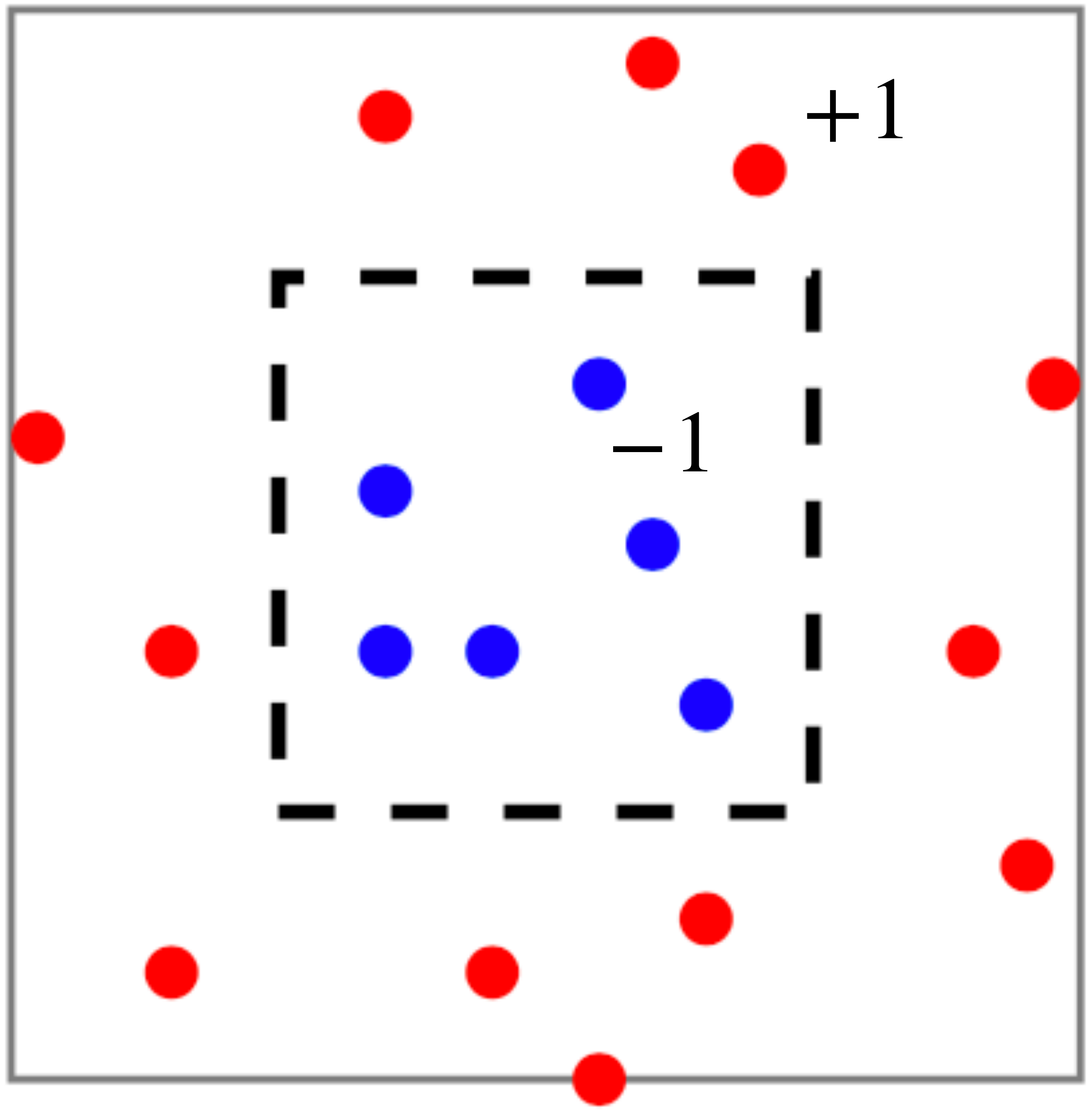


$$\hat{h}(x) = \begin{cases} y_i & \text{if } \exists i, x_i = x \\ +1 & \text{else} \end{cases}$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{h}(x_i), y_i) = 0$$

Example:

P: x uniformly distribution over the square;
Label: blue if inside the dashed square, else red



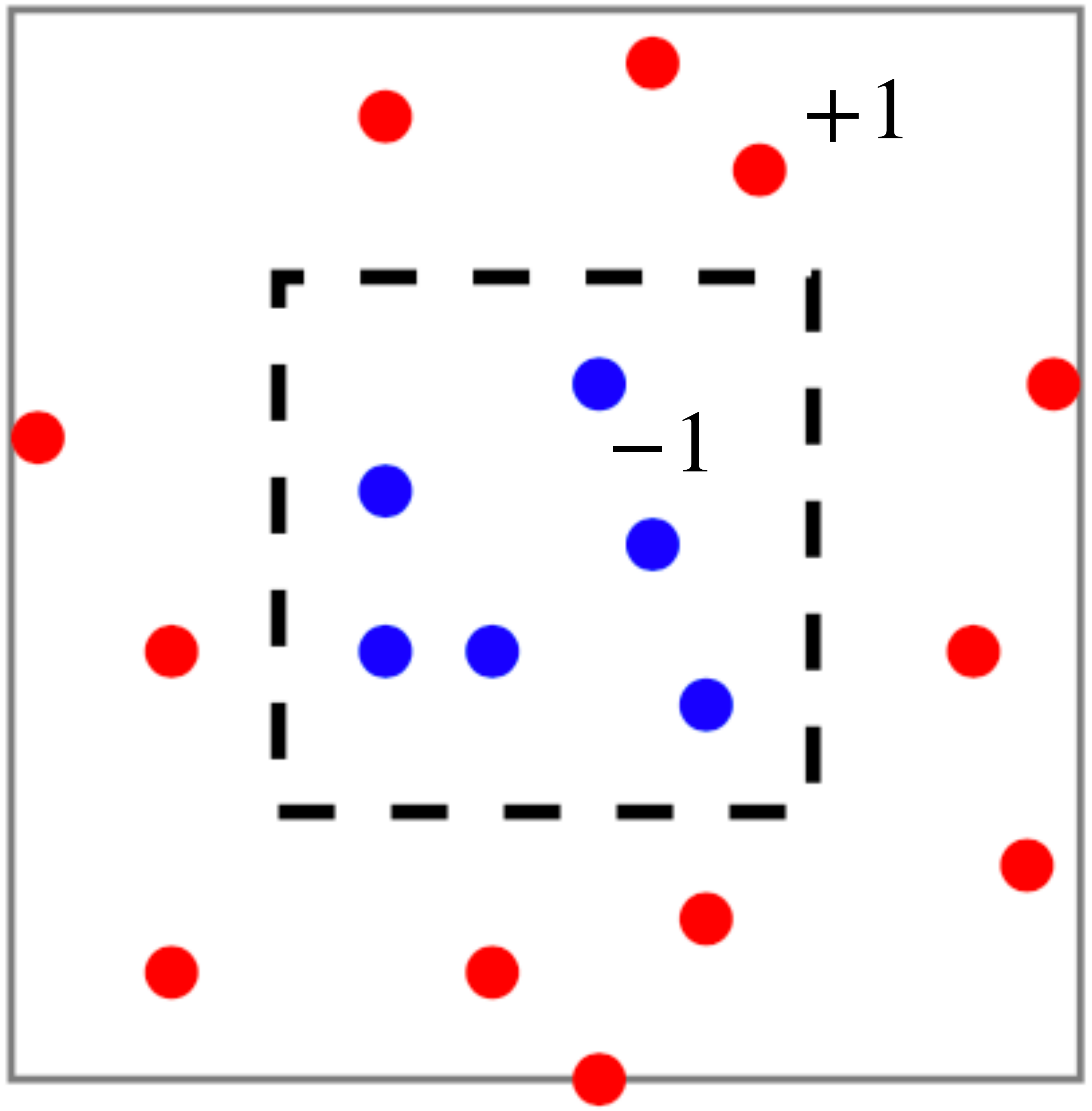
$$\hat{h}(x) = \begin{cases} y_i & \text{if } \exists i, x_i = x \\ +1 & \text{else} \end{cases}$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{h}(x_i), y_i) = 0$$

Q: But what's the true expected error of this $\hat{h}$?

Example:

P: x uniformly distribution over the square;
Label: blue if inside the dashed square, else red



$$\hat{h}(x) = \begin{cases} y_i & \text{if } \exists i, x_i = x \\ +1 & \text{else} \end{cases}$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{h}(x_i), y_i) = 0$$

Q: But what's the true expected error of this $\hat{h}$?

A: area of smaller box / total area

ERM with inductive bias

A common solution is to restrict the search space (i.e., hypothesis class)

$$\hat{h}_{ERM} := \arg\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \left[\ell(h(x_i), y_i)\right]$$

ERM with inductive bias

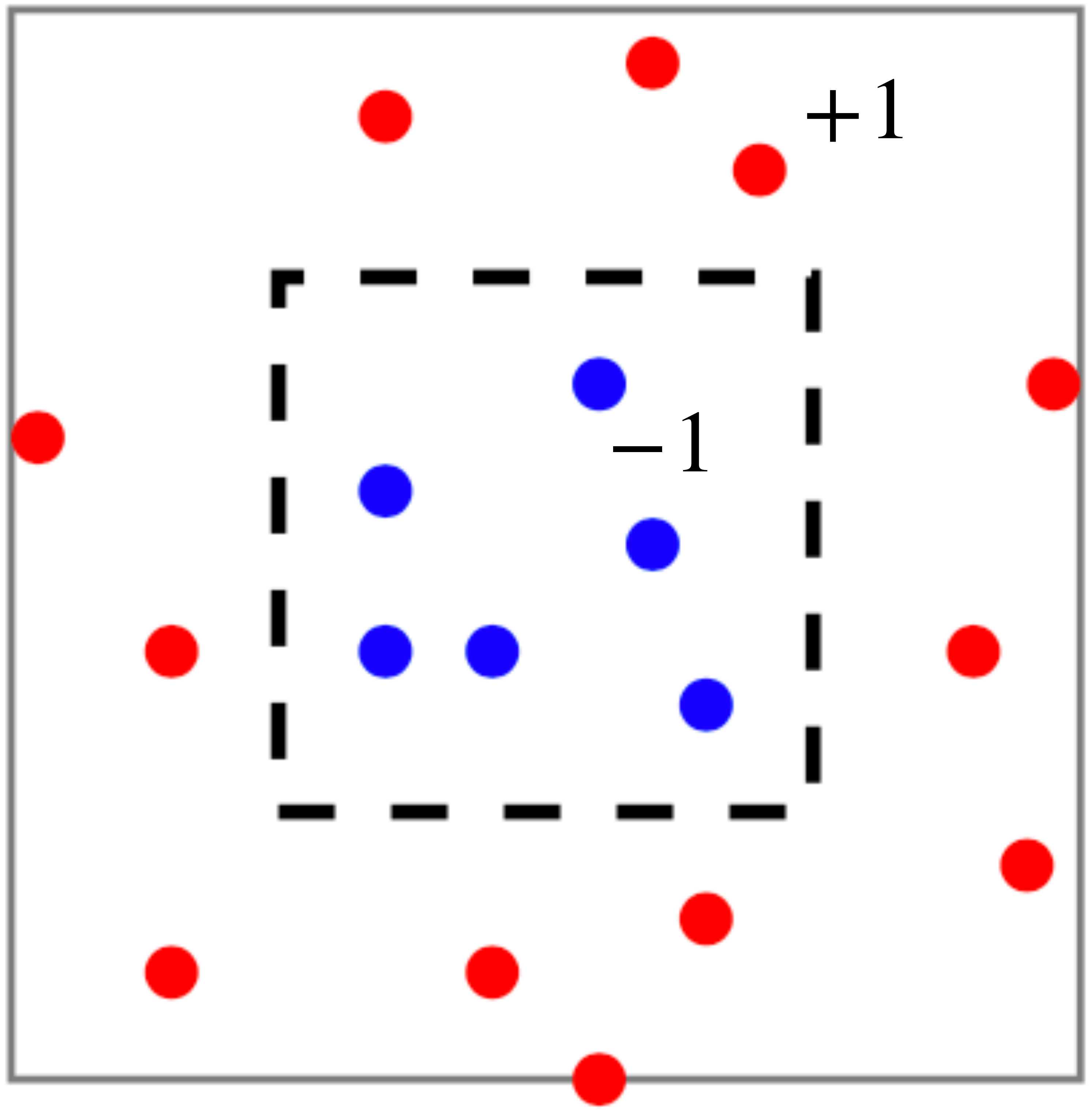
A common solution is to restrict the search space (i.e., hypothesis class)

$$\hat{h}_{ERM} := \arg\min_{h \in \mathscr{H}} \frac{1}{n} \sum_{i=1}^{n} \left[\ell(h(x_i), y_i) \right]$$

By restricting to $\mathscr{H}$, we bias towards solutions from $\mathscr{H}$

Example:

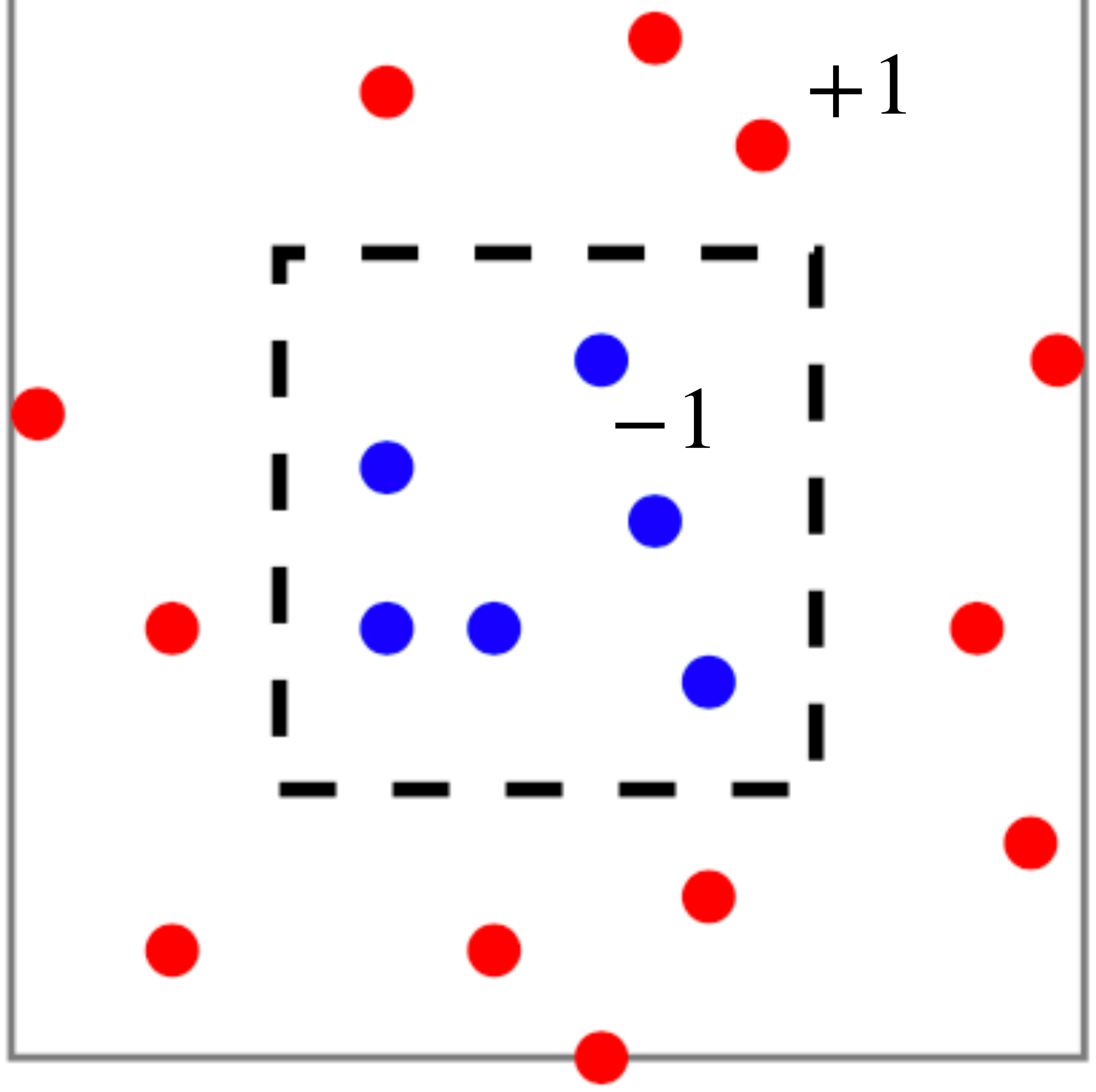
P: x uniformly distribution over the square;
Label: blue if inside the dashed square, else red



Unrestricted hypothesis class did not work;

Example:

P: x uniformly distribution over the square;
Label: blue if inside the dashed square, else red

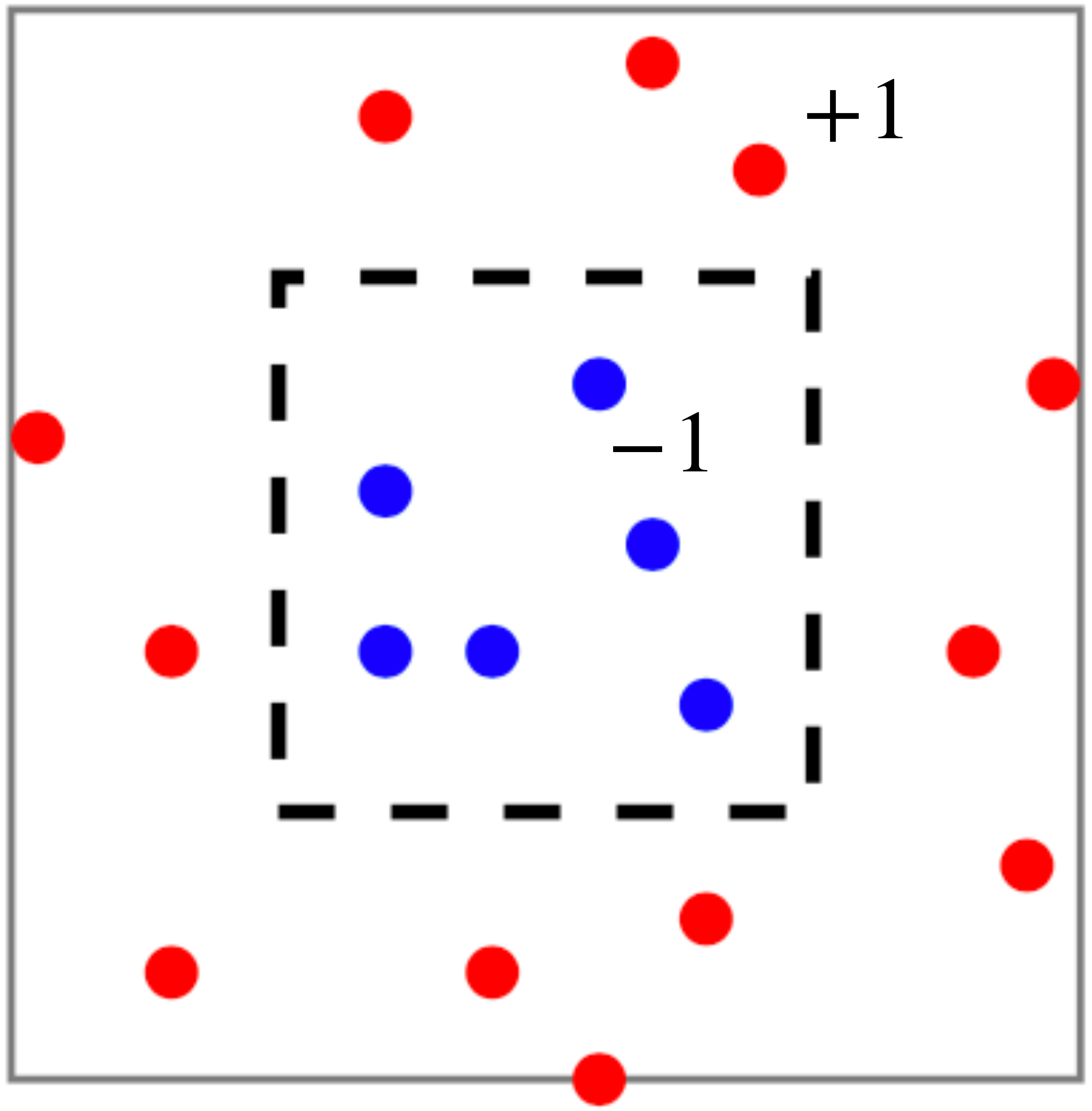


Unrestricted hypothesis class did not work;

However, if we restrict $\mathscr{H}$ to contains ALL axis-aligned rectangles,
then ERM will succeed, i.e.,

Example:

P: x uniformly distribution over the square;
Label: blue if inside the dashed square, else red



Unrestricted hypothesis class did not work;

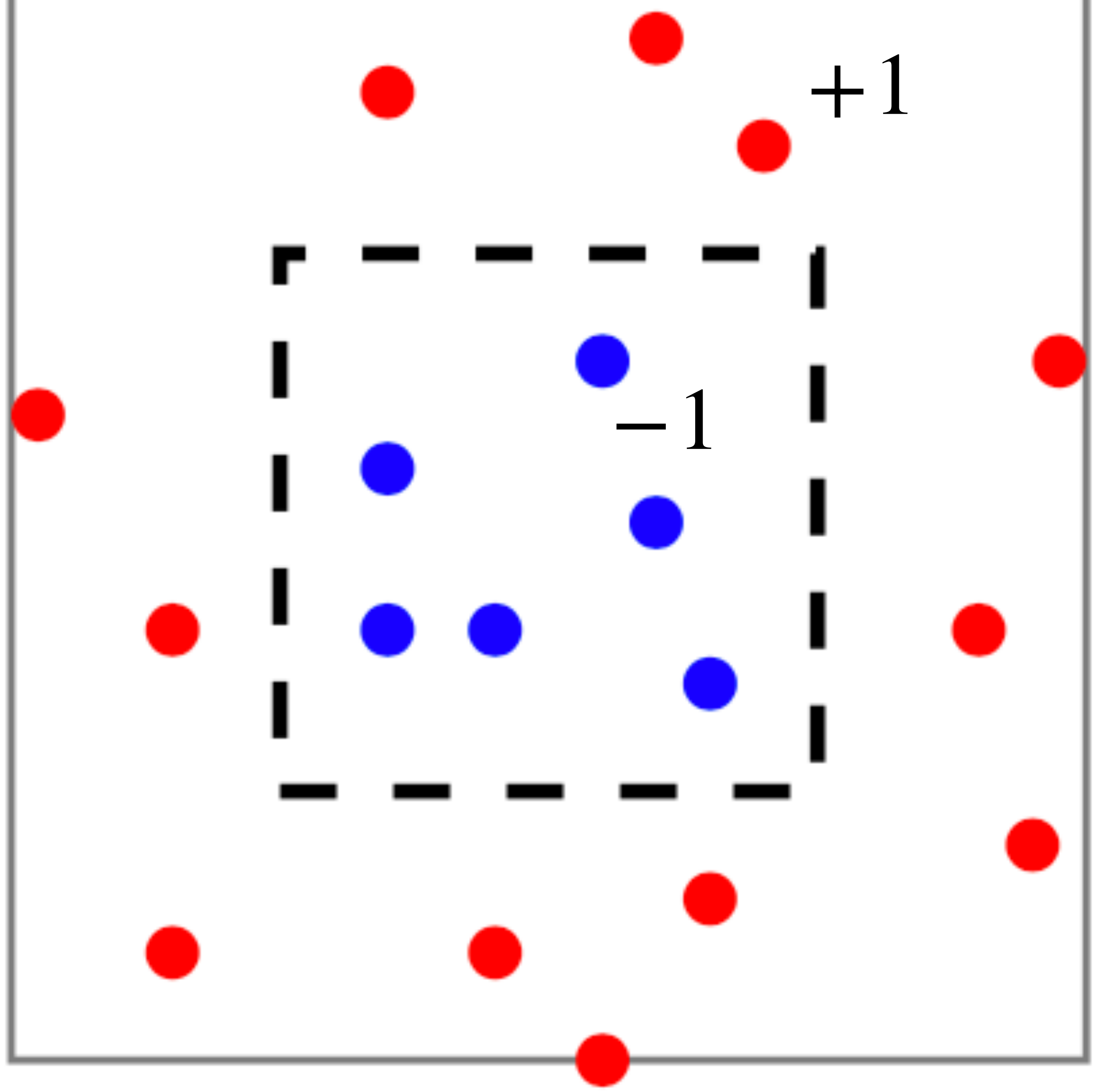
However, if we restrict $\mathscr{H}$ to contains ALL axis-aligned rectangles,
then ERM will succeed, i.e.,

$$\mathbb{E}_{\mathscr{D}}\left[\mathbb{E}_{x,y\sim P}\ell(\hat{h}_{ERM}(x), y)\right]$$

$$\leq \min_{h\in\mathscr{H}} \mathbb{E}_{x,y\sim P}\ell(h(x), y) + O(1/\sqrt{n})$$

Example:

P: x uniformly distribution over the square;
Label: blue if inside the dashed square, else red



Unrestricted hypothesis class did not work;

However, if we restrict $\mathscr{H}$ to contains ALL axis-aligned rectangles,
then ERM will succeed, i.e.,

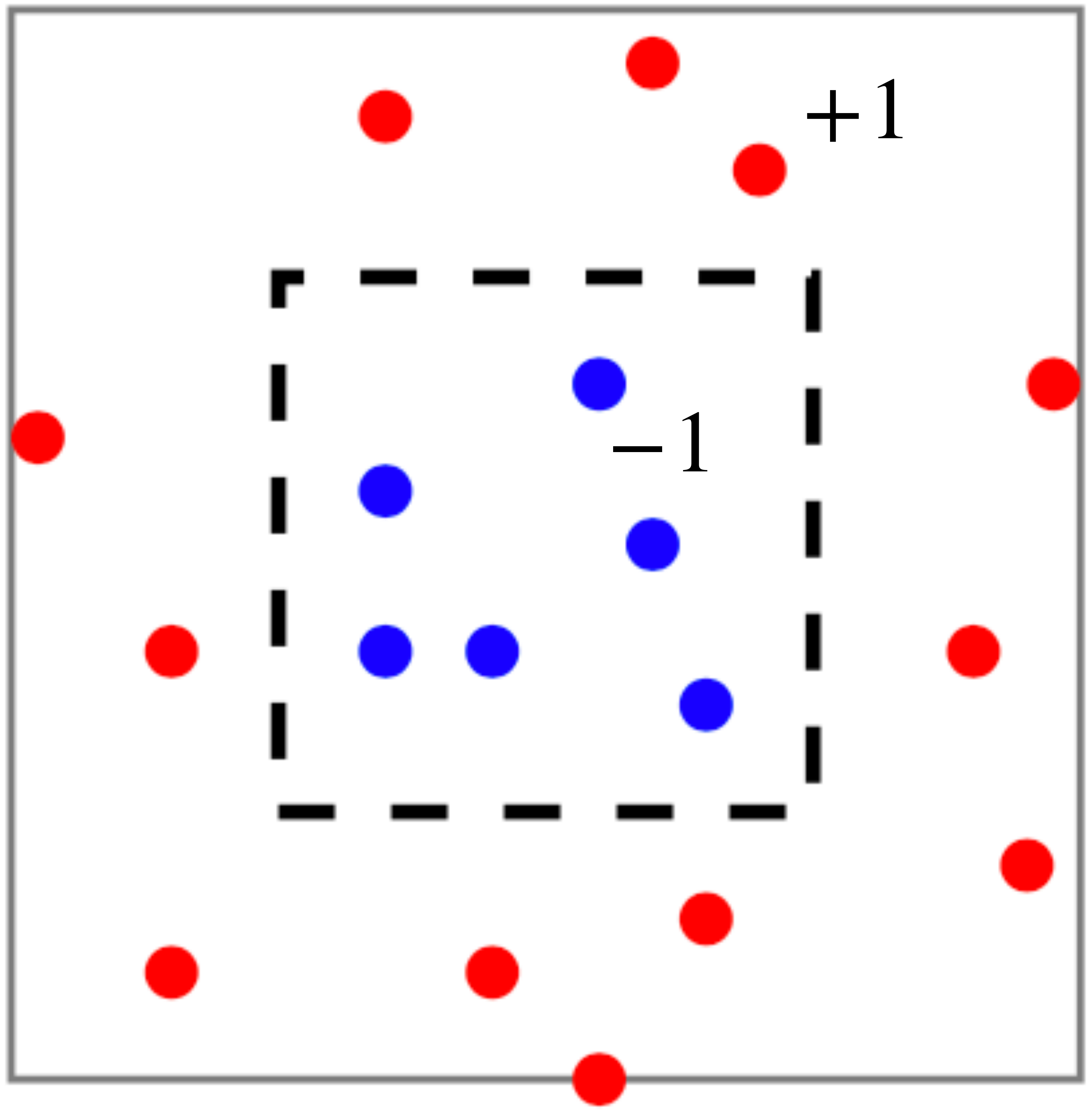
$$\mathbb{E}_{\mathscr{D}}\left[\mathbb{E}_{x,y\sim P}\ell(\hat{h}_{ERM}(x), y)\right]$$

$$\leq \min_{h\in\mathscr{H}} \mathbb{E}_{x,y\sim P}\ell(h(x), y) + O(1/\sqrt{n})$$

$$\leq O(1/\sqrt{n})$$

Example:

P: x uniformly distribution over the square;
Label: blue if inside the dashed square, else red



Unrestricted hypothesis class did not work;

However, if we restrict $\mathscr{H}$ to contains ALL axis-aligned rectangles,
then ERM will succeed, i.e.,

$$\mathbb{E}_{\mathscr{D}}\left[\mathbb{E}_{x,y\sim P}\ell(\hat{h}_{ERM}(x), y)\right]$$

$$\leq \min_{h\in\mathscr{H}} \mathbb{E}_{x,y\sim P}\ell(h(x), y) + O(1/\sqrt{n})$$

$$\leq O(1/\sqrt{n})$$

(Exact proof out of the scope of this class — see CS 4783/5783)

Summary so far

ERM with unrestricted hypothesis class could fail (i.e., overfitting)

To guarantee small test error, we need to restrict $\mathscr{H}$

Outline for Today

1. Empirical Risk Minimization

2. Examples on loss & hypothesis classes

3. Regularization

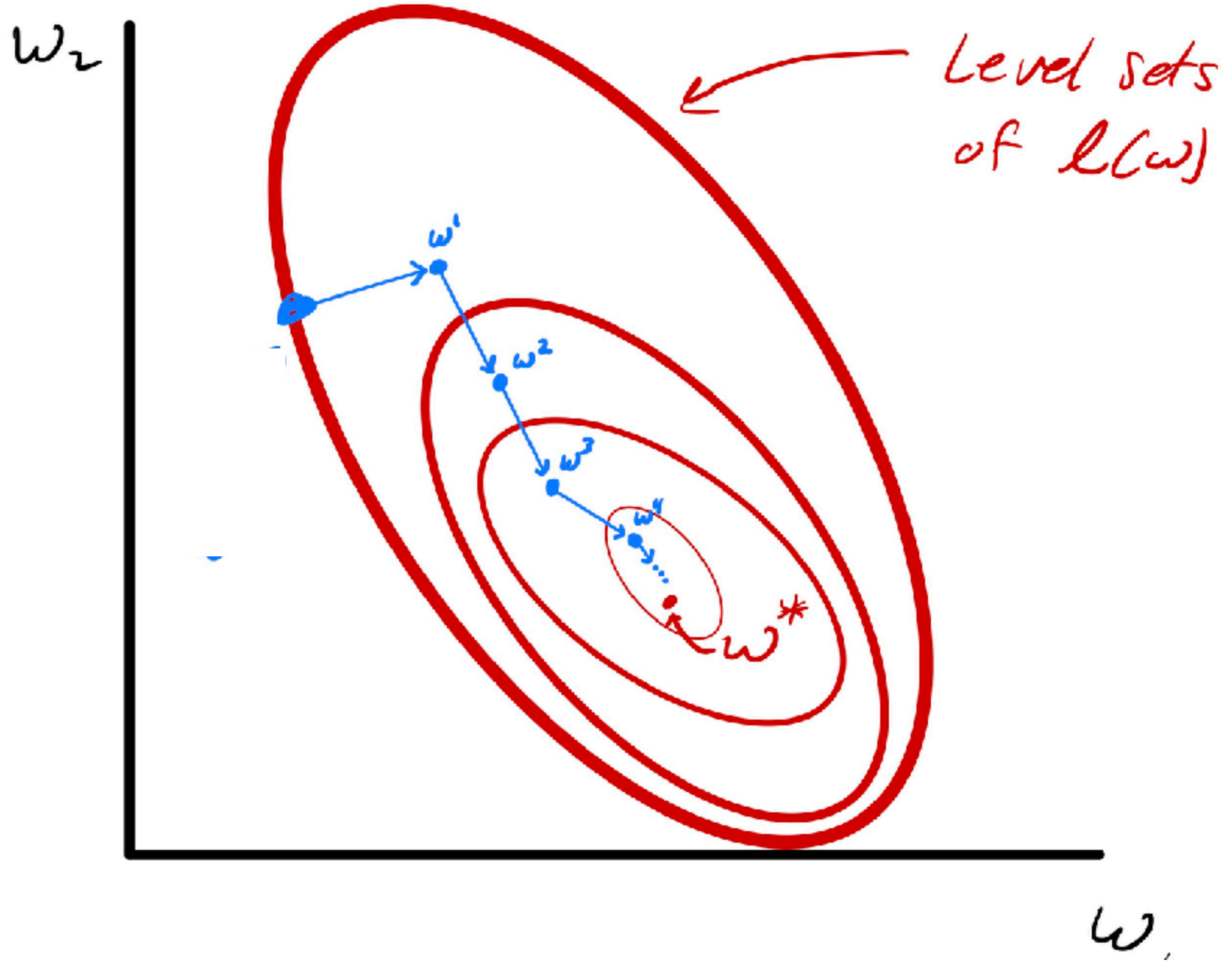
ERM with restricted hypothesis class

$$\min_{h} \frac{1}{n} \sum_{i=1}^{n} \left[\ell(h(x_i), y_i) \right]$$

$$\text{s.t. } h \in \mathcal{H}$$

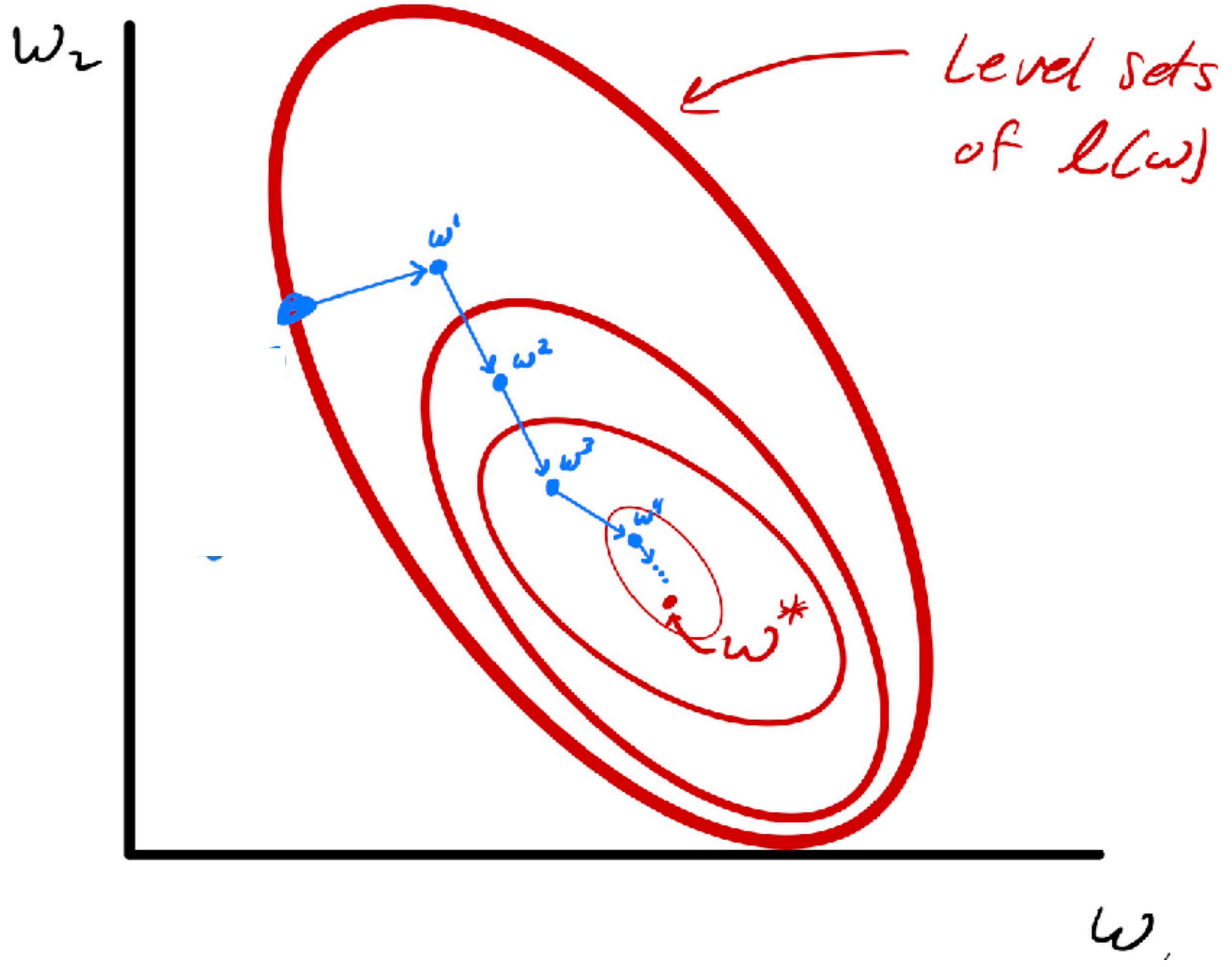
Let's go through several examples on Constraints under the linear regression context

Linear Regression: squared loss + ℓ_2 constraint



$$\min_w \frac{1}{n} \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

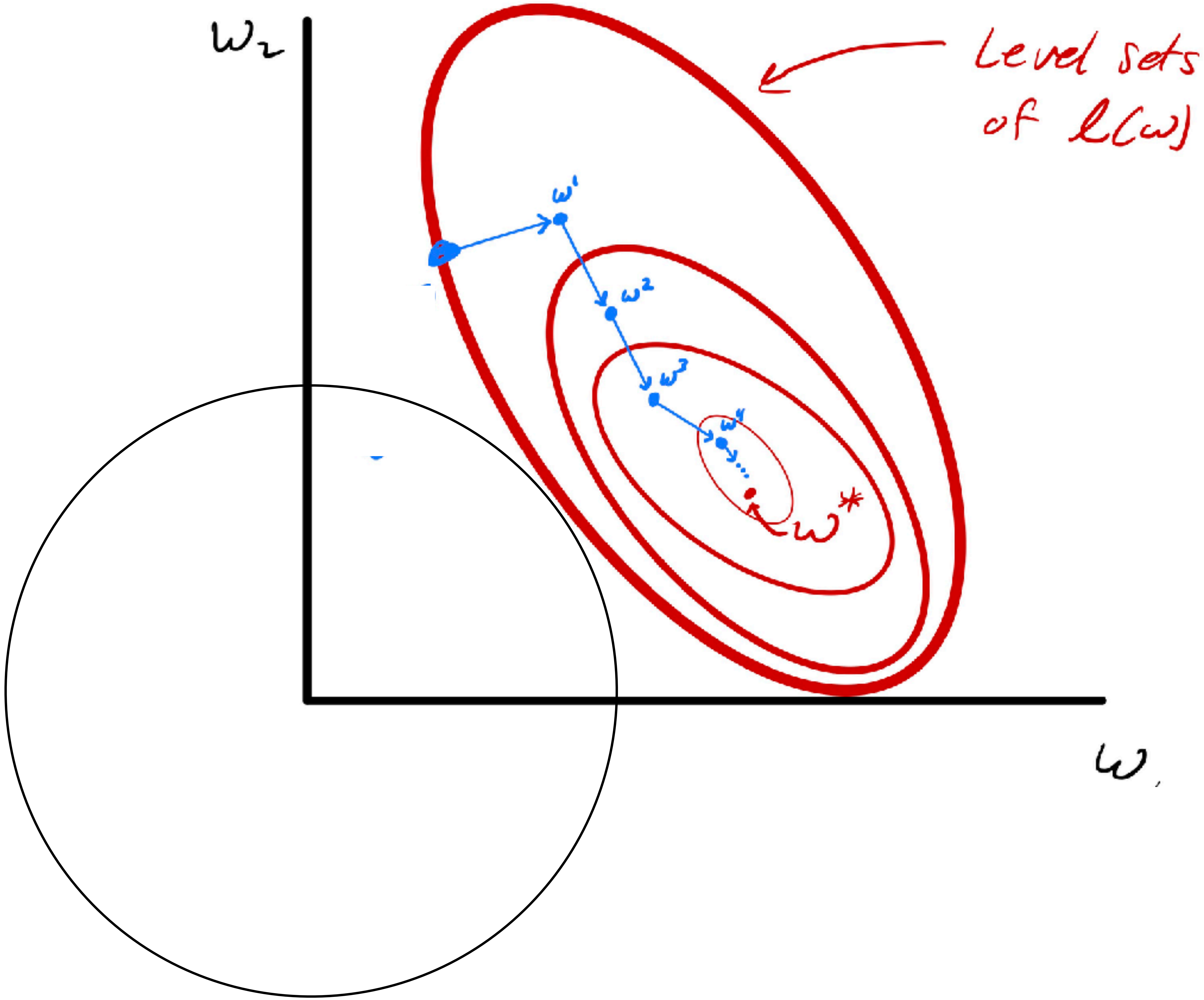
Linear Regression: squared loss + ℓ_2 constraint



$$\min_w \frac{1}{n}\sum_{i=1}^{n}(w^\top x_i - y_i)^2$$

$$\text{s.t. } \|w\|_2^2 \leq B$$

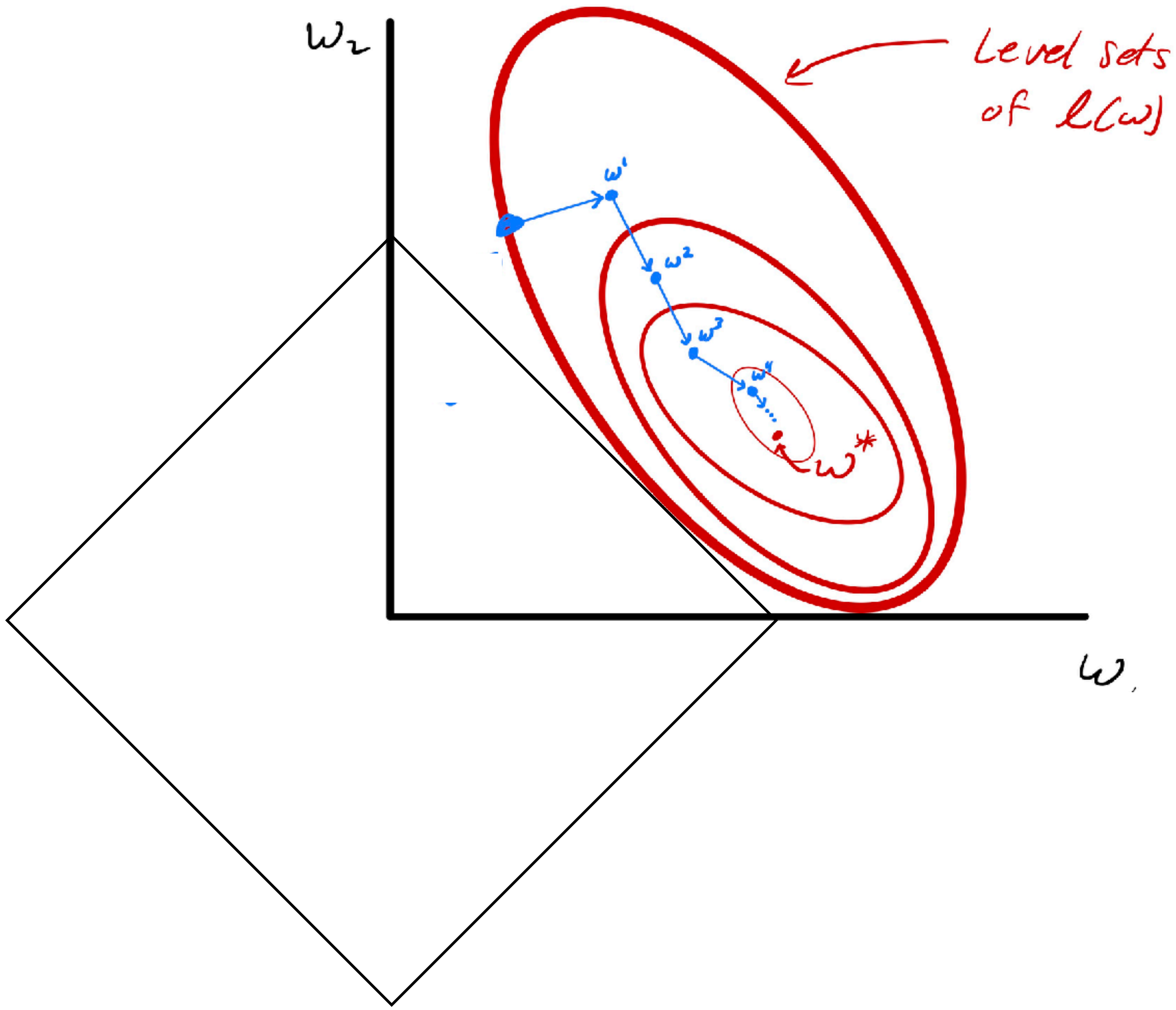
Linear Regression: squared loss + ℓ_2 constraint



$$\min_w \frac{1}{n} \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

$$\text{s.t. } \|w\|_2^2 \leq B$$

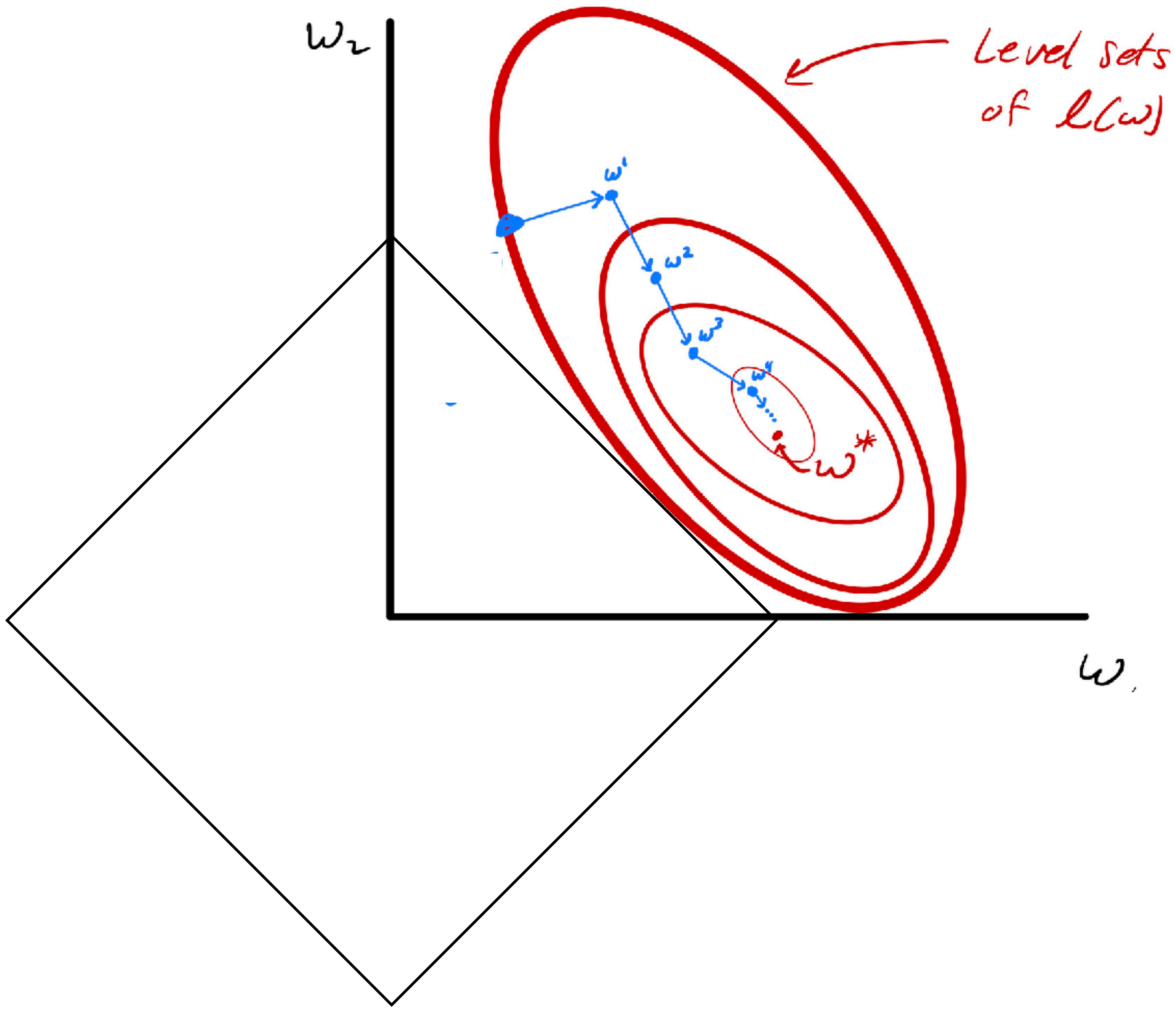
Linear Regression: squared loss + ℓ_1 constraint



$$\min_w \frac{1}{n} \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

$$\text{s.t. } \|w\|_1 \leq B$$

Linear Regression: squared loss + ℓ_1 constraint

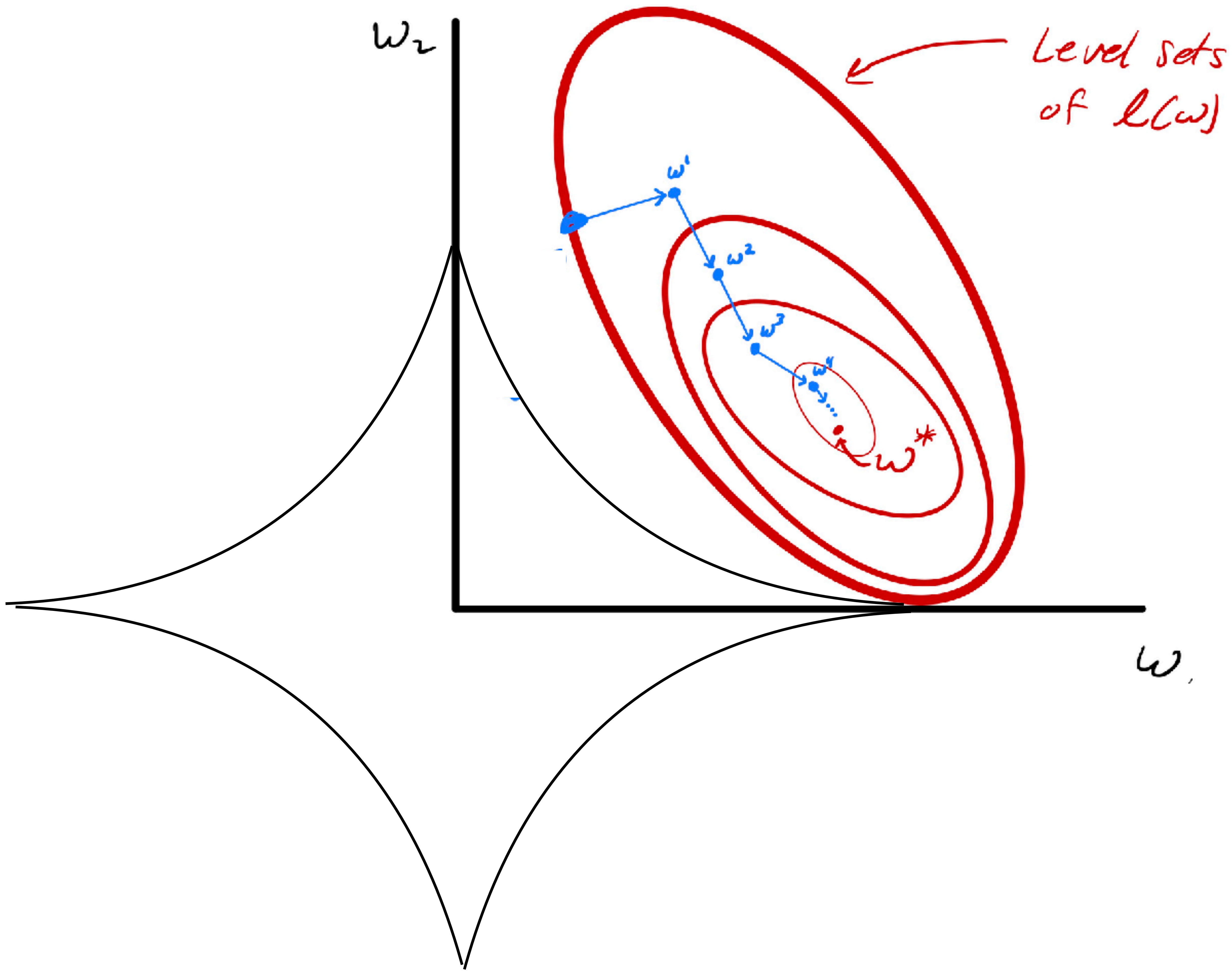


$$\min_w \frac{1}{n} \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

$$\text{s.t. } \|w\|_1 \leq B$$

Advantage: give sparse solution

Linear Regression: squared loss + ℓ_p constraint

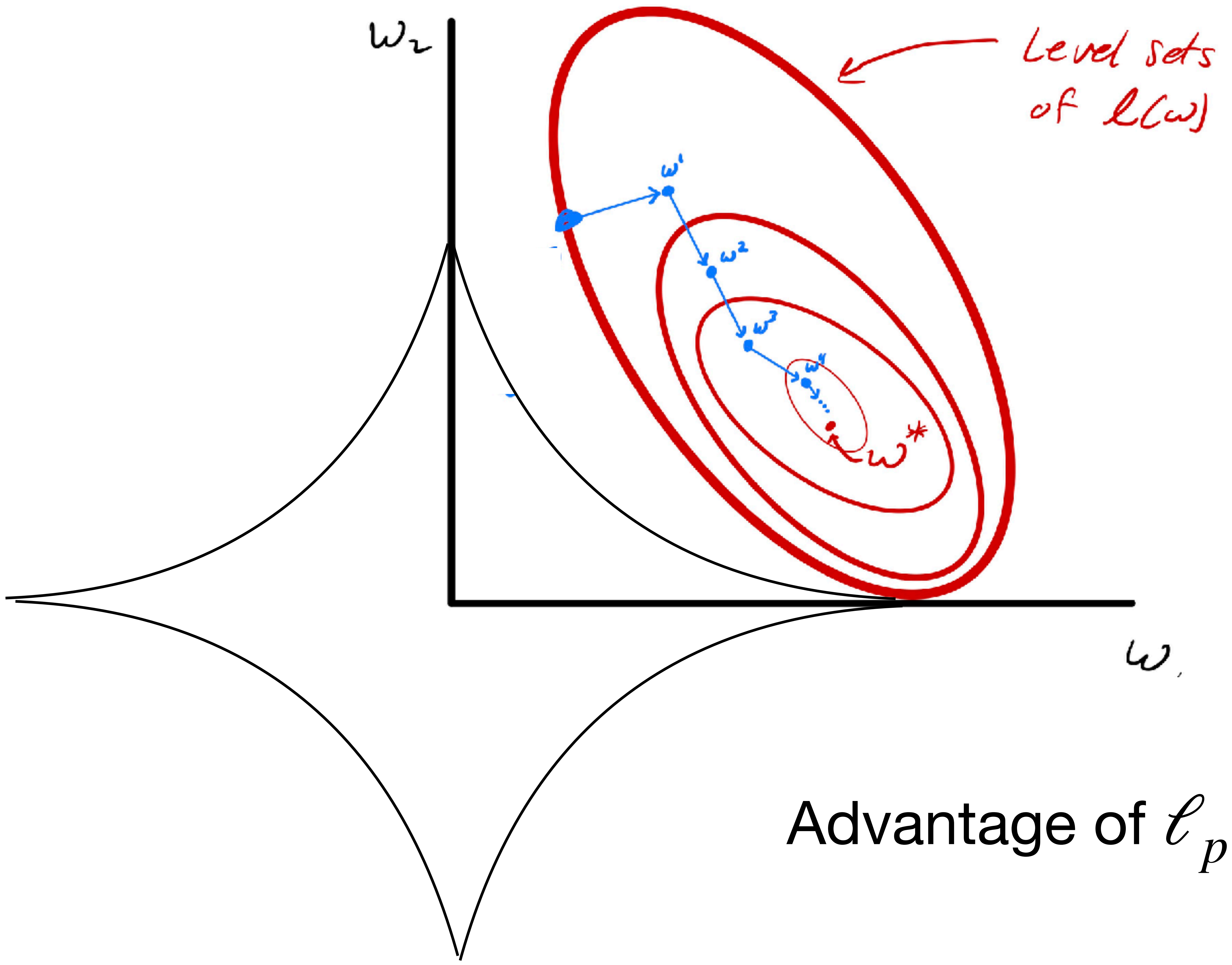


$$\min_w \frac{1}{n} \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

s.t. $\|w\|_p \leq B$

$0 < p < 1$

Linear Regression: squared loss + ℓ_p constraint



$$\min_w \frac{1}{n} \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

s.t. $\|w\|_p \leq B$

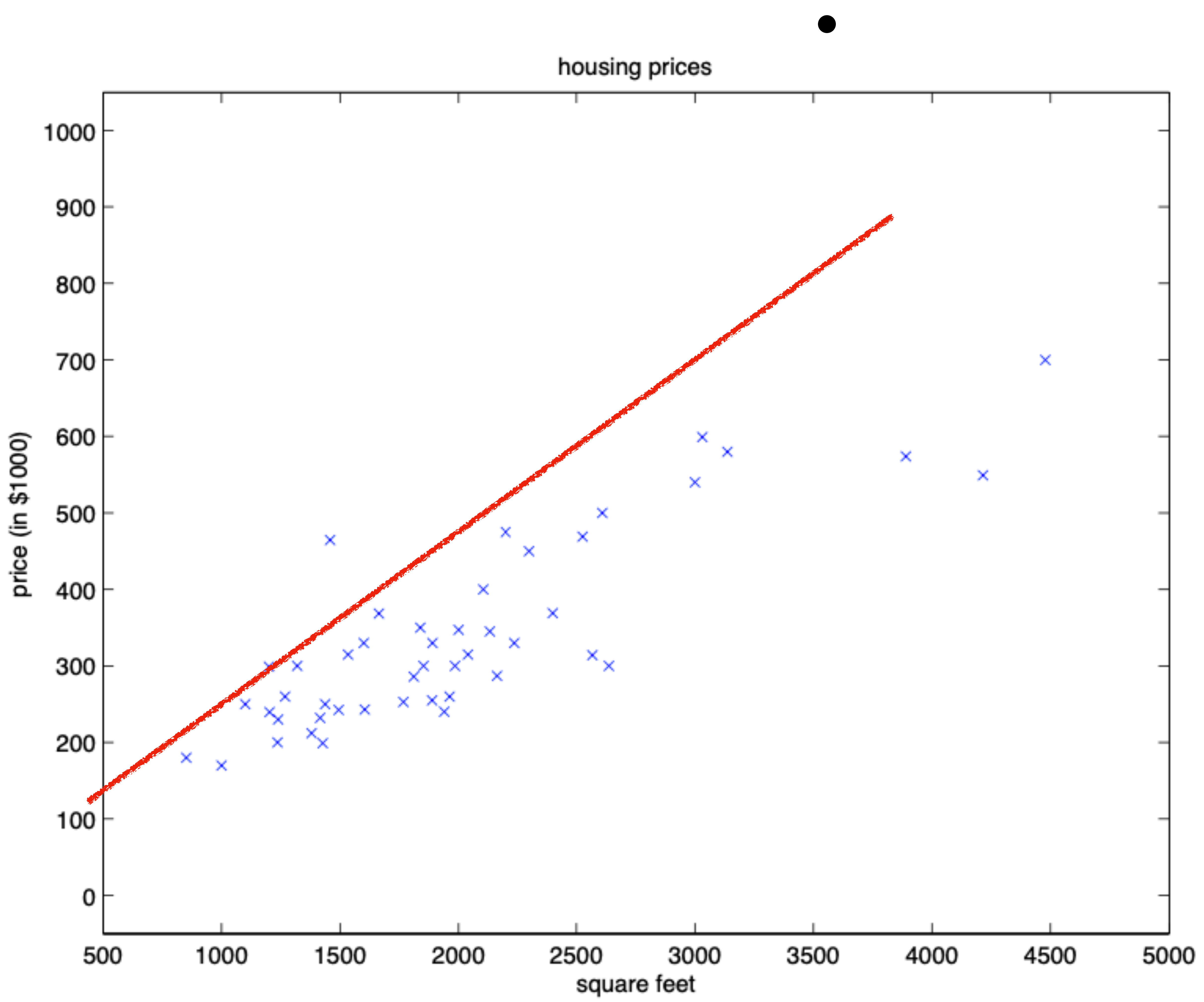
$0 < p < 1$

Advantage of ℓ_p constraint : very sparse solution

Disadvantage: Non-convex

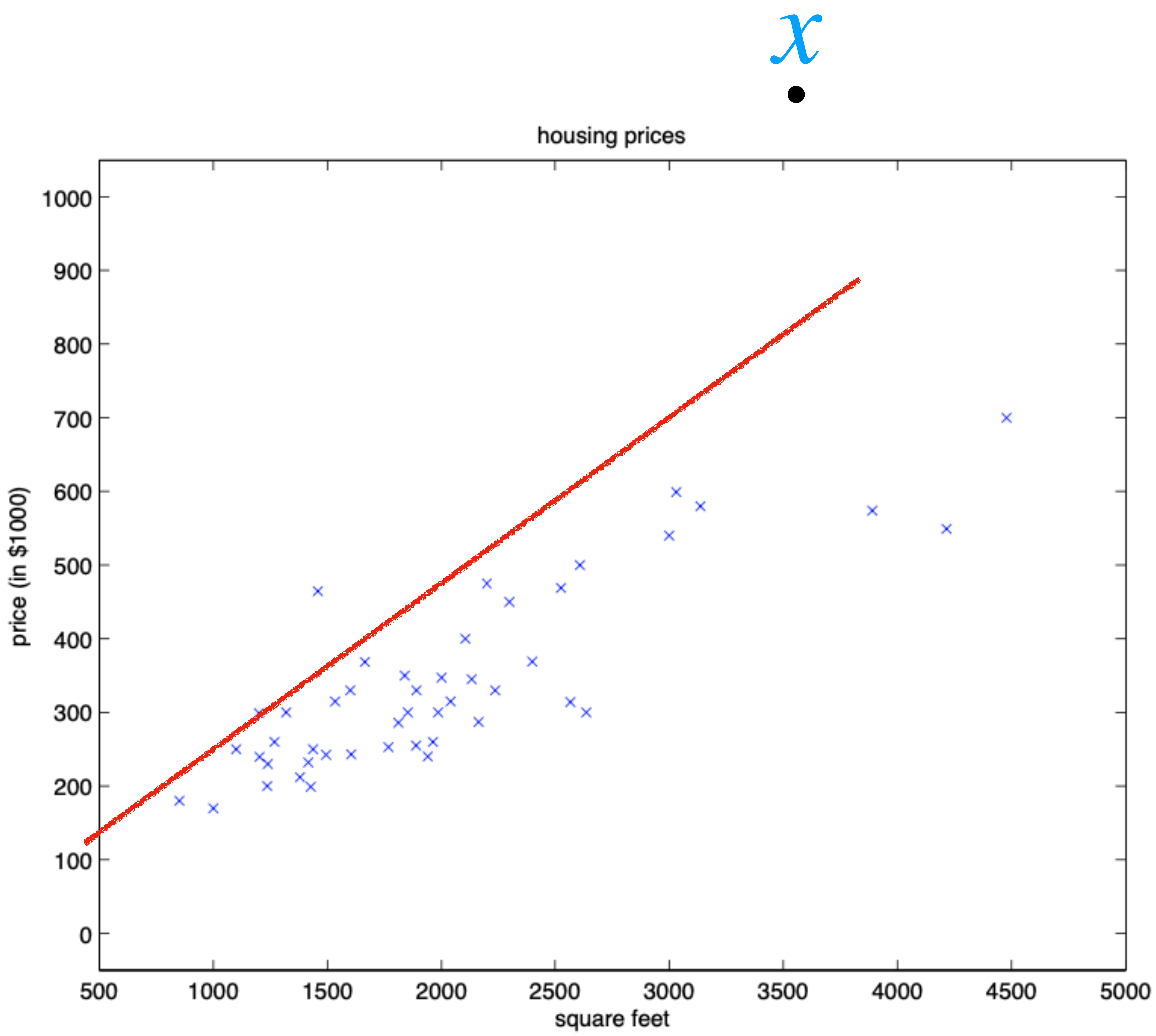
Constraints help avoid overfitting

x



Without constraint, we might overfit to an outlier

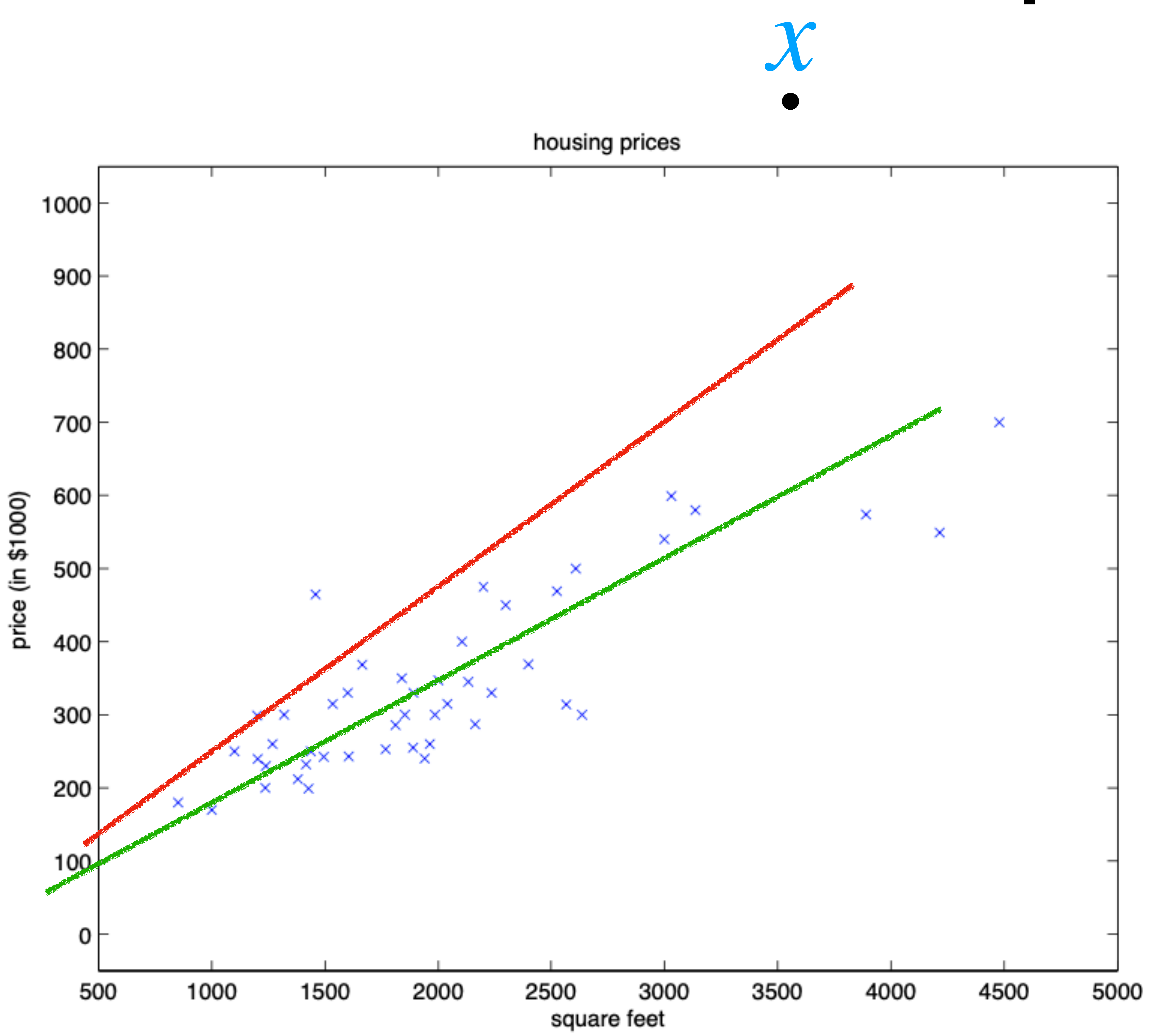
Constraints help avoid overfitting



Without constraint, we might overfit to an outlier

With constraint $||w||_2^2 \leq B$, we can avoid overfitting (i.e., force us to not pay too much attention to minimizing loss)

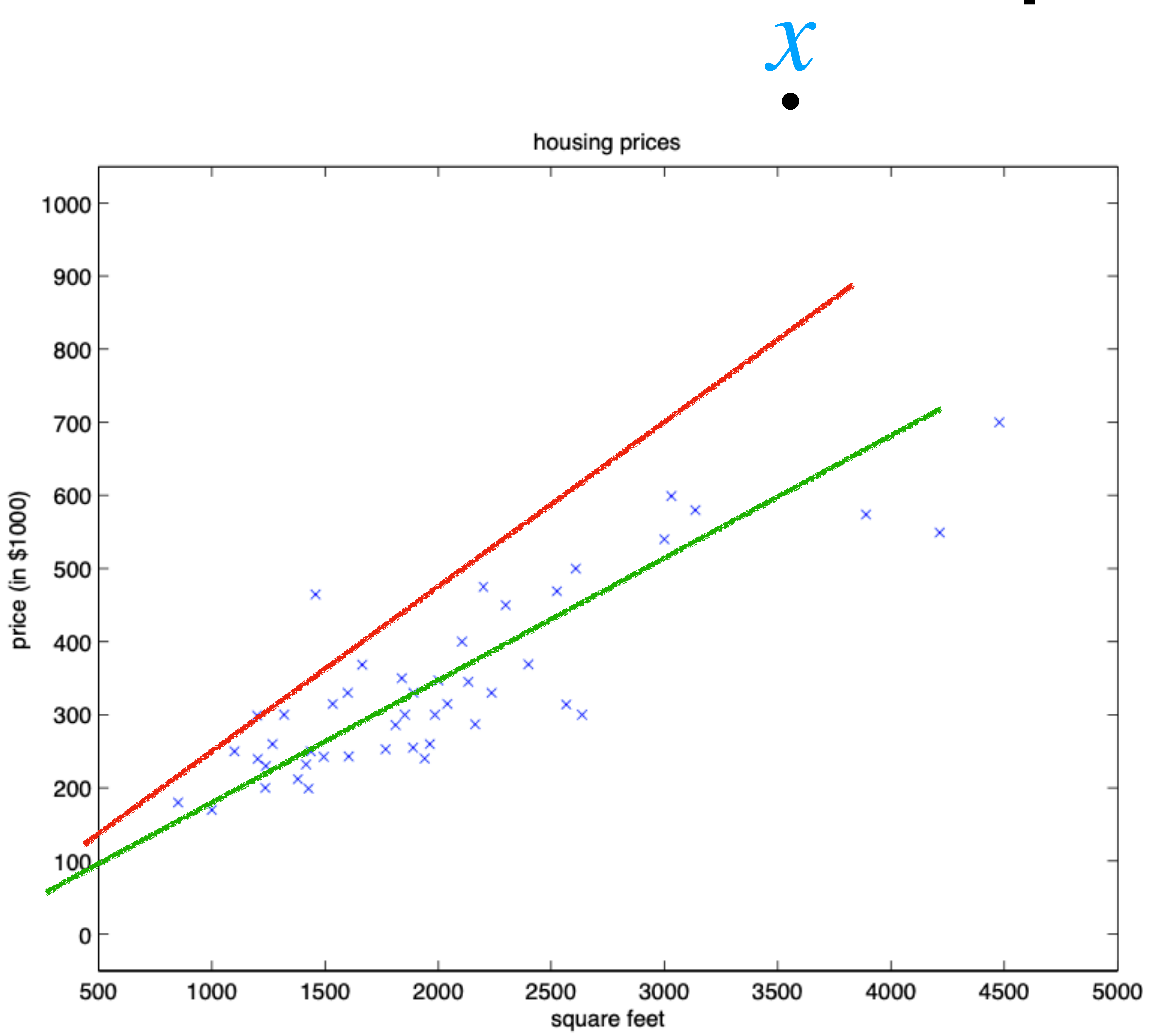
Constraints help avoid overfitting



Without constraint, we might overfit to an outlier

With constraint $\|w\|_2^2 \leq B$, we can avoid overfitting (i.e., force us to not pay too much attention to minimizing loss)

Constraints help avoid overfitting



Without constraint, we might overfit to an outlier

With constraint $||w||_2^2 \leq B$, we can avoid overfitting (i.e., force us to not pay too much attention to minimizing loss)

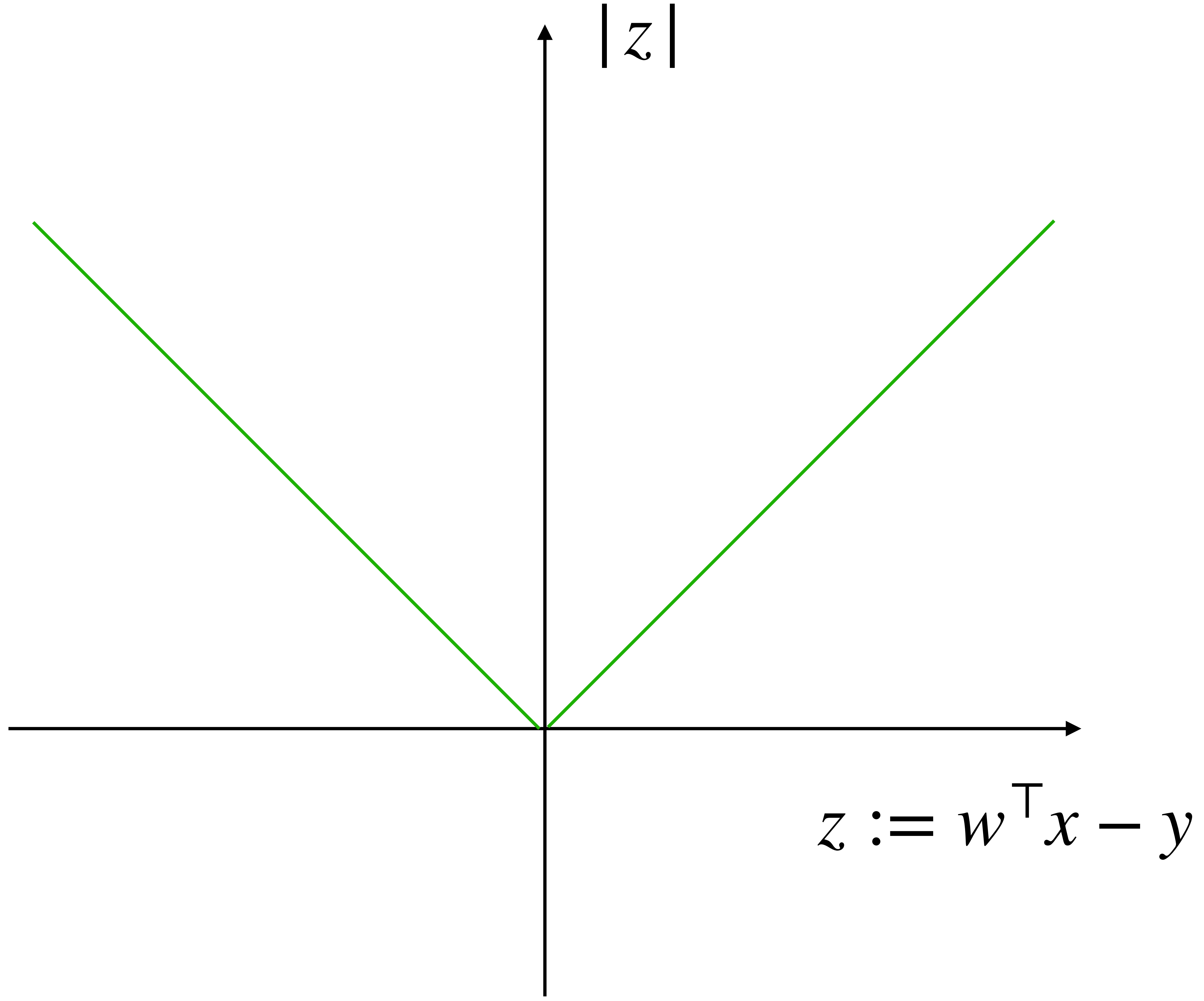
(More details in next lecture)

Other loss functions with linear regression

Absolute loss:

$$\min_w \frac{1}{n} \sum_{i=1}^{n} |w^\top x_i - y_i|$$

s.t. $R(w) \le B$



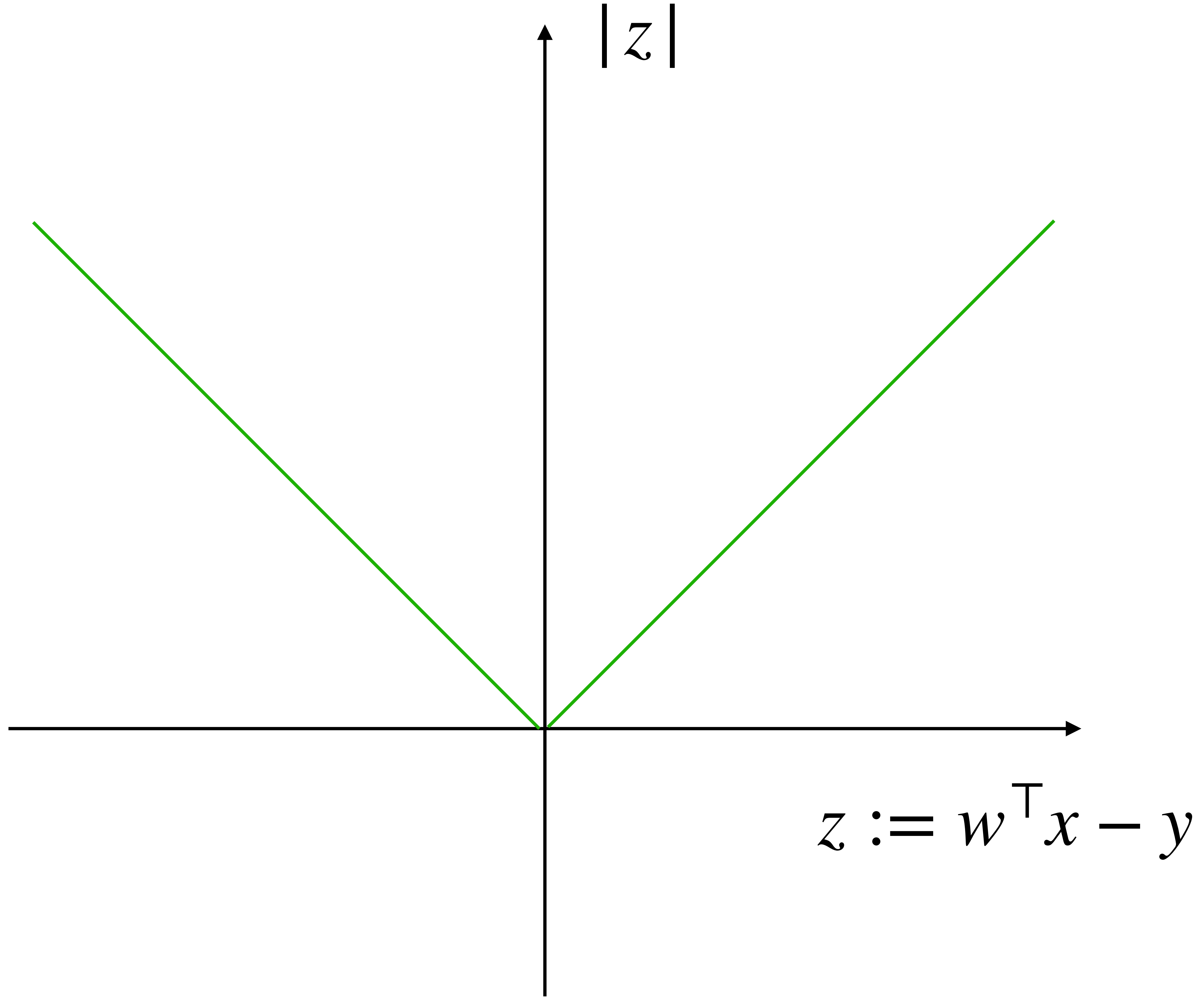
Other loss functions with linear regression

Absolute loss:

$$\min_w \frac{1}{n} \sum_{i=1}^{n} |w^\top x_i - y_i|$$

$$\text{s.t. } R(w) \le B$$

Advantage: less sensitive to outliers

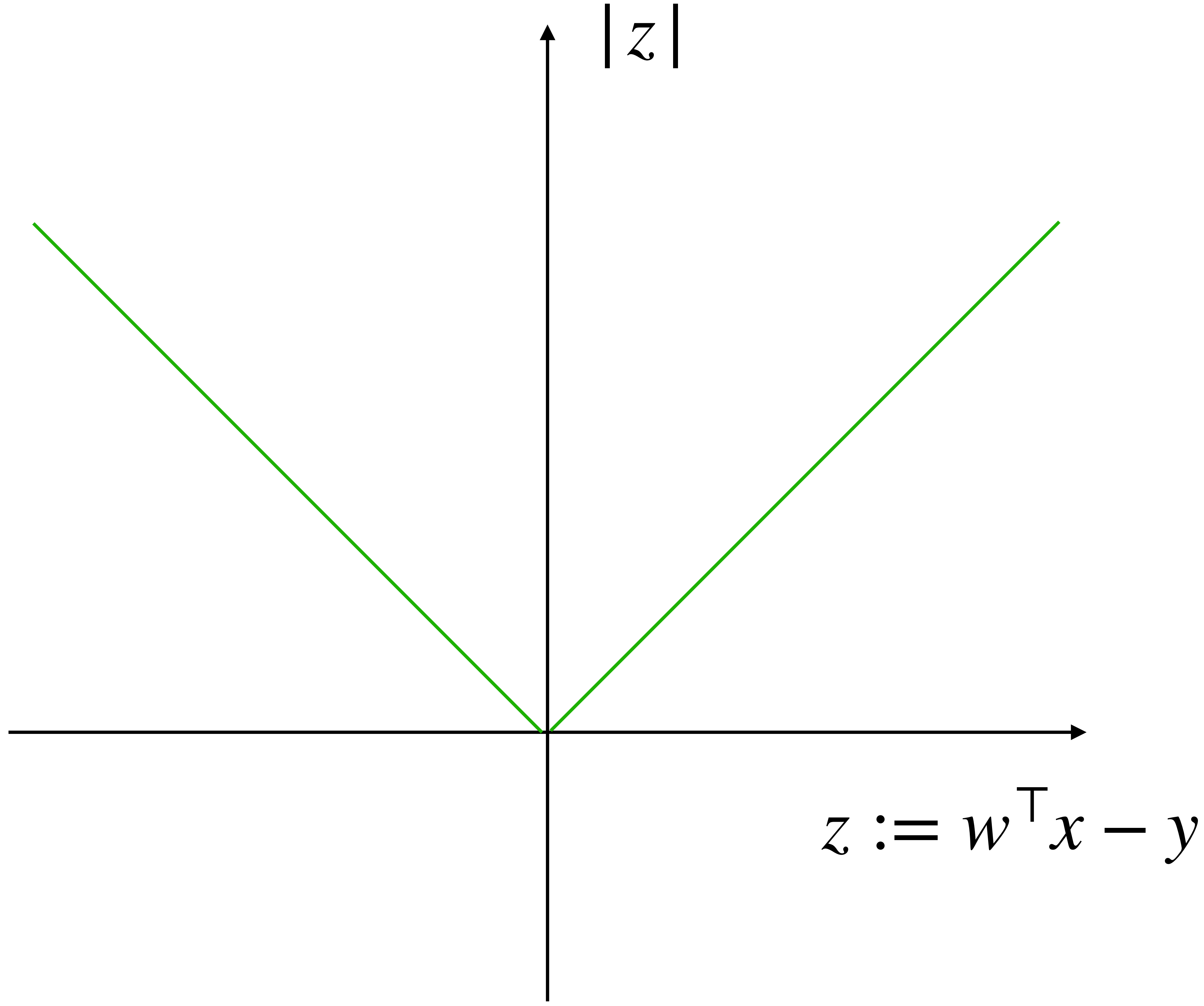


Other loss functions with linear regression

Absolute loss:

$$\min_{w} \frac{1}{n} \sum_{i=1}^{n} |w^\top x_i - y_i|$$

s.t. $R(w) \leq B$



Advantage: less sensitive to outliers

Disadvantage: no closed-form solution, non-differentiable at 0

Other loss functions with linear regression

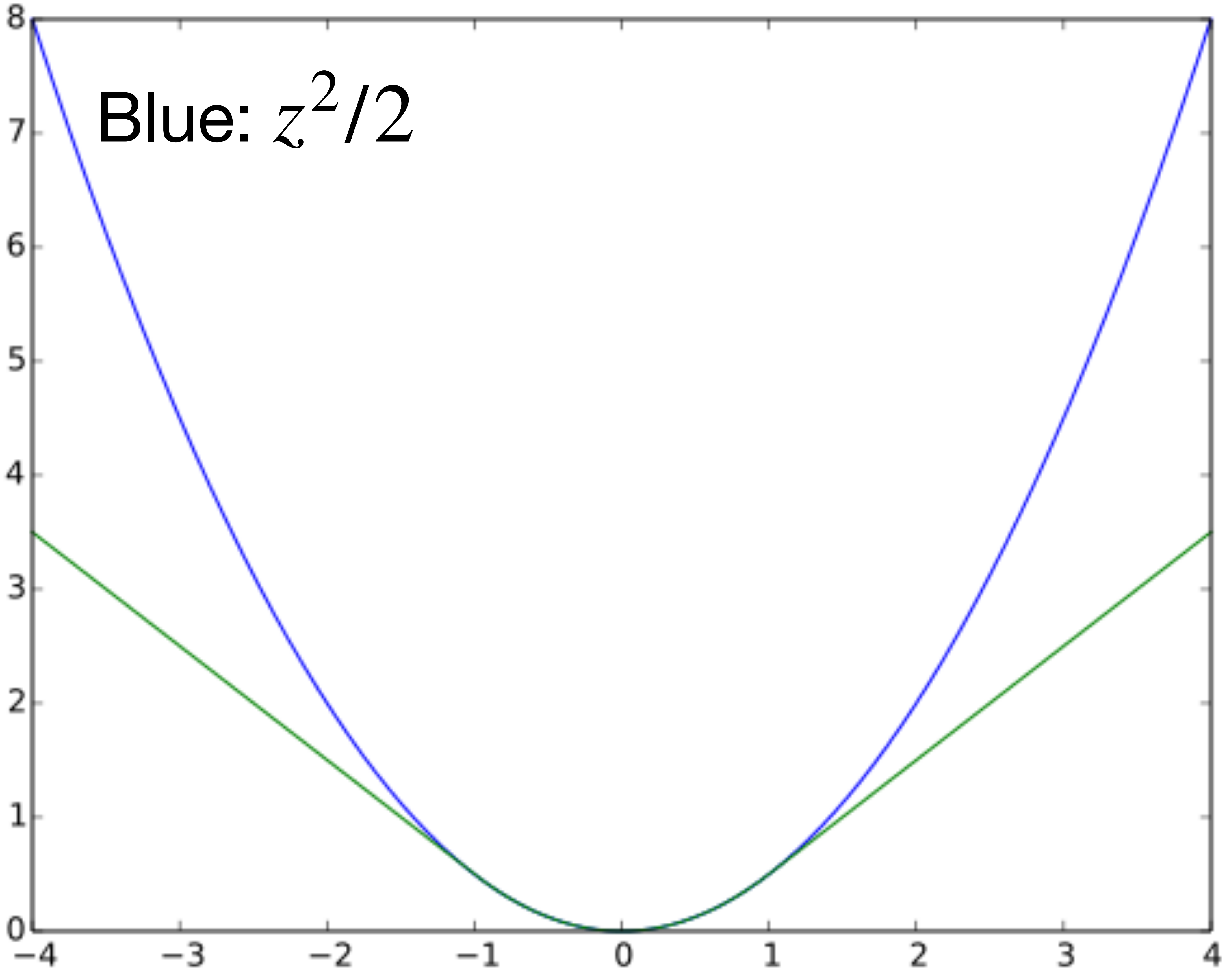
Huber loss:

$$\min_{w} \frac{1}{n} \sum_{i=1}^{n} L_\delta(w^\top x - y)$$

s.t. $R(w) \leq B$

Where

$$L_\delta(z) = \begin{cases} z^2/2 & |z| \leq \delta \\ \delta(|z| - \delta/2) & \text{else} \end{cases}$$



Blue: $z^2/2$

Green: huber with $\delta = 1$

Other loss functions with linear regression

Huber loss:

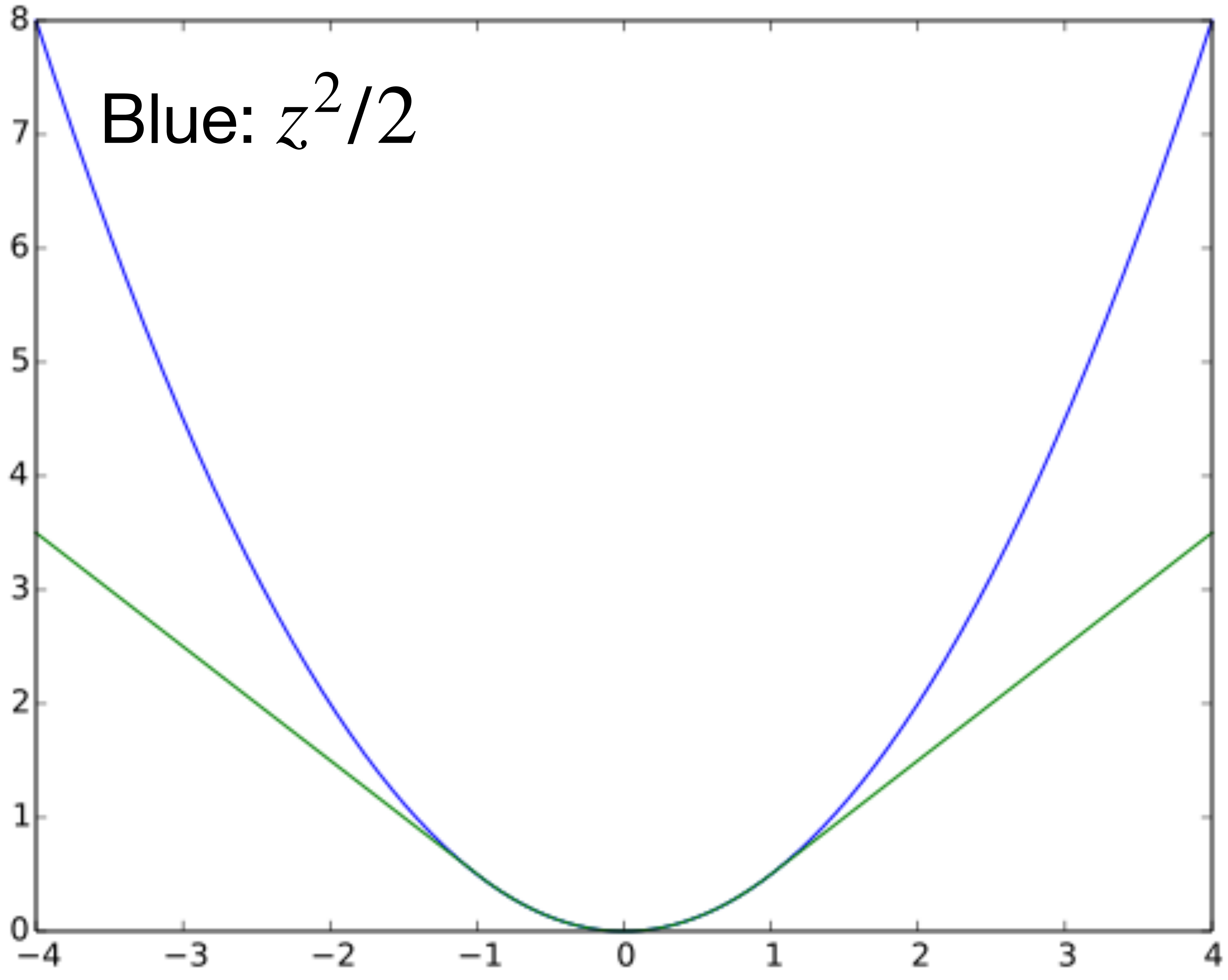
$$\min_w \frac{1}{n} \sum_{i=1}^{n} L_\delta(w^\top x - y)$$

s.t. $R(w) \leq B$

Where

$$L_\delta(z) = \begin{cases} z^2/2 & |z| \leq \delta \\ \delta(|z| - \delta/2) & \text{else} \end{cases}$$

Advantage: best of both worlds



Green: huber with $\delta = 1$

Other loss functions with linear regression

Huber loss:

$$\min_w \frac{1}{n} \sum_{i=1}^{n} L_\delta(w^\top x - y)$$

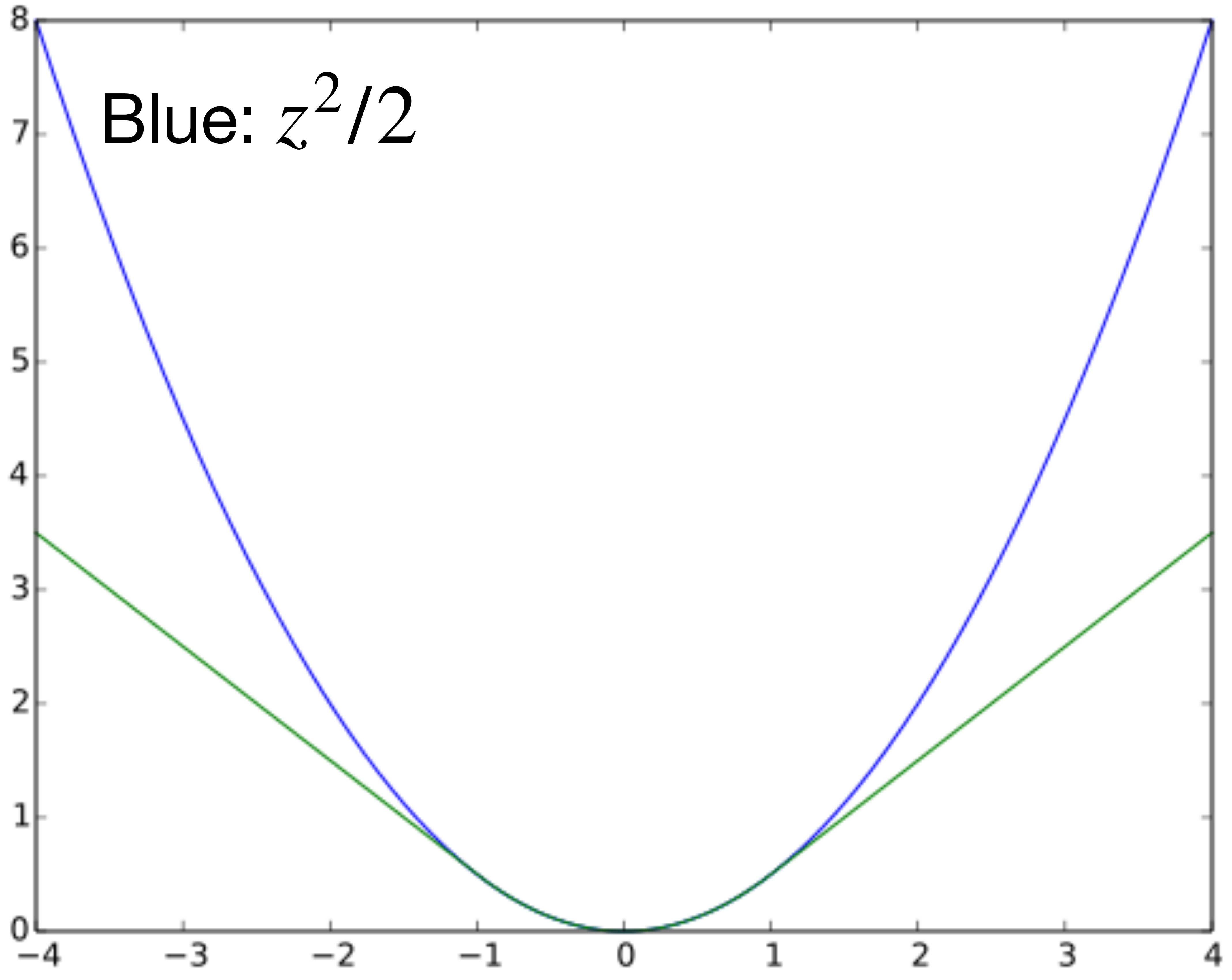
s.t. $R(w) \le B$

Where

$$L_\delta(z) = \begin{cases} z^2/2 & |z| \le \delta \\ \delta(|z| - \delta/2) & \text{else} \end{cases}$$

Advantage: best of both worlds

Disadvantage: additional parameter δ to tune



Green: huber with $\delta = 1$

Linear classification: Hinge loss + constraint

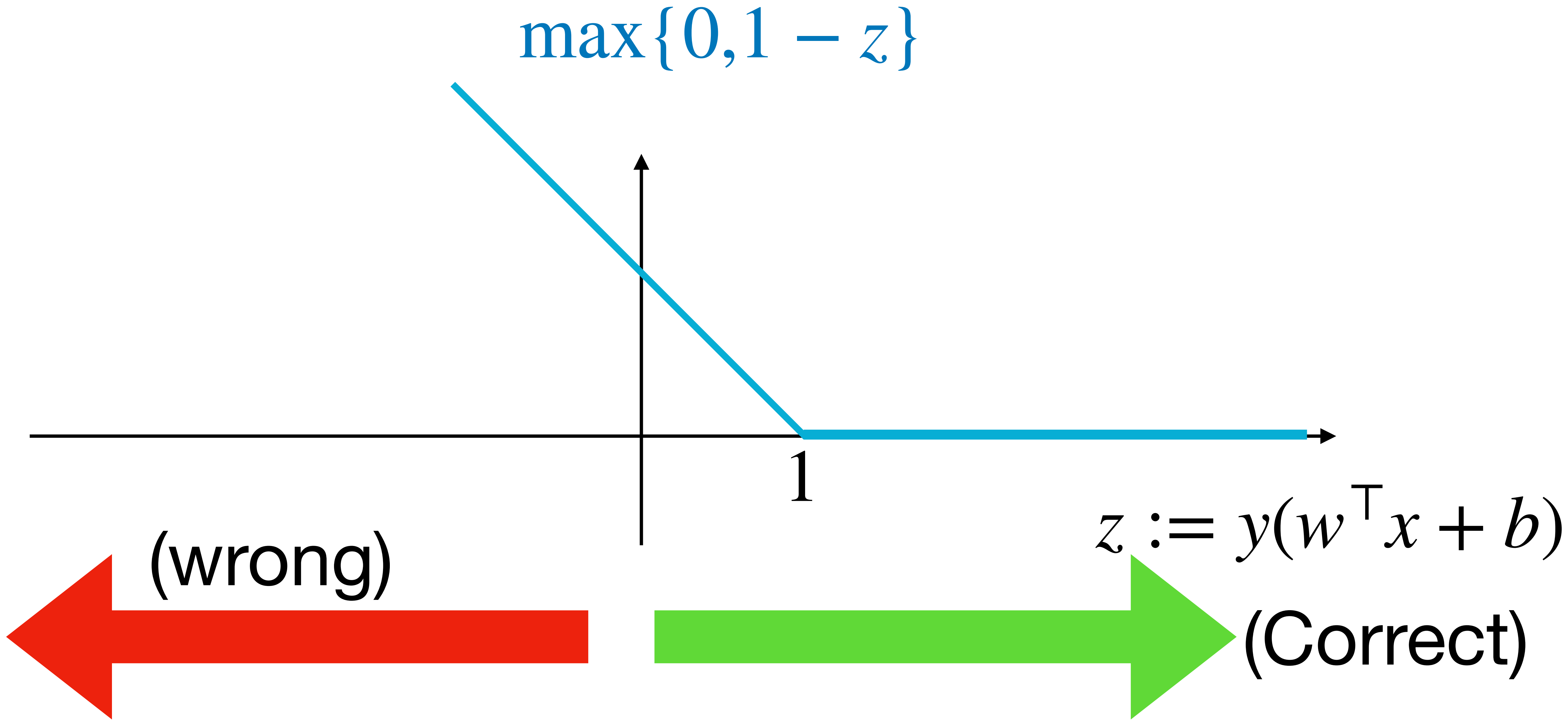
$$\min_{w,b} \frac{1}{n} \sum_{i=1}^{n} \max\left\{0,\, 1 - y_i(w^\top x_i + b)\right\}$$

$$\text{s.t.} \quad \|w\|_2^2 \leq B$$

Linear classification: Hinge loss + constraint

$$\min_{w,b} \frac{1}{n} \sum_{i=1}^{n} \max\left\{0,\, 1 - y_i(w^\top x_i + b)\right\}$$

$$\text{s.t.} \quad \|w\|_2^2 \le B$$

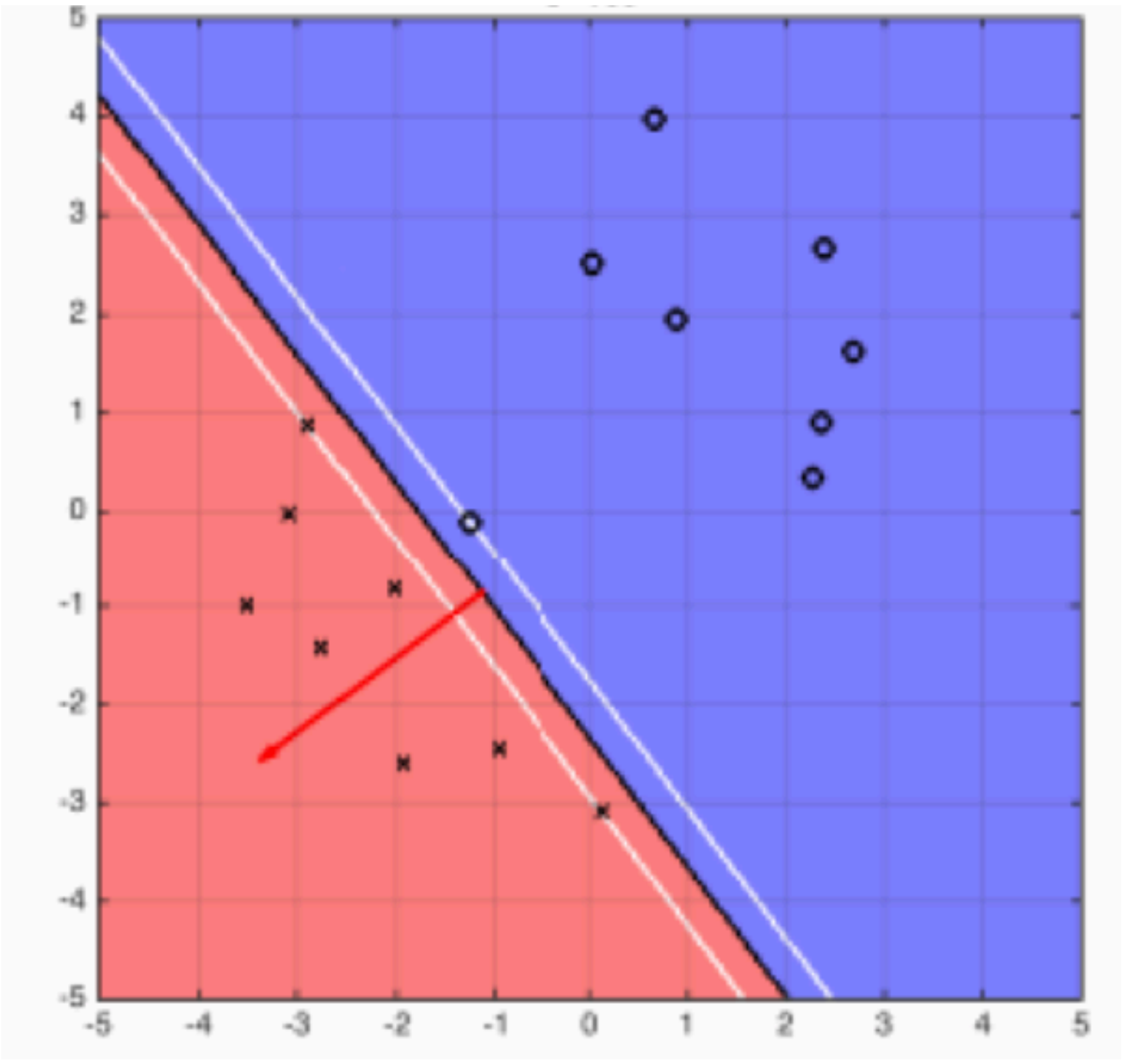


Linear classification: Hinge loss + constraint

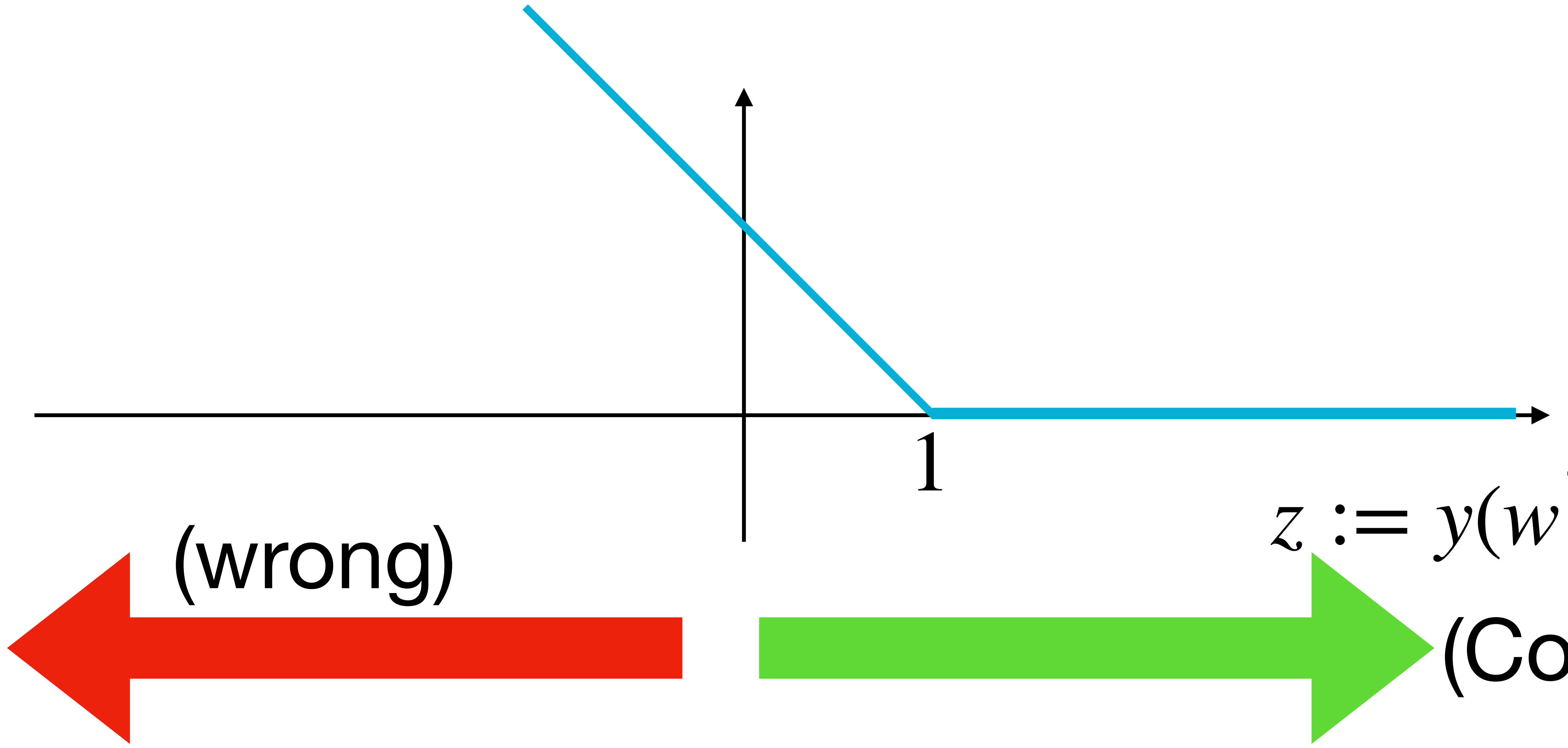
$$\min_{w,b} \frac{1}{n} \sum_{i=1}^{n} \max\left\{0,\ 1 - y_i(w^\top x_i + b)\right\}$$

$$\text{s.t. } \|w\|_2^2 \le B$$

Constraint avoids overfit:
(Recall: small $\|w\|_2$ should have large street width)



$\max\{0,1-z\}$



Linear classification: Log-loss + constraints

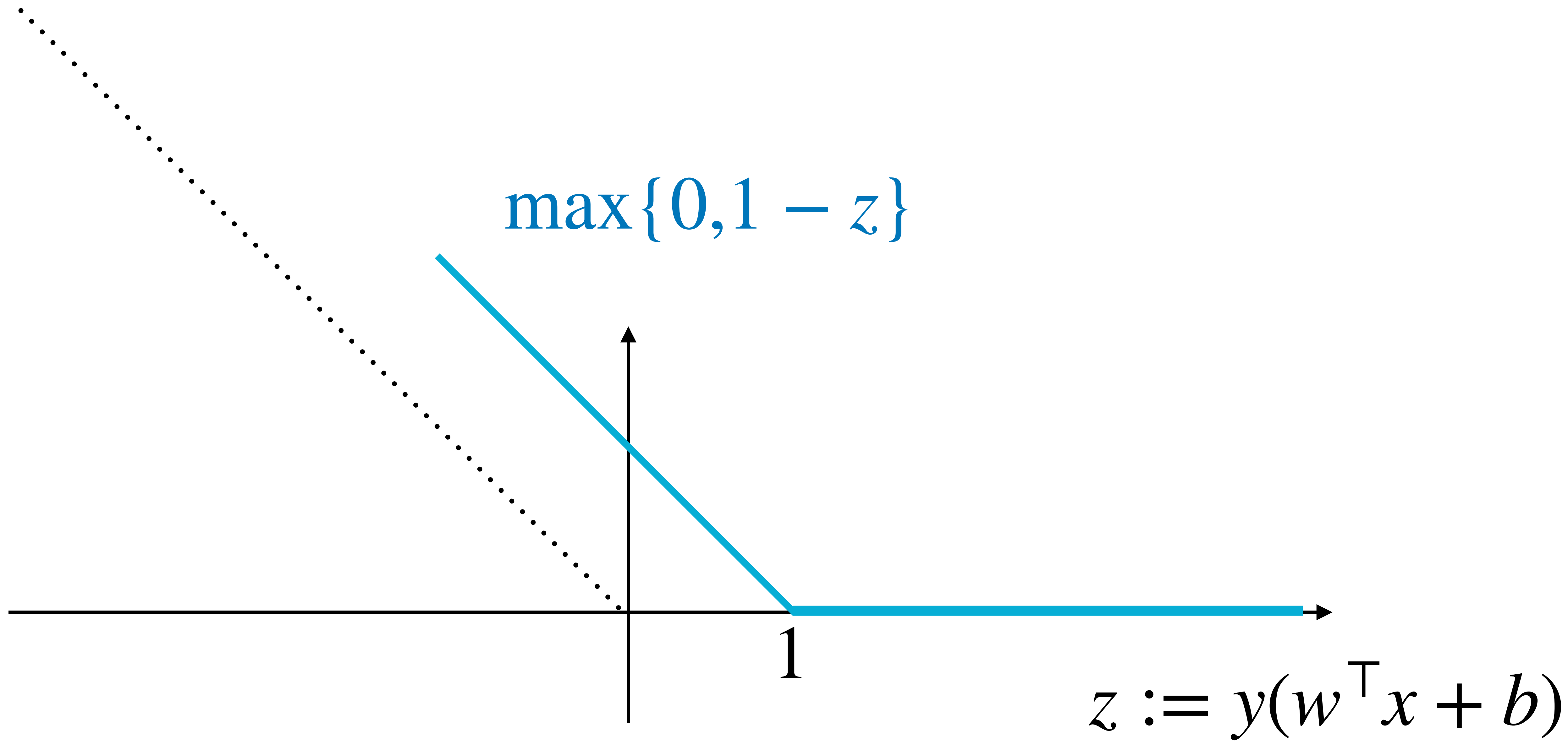
$$\min_{w,b} \frac{1}{n} \sum_{i=1}^{n} \ln \left(1 + \exp(-y_i(w^\top x_i + b))\right)$$

$$\text{s.t.} \quad \|w\|_2^2 \leq B$$

Linear classification: Log-loss + constraints

$$\min_{w,b} \frac{1}{n} \sum_{i=1}^{n} \ln\left(1 + \exp(-y_i(w^\top x_i + b))\right)$$

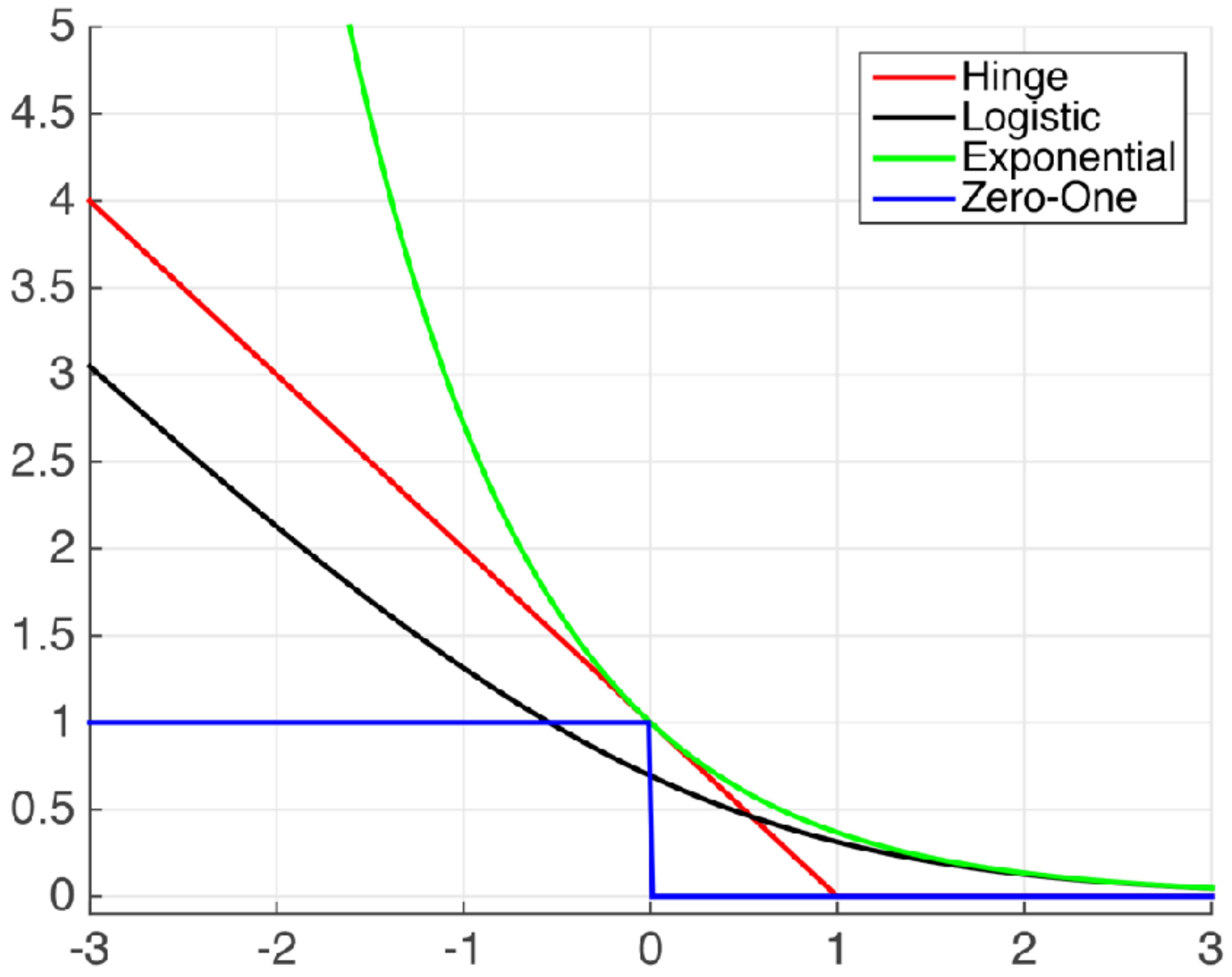
$$\text{s.t.} \quad \|w\|_2^2 \le B$$



Linear classification: Exponential loss + constraints

$$\min_{w,b} \frac{1}{n} \sum_{i=1}^{n} \exp\left(-y_i(w^\top x_i + b)\right)$$

$$\text{s.t. } \|w\|_2^2 \leq B$$

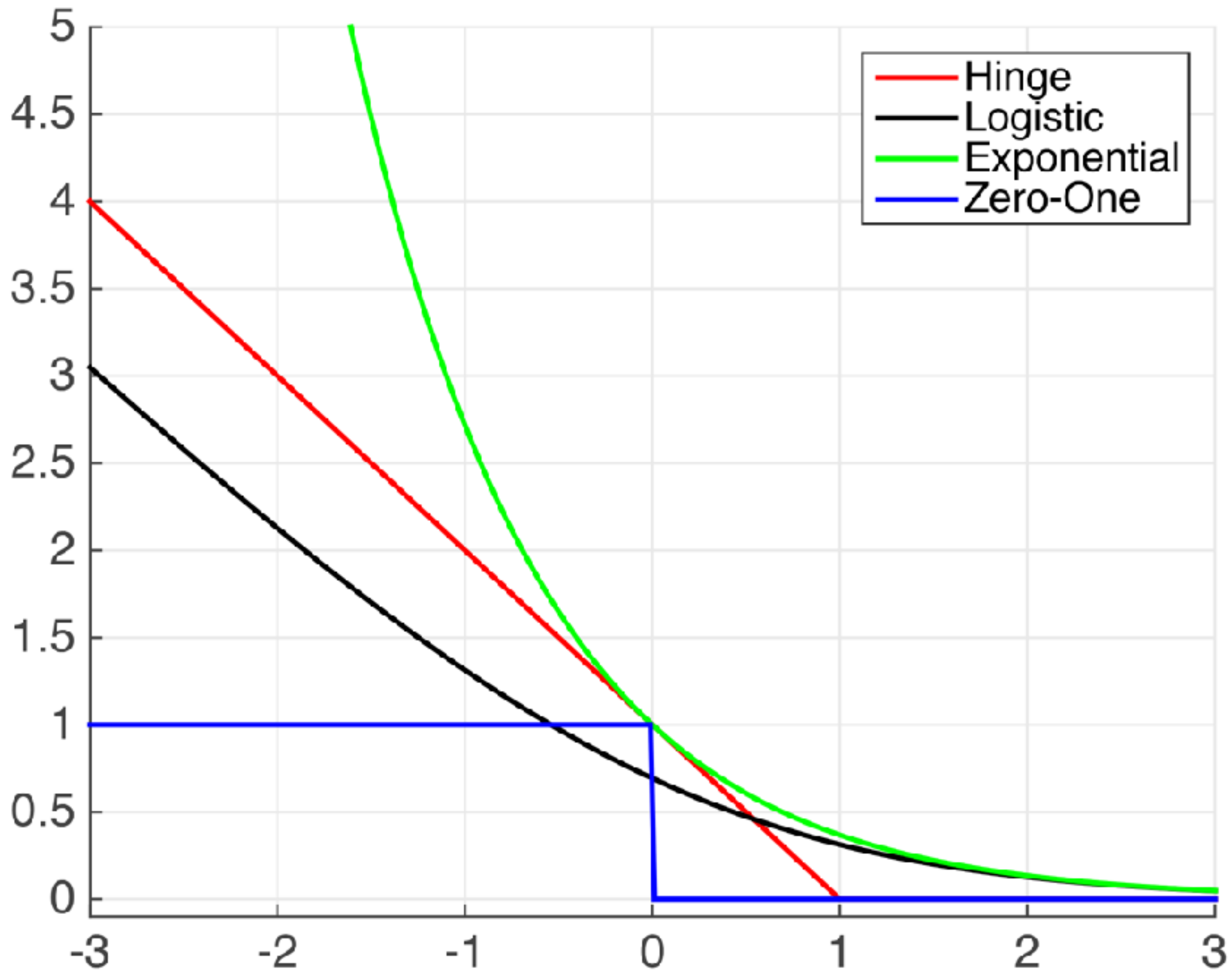


Linear classification: Exponential loss + constraints

$$\min_{w,b} \frac{1}{n} \sum_{i=1}^{n} \exp\left(-y_i(w^\top x_i + b)\right)$$

s.t. $\|w\|_2^2 \leq B$

(Later, AdaBoost uses this loss)



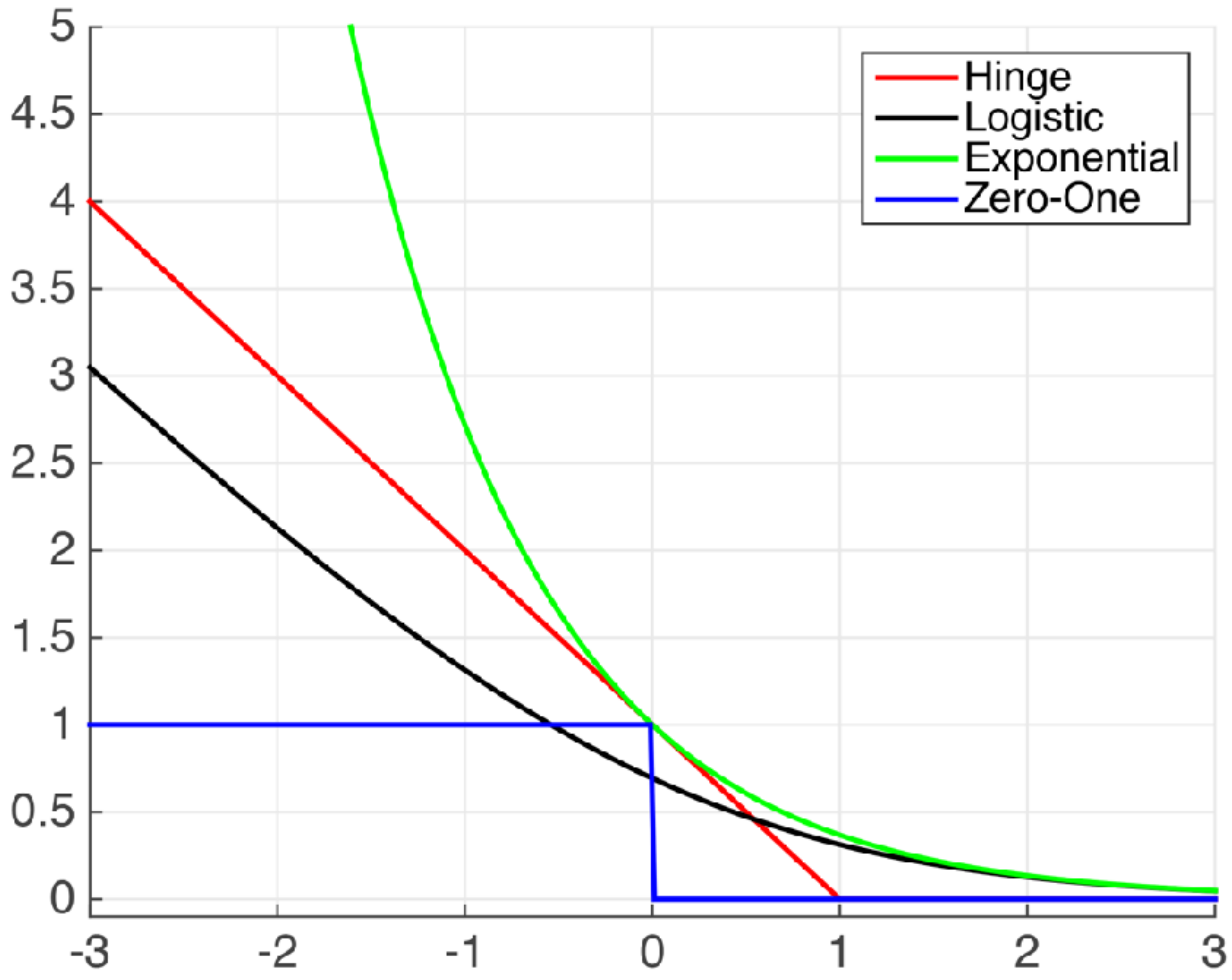
Linear classification: Exponential loss + constraints

$$\min_{w,b} \frac{1}{n} \sum_{i=1}^{n} \exp\left(-y_i(w^\top x_i + b)\right)$$

$$\text{s.t.} \quad \|w\|_2^2 \leq B$$

(Later, AdaBoost uses this loss)

Very aggressive loss (but may overfit w/ noisy data)



Outline for Today

1. Empirical Risk Minimization

2. Examples on loss & hypothesis classes

3. Regularization

Regularization

We can turn constraint optimization problem into unconstrained using Lagrange multiplier

Example:

Regularization

We can turn constraint optimization problem into unconstrained using Lagrange multiplier

Example:

$$\min_w \frac{1}{n} \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

s.t. $||w||_1 \leq B$

Regularization

We can turn constraint optimization problem into unconstrained using Lagrange multiplier

Example:

$$\min_w \frac{1}{n}\sum_{i=1}^{n}(w^\top x_i - y_i)^2$$

s.t. $\|w\|_1 \leq B$

$$\Longleftrightarrow$$

$$\min_w \frac{1}{n}\sum_{i=1}^{n}(w^\top x_i - y_i)^2 + \lambda\|w\|_2^2$$

Regularization

We can turn constraint optimization problem into unconstrained using Lagrange multiplier

Example:

$$\min_w \frac{1}{n} \sum_{i=1}^{n} (w^\top x_i - y_i)^2 \quad \text{s.t. } \|w\|_1 \leq B \quad \Longleftrightarrow \quad \min_w \frac{1}{n} \sum_{i=1}^{n} (w^\top x_i - y_i)^2 + \lambda \|w\|_2^2$$

(More details about Lagrange multiplier in Anil's optimization class CS4220)

Examples:

Soft-margin SVM:

$$\min_{w,b} \sum_{i=1}^{n} \max\left\{0, 1 - y_i(w^\top x_i + b)\right\} + \lambda \|w\|_2^2$$

Examples:

Soft-margin SVM:

$$\min_{w,b} \sum_{i=1}^{n} \max\left\{0, 1 - y_i(w^\top x_i + b)\right\} + \lambda \|w\|_2^2$$

Ridge Linear Regression

$$\min_{w} \sum_{i=1}^{n} (w^\top x_i - y_i)^2 + \lambda \|w\|_2^2$$

Examples:

Soft-margin SVM:

$$\min_{w,b} \sum_{i=1}^{n} \max\left\{0, 1 - y_i(w^\top x_i + b)\right\} + \lambda \|w\|_2^2$$

Ridge Linear Regression

$$\min_{w} \sum_{i=1}^{n} (w^\top x_i - y_i)^2 + \lambda \|w\|_2^2$$

Lasso:

$$\min_{w} \sum_{i=1}^{n} (w^\top x_i - y_i)^2 + \lambda \|w\|_1$$

Examples:

Soft-margin SVM:

$$\min_{w,b} \sum_{i=1}^{n} \max\left\{0, 1 - y_i(w^\top x_i + b)\right\} + \lambda \|w\|_2^2$$

Ridge Linear Regression

$$\min_{w} \sum_{i=1}^{n} (w^\top x_i - y_i)^2 + \lambda \|w\|_2^2$$

Lasso:

$$\min_{w} \sum_{i=1}^{n} (w^\top x_i - y_i)^2 + \lambda \|w\|_1$$

Returned solution is often sparse!

Examples:

Soft-margin SVM:

$$\min_{w,b} \sum_{i=1}^{n} \max\left\{0, 1 - y_i(w^\top x_i + b)\right\} + \lambda \|w\|_2^2$$

Ridge Linear Regression

$$\min_{w} \sum_{i=1}^{n} (w^\top x_i - y_i)^2 + \lambda \|w\|_2^2$$

Lasso:

$$\min_{w} \sum_{i=1}^{n} (w^\top x_i - y_i)^2 + \lambda \|w\|_1$$

Returned solution is often sparse!

Good for feature selection!

Summary for today

1. Empirical risk minimization framework

Summary for today

1. Empirical risk minimization framework

2. Need to restrict our hypothesis class:

Select hypothesis that is simple while can also explain the data reasonably well

Summary for today

1. Empirical risk minimization framework

2. Need to restrict our hypothesis class:

Select hypothesis that is simple while can also explain the data reasonably well

3. Examples of loss functions & Regularizations