Low ℓ_1 norm and guarantees on Sparsifiability

Shai Shalev-Shwartz & Nathan Srebro
Toyota Technologica Institute–Chicago

ICML/COLT/UAI workshop, July 2008

Motivation

Problem 1:

$$\mathbf{w}_0 \;=\; \operatorname*{argmin}_{\mathbf{w}} \quad \mathbb{E}[L(\langle \mathbf{w}, \mathbf{x} \rangle, y)] \quad \text{s.t.} \quad \|\mathbf{w}\|_0 \leq S$$

Motivation

Problem I:

$$\mathbf{w}_0 = \operatorname*{argmin}_{\mathbf{w}} \mathbb{E}[L(\langle \mathbf{w}, \mathbf{x} \rangle, y)] \quad \text{s.t.} \quad \cancel{\|\mathbf{w}\|_0 \leq S}$$

Problem II:

$$\mathbf{w}_1 = \operatorname*{argmin}_{\mathbf{w}} \mathbb{E}[L(\langle \mathbf{w}, \mathbf{x} \rangle, y)] \quad \text{s.t.} \quad \boxed{\|\mathbf{w}\|_1 \leq B}$$

Motivation

Problem I:

$$\mathbf{w}_0 = \operatorname*{argmin}_{\mathbf{w}} \ \mathbb{E}[L(\langle \mathbf{w}, \mathbf{x} \rangle, y)] \quad \text{s.t.} \quad \cancel{\|\mathbf{w}\|_0 \leq S}$$

Problem II:

$$\mathbf{w}_1 = \operatorname*{argmin}_{\mathbf{w}} \ \mathbb{E}[L(\langle \mathbf{w}, \mathbf{x} \rangle, y)] \quad \text{s.t.} \quad \boxed{\|\mathbf{w}\|_1 \leq B}$$

- Strict assumptions on data distribution $\Rightarrow$ $\mathbf{w}_1$ is also sparse
- But, what if $\mathbf{w}_1$ is not sparse ?

Motivation

Problem I:

$$\mathbf{w}_0 = \operatorname*{argmin}_{\mathbf{w}} \; \mathbb{E}[L(\langle \mathbf{w}, \mathbf{x} \rangle, y)] \quad \text{s.t.} \quad \cancel{\|\mathbf{w}\|_0 \leq S}$$

Problem II:

$$\mathbf{w}_1 = \operatorname*{argmin}_{\mathbf{w}} \; \mathbb{E}[L(\langle \mathbf{w}, \mathbf{x} \rangle, y)] \quad \text{s.t.} \quad \boxed{\|\mathbf{w}\|_1 \leq B}$$

features not correlated

- Strict assumptions on data distribution $\Rightarrow$ $\mathbf{w}_1$ is also sparse
- But, what if $\mathbf{w}_1$ is not sparse ?

Sparsification

Predictor $\mathbf{w}$ with $\|\mathbf{w}\|_1 = B$

$\downarrow$

Sparsification procedure

$\downarrow$

Predictor $\tilde{\mathbf{w}}$ with $\|\tilde{\mathbf{w}}\|_0 = S$

Sparsification

$$\text{Predictor } \mathbf{w} \text{ with } \|\mathbf{w}\|_1 = B$$

$$\downarrow$$

Sparsification procedure

$$\downarrow$$

$$\text{Predictor } \tilde{\mathbf{w}} \text{ with } \|\tilde{\mathbf{w}}\|_0 = S$$

- **Constraint:** $\mathbb{E}[L(\langle \tilde{\mathbf{w}}, \mathbf{x} \rangle, y)] \leq \mathbb{E}[L(\langle \mathbf{w}, \mathbf{x} \rangle, y)] + \epsilon$
- **Goal:** Minimal S that satisfies constraint
- **Question:** How S depends on B and ϵ ?

Main Result

- Theorem:
  - For any predictor $\mathbf{w}$, λ-Lipschitz loss function L, distribution D over $\mathcal{X} \times Y$, desired accuracy ϵ
  - Exists $\tilde{\mathbf{w}}$ s.t. $\mathbb{E}[L(\langle \tilde{\mathbf{w}}, \mathbf{x} \rangle, y)] \leq \mathbb{E}[L(\langle \mathbf{w}, \mathbf{x} \rangle, y)] + \epsilon$ and

$$\boxed{\|\tilde{\mathbf{w}}\|_0 = O\left(\left(\frac{\lambda \|\mathbf{w}\|_1}{\epsilon}\right)^2\right)}$$

- Tightness:
  - Data distribution, loss function, dense predictor $\mathbf{w}$ with loss l, but need $\Omega((\|\mathbf{w}\|_1^2/\epsilon)^2)$ features for loss $l + \epsilon$
  - Sparsifying by taking largest weights or following ℓ_1 regularization path might fail
  - Low ℓ_2 norm predictor $\not\Rightarrow$ sparse predictor

Main Result (cont.)

- Distribution D
- Loss L

Main Result (cont.)

- Distribution D
- Loss L

Convex opt.

Low ℓ_1 predictor $\mathbf{w}$

Main Result (cont.)

- Distribution D
- Loss L

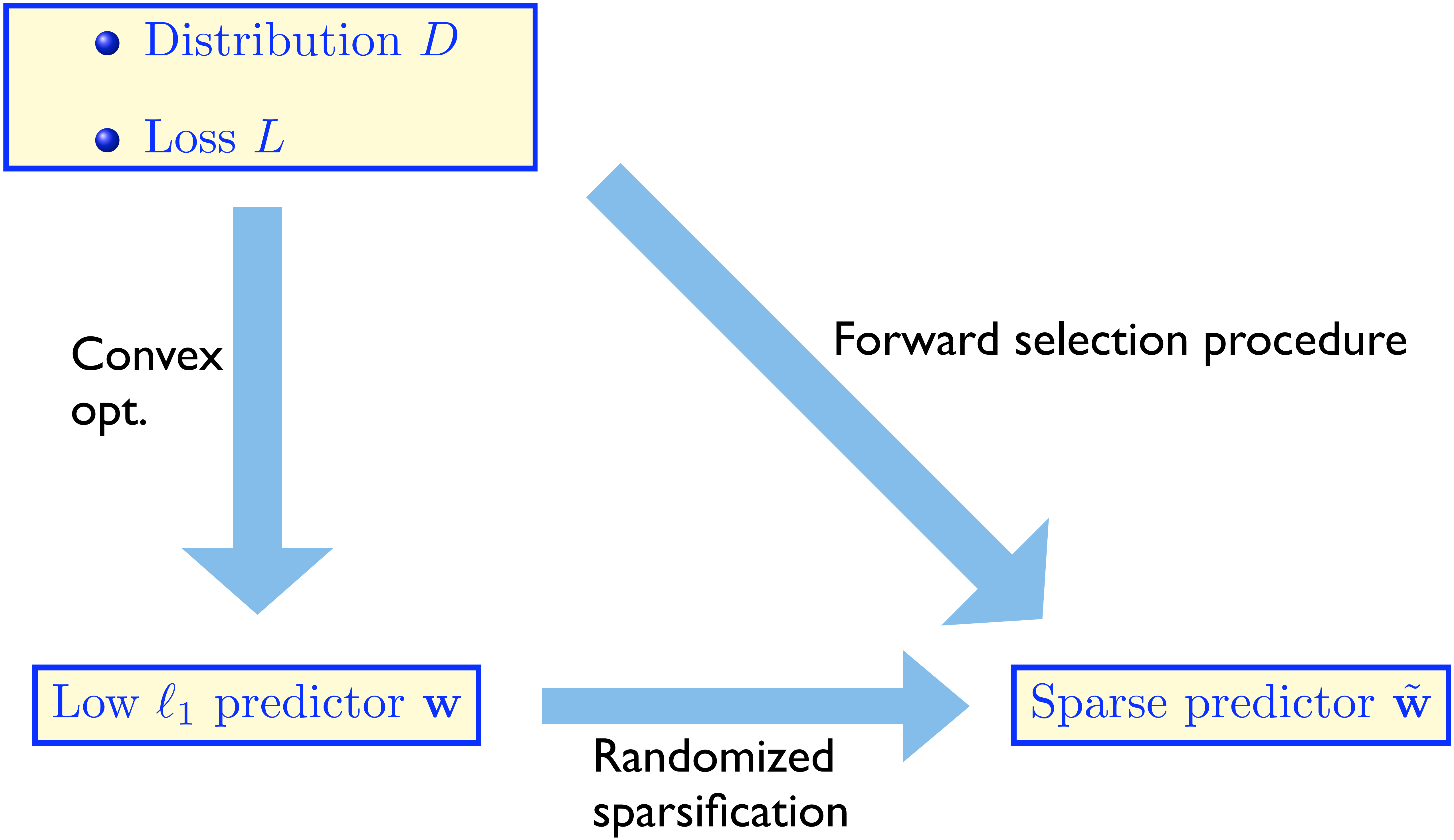
Convex opt.

Low ℓ_1 predictor $\mathbf{w}$

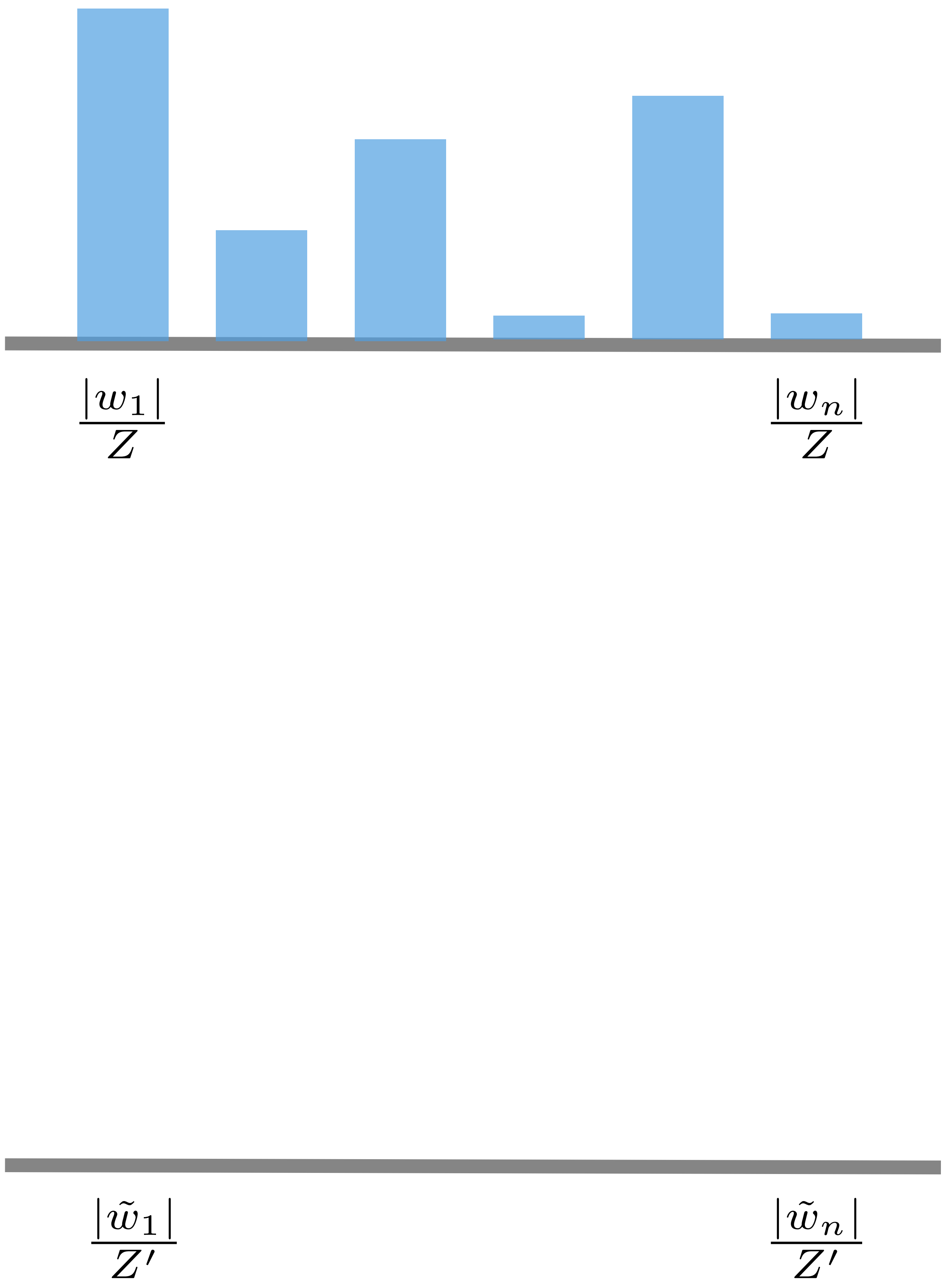
Randomized sparsification

Sparse predictor $\tilde{\mathbf{w}}$

Main Result (cont.)



Randomized Sparsification Procedure



Sparsification Procedure

For $j = 1, \ldots, S$

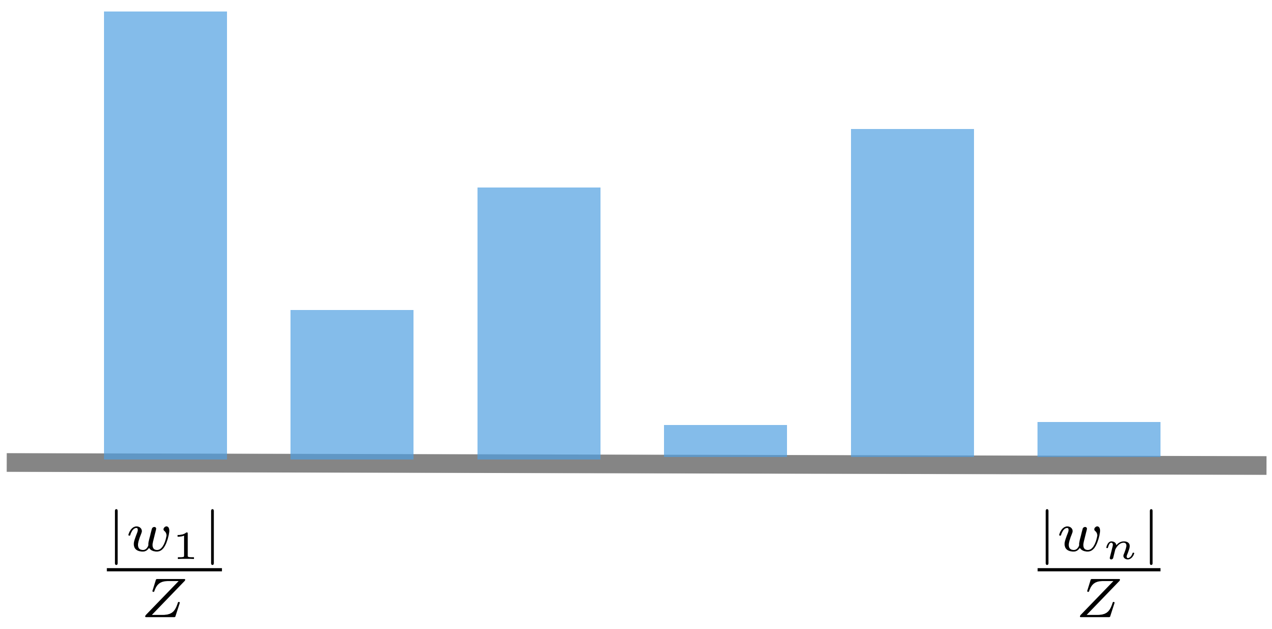
- Sample r_i from distribution $P_i \propto |w_i|$
- Add $|\tilde{w}_i| \leftarrow |\tilde{w}_i| + 1$

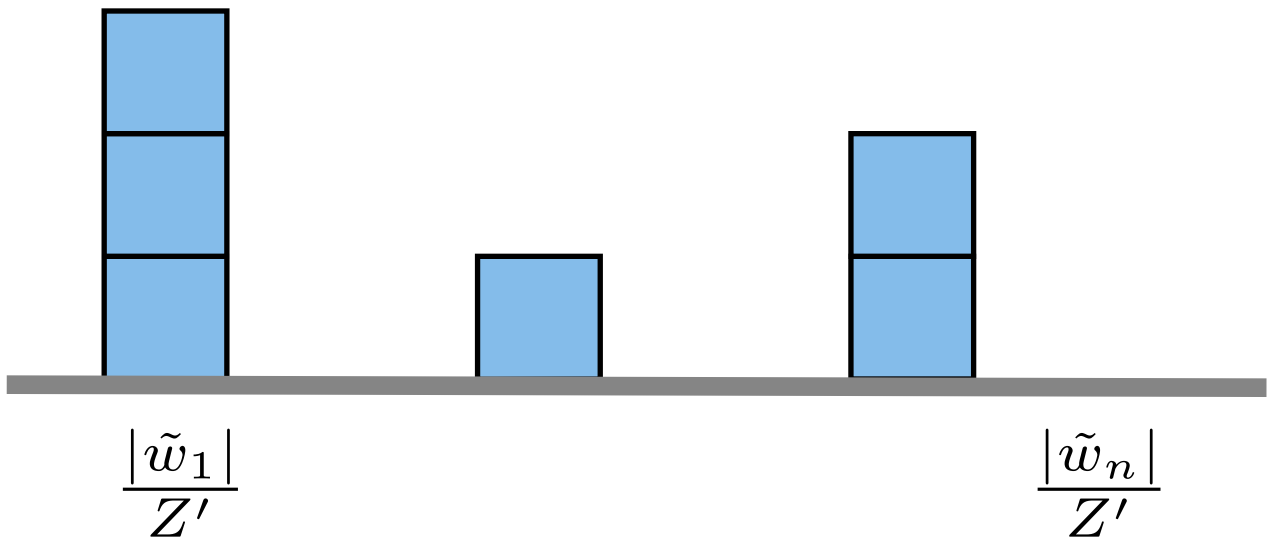
Randomized Sparsification Procedure

Sparsification Procedure

For $j = 1, \ldots, S$

- Sample r_i from distribution $P_i \propto |w_i|$
- Add $|\tilde{w}_i| \leftarrow |\tilde{w}_i| + 1$





Randomized Sparsification Procedure

Sparsification Procedure

For $j = 1, \ldots, S$

- Sample r_i from distribution $P_i \propto |w_i|$
- Add $|\tilde{w}_i| \leftarrow |\tilde{w}_i| + 1$

Guarantee

- Assume: $\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\|_\infty \leq 1\}$, Y = arbitrary set, $\mathcal{D}$ = arbitrary distribution over $\mathcal{X} \times Y$, Loss $L : \mathbb{R} \times Y \to \mathbb{R}$ is λ-Lipschitz w.r.t. 1st argument
- If: $S \geq \Omega\left(\frac{\lambda^2 \|\mathbf{w}\|_1^2 \log(1/\delta)}{\epsilon^2}\right)$
- Then, with probability at least $1 - \delta$, $\mathbb{E}[L(\langle \tilde{\mathbf{w}}, \mathbf{x} \rangle, y)] - \mathbb{E}[L(\langle \mathbf{w}, \mathbf{x} \rangle, y)] \leq \epsilon$

Randomized Sparsification Procedure

- Distribution D
- Loss L

Convex opt.

Low ℓ_1 predictor $\mathbf{w}$

- Requires access to w
- Does not require access to D

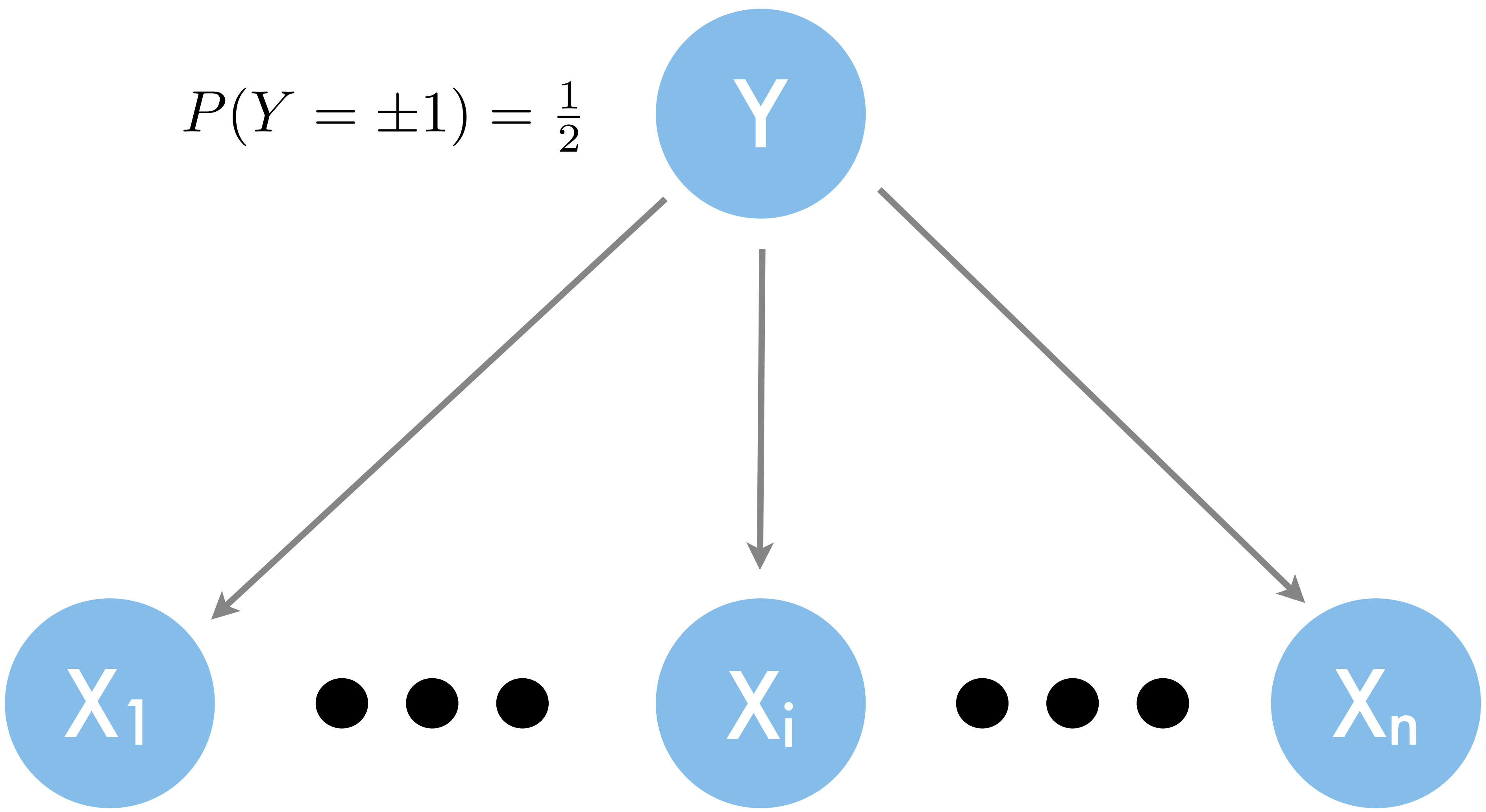
Randomized sparsification

Sparse predictor $\tilde{\mathbf{w}}$

Tightness

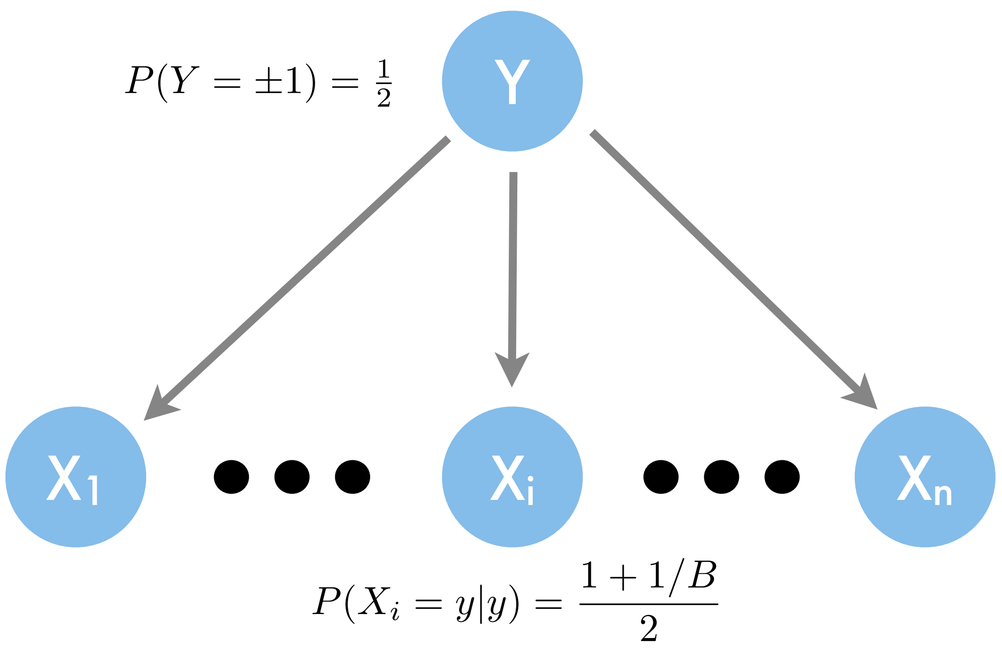
Data distribution: spread ‘information’ about label among all features

$P(Y = \pm 1) = \frac{1}{2}$



$$P(X_i = y|y) = \frac{1 + 1/B}{2}$$

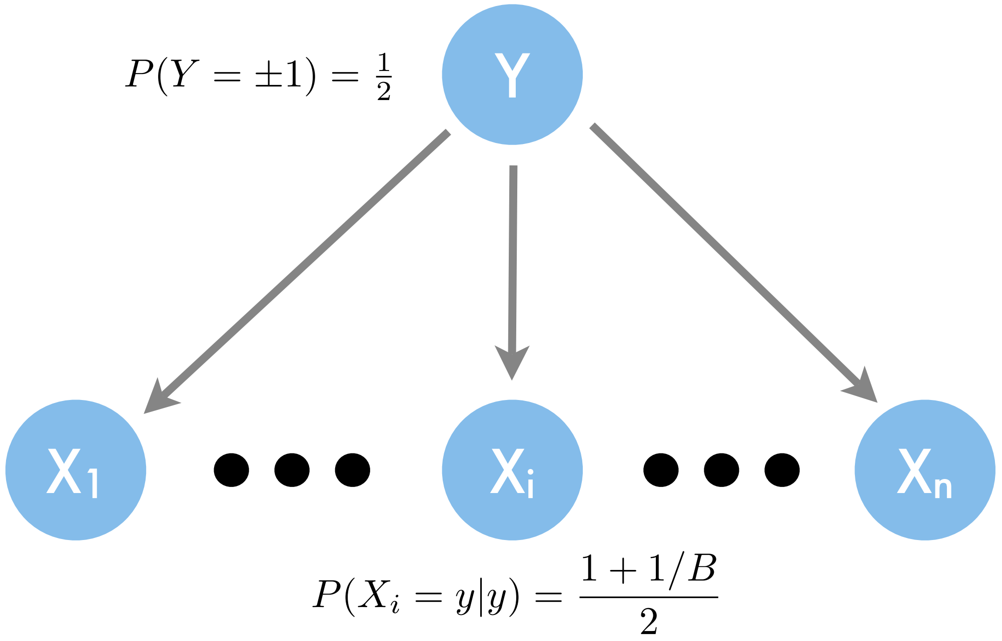
Tightness (cont.)



- Dense predictor:
  - $w_i = \frac{B}{n}$ and thus $\|\mathbf{w}\|_1 = B$
  - $\mathbb{E}[|\langle \mathbf{w}, \mathbf{x}\rangle - y|] \leq \frac{B}{\sqrt{n}}$
- Sparse predictor:
  - Any $\mathbf{u}$ with $\mathbb{E}[|\langle \mathbf{u}, \mathbf{x}\rangle - y|] \leq \epsilon$ must satisfy:

$$\|\mathbf{u}\|_0 = \Omega\left(\frac{B^2}{\epsilon^2}\right)$$

Tightness (cont.)



- Dense predictor:
  - $w_i = \frac{B}{n}$ and thus $\|\mathbf{w}\|_1 = B$
  - $\mathbb{E}[|\langle \mathbf{w}, \mathbf{x} \rangle - y|] \leq \frac{B}{\sqrt{n}}$
- Sparse predictor:
  - Any $\mathbf{u}$ with $\mathbb{E}[|\langle \mathbf{u}, \mathbf{x} \rangle - y|] \leq \epsilon$ must satisfy:

$$\|\mathbf{u}\|_0 = \Omega\left(\frac{B^2}{\epsilon^2}\right)$$

Proof uses a generalization of Khintchine inequality:
If $\mathbf{x} = (x_1, \ldots, x_n)$ are independent random variables with $\mathcal{P}[x_k = 1] \in (5\%, 95\%)$ and Q is degree d polynomial, then:

$$\mathbb{E}[|Q(\mathbf{x})|] \geq (0.2)^d \, \mathbb{E}[|Q(\mathbf{x})|^2]^{\frac{1}{2}}$$

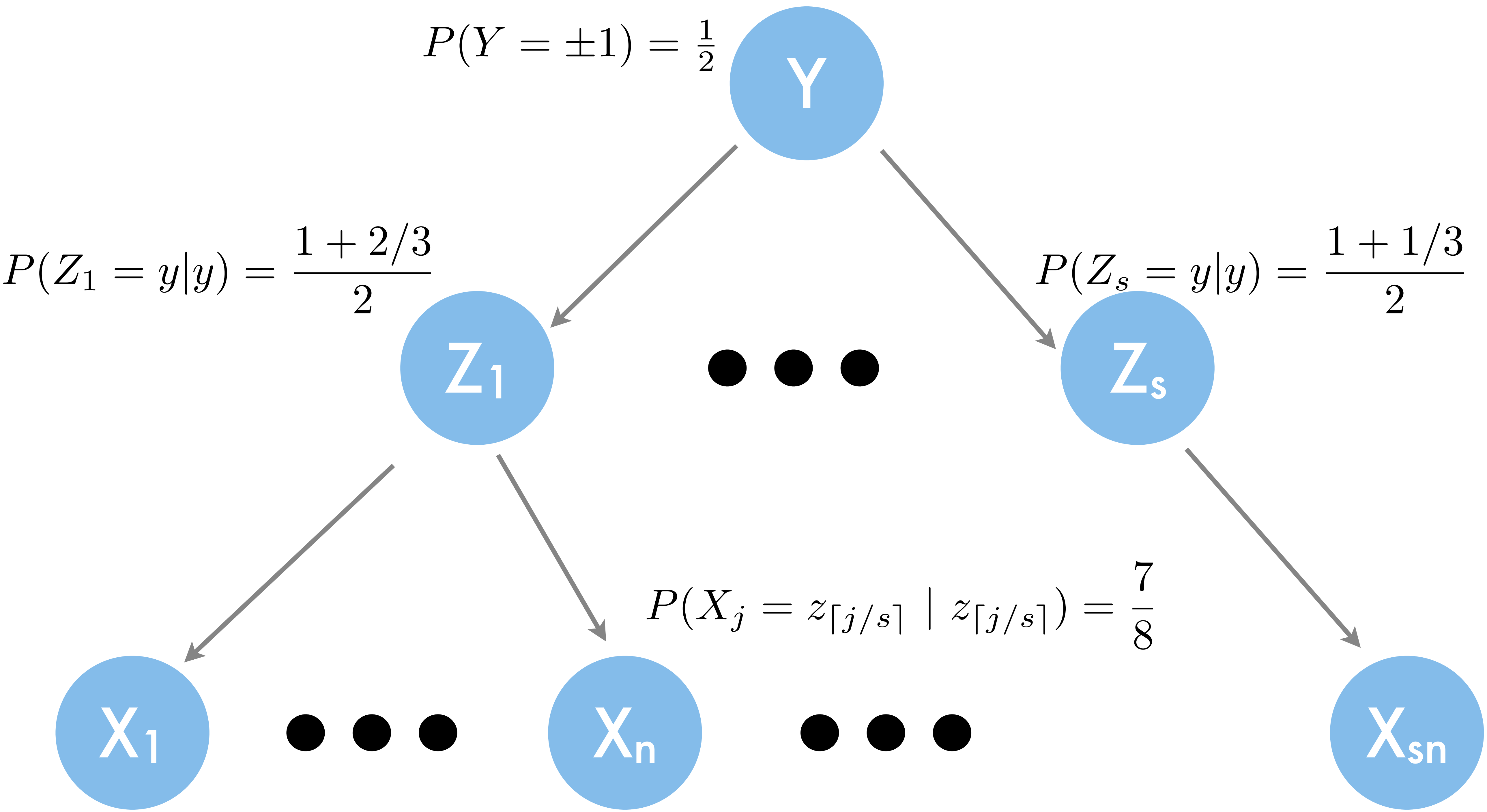
Low L2 norm does not guarantee sparsifiability

- Same data distribution as before with $B = \epsilon\sqrt{n}$
- Dense predictor:
  - $w_i = \frac{B}{n}$
  - $\mathbb{E}[|\langle \mathbf{w}, \mathbf{x}\rangle - y|] \leq \frac{B}{\sqrt{n}} = \epsilon$
  - $\|\mathbf{w}\|_2 = \frac{B}{\sqrt{n}} = \epsilon$
- Sparse predictor:
  - Any $\mathbf{u}$ with $\mathbb{E}[|\langle \mathbf{u}, \mathbf{x}\rangle - y|] \leq 2\,\epsilon$ must use almost all features:

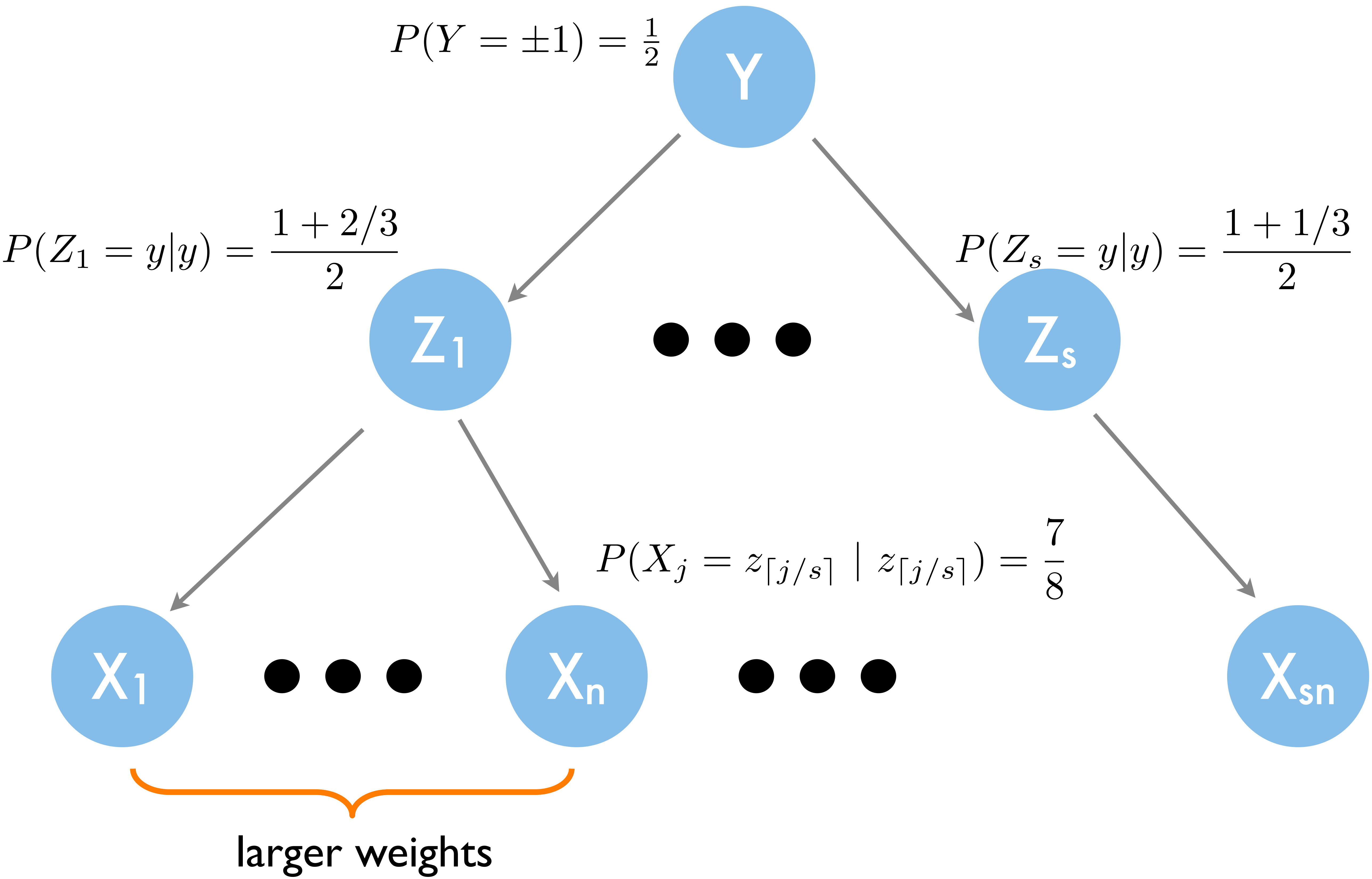
$$\|\mathbf{u}\|_0 = \Omega\left(\frac{B^2}{\epsilon^2}\right) = \Omega(n)$$

- **ℓ_1 captures sparsity but ℓ_2 doesn't !**

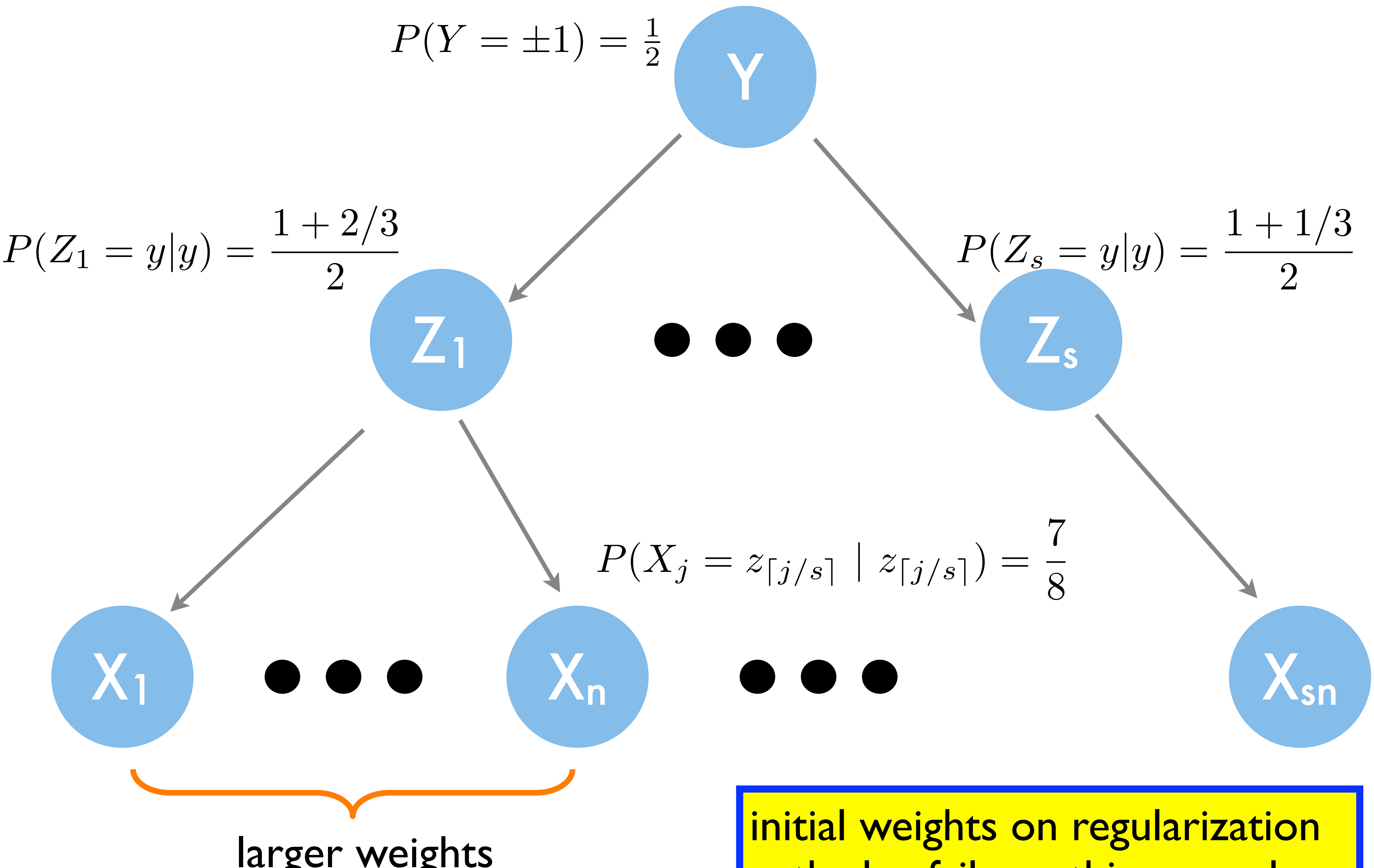
Sparsifying by zeroing small weights fails



Sparsifying by zeroing small weights fails



Sparsifying by zeroing small weights fails



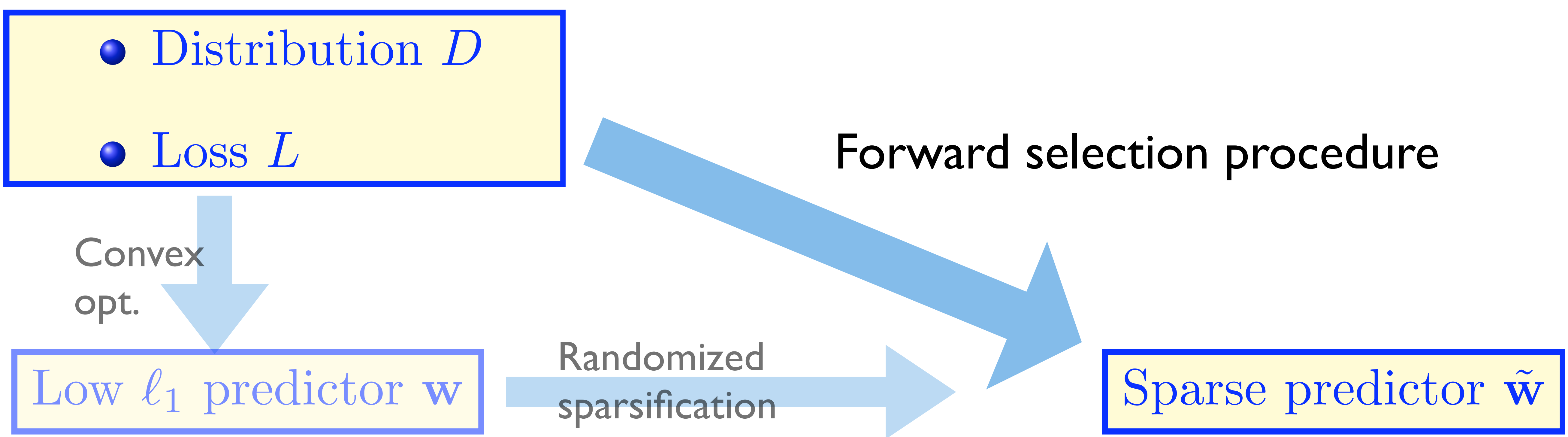
initial weights on regularization path also fails on this example

Intermediate Summary

- We answer a fundamental question:
  How much sparsity does low ℓ_1 norm guarantee ?
  - $\|\tilde{\mathbf{w}}\|_0 \leq O\left(\frac{\|\mathbf{w}\|_1^2}{\epsilon^2}\right)$
  - This is tight
  - Achievable by simple randomized procedure
- Coming next: Direct approach also works !

Intermediate Summary

- We answer a fundamental question:
  How much sparsity does low ℓ_1 norm guarantee ?
  - $\|\tilde{\mathbf{w}}\|_0 \leq O\left(\frac{\|\mathbf{w}\|_1^2}{\epsilon^2}\right)$
  - This is tight
  - Achievable by simple randomized procedure
- Coming next: Direct approach also works !



Greedy Forward Selection

- Step 1: Define a slightly modified loss function

$$\tilde{L}(v, y) = \min_u \frac{\lambda^2}{\epsilon}(u - v)^2 + L(u, y)$$

Using infimal convolution theory, it can be shown that

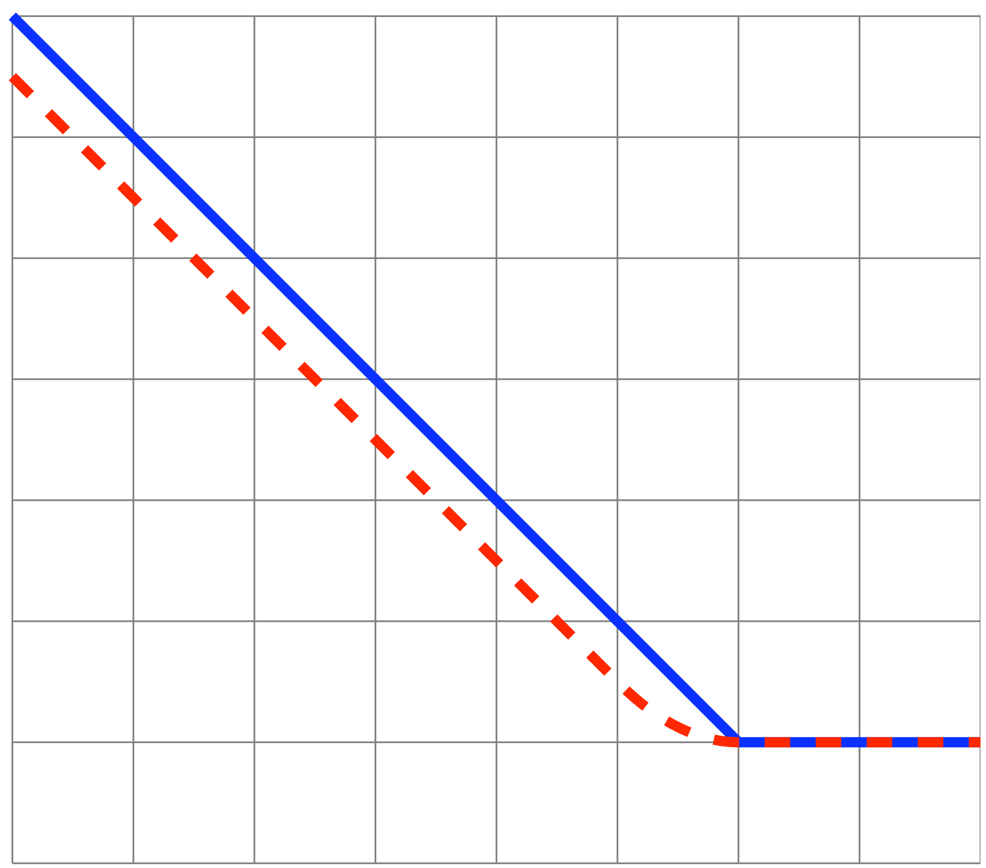
  - $\tilde{L}$ has Lipschitz continuous derivative
  - $\forall v, y \;\; |L(v, y) - \tilde{L}(v, y)| \leq \epsilon/4$

- Step 2: Apply forward greedy selection on $\tilde{L}$
  - Initialize $\mathbf{w}_1 = \mathbf{0}$
  - Choose feature using largest element of gradient
  - Choose step size η_t (closed form solution exists)
  - Update $\mathbf{w}_{t+1} = (1 - \eta_t)\mathbf{w}_t + \eta_t \, B \, \mathbf{e}^{j_t}$

Greedy Forward Selection

Example – Hinge loss:

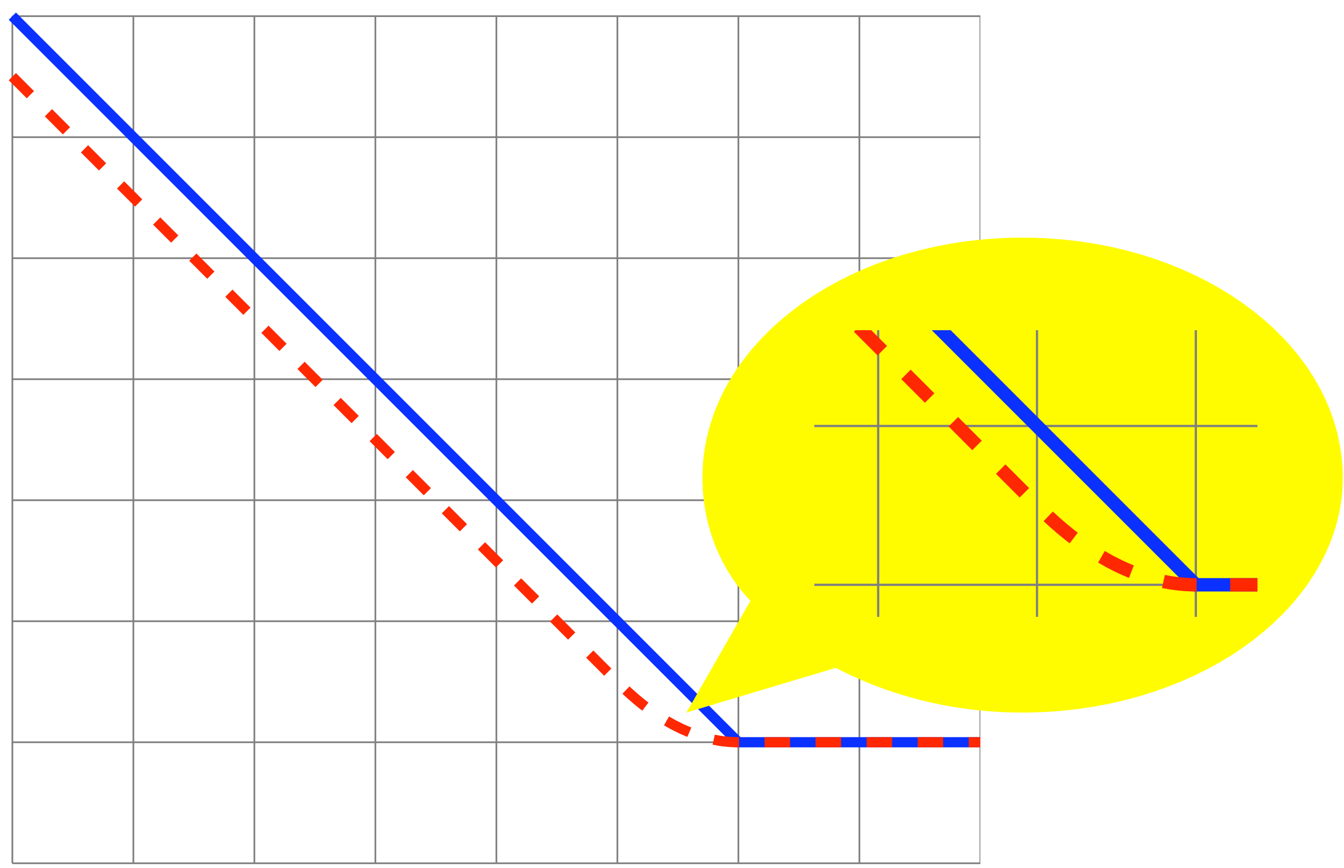
$$L(v,y) = \max\{0, 1-v\} \quad ; \quad \tilde{L}(v,y) = \begin{cases} 0 & \text{if } v > 1 \\ \frac{1}{\epsilon}(v-1)^2 & \text{if } v \in [1-\frac{1}{\epsilon}, 1] \\ (1-\frac{\epsilon}{4}) - v & \text{else} \end{cases}$$



Greedy Forward Selection

Example – Hinge loss:

$$L(v,y) = \max\{0, 1-v\} \quad ; \quad \tilde{L}(v,y) = \begin{cases} 0 & \text{if } v > 1 \\ \frac{1}{\epsilon}(v-1)^2 & \text{if } v \in [1-\frac{1}{\epsilon}, 1] \\ (1-\frac{\epsilon}{4}) - v & \text{else} \end{cases}$$



Guarantees

Theorem

- $\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\|_\infty \leq 1\}$, Y = arbitrary set
- $\mathcal{D}$ = arbitrary distribution over $\mathcal{X} \times Y$
- Loss $L : \mathbb{R} \times Y \to \mathbb{R}$ is proper, convex, and λ-Lipschitz w.r.t. 1st argument
- Forward greedy selection on $\tilde{L}$ finds $\tilde{\mathbf{w}}$ s.t.
  - $\|\tilde{\mathbf{w}}\|_0 = O\left(\frac{\lambda^2 B^2}{\epsilon^2}\right)$
  - For any $\mathbf{w}$ with $\|\mathbf{w}\|_1 \leq B$ we have:

$$\mathbb{E}[L(\langle \tilde{\mathbf{w}}, \mathbf{x} \rangle, y)] - \mathbb{E}[L(\langle \mathbf{w}, \mathbf{x} \rangle, y)] \leq \epsilon$$

Related Work

- ℓ_1 norm and sparsity:
  - Donoho provides sufficient conditions for when minimizer of ℓ_1 norm is also sparse. But, what if these conditions are not met?
  - Compressed sensing: ℓ_1 norm recovers sparse predictor, but only under server assumptions on the design matrix (in our case, the training examples)
- Converse question: Small $\|\tilde{\mathbf{w}}\|_0 \stackrel{?}{\Rightarrow}$ Small $\|\mathbf{w}\|_1$?
  - Servedio: partial answer for the case of linear classification
  - Wainwright: partial answer for the Lasso
- Sparsification:
  - Randomized sparsification procedure previously proposed by Schapire et al. However, their bound depends on training set size
  - Lee, Bartlett, and Williamson addressed similar question for the special case of squared-error loss
  - Zhang presented forward greedy procedure for twice differentiable losses

Summary

- Distribution D
- Loss L

Convex opt.

Low ℓ_1 predictor $\mathbf{w}$

Randomized

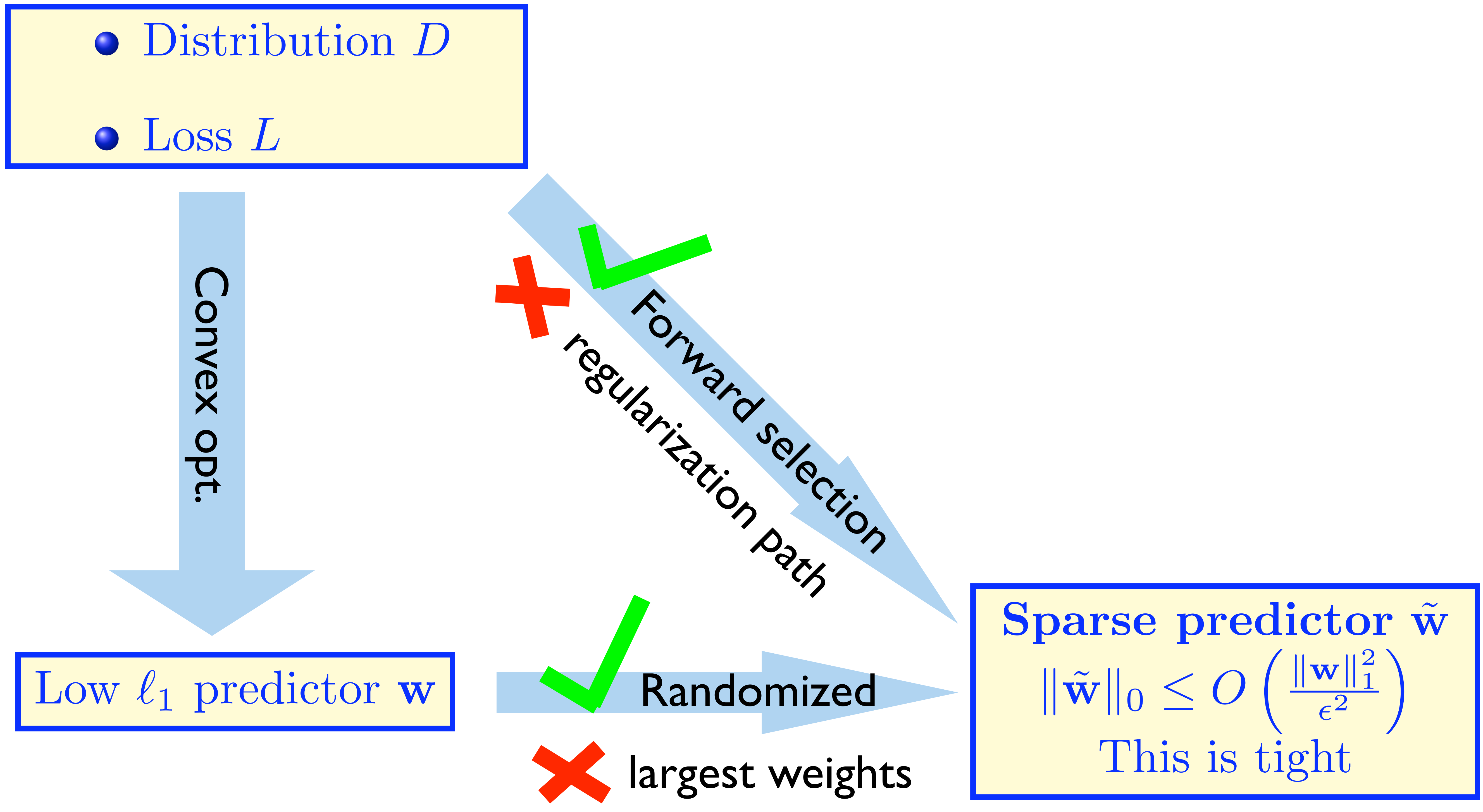
largest weights

Sparse predictor $\tilde{\mathbf{w}}$

$$\|\tilde{\mathbf{w}}\|_0 \leq O\left(\frac{\|\mathbf{w}\|_1^2}{\epsilon^2}\right)$$

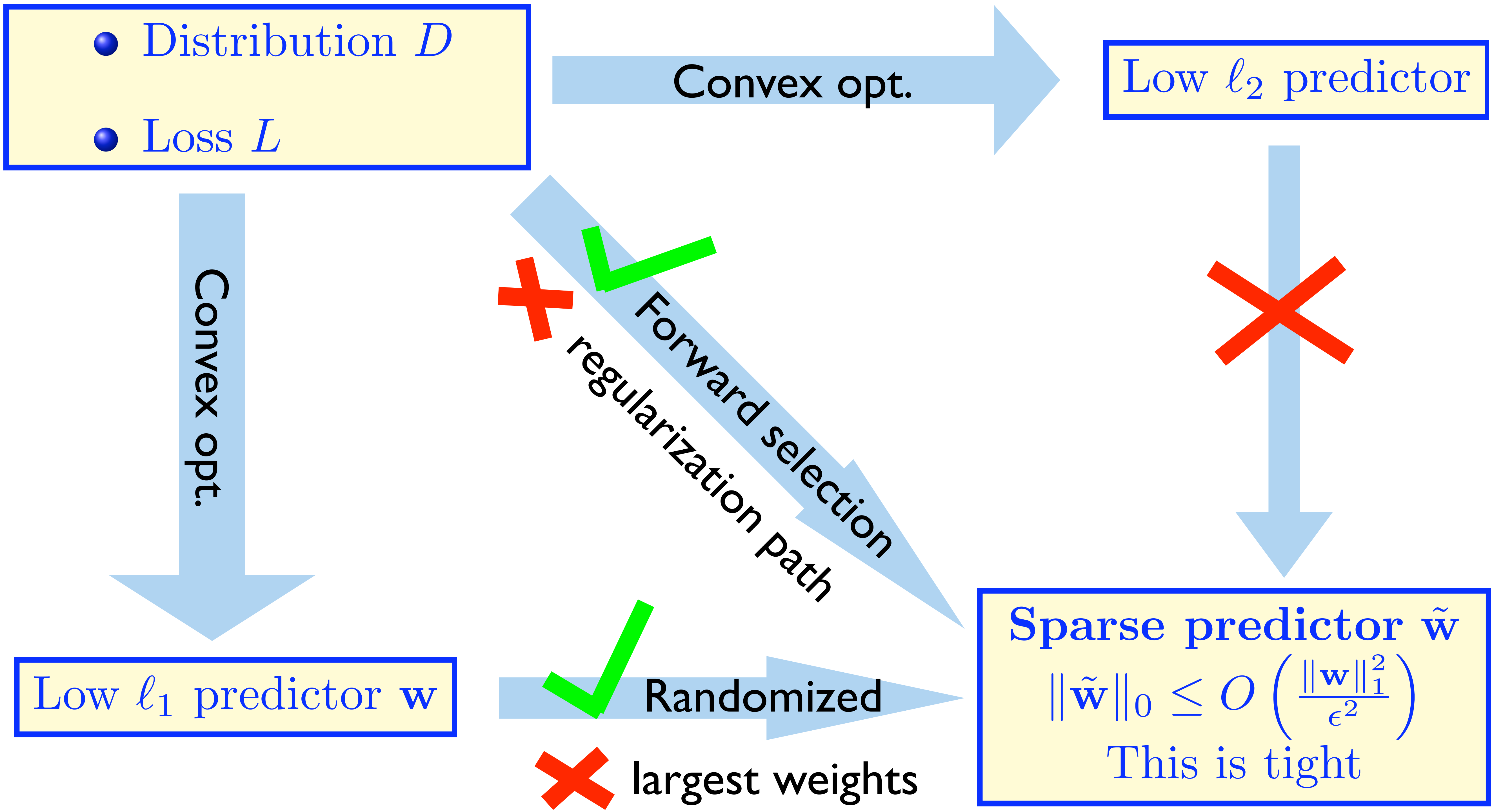
This is tight

Summary



Distribution D
Loss L
Convex opt.
Low ℓ_1 predictor $\mathbf{w}$
Forward selection
regularization path
Randomized
largest weights
Sparse predictor $\tilde{\mathbf{w}}$
$\|\tilde{\mathbf{w}}\|_0 \le O\left(\frac{\|\mathbf{w}\|_1^2}{\epsilon^2}\right)$
This is tight

Summary



Distribution D
Loss L
Convex opt.
Low ℓ_2 predictor
Convex opt.
Forward selection
regularization path
Low ℓ_1 predictor $\mathbf{w}$
Randomized
largest weights
Sparse predictor $\tilde{\mathbf{w}}$
$\|\tilde{\mathbf{w}}\|_0 \leq O\left(\frac{\|\mathbf{w}\|_1^2}{\epsilon^2}\right)$
This is tight