Sparse regression

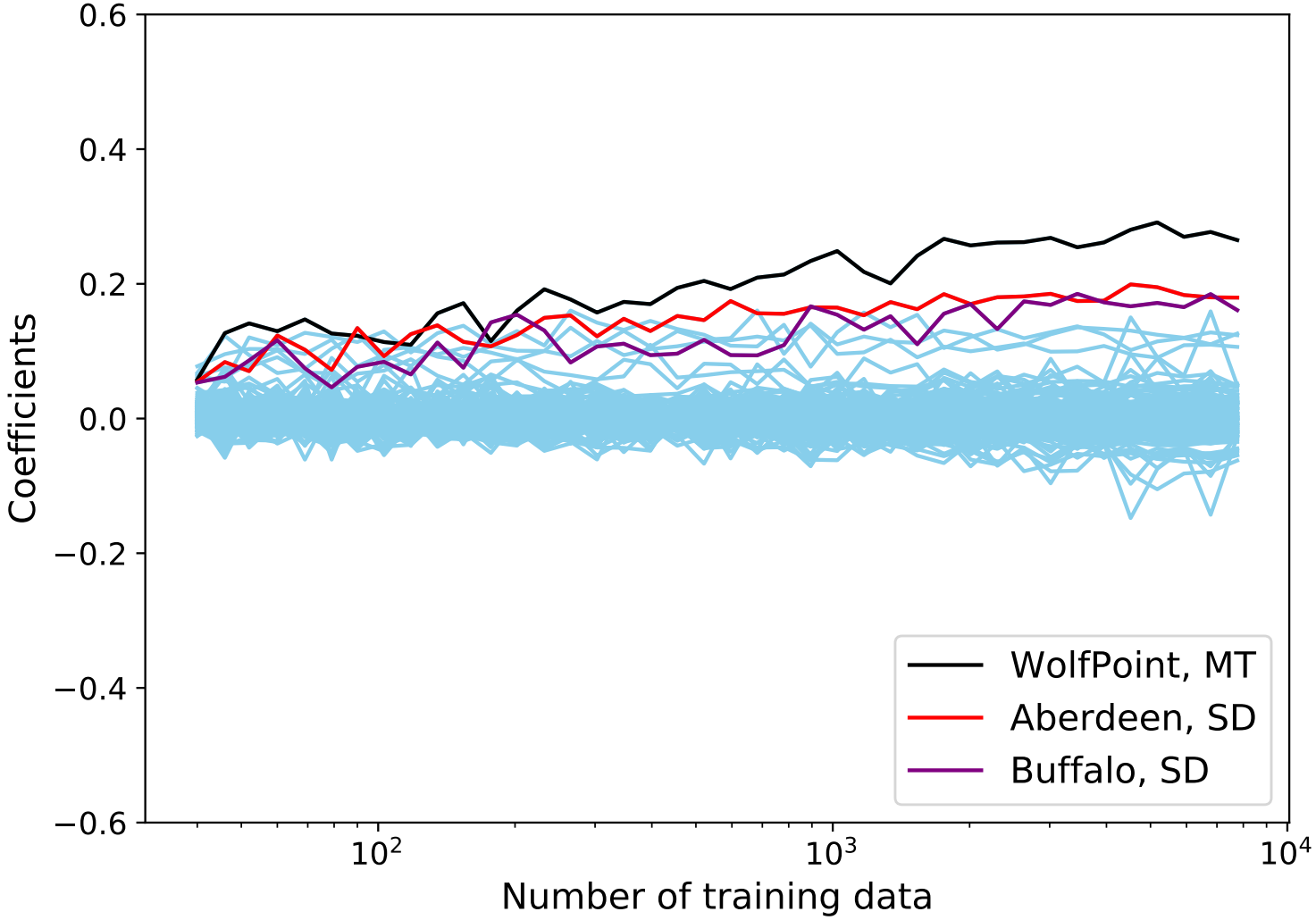
DS-GA 1013 / MATH-GA 2824 Mathematical Tools for Data Science

Carlos Fernandez-Granda

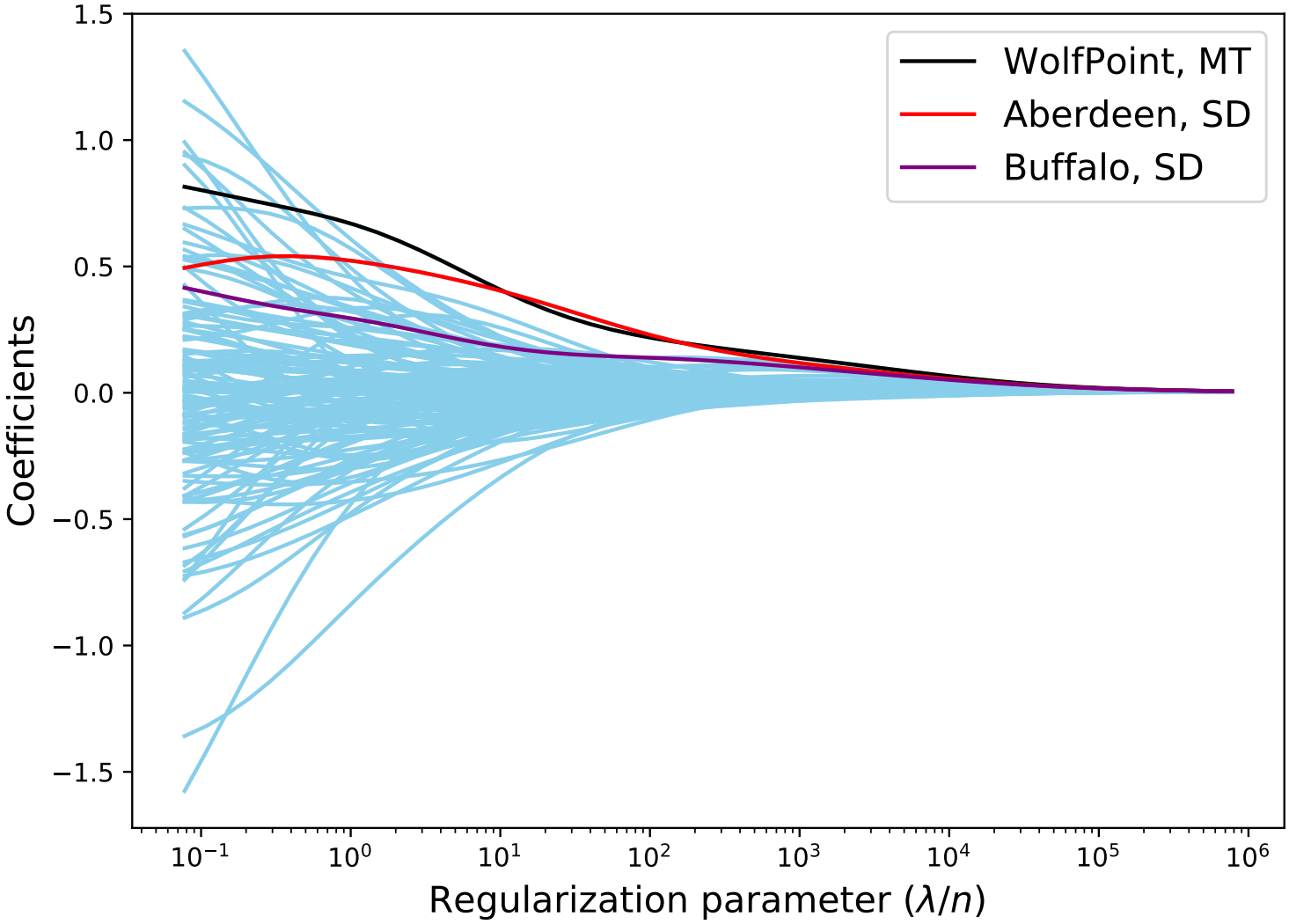
Temperature prediction via linear regression

- Dataset of hourly temperatures measured at weather stations all over the US
- Goal: Predict temperature in Jamestown (North Dakota) from other temperatures
- Response: Temperature in Jamestown
- Features: Temperatures in 133 other stations ($p = 133$) in 2015
- Test set: 10^3 measurements
- Additional test set: All measurements from 2016

Ridge-regression coefficients



Ridge regression $n := 135$

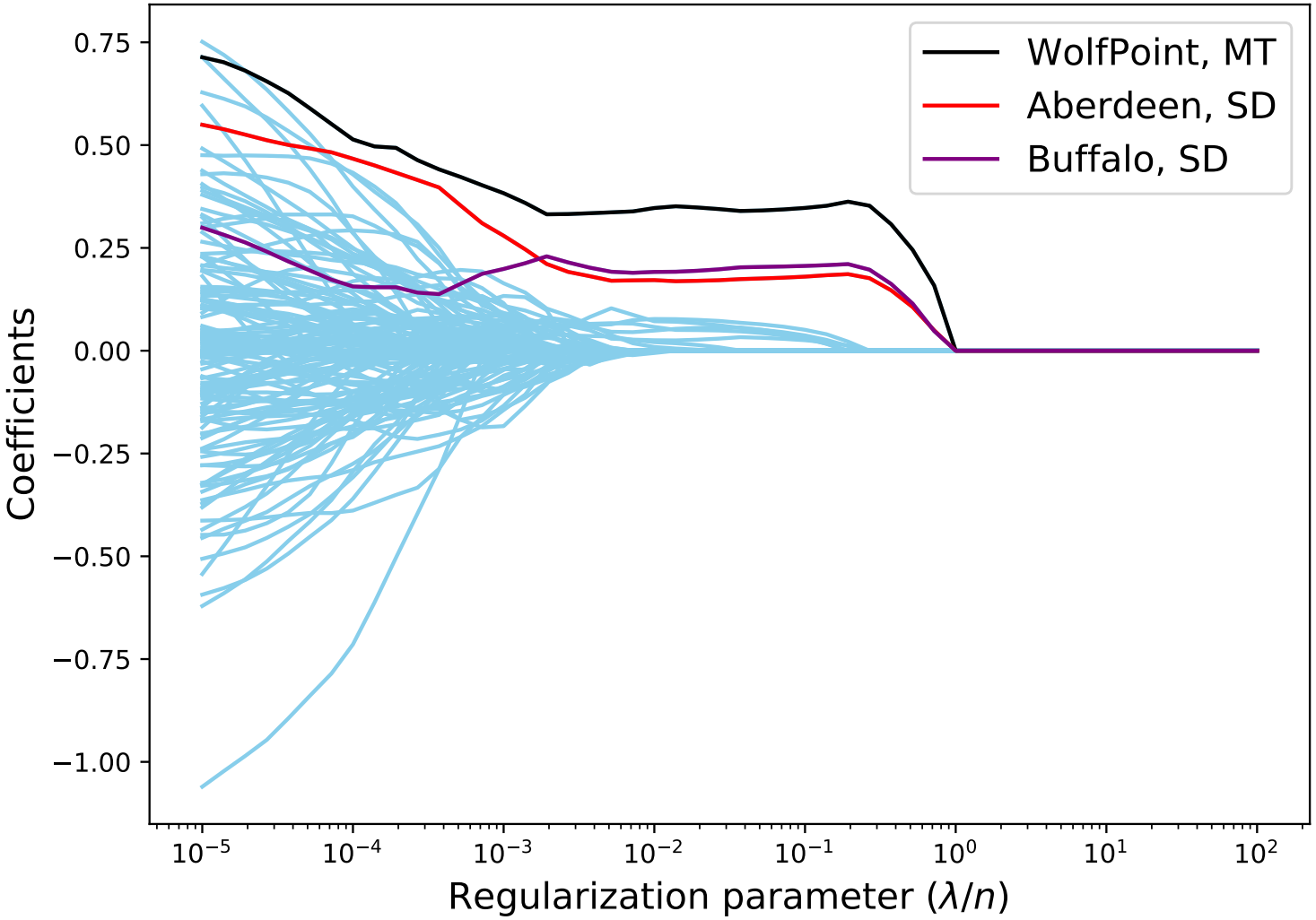


The lasso

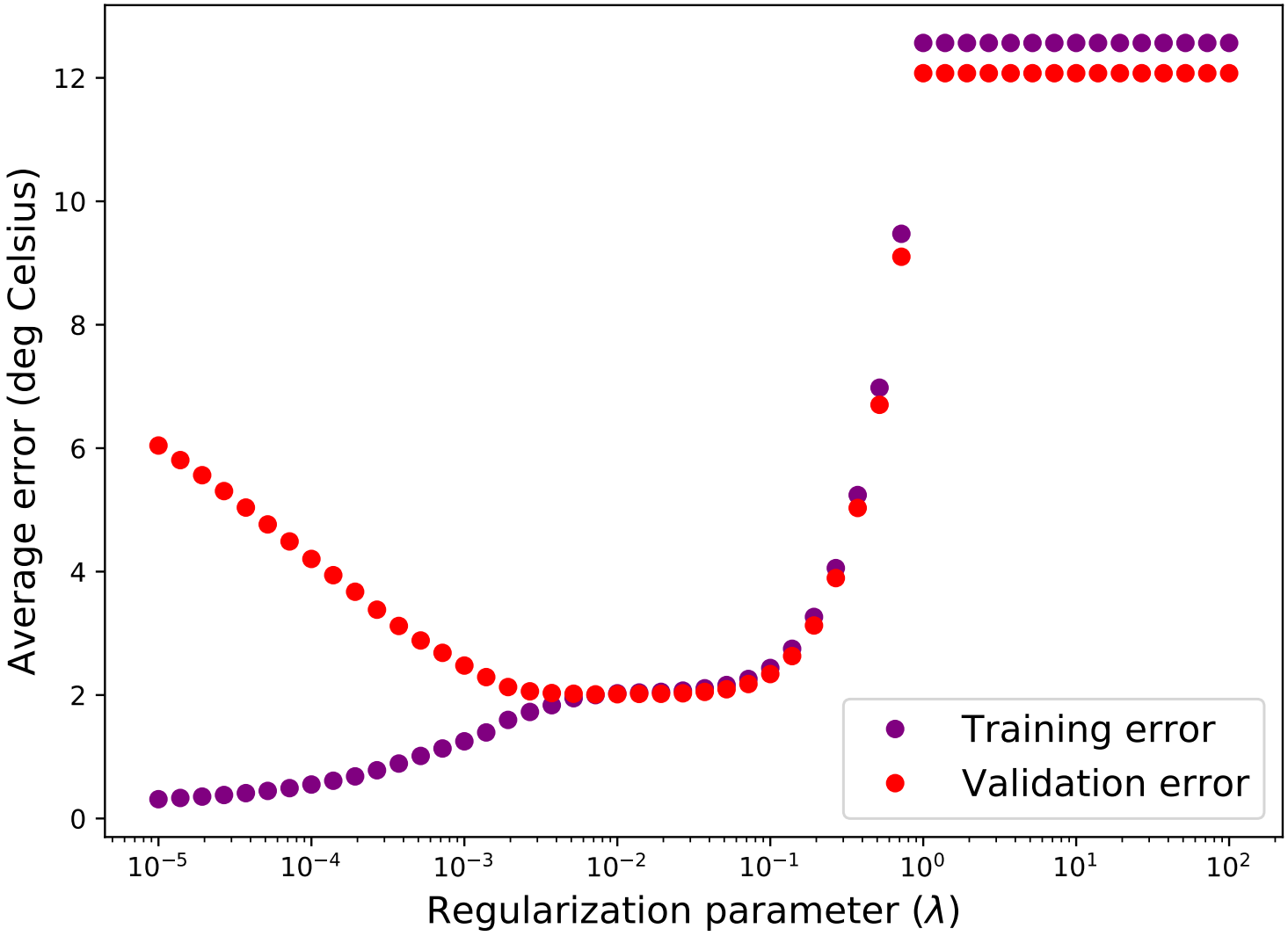
Uses ℓ_1-norm regularization to promote sparse coefficients

$$\beta_{\text{lasso}} := \arg\min_{\beta} \frac{1}{2} \left|\left| y - X^T \beta \right|\right|_2^2 + \lambda \left|\left|\beta\right|\right|_1$$

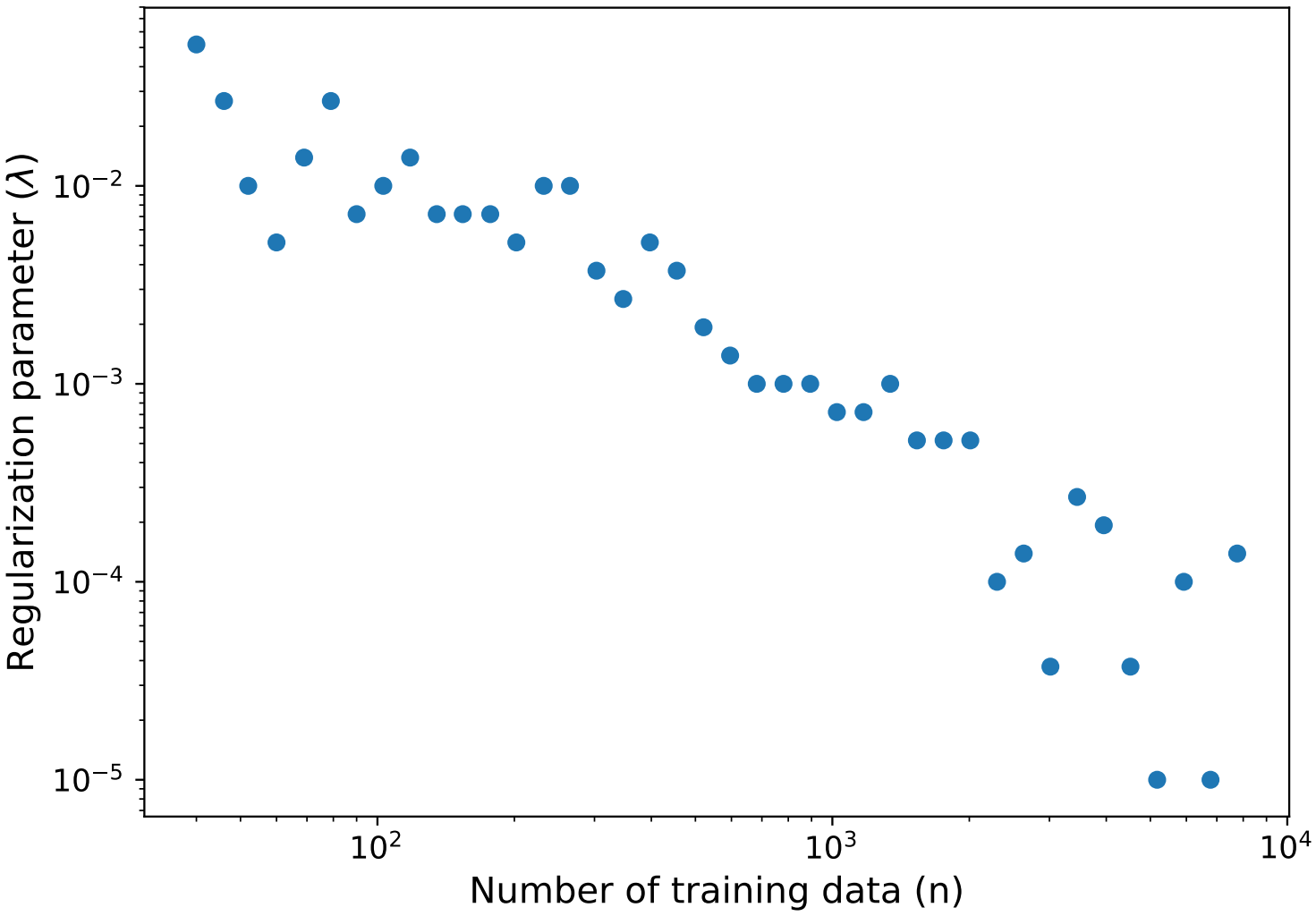
Lasso $n := 135$


WolfPoint, MT
Aberdeen, SD
Buffalo, SD
Coefficients
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
Regularization parameter (λ/n)

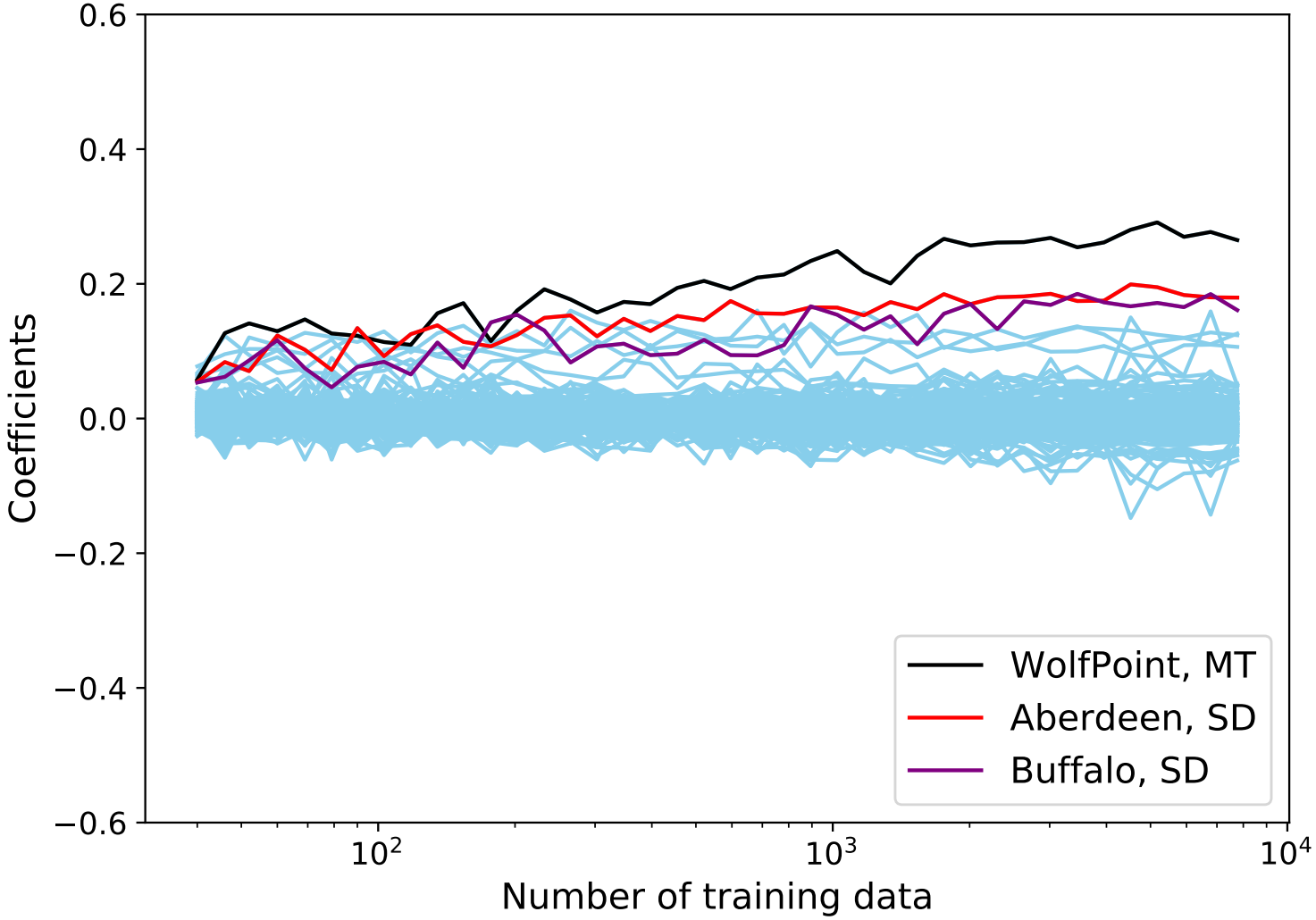
Lasso $n := 135$


Average error (deg Celsius)
0
2
4
6
8
10
12
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
Regularization parameter (λ)
Training error
Validation error

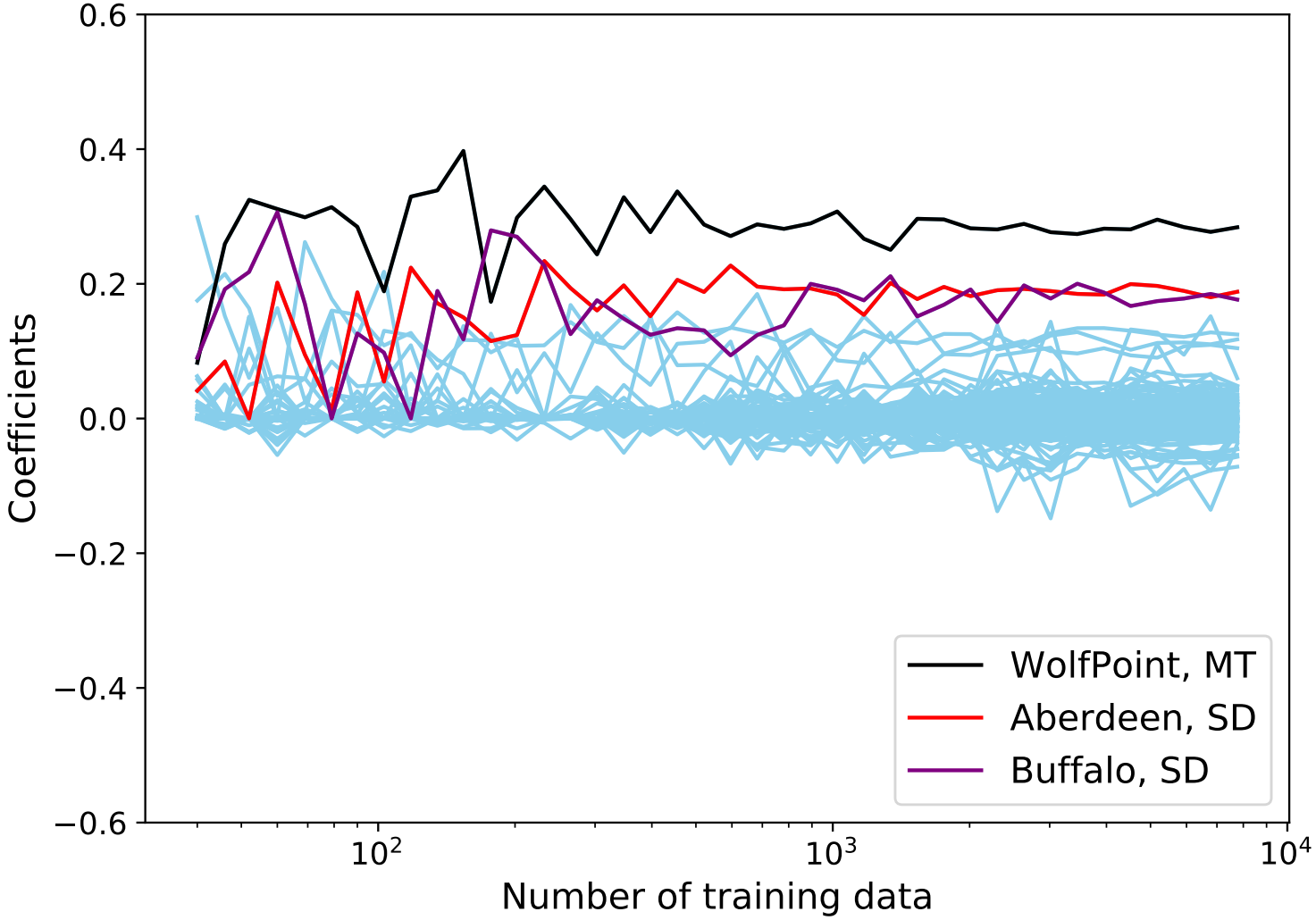
Lasso $n := 135$


Regularization parameter (λ)
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{2}
10^{3}
10^{4}
Number of training data (n)

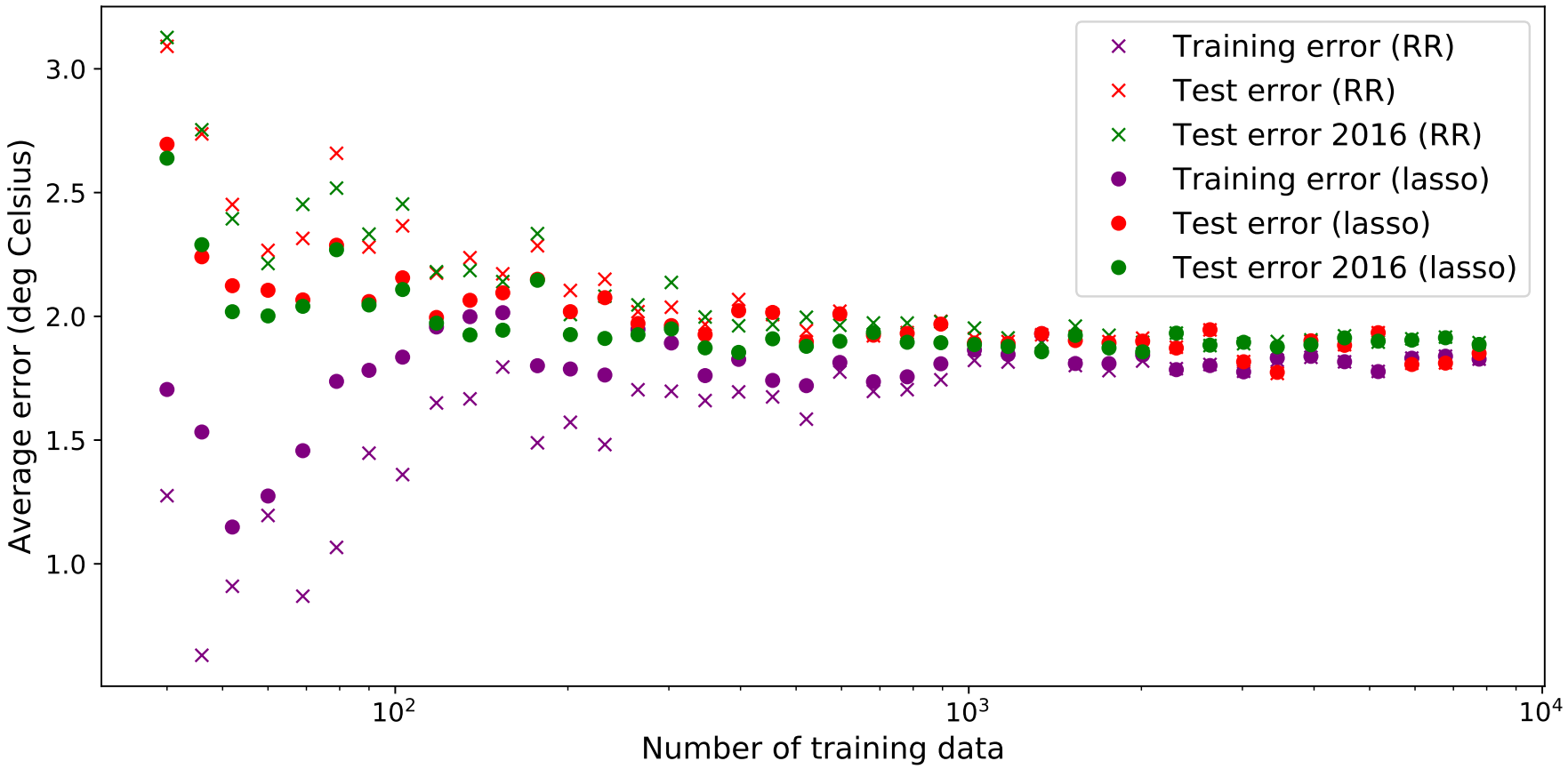
Ridge-regression coefficients



Lasso coefficients



Results


Training error (RR)
Test error (RR)
Test error 2016 (RR)
Training error (lasso)
Test error (lasso)
Test error 2016 (lasso)
Average error (deg Celsius)
3.0
2.5
2.0
1.5
1.0
10^2
10^3
10^4
Number of training data

Sparse regression with two features

Feature vectors and noise are fixed n-dimensional vectors

$$y := a + z$$

We fit a model using an additional feature

$$X := \begin{bmatrix} a & b \end{bmatrix}^T$$

$$\beta_{\text{true}} := \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$||a||_2 = ||b||_2 = 1$

Ridge regression

$$\min_{\beta} \left|\left| X^T \beta - y \right|\right|_2^2 + \lambda \, ||\beta||_2^2$$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Solution

$$\beta_{\text{RR}} =$$

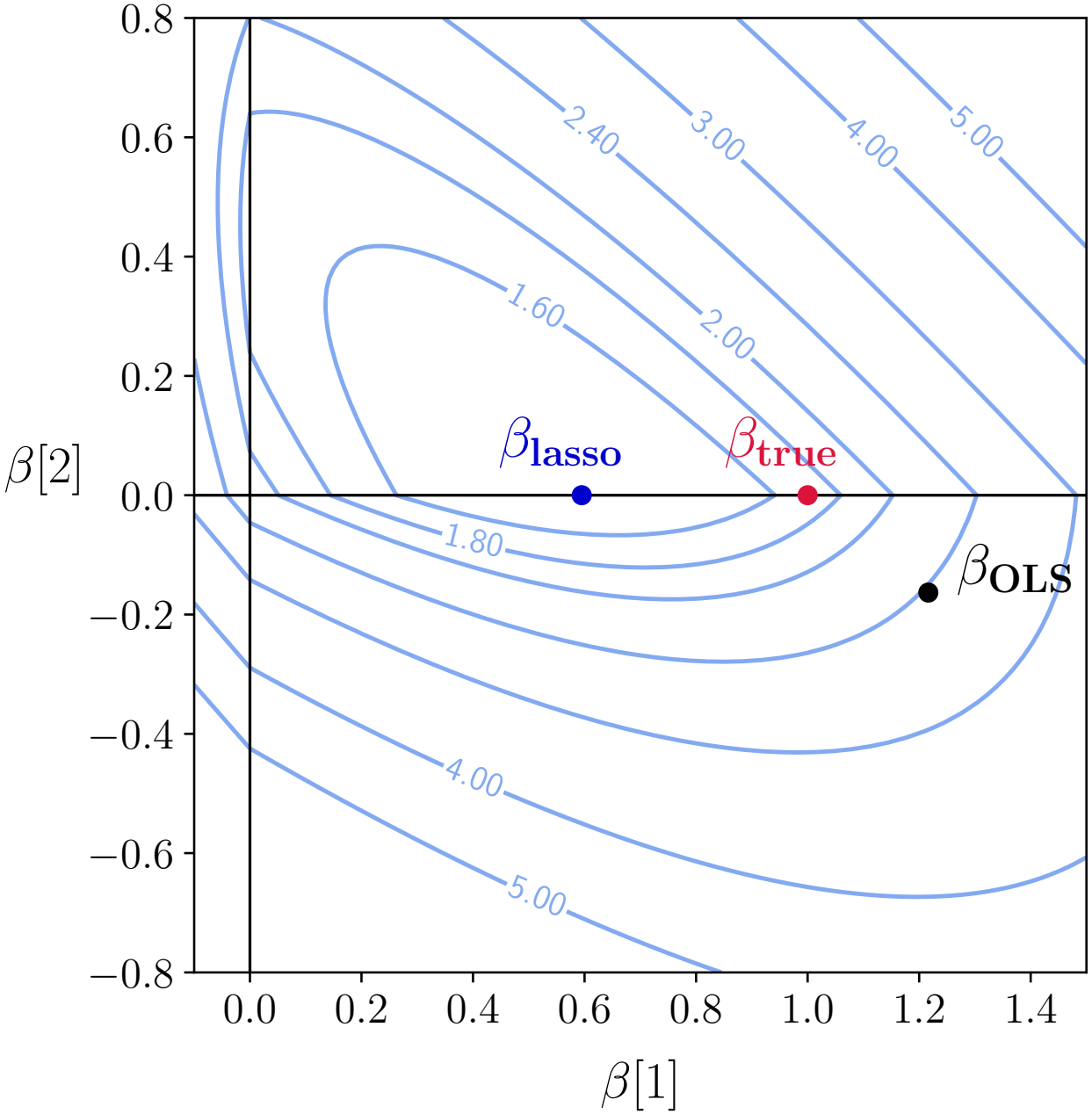
Ridge regression

$$\min_{\beta} \left|\left| X^T\beta - y \right|\right|_2^2 + \lambda \left|\left|\beta\right|\right|_2^2 \qquad \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

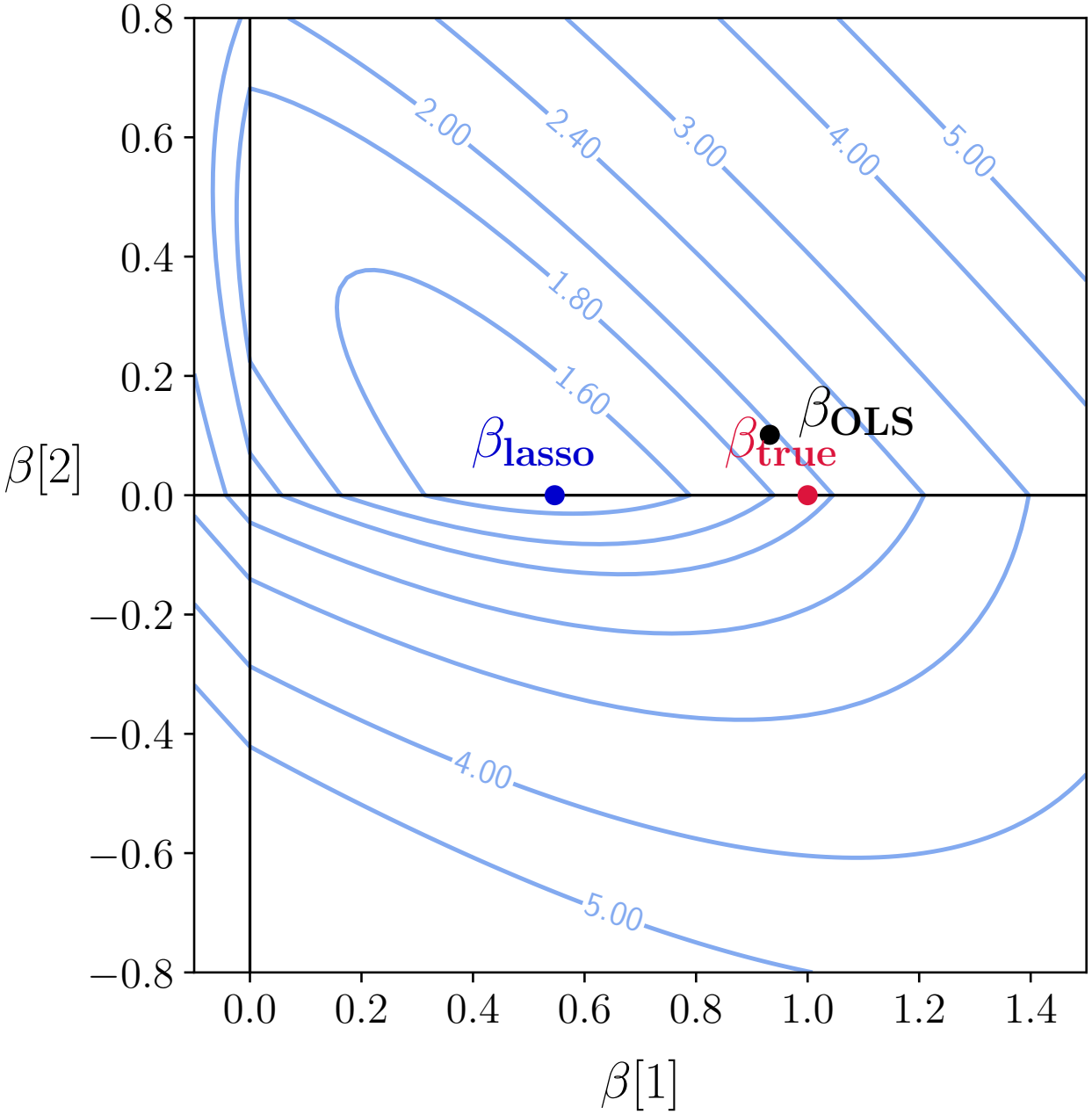
Solution

$$\begin{aligned}
\beta_{\text{RR}} &= (XX^T + \lambda I)^{-1} Xy \\
&= \begin{bmatrix} 1+\lambda & \rho \\ \rho & 1+\lambda \end{bmatrix}^{-1} \begin{bmatrix} 1 + a^T z \\ \rho + b^T z \end{bmatrix} \\
&= \frac{1}{(1+\lambda)^2 - \rho^2} \begin{bmatrix} 1+\lambda & -\rho \\ -\rho & 1+\lambda \end{bmatrix} \begin{bmatrix} 1 + a^T z \\ \rho + b^T z \end{bmatrix} \\
&= \frac{1}{(1+\lambda)^2 - \rho^2} \begin{bmatrix} (1+\lambda)(1 + a^T z) - \rho(\rho + b^T z) \\ b^T z(1+\lambda) + \rho(\lambda - a^T z) \end{bmatrix}
\end{aligned}$$

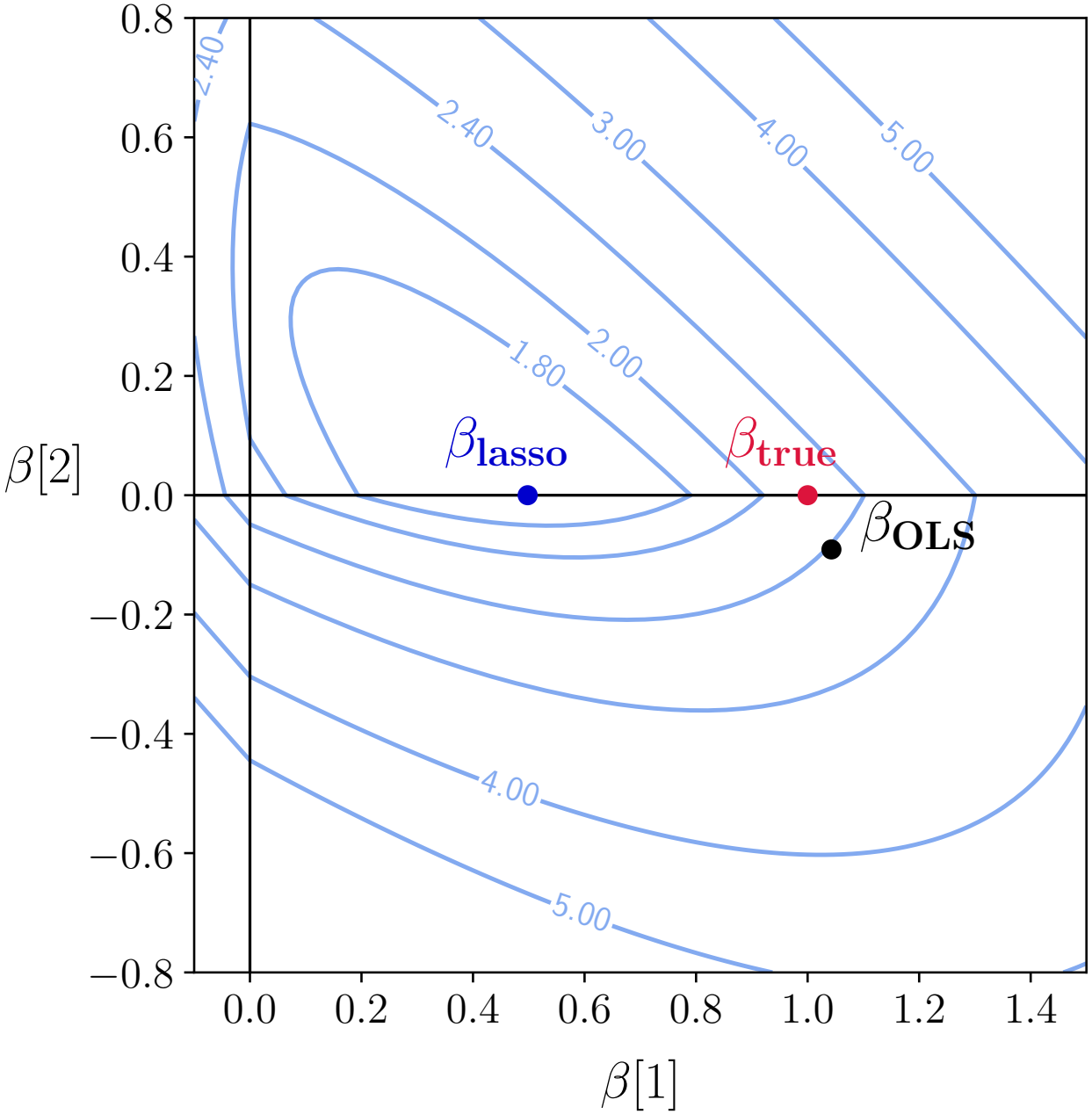
$$\frac{1}{2}\left|\left|X^T\beta - y\right|\right|_2^2 + \lambda\left|\left|\beta\right|\right|_1$$



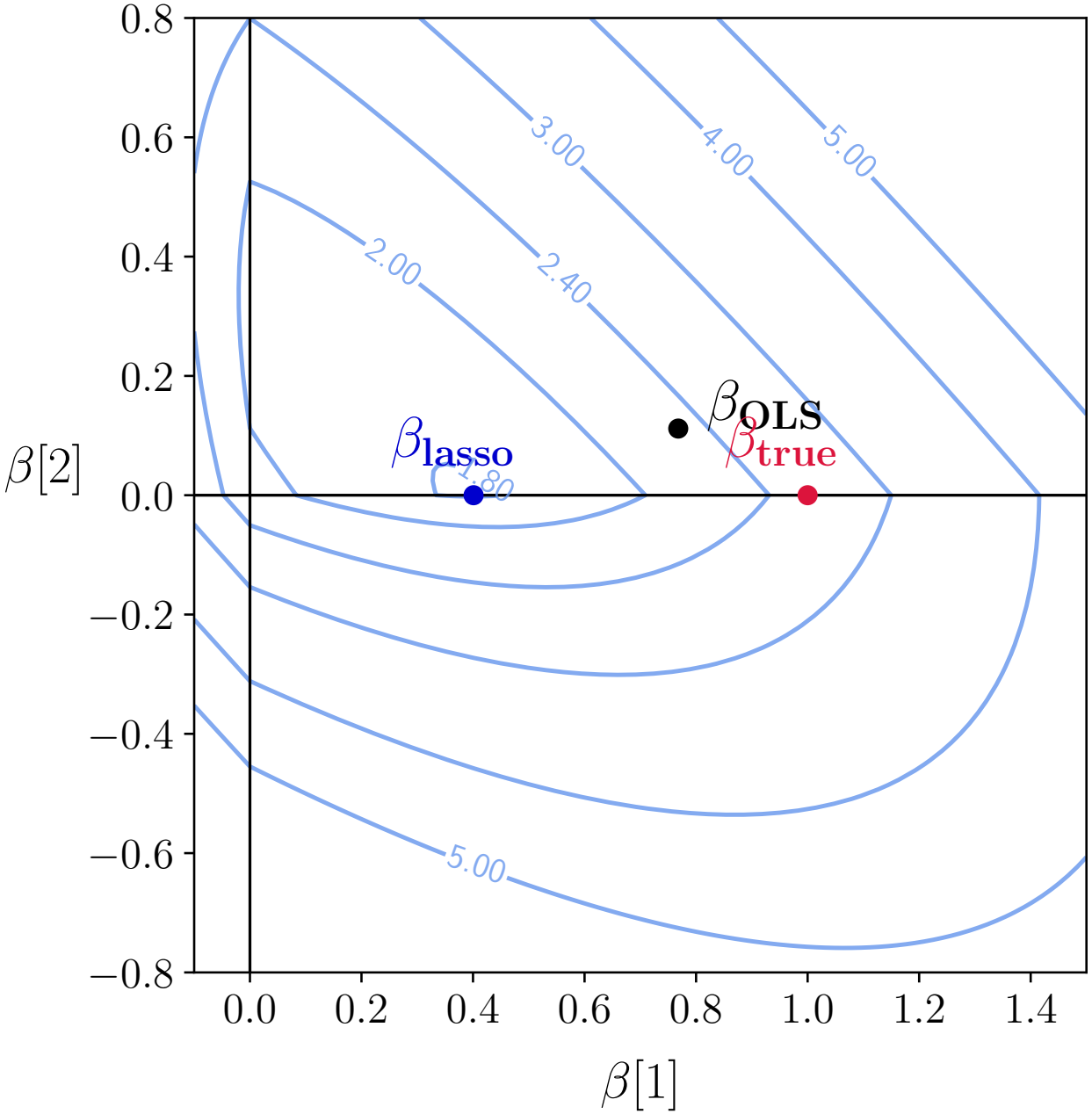
$$\frac{1}{2}\left|\left|X^T\beta - y\right|\right|_2^2 + \lambda\left|\left|\beta\right|\right|_1$$



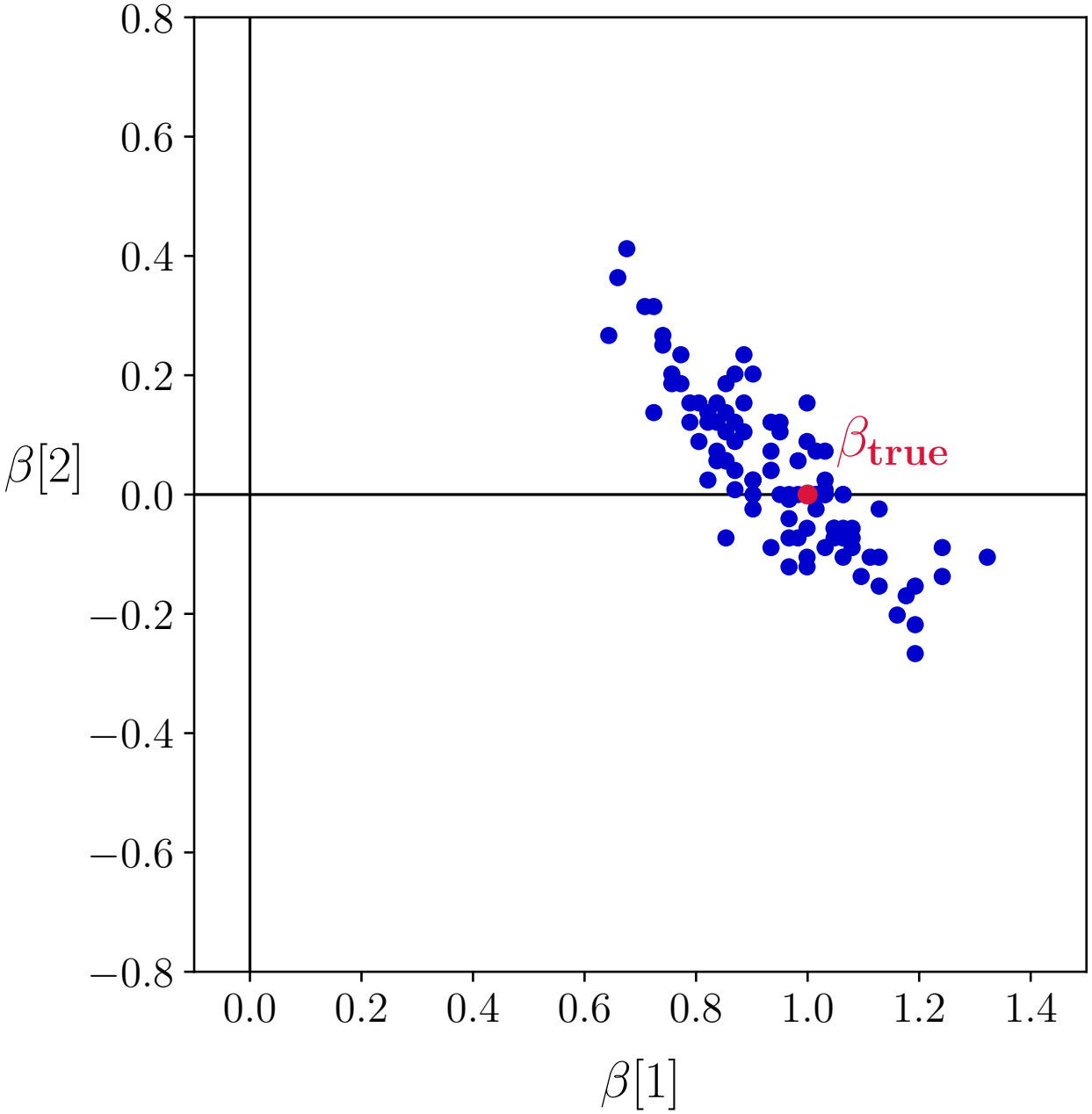
$$\frac{1}{2}\left|\left|X^T\beta - y\right|\right|_2^2 + \lambda\left|\left|\beta\right|\right|_1$$



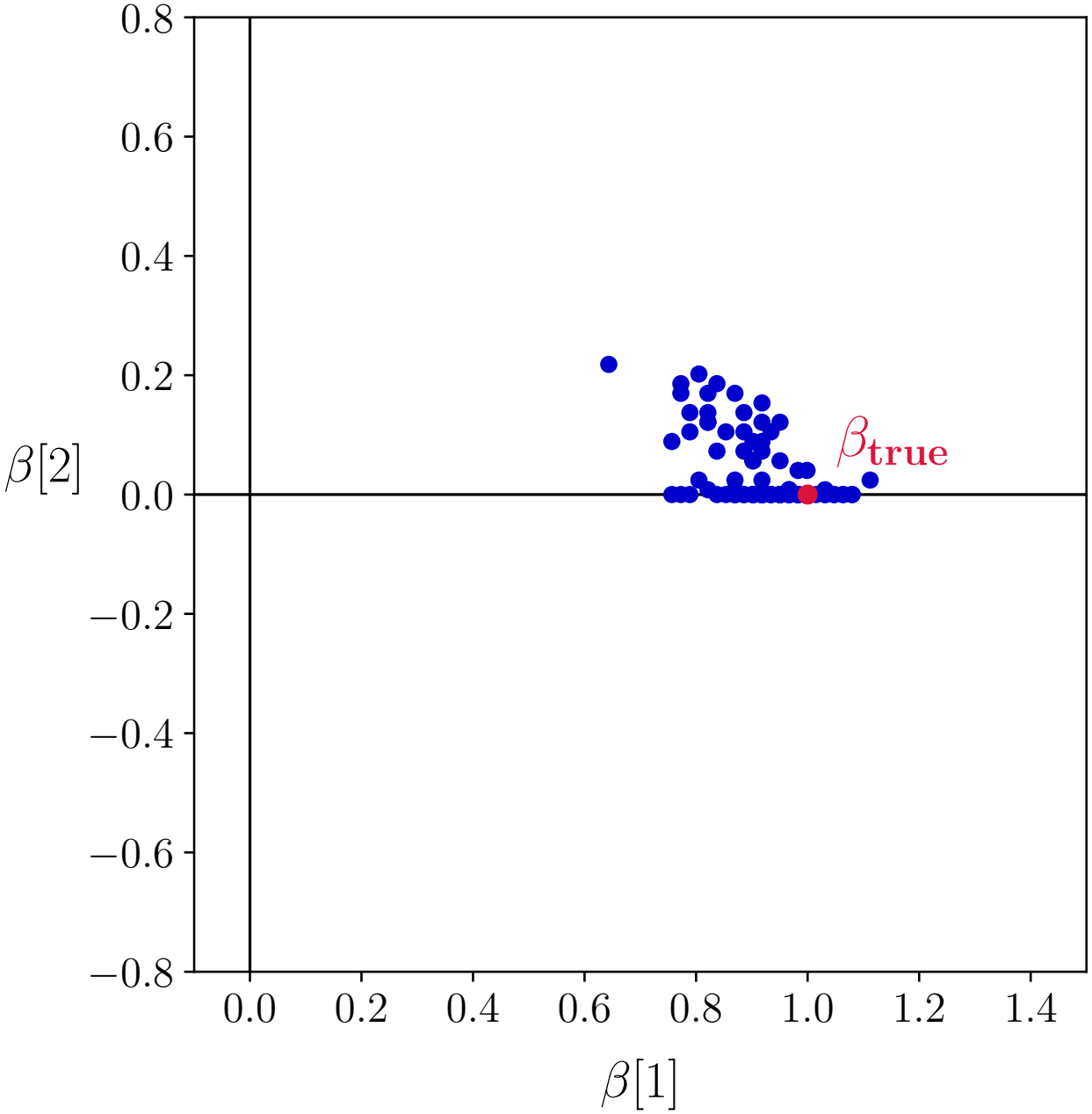
$$\frac{1}{2}\left|\left|X^T\beta - y\right|\right|_2^2 + \lambda\left|\left|\beta\right|\right|_1$$



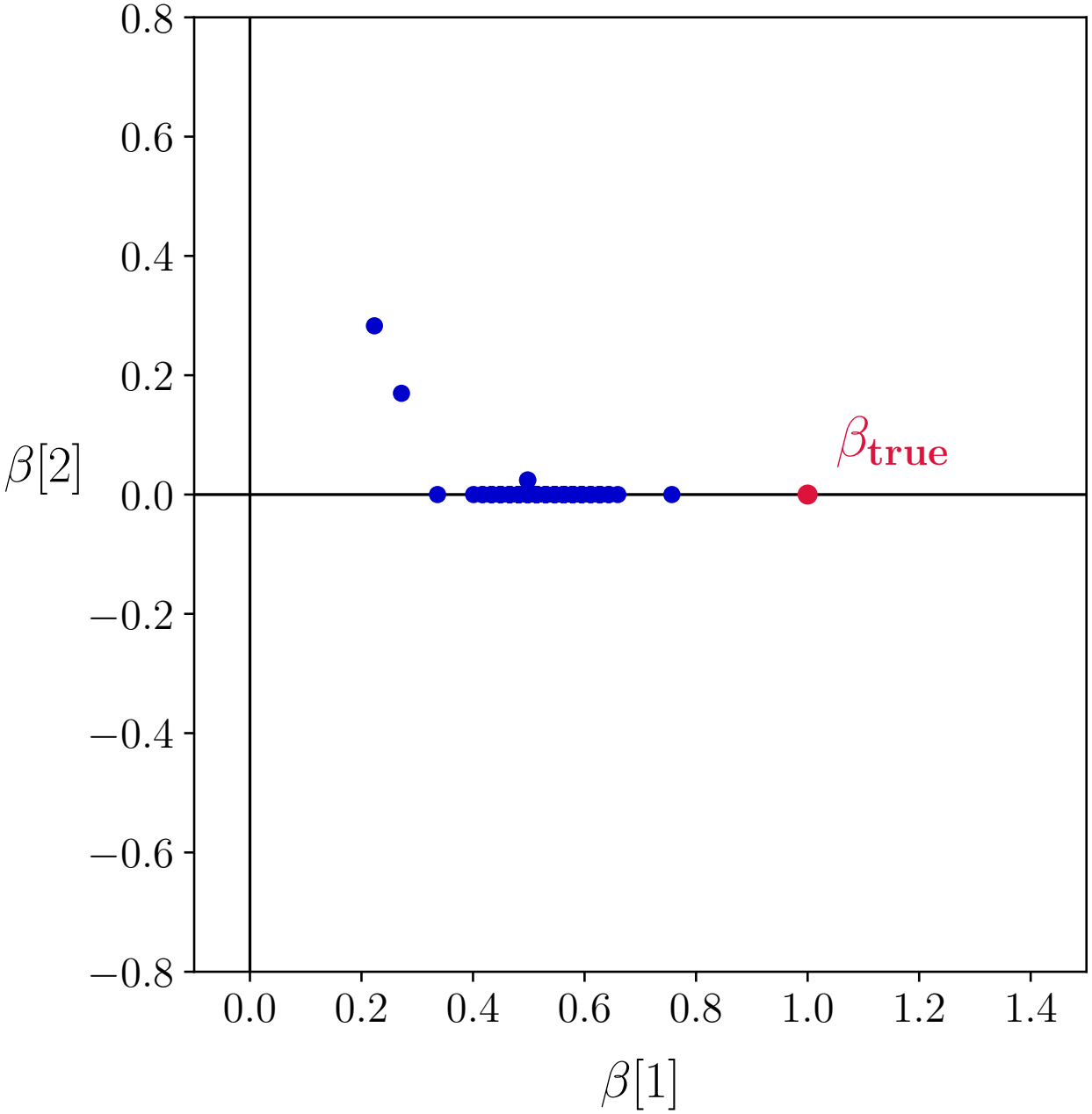
$\lambda = 0.02$


0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
$\beta[2]$
$\beta_{\mathbf{true}}$
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
$\beta[1]$

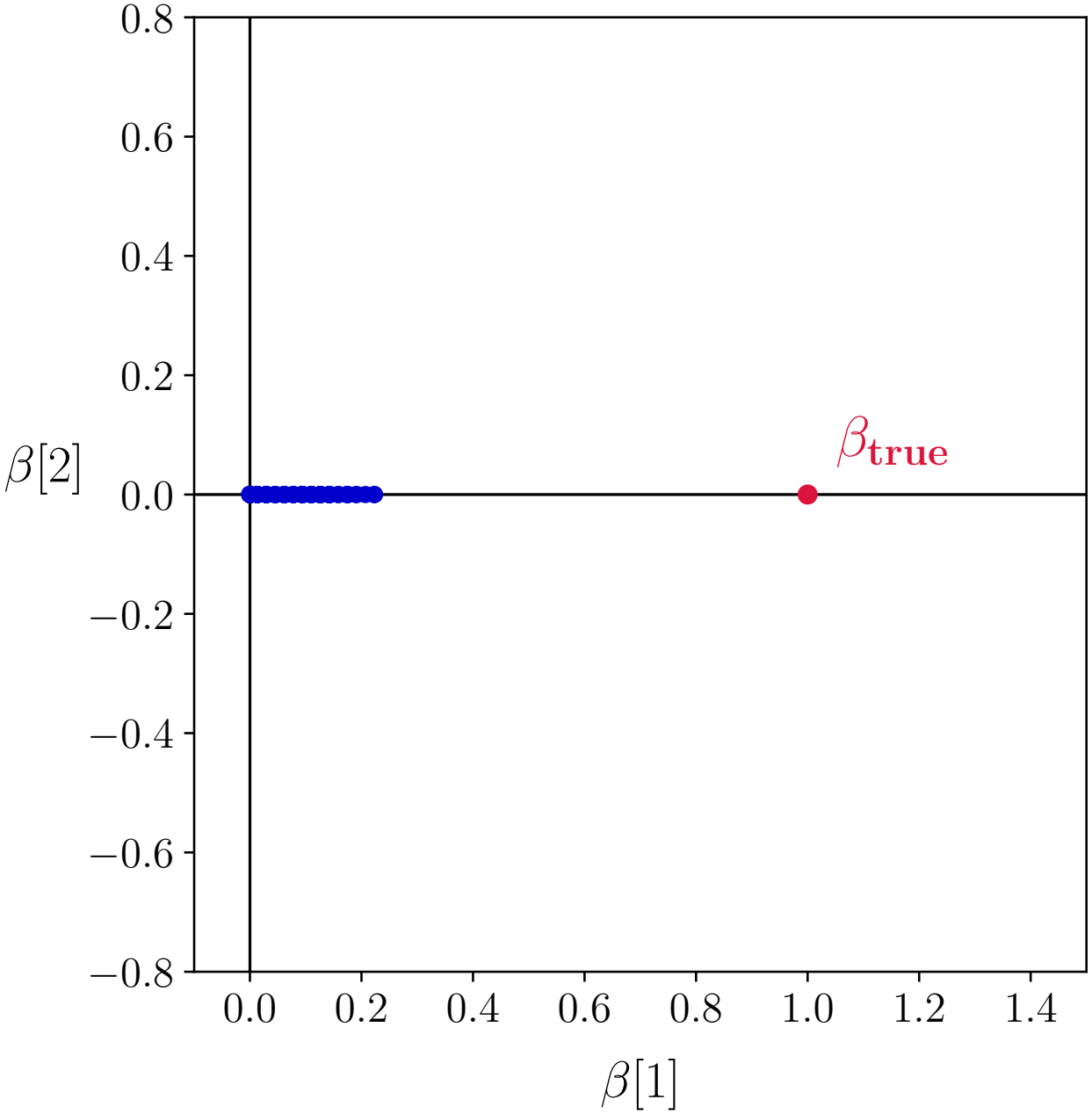
$\lambda = 0.2$


0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
$\beta[2]$
$\beta_{\mathbf{true}}$
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
$\beta[1]$

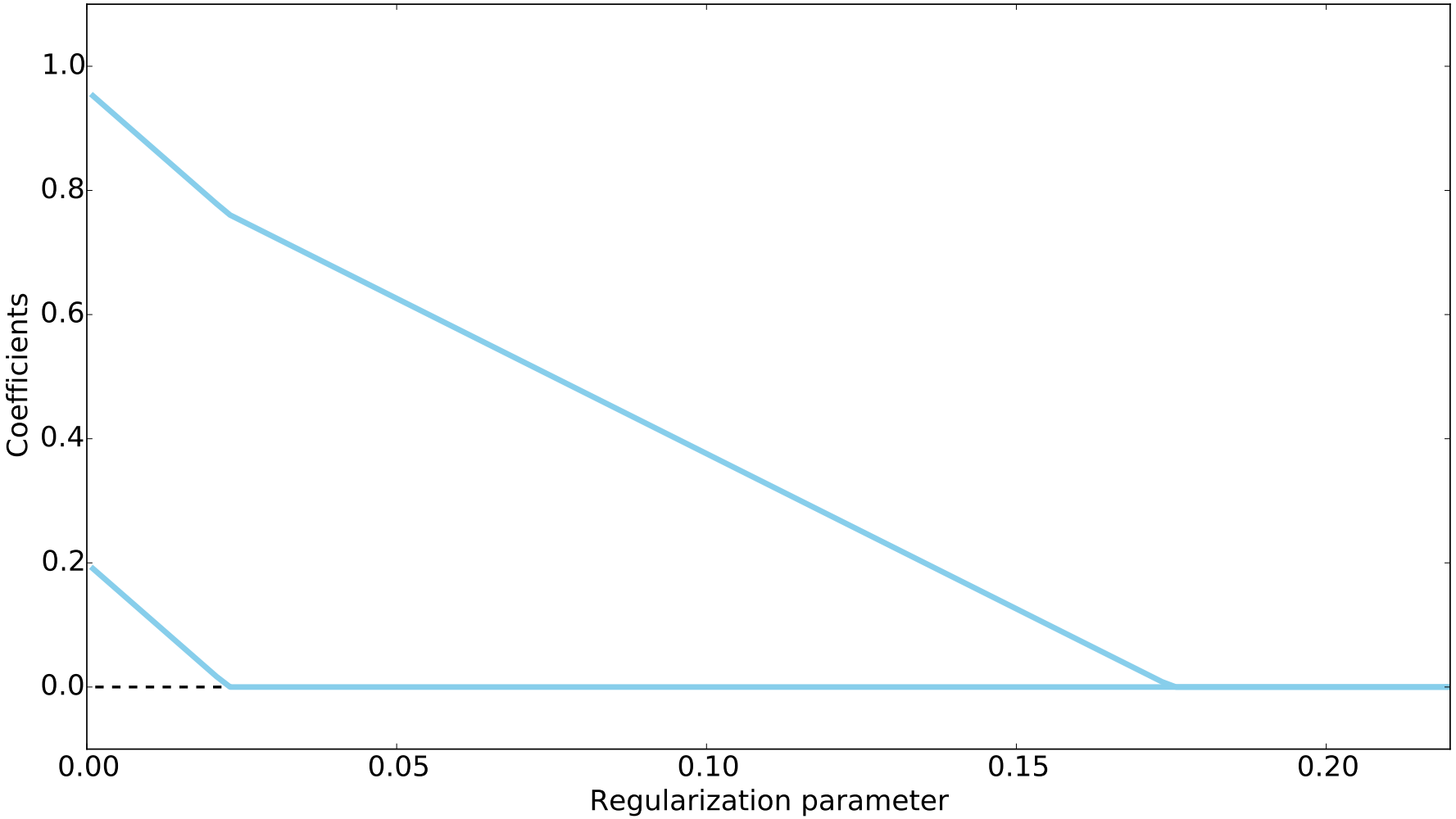
$\lambda = 2$


0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
$\beta[2]$
$\beta_{\mathbf{true}}$
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
$\beta[1]$

$\lambda = 4$


0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
$\beta[2]$
$\beta_{\mathbf{true}}$
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
$\beta[1]$

Coefficients



Lasso

$$\min_{\beta} \frac{1}{2} \left|\left| X^T \beta - y \right|\right|_2^2 + \lambda \left|\left|\beta\right|\right|_1$$

Does this function have a unique minimum? Why?

How can we prove that the minimum is sparse?

Lasso

$$\min_{\beta} \frac{1}{2} \left|\left| X^T \beta - y \right|\right|_2^2 + \lambda \left|\left|\beta\right|\right|_1$$

Does this function have a unique minimum? Why?

Yes if X is full rank and $n \geq p$ because it is strictly convex

How can we prove that the minimum is sparse?

Show that $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ is a subgradient of the cost function at $\begin{bmatrix} \alpha \\ 0 \end{bmatrix}$ for some $\alpha > 0$

Subgradients

The subgradient of $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$ is a vector $g \in \mathbb{R}^n$ such that

$$f(y) \geq f(x) + g^T (y - x), \quad \text{for all } y \in \mathbb{R}^n$$

Subgradients

The subgradient of $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$ is a vector $g \in \mathbb{R}^n$ such that

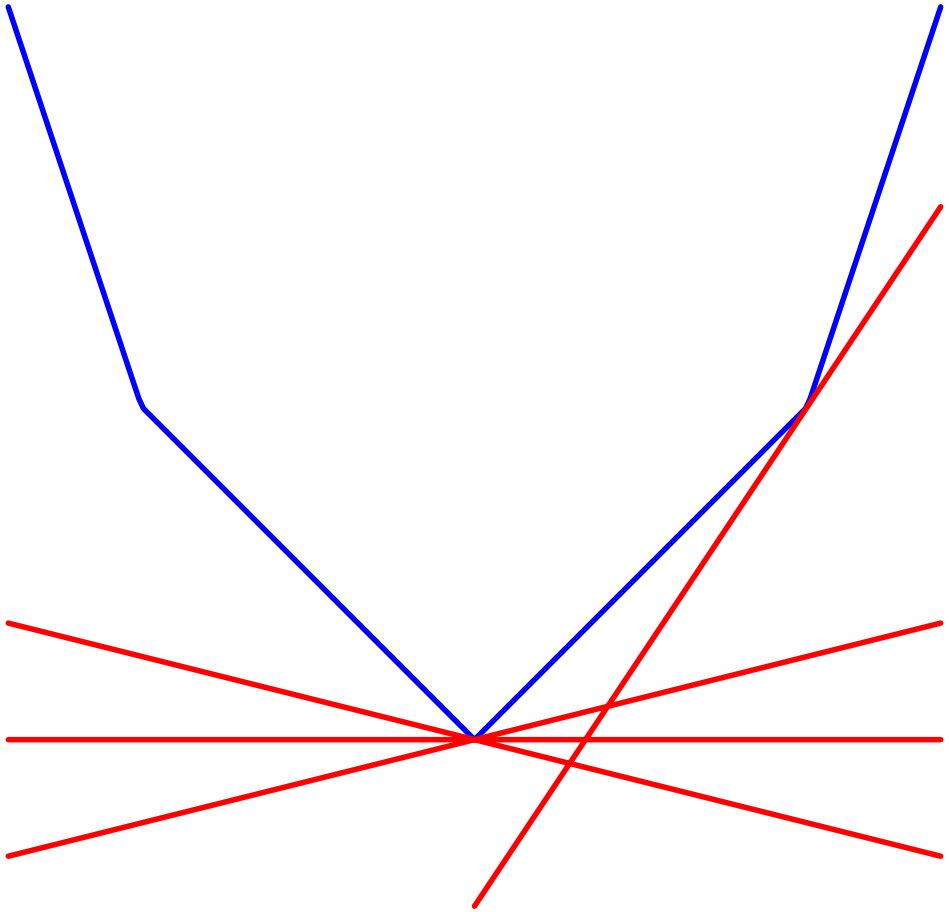
$$f(y) \geq f(x) + g^T (y - x), \quad \text{for all } y \in \mathbb{R}^n$$

The hyperplane

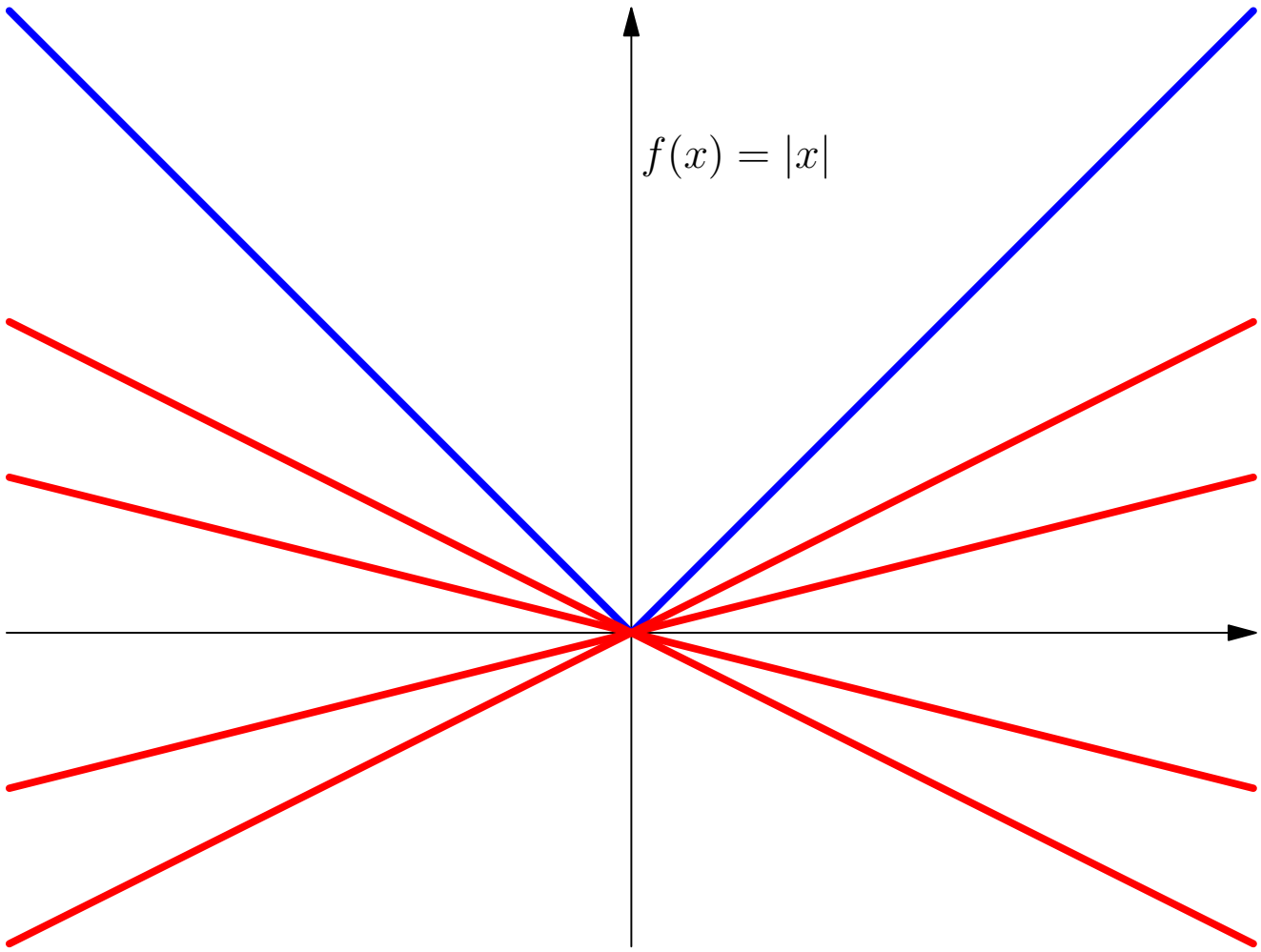
$$\mathcal{H}_g := \left\{ y \mid y[n+1] = f(x) + g^T \left(\begin{bmatrix} y[1] \\ \dots \\ y[n] \end{bmatrix} - x \right) \right\}$$

is a supporting hyperplane of the epigraph of f at $\begin{bmatrix} x \\ f(x) \end{bmatrix}$

Subgradients



Subgradients of absolute value



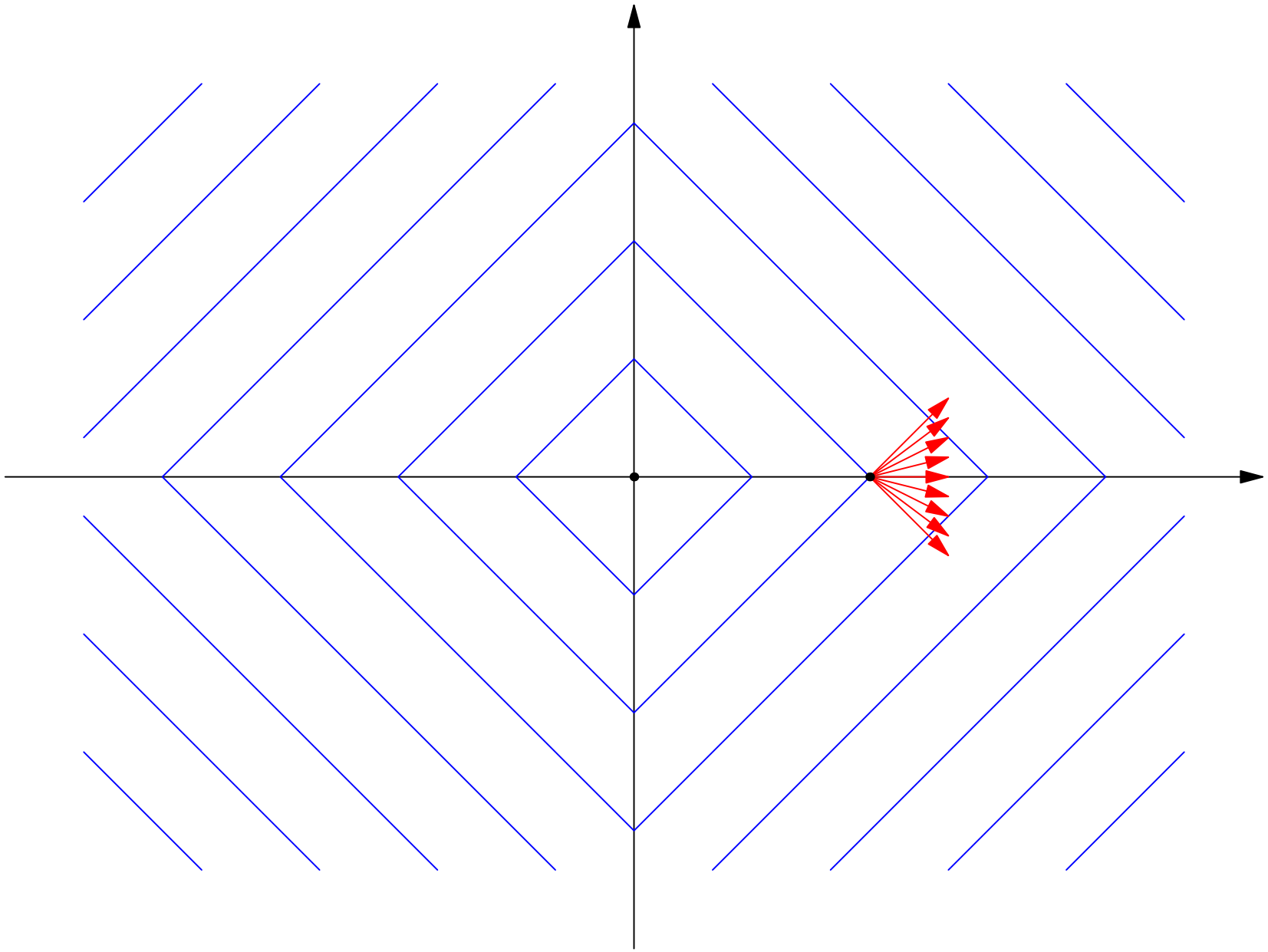
Subgradients of ℓ_1 norm

g is a subgradient of the ℓ_1 norm at $x \in \mathbb{R}^n$ if and only if

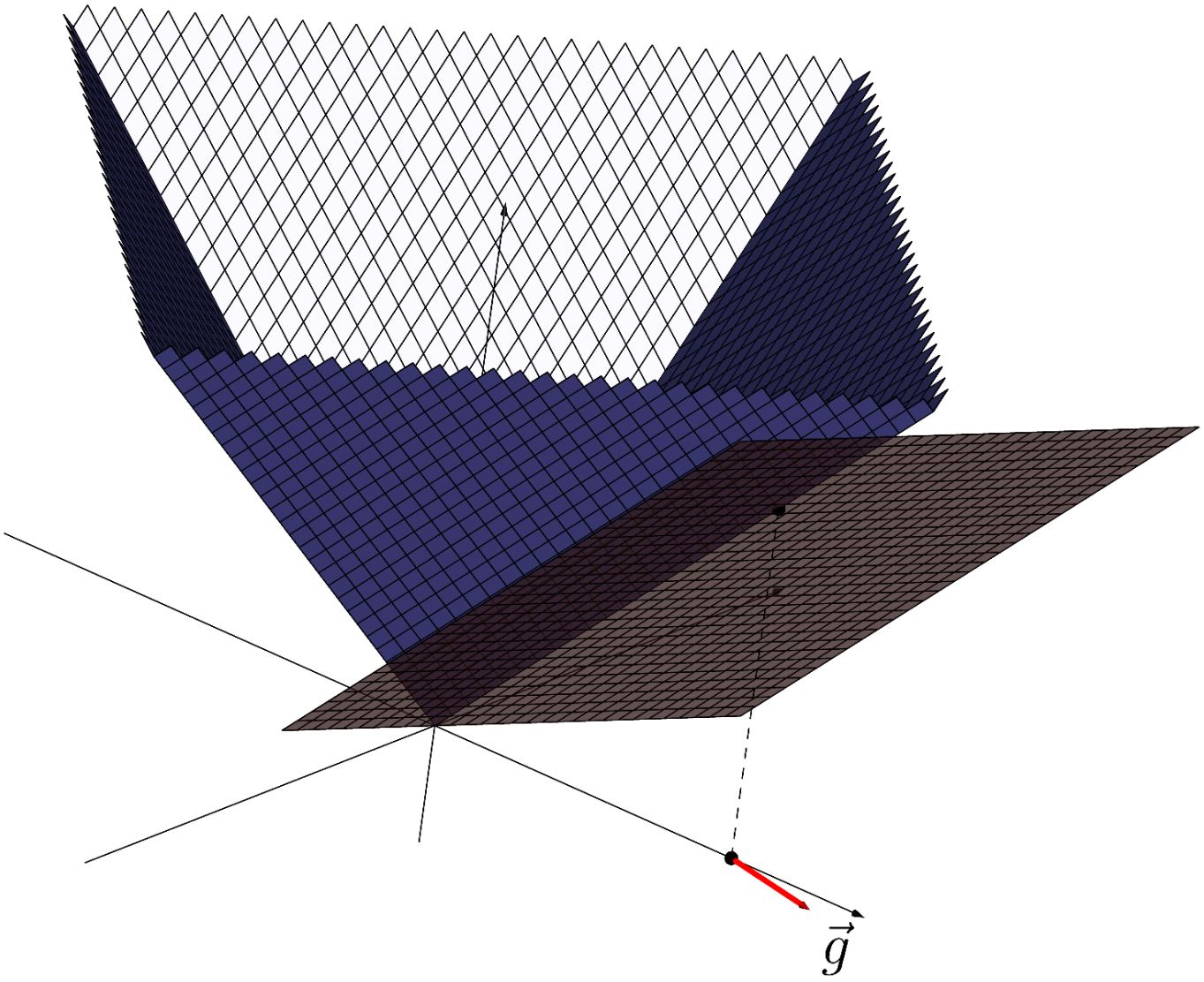
$$g[i] = \text{sign}\left(x[i]\right) \qquad \text{if } x[i] \neq 0$$

$$|g[i]| \leq 1 \qquad \text{if } x[i] = 0$$

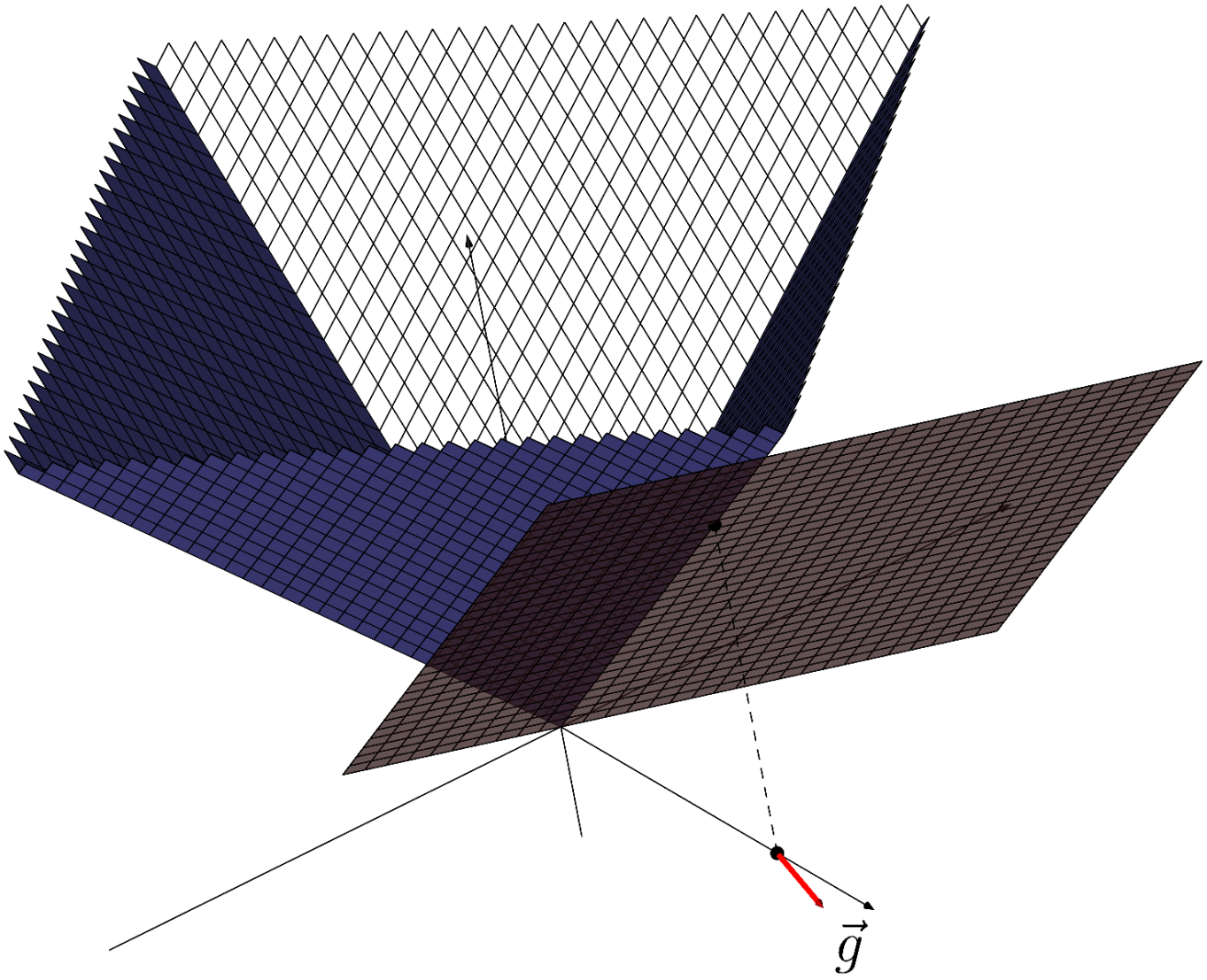
Subgradients of ℓ_1 norm



Subgradients of ℓ_1 norm



Subgradients of ℓ_1 norm



Subgradients of lasso cost function

Gradient of $\frac{1}{2}\left|\left|X^T\beta - y\right|\right|_2^2$ at β_{lasso}:

Subgradient of ℓ_1 norm at $\begin{bmatrix} \alpha \\ 0 \end{bmatrix}$ for some $\alpha > 0$:

Subgradient of lasso cost function at $\begin{bmatrix} \alpha \\ 0 \end{bmatrix}$ for some $\alpha > 0$:

Subgradients of lasso cost function

Gradient of $\frac{1}{2}\left|\left|X^T\beta - y\right|\right|_2^2$ at β_{lasso}:

$$X\left(X^T\beta_{\text{lasso}} - y\right)$$

Subgradient of ℓ_1 norm at $\begin{bmatrix}\alpha \\ 0\end{bmatrix}$ for some $\alpha > 0$:

$$g_{\ell_1} := \begin{bmatrix}1 \\ \gamma\end{bmatrix} \qquad |\gamma| \leq 1$$

Subgradient of lasso cost function at $\begin{bmatrix}\alpha \\ 0\end{bmatrix}$ for some $\alpha > 0$:

$$g_{\text{lasso}} := X\left(X^T\begin{bmatrix}\alpha \\ 0\end{bmatrix} - y\right) + \lambda\begin{bmatrix}1 \\ \gamma\end{bmatrix} \qquad |\gamma| \leq 1$$

Subgradients of lasso cost function

At $\begin{bmatrix} \alpha \\ 0 \end{bmatrix}$ for some $\alpha > 0$

$$g_{\text{lasso}} := X \left(X^T \begin{bmatrix} \alpha \\ 0 \end{bmatrix} - y \right) + \lambda \begin{bmatrix} 1 \\ \gamma \end{bmatrix}$$

Subgradients of lasso cost function

At $\begin{bmatrix} \alpha \\ 0 \end{bmatrix}$ for some $\alpha > 0$

$$
\begin{aligned}
g_{\text{lasso}} &:= X\left(X^T \begin{bmatrix} \alpha \\ 0 \end{bmatrix} - y\right) + \lambda \begin{bmatrix} 1 \\ \gamma \end{bmatrix} \\
&= X\left(\alpha a - a - z\right) + \lambda \begin{bmatrix} 1 \\ \gamma \end{bmatrix} \\
&= \begin{bmatrix} a^T\left((\alpha - 1)a - z\right) + \lambda \\ b^T\left((\alpha - 1)a - z\right) + \lambda\gamma \end{bmatrix} \\
&= \begin{bmatrix} \alpha - 1 - a^T z + \lambda \\ \rho(\alpha - 1) - b^T z + \lambda\gamma \end{bmatrix}
\end{aligned}
$$

Is zero a subgradient for some α?

Is zero a subgradient for some α?

Setting $g_{\text{lasso}} = 0$

$$\begin{aligned}
\alpha &= 1 - \lambda + a^T z \\
\gamma &= \frac{\rho + b^T z - \rho\alpha}{\lambda} \\
&= \frac{b^T z - \rho a^T z}{\lambda} + \rho
\end{aligned}$$

We need $\alpha \geq 0$

$$\lambda \leq 1 + a^T z$$

We need $|\gamma| \leq 1$

$$\frac{\left| b^T z - \rho a^T z \right|}{1 - |\rho|} \leq \lambda$$