Review on Regularization for Linear Regression (blended lecture)

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

Carlos Fernandez-Granda

Goal

Review of coefficient analysis for OLS, ridge regression, and early stopping

Ordinary least squares

Ridge regression

Early stopping

Model for training data

$$\tilde{y}_{\text{train}} := X^T \beta_{\text{true}} + \tilde{z}_{\text{train}}$$

- Feature matrix $X \in \mathbb{R}^{p \times n}$ is deterministic
- Coefficients $\beta_{\text{true}} \in \mathbb{R}^p$ are deterministic
- Noise $\tilde{z}_{\text{train}}$ is an n-dimensional iid Gaussian vector with zero mean and standard deviation $\sigma := 0.2$

Example

$$\tilde{y}_{\text{train}} := X^T \beta_{\text{true}} + \tilde{z}_{\text{train}} = X^T \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \tilde{z}_{\text{train}}$$

$$X := \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} = USV^T$$

$$\Sigma_{\{x_1,\ldots,x_n\}} = XX^T = US^2U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0.01 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\Sigma_{\tilde{z}_{\text{train}}} = 0.04I$$

Decomposition of OLS cost function

$$\arg\min_{\beta} \|\tilde{y}_{\text{train}} - X^T \beta\|_2^2$$

Decomposition of OLS cost function

$$\arg\min_{\beta} \|\tilde{y}_{\text{train}} - X^T\beta\|_2^2$$

$$= \arg\min_{\beta} \|\tilde{z}_{\text{train}} - X^T(\beta - \beta_{\text{true}})\|_2^2$$

$$= \arg\min_{\beta} (\beta - \beta_{\text{true}})^T XX^T(\beta - \beta_{\text{true}}) - 2\tilde{z}_{\text{train}}^T X^T(\beta - \beta_{\text{true}}) + \tilde{z}_{\text{train}}^T \tilde{z}_{\text{train}}$$

$$= \arg\min_{\beta} (\beta - \beta_{\text{true}})^T XX^T(\beta - \beta_{\text{true}}) - 2\tilde{z}_{\text{train}}^T X^T\beta$$

Quadratic component $(\beta - \beta_{\text{true}})^T XX^T (\beta - \beta_{\text{true}})$

Contour lines?

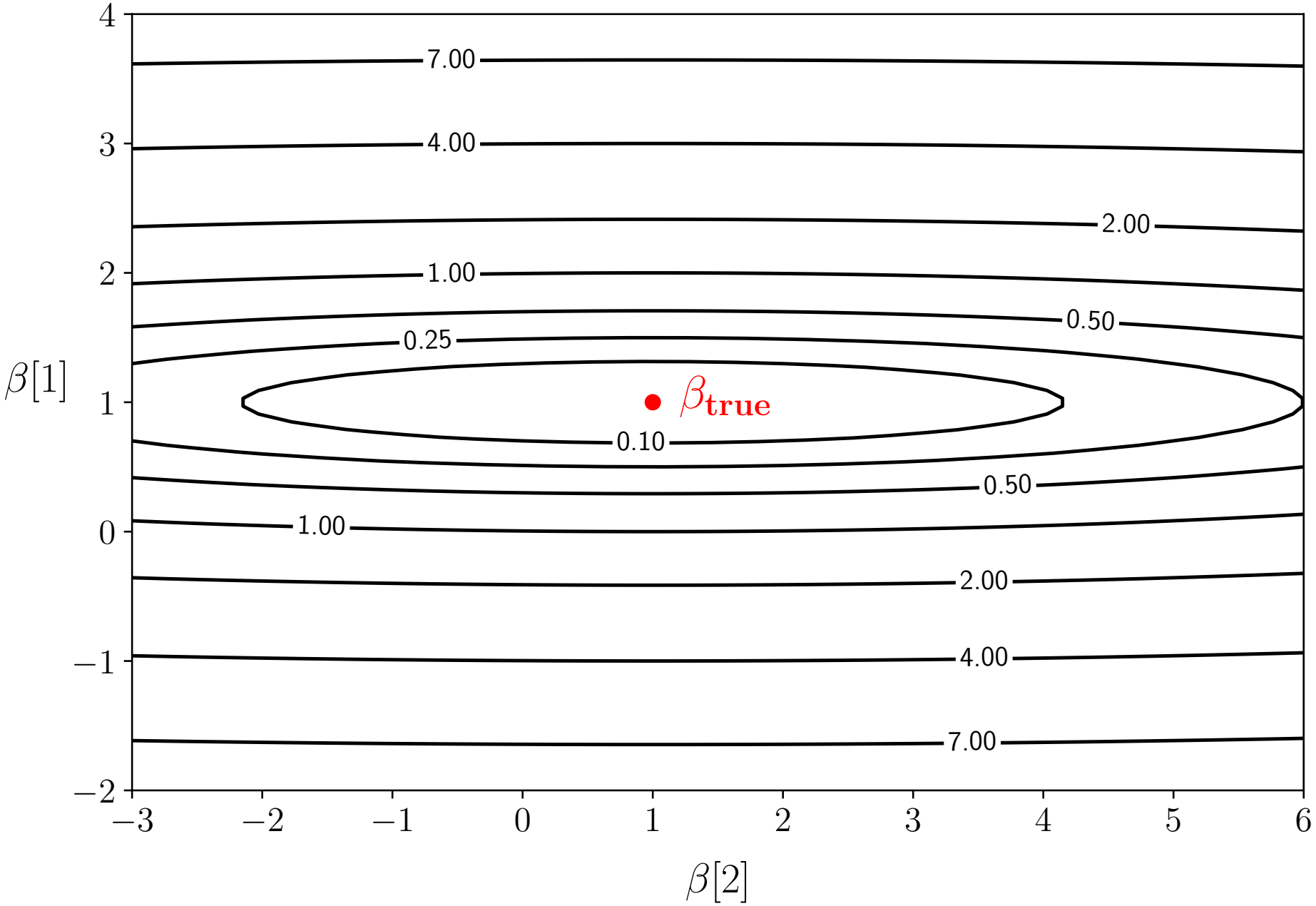
$$c^2 = (\beta - \beta_{\text{true}})^T XX^T (\beta - \beta_{\text{true}})$$

Quadratic component $(\beta - \beta_{\text{true}})^T XX^T (\beta - \beta_{\text{true}})$

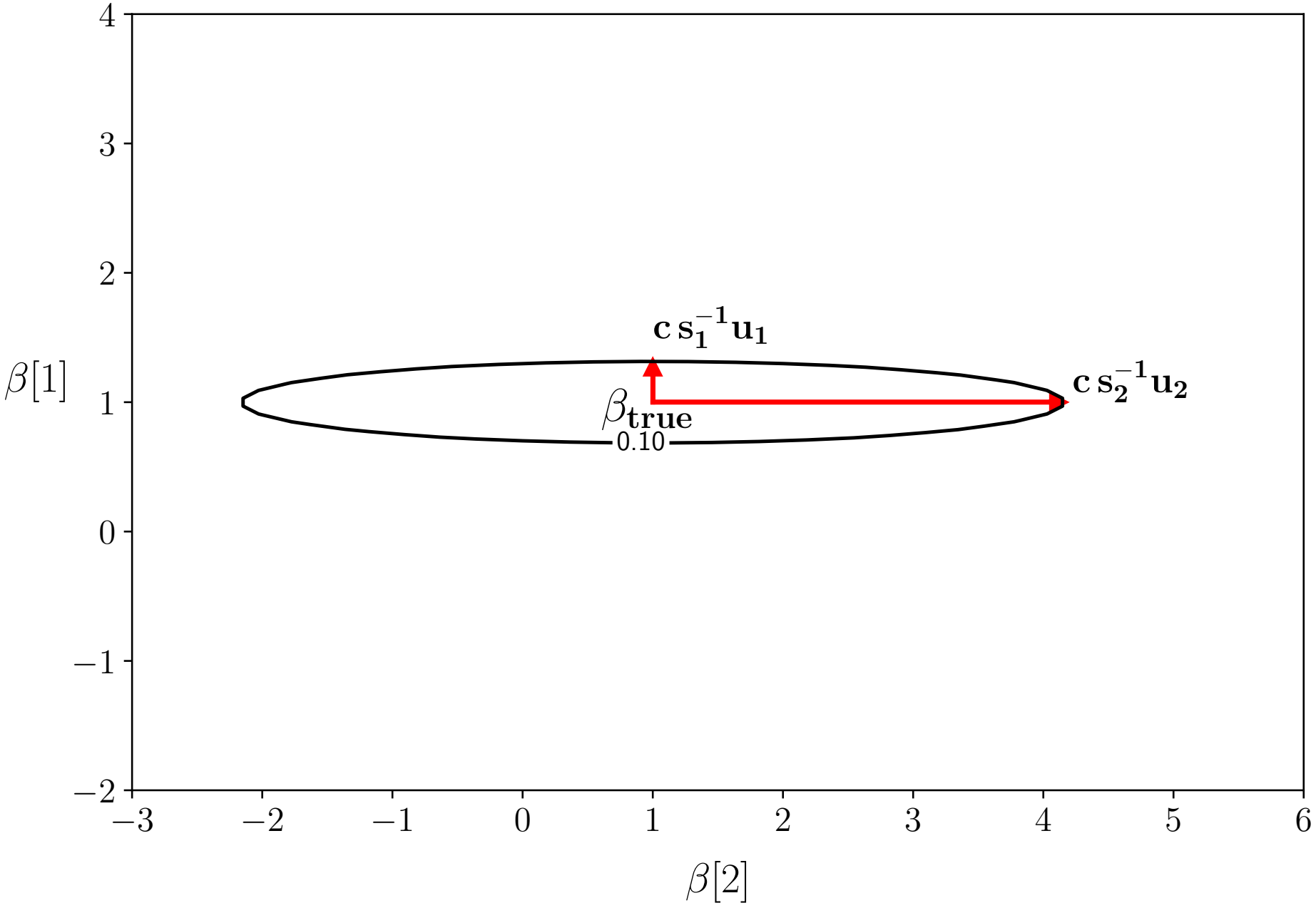
Contour lines?

$$
\begin{aligned}
c^2 &= (\beta - \beta_{\text{true}})^T XX^T (\beta - \beta_{\text{true}}) \\
&= (\beta - \beta_{\text{true}})^T US^2 U^T (\beta - \beta_{\text{true}}) \\
&= \begin{bmatrix} \beta[1] - 1 \\ \beta[2] - 1 \end{bmatrix}^T \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0.01 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta[1] - 1 \\ \beta[2] - 1 \end{bmatrix} \\
&= (\beta[1] - 1)^2 + \frac{(\beta[2] - 1)^2}{100}
\end{aligned}
$$

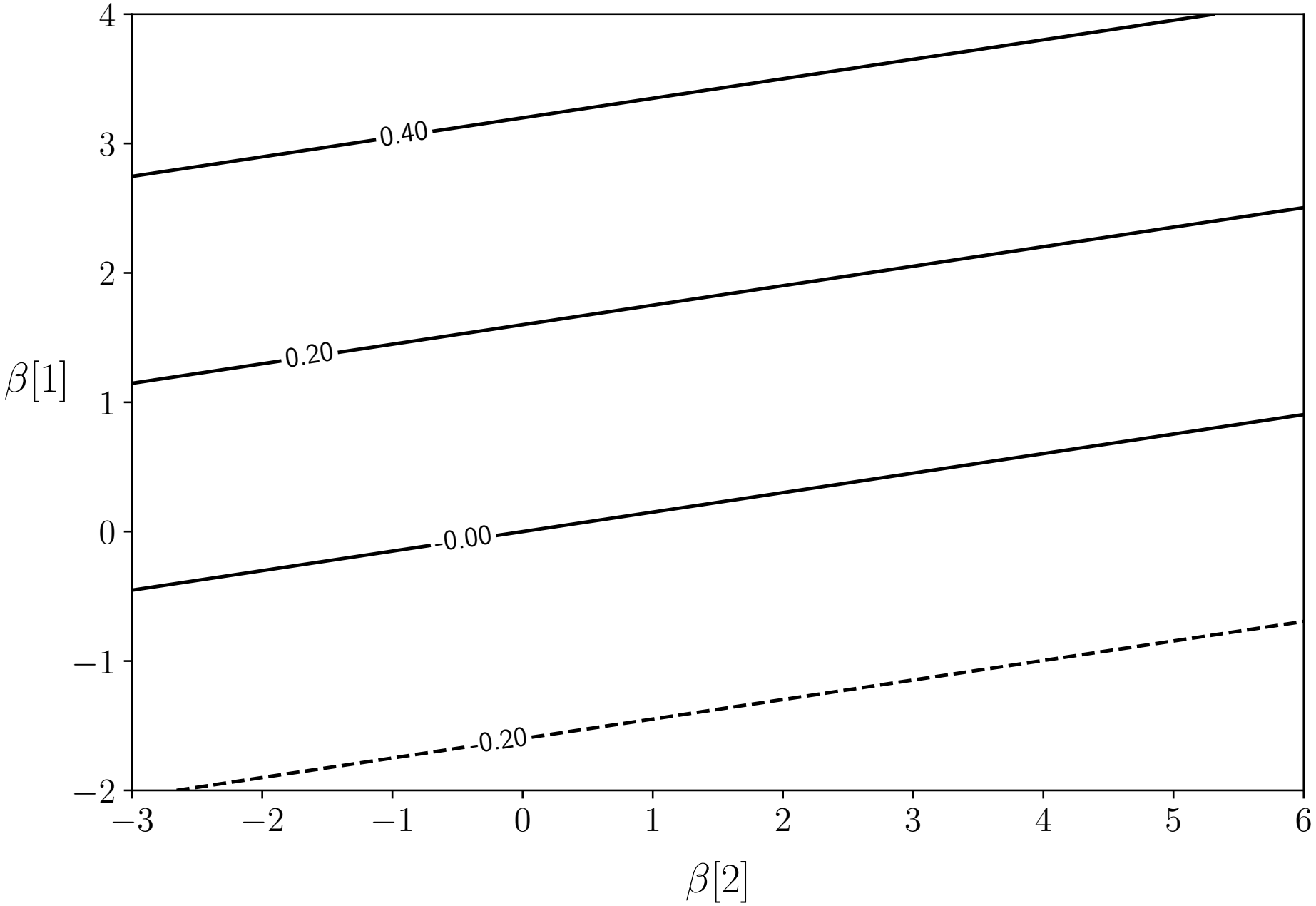
$$(\beta - \beta_{\text{true}})^T XX^T (\beta - \beta_{\text{true}})$$



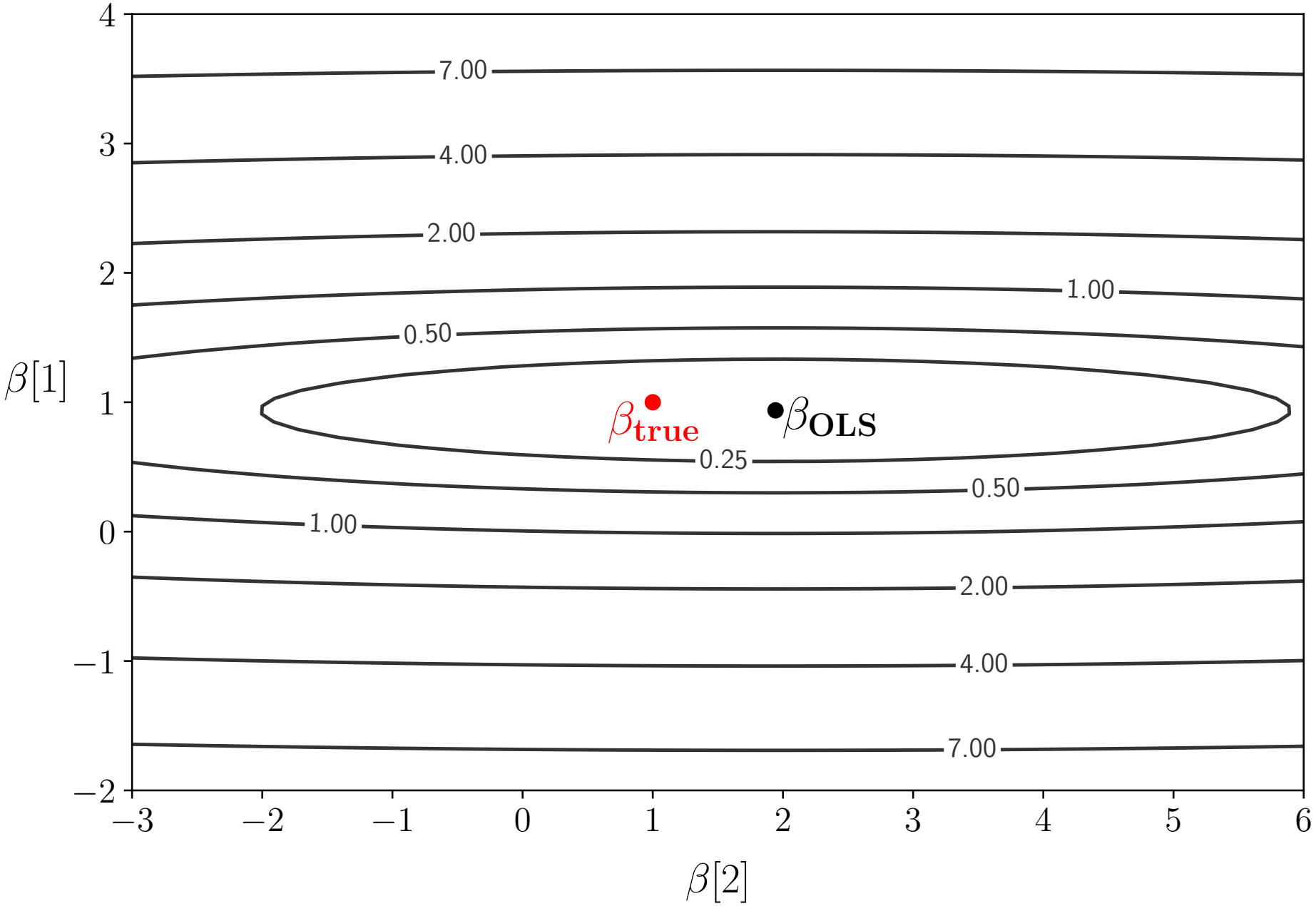
$$(\beta - \beta_{\text{true}})^T XX^T (\beta - \beta_{\text{true}})$$



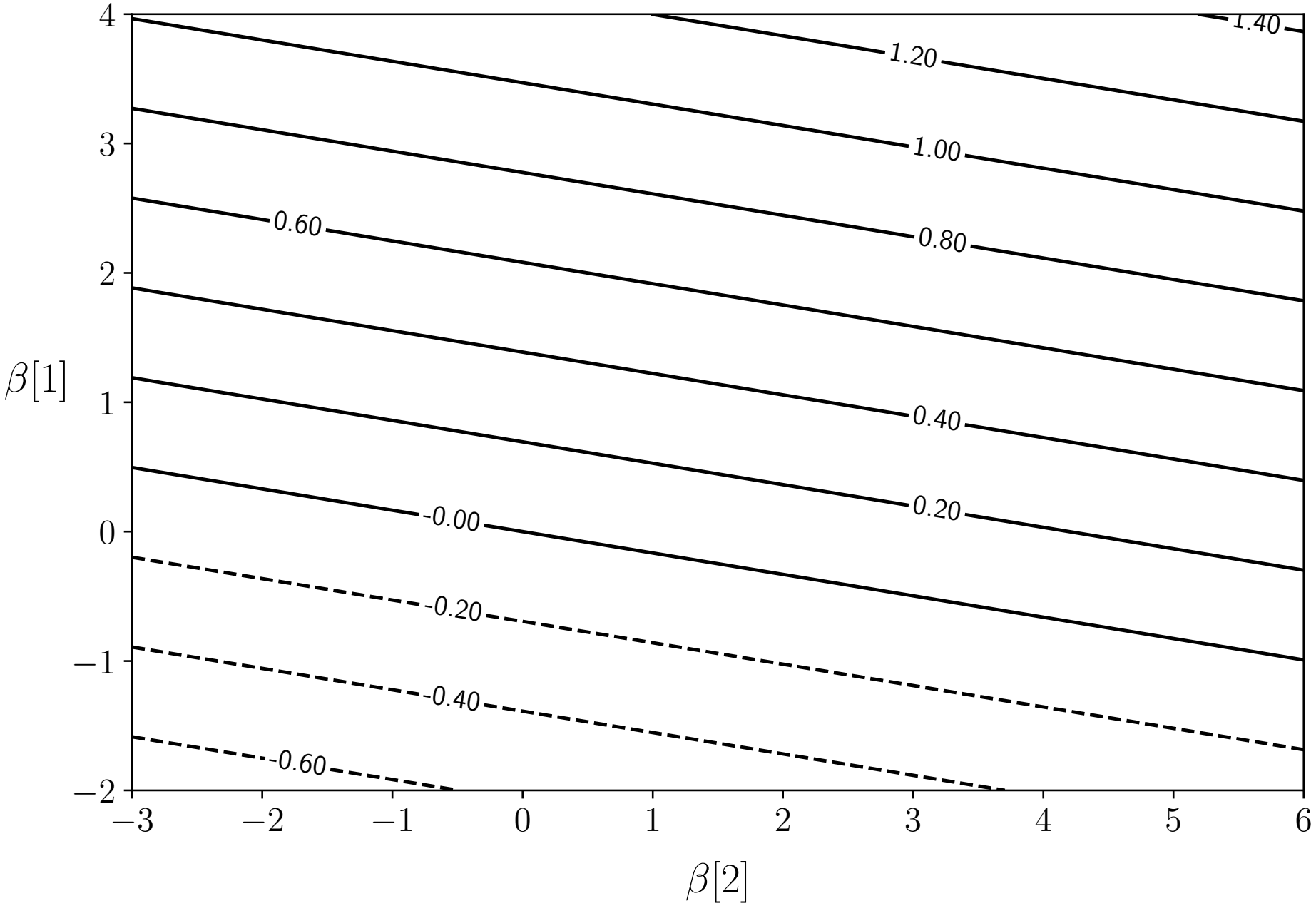
$-2\tilde{z}_{\text{train}}^T X^T \beta$



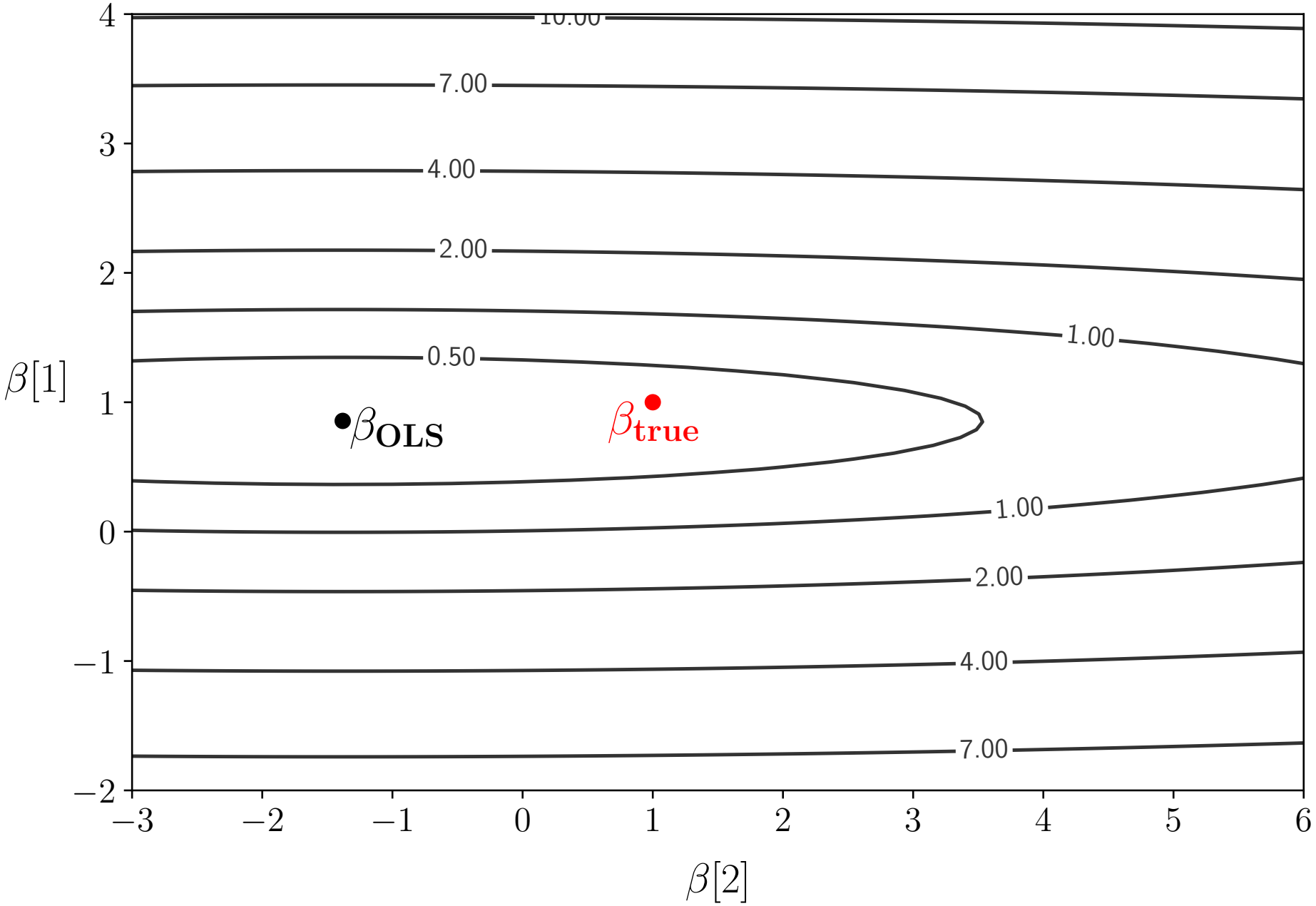
$$(\beta - \beta_{\text{true}})^T XX^T(\beta - \beta_{\text{true}}) - 2\tilde{z}_{\text{train}}^T X^T \beta$$



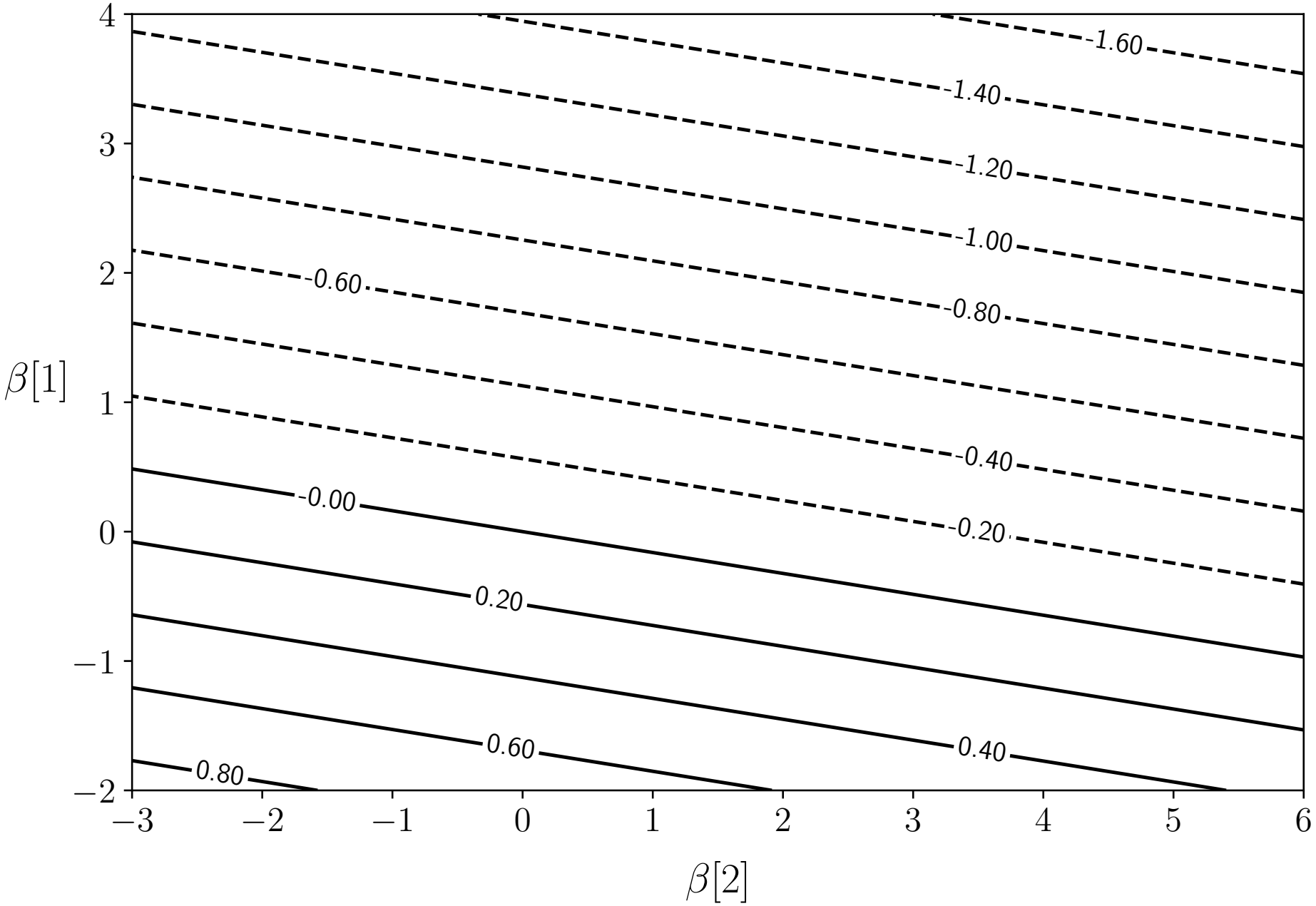
$-2\tilde{z}_{\text{train}}^T X^T \beta$



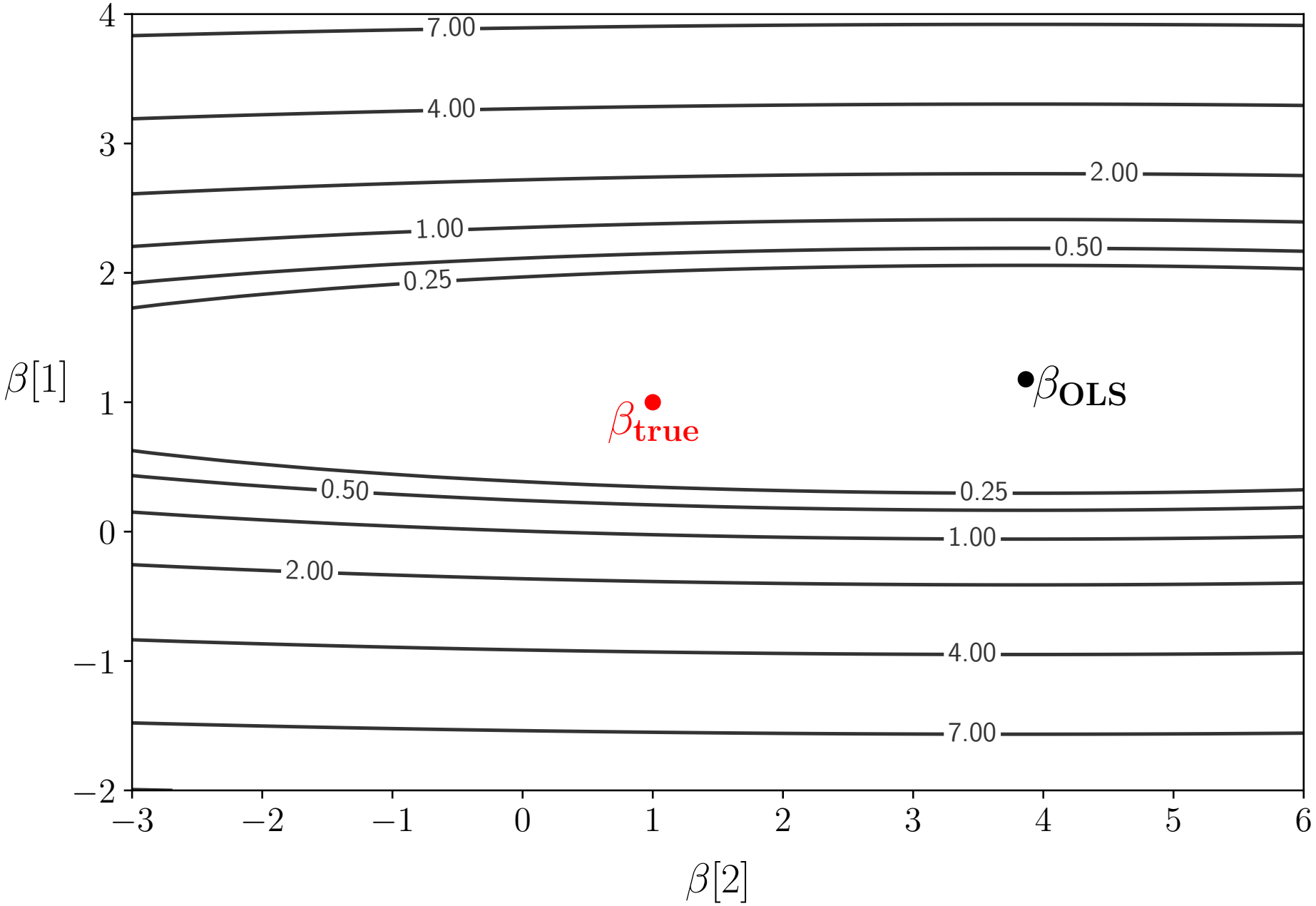
$$(\beta - \beta_{\text{true}})^T XX^T(\beta - \beta_{\text{true}}) - 2\tilde{z}_{\text{train}}^T X^T \beta$$



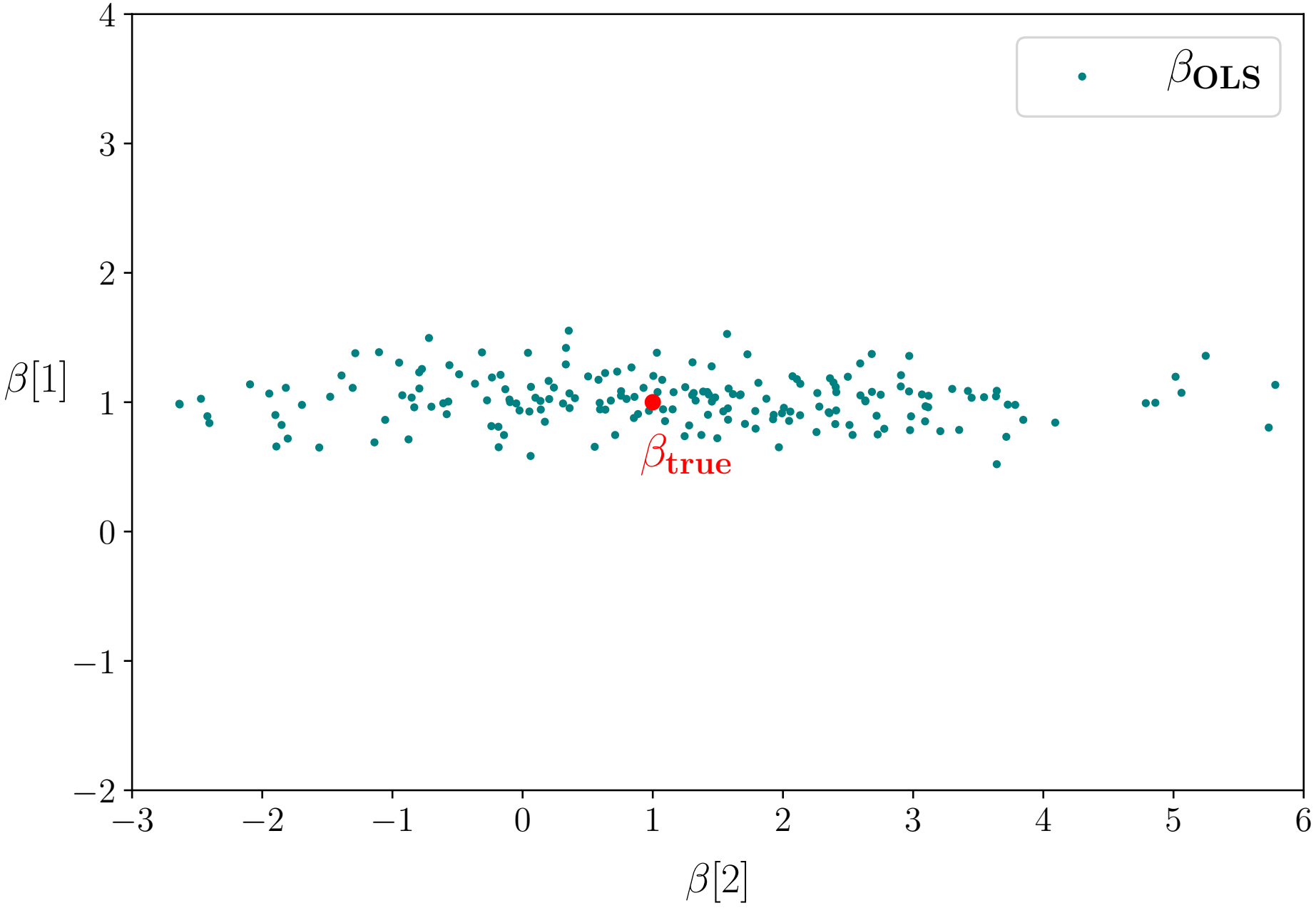
$-2\tilde{z}_{\text{train}}^T X^T \beta$



$$(\beta - \beta_{\text{true}})^T XX^T(\beta - \beta_{\text{true}}) - 2\tilde{z}_{\text{train}}^T X^T \beta$$



Minima for 200 realizations



Minima

$$\tilde{\beta}_{\mathsf{OLS}} = (XX^T)^{-1} X \tilde{y}_{\mathsf{train}}$$

Minima

$$
\begin{aligned}
\tilde{\beta}_{\text{OLS}} &= (XX^T)^{-1} X \tilde{y}_{\text{train}} \\
&= (XX^T)^{-1} XX^T \beta_{\text{true}} + (XX^T)^{-1} X \tilde{z}_{\text{train}} \\
&= \beta_{\text{true}} + (XX^T)^{-1} X \tilde{z}_{\text{train}} \\
&= \beta_{\text{true}} + US^{-1} V^T \tilde{z}_{\text{train}}
\end{aligned}
$$

Distribution?

Linear transformations of Gaussian random vectors

Let $\tilde{x}$ be a Gaussian random vector of dimension d with mean μ and covariance matrix Σ

For any matrix $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$ $\tilde{y} = A\tilde{x} + b$ is Gaussian with mean $A\mu + b$ and covariance matrix $A\Sigma A^T$ (as long as it is full rank)

Minima

$$
\begin{aligned}
\beta_{\text{OLS}} &= (XX^T)^{-1} X \tilde{y}_{\text{train}} \\
&= (XX^T)^{-1} XX^T \beta_{\text{true}} + (XX^T)^{-1} X \tilde{z}_{\text{train}} \\
&= \beta_{\text{true}} + (XX^T)^{-1} X \tilde{z}_{\text{train}} \\
&= \beta_{\text{true}} + US^{-1} V^T \tilde{z}_{\text{train}}
\end{aligned}
$$

Distribution?

Minima

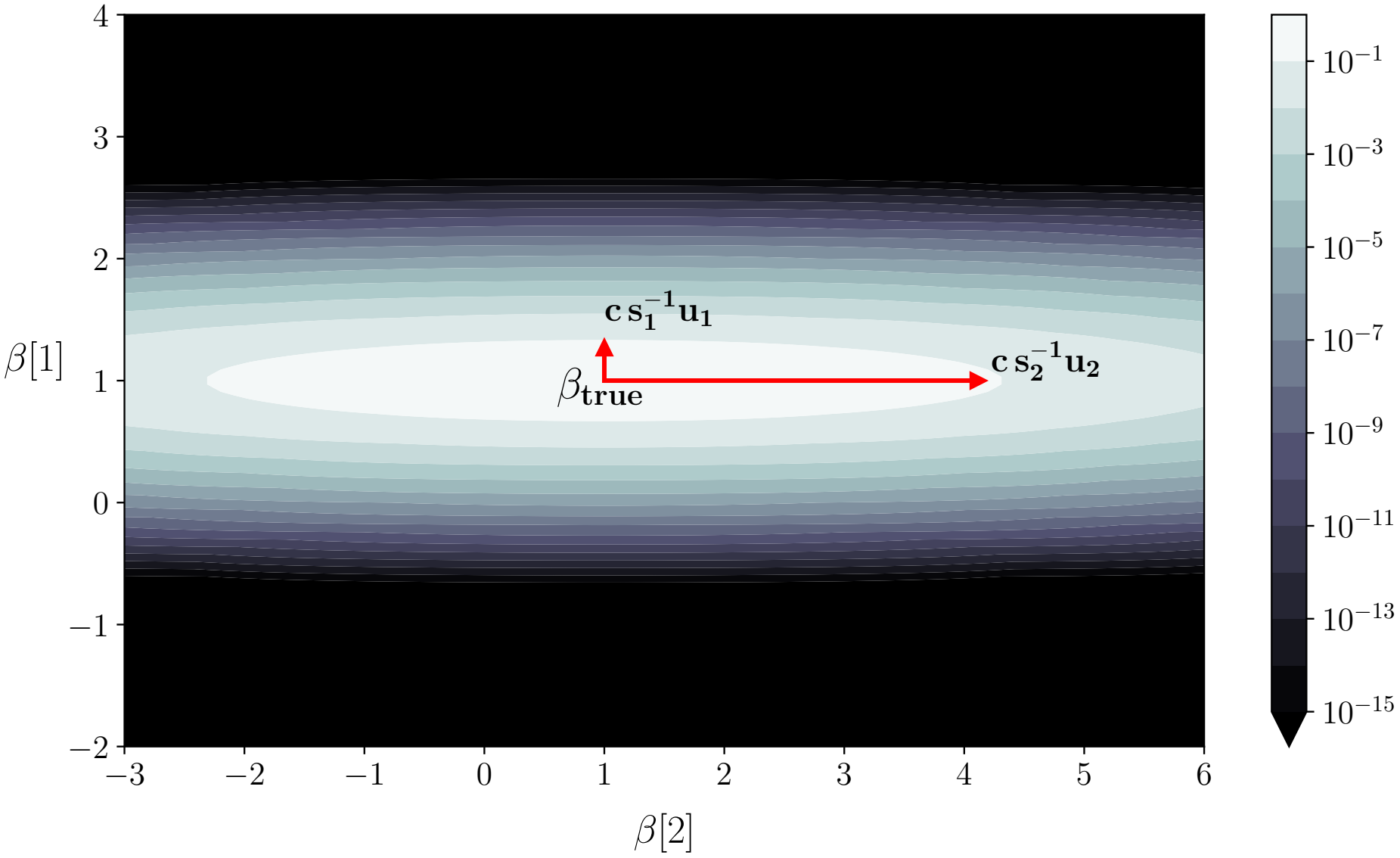
$$
\begin{aligned}
\beta_{\text{OLS}} &= (XX^T)^{-1} X \tilde{y}_{\text{train}} \\
&= (XX^T)^{-1} XX^T \beta_{\text{true}} + (XX^T)^{-1} X \tilde{z}_{\text{train}} \\
&= \beta_{\text{true}} + (XX^T)^{-1} X \tilde{z}_{\text{train}} \\
&= \beta_{\text{true}} + US^{-1} V^T \tilde{z}_{\text{train}}
\end{aligned}
$$

Distribution?

Gaussian with mean $\beta_{\text{true}} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and covariance matrix

$$\sigma^2 US^{-2} U^T = \begin{bmatrix} 0.04 & 0 \\ 0 & 4 \end{bmatrix}$$

Minima


$\mathbf{c\,s_1^{-1}u_1}$
$\mathbf{c\,s_2^{-1}u_2}$
$\beta_{\mathbf{true}}$
$\beta[1]$
$\beta[2]$
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4
5
6
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
10^{-11}
10^{-13}
10^{-15}

Ordinary least squares

Ridge regression

Early stopping

Decomposition of ridge-regression cost function

$$\arg\min_{\beta} \|\tilde{y}_{\text{train}} - X^T\beta\|_2^2 + \lambda\|\beta\|_2^2$$

Decomposition of ridge-regression cost function

$$\arg\min_{\beta} \|\tilde{y}_{\text{train}} - X^T\beta\|_2^2 + \lambda\|\beta\|_2^2$$

$$= \arg\min_{\beta} (\beta - \beta_{\text{true}})^T XX^T(\beta - \beta_{\text{true}}) + \lambda\beta^T\beta - 2\tilde{z}_{\text{train}}^T X^T\beta$$

Quadratic component $(\beta - \beta_{\text{true}})^T XX^T(\beta - \beta_{\text{true}}) + \lambda\beta^T\beta$

Contour lines? To simplify notation $a := \beta[1]$, $b := \beta[2]$

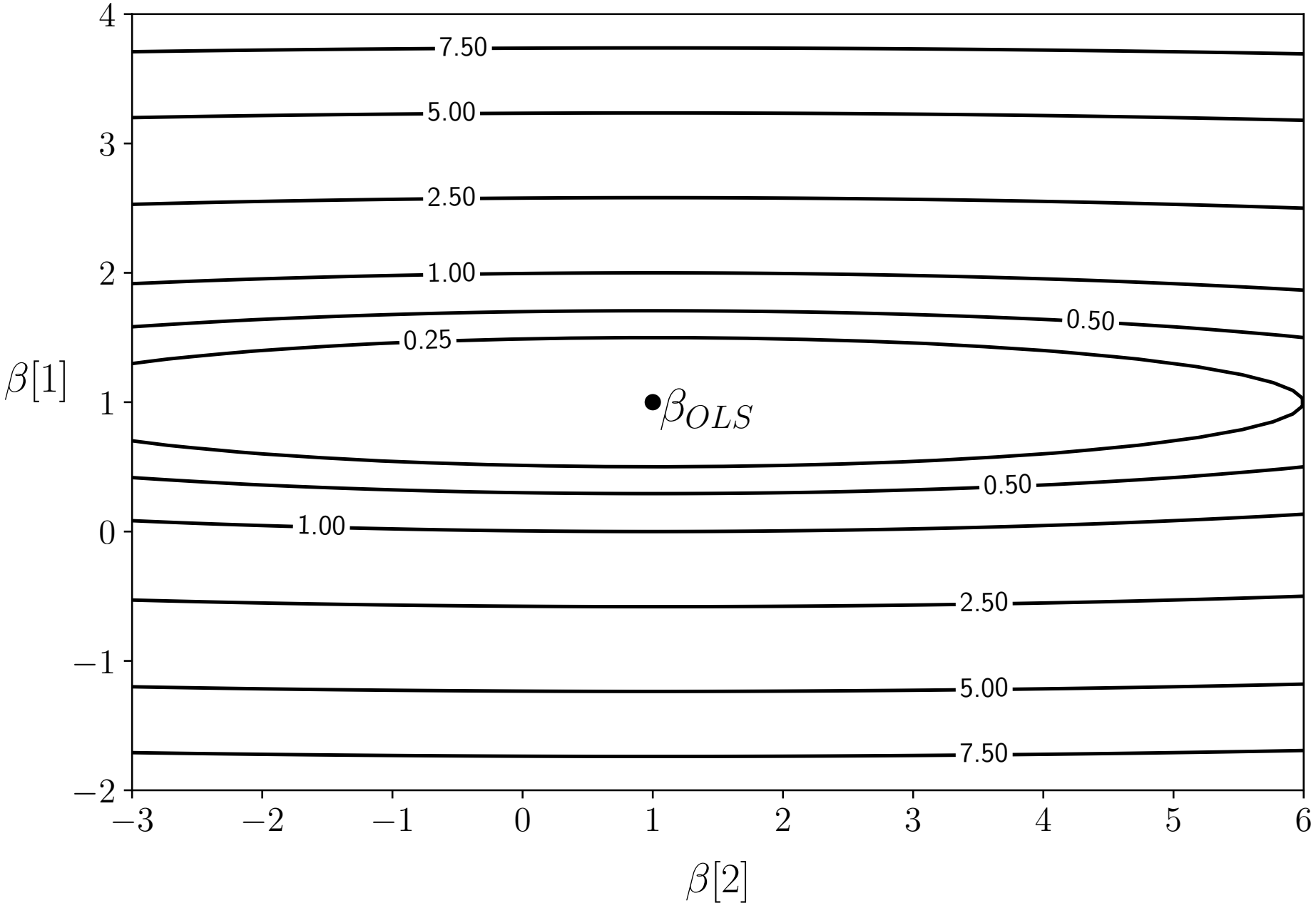
$$c^2 = (\beta - \beta_{\text{true}})^T XX^T(\beta - \beta_{\text{true}}) + \lambda\beta^T\beta$$

Quadratic component $(\beta - \beta_{\text{true}})^T XX^T(\beta - \beta_{\text{true}}) + \lambda\beta^T\beta$

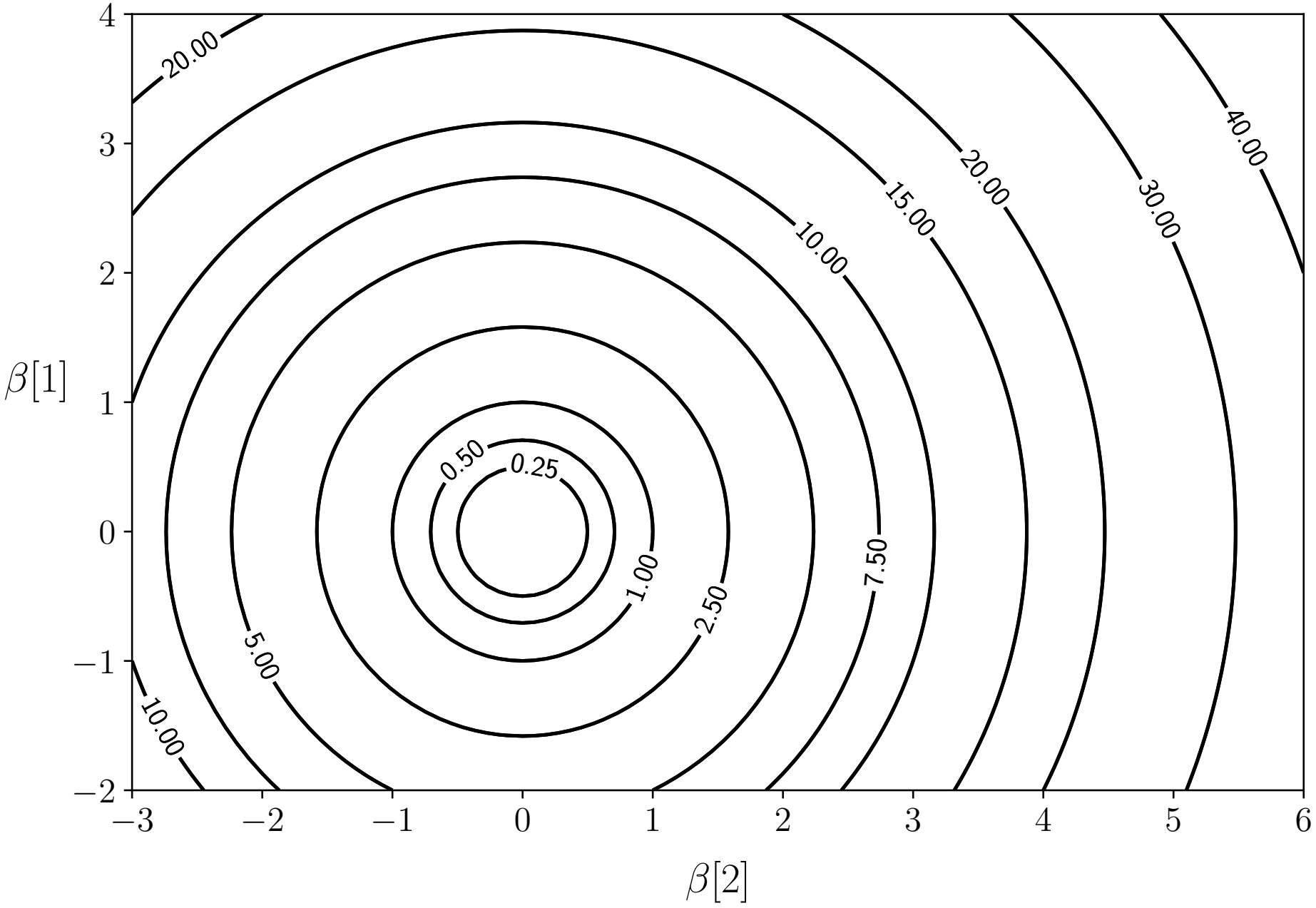
Contour lines? To simplify notation $a := \beta[1]$, $b := \beta[2]$

$$
\begin{aligned}
c^2 &= (\beta - \beta_{\text{true}})^T XX^T(\beta - \beta_{\text{true}}) + \lambda\beta^T\beta \\
&= \begin{bmatrix} a-1 \\ b-1 \end{bmatrix}^T \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0.01 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a-1 \\ b-1 \end{bmatrix} + \lambda\beta^T\beta \\
&= (a-1)^2 + \frac{(b-1)^2}{100} + \lambda(a^2 + b^2) \\
&= (1+\lambda)a^2 - 2a + 1 + \frac{(1+100\lambda)\, b^2}{100} - \frac{2b+1}{100} \\
f(c,\lambda) &= (1+\lambda)\left(a - \frac{1}{1+\lambda}\right)^2 + \frac{1+100\lambda}{100}\left(b - \frac{1}{1+100\lambda}\right)^2
\end{aligned}
$$

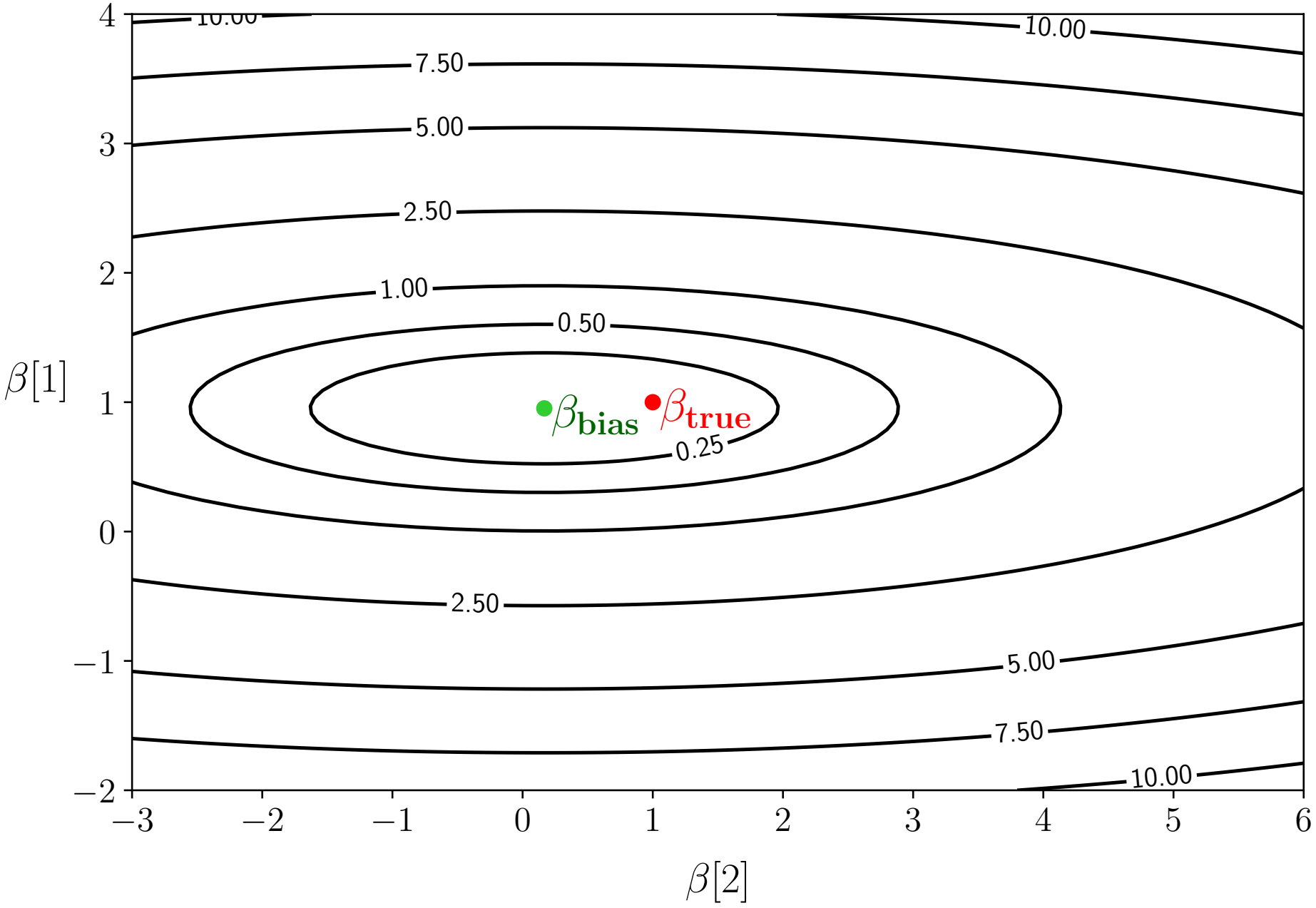
$$(\beta - \beta_{\text{true}})^T XX^T (\beta - \beta_{\text{true}})$$



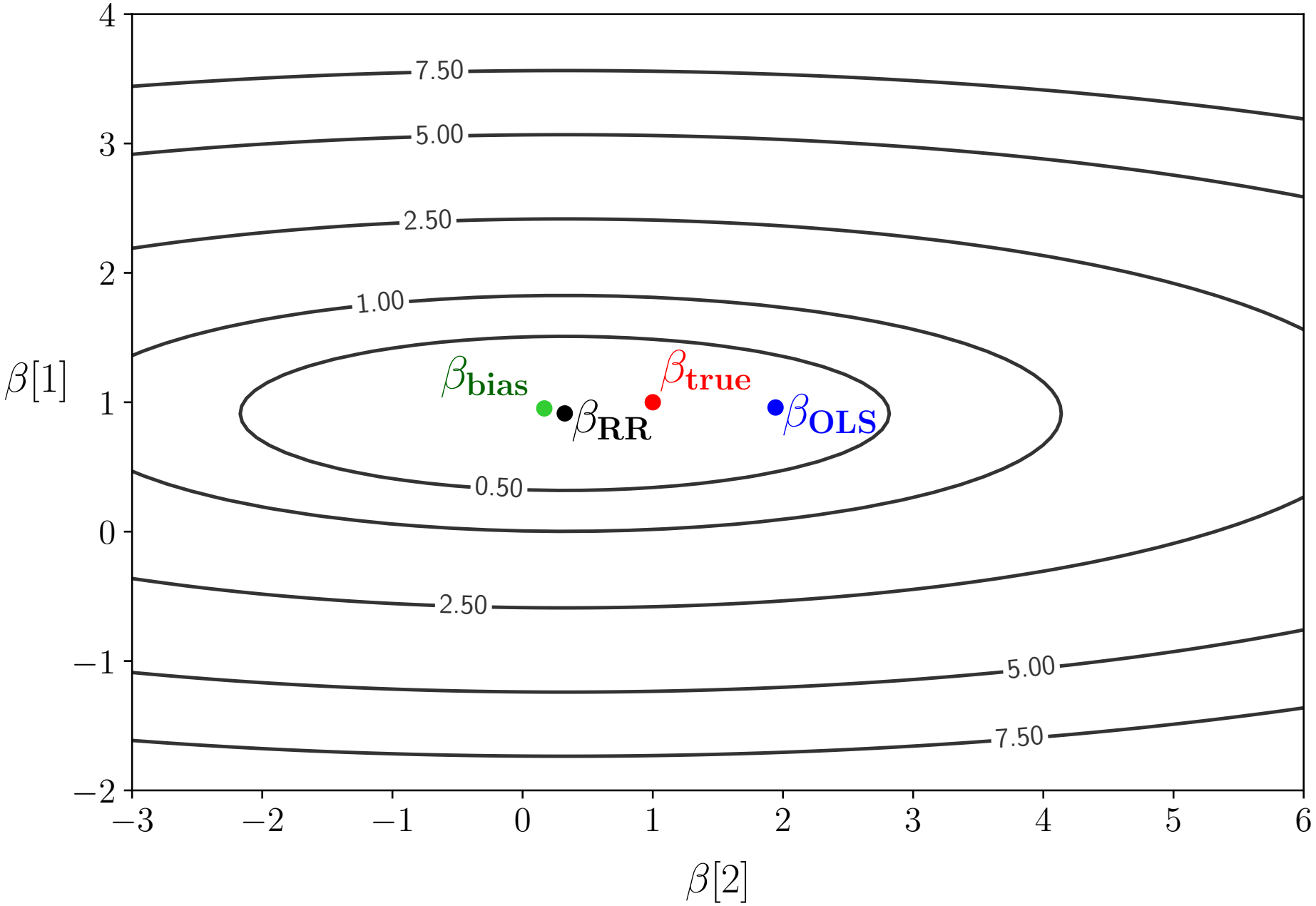
$\beta^T \beta$


20.00
40.00
20.00
30.00
15.00
10.00
0.50
0.25
1.00
2.50
7.50
5.00
10.00
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4
5
6
$\beta[1]$
$\beta[2]$

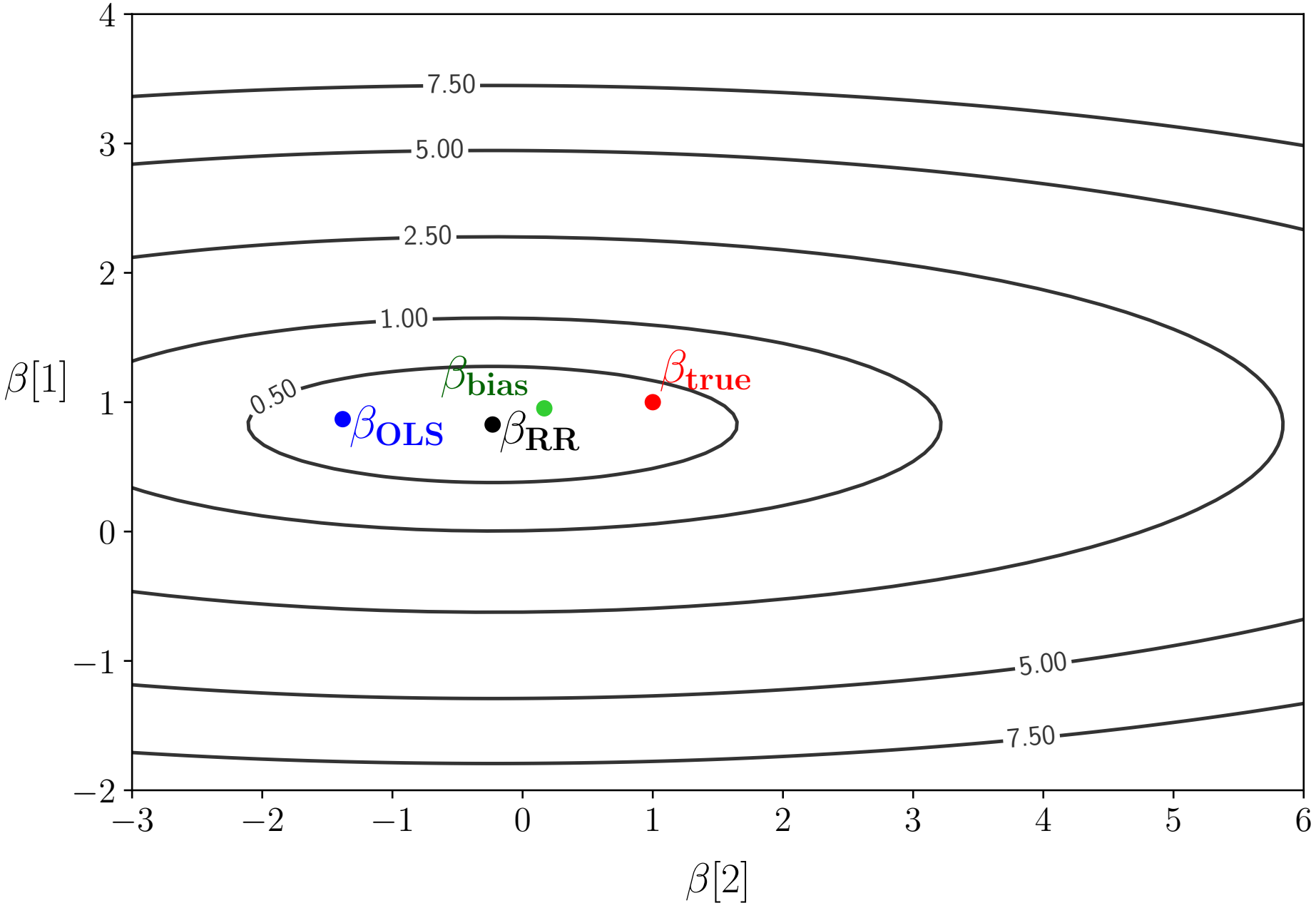
$$(\beta - \beta_{\text{true}})^T XX^T (\beta - \beta_{\text{true}}) + \lambda \beta^T \beta$$



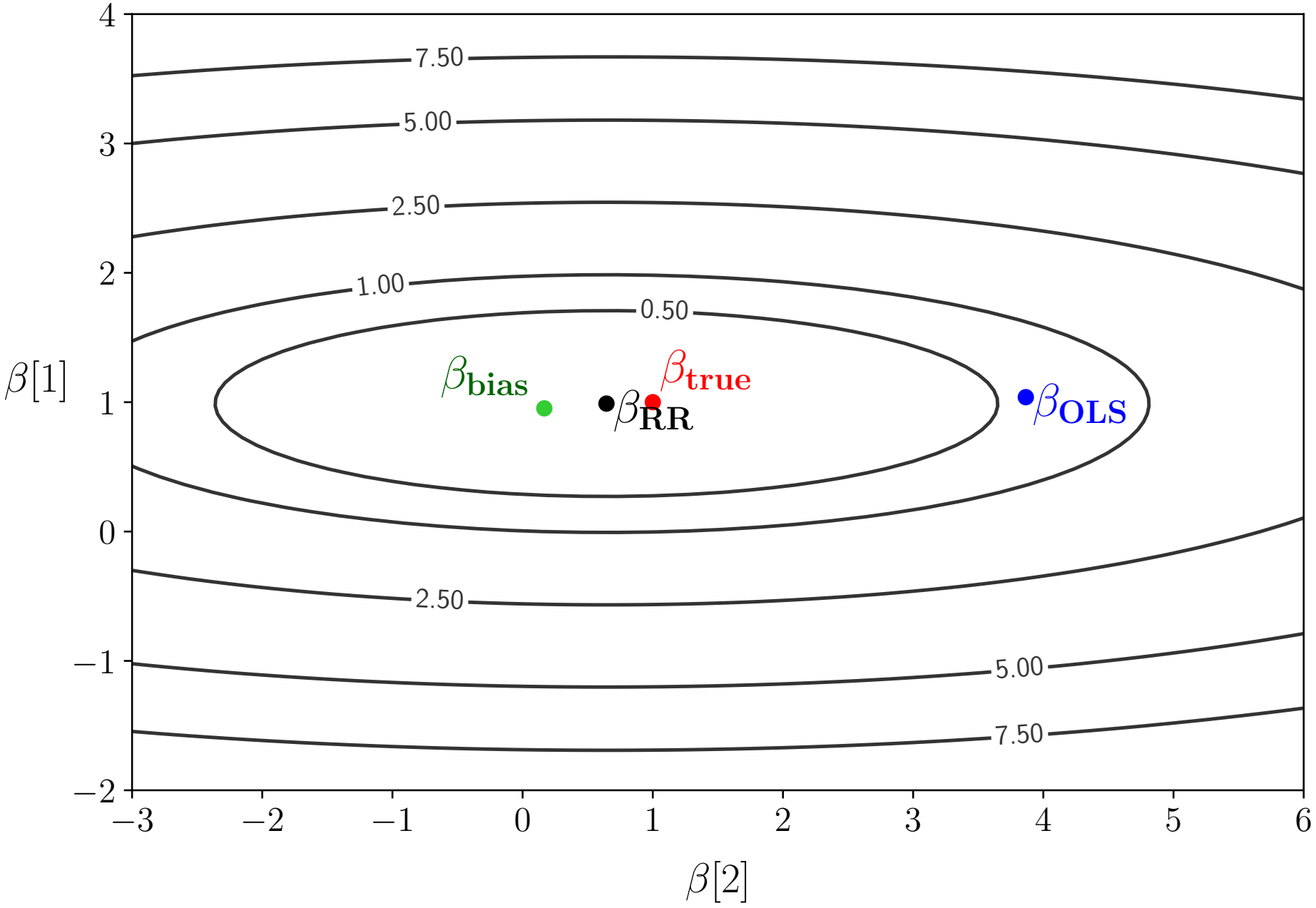
$$(\beta - \beta_{\text{true}})^T XX^T(\beta - \beta_{\text{true}}) + \lambda\beta^T\beta - 2\tilde{z}_{\text{train}}^T X^T\beta$$



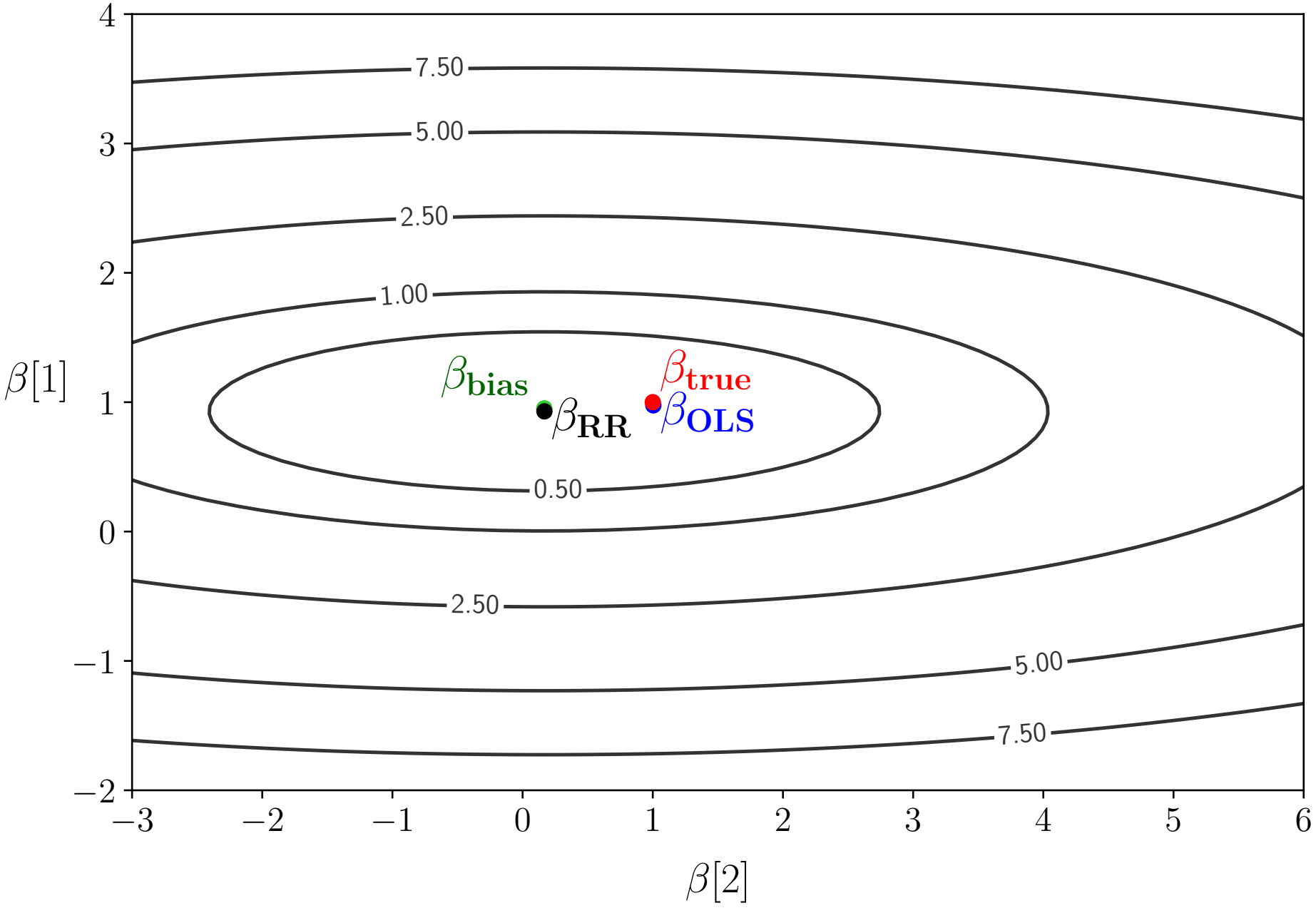
$$(\beta - \beta_{\text{true}})^T X X^T (\beta - \beta_{\text{true}}) + \lambda \beta^T \beta - 2\tilde{z}_{\text{train}}^T X^T \beta$$



$$(\beta - \beta_{\text{true}})^T XX^T(\beta - \beta_{\text{true}}) + \lambda\beta^T\beta - 2\tilde{z}_{\text{train}}^T X^T\beta$$



$$(\beta - \beta_{\text{true}})^T XX^T(\beta - \beta_{\text{true}}) + \lambda\beta^T\beta - 2\tilde{z}_{\text{train}}^T X^T\beta$$



Ridge-regression coefficient estimate

$$\tilde{\beta}_{\mathsf{RR}} = \left(XX^T + \lambda I\right)^{-1} X \left(X^T \beta_{\mathsf{true}} + \tilde{z}_{\mathsf{train}}\right)$$

Ridge-regression coefficient estimate

$$\tilde{\beta}_{\mathsf{RR}} = \left(XX^T + \lambda I\right)^{-1} X \left(X^T \beta_{\mathsf{true}} + \tilde{z}_{\mathsf{train}}\right)$$

$$= \begin{bmatrix} \frac{1}{1+\lambda} \\ \frac{0.01}{0.01+\lambda} \end{bmatrix} + \begin{bmatrix} \frac{1}{1+\lambda} & 0 \\ 0 & \frac{1}{0.01+\lambda} \end{bmatrix} X \tilde{z}_{\mathsf{train}}$$

Ridge-regression coefficient estimate

$$\tilde{\beta}_{\mathsf{RR}} = \begin{bmatrix} \frac{1}{1+\lambda} \\ \frac{0.01}{0.01+\lambda} \end{bmatrix} + \begin{bmatrix} \frac{1}{1+\lambda} & 0 \\ 0 & \frac{1}{0.01+\lambda} \end{bmatrix} X \tilde{z}_{\mathsf{train}}$$

Distribution?

Ridge-regression coefficient estimate

$$\tilde{\beta}_{\mathsf{RR}} = \begin{bmatrix} \frac{1}{1+\lambda} \\ \frac{0.01}{0.01+\lambda} \end{bmatrix} + \begin{bmatrix} \frac{1}{1+\lambda} & 0 \\ 0 & \frac{1}{0.01+\lambda} \end{bmatrix} X \tilde{z}_{\mathsf{train}}$$

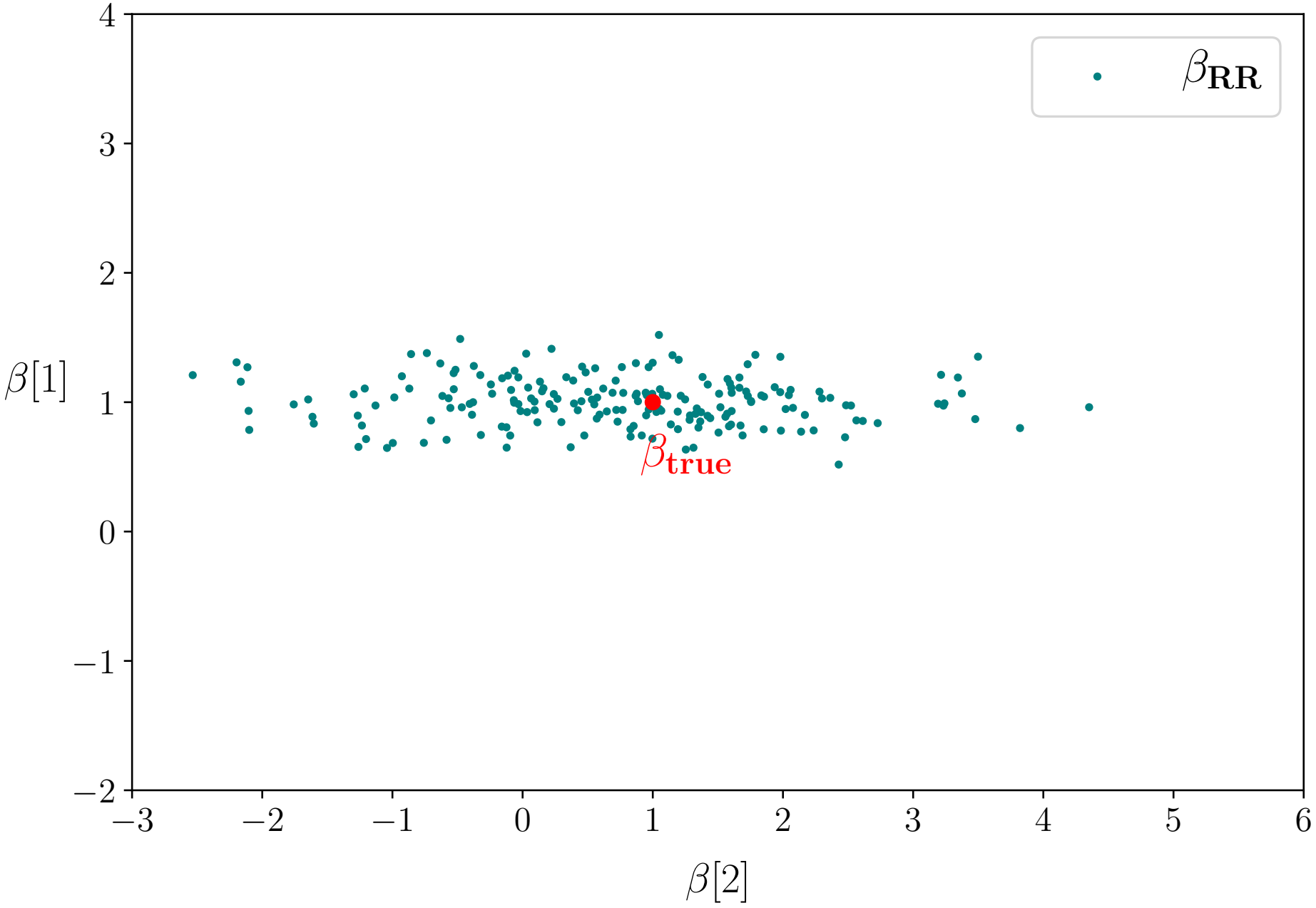
Distribution? Gaussian with mean

$$\beta_{\mathsf{bias}} := \begin{bmatrix} \frac{1}{1+\lambda} \\ \frac{0.01}{0.01+\lambda} \end{bmatrix}$$

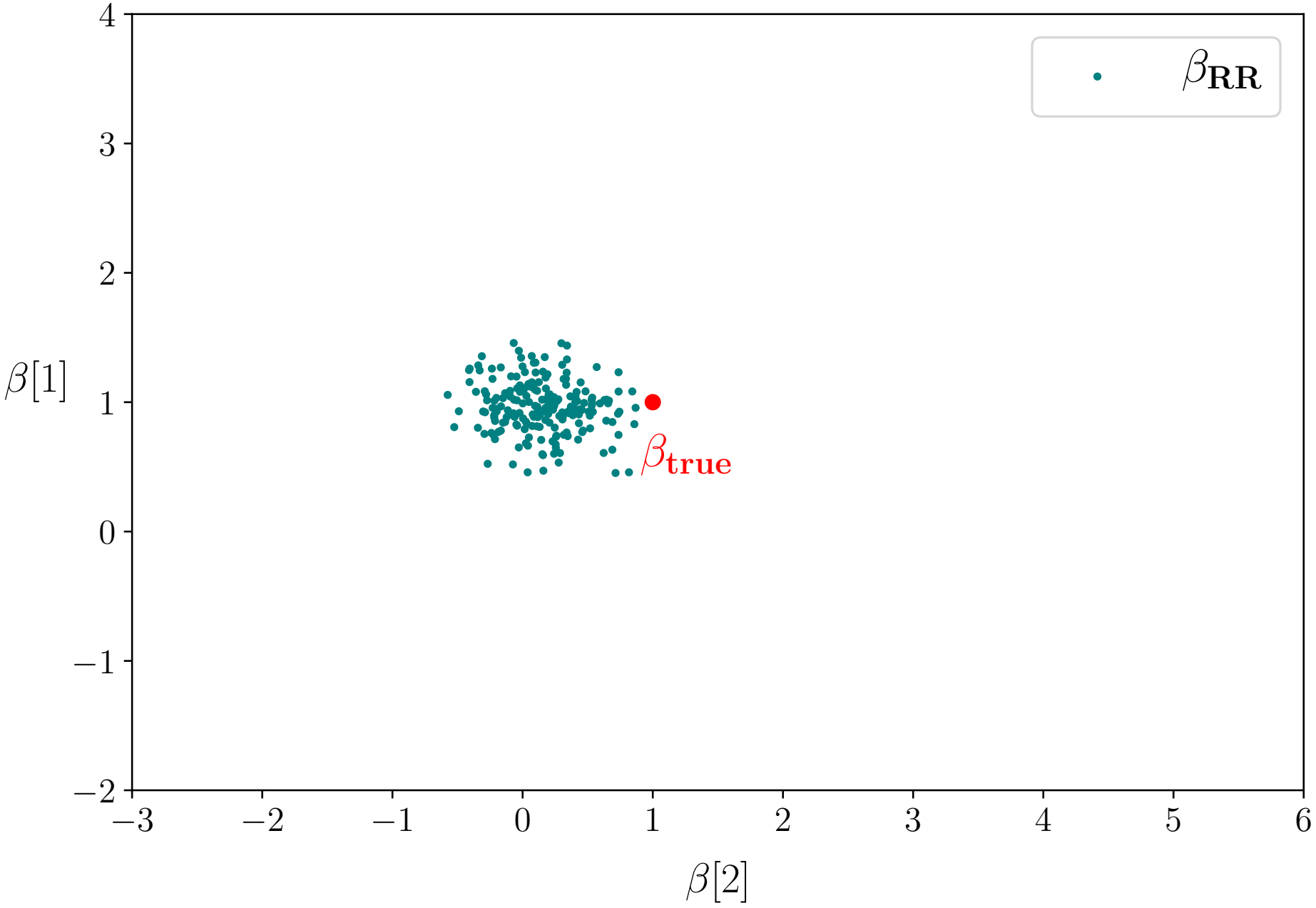
and covariance matrix

$$\Sigma_{\mathsf{RR}} := 0.04 \begin{bmatrix} \frac{1}{(1+\lambda)^2} & 0 \\ 0 & \frac{0.01}{(0.01+\lambda)^2} \end{bmatrix}$$

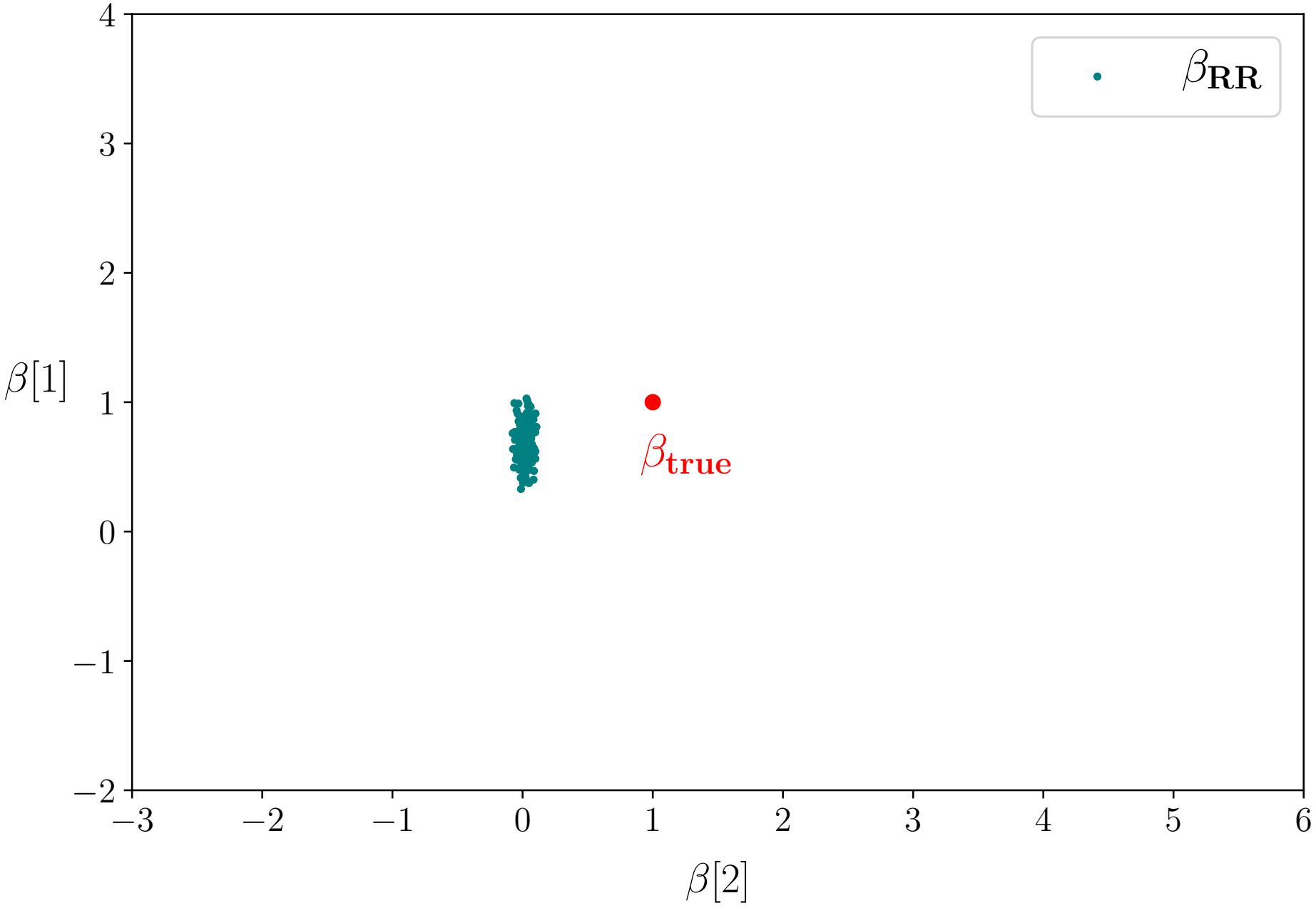
$\lambda = 0.005$


$\beta_{\mathbf{RR}}$
$\beta_{\mathbf{true}}$
$\beta[1]$
$\beta[2]$
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4
5
6

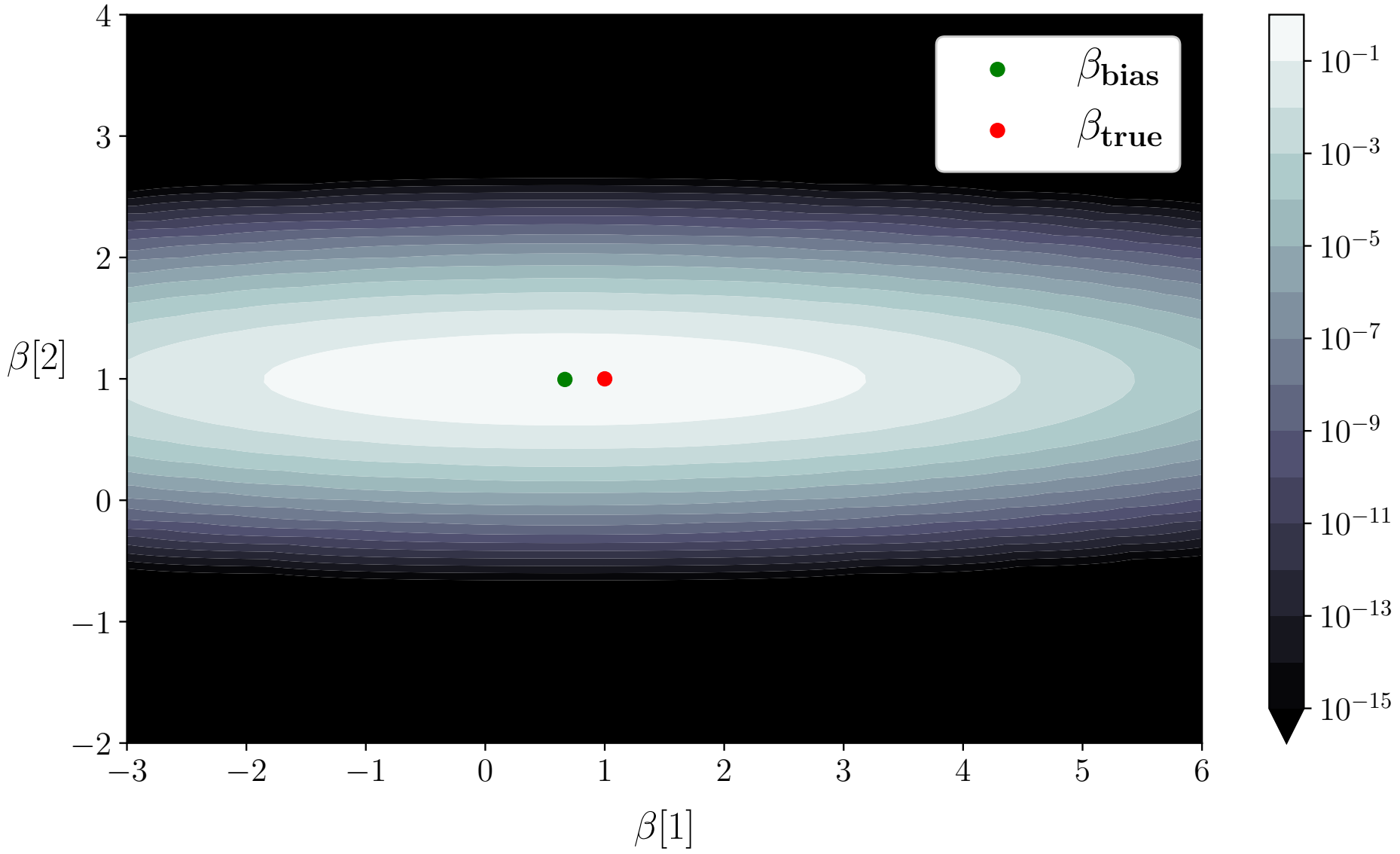
$\lambda = 0.05$


$\beta_{\mathbf{RR}}$
$\beta_{\mathbf{true}}$
$\beta[1]$
$\beta[2]$
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4
5
6

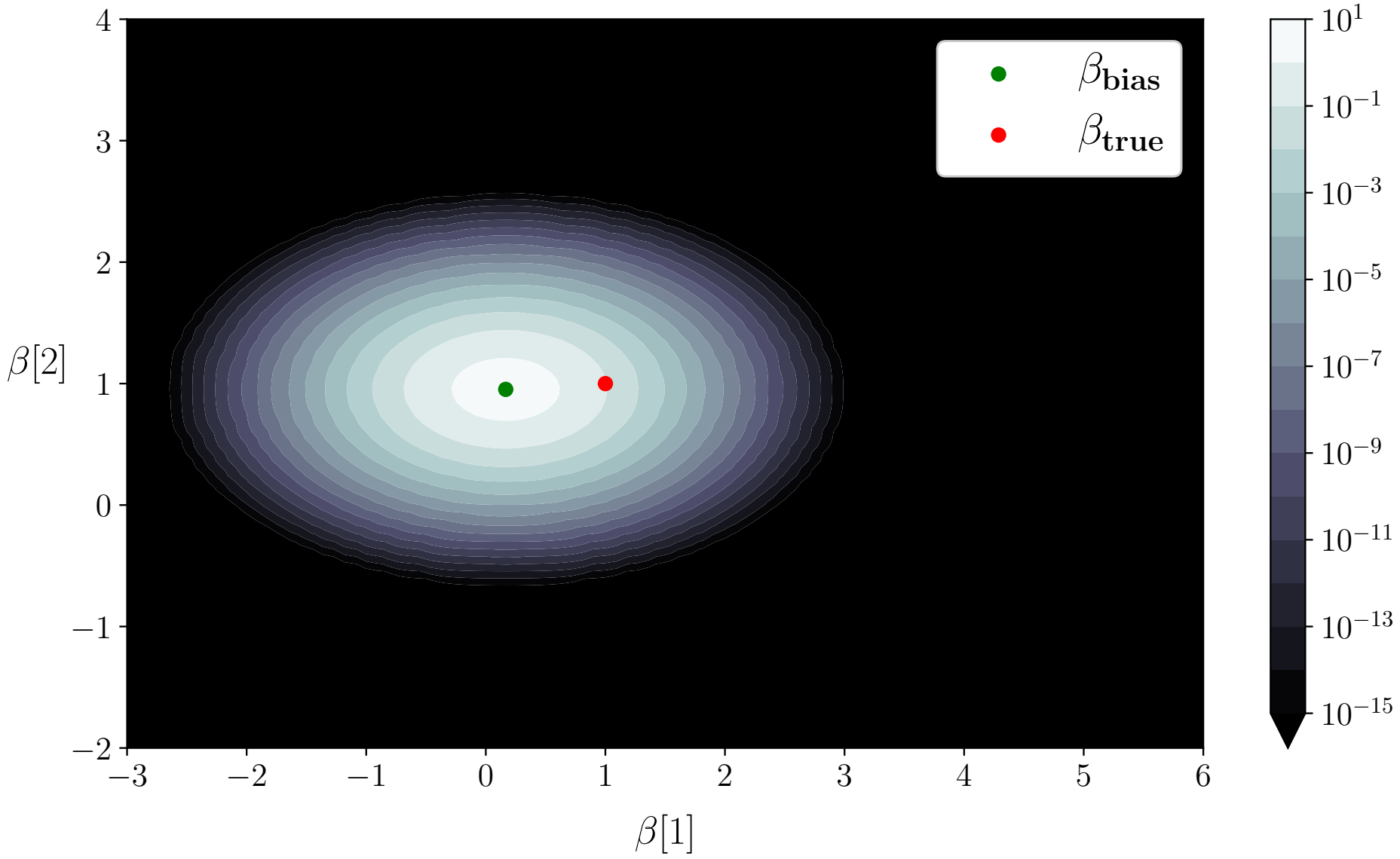
$\lambda = 0.5$


$\beta_{\mathbf{RR}}$
$\beta_{\mathbf{true}}$
$\beta[1]$
$\beta[2]$
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4
5
6

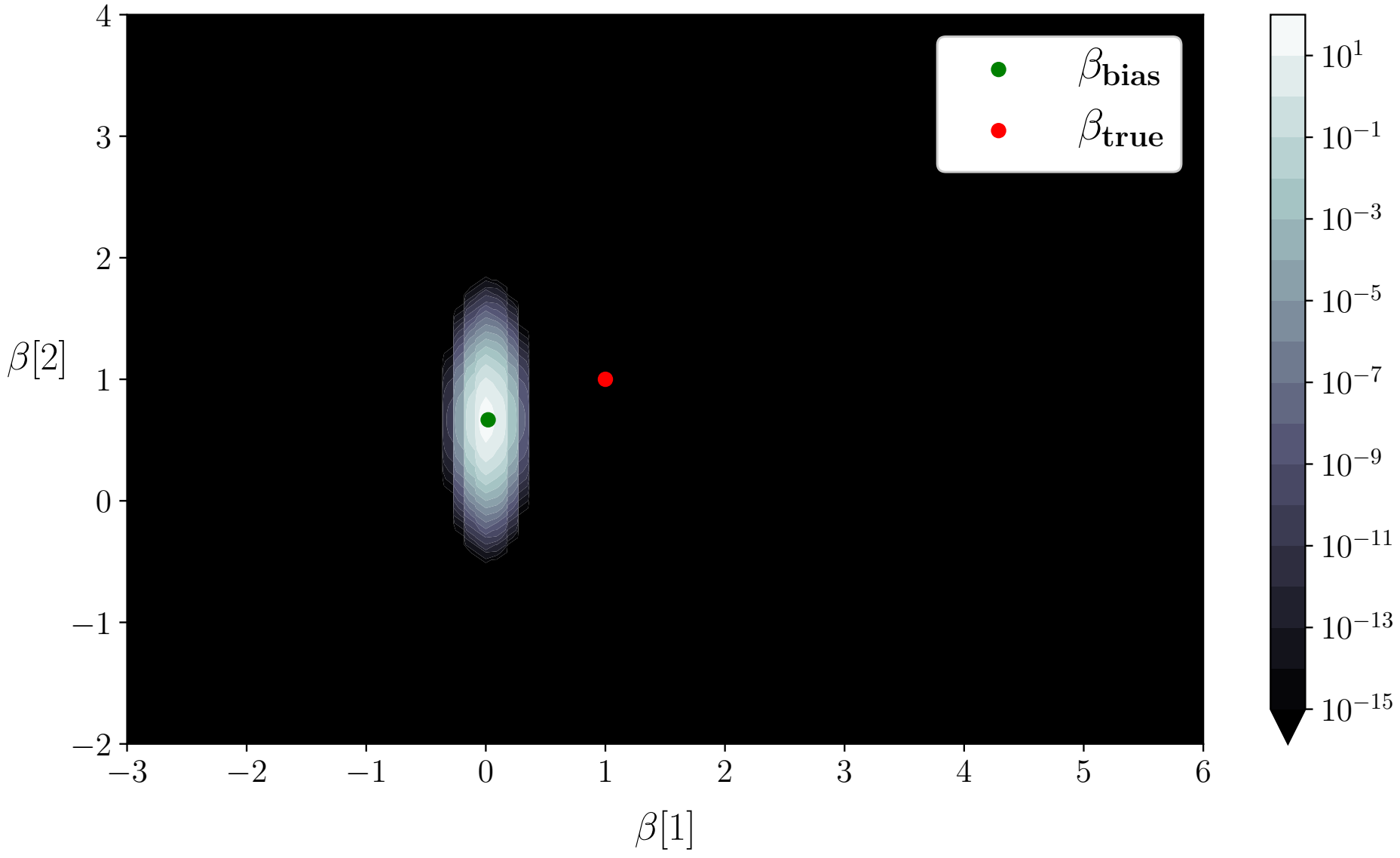
$\lambda = 0.005$


$\beta_{\mathbf{bias}}$
$\beta_{\mathbf{true}}$
$\beta[2]$
$\beta[1]$
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4
5
6
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
10^{-11}
10^{-13}
10^{-15}

$\lambda = 0.05$


$\beta_{\mathbf{bias}}$
$\beta_{\mathbf{true}}$
$\beta[2]$
$\beta[1]$
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4
5
6
10^{1}
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
10^{-11}
10^{-13}
10^{-15}

$\lambda = 0.5$


$\beta_{\mathbf{bias}}$
$\beta_{\mathbf{true}}$
$\beta[2]$
$\beta[1]$
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4
5
6
10^{1}
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
10^{-11}
10^{-13}
10^{-15}

Ordinary least squares

Ridge regression

Early stopping

Initialization at origin with constant step size

$$\beta^{(k+1)} = \beta^{(k)} - \alpha X \left(X^T \beta^{(k)} - y \right)$$

$$= \left(I - \alpha X X^T \right) \beta^{(k)} + \alpha X \tilde{y}$$

$$\beta^{(0)} =$$

$$\beta^{(1)} =$$

$$\beta^{(2)} =$$

$$\beta^{(3)} =$$

$$\beta^{(k+1)} =$$

Initialization at origin with constant step size

$$
\begin{aligned}
\beta^{(k+1)} &= \beta^{(k)} - \alpha X \left(X^T \beta^{(k)} - y \right) \\
&= \left(I - \alpha X X^T \right) \beta^{(k)} + \alpha X \tilde{y} \\
\beta^{(0)} &= 0 \\
\beta^{(1)} &= \alpha X y \\
\beta^{(2)} &= \left(I - \alpha X X^T \right) \beta^{(1)} + \alpha X y \\
&= \left(I - \alpha X X^T \right) \alpha X y + \alpha X y \\
\beta^{(3)} &= \left(I - \alpha X X^T \right) \beta^{(2)} + \alpha X y \\
&= \left(I - \alpha X X^T \right)^2 \alpha X y + \left(I - \alpha X X^T \right) \alpha X y + \alpha X y \\
\beta^{(k+1)} &= \sum_{i=0}^{k} \left(I - \alpha X X^T \right)^i \alpha X y
\end{aligned}
$$

Gradient descent iterates, starting at origin

$$\beta^{(k+1)} = \sum_{i=0}^{k} \left(I - \alpha X X^T\right)^i \alpha X \tilde{y}_{\text{train}} \qquad \left(\sum_{i=0}^{k} r^i = \frac{1 - r^{k+1}}{1 - r}\right)$$

Gradient descent iterates, starting at origin

$$\beta^{(k+1)} = \sum_{i=0}^{k} \left(I - \alpha X X^T\right)^i \alpha X \tilde{y}_{\text{train}} \qquad \left(\sum_{i=0}^{k} r^i = \frac{1 - r^{k+1}}{1 - r}\right)$$

$$= \begin{bmatrix} \sum_{i=0}^{k} (1-\alpha)^i & 0 \\ 0 & \sum_{i=0}^{k} (1 - 0.01\alpha)^i \end{bmatrix} \alpha X (X^T \beta_{\text{true}} + \tilde{z}_{\text{train}})$$

$$= \begin{bmatrix} 1 - (1-\alpha)^{k+1} & 0 \\ 0 & 100\left(1 - (1 - 0.01\alpha)^{k+1}\right) \end{bmatrix} X \tilde{y}_{\text{train}}$$

Gradient descent iterates, starting at origin

$$\beta^{(k+1)} = \begin{bmatrix} 1-(1-\alpha)^{k+1} & 0 \\ 0 & 100\left(1-(1-0.01\alpha)^{k+1}\right) \end{bmatrix} X(X^T\beta_{\text{true}} + \tilde{z}_{\text{train}})$$

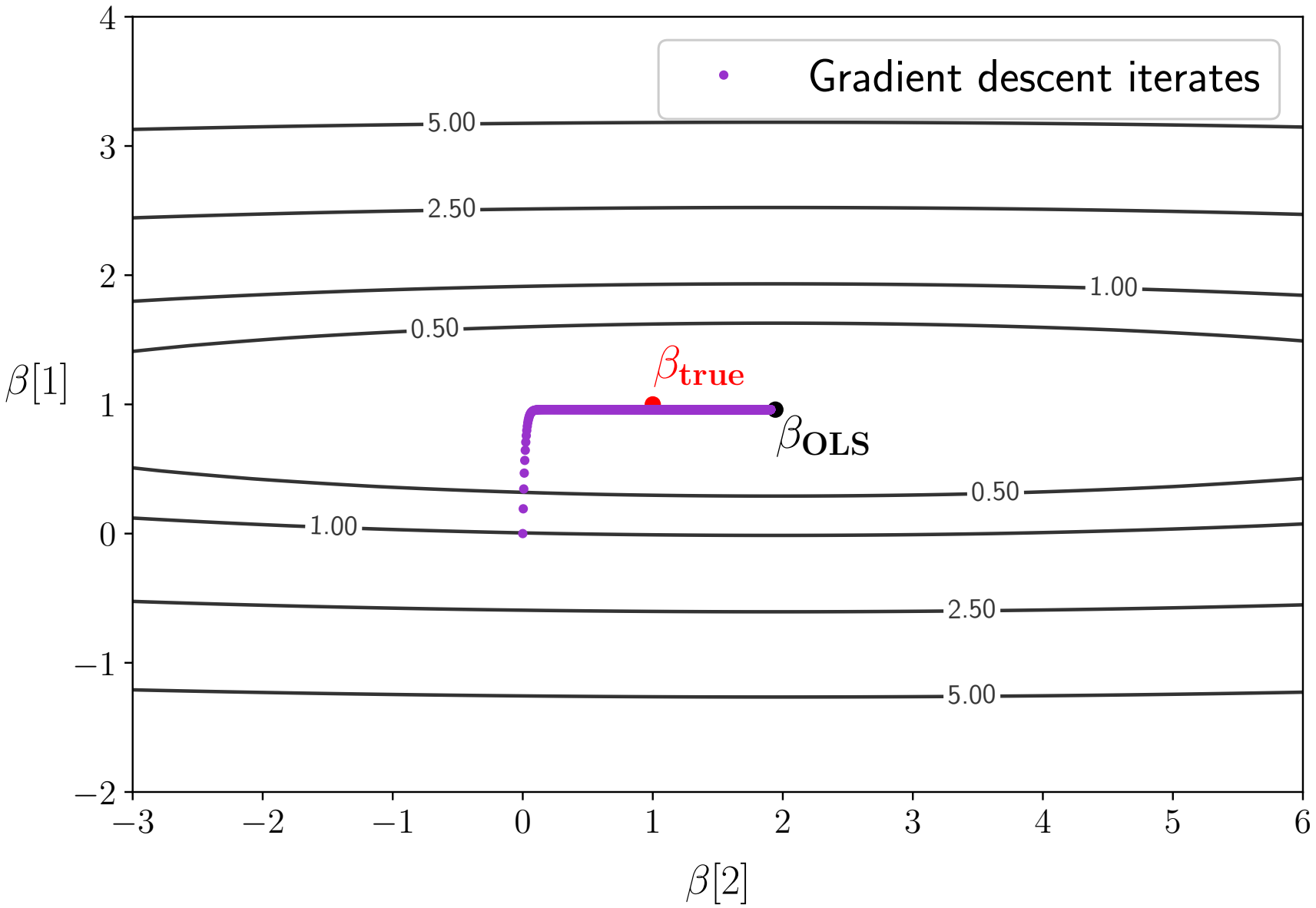
Gradient descent iterates, starting at origin

$$\beta^{(k+1)} = \begin{bmatrix} 1-(1-\alpha)^{k+1} & 0 \\ 0 & 100\left(1-(1-0.01\alpha)^{k+1}\right) \end{bmatrix} X(X^T\beta_{\text{true}} + \tilde{z}_{\text{train}})$$

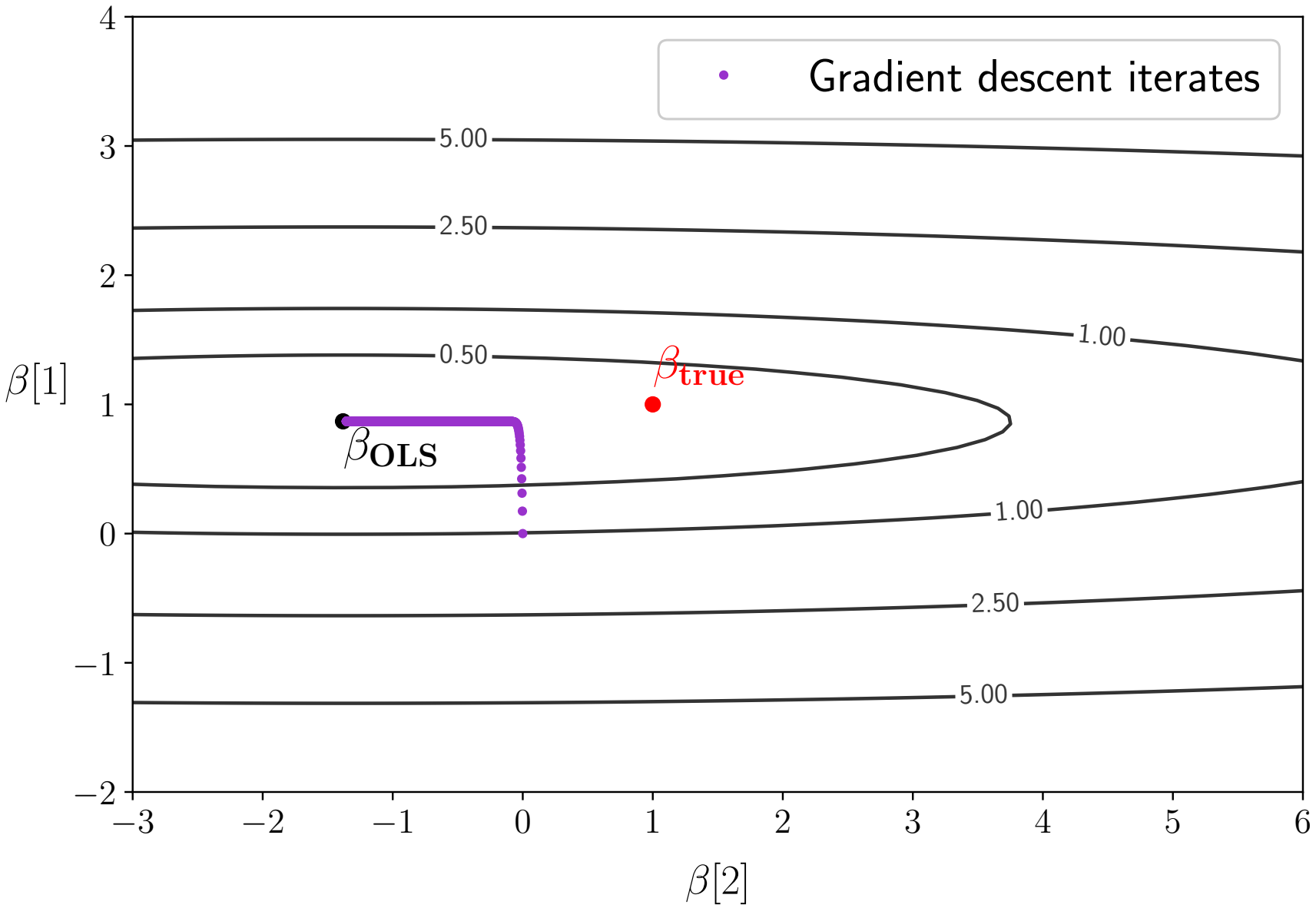
$$= \begin{bmatrix} 1-(1-\alpha)^{k+1} \\ 1-(1-0.01\alpha)^{k+1} \end{bmatrix}$$

$$+ \begin{bmatrix} 1-(1-\alpha)^{k+1} & 0 \\ 0 & 100\left(1-(1-0.01\alpha)^{k+1}\right) \end{bmatrix} X\tilde{z}_{\text{train}}$$

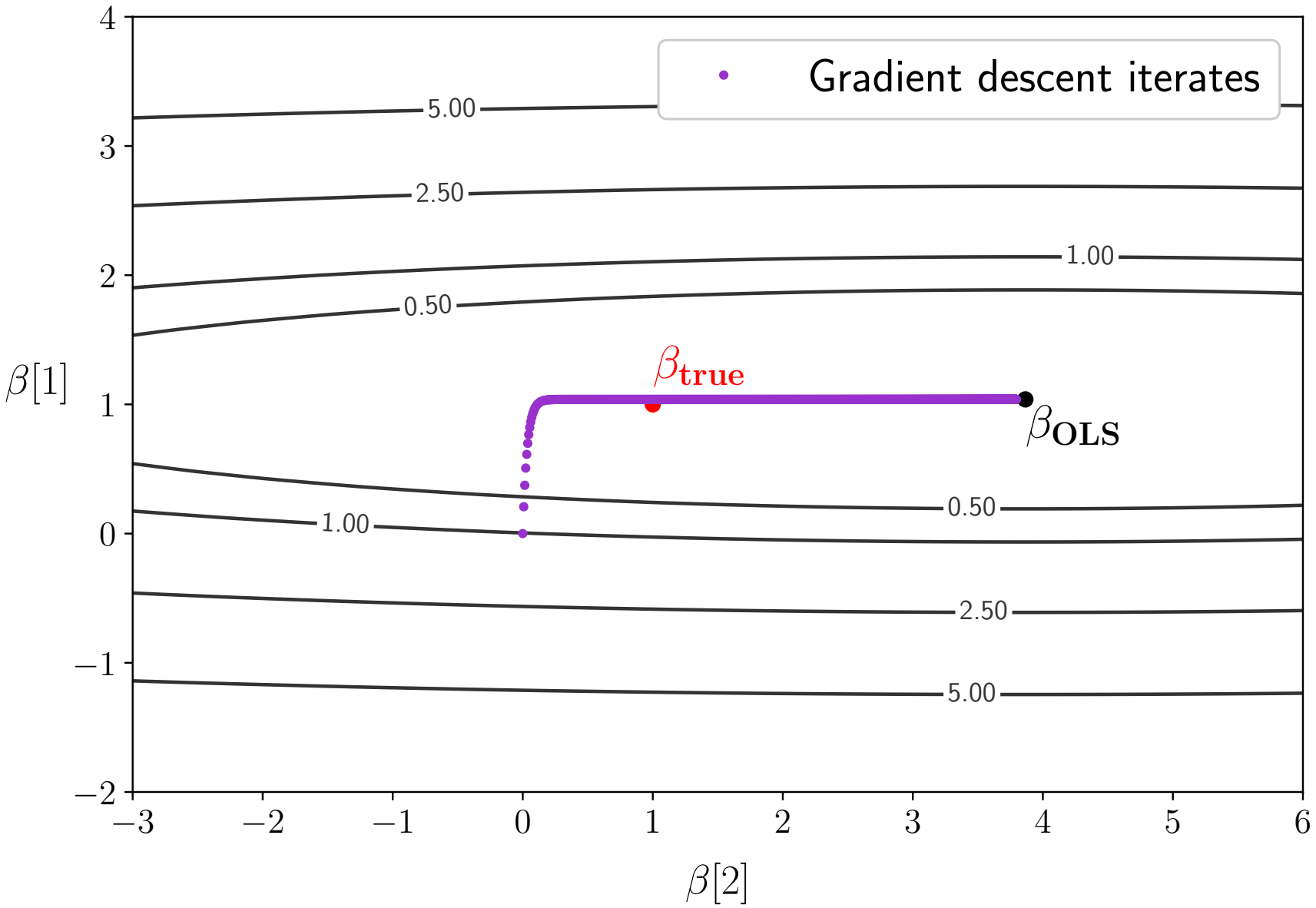
Additive model ($s_1 = 1$, $s_2 = 0.1$)



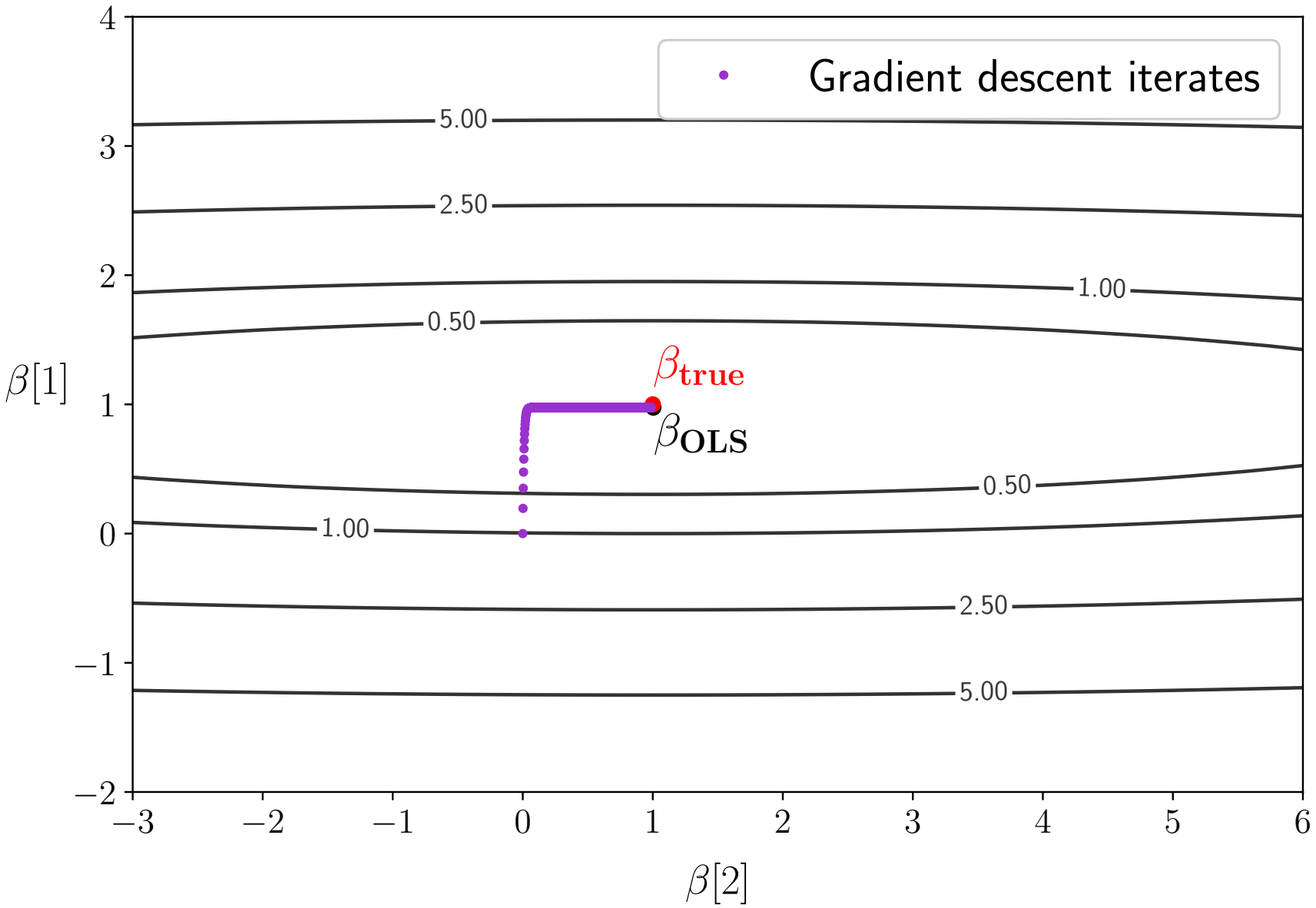
Additive model ($s_1 = 1$, $s_2 = 0.1$)



Additive model ($s_1 = 1$, $s_2 = 0.1$)



Additive model ($s_1 = 1$, $s_2 = 0.1$)



Gradient descent coefficient estimate

At iteration k

$$\tilde{\beta}_{\mathsf{GD}} = \begin{bmatrix} 1-(1-\alpha)^{k+1} \\ 1-(1-0.01\alpha)^{k+1} \end{bmatrix} + \begin{bmatrix} 1-(1-\alpha)^{k} & 0 \\ 0 & 100\left(1-(1-0.01\alpha)^{k}\right) \end{bmatrix} X\tilde{z}_{\mathsf{train}}$$

Distribution?

Gradient descent coefficient estimate

At iteration k

$$\tilde{\beta}_{\text{GD}} = \begin{bmatrix} 1-(1-\alpha)^{k+1} \\ 1-(1-0.01\alpha)^{k+1} \end{bmatrix} + \begin{bmatrix} 1-(1-\alpha)^{k} & 0 \\ 0 & 100\left(1-(1-0.01\alpha)^{k}\right) \end{bmatrix} X\tilde{z}_{\text{train}}$$

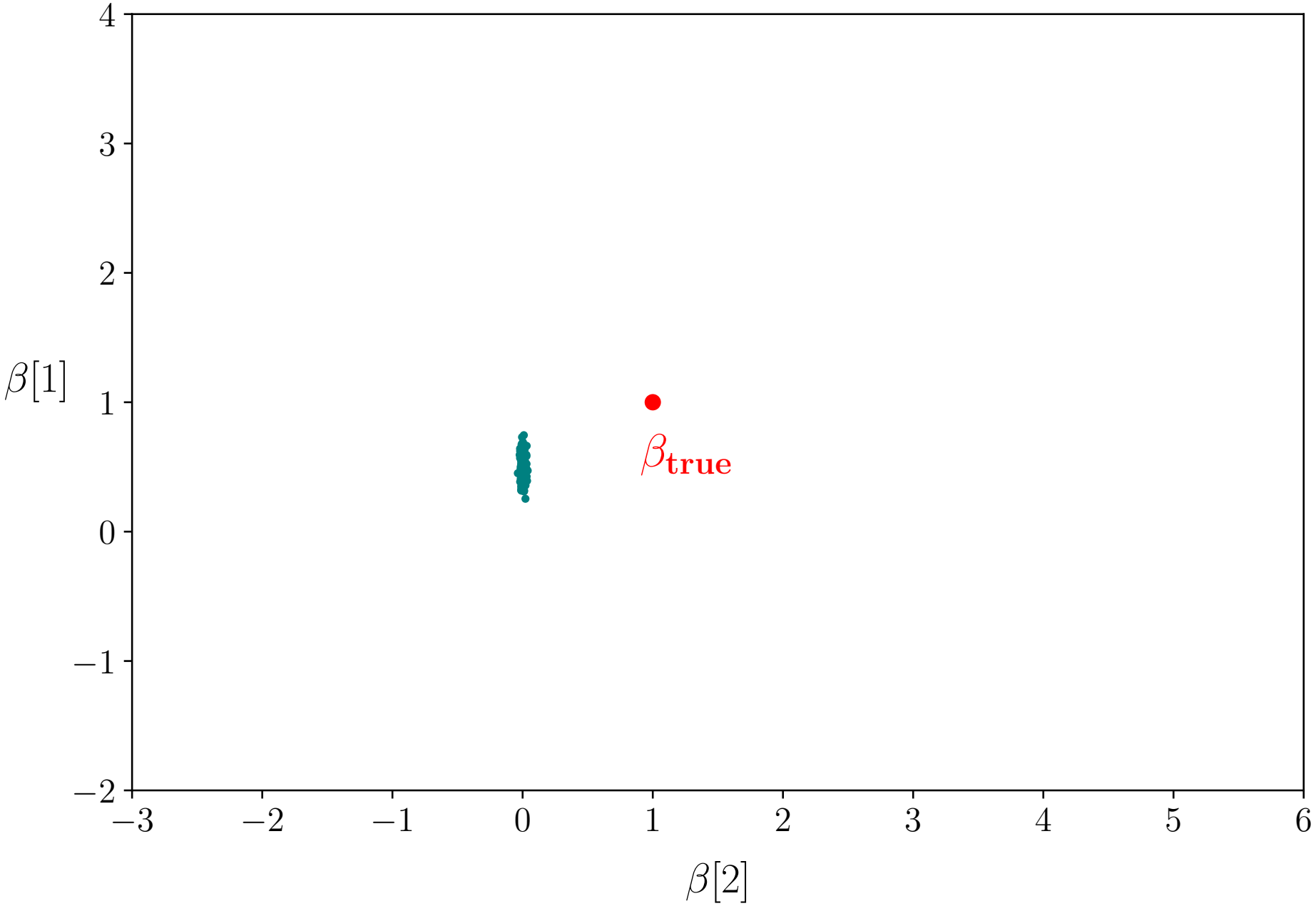
Distribution? Gaussian with mean

$$\beta_{\text{bias}} := \begin{bmatrix} 1-(1-\alpha)^{k} & 0 \\ 0 & 1-(1-0.01\alpha)^{k} \end{bmatrix} \beta_{\text{true}}$$

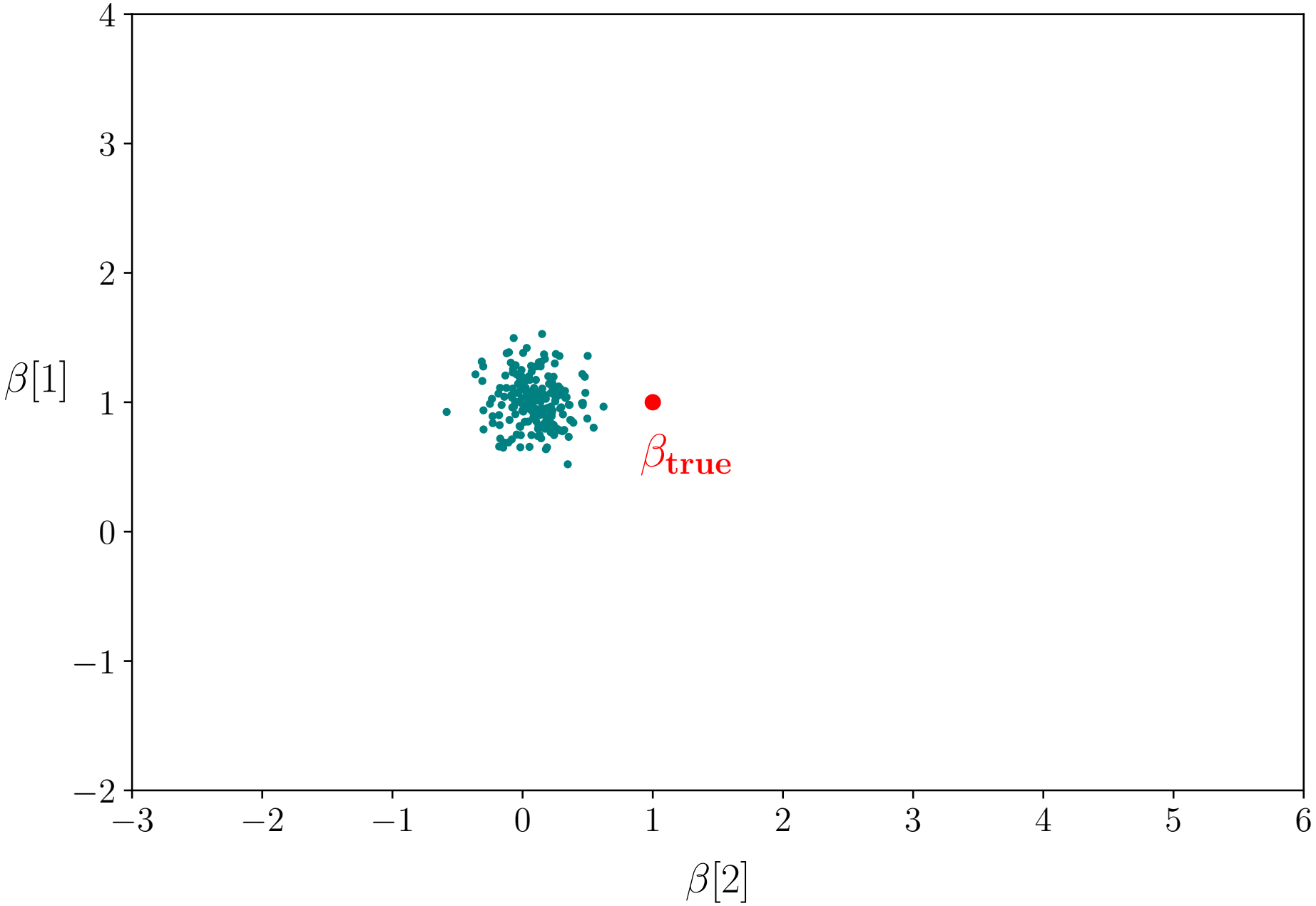
and covariance matrix

$$\Sigma_{\text{GD}} := \begin{bmatrix} 0.04\left(1-(1-\alpha)^{k}\right)^2 & 0 \\ 0 & 4\left(1-(1-0.01\alpha)^{k}\right)^2 \end{bmatrix} X\tilde{z}_{\text{train}}$$

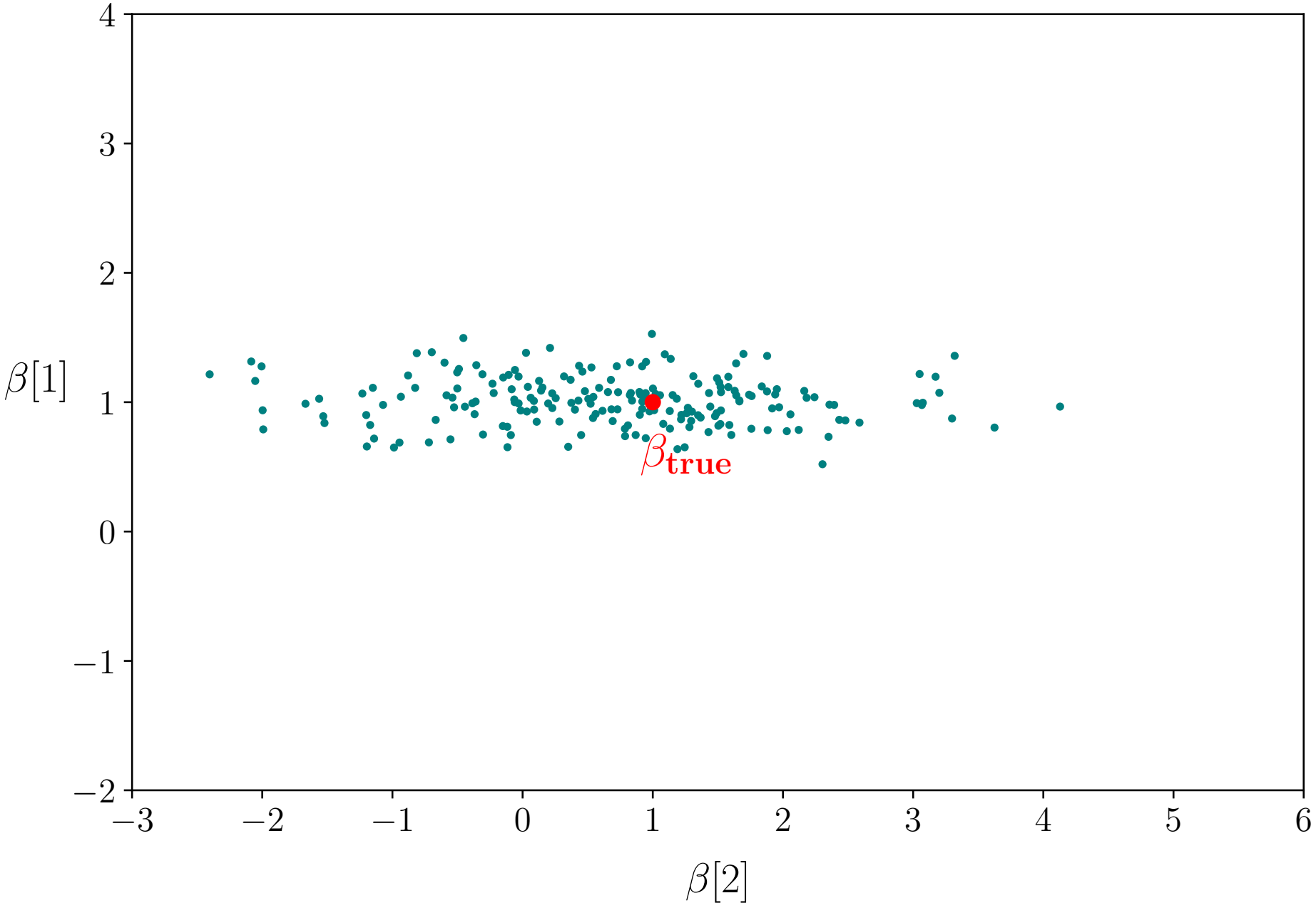
$k = 3$



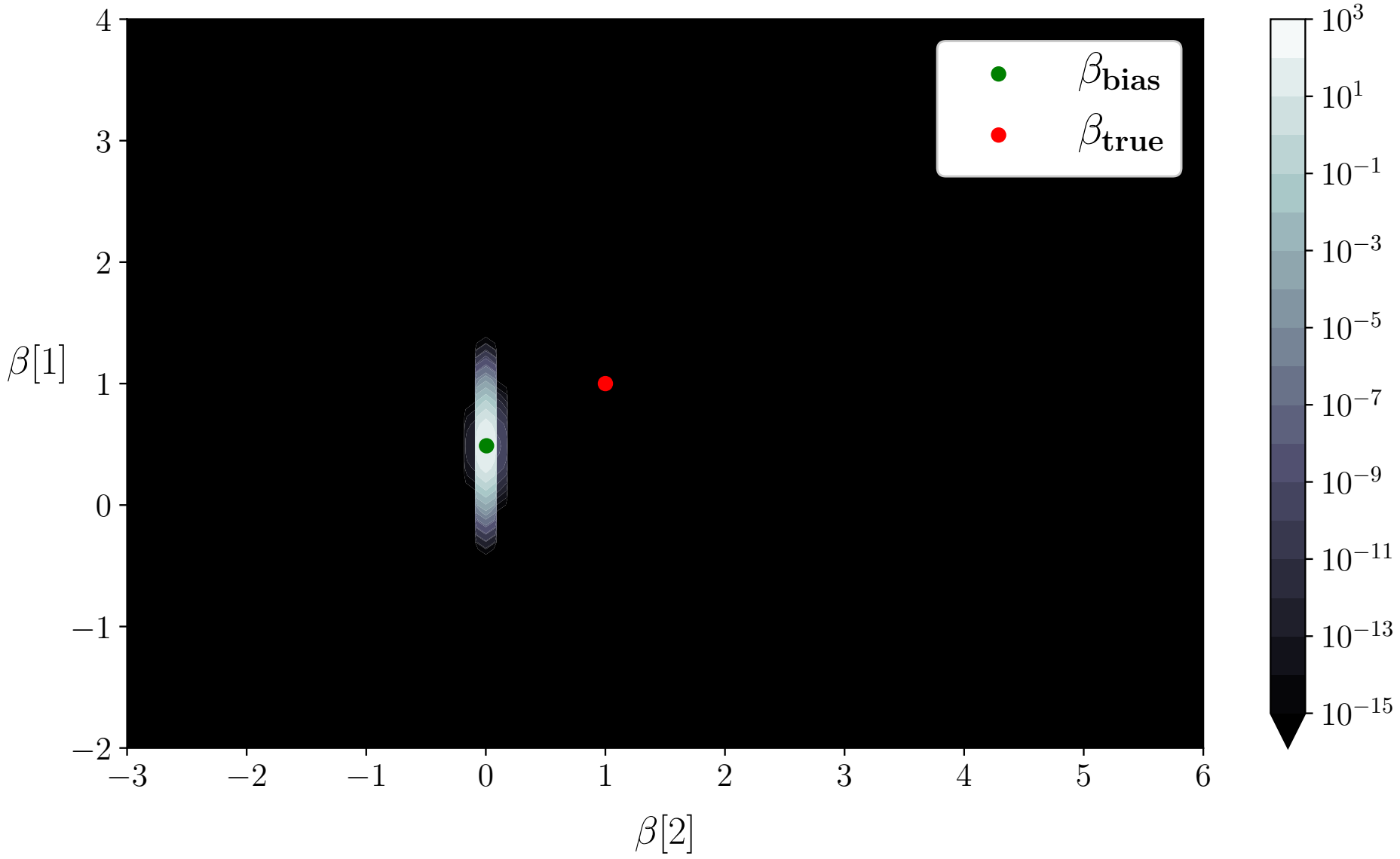
$k = 50$



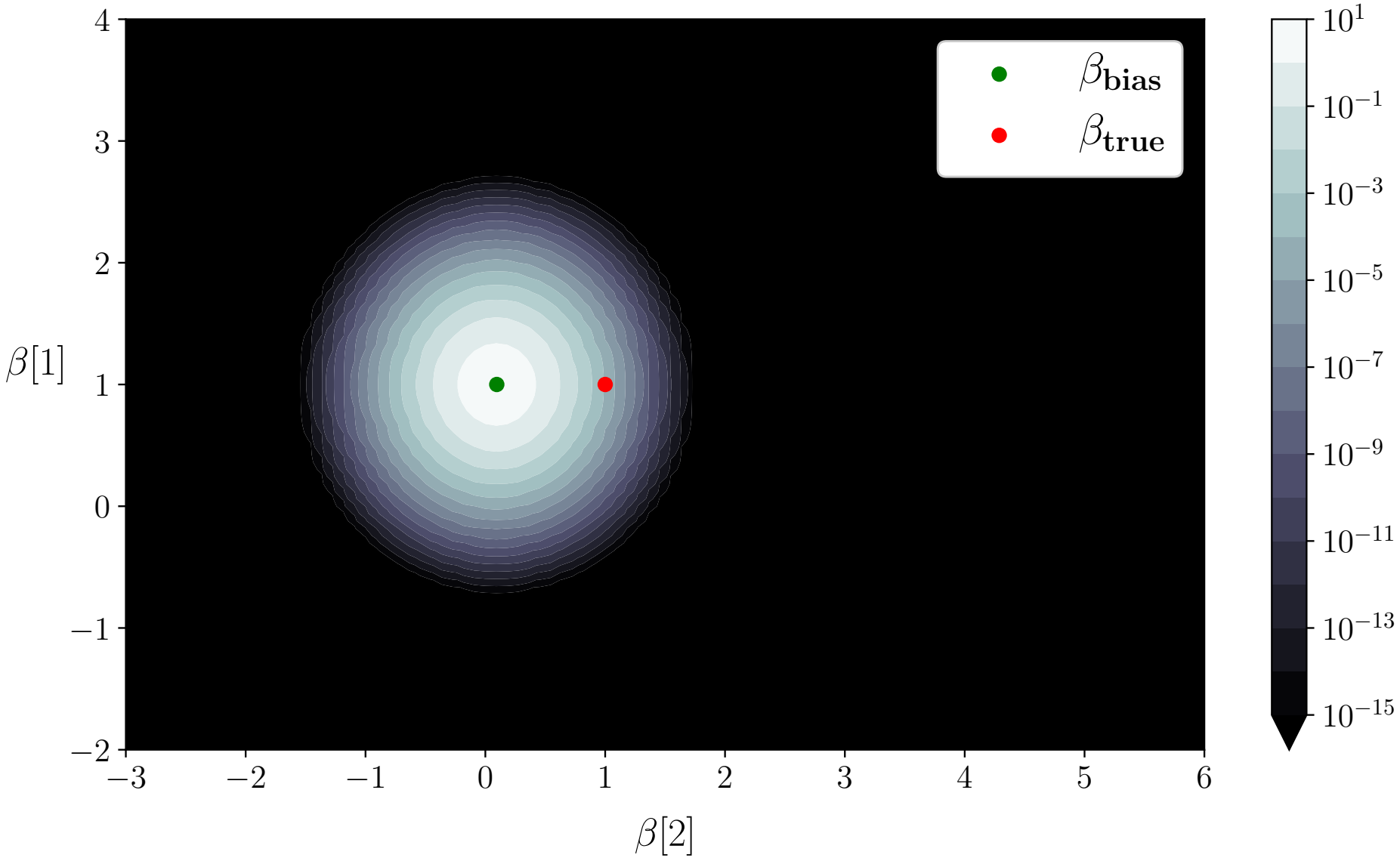
$k = 500$



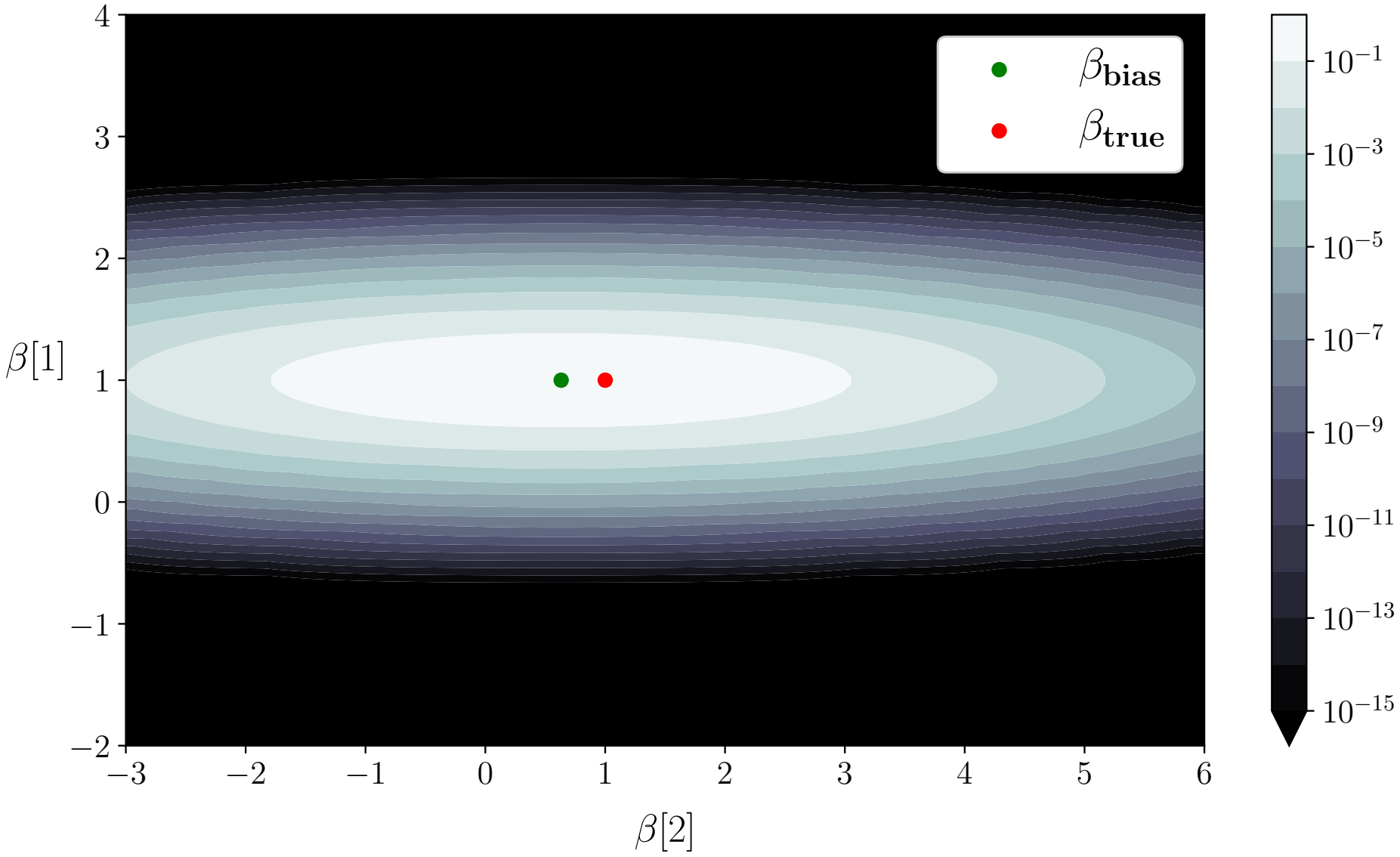
$k = 3$


$\beta_{\mathbf{bias}}$
$\beta_{\mathbf{true}}$
$\beta[1]$
$\beta[2]$
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4
5
6
10^{3}
10^{1}
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
10^{-11}
10^{-13}
10^{-15}

$k = 50$


$\beta_{\mathbf{bias}}$
$\beta_{\mathbf{true}}$
$\beta[1]$
$\beta[2]$
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4
5
6
10^{1}
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
10^{-11}
10^{-13}
10^{-15}

$k = 500$


$\beta_{\mathbf{bias}}$
$\beta_{\mathbf{true}}$
$\beta[1]$
$\beta[2]$
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
4
5
6
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
10^{-11}
10^{-13}
10^{-15}