Convex Optimization

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

https://cims.nyu.edu/~cfgranda/pages/MTDS_spring19/index.html

Carlos Fernandez-Granda

Motivating applications

Convex functions

Optimality conditions

Convex Regularization

Optimization algorithms

Sparse regression

Linear regression is challenging when the number of features p is large

Problem: How to choose subset of features $\mathcal{I} \subset \{1, \ldots, p\}$, such that

$$y \approx \sum_{i \in \mathcal{I}} \vec{\beta}[i] \vec{x}[i] + \beta_0$$

Equivalently, find sparse coefficient vector $\vec{\beta} \in \mathbb{R}^p$ such that

$$y \approx \left\langle \vec{x}, \vec{\beta} \right\rangle + \beta_0$$

Collaborative filtering

Quantity $y[i, j]$ depends on indices i and j

We observe examples and want to predict new instances

For example, $y[i, j]$ is rating given to a movie i by a user j

Collaborative filtering

Rank-r bilinear model

Certain people like certain movies: r factors

$$y[i,j] \approx \sum_{l=1}^{r} a_l[i] b_l[j]$$

For each factor l

- $a_l[i]$: movie i is positively (> 0), negatively (< 0) or not (≈ 0) associated to factor l
- $b_l[j]$: user j is positively (> 0), negatively (< 0) or not (≈ 0) associated to factor l

SVD can be used to fit the model if all entries are observed

Motivating applications

Convex functions

Optimality conditions

Convex Regularization

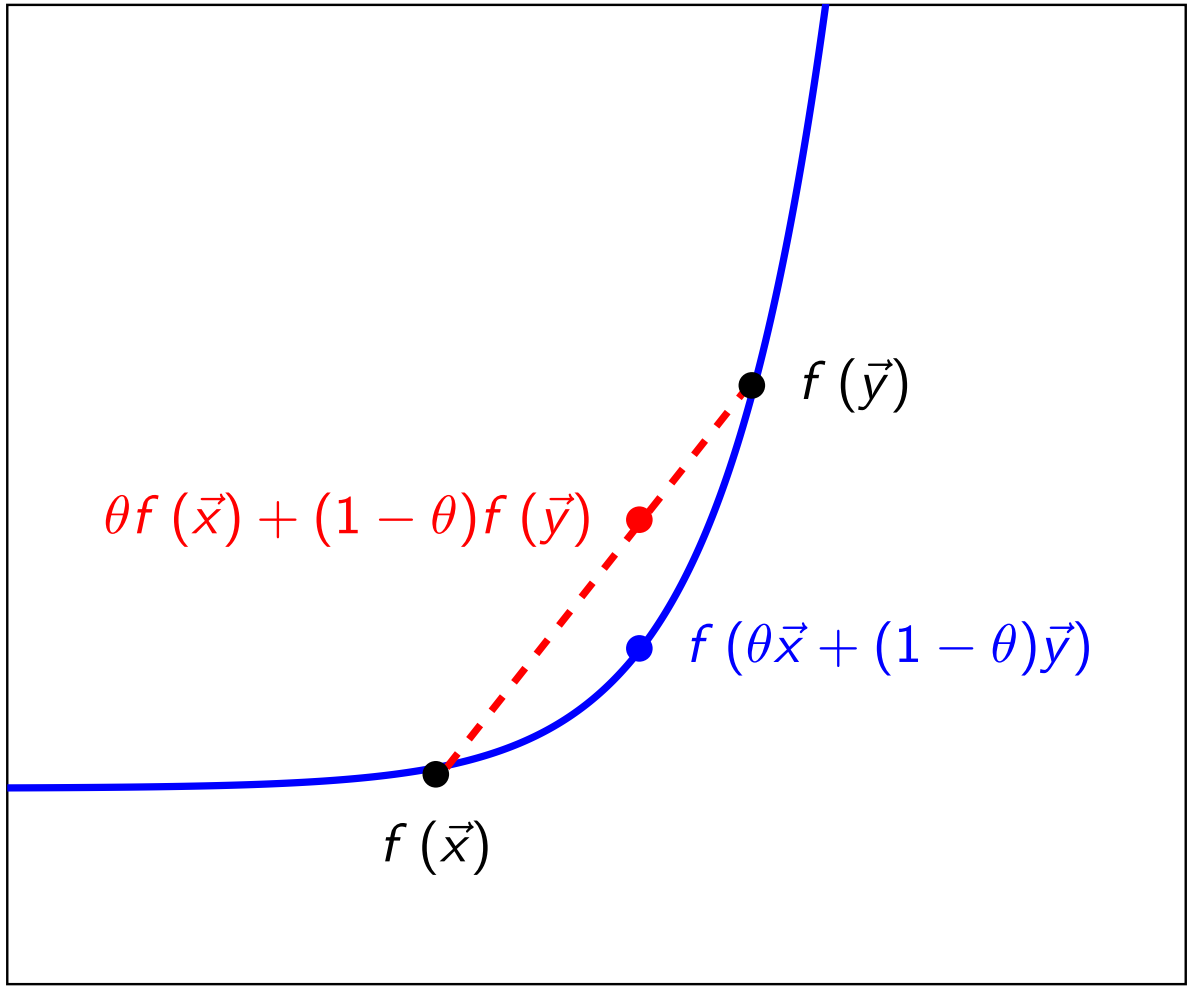
Optimization algorithms

Convex functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for any $\vec{x}, \vec{y} \in \mathbb{R}^n$ and any $\theta \in (0, 1)$

$$\theta f(\vec{x}) + (1 - \theta) f(\vec{y}) \geq f(\theta \vec{x} + (1 - \theta) \vec{y})$$

Convex functions



Linear functions are convex

If f is linear

$$f\left(\theta\vec{x} + (1-\theta)\,\vec{y}\right)$$

Linear functions are convex

If f is linear

$$f\left(\theta\vec{x}+\left(1-\theta\right)\vec{y}\right)=\theta f\left(\vec{x}\right)+\left(1-\theta\right)f\left(\vec{y}\right)$$

Strictly convex functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is strictly convex if for any $\vec{x}, \vec{y} \in \mathbb{R}^n$ and any $\theta \in (0, 1)$

$$\theta f(\vec{x}) + (1 - \theta) f(\vec{y}) > f(\theta \vec{x} + (1 - \theta) \vec{y})$$

Local minima are global

Any local minimum of a convex function is also a global minimum

Proof

Let $\vec{x}_{\text{loc}}$ be a local minimum: for all $\vec{x} \in \mathbb{R}^n$ such that $||\vec{x} - \vec{x}_{\text{loc}}||_2 \leq \gamma$

$$f\left(\vec{x}_{\text{loc}}\right) \leq f\left(\vec{x}\right)$$

Let $\vec{x}_{\text{glob}}$ be a global minimum

$$f\left(\vec{x}_{\text{glob}}\right) < f\left(\vec{x}_{\text{loc}}\right)$$

Proof

Choose θ so that $\vec{x}_\theta := \theta \vec{x}_{\text{loc}} + (1-\theta)\, \vec{x}_{\text{glob}}$ satisfies

$$||\vec{x}_\theta - \vec{x}_{\text{loc}}||_2 \leq \gamma$$

then

$$f\left(\vec{x}_{\text{loc}}\right) \leq f\left(\vec{x}_\theta\right)$$

Proof

Choose θ so that $\vec{x}_\theta := \theta\vec{x}_{\text{loc}} + (1-\theta)\,\vec{x}_{\text{glob}}$ satisfies

$$||\vec{x}_\theta - \vec{x}_{\text{loc}}||_2 \leq \gamma$$

then

$$\begin{aligned} f\left(\vec{x}_{\text{loc}}\right) &\leq f\left(\vec{x}_\theta\right) \\ &= f\left(\theta\vec{x}_{\text{loc}} + (1-\theta)\,\vec{x}_{\text{glob}}\right) \end{aligned}$$

Proof

Choose θ so that $\vec{x}_\theta := \theta\vec{x}_{\text{loc}} + (1-\theta)\,\vec{x}_{\text{glob}}$ satisfies

$$||\vec{x}_\theta - \vec{x}_{\text{loc}}||_2 \leq \gamma$$

then

$$\begin{aligned} f\left(\vec{x}_{\text{loc}}\right) &\leq f\left(\vec{x}_\theta\right) \\ &= f\left(\theta\vec{x}_{\text{loc}} + (1-\theta)\,\vec{x}_{\text{glob}}\right) \\ &\leq \theta f\left(\vec{x}_{\text{loc}}\right) + (1-\theta)\,f\left(\vec{x}_{\text{glob}}\right) \quad \text{by convexity of } f \end{aligned}$$

Proof

Choose θ so that $\vec{x}_\theta := \theta\vec{x}_{\text{loc}} + (1-\theta)\,\vec{x}_{\text{glob}}$ satisfies

$$||\vec{x}_\theta - \vec{x}_{\text{loc}}||_2 \leq \gamma$$

then

$$\begin{aligned} f\left(\vec{x}_{\text{loc}}\right) &\leq f\left(\vec{x}_\theta\right) \\ &= f\left(\theta\vec{x}_{\text{loc}} + (1-\theta)\,\vec{x}_{\text{glob}}\right) \\ &\leq \theta f\left(\vec{x}_{\text{loc}}\right) + (1-\theta)\,f\left(\vec{x}_{\text{glob}}\right) \quad \text{by convexity of } f \\ &< f\left(\vec{x}_{\text{loc}}\right) \quad \text{because } f\left(\vec{x}_{\text{glob}}\right) < f\left(\vec{x}_{\text{loc}}\right) \end{aligned}$$

Strictly convex functions

Strictly convex functions have at most one global minimum

Strictly convex functions

Strictly convex functions have at most one global minimum

Proof: Assume two minima exist at $\vec{x} \neq \vec{y}$ with value v_{min}

Strictly convex functions

Strictly convex functions have at most one global minimum

Proof: Assume two minima exist at $\vec{x} \neq \vec{y}$ with value v_{min}

$$f\left(0.5\vec{x} + 0.5\vec{y}\right) < 0.5f(\vec{x}) + 0.5f(\vec{y})$$

Strictly convex functions

Strictly convex functions have at most one global minimum

Proof: Assume two minima exist at $\vec{x} \neq \vec{y}$ with value v_{min}

$$f\left(0.5\vec{x} + 0.5\vec{y}\right) < 0.5f(\vec{x}) + 0.5f(\vec{y})$$
$$= v_{\text{min}}$$

Norm

Let $\mathcal{V}$ be a vector space, a norm is a function $||\cdot||$ from $\mathcal{V}$ to $\mathbb{R}$ with the following properties

- It is homogeneous. For any scalar α and any $\vec{x} \in \mathcal{V}$

$$||\alpha\,\vec{x}|| = |\alpha|\,||\vec{x}||\,.$$

- It satisfies the triangle inequality

$$||\vec{x} + \vec{y}|| \leq ||\vec{x}|| + ||\vec{y}||\,.$$

  In particular, $||\vec{x}|| \geq 0$

- $||\vec{x}|| = 0$ implies $\vec{x} = \vec{0}$

Norms are convex

For any $\vec{x}, \vec{y} \in \mathbb{R}^n$ and any $\theta \in (0, 1)$

$$||\theta\vec{x} + (1 - \theta)\,\vec{y}||$$

Norms are convex

For any $\vec{x}, \vec{y} \in \mathbb{R}^n$ and any $\theta \in (0, 1)$

$$||\theta\vec{x} + (1-\theta)\,\vec{y}|| \leq ||\theta\vec{x}|| + ||(1-\theta)\,\vec{y}||$$

Norms are convex

For any $\vec{x}, \vec{y} \in \mathbb{R}^n$ and any $\theta \in (0, 1)$

$$\begin{aligned} ||\theta\vec{x} + (1-\theta)\,\vec{y}|| &\leq ||\theta\vec{x}|| + ||(1-\theta)\,\vec{y}|| \\ &= \theta\,||\vec{x}|| + (1-\theta)\,||\vec{y}|| \end{aligned}$$

Composition of convex and affine function

If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then for any $A \in \mathbb{R}^{n \times m}$ and $\vec{b} \in \mathbb{R}^n$

$$h(\vec{x}) := f\left(A\vec{x} + \vec{b}\right)$$

is convex

Consequence:

$$f(\vec{x}) := \left|\left|A\vec{x} + \vec{b}\right|\right|$$

is convex for any A and $\vec{b}$

Least-squares cost function is convex

Composition of convex and affine function

$$h\left(\theta\vec{x} + (1-\theta)\,\vec{y}\right)$$

Composition of convex and affine function

$$h\left(\theta\vec{x}+\left(1-\theta\right)\vec{y}\right)=f\left(\theta\left(A\vec{x}+\vec{b}\right)+\left(1-\theta\right)\left(A\vec{y}+\vec{b}\right)\right)$$

Composition of convex and affine function

$$
\begin{aligned}
h\left(\theta\vec{x}+(1-\theta)\,\vec{y}\right) &= f\left(\theta\left(A\vec{x}+\vec{b}\right)+(1-\theta)\left(A\vec{y}+\vec{b}\right)\right) \\
&\leq \theta f\left(A\vec{x}+\vec{b}\right)+(1-\theta)\,f\left(A\vec{y}+\vec{b}\right)
\end{aligned}
$$

Composition of convex and affine function

$$
\begin{aligned}
h\left(\theta\vec{x}+(1-\theta)\,\vec{y}\right) &= f\left(\theta\left(A\vec{x}+\vec{b}\right)+(1-\theta)\left(A\vec{y}+\vec{b}\right)\right) \\
&\leq \theta f\left(A\vec{x}+\vec{b}\right)+(1-\theta)\,f\left(A\vec{y}+\vec{b}\right) \\
&= \theta\, h\left(\vec{x}\right)+(1-\theta)\,h\left(\vec{y}\right)
\end{aligned}
$$

ℓ_0 "norm"

Number of nonzero entries in a vector

Not a norm!

$$||2\vec{x}||_0$$

ℓ_0 "norm"

Number of nonzero entries in a vector

Not a norm!

$$||2\vec{x}||_0 = ||\vec{x}||_0$$

ℓ_0 "norm"

Number of nonzero entries in a vector

Not a norm!

$$\begin{aligned} ||2\vec{x}||_0 &= ||\vec{x}||_0 \\ &\neq 2\,||\vec{x}||_0 \end{aligned}$$

ℓ_0 "norm"

Number of nonzero entries in a vector

Not a norm!

$$\begin{aligned} ||2\vec{x}||_0 &= ||\vec{x}||_0 \\ &\neq 2\,||\vec{x}||_0 \end{aligned}$$

Not convex

ℓ_0 "norm"

Number of nonzero entries in a vector

Not a norm!

$$\begin{aligned} ||2\vec{x}||_0 &= ||\vec{x}||_0 \\ &\neq 2\,||\vec{x}||_0 \end{aligned}$$

Not convex

Let $\vec{x} := \left(\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right)$ and $\vec{y} := \left(\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right)$, for any $\theta \in (0, 1)$

$$||\theta\vec{x} + (1-\theta)\,\vec{y}||_0$$

$$\theta\,||\vec{x}||_0 + (1-\theta)\,||\vec{y}||_0$$

ℓ_0 "norm"

Number of nonzero entries in a vector

Not a norm!

$$\begin{aligned} ||2\vec{x}||_0 &= ||\vec{x}||_0 \\ &\neq 2\,||\vec{x}||_0 \end{aligned}$$

Not convex

Let $\vec{x} := \left(\begin{smallmatrix}1\\0\end{smallmatrix}\right)$ and $\vec{y} := \left(\begin{smallmatrix}0\\1\end{smallmatrix}\right)$, for any $\theta \in (0,1)$

$$||\theta\vec{x} + (1-\theta)\,\vec{y}||_0 = 2$$

$$\theta\,||\vec{x}||_0 + (1-\theta)\,||\vec{y}||_0$$

ℓ_0 "norm"

Number of nonzero entries in a vector

Not a norm!

$$\begin{aligned} ||2\vec{x}||_0 &= ||\vec{x}||_0 \\ &\neq 2\,||\vec{x}||_0 \end{aligned}$$

Not convex

Let $\vec{x} := \left(\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right)$ and $\vec{y} := \left(\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right)$, for any $\theta \in (0, 1)$

$$||\theta\vec{x} + (1-\theta)\,\vec{y}||_0 = 2$$

$$\theta\,||\vec{x}||_0 + (1-\theta)\,||\vec{y}||_0 = 1$$

Promoting sparsity

Toy problem: Find t such that

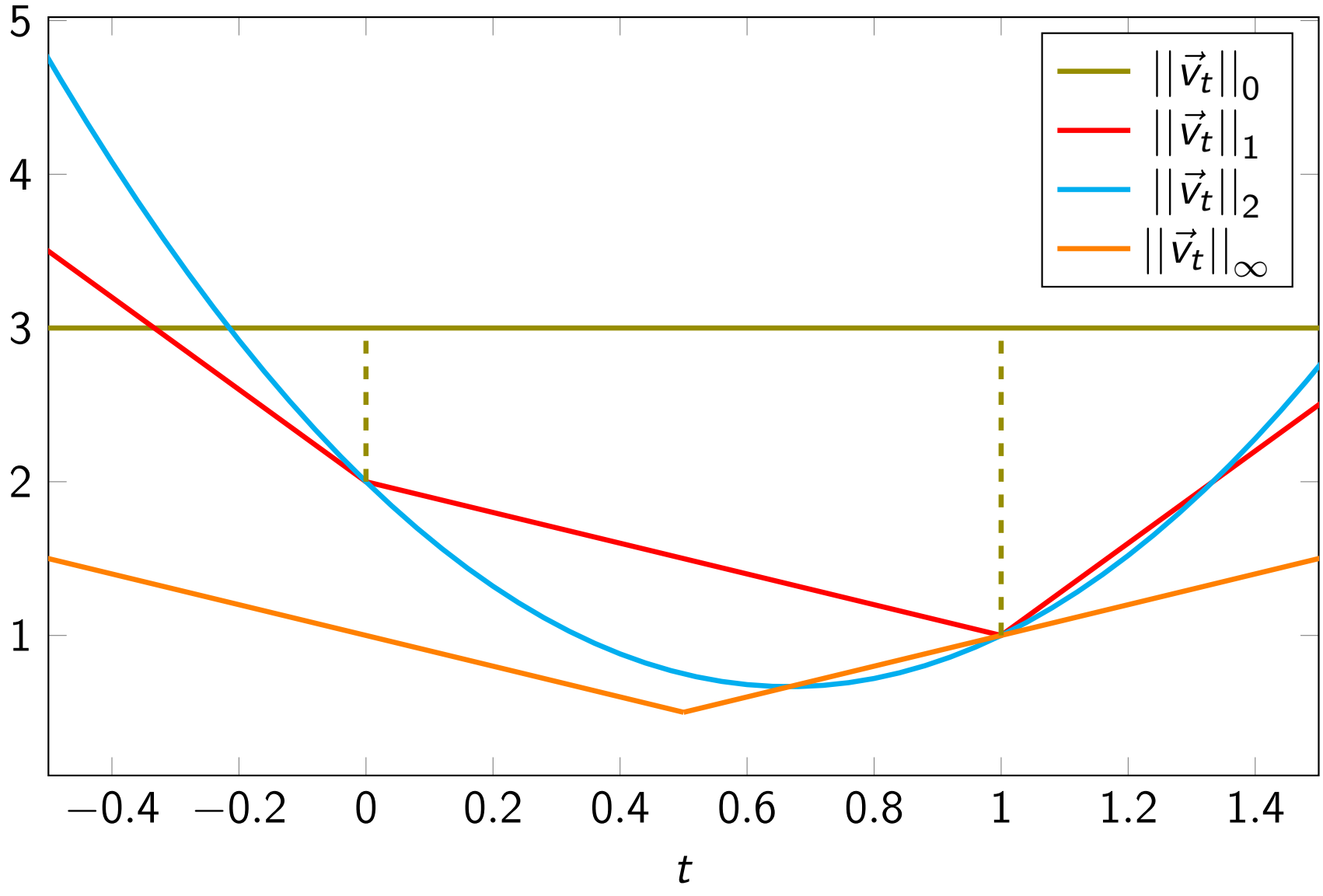
$$\vec{v}_t := \begin{bmatrix} t \\ t-1 \\ t-1 \end{bmatrix}$$

is sparse

Strategy: Minimize

$$f(t) := ||\vec{v}_t||$$

Promoting sparsity



The rank is not convex

The rank of matrices in $\mathbb{R}^{n\times n}$ interpreted as a function from $\mathbb{R}^{n\times n}$ to $\mathbb{R}$ is not convex

The rank is not convex

The rank of matrices in $\mathbb{R}^{n\times n}$ interpreted as a function from $\mathbb{R}^{n\times n}$ to $\mathbb{R}$ is not convex

$$X := \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \qquad Y := \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

For any $\theta \in (0, 1)$

$$\operatorname{rank}\left(\theta X + (1-\theta)\, Y\right)$$

$$\theta \operatorname{rank}\left(X\right) + (1-\theta) \operatorname{rank}\left(Y\right)$$

The rank is not convex

The rank of matrices in $\mathbb{R}^{n\times n}$ interpreted as a function from $\mathbb{R}^{n\times n}$ to $\mathbb{R}$ is not convex

$$X := \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \qquad Y := \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

For any $\theta \in (0, 1)$

$$\operatorname{rank}\left(\theta X + (1-\theta)\, Y\right) = 2$$

$$\theta \operatorname{rank}\left(X\right) + (1-\theta) \operatorname{rank}\left(Y\right)$$

The rank is not convex

The rank of matrices in $\mathbb{R}^{n\times n}$ interpreted as a function from $\mathbb{R}^{n\times n}$ to $\mathbb{R}$ is not convex

$$X := \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \qquad Y := \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

For any $\theta \in (0, 1)$

$$\operatorname{rank}\left(\theta X + (1-\theta)\, Y\right) = 2$$

$$\theta \operatorname{rank}\left(X\right) + (1-\theta) \operatorname{rank}\left(Y\right) = 1$$

Matrix norms

Frobenius norm

$$||A||_{\mathsf{F}} := \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2}$$

Operator norm

$$||A|| := \max_{\left\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\right\}} ||A\,\vec{x}||_2 = \sigma_1$$

Nuclear norm

$$||A||_* := \sum_{i=1}^{\min\{m,n\}} \sigma_i$$

Hölder's inequality for matrices

For any matrix $A \in \mathbb{R}^{m \times n}$,

$$||A||_* = \sup_{\{||B|| \leq 1 \mid B \in \mathbb{R}^{m \times n}\}} \langle A, B \rangle .$$

Hölder's inequality for matrices

For any matrix $A \in \mathbb{R}^{m \times n}$,

$$||A||_* = \sup_{\{||B|| \leq 1 \mid B \in \mathbb{R}^{m \times n}\}} \langle A, B \rangle .$$

Consequence: nuclear norm satisfies triangle inequality

$$||A + B||_* = \sup_{\{||C|| \leq 1 \mid C \in \mathbb{R}^{m \times n}\}} \langle A + B, C \rangle$$

Hölder's inequality for matrices

For any matrix $A \in \mathbb{R}^{m \times n}$,

$$||A||_* = \sup_{\{||B|| \leq 1 \mid B \in \mathbb{R}^{m \times n}\}} \langle A, B \rangle .$$

Consequence: nuclear norm satisfies triangle inequality

$$\begin{aligned} ||A + B||_* &= \sup_{\{||C|| \leq 1 \mid C \in \mathbb{R}^{m \times n}\}} \langle A + B, C \rangle \\ &\leq \sup_{\{||C|| \leq 1 \mid C \in \mathbb{R}^{m \times n}\}} \langle A, C \rangle + \sup_{\{||D|| \leq 1 \mid D \in \mathbb{R}^{m \times n}\}} \langle B, D \rangle \end{aligned}$$

Hölder's inequality for matrices

For any matrix $A \in \mathbb{R}^{m \times n}$,

$$||A||_* = \sup_{\{||B|| \leq 1 \mid B \in \mathbb{R}^{m \times n}\}} \langle A, B \rangle .$$

Consequence: nuclear norm satisfies triangle inequality

$$\begin{aligned} ||A + B||_* &= \sup_{\{||C|| \leq 1 \mid C \in \mathbb{R}^{m \times n}\}} \langle A + B, C \rangle \\ &\leq \sup_{\{||C|| \leq 1 \mid C \in \mathbb{R}^{m \times n}\}} \langle A, C \rangle + \sup_{\{||D|| \leq 1 \mid D \in \mathbb{R}^{m \times n}\}} \langle B, D \rangle \\ &= ||A||_* + ||B||_* \end{aligned}$$

Promoting low-rank structure

Finding low-rank matrices consistent with data is often very useful

Toy problem: Find t such that

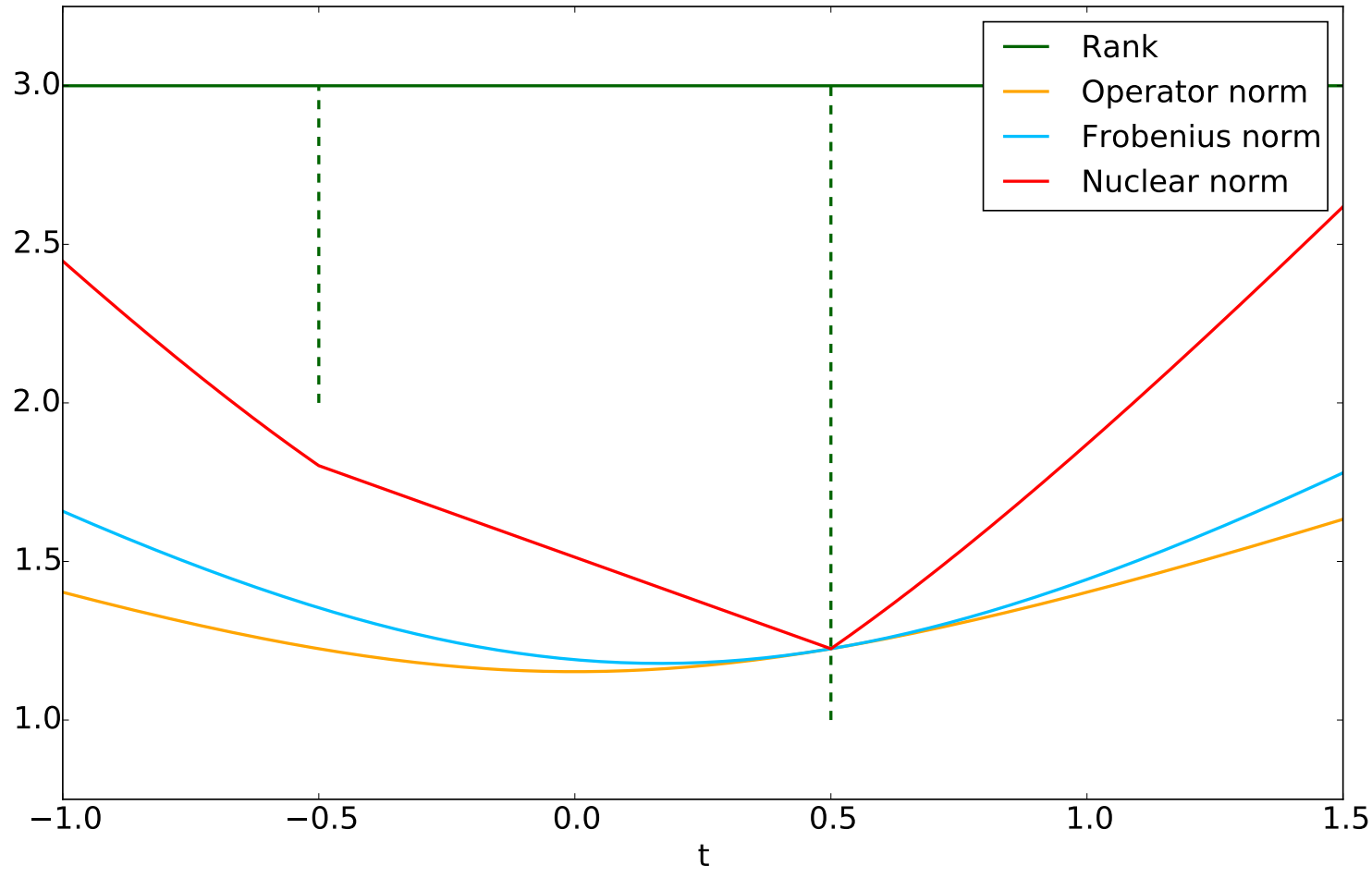
$$M(t) := \begin{bmatrix} 0.5+t & 1 & 1 \\ 0.5 & 0.5 & t \\ 0.5 & 1-t & 0.5 \end{bmatrix},$$

is low rank

Strategy: Minimize

$$f(t) := ||M(t)||$$

Promoting low-rank structure



Motivating applications

Convex functions

Optimality conditions

Convex Regularization

Optimization algorithms

Gradient

$$\nabla f(\vec{x}) = \begin{bmatrix} \frac{\partial f(\vec{x})}{\partial \vec{x}[1]} \\ \frac{\partial f(\vec{x})}{\partial \vec{x}[2]} \\ \cdots \\ \frac{\partial f(\vec{x})}{\partial \vec{x}[n]} \end{bmatrix}$$

If the gradient exists at every point, the function is said to be differentiable

Directional derivative

Encodes first-order rate of change in a particular direction

$$f'_{\vec{u}}(\vec{x}) := \lim_{h \to 0} \frac{f(\vec{x} + h\vec{u}) - f(\vec{x})}{h}$$
$$= \langle \nabla f(\vec{x}), \vec{u} \rangle$$

where $||u||_2 = 1$

Direction of maximum variation

∇f is direction of maximum increase

-∇f is direction of maximum decrease

$$\left|f'_{\vec{u}}(\vec{x})\right| = \left|\nabla f\left(\vec{x}\right)^T \vec{u}\right|$$

Direction of maximum variation

∇f is direction of maximum increase

-∇f is direction of maximum decrease

$$\begin{aligned} \left|f'_{\vec{u}}(\vec{x})\right| &= \left|\nabla f(\vec{x})^T \vec{u}\right| \\ &\leq ||\nabla f(\vec{x})||_2 \, ||\vec{u}||_2 \end{aligned}$$

Cauchy-Schwarz inequality

Direction of maximum variation

∇f is direction of maximum increase

-∇f is direction of maximum decrease

$$\begin{aligned} \left|f'_{\vec{u}}(\vec{x})\right| &= \left|\nabla f(\vec{x})^T \vec{u}\right| \\ &\leq ||\nabla f(\vec{x})||_2 \, ||\vec{u}||_2 \qquad \text{Cauchy-Schwarz inequality} \\ &= ||\nabla f(\vec{x})||_2 \end{aligned}$$

Direction of maximum variation

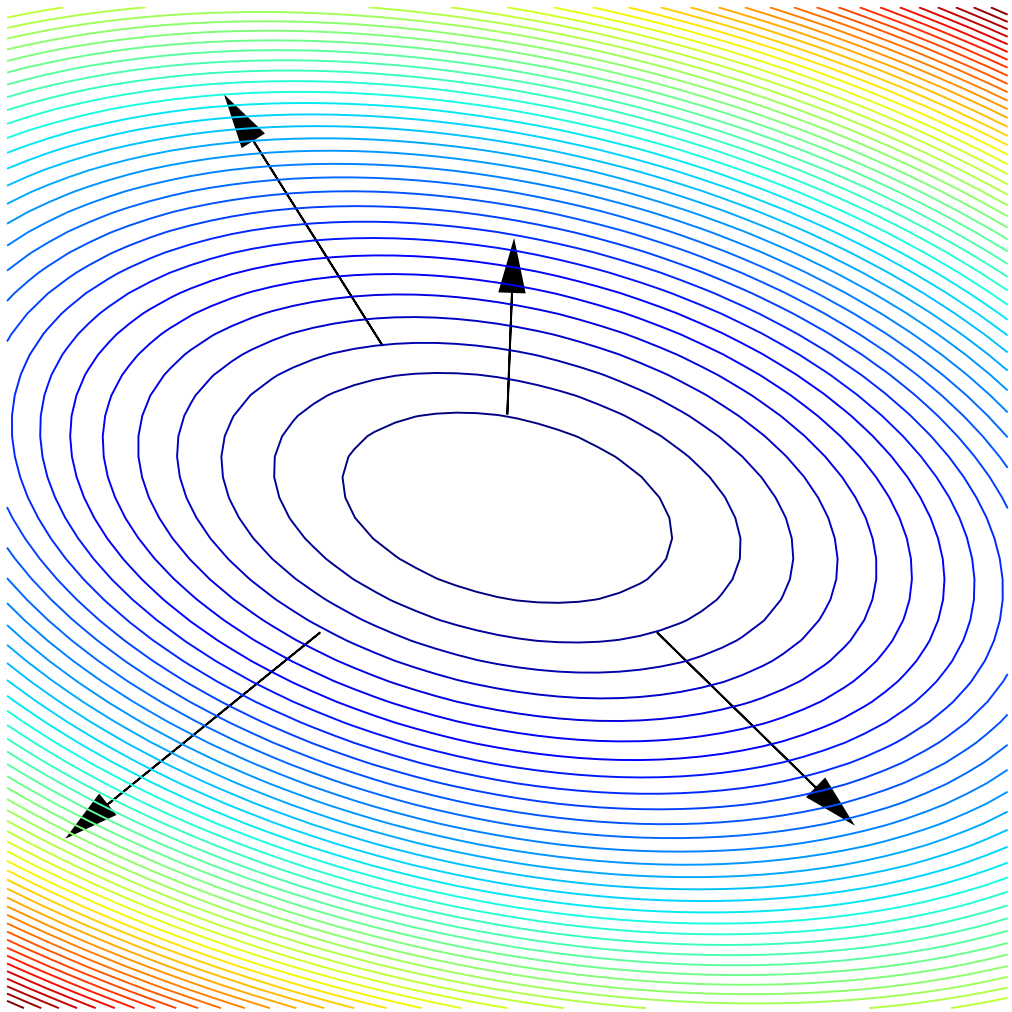
∇f is direction of maximum increase

-∇f is direction of maximum decrease

$$\begin{aligned} \left|f'_{\vec{u}}(\vec{x})\right| &= \left|\nabla f(\vec{x})^T \vec{u}\right| \\ &\leq ||\nabla f(\vec{x})||_2 \, ||\vec{u}||_2 \qquad \text{Cauchy-Schwarz inequality} \\ &= ||\nabla f(\vec{x})||_2 \end{aligned}$$

equality holds if and only if $\vec{u} = \pm \frac{\nabla f(\vec{x})}{||\nabla f(\vec{x})||_2}$

Gradient



Gradients of simple functions

Let $A \in \mathbb{R}^{a \times b}$, $M \in \mathbb{R}^{b \times b}$ and symmetric, $\vec{x} \in \mathbb{R}^b$,$\vec{c} \in \mathbb{R}^b$, $d \in \mathbb{R}$

$$f_1(\vec{x}) := d,$$
$$f_2(\vec{x}) := \vec{c}^T \vec{x},$$
$$f_3(\vec{x}) := \vec{x}^T M \vec{x},$$

$$\nabla f_1(\vec{x}) =$$
$$\nabla f_2(\vec{x}) =$$
$$\nabla f_3(\vec{x}) =$$

Gradients of simple functions

Let $A \in \mathbb{R}^{a \times b}$, $M \in \mathbb{R}^{b \times b}$ and symmetric, $\vec{x} \in \mathbb{R}^b$,$\vec{c} \in \mathbb{R}^b$, $d \in \mathbb{R}$

$$f_1(\vec{x}) := d,$$
$$f_2(\vec{x}) := \vec{c}^T \vec{x},$$
$$f_3(\vec{x}) := \vec{x}^T M \vec{x},$$

$$\nabla f_1(\vec{x}) = 0$$
$$\nabla f_2(\vec{x}) =$$
$$\nabla f_3(\vec{x}) =$$

Gradients of simple functions

Let $A \in \mathbb{R}^{a \times b}$, $M \in \mathbb{R}^{b \times b}$ and symmetric, $\vec{x} \in \mathbb{R}^b$,$\vec{c} \in \mathbb{R}^b$, $d \in \mathbb{R}$

$$
\begin{aligned}
f_1(\vec{x}) &:= d, \\
f_2(\vec{x}) &:= \vec{c}^T \vec{x}, \\
f_3(\vec{x}) &:= \vec{x}^T M \vec{x},
\end{aligned}
$$

$$
\begin{aligned}
\nabla f_1(\vec{x}) &= 0 \\
\nabla f_2(\vec{x}) &= \vec{c} \\
\nabla f_3(\vec{x}) &=
\end{aligned}
$$

Gradients of simple functions

Let $A \in \mathbb{R}^{a \times b}$, $M \in \mathbb{R}^{b \times b}$ and symmetric, $\vec{x} \in \mathbb{R}^b$,$\vec{c} \in \mathbb{R}^b$, $d \in \mathbb{R}$

$$
\begin{aligned}
f_1(\vec{x}) &:= d, \\
f_2(\vec{x}) &:= \vec{c}^T \vec{x}, \\
f_3(\vec{x}) &:= \vec{x}^T M \vec{x},
\end{aligned}
$$

$$
\begin{aligned}
\nabla f_1(\vec{x}) &= 0 \\
\nabla f_2(\vec{x}) &= \vec{c} \\
\nabla f_3(\vec{x}) &= 2M\vec{x}
\end{aligned}
$$

Least squares

Let $\vec{y} \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, $\vec{\beta} \in \mathbb{R}^p$

The gradient of the least-squares cost function

$$f(\vec{\beta}) := \frac{1}{2} \left|\left| \vec{y} - X\vec{\beta} \right|\right|_2^2 = \frac{1}{2}\vec{y}^T\vec{y} + \frac{1}{2}\vec{\beta}^T X^T X\vec{\beta} - \vec{y}^T X\vec{\beta}$$

equals

$$\nabla f(\vec{\beta}) =$$

Least squares

Let $\vec{y} \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, $\vec{\beta} \in \mathbb{R}^p$

The gradient of the least-squares cost function

$$f(\vec{\beta}) := \frac{1}{2} \left|\left| \vec{y} - X\vec{\beta} \right|\right|_2^2 = \frac{1}{2}\vec{y}^T\vec{y} + \frac{1}{2}\vec{\beta}^T X^T X\vec{\beta} - \vec{y}^T X\vec{\beta}$$

equals

$$\nabla f(\vec{\beta}) = X^T(X\vec{\beta} - \vec{y})$$

First-order approximation

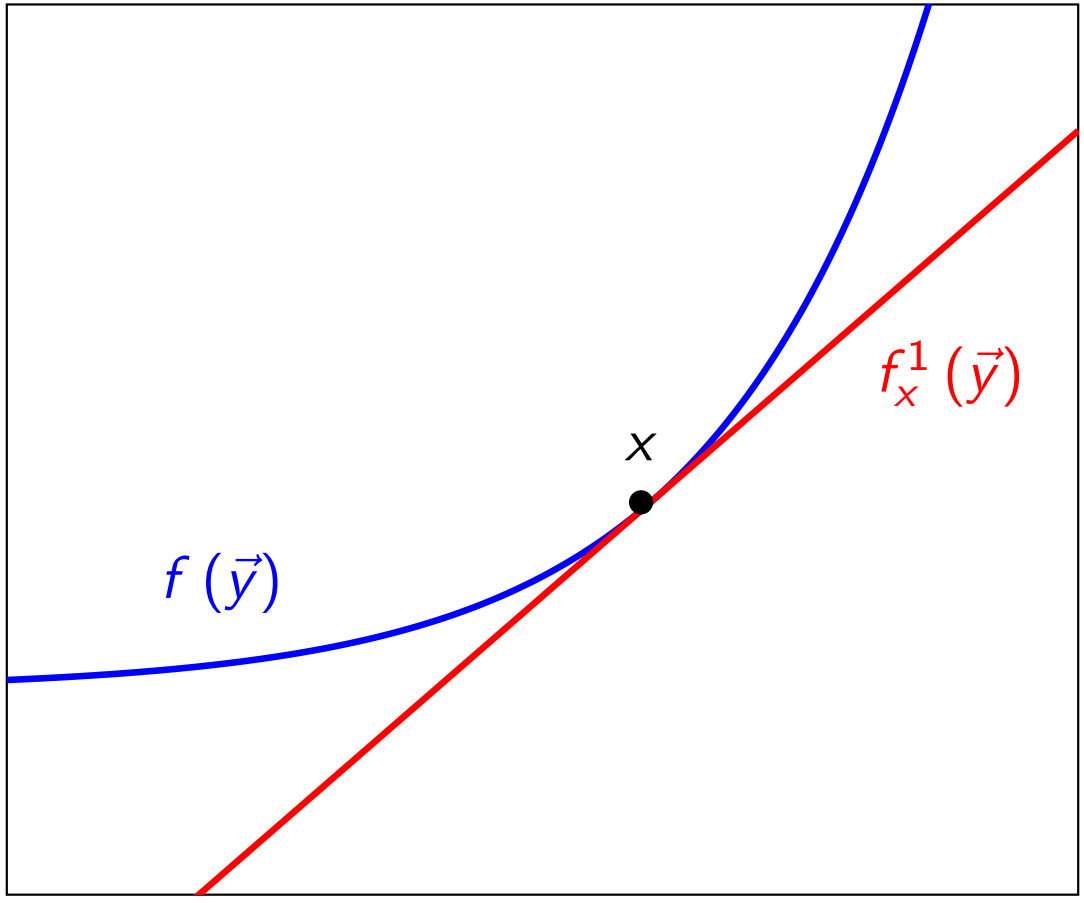
The first-order or linear approximation of $f : \mathbb{R}^n \to \mathbb{R}$ at $\vec{x}$ is

$$f_{\vec{x}}^1(\vec{y}) := f(\vec{x}) + \nabla f(\vec{x})^T (\vec{y} - \vec{x})$$

If f is differentiable at $\vec{x}$

$$\lim_{\vec{y} \to \vec{x}} \frac{f(\vec{y}) - f_{\vec{x}}^1(\vec{y})}{||\vec{y} - \vec{x}||_2} = 0$$

First-order approximation



Convexity

A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if for every $\vec{x}, \vec{y} \in \mathbb{R}^n$

$$f(\vec{y}) \geq f(\vec{x}) + \nabla f(\vec{x})^T (\vec{y} - \vec{x})$$

It is strictly convex if and only if

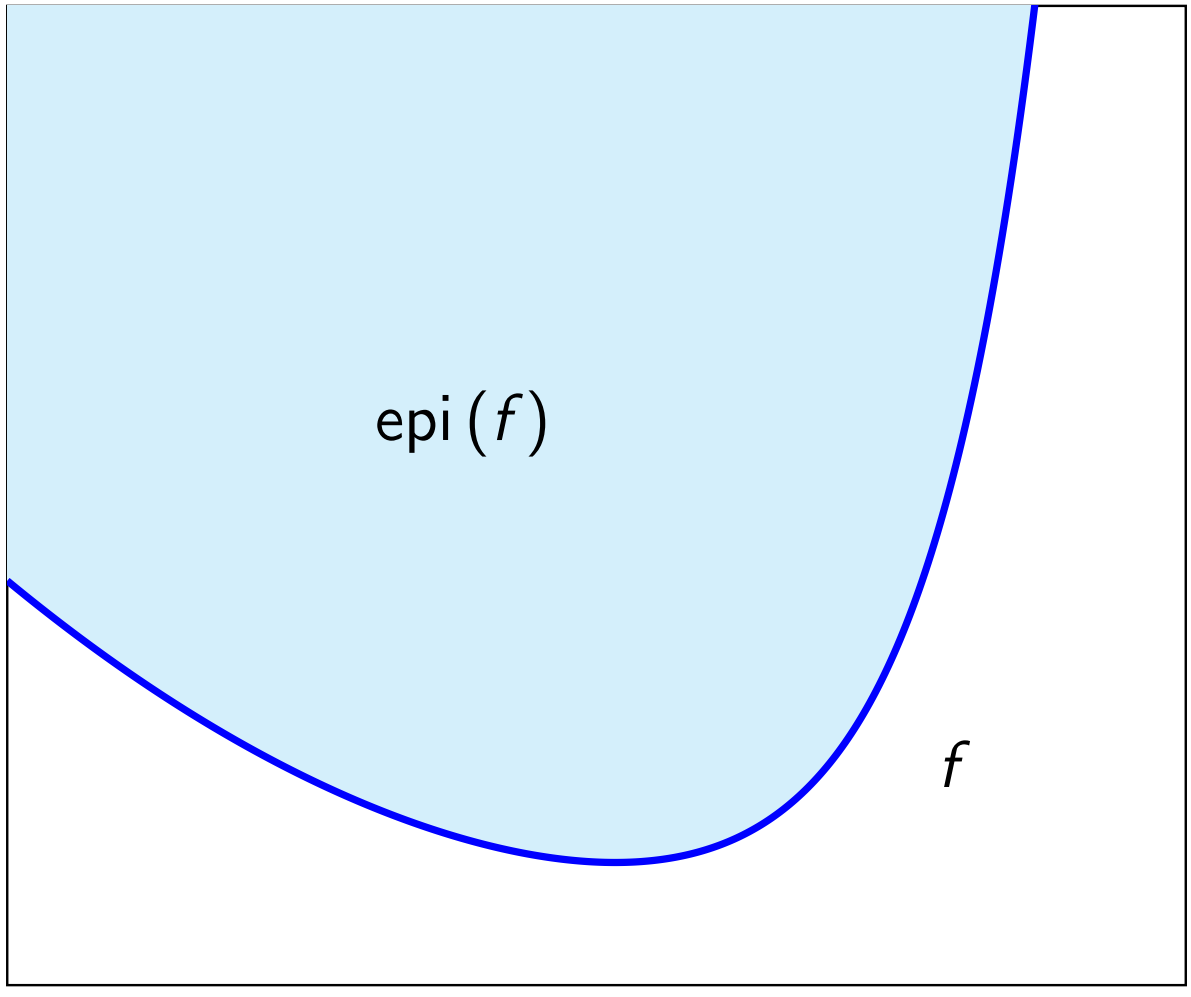
$$f(\vec{y}) > f(\vec{x}) + \nabla f(\vec{x})^T (\vec{y} - \vec{x})$$

Epigraph

The epigraph of $f : \mathbb{R}^n \to \mathbb{R}$ is

$$\text{epi}\,(f) := \left\{ \vec{x} \mid f\left(\begin{bmatrix} \vec{x}[1] \\ \cdots \\ \vec{x}[n] \end{bmatrix} \right) \leq \vec{x}[n+1] \right\}$$

Epigraph



Supporting hyperplane

A hyperplane $\mathcal{H}$ is a supporting hyperplane of a set $\mathcal{S}$ at $\vec{x}$ if

- $\mathcal{H}$ and $\mathcal{S}$ intersect at $\vec{x}$
- $\mathcal{S}$ is contained in one of the half-spaces bounded by $\mathcal{H}$

Geometric intuition

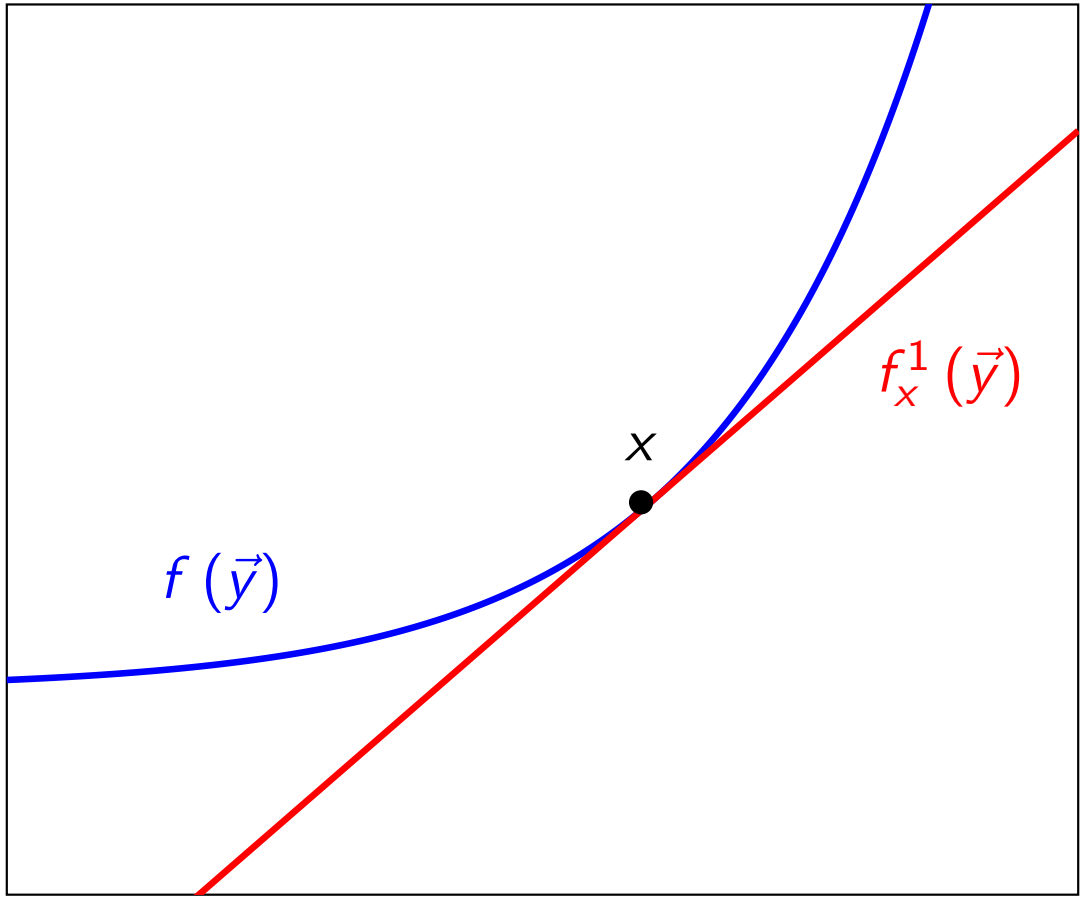
Geometrically, f is convex if and only if for every $\vec{x}$ the plane

$$\mathcal{H}_{f,\vec{x}} := \left\{ \vec{y} \mid \vec{y}[n+1] = f^1_{\vec{x}} \left(\begin{bmatrix} \vec{y}[1] \\ \cdots \\ \vec{y}[n] \end{bmatrix} \right) \right\}$$

is a supporting hyperplane of the epigraph at $\vec{x}$

If $\nabla f(\vec{x}) = 0$ the hyperplane is horizontal

Convexity



Optimality condition

If f is convex, $\nabla f(\vec{x}) = 0$ if and only if $\vec{x}$ is a minimum

For any $\vec{y}$

$$\begin{aligned} f(\vec{y}) &\geq f(\vec{x}) + \nabla f(\vec{x})^T (\vec{y} - \vec{x}) \\ &\geq f(\vec{x}) \end{aligned}$$

Least squares

Let $\vec{y} \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, $\vec{\beta} \in \mathbb{R}^p$

$$f(\vec{\beta}) := \frac{1}{2} \left|\left| \vec{y} - X\vec{\beta} \right|\right|_2^2$$
$$\nabla f(\vec{\beta}) = X^T(X\vec{\beta} - \vec{y})$$

Setting gradient to zero yields normal equations

$$X^T X\vec{\beta} = X^T\vec{y}$$

Least squares

Let $\vec{y} \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, $\vec{\beta} \in \mathbb{R}^p$

$$f(\vec{\beta}) := \frac{1}{2} \left\| \vec{y} - X\vec{\beta} \right\|_2^2$$
$$\nabla f(\vec{\beta}) = X^T(X\vec{\beta} - \vec{y})$$

Setting gradient to zero yields normal equations

$$X^T X \vec{\beta} = X^T \vec{y}$$

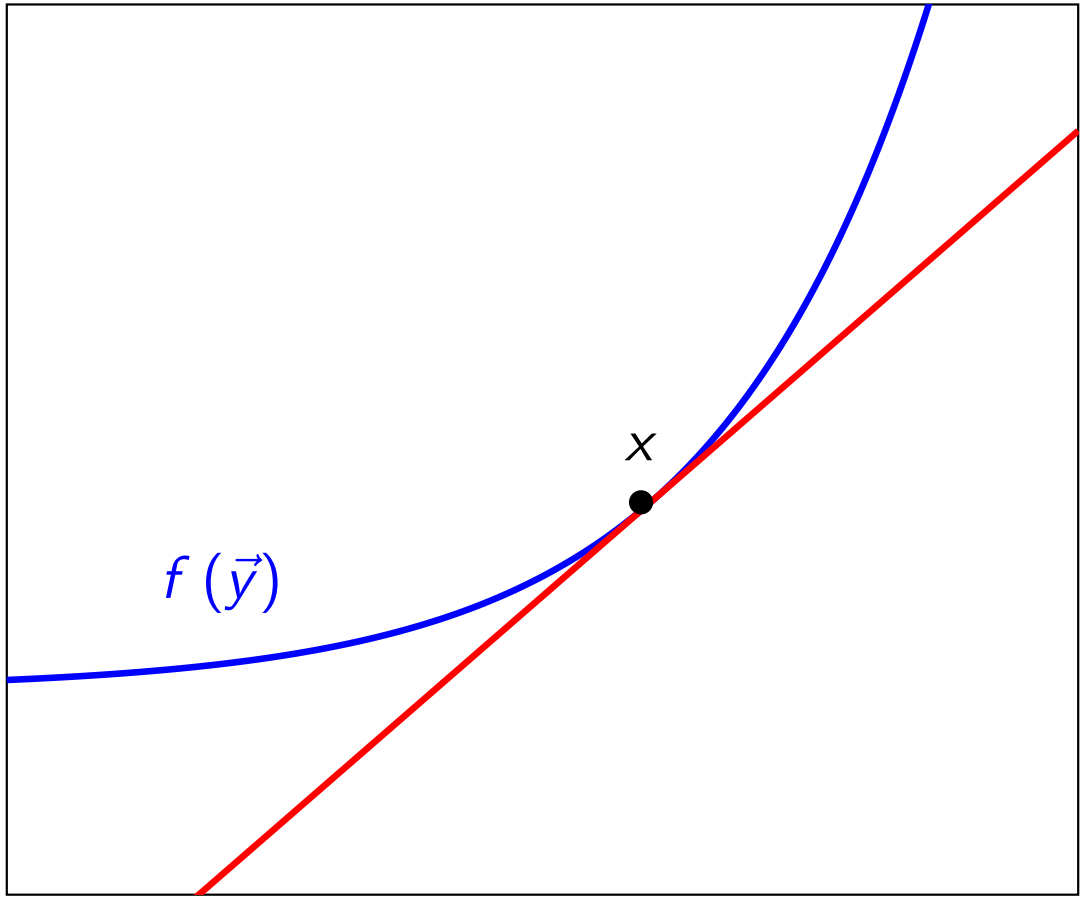
If $X^T X$ is invertible $\vec{\beta}_{\mathsf{LS}} := (X^T X)^{-1} X^T \vec{y}$

Gradient

A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if for every $\vec{x}, \vec{y} \in \mathbb{R}^n$

$$f(\vec{y}) \geq f(\vec{x}) + \nabla f(\vec{x})^T (\vec{y} - \vec{x})$$

Gradient



Subgradient

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

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

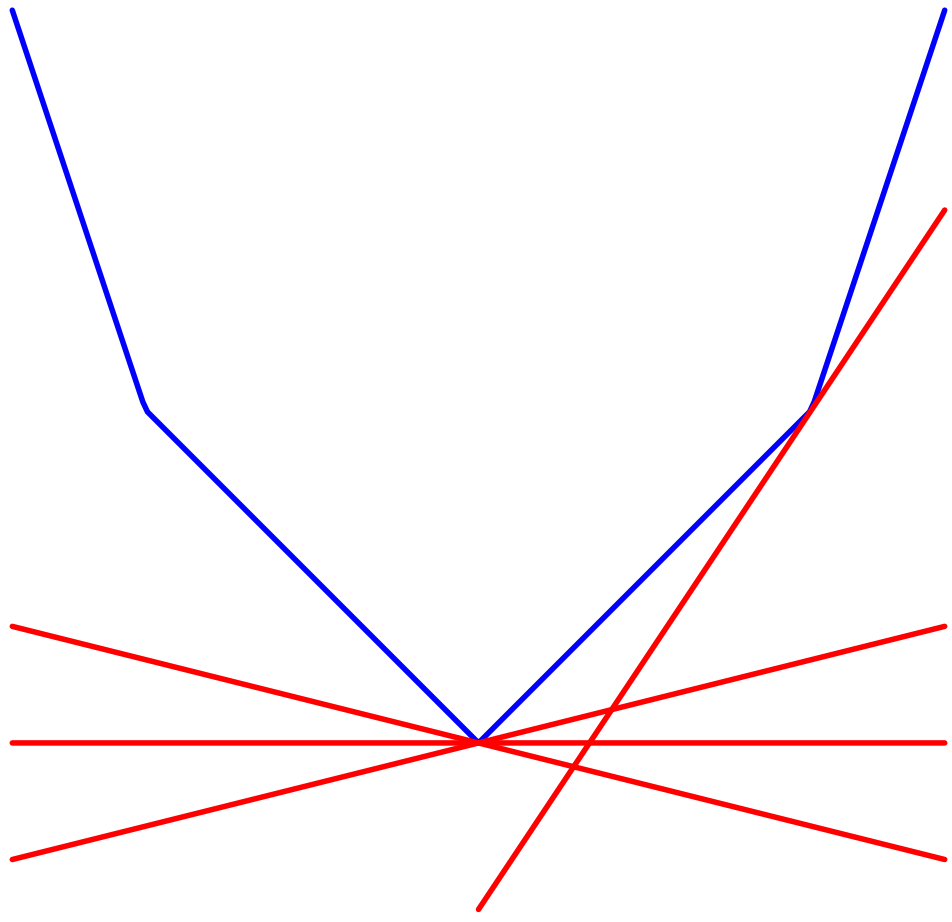
Geometrically, the hyperplane

$$\mathcal{H}_{\vec{g}} := \left\{ \vec{y} \mid \vec{y}[n+1] = \vec{g}^T \begin{pmatrix} \begin{bmatrix} \vec{y}[1] \\ \cdots \\ \vec{y}[n] \end{bmatrix} \end{pmatrix} \right\}$$

is a supporting hyperplane of the epigraph at $\vec{x}$

The set of all subgradients at $\vec{x}$ is called the subdifferential

Subgradients



Subgradient of differentiable function

If a function is differentiable, the only subgradient at each point is the gradient

Proof

Assume $\vec{g}$ is a subgradient at $\vec{x}$, for any $\alpha \geq 0$

$$\begin{aligned} f\left(\vec{x} + \alpha\, \vec{e}_i\right) &\geq f\left(\vec{x}\right) + \vec{g}^T \alpha\, \vec{e}_i \\ &= f\left(\vec{x}\right) + \vec{g}[i]\, \alpha \\ f\left(\vec{x}\right) &\leq f\left(\vec{x} - \alpha\, \vec{e}_i\right) + \vec{g}^T \alpha\, \vec{e}_i \\ &= f\left(\vec{x} - \alpha\, \vec{e}_i\right) + \vec{g}[i]\, \alpha \end{aligned}$$

Proof

Assume $\vec{g}$ is a subgradient at $\vec{x}$, for any $\alpha \geq 0$

$$\begin{aligned} f\left(\vec{x}+\alpha\,\vec{e}_i\right) &\geq f\left(\vec{x}\right)+\vec{g}^{\,T}\alpha\,\vec{e}_i \\ &= f\left(\vec{x}\right)+\vec{g}[i]\,\alpha \\ f\left(\vec{x}\right) &\leq f\left(\vec{x}-\alpha\,\vec{e}_i\right)+\vec{g}^{\,T}\alpha\,\vec{e}_i \\ &= f\left(\vec{x}-\alpha\,\vec{e}_i\right)+\vec{g}[i]\,\alpha \end{aligned}$$

Combining both inequalities

$$\frac{f\left(\vec{x}\right)-f\left(\vec{x}-\alpha\,\vec{e}_i\right)}{\alpha} \leq \vec{g}[i] \leq \frac{f\left(\vec{x}+\alpha\,\vec{e}_i\right)-f\left(\vec{x}\right)}{\alpha}$$

Proof

Assume $\vec{g}$ is a subgradient at $\vec{x}$, for any $\alpha \geq 0$

$$\begin{aligned} f\left(\vec{x} + \alpha\, \vec{e_i}\right) &\geq f\left(\vec{x}\right) + \vec{g}^T \alpha\, \vec{e_i} \\ &= f\left(\vec{x}\right) + \vec{g}[i]\, \alpha \\ f\left(\vec{x}\right) &\leq f\left(\vec{x} - \alpha\, \vec{e_i}\right) + \vec{g}^T \alpha\, \vec{e_i} \\ &= f\left(\vec{x} - \alpha\, \vec{e_i}\right) + \vec{g}[i]\, \alpha \end{aligned}$$

Combining both inequalities

$$\frac{f\left(\vec{x}\right) - f\left(\vec{x} - \alpha\, \vec{e_i}\right)}{\alpha} \leq \vec{g}[i] \leq \frac{f\left(\vec{x} + \alpha\, \vec{e_i}\right) - f\left(\vec{x}\right)}{\alpha}$$

Letting $\alpha \to 0$, implies $\vec{g}[i] = \frac{\partial f(\vec{x})}{\partial \vec{x}[i]}$

Subgradient

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if it has a subgradient at every point

It is strictly convex if and only for all $\vec{x} \in \mathbb{R}^n$ there exists $\vec{g} \in \mathbb{R}^n$ such that

$$f(\vec{y}) > f(\vec{x}) + \vec{g}^{\,T}(\vec{y} - \vec{x}), \quad \text{for all } \vec{y} \neq \vec{x}.$$

Optimality condition for nondifferentiable functions

$\vec{x}$ is a minimum of f if and only if the zero vector is a subgradient of f at $\vec{x}$

$$f(\vec{y}) \geq f(\vec{x}) + \vec{0}^T(\vec{y} - \vec{x})$$

Optimality condition for nondifferentiable functions

$\vec{x}$ is a minimum of f if and only if the zero vector is a subgradient of f at $\vec{x}$

$$\begin{aligned} f(\vec{y}) &\geq f(\vec{x}) + \vec{0}^T (\vec{y} - \vec{x}) \\ &= f(\vec{x}) \end{aligned}$$

for all $\vec{y} \in \mathbb{R}^n$

Sum of subgradients

Let $\vec{g}_1$ and $\vec{g}_2$ be subgradients at $\vec{x} \in \mathbb{R}^n$ of $f_1 : \mathbb{R}^n \to \mathbb{R}$ and $f_2 : \mathbb{R}^n \to \mathbb{R}$

$\vec{g} := \vec{g}_1 + \vec{g}_2$ is a subgradient of $f := f_1 + f_2$ at $\vec{x}$

Sum of subgradients

Let $\vec{g}_1$ and $\vec{g}_2$ be subgradients at $\vec{x} \in \mathbb{R}^n$ of $f_1 : \mathbb{R}^n \to \mathbb{R}$ and $f_2 : \mathbb{R}^n \to \mathbb{R}$

$\vec{g} := \vec{g}_1 + \vec{g}_2$ is a subgradient of $f := f_1 + f_2$ at $\vec{x}$

Proof: For any $\vec{y} \in \mathbb{R}^n$

$$f(\vec{y}) = f_1(\vec{y}) + f_2(\vec{y})$$

Sum of subgradients

Let $\vec{g}_1$ and $\vec{g}_2$ be subgradients at $\vec{x} \in \mathbb{R}^n$ of $f_1 : \mathbb{R}^n \to \mathbb{R}$ and $f_2 : \mathbb{R}^n \to \mathbb{R}$

$\vec{g} := \vec{g}_1 + \vec{g}_2$ is a subgradient of $f := f_1 + f_2$ at $\vec{x}$

Proof: For any $\vec{y} \in \mathbb{R}^n$

$$\begin{aligned} f(\vec{y}) &= f_1(\vec{y}) + f_2(\vec{y}) \\ &\geq f_1(\vec{x}) + \vec{g}_1^T(\vec{y} - \vec{x}) + f_2(\vec{y}) + \vec{g}_2^T(\vec{y} - \vec{x}) \end{aligned}$$

Sum of subgradients

Let $\vec{g}_1$ and $\vec{g}_2$ be subgradients at $\vec{x} \in \mathbb{R}^n$ of $f_1 : \mathbb{R}^n \to \mathbb{R}$ and $f_2 : \mathbb{R}^n \to \mathbb{R}$

$\vec{g} := \vec{g}_1 + \vec{g}_2$ is a subgradient of $f := f_1 + f_2$ at $\vec{x}$

Proof: For any $\vec{y} \in \mathbb{R}^n$

$$\begin{aligned} f(\vec{y}) &= f_1(\vec{y}) + f_2(\vec{y}) \\ &\geq f_1(\vec{x}) + \vec{g}_1^T(\vec{y} - \vec{x}) + f_2(\vec{y}) + \vec{g}_2^T(\vec{y} - \vec{x}) \\ &\geq f(\vec{x}) + \vec{g}^T(\vec{y} - \vec{x}) \end{aligned}$$

Subgradient of scaled function

Let $\vec{g}_1$ be a subgradient at $\vec{x} \in \mathbb{R}^n$ of $f_1 : \mathbb{R}^n \to \mathbb{R}$

For any $\alpha \geq 0$ $\vec{g}_2 := \alpha \vec{g}_1$ is a subgradient of $f_2 := \alpha f_1$ at $\vec{x}$

Subgradient of scaled function

Let $\vec{g}_1$ be a subgradient at $\vec{x} \in \mathbb{R}^n$ of $f_1 : \mathbb{R}^n \to \mathbb{R}$

For any $\alpha \geq 0$ $\vec{g}_2 := \alpha \vec{g}_1$ is a subgradient of $f_2 := \alpha f_1$ at $\vec{x}$

Proof: For any $\vec{y} \in \mathbb{R}^n$

$$f_2(\vec{y}) = \alpha f_1(\vec{y})$$

Subgradient of scaled function

Let $\vec{g}_1$ be a subgradient at $\vec{x} \in \mathbb{R}^n$ of $f_1 : \mathbb{R}^n \to \mathbb{R}$

For any $\alpha \geq 0$ $\vec{g}_2 := \alpha \vec{g}_1$ is a subgradient of $f_2 := \alpha f_1$ at $\vec{x}$

Proof: For any $\vec{y} \in \mathbb{R}^n$

$$\begin{aligned} f_2(\vec{y}) &= \alpha f_1(\vec{y}) \\ &\geq \alpha \left(f_1(\vec{x}) + \vec{g}_1^T (\vec{y} - \vec{x}) \right) \end{aligned}$$

Subgradient of scaled function

Let $\vec{g}_1$ be a subgradient at $\vec{x} \in \mathbb{R}^n$ of $f_1 : \mathbb{R}^n \to \mathbb{R}$

For any $\alpha \geq 0$ $\vec{g}_2 := \alpha \vec{g}_1$ is a subgradient of $f_2 := \alpha f_1$ at $\vec{x}$

Proof: For any $\vec{y} \in \mathbb{R}^n$

$$\begin{aligned} f_2(\vec{y}) &= \alpha f_1(\vec{y}) \\ &\geq \alpha \left(f_1(\vec{x}) + \vec{g}_1^T (\vec{y} - \vec{x}) \right) \\ &\geq f_2(\vec{x}) + \vec{g}_2^T (\vec{y} - \vec{x}) \end{aligned}$$

Subdifferential of absolute value

At $x \neq 0$, $f(x) = |x|$ is differentiable, so $g = \text{sign}(x)$

At $x = 0$, we need

$$f(0 + y) \geq f(0) + g(y - 0)$$

Subdifferential of absolute value

At $x \neq 0$, $f(x) = |x|$ is differentiable, so $g = \text{sign}(x)$

At $x = 0$, we need

$$f(0 + y) \geq f(0) + g(y - 0)$$

$$|y| \geq gy$$

Subdifferential of absolute value

At $x \neq 0$, $f(x) = |x|$ is differentiable, so $g = \text{sign}(x)$

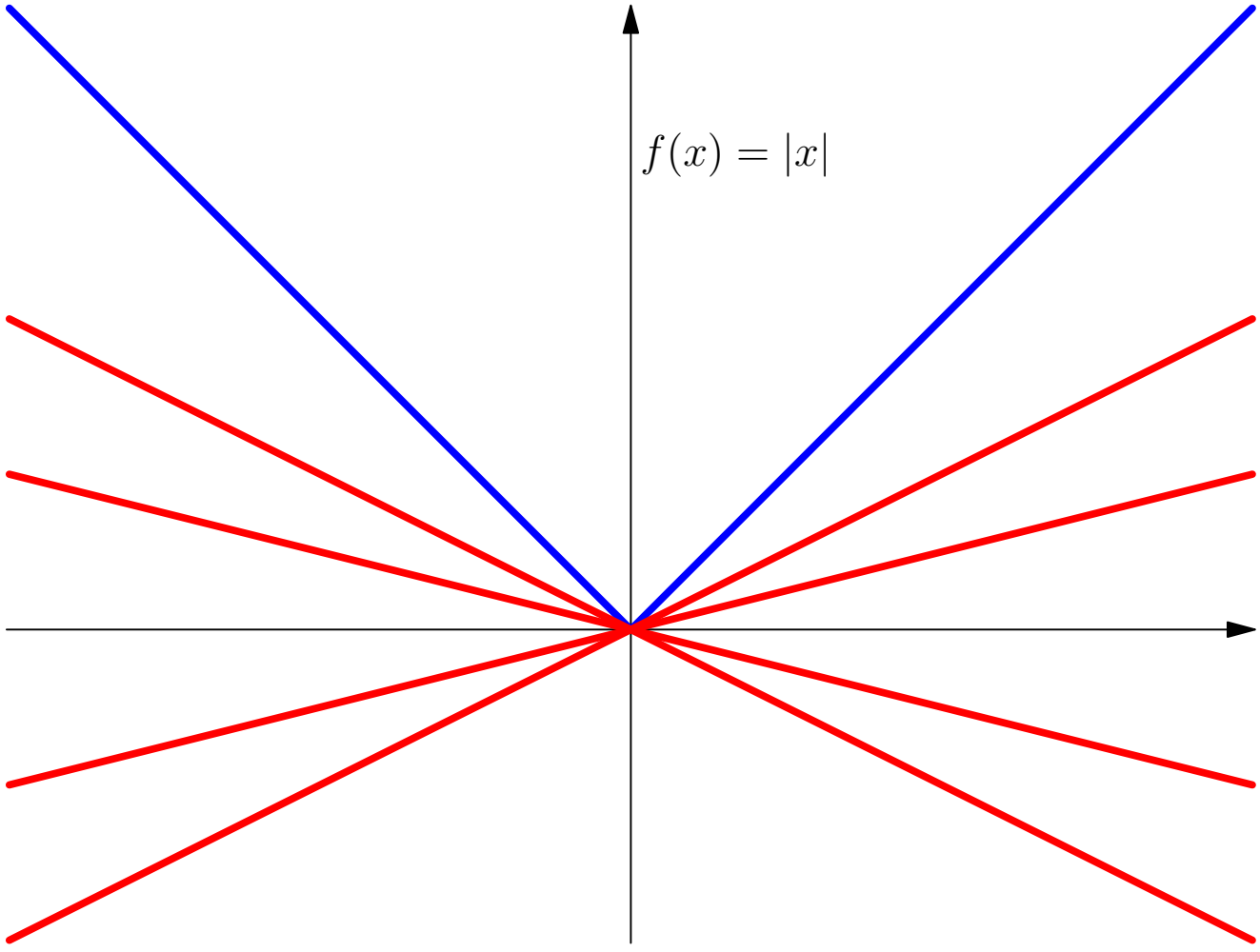
At $x = 0$, we need

$$f(0 + y) \geq f(0) + g(y - 0)$$

$$|y| \geq gy$$

Holds if and only if $|g| \leq 1$

Subdifferential of absolute value



Subdifferential of ℓ_1 norm

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

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

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

Proof (one direction)

$$||\vec{y}||_1 \geq \langle g, \vec{y} \rangle \quad \text{by Hölder's inequality because } ||g||_\infty \leq 1$$

Proof (one direction)

$$
\begin{aligned}
||\vec{y}||_1 &\geq \langle g, \vec{y} \rangle \quad \text{by Hölder's inequality because } ||g||_\infty \leq 1 \\
&= \langle g, \vec{x} \rangle + \langle g, \vec{y} - \vec{x} \rangle
\end{aligned}
$$

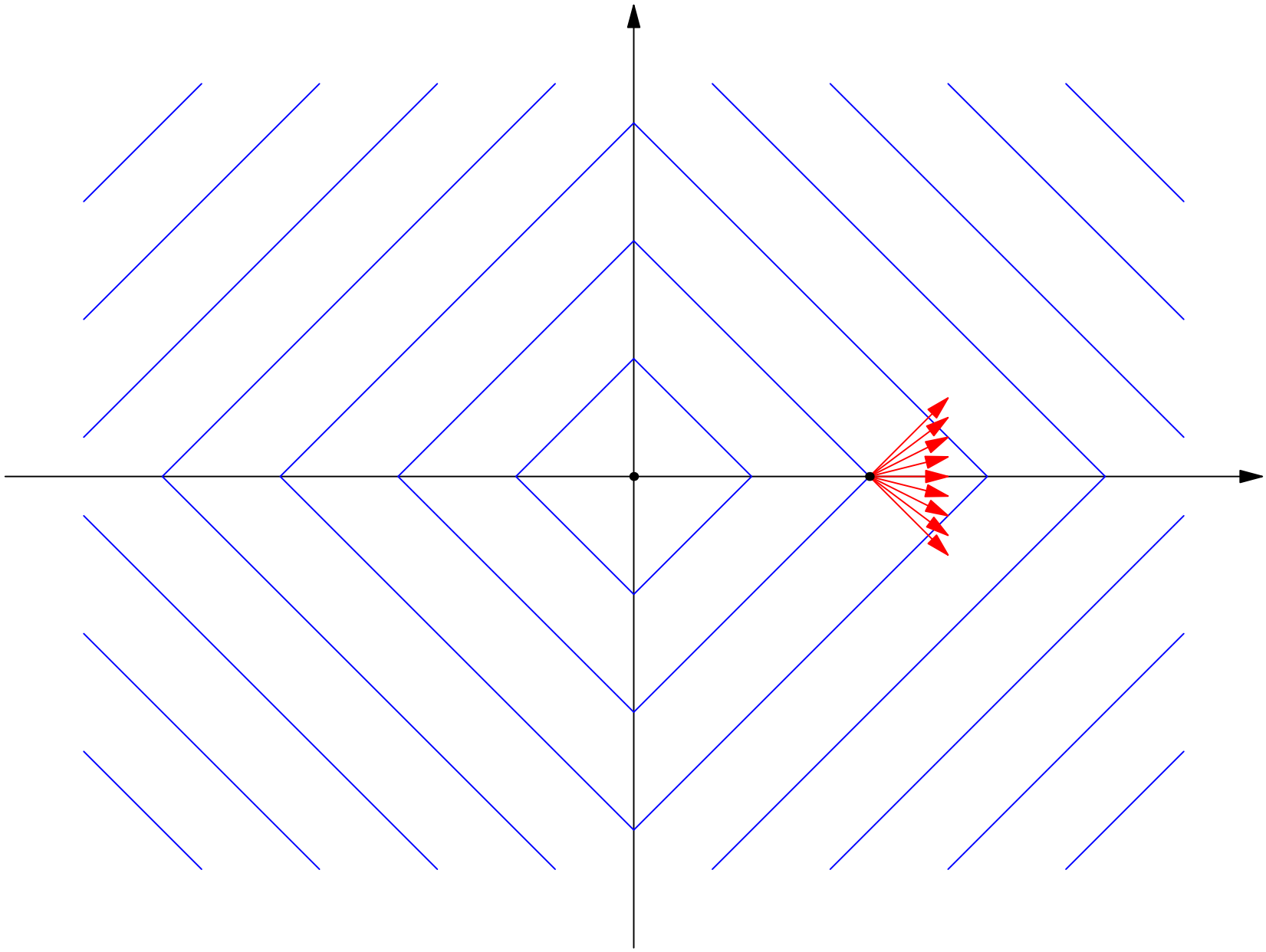
Proof (one direction)

$$
\begin{aligned}
||\vec{y}||_1 &\geq \langle g, \vec{y} \rangle \quad \text{by Hölder's inequality because } ||g||_\infty \leq 1 \\
&= \langle g, \vec{x} \rangle + \langle g, \vec{y} - \vec{x} \rangle \\
&= \sum_{i=1}^{n} \vec{x}[i] \operatorname{sign}(\vec{x}[i]) + \langle g, \vec{y} - \vec{x} \rangle
\end{aligned}
$$

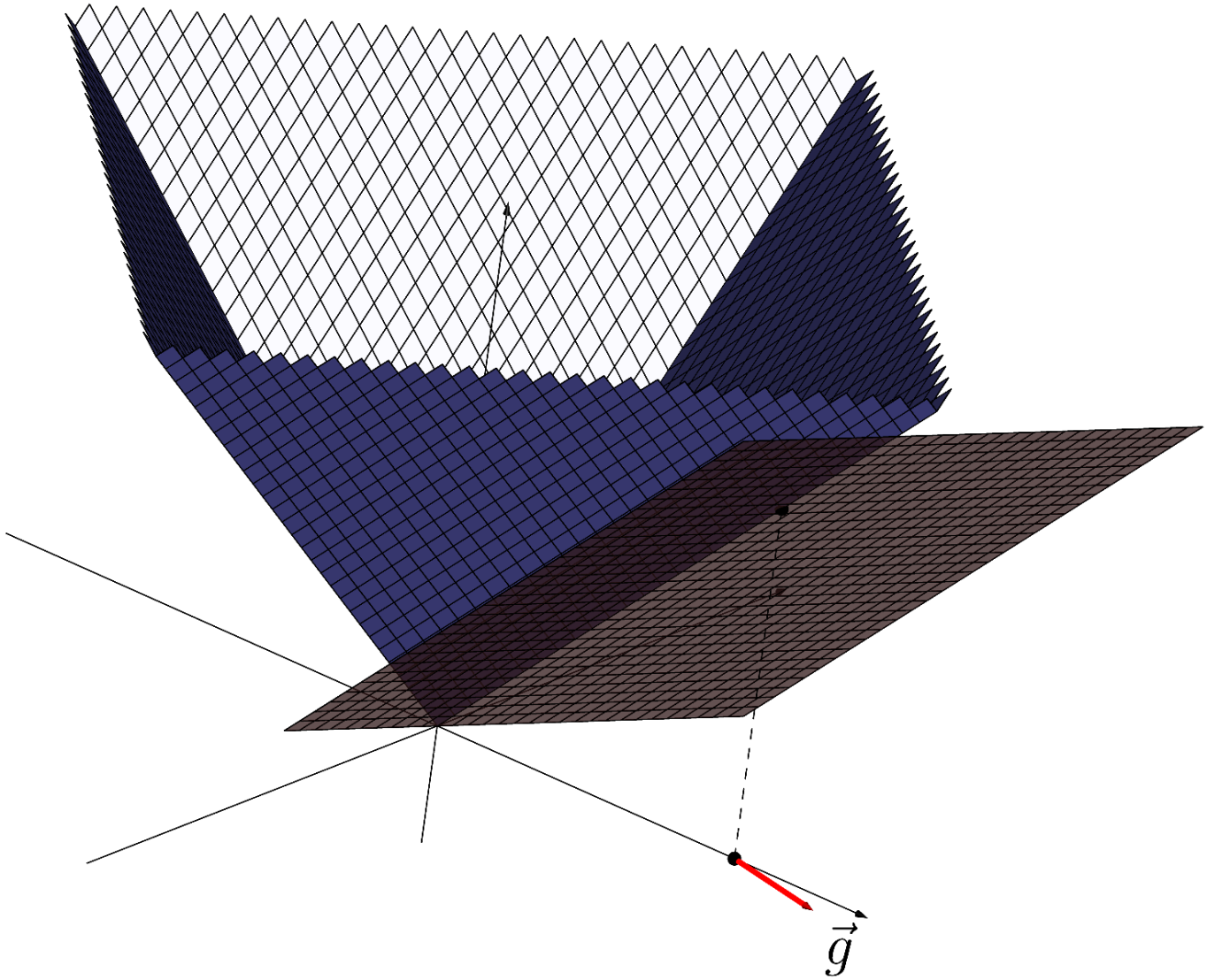
Proof (one direction)

$$
\begin{aligned}
||\vec{y}||_1 &\geq \langle g, \vec{y} \rangle \quad \text{by Hölder's inequality because } ||g||_\infty \leq 1 \\
&= \langle g, \vec{x} \rangle + \langle g, \vec{y} - \vec{x} \rangle \\
&= \sum_{i=1}^{n} \vec{x}[i] \operatorname{sign}(\vec{x}[i]) + \langle g, \vec{y} - \vec{x} \rangle \\
&= ||\vec{x}||_1 + \langle g, \vec{y} - \vec{x} \rangle
\end{aligned}
$$

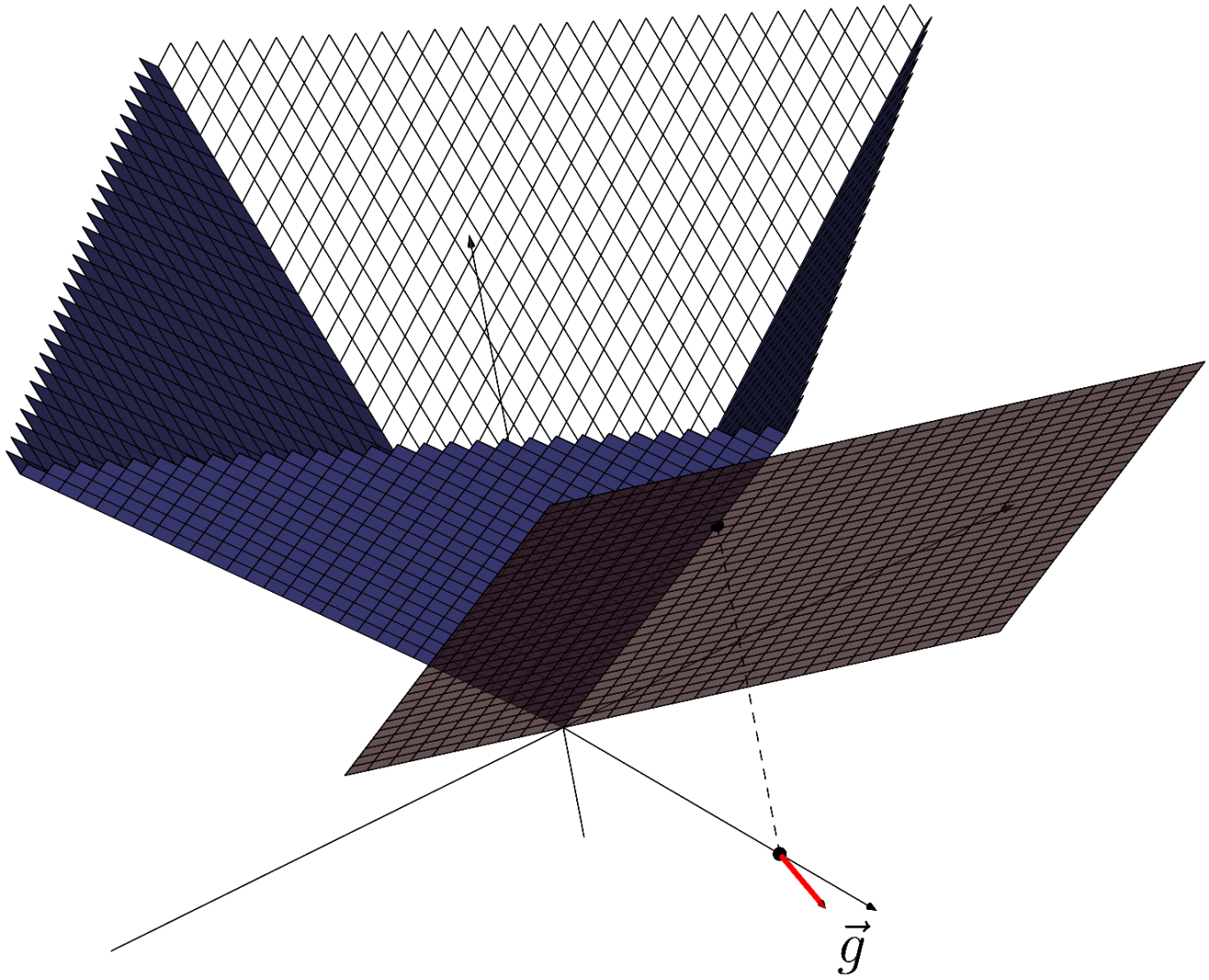
Subdifferential of ℓ_1 norm



Subdifferential of ℓ_1 norm



Subdifferential of ℓ_1 norm



Subdifferential of the nuclear norm

Let $X \in \mathbb{R}^{m \times n}$ be a rank-r matrix with SVD USV^T, where $U \in \mathbb{R}^{m \times r}$, $V \in \mathbb{R}^{n \times r}$ and $S \in \mathbb{R}^{r \times r}$

A matrix G is a subgradient of the nuclear norm at X if and only if

$$G := UV^T + W$$

where W satisfies

$$\begin{aligned} ||W|| &\leq 1 \\ U^T W &= 0 \\ W V &= 0 \end{aligned}$$

Proof

Analogy with ℓ_1 norm

- ℓ_1 norm $\rightarrow$ nuclear norm
- ℓ_∞ norm $\rightarrow$ operator norm
- $UV^T \rightarrow$ sign vector

Proof

First step: prove that $||G|| \leq 1$

By Pythagoras' Theorem, for any $\vec{x} \in \mathbb{R}^m$ with unit ℓ_2 norm we have

$$\left|\left|\mathcal{P}_{\mathsf{row}(X)}\,\vec{x}\right|\right|_2^2 + \left|\left|\mathcal{P}_{\mathsf{row}(X)^\perp}\,\vec{x}\right|\right|_2^2 = ||\vec{x}||_2^2 = 1$$

Proof

First step: prove that $||G|| \leq 1$

By Pythagoras' Theorem, for any $\vec{x} \in \mathbb{R}^m$ with unit ℓ_2 norm we have

$$\left|\left|\mathcal{P}_{\text{row}(X)}\,\vec{x}\right|\right|_2^2 + \left|\left|\mathcal{P}_{\text{row}(X)^\perp}\,\vec{x}\right|\right|_2^2 = ||\vec{x}||_2^2 = 1$$

The rows of UV^T are in $\text{row}\,(X)$ and the rows of W in $\text{row}\,(X)^\perp$, so

$$||G||^2 := \max_{\left\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\right\}} ||G\,\vec{x}||_2^2$$

Proof

First step: prove that $||G|| \leq 1$

By Pythagoras' Theorem, for any $\vec{x} \in \mathbb{R}^m$ with unit ℓ_2 norm we have

$$\left|\left|\mathcal{P}_{\text{row}(X)}\,\vec{x}\right|\right|_2^2 + \left|\left|\mathcal{P}_{\text{row}(X)^\perp}\,\vec{x}\right|\right|_2^2 = ||\vec{x}||_2^2 = 1$$

The rows of UV^T are in $\text{row}\,(X)$ and the rows of W in $\text{row}\,(X)^\perp$, so

$$||G||^2 := \max_{\left\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\right\}} ||G\,\vec{x}||_2^2$$

$$= \max_{\left\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\right\}} \left|\left|UV^T\,\vec{x}\right|\right|_2^2 + ||W\,\vec{x}||_2^2$$

Proof

First step: prove that $||G|| \leq 1$

By Pythagoras' Theorem, for any $\vec{x} \in \mathbb{R}^m$ with unit ℓ_2 norm we have

$$\left|\left|\mathcal{P}_{\text{row}(X)}\,\vec{x}\right|\right|_2^2 + \left|\left|\mathcal{P}_{\text{row}(X)^\perp}\,\vec{x}\right|\right|_2^2 = ||\vec{x}||_2^2 = 1$$

The rows of UV^T are in $\text{row}\,(X)$ and the rows of W in $\text{row}\,(X)^\perp$, so

$$\begin{aligned}
||G||^2 &:= \max_{\left\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\right\}} ||G\,\vec{x}||_2^2 \\
&= \max_{\left\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\right\}} \left|\left|UV^T\,\vec{x}\right|\right|_2^2 + ||W\,\vec{x}||_2^2 \\
&= \max_{\left\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\right\}} \left|\left|UV^T\,\mathcal{P}_{\text{row}(X)}\,\vec{x}\right|\right|_2^2 + \left|\left|W\,\mathcal{P}_{\text{row}(X)^\perp}\,\vec{x}\right|\right|_2^2
\end{aligned}$$

Proof

First step: prove that $||G|| \leq 1$

By Pythagoras' Theorem, for any $\vec{x} \in \mathbb{R}^m$ with unit ℓ_2 norm we have

$$\left|\left|\mathcal{P}_{\text{row}(X)}\,\vec{x}\right|\right|_2^2 + \left|\left|\mathcal{P}_{\text{row}(X)^\perp}\,\vec{x}\right|\right|_2^2 = ||\vec{x}||_2^2 = 1$$

The rows of UV^T are in $\text{row}\,(X)$ and the rows of W in $\text{row}\,(X)^\perp$, so

$$\begin{aligned}
||G||^2 &:= \max_{\left\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\right\}} ||G\,\vec{x}||_2^2 \\
&= \max_{\left\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\right\}} \left|\left|UV^T\,\vec{x}\right|\right|_2^2 + ||W\,\vec{x}||_2^2 \\
&= \max_{\left\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\right\}} \left|\left|UV^T\,\mathcal{P}_{\text{row}(X)}\,\vec{x}\right|\right|_2^2 + \left|\left|W\,\mathcal{P}_{\text{row}(X)^\perp}\,\vec{x}\right|\right|_2^2 \\
&\leq \left|\left|UV^T\right|\right|^2 \left|\left|\mathcal{P}_{\text{row}(X)}\,\vec{x}\right|\right|_2^2 + ||W||^2 \left|\left|\mathcal{P}_{\text{row}(X)^\perp}\,\vec{x}\right|\right|_2^2
\end{aligned}$$

Proof

First step: prove that $||G|| \leq 1$

By Pythagoras' Theorem, for any $\vec{x} \in \mathbb{R}^m$ with unit ℓ_2 norm we have

$$\left|\left|\mathcal{P}_{\text{row}(X)}\,\vec{x}\right|\right|_2^2 + \left|\left|\mathcal{P}_{\text{row}(X)^\perp}\,\vec{x}\right|\right|_2^2 = ||\vec{x}||_2^2 = 1$$

The rows of UV^T are in $\text{row}\,(X)$ and the rows of W in $\text{row}\,(X)^\perp$, so

$$\begin{aligned}
||G||^2 &:= \max_{\left\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\right\}} ||G\,\vec{x}||_2^2 \\
&= \max_{\left\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\right\}} \left|\left|UV^T\,\vec{x}\right|\right|_2^2 + ||W\,\vec{x}||_2^2 \\
&= \max_{\left\{||\vec{x}||_2=1 \mid \vec{x}\in\mathbb{R}^n\right\}} \left|\left|UV^T\,\mathcal{P}_{\text{row}(X)}\,\vec{x}\right|\right|_2^2 + \left|\left|W\,\mathcal{P}_{\text{row}(X)^\perp}\,\vec{x}\right|\right|_2^2 \\
&\leq \left|\left|UV^T\right|\right|^2 \left|\left|\mathcal{P}_{\text{row}(X)}\,\vec{x}\right|\right|_2^2 + ||W||^2 \left|\left|\mathcal{P}_{\text{row}(X)^\perp}\,\vec{x}\right|\right|_2^2 \\
&\leq 1
\end{aligned}$$

Hölder's inequality for matrices

Hölder's inequality:

$$||\vec{x}||_1 = \sup_{\left\{||\vec{y}||_\infty \leq 1 \mid \vec{y} \in \mathbb{R}^n\right\}} \langle \vec{x}, \vec{y} \rangle$$

For any matrix $A \in \mathbb{R}^{m \times n}$,

$$||A||_* = \sup_{\{||B|| \leq 1 \mid B \in \mathbb{R}^{m \times n}\}} \langle A, B \rangle$$

Proof

For any matrix $Y \in \mathbb{R}^{m \times n}$

$$\begin{aligned} ||Y||_* &\geq \langle G, Y \rangle \\ &= \langle G, X \rangle + \langle G, Y - X \rangle \\ &= \left\langle UV^T, X \right\rangle + \langle W, X \rangle + \langle G, Y - X \rangle \end{aligned}$$

Proof

$$U^T W = 0 \text{ implies } \langle W, X \rangle = \left\langle W, USV^T \right\rangle = \left\langle U^T W, SV^T \right\rangle = 0$$

Proof

For any matrix $Y \in \mathbb{R}^{m \times n}$

$$\begin{aligned} ||Y||_* &\geq \langle G, Y \rangle \\ &= \langle G, X \rangle + \langle G, Y - X \rangle \\ &= \left\langle UV^T, X \right\rangle + \langle W, X \rangle + \langle G, Y - X \rangle \\ &= \left\langle UV^T, X \right\rangle + \langle G, Y - X \rangle \end{aligned}$$

Proof

$$\left\langle UV^T, X \right\rangle$$

Proof

$$\left\langle UV^T, X \right\rangle = \mathsf{tr}\left(VU^T X\right)$$

Proof

$$\begin{aligned}\left\langle UV^T, X \right\rangle &= \mathsf{tr}\left(VU^T X\right) \\ &= \mathsf{tr}\left(VU^T USV^T\right)\end{aligned}$$

Proof

$$\begin{aligned}\left\langle UV^T, X \right\rangle &= \mathsf{tr}\left(VU^T X\right) \\ &= \mathsf{tr}\left(VU^T USV^T\right) \\ &= \mathsf{tr}\left(V^T V\, S\right)\end{aligned}$$

Proof

$$\begin{aligned}
\left\langle UV^T, X \right\rangle &= \mathrm{tr}\left(VU^T X\right) \\
&= \mathrm{tr}\left(VU^T USV^T\right) \\
&= \mathrm{tr}\left(V^T V\, S\right) \\
&= \mathrm{tr}\,(S)
\end{aligned}$$

Proof

$$
\begin{aligned}
\left\langle UV^T, X \right\rangle &= \mathrm{tr}\left(VU^T X\right) \\
&= \mathrm{tr}\left(VU^T USV^T\right) \\
&= \mathrm{tr}\left(V^T V\, S\right) \\
&= \mathrm{tr}\left(S\right) \\
&= ||X||_*
\end{aligned}
$$

Proof

For any matrix $Y \in \mathbb{R}^{m \times n}$

$$
\begin{aligned}
||Y||_* &\geq \langle G, Y \rangle \\
&= \langle G, X \rangle + \langle G, Y - X \rangle \\
&= \left\langle UV^T, X \right\rangle + \langle G, Y - X \rangle \\
&= \left\langle UV^T, X \right\rangle + \langle W, X \rangle + \langle G, Y - X \rangle \\
&= ||X||_* + \langle G, Y - X \rangle
\end{aligned}
$$

Motivating applications

Convex functions

Optimality conditions

Convex Regularization

Optimization algorithms

Sparse regression

Linear regression is challenging when the number of features p is large

Problem: How to choose subset of features $\mathcal{I} \subset \{1, \ldots, p\}$, such that

$$y \approx \sum_{i \in \mathcal{I}} \vec{\beta}[i] \vec{x}[i] + \beta_0$$

Equivalently, find sparse coefficient vector $\vec{\beta} \in \mathbb{R}^p$ such that

$$y \approx \left\langle \vec{x}, \vec{\beta} \right\rangle + \beta_0$$

Sparse linear regression

Find a small subset of useful features

Model selection problem

Two objectives:

- Good fit to the data; $\left|\left|X\vec{\beta} - \vec{y}\right|\right|_2^2$ should be as small as possible
- Using a small number of features; $\vec{\beta}$ should be as sparse as possible

The lasso

Uses ℓ_1-norm regularization to promote sparse coefficients

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

Nonnegative weighted sums

The weighted sum of m convex functions $f_1, \ldots, f_m$

$$f := \sum_{i=1}^{m} \alpha_i f_i$$

is convex if $\alpha_1, \ldots, \alpha \in \mathbb{R}$ are nonnegative

Nonnegative weighted sums

The weighted sum of m convex functions $f_1, \ldots, f_m$

$$f := \sum_{i=1}^{m} \alpha_i f_i$$

is convex if $\alpha_1, \ldots, \alpha \in \mathbb{R}$ are nonnegative

Proof:

$$f\left(\theta \vec{x} + (1-\theta)\, \vec{y}\right)$$

Nonnegative weighted sums

The weighted sum of m convex functions $f_1, \ldots, f_m$

$$f := \sum_{i=1}^{m} \alpha_i f_i$$

is convex if $\alpha_1, \ldots, \alpha \in \mathbb{R}$ are nonnegative

Proof:

$$f\left(\theta \vec{x} + (1-\theta)\vec{y}\right) = \sum_{i=1}^{m} \alpha_i f_i\left(\theta \vec{x} + (1-\theta)\vec{y}\right)$$

Nonnegative weighted sums

The weighted sum of m convex functions $f_1, \ldots, f_m$

$$f := \sum_{i=1}^{m} \alpha_i f_i$$

is convex if $\alpha_1, \ldots, \alpha \in \mathbb{R}$ are nonnegative

Proof:

$$\begin{aligned} f\left(\theta \vec{x} + (1-\theta)\vec{y}\right) &= \sum_{i=1}^{m} \alpha_i f_i\left(\theta \vec{x} + (1-\theta)\vec{y}\right) \\ &\leq \sum_{i=1}^{m} \alpha_i \left(\theta f_i\left(\vec{x}\right) + (1-\theta) f_i\left(\vec{y}\right)\right) \end{aligned}$$

Nonnegative weighted sums

The weighted sum of m convex functions $f_1, \ldots, f_m$

$$f := \sum_{i=1}^{m} \alpha_i f_i$$

is convex if $\alpha_1, \ldots, \alpha \in \mathbb{R}$ are nonnegative

Proof:

$$\begin{aligned} f\left(\theta \vec{x} + (1-\theta)\vec{y}\right) &= \sum_{i=1}^{m} \alpha_i f_i\left(\theta \vec{x} + (1-\theta)\vec{y}\right) \\ &\leq \sum_{i=1}^{m} \alpha_i \left(\theta f_i\left(\vec{x}\right) + (1-\theta) f_i\left(\vec{y}\right)\right) \\ &= \theta f\left(\vec{x}\right) + (1-\theta) f\left(\vec{y}\right) \end{aligned}$$

Regularized least-squares

Regularized least-squares cost functions

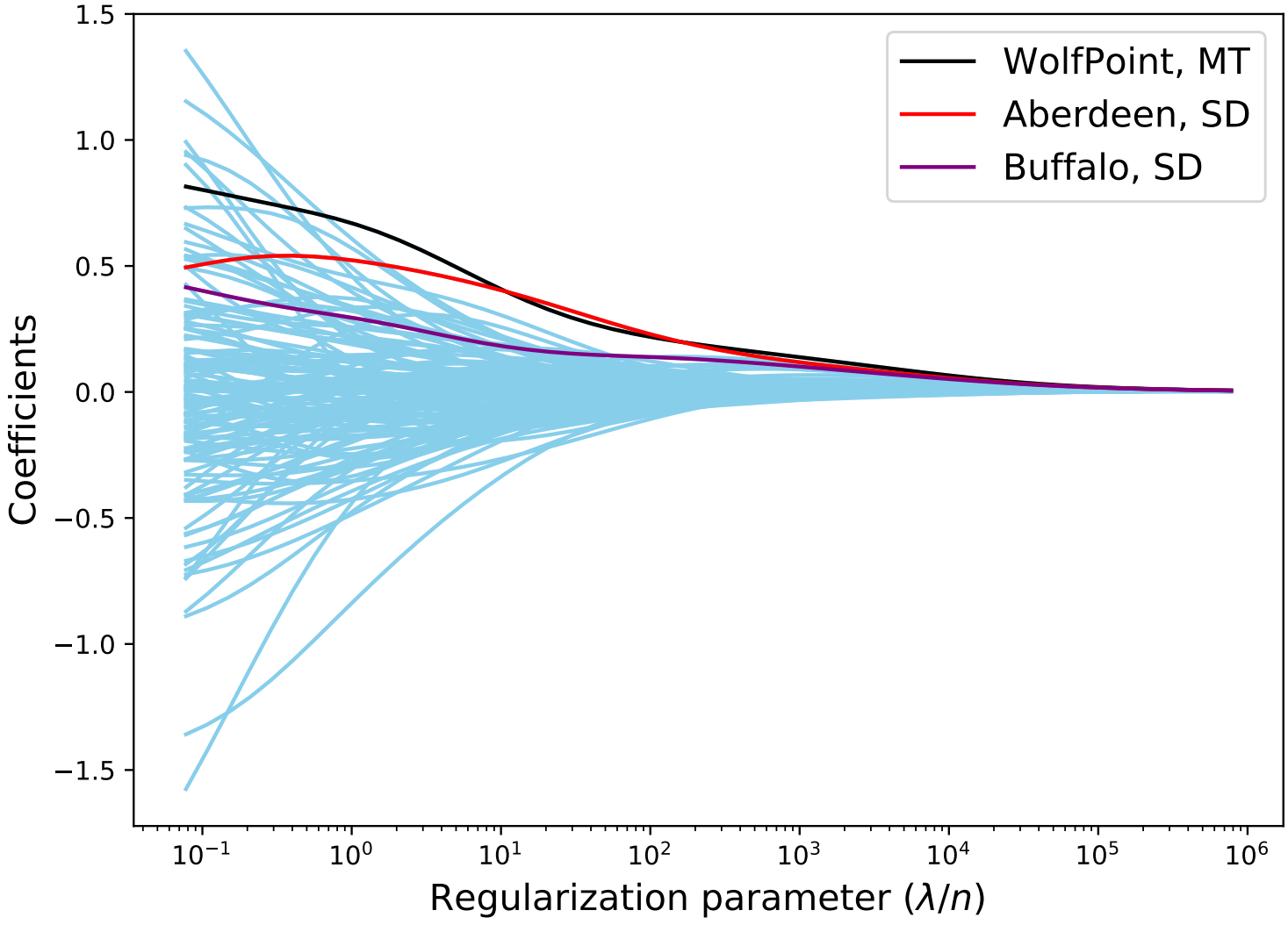
$$||A\vec{x} - \vec{y}||_2^2 + ||\vec{x}||$$

are convex

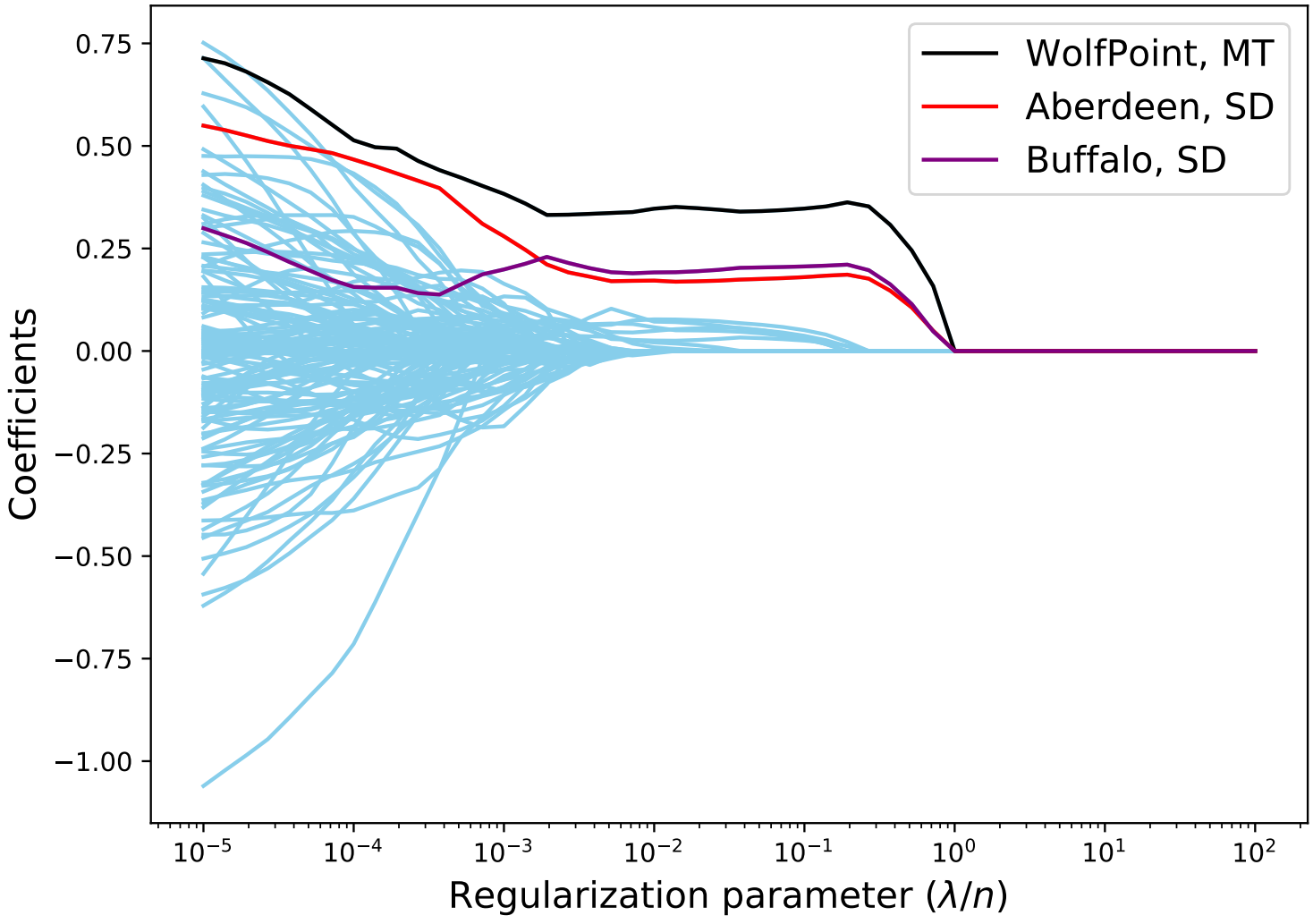
Temperature prediction via linear regression

- 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

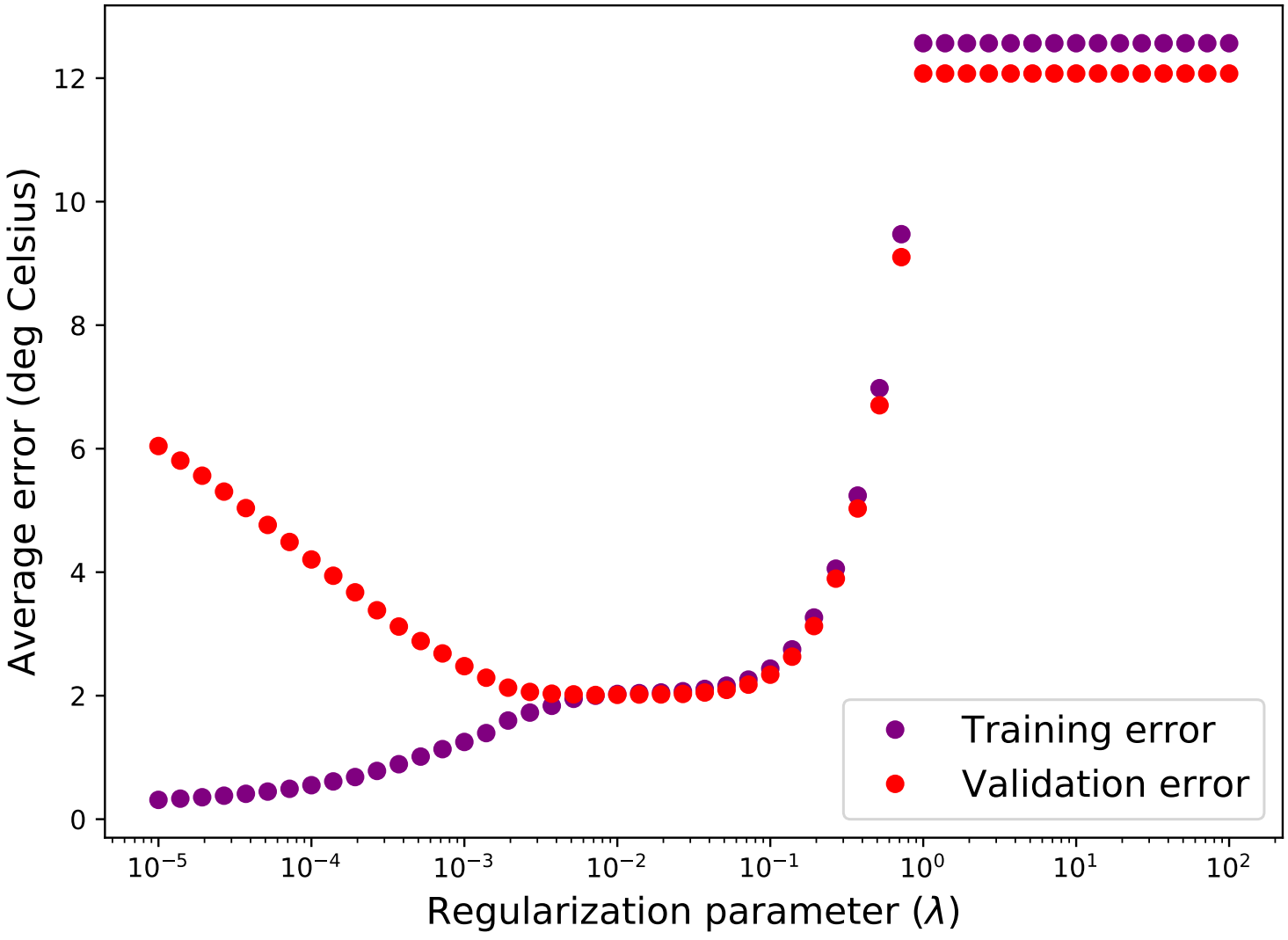
Ridge regression $n := 135$



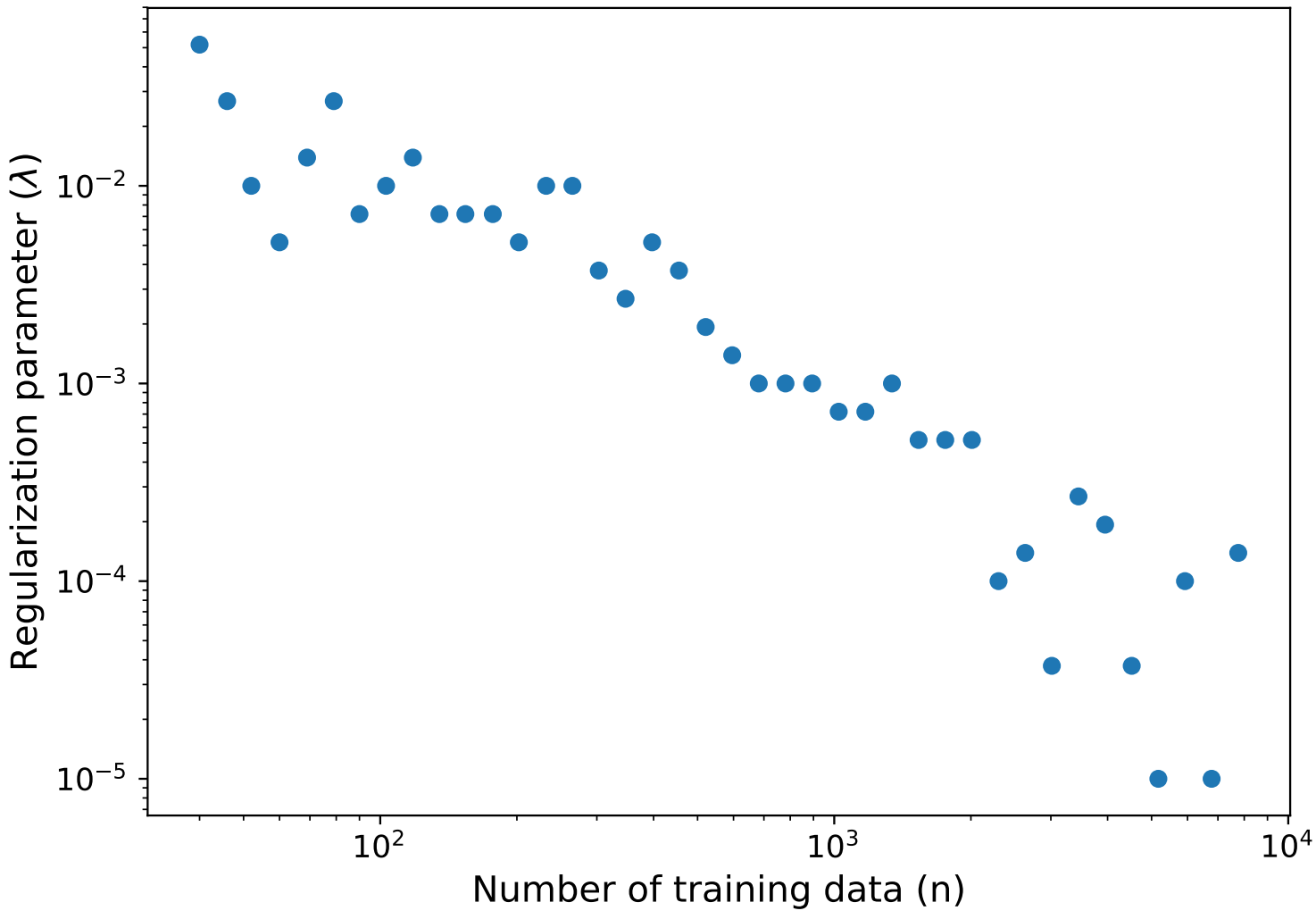
Lasso $n := 135$


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)

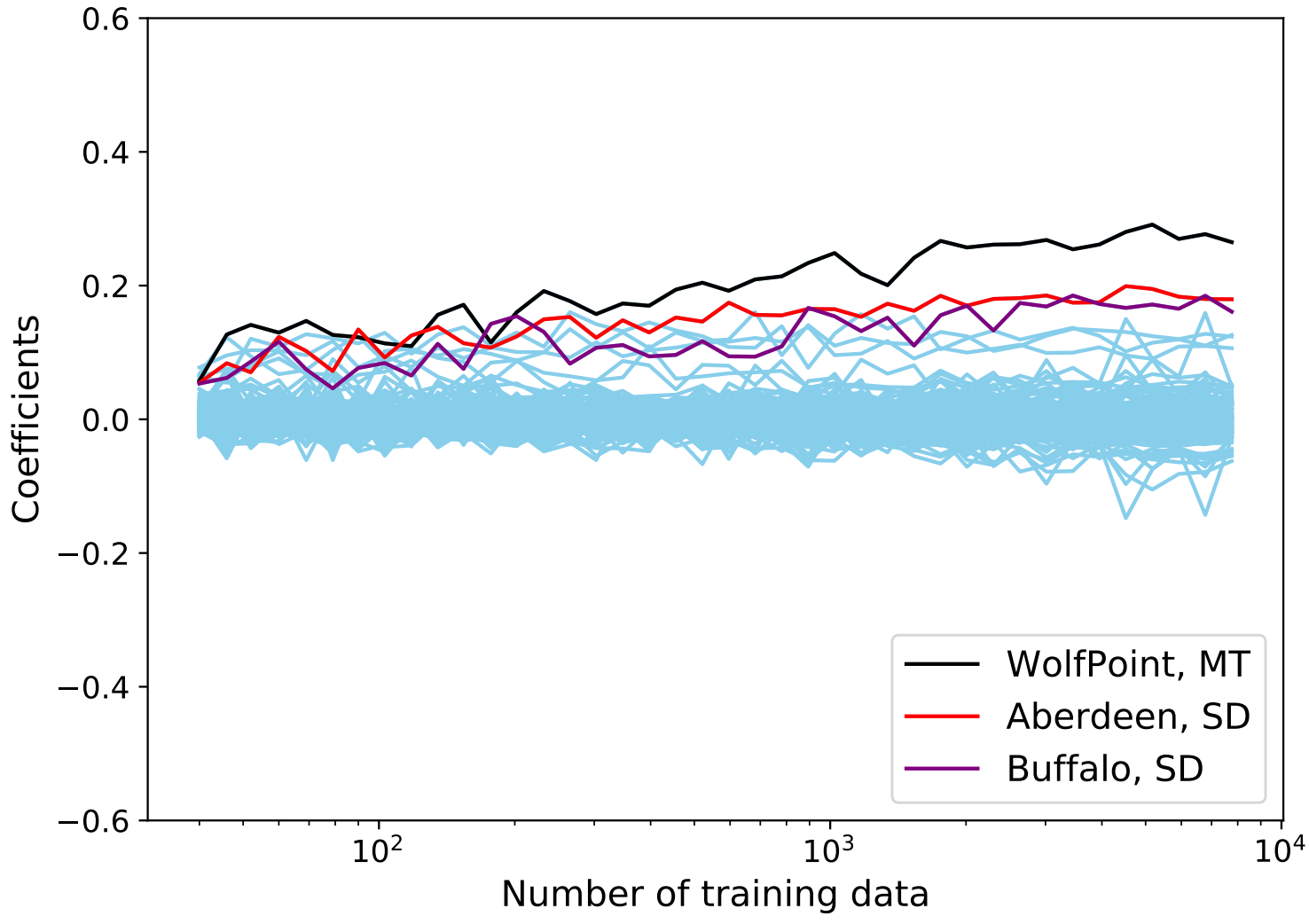
Lasso $n := 135$


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

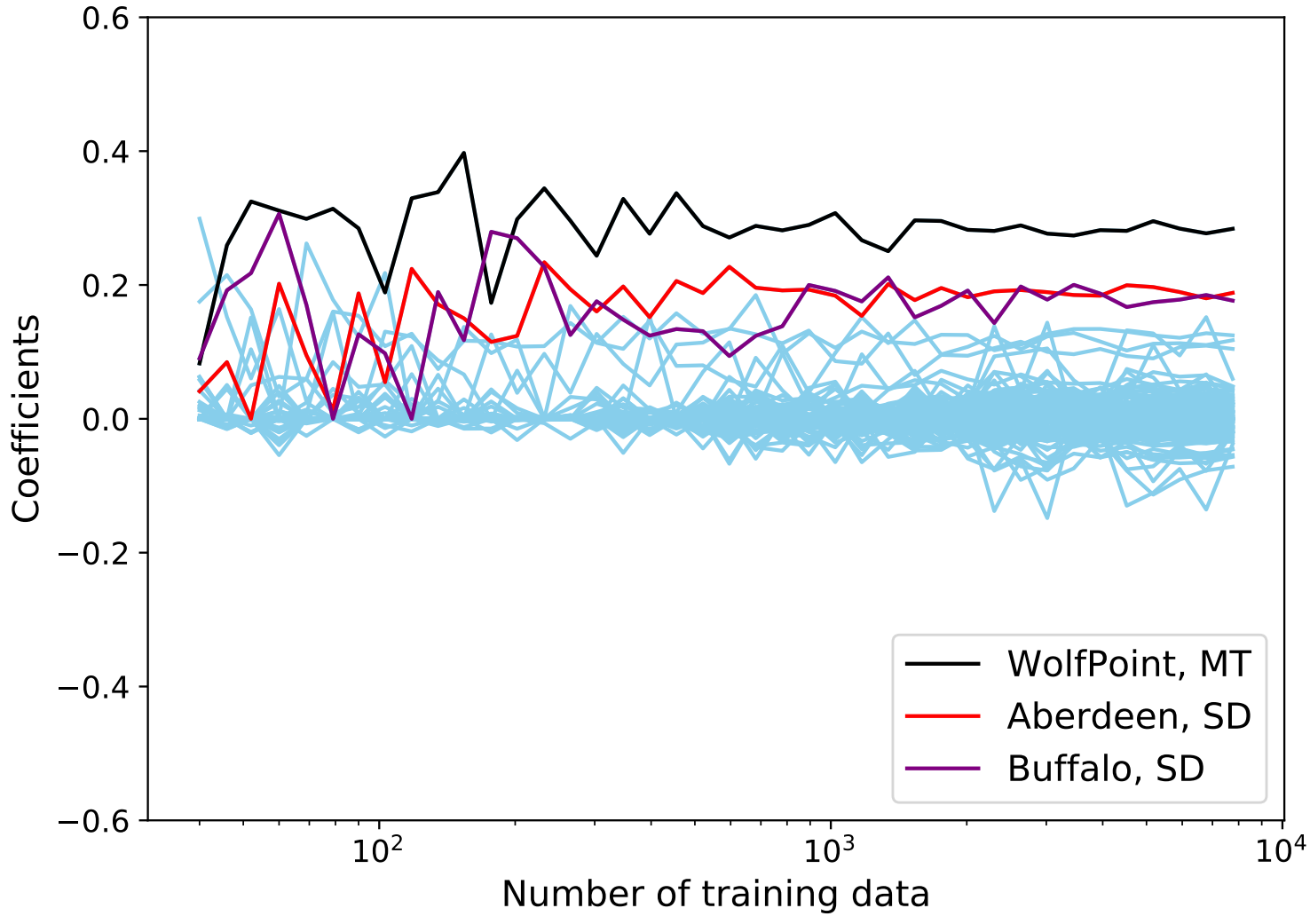
Lasso $n := 135$


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

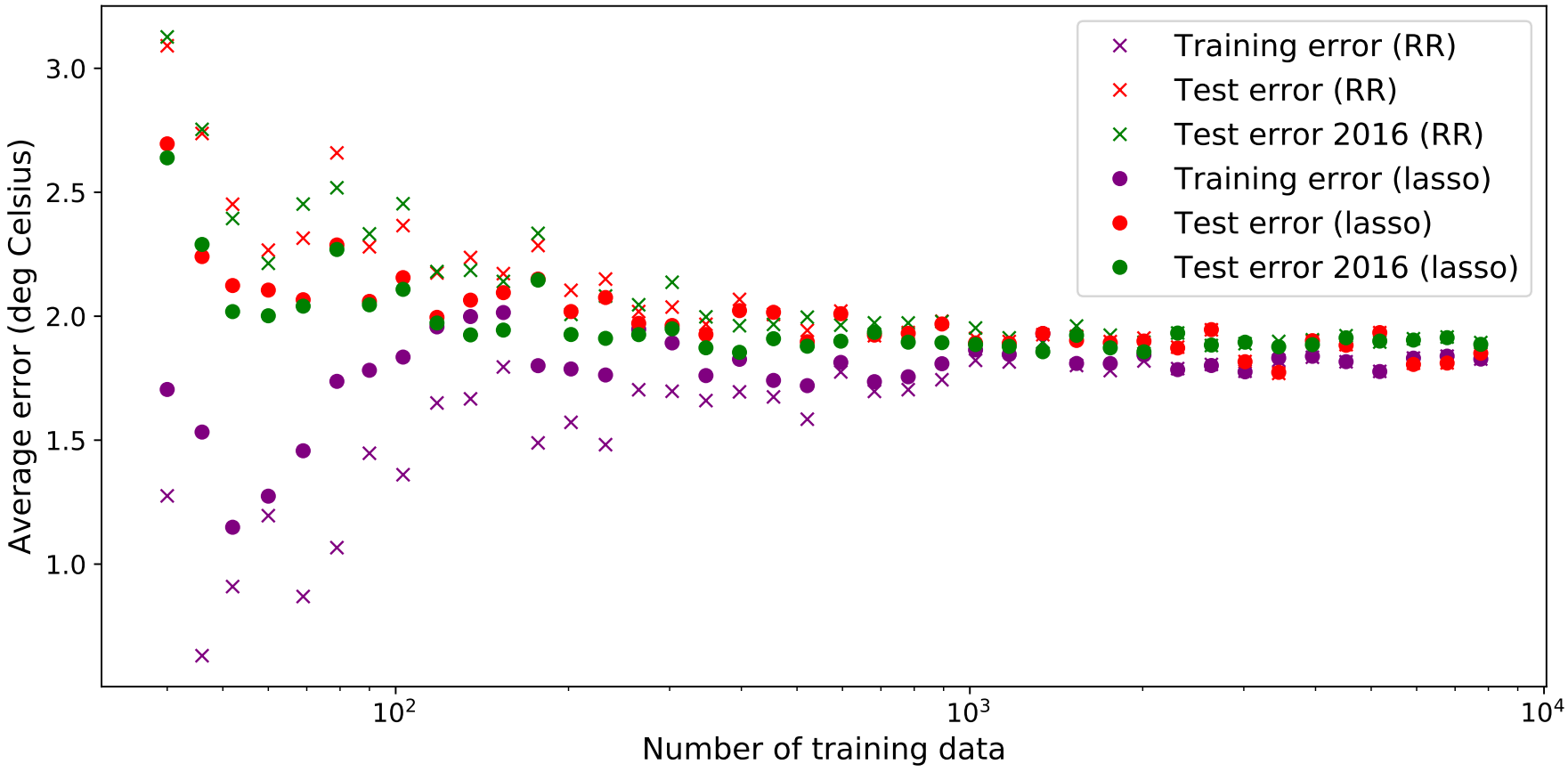
Ridge-regression coefficients



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

Proximal operator

Linear combination of function and least-squares term

$$\text{prox}_f\left(\vec{y}\right) := \arg\min_{\vec{x}} f\left(\vec{x}\right) + \frac{1}{2}\left|\left|\vec{x} - \vec{y}\right|\right|_2^2$$

Proximal operator of ℓ_2 norm

The proximal operator of the squared ℓ_2 norm is

$$\text{prox}_{\alpha ||\cdot||_2^2}(\vec{y}) = \frac{\vec{y}}{1 + 2\alpha}$$

for all $\alpha > 0$

Proof

The gradient of the function

$$f(\vec{x}) := \alpha \, ||\vec{x}||_2^2 + \frac{1}{2} \, ||\vec{x} - \vec{y}||_2^2$$
$$= \left(\frac{1}{2} + \alpha \right) \vec{x}^T \vec{x} + \frac{1}{2} \vec{y}^T \vec{y} - \vec{y}^T \vec{x}$$

equals

$$\nabla f(\vec{x}) =$$

Proof

The gradient of the function

$$f(\vec{x}) := \alpha \, ||\vec{x}||_2^2 + \frac{1}{2} \, ||\vec{x} - \vec{y}||_2^2$$
$$= \left(\frac{1}{2} + \alpha \right) \vec{x}^T \vec{x} + \frac{1}{2} \vec{y}^T \vec{y} - \vec{y}^T \vec{x}$$

equals

$$\nabla f(\vec{x}) = \, (1 + 2\alpha) \, \vec{x} - \vec{y}$$

Proximal operator of ℓ_1 norm

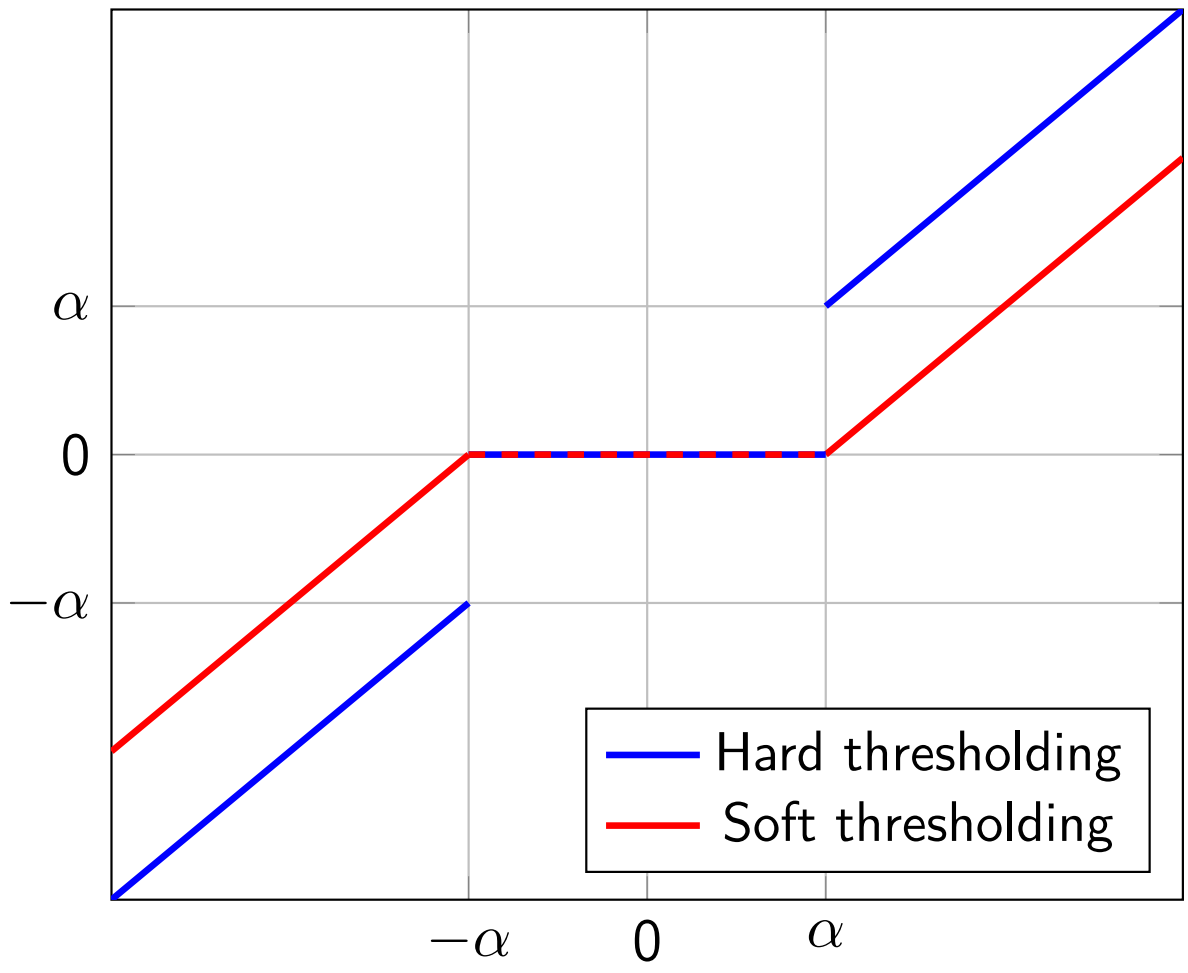
The proximal operator of the ℓ_1 norm is the soft-thresholding operator

$$\text{prox}_{\alpha \|\cdot\|_1}(y) = \mathcal{S}_\alpha(y)$$

where $\alpha > 0$ and

$$\mathcal{S}_\alpha(y)_i := \begin{cases} y_i - \text{sign}(y_i)\,\alpha & \text{if } |y_i| \geq \alpha \\ 0 & \text{otherwise} \end{cases}$$

Thresholding



Proof

The subgradients of the function

$$f(\vec{x}) := \alpha \left|\left|\vec{x}\right|\right|_1 + \frac{1}{2} \left|\left|\vec{x} - \vec{y}\right|\right|_2^2$$

at $\vec{x}$ equal

$$\vec{g}_{\text{prox}} = \alpha \vec{g}_{\ell_1}(\vec{x}) + \vec{x} - \vec{y}$$

where $\vec{g}_{\ell_1}(\vec{x})$ is any subgradient of the ℓ_1 norm at $\vec{x}$

Setting it to zero

$$\vec{x}_{\text{min}} + \alpha \vec{g}_{\ell_1}(\vec{x}_{\text{min}}) = \vec{y}$$

Proof

If $|\vec{y}[j]| \geq \alpha$

$$\vec{x}_{\mathsf{min}}[j] = \vec{y}[j] - \alpha \vec{g}_{\ell_1}(\vec{x}_{\mathsf{min}})[j]$$

Proof

If $|\vec{y}[j]| \geq \alpha$

$$\begin{aligned}\vec{x}_{\mathsf{min}}[j] &= \vec{y}[j] - \alpha \vec{g}_{\ell_1}(\vec{x}_{\mathsf{min}})[j] \\ &= \vec{y}[j] - \alpha \,\mathsf{sign}(\vec{x}[j])\end{aligned}$$

Proof

If $|\vec{y}[j]| \geq \alpha$

$$\begin{aligned} \vec{x}_{\text{min}}[j] &= \vec{y}[j] - \alpha \vec{g}_{\ell_1}(\vec{x}_{\text{min}})[j] \\ &= \vec{y}[j] - \alpha \, \text{sign}(\vec{x}[j]) \end{aligned}$$

If $|\vec{y}[j]| < \alpha$ the only possible solution is zero

Linear regression with orthonormal features

Let $U \in \mathbb{R}^{n \times p}$ be a matrix with orthonormal columns and $\vec{y} \in \mathbb{R}^n$

For any f

$$\arg\min_{\vec{\beta}} \frac{1}{2} \left|\left|\vec{y} - U\vec{\beta}\right|\right|_2^2 + f(\vec{\beta}) = \arg\min_{\vec{\beta}} \frac{1}{2} \left|\left|U^T\vec{y} - \vec{\beta}\right|\right|_2^2 + f(\vec{\beta}).$$

Proof

$$\frac{1}{2}\left|\left|\vec{y} - U\vec{\beta}\right|\right|_2^2 + f(\vec{\beta}) = \frac{1}{2}\vec{\beta}^T U^T U\vec{\beta} + \frac{1}{2}\vec{y}^T\vec{y} - \vec{y}^T U\vec{\beta} + f(\vec{\beta})$$

Proof

$$\begin{aligned}\frac{1}{2}\left|\left|\vec{y}-U\vec{\beta}\right|\right|_2^2+f(\vec{\beta}) &= \frac{1}{2}\vec{\beta}^T U^T U\vec{\beta}+\frac{1}{2}\vec{y}^T\vec{y}-\vec{y}^T U\vec{\beta}+f(\vec{\beta}) \\ &= \frac{1}{2}\vec{\beta}^T\vec{\beta}+\frac{1}{2}\vec{y}^T\vec{y}-(U^T\vec{y})^T\vec{\beta}+f(\vec{\beta})\end{aligned}$$

Proof

$$\begin{aligned}
\frac{1}{2}\left|\left|\vec{y} - U\vec{\beta}\right|\right|_2^2 + f(\vec{\beta}) &= \frac{1}{2}\vec{\beta}^T U^T U\vec{\beta} + \frac{1}{2}\vec{y}^T\vec{y} - \vec{y}^T U\vec{\beta} + f(\vec{\beta}) \\
&= \frac{1}{2}\vec{\beta}^T\vec{\beta} + \frac{1}{2}\vec{y}^T\vec{y} - (U^T\vec{y})^T\vec{\beta} + f(\vec{\beta}) \\
&= \frac{1}{2}\left|\left|U^T\vec{y} - \vec{\beta}\right|\right|_2^2 + f(\vec{\beta}) + \frac{1}{2}\vec{y}^T\vec{y} - \frac{1}{2}\vec{y}^T UU^T\vec{y}
\end{aligned}$$

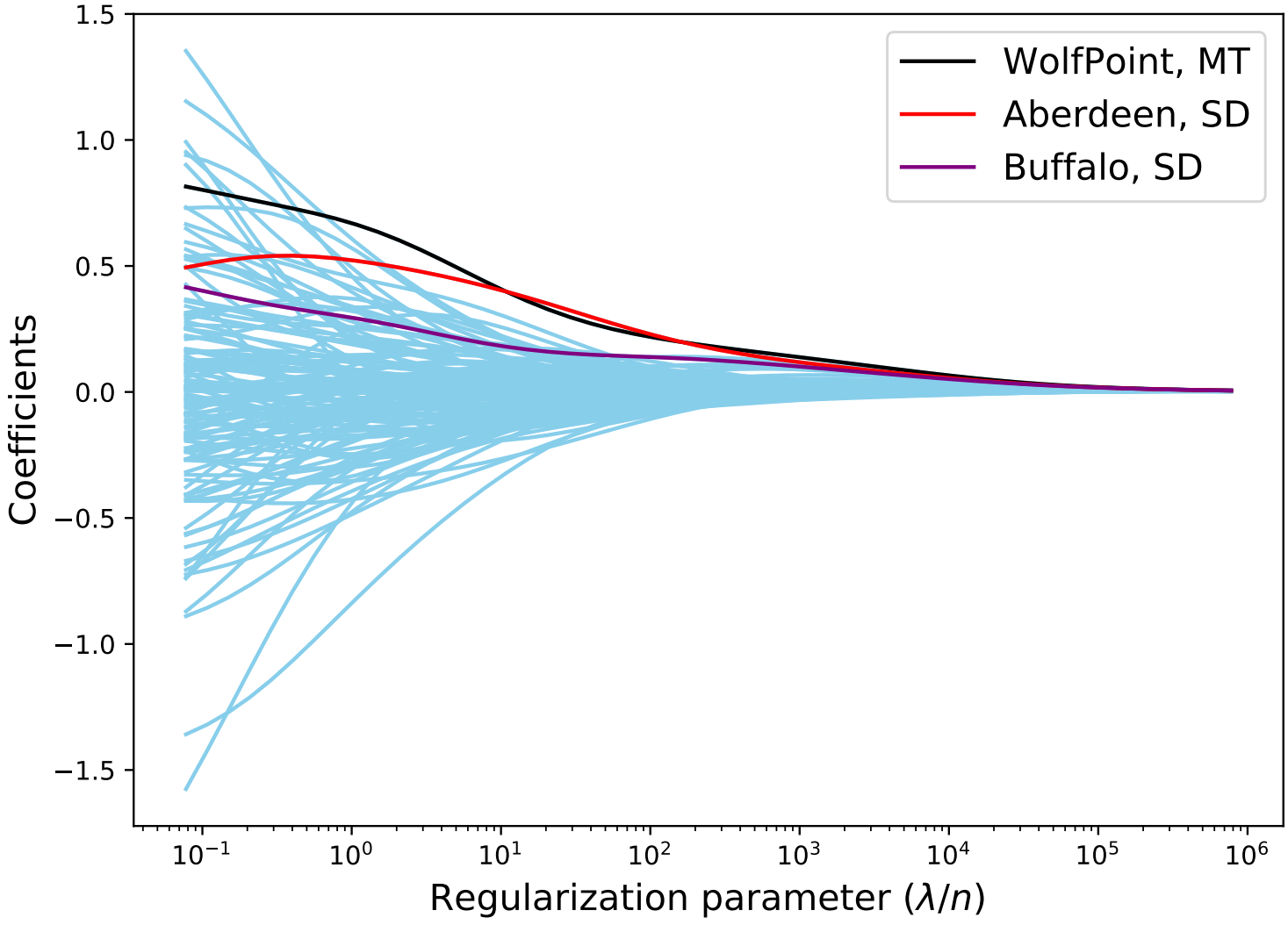
Consequence

$U^T \vec{y}$ is the least-squares solution

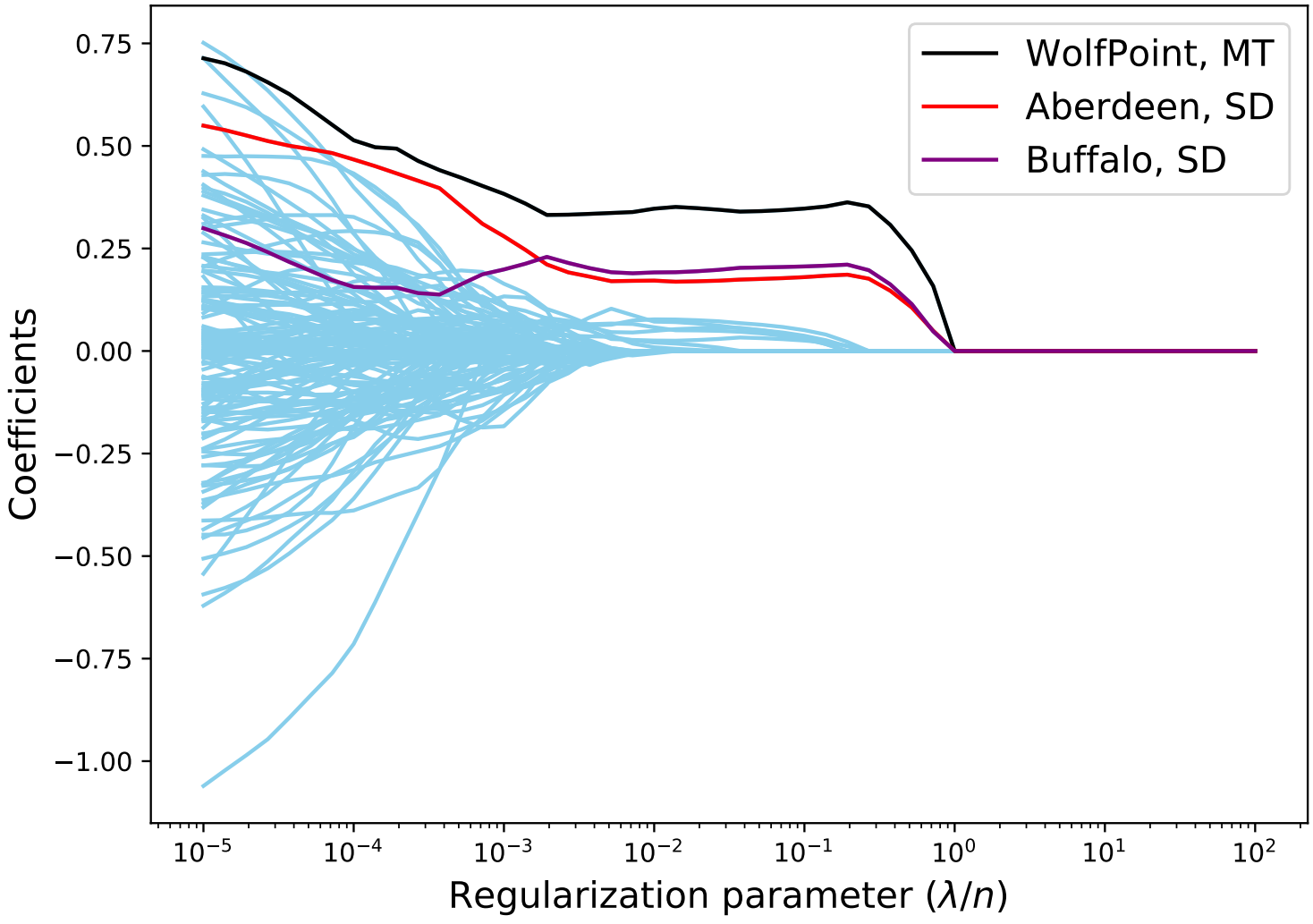
Ridge regression scales all entries equally

The lasso soft-thresholds the entries

Ridge regression $n := 135$



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)

Collaborative filtering

Low-rank matrix estimation

If all entries are observed, truncating the SVD is optimal

Let USV^T be the SVD of a matrix $A \in \mathbb{R}^{m \times n}$

The truncated SVD $U_{:,1:r} S_{1:r,1:r} V_{:,1:r}^T$ is the best rank-r approximation

$$U_{:,1:r} S_{1:r,1:r} V_{:,1:r}^T = \underset{\left\{\widetilde{A} \mid \text{ rank}\left(\widetilde{A}\right)=r\right\}}{\arg\min} \left|\left|A - \widetilde{A}\right|\right|_F$$

Low-rank matrix estimation

First idea:

$$\min_{X \in \mathbb{R}^{m \times n}} \operatorname{rank}(X) \quad \text{such that } X_\Omega = \vec{y}$$

Ω: indices of revealed entries

$\vec{y}$: revealed entries

Low-rank matrix estimation

First idea:

$$\min_{X \in \mathbb{R}^{m \times n}} \operatorname{rank}(X) \quad \text{such that } X_\Omega = \vec{y}$$

Ω: indices of revealed entries
$\vec{y}$: revealed entries

Computationally intractable because of missing entries

Low-rank matrix estimation

First idea:

$$\min_{X \in \mathbb{R}^{m \times n}} \operatorname{rank}(X) \quad \text{such that } X_\Omega = \vec{y}$$

Ω: indices of revealed entries
$\vec{y}$: revealed entries

Computationally intractable because of missing entries

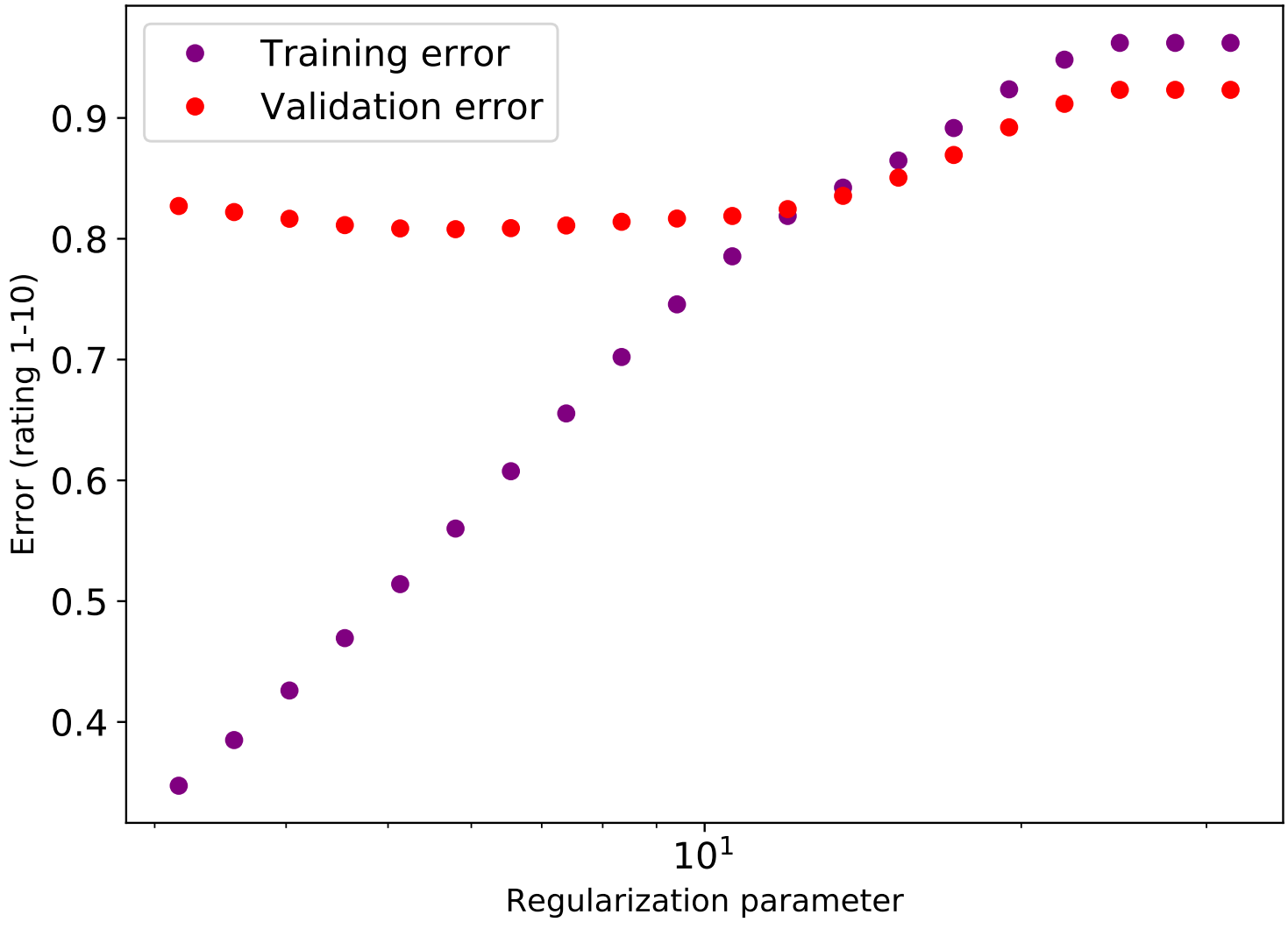
Tractable alternative:

$$\min_{X \in \mathbb{R}^{n_1 \times n_2}} \frac{1}{2} ||X_\Omega - \vec{y}||_2^2 + \lambda ||X||_*$$

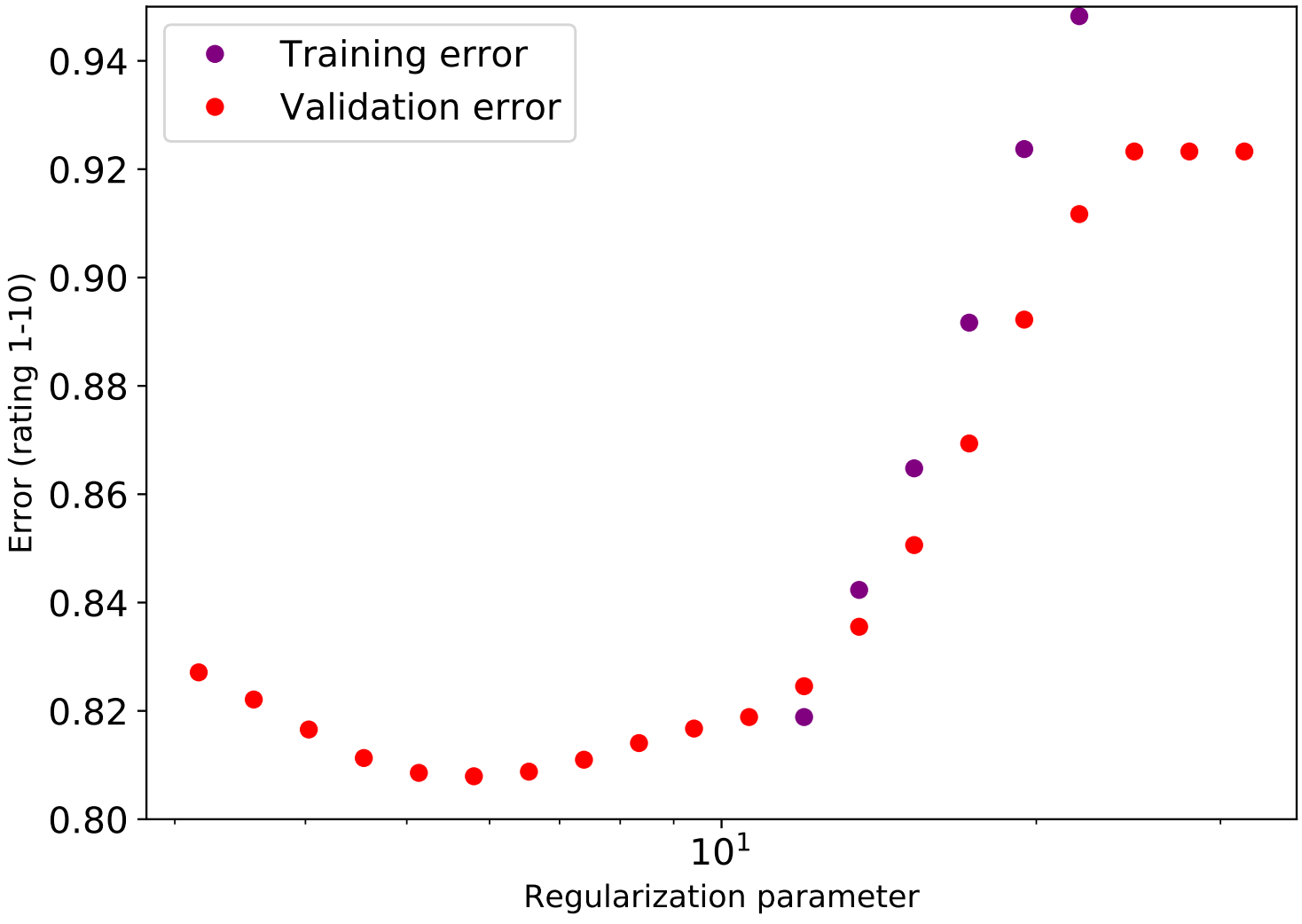
Collaborative filtering

- Movielens data
- Ratings between 1 and 10
- 100 users and movies with more ratings
- 6,031 out of 10^4 ratings are observed
- Test set with 10^3
- Validation set with $\max\{n_{\text{train}}, 400\}$

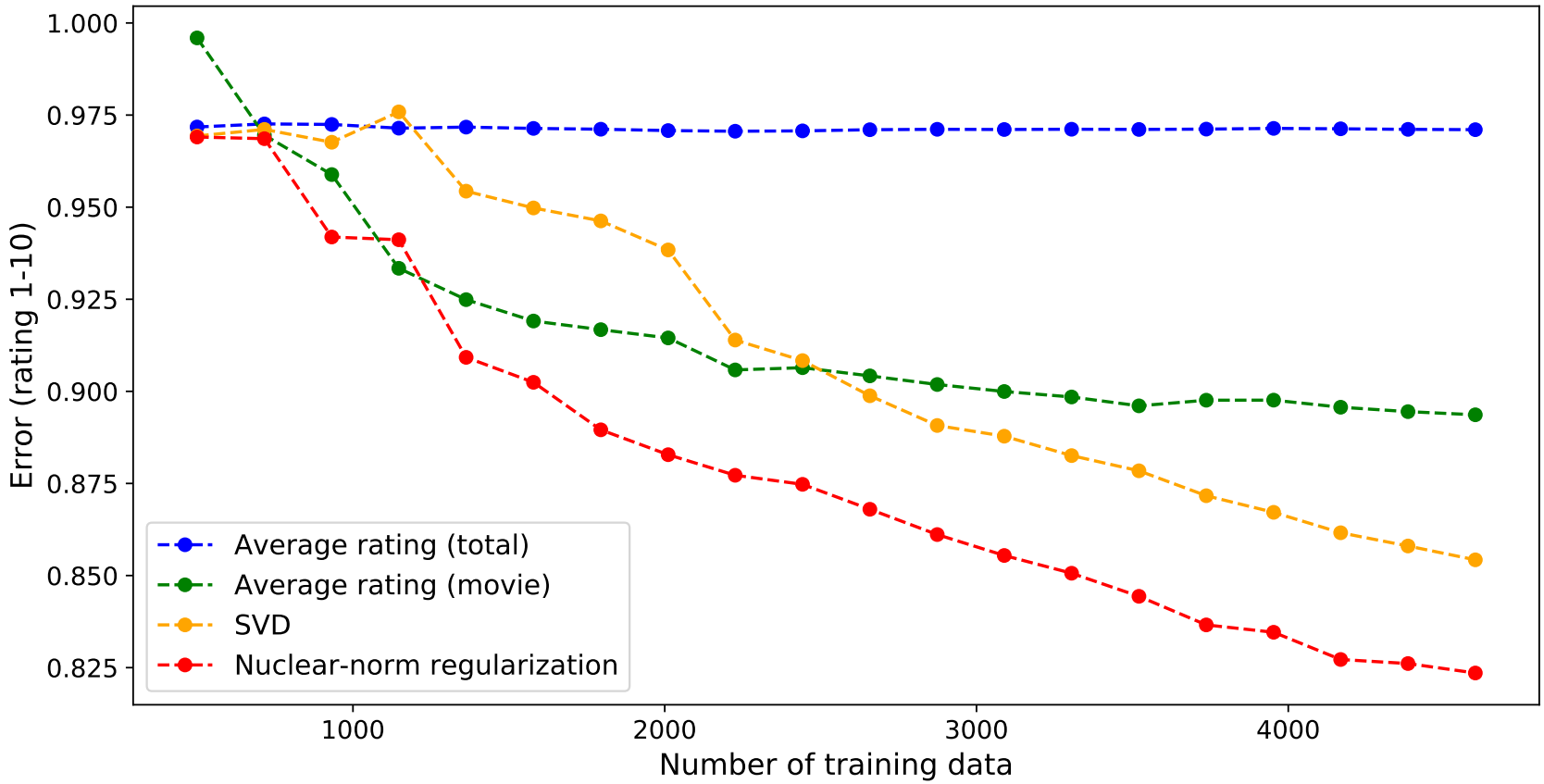
$n_{\text{train}} := 4{,}600$



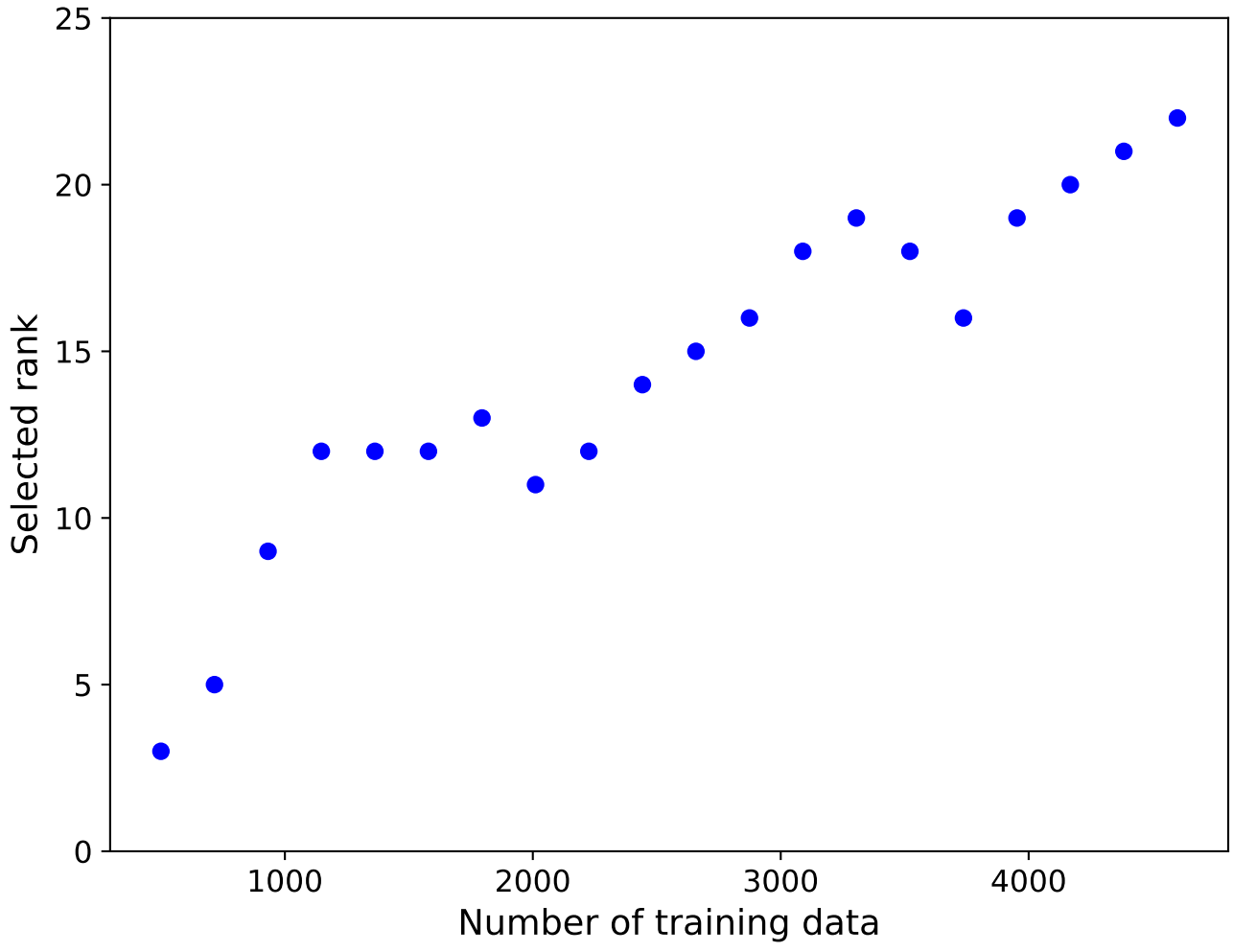
$n_{\text{train}} := 4{,}600$



Results


1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
Error (rating 1-10)
1000
2000
3000
4000
Number of training data
Average rating (total)
Average rating (movie)
SVD
Nuclear-norm regularization

Rank


25
20
15
10
5
0
Selected rank
1000
2000
3000
4000
Number of training data

Proximal operator of nuclear norm

The solution X to

$$\min_{X \in \mathbb{R}^{m \times n}} \frac{1}{2} ||Y - X||_F^2 + \alpha ||X||_*$$

is obtained by soft-thresholding the SVD of Y

$$X_{\text{prox}} = \mathcal{D}_\alpha (Y)$$

$$\mathcal{D}_\alpha (M) := U \mathcal{S}_\alpha (S) V^T \qquad \text{where } M = U S V^T$$

$$\mathcal{S}_\alpha (S)_{ii} := \begin{cases} S_{ii} - \alpha & \text{if } S_{ii} > \alpha \\ 0 & \text{otherwise} \end{cases}$$

Proximal operator of nuclear norm

The subgradients of

$$\frac{1}{2}\left|\left|X-Y\right|\right|_{\mathrm{F}}^{2}+\alpha\left|\left|X\right|\right|_{*}$$

are of the form

$$X-Y+\alpha G$$

where G is a subgradient of the nuclear norm at X

$\mathcal{D}_{\alpha}\left(Y\right)$ is a minimizer if and only if

$$G=\frac{1}{\alpha}\left(Y-D_{\alpha}\left(Y\right)\right)$$

is a subgradient of the nuclear norm at $\mathcal{D}_{\alpha}\left(Y\right)$

Subdifferential of the nuclear norm

Let $X \in \mathbb{R}^{m \times n}$ be a rank-r matrix with SVD USV^T, where $U \in \mathbb{R}^{m \times r}$, $V \in \mathbb{R}^{n \times r}$ and $S \in \mathbb{R}^{r \times r}$

A matrix G is a subgradient of the nuclear norm at X if and only if

$$G := UV^T + W$$

where W satisfies

$$\begin{aligned} ||W|| &\leq 1 \\ U^T W &= 0 \\ W V &= 0 \end{aligned}$$

Proximal operator of nuclear norm

Separate SVD of Y into singular values greater or smaller than α

$$Y = U\,SV^T$$

$$= \begin{bmatrix} U_0 & U_1 \end{bmatrix} \begin{bmatrix} S_0 & 0 \\ 0 & S_1 \end{bmatrix} \begin{bmatrix} V_0 & V_1 \end{bmatrix}^T$$

$D_\alpha(Y) = U_0 (S_0 - \alpha I) V_0^T$, so

$$Y - D_\alpha(Y) =$$

Proximal operator of nuclear norm

Separate SVD of Y into singular values greater or smaller than α

$$\begin{aligned} Y &= U\,SV^T \\ &= \begin{bmatrix} U_0 & U_1 \end{bmatrix} \begin{bmatrix} S_0 & 0 \\ 0 & S_1 \end{bmatrix} \begin{bmatrix} V_0 & V_1 \end{bmatrix}^T \end{aligned}$$

$D_\alpha(Y) = U_0 (S_0 - \alpha I) V_0^T$, so

$$Y - D_\alpha(Y) = \alpha U_0 V_0^T + U_1 S_1 V_1^T$$

Proximal operator of nuclear norm

Separate SVD of Y into singular values greater or smaller than α

$$\begin{aligned} Y &= U S V^T \\ &= \begin{bmatrix} U_0 & U_1 \end{bmatrix} \begin{bmatrix} S_0 & 0 \\ 0 & S_1 \end{bmatrix} \begin{bmatrix} V_0 & V_1 \end{bmatrix}^T \end{aligned}$$

$D_\alpha(Y) = U_0 (S_0 - \alpha I) V_0^T$, so

$$\begin{aligned} Y - D_\alpha(Y) &= \alpha U_0 V_0^T + U_1 S_1 V_1^T \\ \frac{1}{\alpha}(Y - D_\alpha(Y)) &= U_0 V_0^T + \frac{1}{\alpha} U_1 S_1 V_1^T \end{aligned}$$

Motivating applications

Convex functions

Optimality conditions

Convex Regularization

Optimization algorithms

Problem

Aim: Minimizing differentiable convex functions

$$\min_{\vec{x} \in \mathbb{R}^n} \quad f\left(\vec{x}\right)$$

Gradient descent

Intuition: Make local progress in the steepest direction $-\nabla f(\vec{x})$

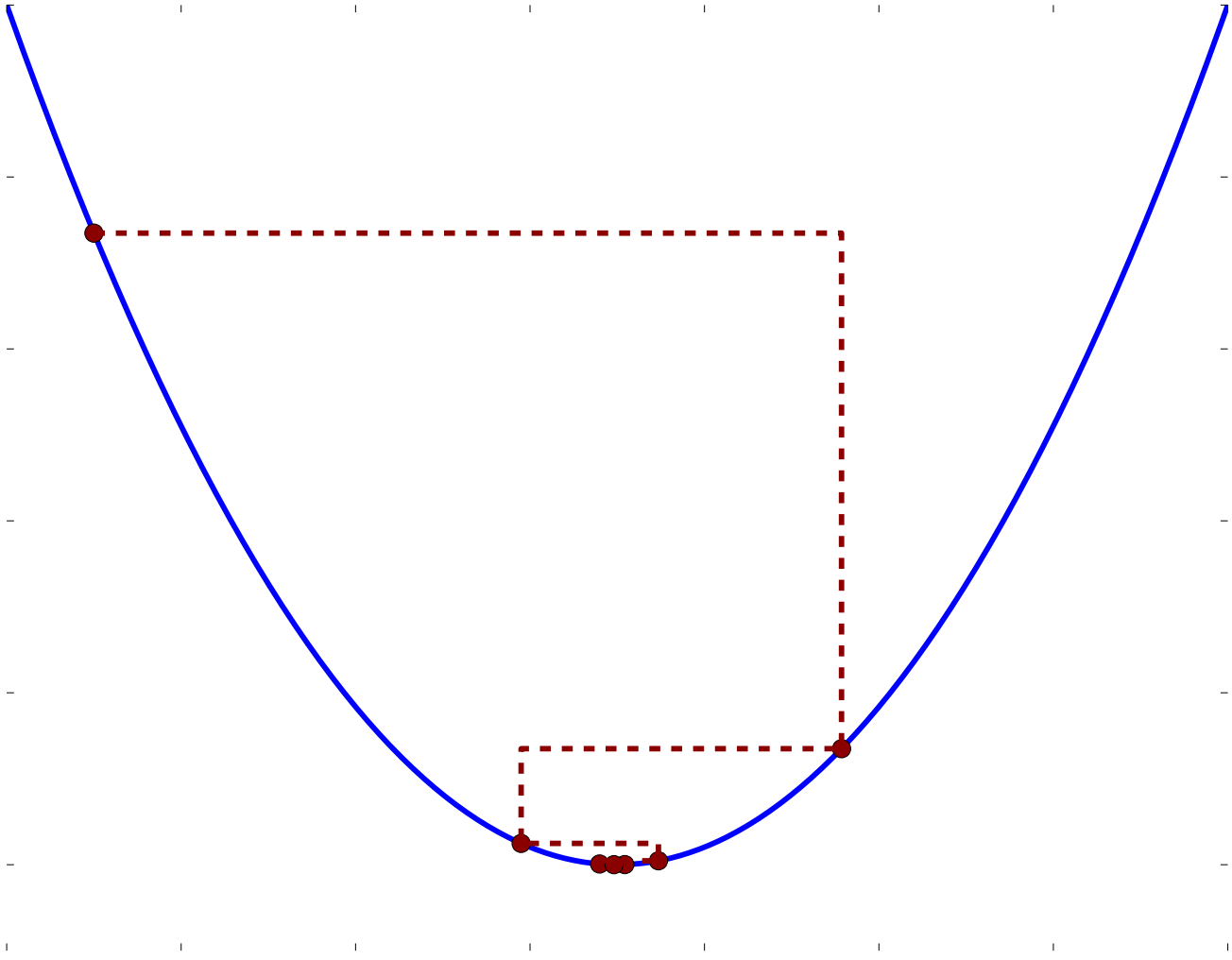
Set the initial point $\vec{x}^{(0)}$ to an arbitrary value

Update by setting

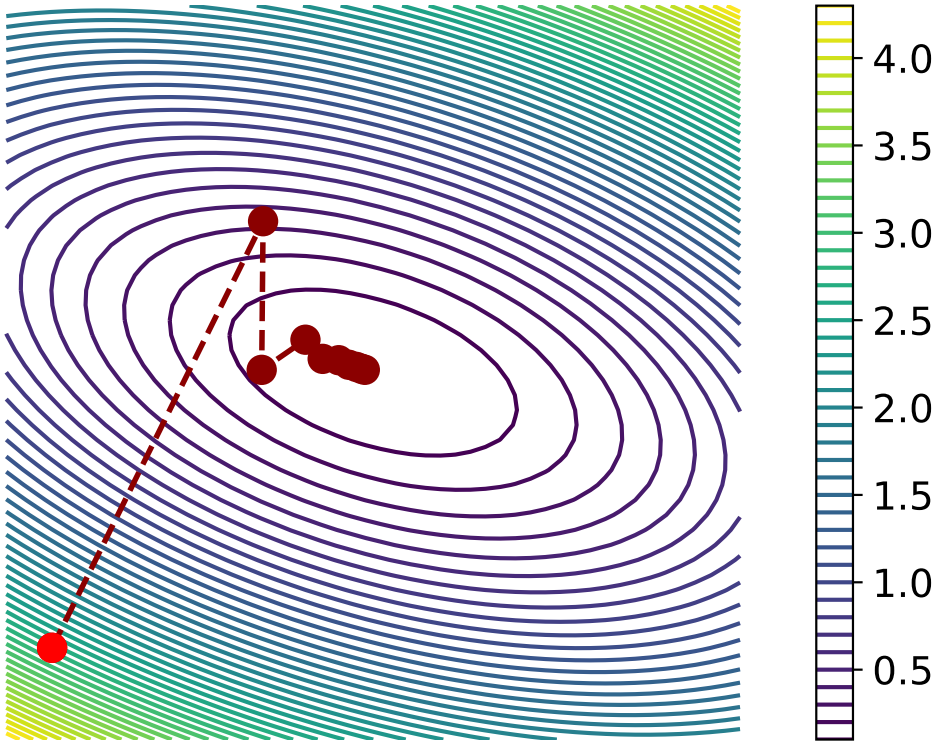
$$\vec{x}^{(k+1)} := \vec{x}^{(k)} - \alpha_k \nabla f\left(\vec{x}^{(k)}\right)$$

where $\alpha_k > 0$ is the step size, until a stopping criterion is met

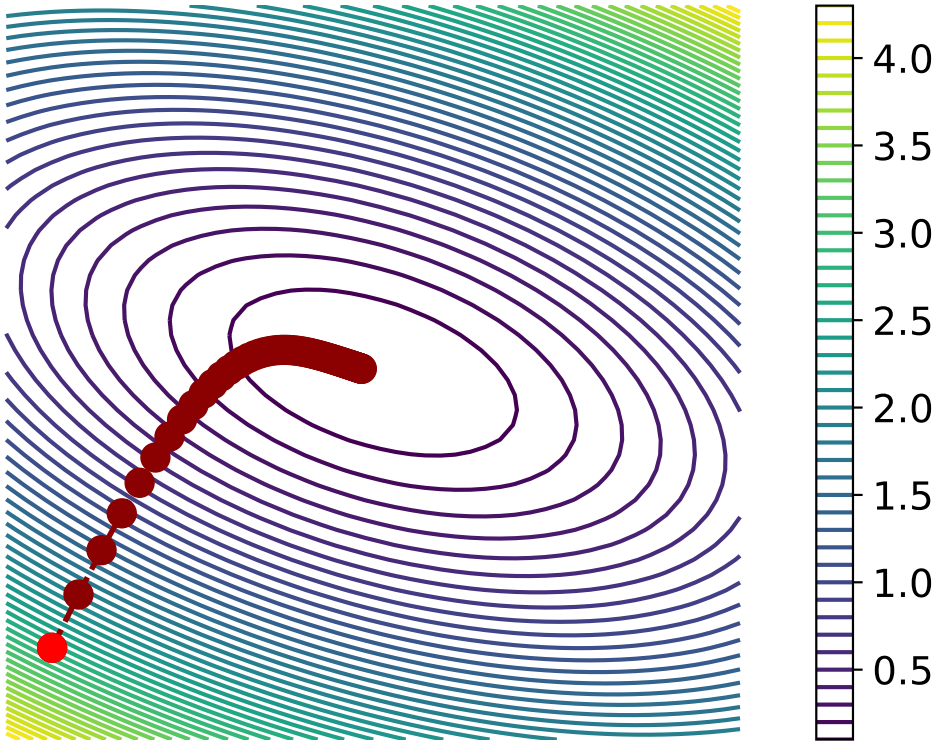
Gradient descent



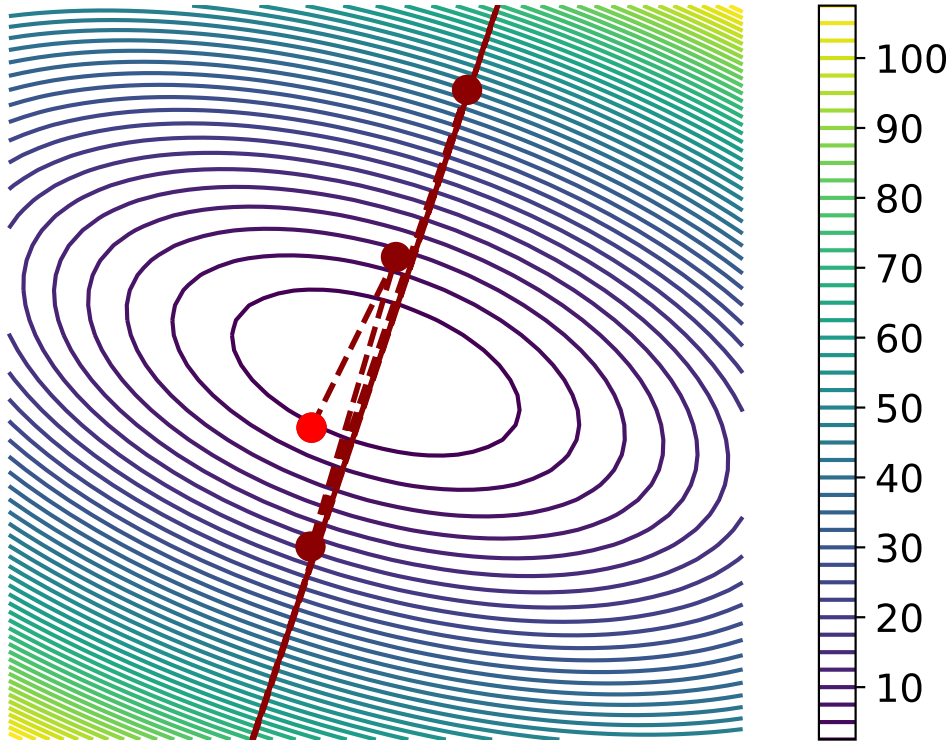
Gradient descent



Small step size



Large step size



Line search

Idea: Find minimum of

$$\begin{aligned}\alpha_k &:= \arg\min_{\alpha} h(\alpha) \\ &= \arg\min_{\alpha \in \mathbb{R}} f\left(\vec{x}^{(k)} - \alpha \nabla f\left(\vec{x}^{(k)}\right)\right)\end{aligned}$$

Problem: Solving 1D convex optimization problem can still be costly

Armijo rule

By convexity of $h := f\left(\vec{x}^{(k)} - \alpha \nabla f\left(\vec{x}^{(k)}\right)\right)$

$$h(\alpha) \geq h(0) + \alpha h'(0)$$

Armijo rule

By convexity of $h := f\left(\vec{x}^{(k)} - \alpha \nabla f\left(\vec{x}^{(k)}\right)\right)$

$$h(\alpha) \geq h(0) + \alpha h'(0)$$

By differentiability, this is the only supporting hyperplane (line)

Armijo rule

By convexity of $h := f\left(\vec{x}^{(k)} - \alpha \nabla f\left(\vec{x}^{(k)}\right)\right)$

$$h(\alpha) \geq h(0) + \alpha h'(0)$$

By differentiability, this is the only supporting hyperplane (line)

For α small enough

$$h(\alpha) \leq h(0) + \frac{\alpha}{2} h'(0)$$

Armijo rule

By convexity of $h := f\left(\vec{x}^{(k)} - \alpha \nabla f\left(\vec{x}^{(k)}\right)\right)$

$$h(\alpha) \geq h(0) + \alpha h'(0)$$

By differentiability, this is the only supporting hyperplane (line)

For α small enough

$$\begin{aligned} h(\alpha) &\leq h(0) + \frac{\alpha}{2} h'(0) \\ &= f\left(\vec{x}^{(k)}\right) + \frac{\alpha}{2} \nabla f\left(\vec{x}^{(k)}\right)^T \left(-\nabla f\left(\vec{x}^{(k)}\right)\right) \quad \text{(chain rule)} \end{aligned}$$

Armijo rule

By convexity of $h := f\left(\vec{x}^{(k)} - \alpha \nabla f\left(\vec{x}^{(k)}\right)\right)$

$$h(\alpha) \geq h(0) + \alpha h'(0)$$

By differentiability, this is the only supporting hyperplane (line)

For α small enough

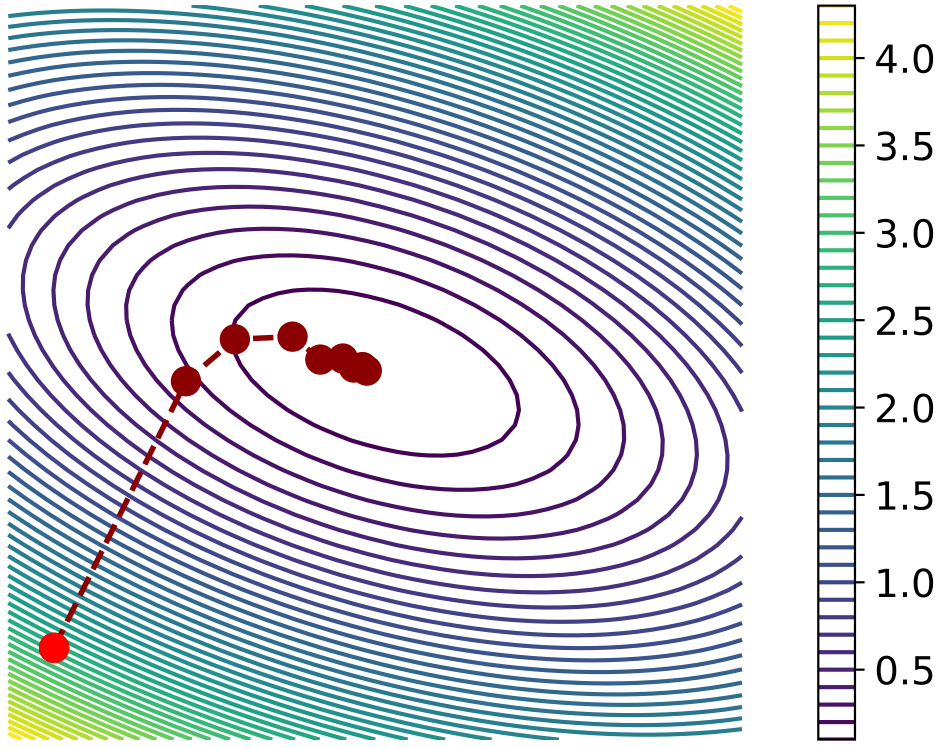
$$\begin{aligned} h(\alpha) &\leq h(0) + \frac{\alpha}{2} h'(0) \\ &= f\left(\vec{x}^{(k)}\right) + \frac{\alpha}{2} \nabla f\left(\vec{x}^{(k)}\right)^T \left(-\nabla f\left(\vec{x}^{(k)}\right)\right) \quad \text{(chain rule)} \\ &= f\left(\vec{x}^{(k)}\right) - \frac{\alpha}{2} \left|\left|\nabla f\left(\vec{x}^{(k)}\right)\right|\right|_2^2 \end{aligned}$$

Backtracking line search with Armijo rule

Given $\alpha_0 \geq 0$, set $\alpha_k := 0.5^i \alpha_0$ for smallest i such that

$$f\left(\vec{x}^{(k)} - \alpha_k \nabla f\left(\vec{x}^{(k)}\right)\right) \leq f\left(\vec{x}^{(k)}\right) - \frac{\alpha_k}{2}\left|\left|\nabla f\left(\vec{x}^{(k)}\right)\right|\right|_2^2$$

Backtracking line search with Armijo rule



Least squares

Let $\vec{y} \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, $\vec{\beta} \in \mathbb{R}^p$

The gradient of the least-squares cost function

$$f(\vec{\beta}) := \frac{1}{2} \left|\left| \vec{y} - X\vec{\beta} \right|\right|_2^2 = \frac{1}{2}\vec{y}^T\vec{y} + \frac{1}{2}\vec{\beta}^T X^T X\vec{\beta} - \vec{y}^T X\vec{\beta}$$

equals

$$\nabla f(\vec{\beta}) = X^T(X\vec{\beta} - \vec{y})$$

Gradient descent for least squares

Gradient descent updates are

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

Gradient descent for least squares

Gradient descent updates are

$$\begin{aligned}\vec{\beta}^{(k+1)} &= \vec{\beta}^{(k)} + \alpha_k X^T \left(\vec{y} - X\vec{\beta}^{(k)}\right) \\ &= \vec{\beta}^{(k)} + \alpha_k \sum_{i=1}^{n} \left(y^{(i)} - \langle \vec{x}^{(i)}, \vec{\beta}^{(k)} \rangle\right) \vec{x}^{(i)}\end{aligned}$$

Gradient descent for least squares

Let $X^{n \times p}$, $n \geq p$, be full rank with SVD USV^T

The $k+1$th iteration of gradient descent with step size $\alpha > 0$ is

$$\begin{aligned}\vec{\beta}^{(k+1)} =& V \operatorname{diag}_{j=1}^{p} \left(\left(1 - \alpha s_j^2\right)^{k+1} \right) V^T \vec{\beta}^{(0)} \\ &+ V \operatorname{diag}_{j=1}^{p} \left(\frac{1 - \left(1 - \alpha s_j^2\right)^{k+1}}{s_j} \right) U^T \vec{y}\end{aligned}$$

where $\operatorname{diag}_{j=1}^{p}(d_i)$ is a diagonal matrix with entries $d_1, \ldots, d_p$

Proof

$$VV^T = V^T V = I$$

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

Proof

$$VV^T = V^T V = I$$

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

Proof

$$VV^T = V^T V = I$$

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

Proof

$$VV^T = V^T V = I$$

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

Proof

$$\vec{\beta}^{(k+1)} = V\left(I - \alpha S^2\right)^{k+1} V^T \vec{\beta}^{(0)} + \alpha V \sum_{i=0}^{k} \left(I - \alpha S^2\right)^i V^T V S U^T \vec{y}$$

Proof

$$\begin{aligned}
\vec{\beta}^{(k+1)} &= V\left(I-\alpha S^2\right)^{k+1} V^T \vec{\beta}^{(0)} + \alpha V \sum_{i=0}^{k} \left(I-\alpha S^2\right)^i V^T V S U^T \vec{y} \\
&= V \operatorname{diag}_{j=1}^{p}\left(\left(1-\alpha s_j^2\right)^{k+1}\right) V^T \vec{\beta}^{(0)} \\
&\quad + \alpha V \operatorname{diag}_{j=1}^{p}\left(\sum_{i=0}^{k}\left(1-\alpha s_j^2\right)^i\right) S U^T \vec{y}
\end{aligned}$$

Proof

$$
\begin{aligned}
\vec{\beta}^{(k+1)} &= V\left(I-\alpha S^2\right)^{k+1} V^T \vec{\beta}^{(0)} + \alpha V \sum_{i=0}^{k} \left(I-\alpha S^2\right)^i V^T V S U^T \vec{y} \\
&= V \operatorname{diag}_{j=1}^{p}\left(\left(1-\alpha s_j^2\right)^{k+1}\right) V^T \vec{\beta}^{(0)} \\
&\quad + \alpha V \operatorname{diag}_{j=1}^{p}\left(\sum_{i=0}^{k}\left(1-\alpha s_j^2\right)^i\right) S U^T \vec{y} \\
&= V \operatorname{diag}_{j=1}^{p}\left(\left(1-\alpha s_j^2\right)^{k+1}\right) V^T \vec{\beta}^{(0)} \\
&\quad + \alpha V \operatorname{diag}_{j=1}^{p}\left(\frac{1-\left(1-\alpha s_j^2\right)^{k+1}}{\alpha s_j^2}\right) S U^T \vec{y}
\end{aligned}
$$

Proof

$$
\begin{aligned}
\vec{\beta}^{(k+1)} &= V\left(I-\alpha S^2\right)^{k+1} V^T \vec{\beta}^{(0)} + \alpha V \sum_{i=0}^{k} \left(I-\alpha S^2\right)^i V^T V S U^T \vec{y} \\
&= V \operatorname{diag}_{j=1}^{p}\left(\left(1-\alpha s_j^2\right)^{k+1}\right) V^T \vec{\beta}^{(0)} \\
&\quad + \alpha V \operatorname{diag}_{j=1}^{p}\left(\sum_{i=0}^{k}\left(1-\alpha s_j^2\right)^i\right) S U^T \vec{y} \\
&= V \operatorname{diag}_{j=1}^{p}\left(\left(1-\alpha s_j^2\right)^{k+1}\right) V^T \vec{\beta}^{(0)} \\
&\quad + \alpha V \operatorname{diag}_{j=1}^{p}\left(\frac{1-\left(1-\alpha s_j^2\right)^{k+1}}{\alpha s_j^2}\right) S U^T \vec{y} \\
&= V \operatorname{diag}_{j=1}^{p}\left(\left(1-\alpha s_j^2\right)^{k+1}\right) V^T \vec{\beta}^{(0)} \\
&\quad + \alpha V \operatorname{diag}_{j=1}^{p}\left(\frac{1-\left(1-\alpha s_j^2\right)^{k+1}}{\alpha s_j}\right) U^T \vec{y}
\end{aligned}
$$

Convergence

If $0 < \alpha < 2/s_1^2 \leq 2/s_j^2$ then $\left|1 - \alpha s_j^2\right| < 1$

and $\lim_{k\to\infty} \left(1 - \alpha s_j^2\right)^k = 0$

Convergence

If $0 < \alpha < 2/s_1^2 \leq 2/s_j^2$ then $\left|1 - \alpha s_j^2\right| < 1$

and $\lim_{k\to\infty} \left(1 - \alpha s_j^2\right)^k = 0$

This implies

$$\begin{aligned}
\lim_{k\to\infty} \vec{\beta}^{(k)} &= \lim_{k\to\infty} V \operatorname{diag}_{j=1}^{p} \left(\left(1 - \alpha s_j^2\right)^k\right) V^T \vec{\beta}^{(0)} \\
&\quad + V \operatorname{diag}_{j=1}^{p} \left(\frac{1 - \left(1 - \alpha s_j^2\right)^k}{s_j}\right) U^T \vec{y} \\
&= V S^{-1} U^T \vec{y}
\end{aligned}$$

Convergence rate

If $\alpha := 1/s_1^2$ then

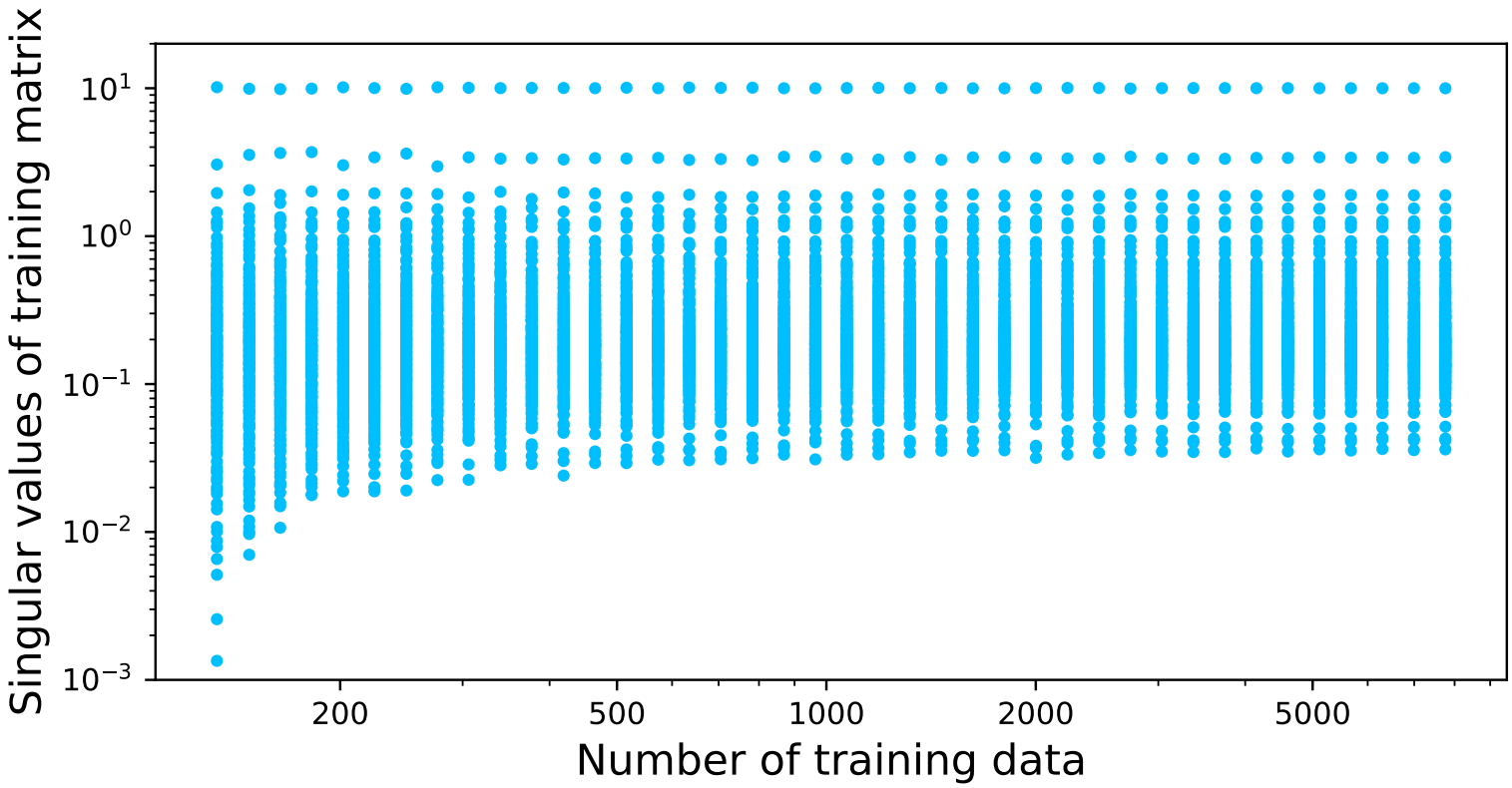
$$\frac{\left|\left|\vec{y}_{\mathrm{LS}} - \vec{y}^{(k)}\right|\right|_2}{\left|\left|\vec{y}\right|\right|_2} \leq \left(1 - \frac{s_p^2}{s_1^2}\right)^k,$$

s_1 is largest singular value of X and s_p is smallest

Convergence rate is slow if matrix is not well conditioned

Conjugate gradients converges much faster

Temperature prediction via linear regression



Proof

$$\begin{aligned}\vec{y}^{(k)} &= X\vec{\beta}^{(k)} \\ &= USV^T V \operatorname{diag}_{j=1}^{p} \left(\frac{1 - \left(1 - \alpha s_j^2\right)^k}{s_j} \right) U^T \vec{y}\end{aligned}$$

Proof

$$
\begin{aligned}
\vec{y}^{(k)} &= X\vec{\beta}^{(k)} \\
&= USV^T V \operatorname{diag}_{j=1}^{p} \left(\frac{1 - \left(1 - \alpha s_j^2\right)^k}{s_j} \right) U^T \vec{y} \\
&= USV^T VS^{-1}U^T \vec{y} - U \operatorname{diag}_{j=1}^{p} \left(\left(1 - \alpha s_j^2\right)^k \right) U^T \vec{y}
\end{aligned}
$$

Proof

$$
\begin{aligned}
\vec{y}^{(k)} &= X\vec{\beta}^{(k)} \\
&= USV^T V \operatorname{diag}_{j=1}^{p} \left(\frac{1 - \left(1 - \alpha s_j^2\right)^k}{s_j} \right) U^T \vec{y} \\
&= USV^T VS^{-1}U^T \vec{y} - U \operatorname{diag}_{j=1}^{p} \left(\left(1 - \alpha s_j^2\right)^k \right) U^T \vec{y} \\
&= UU^T \vec{y} - U \operatorname{diag}_{j=1}^{p} \left(\left(1 - \alpha s_j^2\right)^k \right) U^T \vec{y}
\end{aligned}
$$

Proof

$$\begin{aligned}
\vec{y}^{(k)} &= X\vec{\beta}^{(k)} \\
&= USV^T V \operatorname{diag}_{j=1}^{p} \left(\frac{1-\left(1-\alpha s_j^2\right)^k}{s_j} \right) U^T \vec{y} \\
&= USV^T VS^{-1}U^T\vec{y} - U \operatorname{diag}_{j=1}^{p} \left(\left(1-\alpha s_j^2\right)^k \right) U^T \vec{y} \\
&= UU^T\vec{y} - U \operatorname{diag}_{j=1}^{p} \left(\left(1-\alpha s_j^2\right)^k \right) U^T \vec{y}
\end{aligned}$$

Since $\vec{y}_{\mathsf{LS}} = UU^T\vec{y}$

$$\left|\left|\vec{y}_{\mathsf{LS}} - \vec{y}^{(k)}\right|\right|_2 = \left|\left|U \operatorname{diag}_{j=1}^{p} \left(\left(1-\alpha s_j^2\right)^k \right) U^T \vec{y}\right|\right|_2$$

Proof

$$\begin{aligned}
\vec{y}^{(k)} &= X\vec{\beta}^{(k)} \\
&= USV^T V \operatorname{diag}_{j=1}^{p} \left(\frac{1-\left(1-\alpha s_j^2\right)^k}{s_j} \right) U^T \vec{y} \\
&= USV^T VS^{-1}U^T \vec{y} - U \operatorname{diag}_{j=1}^{p} \left(\left(1-\alpha s_j^2\right)^k \right) U^T \vec{y} \\
&= UU^T \vec{y} - U \operatorname{diag}_{j=1}^{p} \left(\left(1-\alpha s_j^2\right)^k \right) U^T \vec{y}
\end{aligned}$$

Since $\vec{y}_{\mathsf{LS}} = UU^T\vec{y}$

$$\begin{aligned}
\left|\left|\vec{y}_{\mathsf{LS}} - \vec{y}^{(k)}\right|\right|_2 &= \left|\left| U \operatorname{diag}_{j=1}^{p} \left(\left(1-\alpha s_j^2\right)^k \right) U^T \vec{y} \right|\right|_2 \\
&\leq ||U|| \left|\left| \operatorname{diag}_{j=1}^{p} \left(\left(1-\alpha s_j^2\right)^k \right) \right|\right| \left|\left| U^T \vec{y} \right|\right|_2
\end{aligned}$$

Proof

$$\begin{aligned}
\vec{y}^{(k)} &= X\vec{\beta}^{(k)} \\
&= USV^T V \operatorname{diag}_{j=1}^{p} \left(\frac{1-\left(1-\alpha s_j^2\right)^k}{s_j} \right) U^T \vec{y} \\
&= USV^T VS^{-1}U^T\vec{y} - U \operatorname{diag}_{j=1}^{p} \left(\left(1-\alpha s_j^2\right)^k \right) U^T\vec{y} \\
&= UU^T\vec{y} - U \operatorname{diag}_{j=1}^{p} \left(\left(1-\alpha s_j^2\right)^k \right) U^T\vec{y}
\end{aligned}$$

Since $\vec{y}_{\mathsf{LS}} = UU^T\vec{y}$

$$\begin{aligned}
\left|\left|\vec{y}_{\mathsf{LS}} - \vec{y}^{(k)}\right|\right|_2 &= \left|\left| U \operatorname{diag}_{j=1}^{p} \left(\left(1-\alpha s_j^2\right)^k \right) U^T\vec{y} \right|\right|_2 \\
&\leq ||U|| \left|\left| \operatorname{diag}_{j=1}^{p} \left(\left(1-\alpha s_j^2\right)^k \right) \right|\right| \left|\left| U^T\vec{y} \right|\right|_2 \\
&\leq \left| 1 - \frac{s_p^2}{s_1^2} \right|^k ||\vec{y}||_2
\end{aligned}$$

Linear regression

Assume additive model for regression problem

$$\vec{y}_{\text{train}} := X_{\text{train}} \vec{\beta}_{\text{true}} + \vec{z}_{\text{train}}$$

Estimate coefficients via gradient descent up to iteration k

Coefficient error equals

$$\begin{aligned}\vec{\beta}_{\text{true}} - \vec{\beta}^{(k+1)} = & V \operatorname{diag}_{j=1}^{p} \left(\left(1 - \alpha s_j^2\right)^{k+1} \right) V^T \vec{\beta}_{\text{true}} \\ & - V \operatorname{diag}_{j=1}^{p} \left(\frac{1 - \left(1 - \alpha s_j^2\right)^{k+1}}{s_j} \right) U^T \vec{z}_{\text{train}}\end{aligned}$$

Proof

$$\vec{\beta}^{(k+1)} = V \operatorname{diag}_{j=1}^{p} \left(\frac{1 - \left(1 - \alpha s_j^2\right)^{k+1}}{s_j} \right) U^T \left(X_{\text{train}} \vec{\beta}_{\text{true}} + \vec{z}_{\text{train}} \right)$$

Proof

$$\begin{aligned}
\vec{\beta}^{(k+1)} &= V \operatorname{diag}_{j=1}^{p} \left(\frac{1 - \left(1 - \alpha s_j^2\right)^{k+1}}{s_j} \right) U^T \left(X_{\text{train}} \vec{\beta}_{\text{true}} + \vec{z}_{\text{train}} \right) \\
&= V \operatorname{diag}_{j=1}^{p} \left(\frac{1 - \left(1 - \alpha s_j^2\right)^{k+1}}{s_j} \right) U^T \left(USV^T \vec{\beta}_{\text{true}} + \vec{z}_{\text{train}} \right)
\end{aligned}$$

Proof

$$
\begin{aligned}
\vec{\beta}^{(k+1)} &= V \operatorname{diag}_{j=1}^{p} \left(\frac{1 - \left(1 - \alpha s_j^2\right)^{k+1}}{s_j} \right) U^T \left(X_{\text{train}} \vec{\beta}_{\text{true}} + \vec{z}_{\text{train}} \right) \\
&= V \operatorname{diag}_{j=1}^{p} \left(\frac{1 - \left(1 - \alpha s_j^2\right)^{k+1}}{s_j} \right) U^T \left(USV^T \vec{\beta}_{\text{true}} + \vec{z}_{\text{train}} \right) \\
&= VS^{-1} U^T USV^T \vec{\beta}_{\text{true}} - V \operatorname{diag}_{j=1}^{p} \left(\left(1 - \alpha s_j^2\right)^{k+1} \right) V^T \vec{\beta}_{\text{true}} \\
&\quad + V \operatorname{diag}_{j=1}^{p} \left(\frac{1 - \left(1 - \alpha s_j^2\right)^{k+1}}{s_j} \right) U^T \vec{z}_{\text{train}} \\
&= \vec{\beta}_{\text{true}} - V \operatorname{diag}_{j=1}^{p} \left(\left(1 - \alpha s_j^2\right)^{k+1} \right) V^T \vec{\beta}_{\text{true}} \\
&\quad + V \operatorname{diag}_{j=1}^{p} \left(\frac{1 - \left(1 - \alpha s_j^2\right)^{k+1}}{s_j} \right) U^T \vec{z}_{\text{train}}
\end{aligned}
$$

Early stopping

$$\vec{\beta}_{\text{true}} - \vec{\beta}^{(k+1)} = V \,\text{diag}_{j=1}^{p} \left(\left(1 - \alpha s_j^2\right)^{k+1} \right) V^T \vec{\beta}_{\text{true}} - V \,\text{diag}_{j=1}^{p} \left(\frac{1 - \left(1 - \alpha s_j^2\right)^{k+1}}{s_j} \right) U^T \vec{z}_{\text{train}}$$

If sample covariance matrix not close true covariance matrix then 2nd term can produce noise amplification

Early stopping

$$\vec{\beta}_{\text{true}} - \vec{\beta}^{(k+1)} = V \,\text{diag}_{j=1}^{p} \left(\left(1 - \alpha s_j^2\right)^{k+1} \right) V^T \vec{\beta}_{\text{true}} - V \,\text{diag}_{j=1}^{p} \left(\frac{1 - \left(1 - \alpha s_j^2\right)^{k+1}}{s_j} \right) U^T \vec{z}_{\text{train}}$$

If sample covariance matrix not close true covariance matrix then 2nd term can produce noise amplification

Selecting small k reduces effect, but increases 1st term

Early stopping

$$\vec{\beta}_{\text{true}} - \vec{\beta}^{(k+1)} = V \,\text{diag}_{j=1}^{p} \left(\left(1 - \alpha s_j^2\right)^{k+1} \right) V^T \vec{\beta}_{\text{true}} - V \,\text{diag}_{j=1}^{p} \left(\frac{1 - \left(1 - \alpha s_j^2\right)^{k+1}}{s_j} \right) U^T \vec{z}_{\text{train}}$$

If sample covariance matrix not close true covariance matrix then 2nd term can produce noise amplification

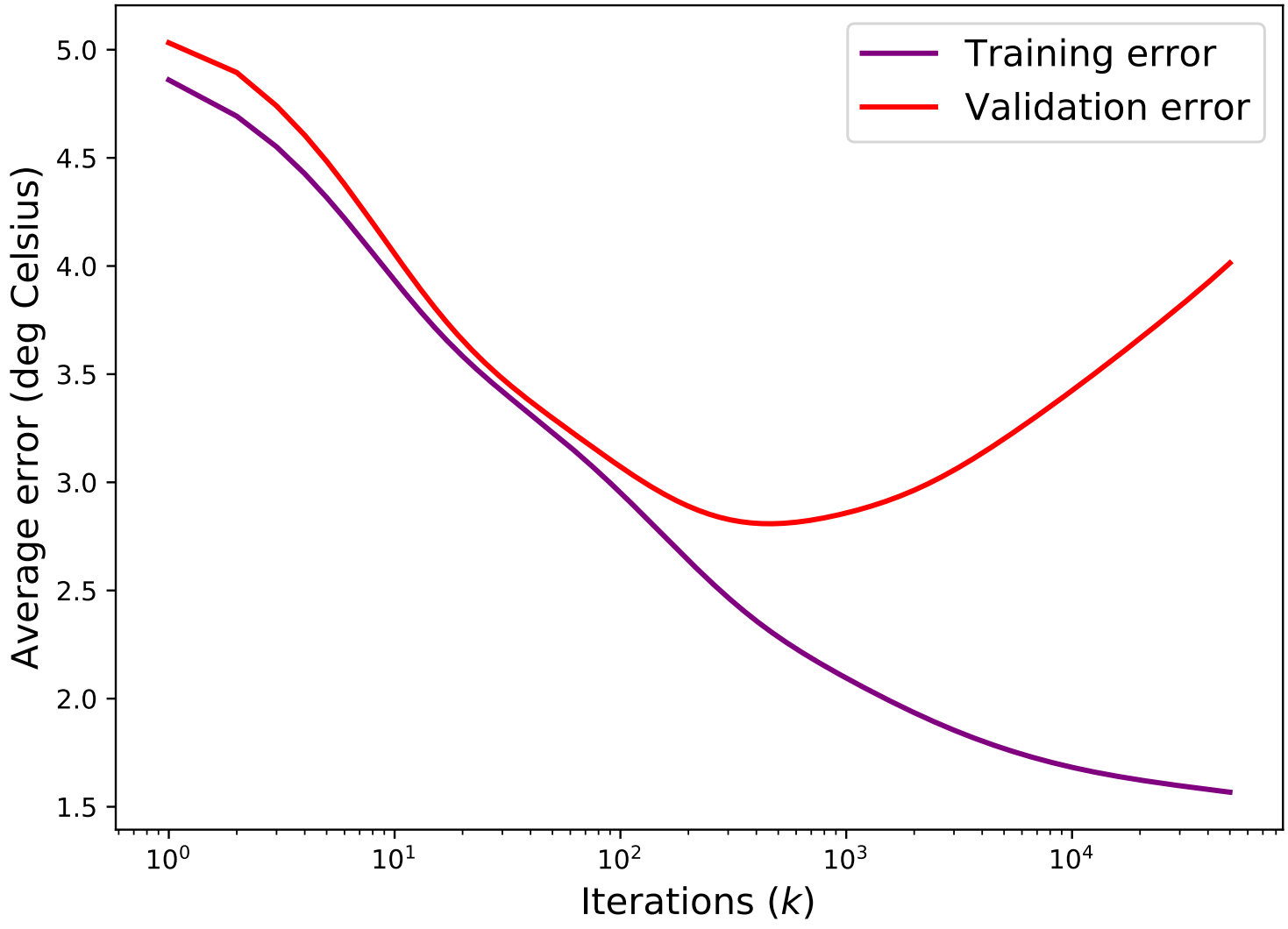
Selecting small k reduces effect, but increases 1st term

Solution: Select best k by minimizing validation error

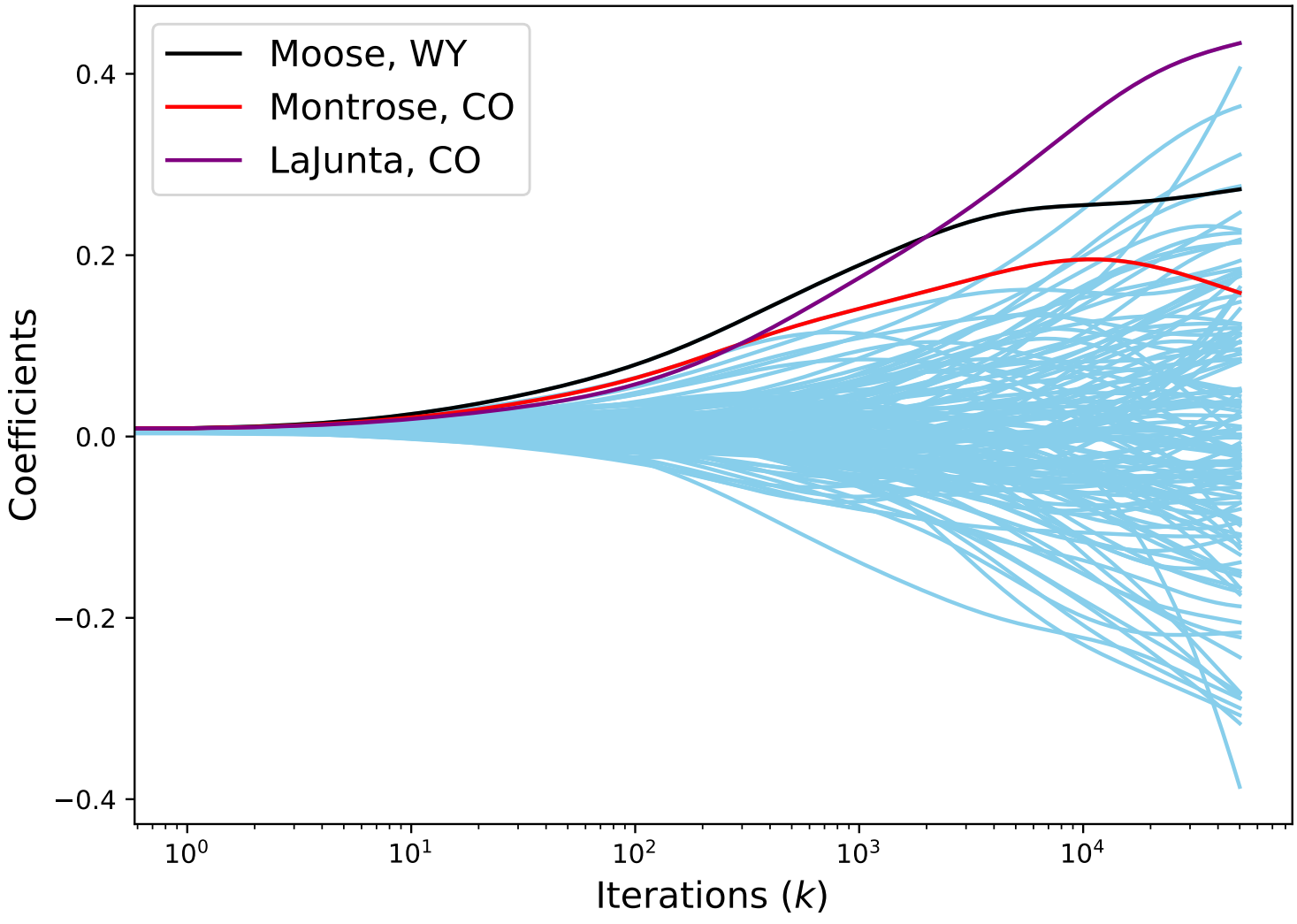
Temperature prediction via linear regression

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

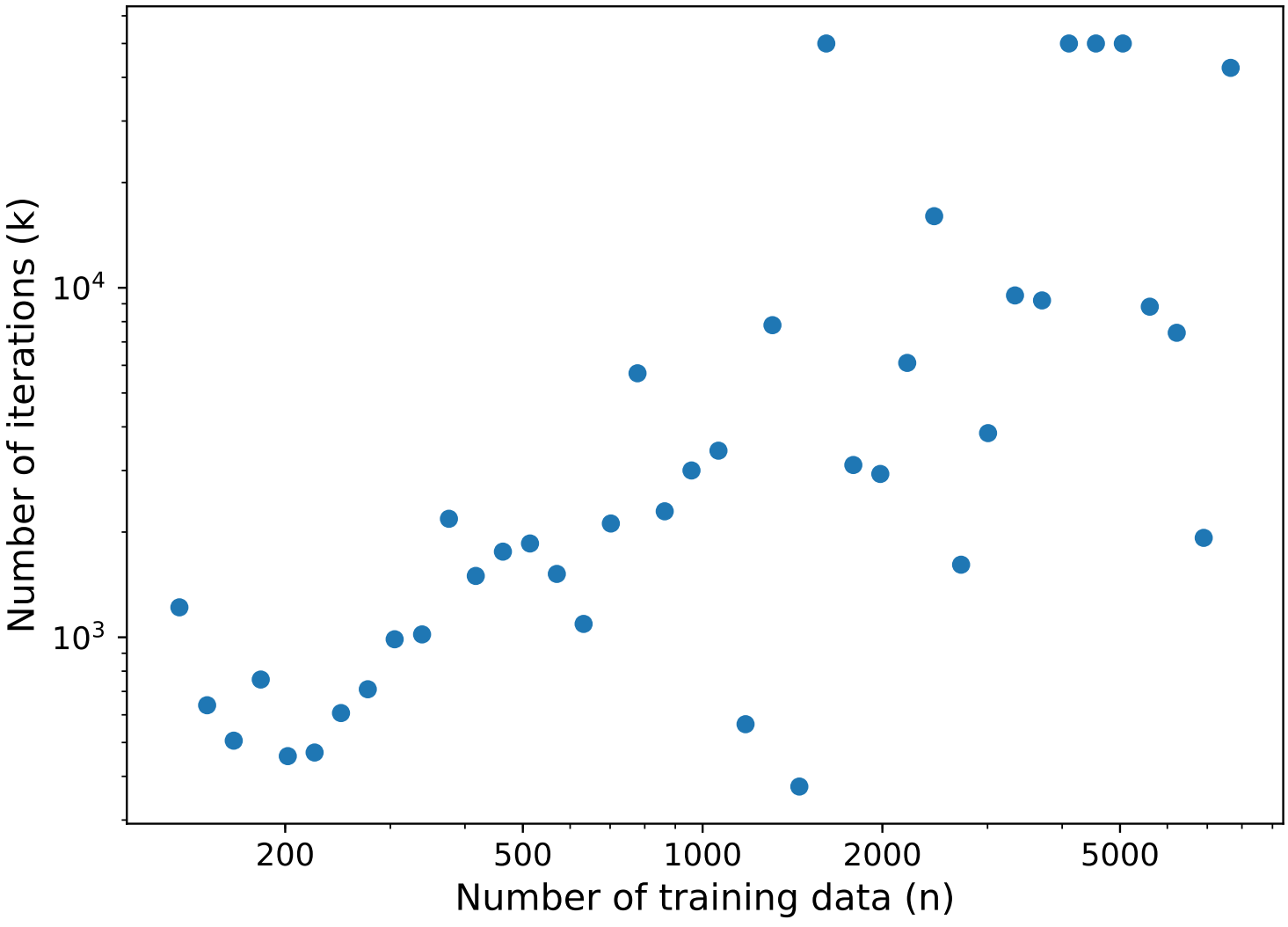
Gradient-descent estimator ($n = 200$)



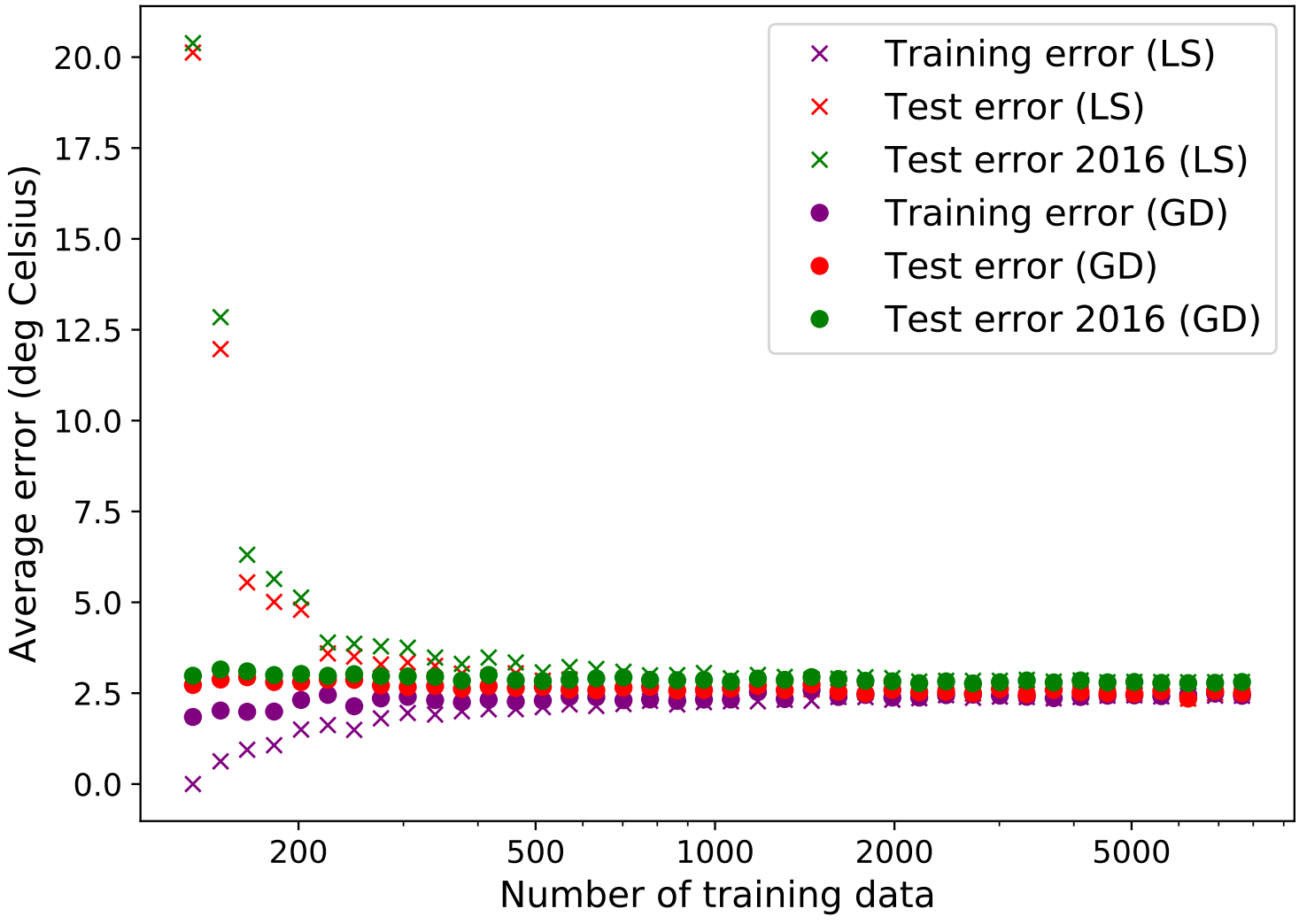
Gradient-descent estimator ($n = 200$)



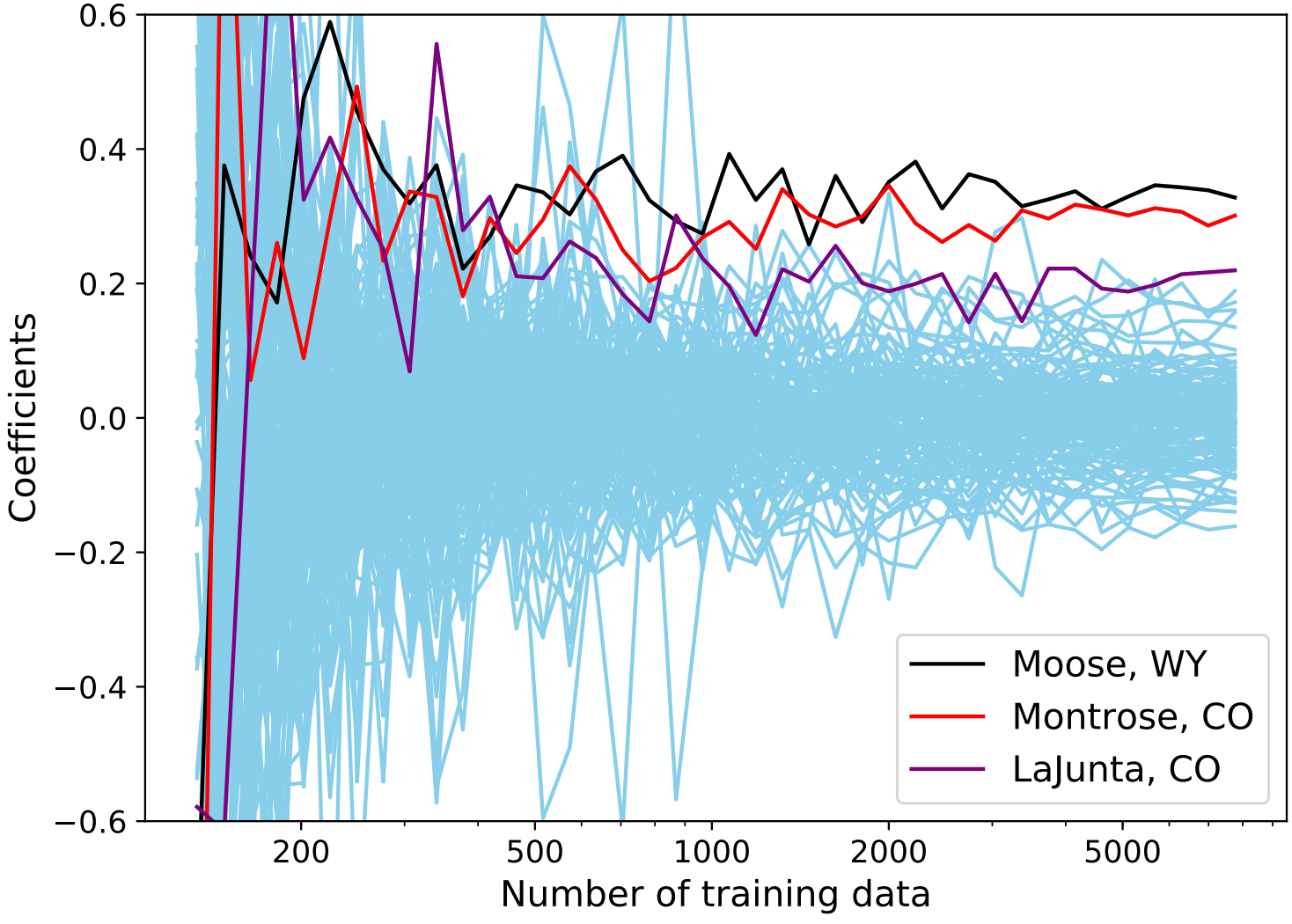
Selected number of iterations



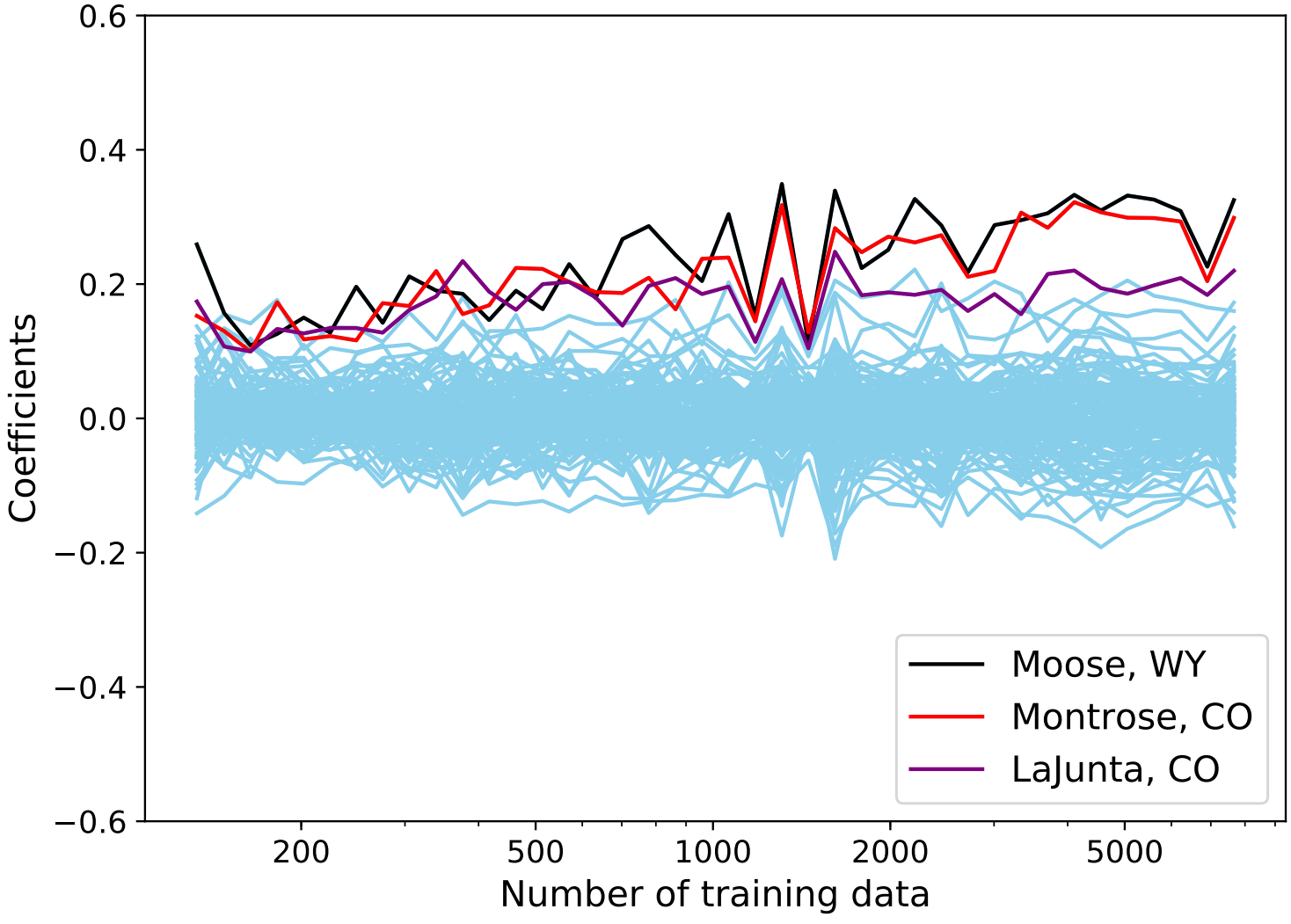
Comparison to least squares



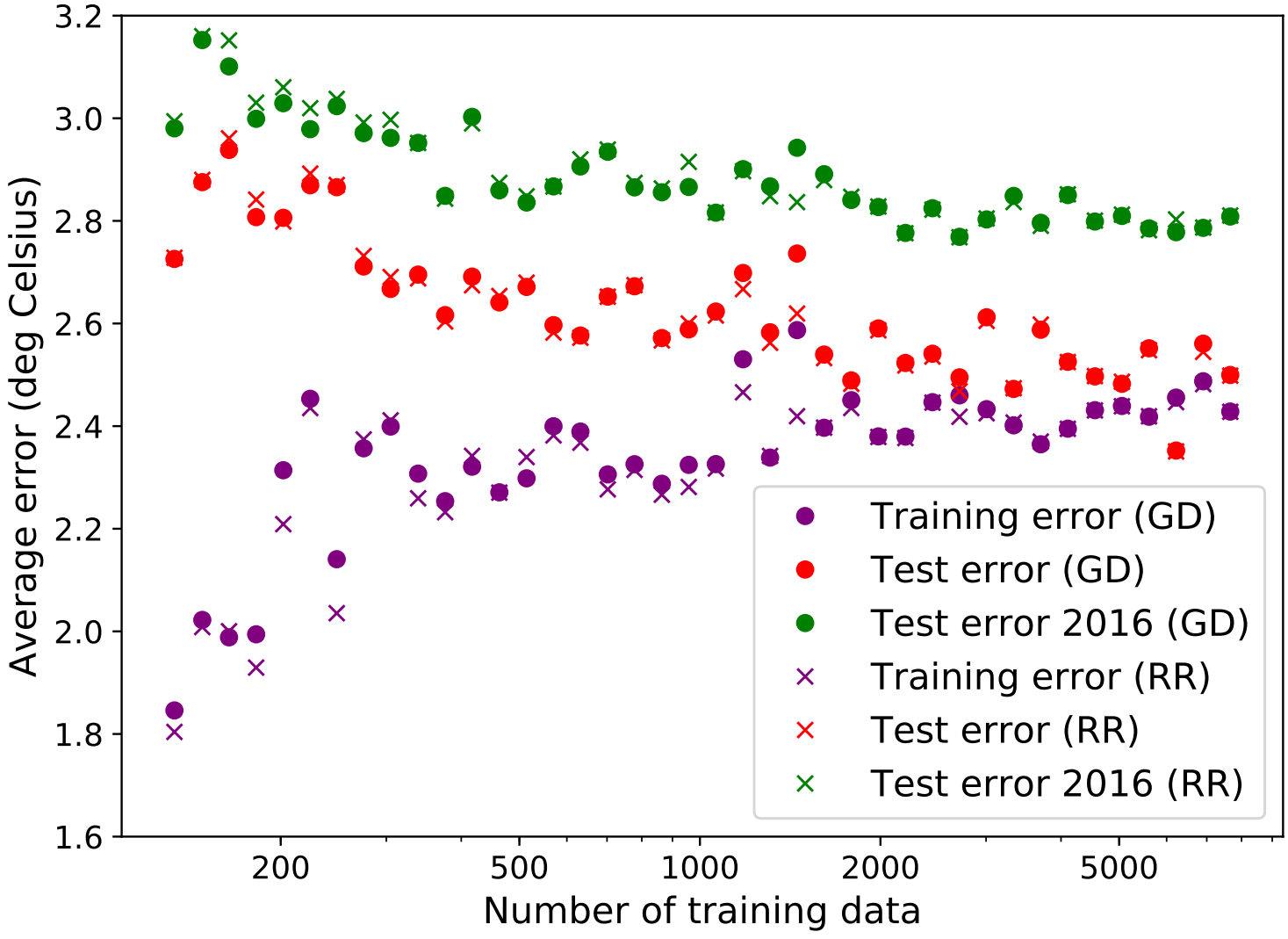
Least-squares coefficients



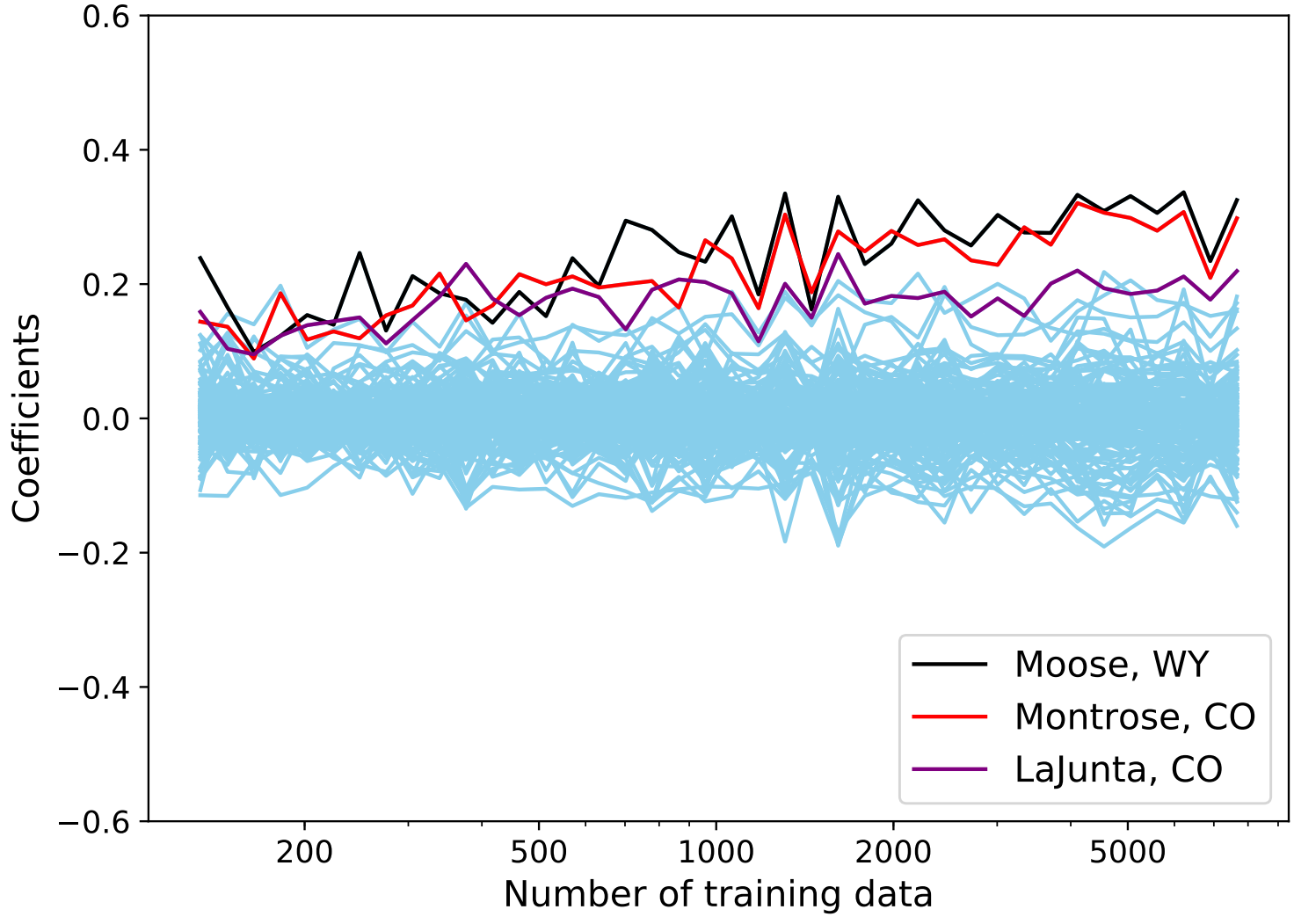
Gradient-descent coefficients



Comparison to ridge regression



Ridge-regression coefficients



Convergence of gradient descent

Functions bounded locally by quadratic with fixed curvature

Satisfied by most reasonable functions

Lipschitz continuity

A function $g : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous if for any $\vec{x}, \vec{y} \in \mathbb{R}^n$

$$||g(\vec{y}) - g(\vec{x})||_2 \leq L ||\vec{y} - \vec{x}||_2 .$$

L is the Lipschitz constant

Lipschitz-continuous gradients

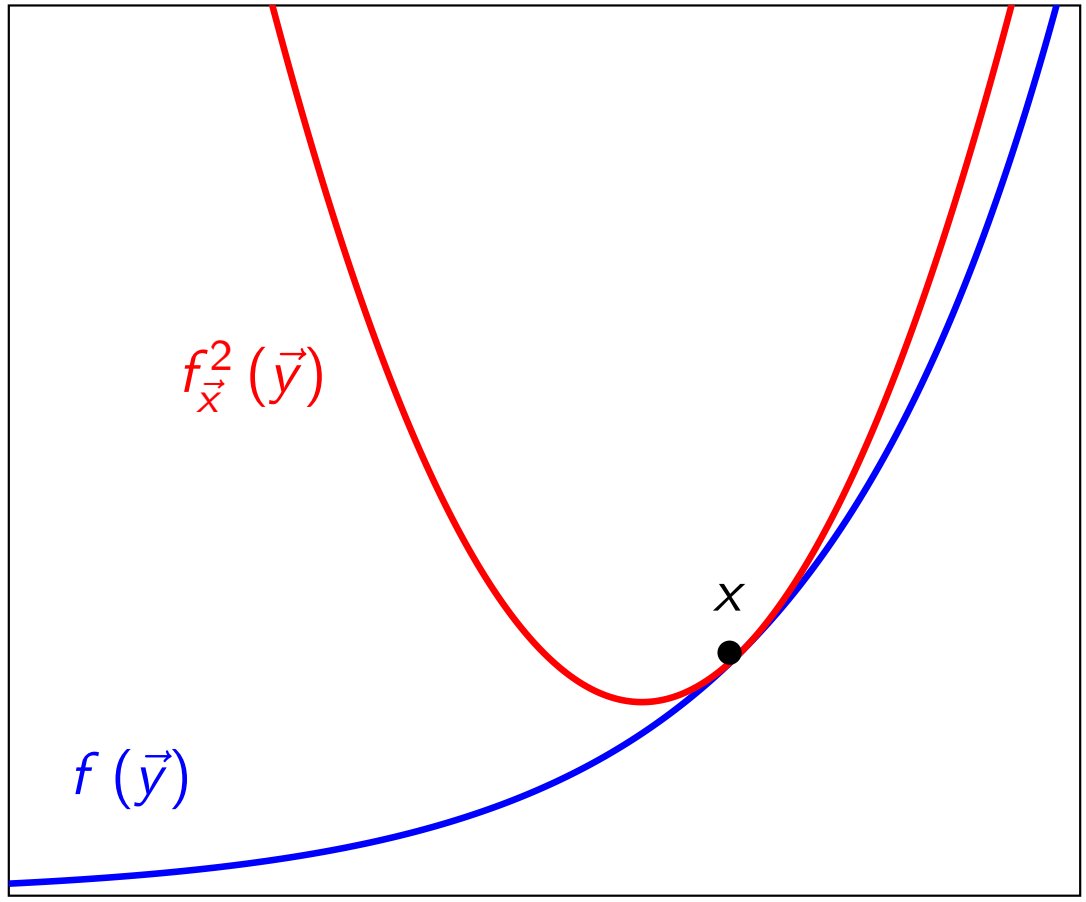
If ∇f is Lipschitz continuous with Lipschitz constant L

$$||\nabla f(\vec{y}) - \nabla f(\vec{x})||_2 \leq L ||\vec{y} - \vec{x}||_2$$

then for any $\vec{x}, \vec{y} \in \mathbb{R}^n$ we have a quadratic upper bound

$$f(\vec{y}) \leq f_{\vec{x}}^2(\vec{y}) := f(\vec{x}) + \nabla f(\vec{x})^T (\vec{y} - \vec{x}) + \frac{L}{2} ||\vec{y} - \vec{x}||_2^2$$

Quadratic upper bound



Least squares

For any $\vec{\beta}_1, \vec{\beta}_2 \in \mathbb{R}^p$,

$$\left|\left|\nabla f\left(\vec{\beta}_2\right) - \nabla f\left(\vec{\beta}_1\right)\right|\right|_2 = \left|\left|X^T X\left(\vec{\beta}_2 - \vec{\beta}_1\right)\right|\right|_2 \tag{1}$$

$$\leq s_1^2 \left|\left|\vec{\beta}_2 - \vec{\beta}_1\right|\right|_2 \tag{2}$$

so $L = s_1^2$

Local progress of gradient descent

$$\vec{x}^{(k+1)} := \vec{x}^{(k)} - \alpha_k \nabla f\left(\vec{x}^{(k)}\right)$$

$$f\left(\vec{x}^{(k+1)}\right)$$

Local progress of gradient descent

$$\vec{x}^{(k+1)} := \vec{x}^{(k)} - \alpha_k \nabla f\left(\vec{x}^{(k)}\right)$$

$$f\left(\vec{x}^{(k+1)}\right)$$
$$\leq f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(\vec{x}^{(k+1)} - \vec{x}^{(k)}\right) + \frac{L}{2}\left|\left|\vec{x}^{(k+1)} - \vec{x}^{(k)}\right|\right|_2^2$$

Local progress of gradient descent

$$\vec{x}^{(k+1)} := \vec{x}^{(k)} - \alpha_k \nabla f\left(\vec{x}^{(k)}\right)$$

$$\begin{aligned} & f\left(\vec{x}^{(k+1)}\right) \\ & \leq f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(\vec{x}^{(k+1)} - \vec{x}^{(k)}\right) + \frac{L}{2}\left|\left|\vec{x}^{(k+1)} - \vec{x}^{(k)}\right|\right|_2^2 \\ & = f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(-\alpha_k \nabla f\left(\vec{x}^{(k)}\right)\right) + \frac{L}{2}\left|\left|-\alpha_k \nabla f\left(\vec{x}^{(k)}\right)\right|\right|_2^2 \end{aligned}$$

Local progress of gradient descent

$$\vec{x}^{(k+1)} := \vec{x}^{(k)} - \alpha_k \nabla f\left(\vec{x}^{(k)}\right)$$

$$\begin{aligned}
& f\left(\vec{x}^{(k+1)}\right) \\
& \leq f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(\vec{x}^{(k+1)} - \vec{x}^{(k)}\right) + \frac{L}{2}\left|\left|\vec{x}^{(k+1)} - \vec{x}^{(k)}\right|\right|_2^2 \\
& = f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(-\alpha_k \nabla f\left(\vec{x}^{(k)}\right)\right) + \frac{L}{2}\left|\left|-\alpha_k \nabla f\left(\vec{x}^{(k)}\right)\right|\right|_2^2 \\
& = f\left(\vec{x}^{(k)}\right) - \alpha_k \left(1 - \frac{\alpha_k L}{2}\right)\left|\left|\nabla f\left(\vec{x}^{(k)}\right)\right|\right|_2^2
\end{aligned}$$

Local progress of gradient descent

$$\vec{x}^{(k+1)} := \vec{x}^{(k)} - \alpha_k \nabla f\left(\vec{x}^{(k)}\right)$$

$$\begin{aligned}
& f\left(\vec{x}^{(k+1)}\right) \\
& \leq f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(\vec{x}^{(k+1)} - \vec{x}^{(k)}\right) + \frac{L}{2}\left|\left|\vec{x}^{(k+1)} - \vec{x}^{(k)}\right|\right|_2^2 \\
& = f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(-\alpha_k \nabla f\left(\vec{x}^{(k)}\right)\right) + \frac{L}{2}\left|\left|-\alpha_k \nabla f\left(\vec{x}^{(k)}\right)\right|\right|_2^2 \\
& = f\left(\vec{x}^{(k)}\right) - \alpha_k\left(1 - \frac{\alpha_k L}{2}\right)\left|\left|\nabla f\left(\vec{x}^{(k)}\right)\right|\right|_2^2
\end{aligned}$$

If $\alpha_k \leq \frac{1}{L}$

$$f\left(\vec{x}^{(k+1)}\right) \leq f\left(\vec{x}^{(k)}\right) - \frac{\alpha_k}{2}\left|\left|\nabla f\left(\vec{x}^{(k)}\right)\right|\right|_2^2$$

Convergence of gradient descent

- f is convex
- ∇f is L-Lipschitz continuous
- There exists a point $\vec{x}^*$ at which f achieves a finite minimum
- The step size is set to $\alpha_k := \alpha \leq 1/L$

$$f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^*\right) \leq \frac{\left|\left|\vec{x}^{(0)} - \vec{x}^*\right|\right|_2^2}{2\,\alpha\,k}$$

Convergence of gradient descent

$$f\left(\vec{x}^{(k)}\right) \leq f\left(\vec{x}^{(k-1)}\right) - \frac{\alpha_k}{2}\left|\left|\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2$$

By convexity

$$f\left(\vec{x}^{*}\right) \geq f\left(\vec{x}^{(k-1)}\right) + \nabla f\left(\vec{x}^{(k-1)}\right)^T\left(\vec{x}^{*} - \vec{x}^{(k-1)}\right)$$

$$f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right)$$

Convergence of gradient descent

$$f\left(\vec{x}^{(k)}\right) \leq f\left(\vec{x}^{(k-1)}\right) - \frac{\alpha_k}{2}\left|\left|\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2$$

By convexity

$$f\left(\vec{x}^*\right) \geq f\left(\vec{x}^{(k-1)}\right) + \nabla f\left(\vec{x}^{(k-1)}\right)^T\left(\vec{x}^* - \vec{x}^{(k-1)}\right)$$

$$f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^*\right)$$
$$\leq f\left(\vec{x}^{(k-1)}\right) - f\left(\vec{x}^*\right) - \frac{\alpha_k}{2}\left|\left|\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2$$

Convergence of gradient descent

$$f\left(\vec{x}^{(k)}\right) \leq f\left(\vec{x}^{(k-1)}\right) - \frac{\alpha_k}{2}\left|\left|\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2$$

By convexity

$$f\left(\vec{x}^{*}\right) \geq f\left(\vec{x}^{(k-1)}\right) + \nabla f\left(\vec{x}^{(k-1)}\right)^T\left(\vec{x}^{*} - \vec{x}^{(k-1)}\right)$$

$$\begin{aligned}
&f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right) \\
&\leq f\left(\vec{x}^{(k-1)}\right) - f\left(\vec{x}^{*}\right) - \frac{\alpha_k}{2}\left|\left|\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2 \\
&\leq \nabla f\left(\vec{x}^{(k-1)}\right)^T\left(\vec{x}^{(k-1)} - \vec{x}^{*}\right) - \frac{\alpha}{2}\left|\left|\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2
\end{aligned}$$

Convergence of gradient descent

$$f\left(\vec{x}^{(k)}\right) \leq f\left(\vec{x}^{(k-1)}\right) - \frac{\alpha_k}{2}\left|\left|\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2$$

By convexity

$$f\left(\vec{x}^{*}\right) \geq f\left(\vec{x}^{(k-1)}\right) + \nabla f\left(\vec{x}^{(k-1)}\right)^T\left(\vec{x}^{*} - \vec{x}^{(k-1)}\right)$$

$$\begin{aligned}
&f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right) \\
&\leq f\left(\vec{x}^{(k-1)}\right) - f\left(\vec{x}^{*}\right) - \frac{\alpha_k}{2}\left|\left|\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2 \\
&\leq \nabla f\left(\vec{x}^{(k-1)}\right)^T\left(\vec{x}^{(k-1)} - \vec{x}^{*}\right) - \frac{\alpha}{2}\left|\left|\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2 \\
&= \frac{1}{2\,\alpha}\left(\left|\left|\vec{x}^{(k-1)} - \vec{x}^{*}\right|\right|_2^2 - \left|\left|\vec{x}^{(k-1)} - \vec{x}^{*} - \alpha\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2\right)
\end{aligned}$$

Convergence of gradient descent

$$f\left(\vec{x}^{(k)}\right) \leq f\left(\vec{x}^{(k-1)}\right) - \frac{\alpha_k}{2}\left|\left|\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2$$

By convexity

$$f\left(\vec{x}^{*}\right) \geq f\left(\vec{x}^{(k-1)}\right) + \nabla f\left(\vec{x}^{(k-1)}\right)^T\left(\vec{x}^{*} - \vec{x}^{(k-1)}\right)$$

$$\begin{aligned}
&f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right) \\
&\leq f\left(\vec{x}^{(k-1)}\right) - f\left(\vec{x}^{*}\right) - \frac{\alpha_k}{2}\left|\left|\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2 \\
&\leq \nabla f\left(\vec{x}^{(k-1)}\right)^T\left(\vec{x}^{(k-1)} - \vec{x}^{*}\right) - \frac{\alpha}{2}\left|\left|\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2 \\
&= \frac{1}{2\,\alpha}\left(\left|\left|\vec{x}^{(k-1)} - \vec{x}^{*}\right|\right|_2^2 - \left|\left|\vec{x}^{(k-1)} - \vec{x}^{*} - \alpha\nabla f\left(\vec{x}^{(k-1)}\right)\right|\right|_2^2\right) \\
&= \frac{1}{2\,\alpha}\left(\left|\left|\vec{x}^{(k-1)} - \vec{x}^{*}\right|\right|_2^2 - \left|\left|\vec{x}^{(k)} - \vec{x}^{*}\right|\right|_2\right)
\end{aligned}$$

Convergence of gradient descent

$$f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right)$$

Convergence of gradient descent

$$f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right) \leq \frac{1}{k} \sum_{i=1}^{k} f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right)$$

Convergence of gradient descent

$$f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right) \leq \frac{1}{k}\sum_{i=1}^{k} f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right) \qquad \text{never increases}$$

Convergence of gradient descent

$$f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right) \leq \frac{1}{k}\sum_{i=1}^{k} f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right) \qquad \text{never increases}$$

$$= \frac{1}{2\,\alpha\,k}\sum_{i=1}^{k} \left|\left|\vec{x}^{(k-1)} - \vec{x}^{*}\right|\right|_2^2 - \left|\left|\vec{x}^{(k)} - \vec{x}^{*}\right|\right|_2$$

Convergence of gradient descent

$$f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right) \leq \frac{1}{k}\sum_{i=1}^{k} f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right) \qquad \text{never increases}$$

$$= \frac{1}{2\,\alpha\,k}\sum_{i=1}^{k} \left|\left|\vec{x}^{(k-1)} - \vec{x}^{*}\right|\right|_2^2 - \left|\left|\vec{x}^{(k)} - \vec{x}^{*}\right|\right|_2$$

$$= \frac{1}{2\,\alpha\,k}\left(\left|\left|\vec{x}^{(0)} - \vec{x}^{*}\right|\right|_2^2 - \left|\left|\vec{x}^{(k)} - \vec{x}^{*}\right|\right|_2^2\right)$$

Convergence of gradient descent

$$f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right) \leq \frac{1}{k}\sum_{i=1}^{k} f\left(\vec{x}^{(k)}\right) - f\left(\vec{x}^{*}\right) \qquad \text{never increases}$$

$$= \frac{1}{2\,\alpha\,k}\sum_{i=1}^{k}\left|\left|\vec{x}^{(k-1)} - \vec{x}^{*}\right|\right|_2^2 - \left|\left|\vec{x}^{(k)} - \vec{x}^{*}\right|\right|_2$$

$$= \frac{1}{2\,\alpha\,k}\left(\left|\left|\vec{x}^{(0)} - \vec{x}^{*}\right|\right|_2^2 - \left|\left|\vec{x}^{(k)} - \vec{x}^{*}\right|\right|_2^2\right)$$

$$\leq \frac{\left|\left|\vec{x}^{(0)} - \vec{x}^{*}\right|\right|_2^2}{2\,\alpha\,k}$$

Accelerated gradient descent

- Gradient descent takes $\mathcal{O}(1/\epsilon)$ to achieve an error of ϵ
- The optimal rate is $\mathcal{O}(1/\sqrt{\epsilon})$
- Gradient descent can be accelerated by adding a momentum term

Stochastic gradient descent

Cost functions to fit models are often additive

$$f\left(\vec{x}\right) = \frac{1}{m} \sum_{i=1}^{m} f_i\left(\vec{x}\right).$$

- Linear regression

$$\sum_{i=1}^{n} \left(y^{(i)} - \vec{x}^{(i)\,T} \vec{\beta}\right)^2 = \left|\left|\vec{y} - X\vec{\beta}\right|\right|_2^2$$

Stochastic gradient descent

In *big data* regime (very large n), gradient descent is too slow

In some cases, data is acquired sequentially (online setting)

Stochastic gradient descent: update solution using a subset of the data

Stochastic gradient descent

Set the initial point $\vec{x}^{(0)}$ to an arbitrary value

Update by

1. Choosing a random subset of b indices $\mathcal{B}$ ($b \ll m$ is the batch size)
2. Setting

$$\vec{\mathbf{x}}^{(k+1)} := \vec{\mathbf{x}}^{(k)} - \alpha_k m \sum_{i \in \mathcal{B}} \nabla f_i \left(\vec{\mathbf{x}}^{(k)} \right)$$

where α_k is the step size

Stochastic gradient descent

For fixed $\vec{x}^{(k)}$ we replace $\nabla f(\vec{x}^{(k)})$ by

$$\sum_{i \in \mathcal{B}} \nabla f_i \left(\vec{x}^{(k)} \right) = \sum_{i=1}^{m} 1_{i \in \mathcal{B}} \nabla f_i \left(\vec{x}^{(k)} \right)$$

Noisy estimate of ∇f

Stochastic gradient descent

If samples are uniform with replacement, then estimate is aligned with ∇f on average

$$\mathrm{E}\left(\sum_{i=1}^{m} 1_{i\in\mathcal{B}} \nabla f_i\left(\vec{x}^{(k)}\right)\right)$$

Stochastic gradient descent

If samples are uniform with replacement, then estimate is aligned with ∇f on average

$$\mathrm{E}\left(\sum_{i=1}^{m} 1_{i\in\mathcal{B}} \nabla f_i\left(\vec{x}^{(k)}\right)\right) = \sum_{i=1}^{m} \mathrm{E}\left(1_{i\in\mathcal{B}}\right) \nabla f_i\left(\vec{x}^{(k)}\right)$$

Stochastic gradient descent

If samples are uniform with replacement, then estimate is aligned with ∇f on average

$$\begin{aligned}\mathrm{E}\left(\sum_{i=1}^{m} 1_{i\in\mathcal{B}} \nabla f_i\left(\vec{x}^{(k)}\right)\right) &= \sum_{i=1}^{m} \mathrm{E}\left(1_{i\in\mathcal{B}}\right) \nabla f_i\left(\vec{x}^{(k)}\right) \\ &= \sum_{i=1}^{m} \mathrm{P}\left(i \in \mathcal{B}\right) \nabla f_i\left(\vec{x}^{(k)}\right)\end{aligned}$$

Stochastic gradient descent

If samples are uniform with replacement, then estimate is aligned with ∇f on average

$$\begin{aligned}
\mathrm{E}\left(\sum_{i=1}^{m} 1_{i\in\mathcal{B}} \nabla f_i\left(\vec{x}^{(k)}\right)\right) &= \sum_{i=1}^{m} \mathrm{E}\left(1_{i\in\mathcal{B}}\right) \nabla f_i\left(\vec{x}^{(k)}\right) \\
&= \sum_{i=1}^{m} \mathrm{P}\left(i \in \mathcal{B}\right) \nabla f_i\left(\vec{x}^{(k)}\right) \\
&= \frac{m}{b} \nabla f\left(\vec{x}^{(k)}\right)
\end{aligned}$$

Stochastic gradient descent

If samples are uniform with replacement, then estimate is aligned with ∇f on average

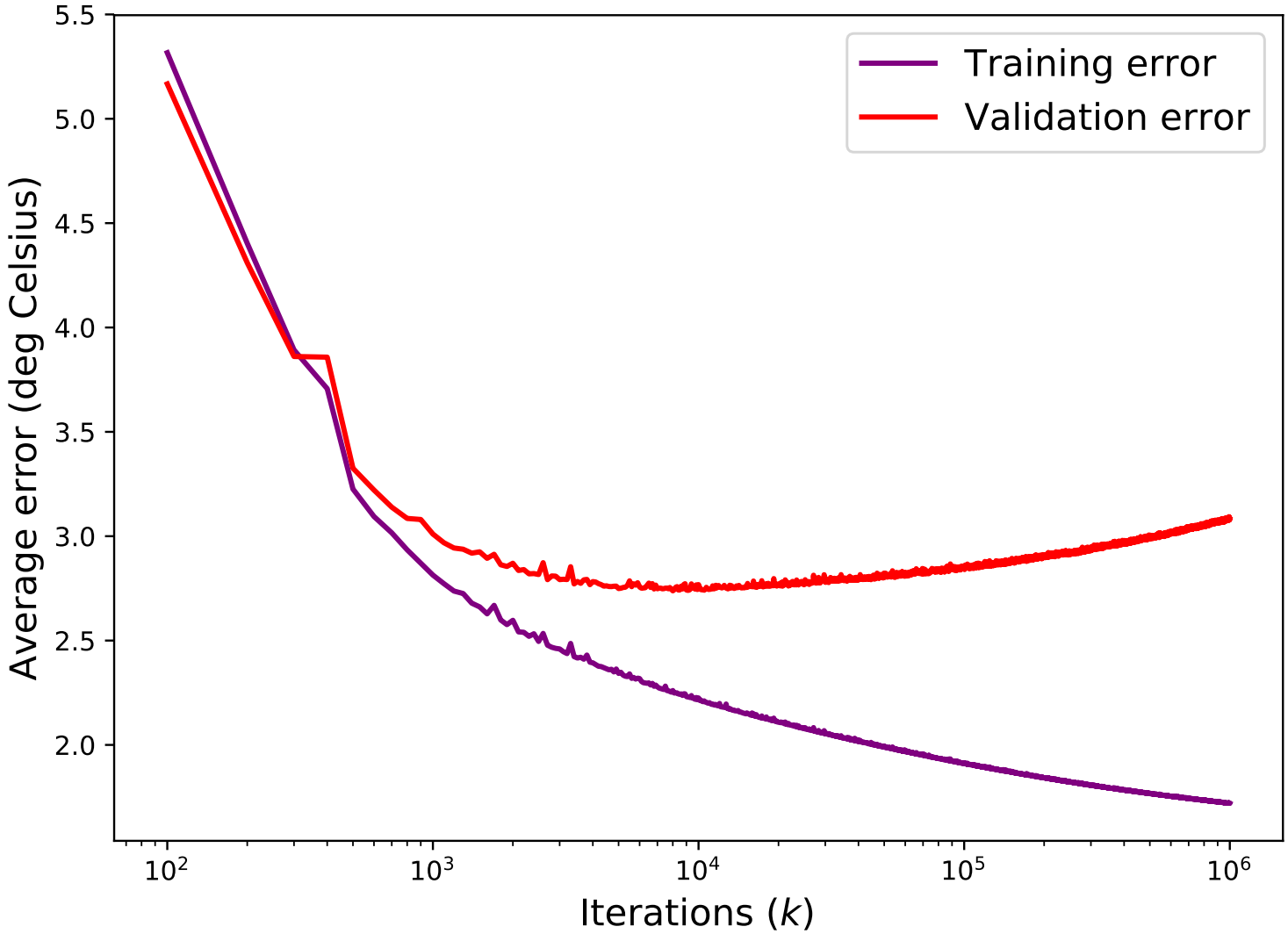
$$\begin{aligned}
\mathrm{E}\left(\sum_{i=1}^{m} 1_{i\in\mathcal{B}} \nabla f_i\left(\vec{x}^{(k)}\right)\right) &= \sum_{i=1}^{m} \mathrm{E}\left(1_{i\in\mathcal{B}}\right) \nabla f_i\left(\vec{x}^{(k)}\right) \\
&= \sum_{i=1}^{m} \mathrm{P}\left(i \in \mathcal{B}\right) \nabla f_i\left(\vec{x}^{(k)}\right) \\
&= \frac{m}{b} \nabla f\left(\vec{x}^{(k)}\right)
\end{aligned}$$

Due to variance, α_k needs to be decreasing in k

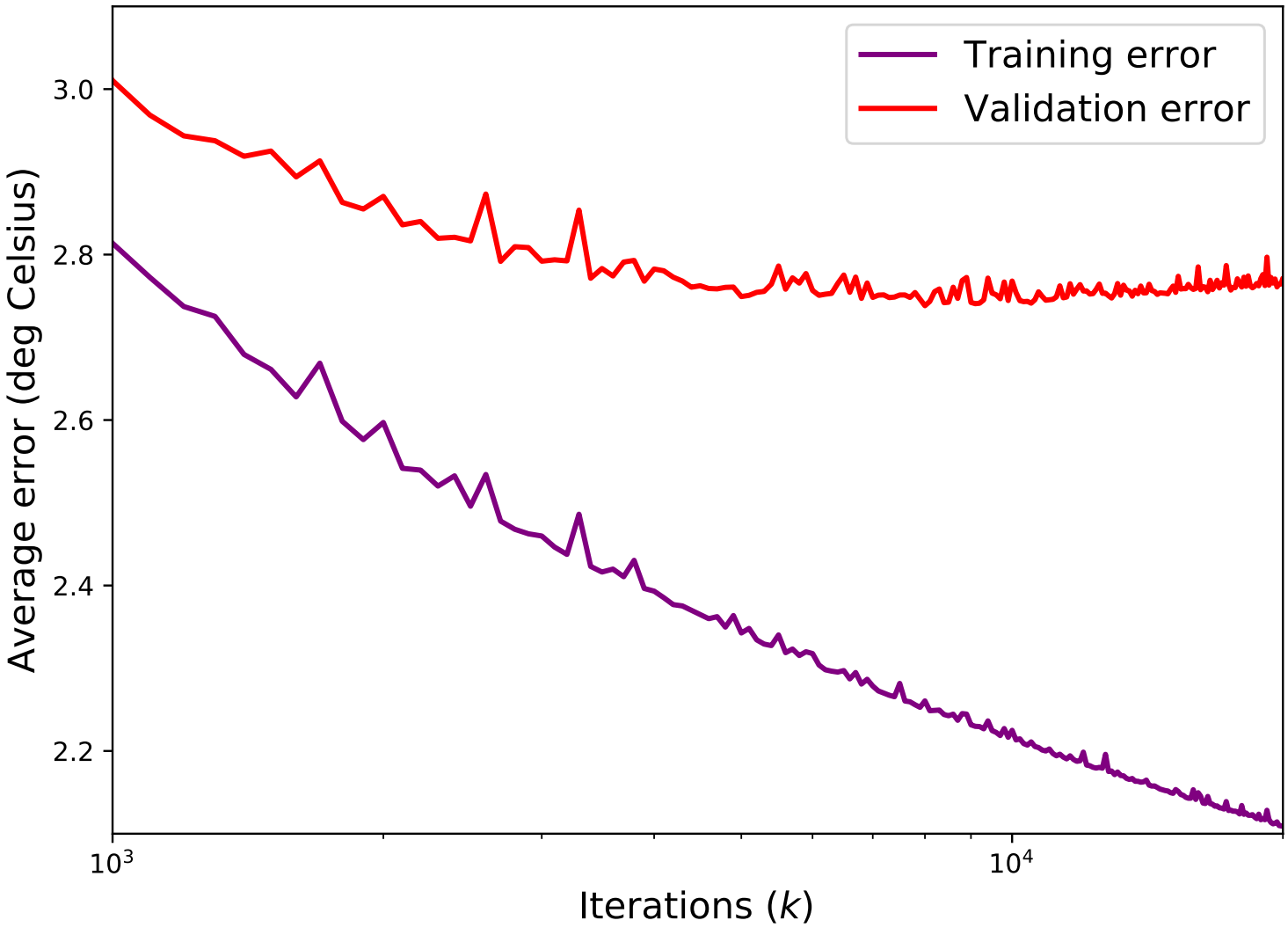
Linear regression

$$\vec{\beta}^{(k+1)} := \vec{\beta}^{(k)} + 2\alpha_k \sum_{i \in \mathcal{B}} \left(\vec{y}^{(i)} - \langle x^{(i)}, \vec{\beta}^{(k)} \rangle \right) x^{(i)}$$

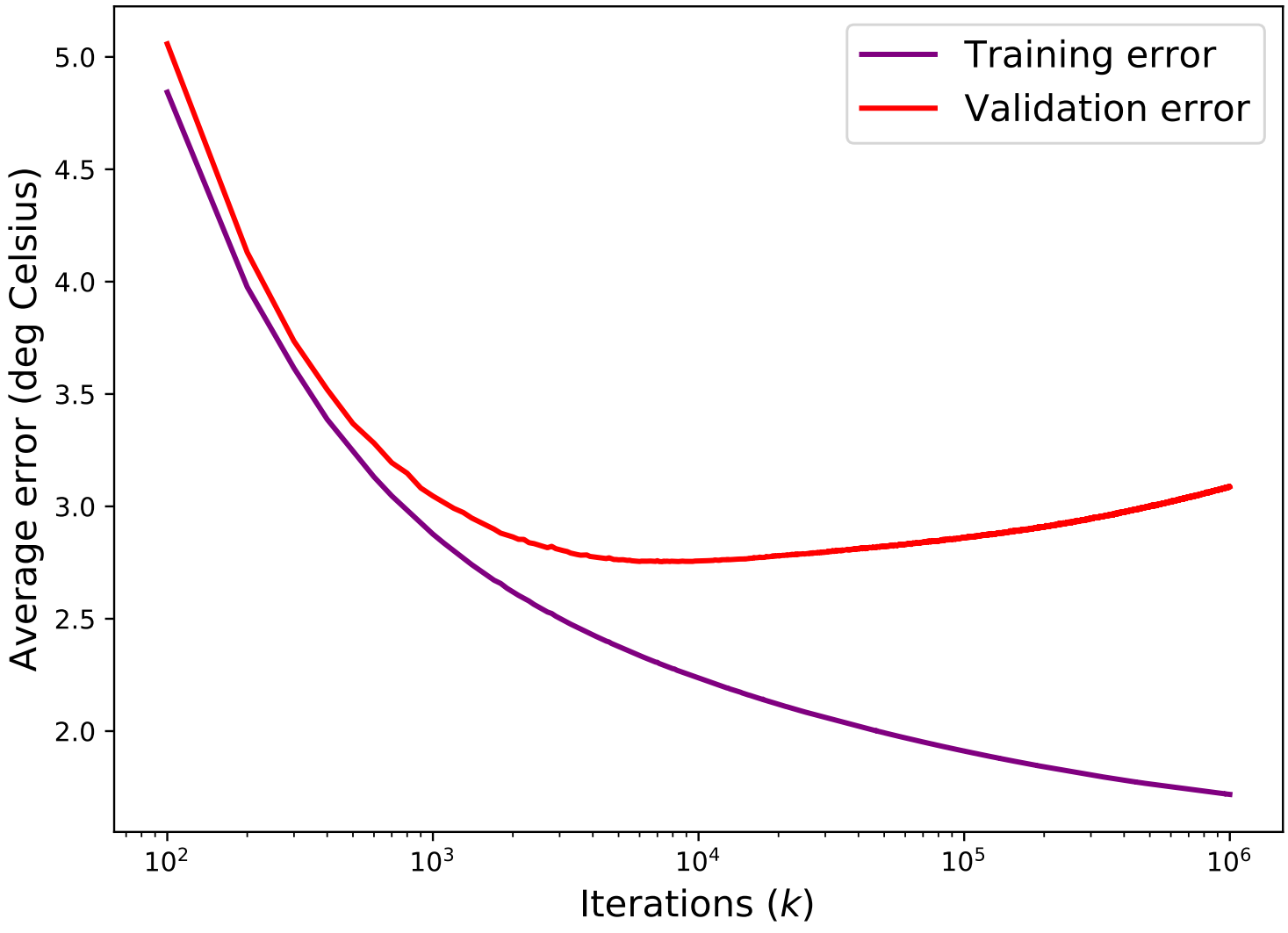
Temperature prediction, batch size = 10



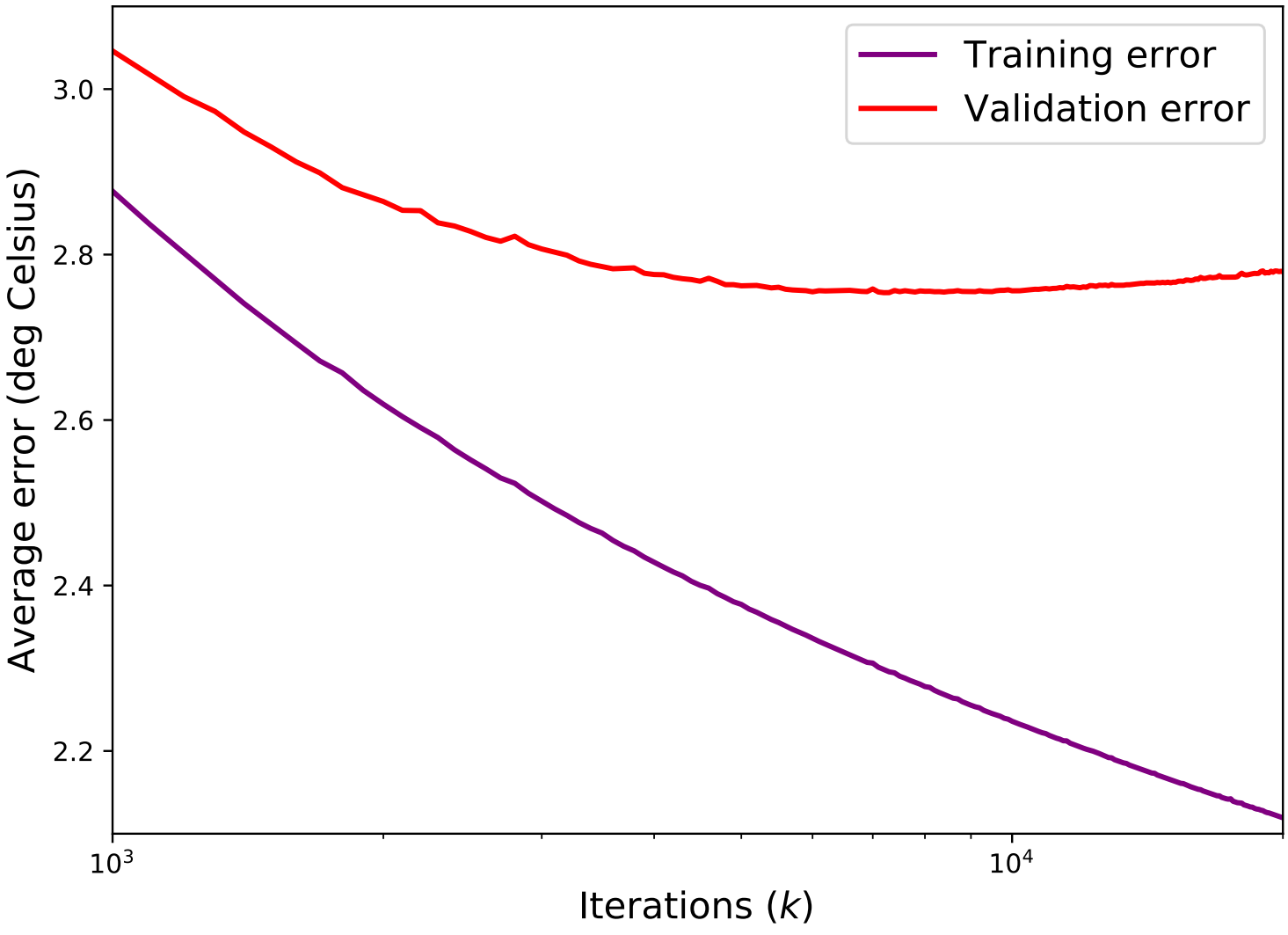
Temperature prediction, batch size = 10



Temperature prediction, batch size = 100



Temperature prediction, batch size = 100



Composite functions

Interesting class of functions for data analysis

$$f\left(\vec{x}\right) + h\left(\vec{x}\right)$$

f convex and differentiable, h convex but not differentiable

Example: Least squares with ℓ_1-norm or nuclear-norm regularization

Motivation

Aim: Minimize convex differentiable function f

Idea: Iteratively minimize first-order approximation, while staying close to current point

$\vec{x}^{(0)} =$ arbitrary initialization

$$\vec{x}^{(k+1)} = \arg\min_{\vec{x}} f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(\vec{x} - \vec{x}^{(k)}\right) + \frac{1}{2\,\alpha_k} \left|\left|\vec{x} - \vec{x}^{(k)}\right|\right|_2^2$$

where α_k is a parameter that determines how close we stay

Motivation

Linear approximation+ ℓ_2 term is convex

$$\nabla \left(f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(\vec{x} - \vec{x}^{(k)}\right) + \frac{1}{2\,\alpha_k} \left|\left|\vec{x} - \vec{x}^{(k)}\right|\right|_2^2 \right)$$

Motivation

Linear approximation+ ℓ_2 term is convex

$$\nabla \left(f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(\vec{x} - \vec{x}^{(k)}\right) + \frac{1}{2\,\alpha_k} \left|\left|\vec{x} - \vec{x}^{(k)}\right|\right|_2^2 \right)$$

$$= \nabla f\left(\vec{x}^{(k)}\right) + \frac{\vec{x} - \vec{x}^{(k)}}{\alpha_k}$$

Motivation

Linear approximation+ ℓ_2 term is convex

$$\nabla \left(f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(\vec{x} - \vec{x}^{(k)}\right) + \frac{1}{2\,\alpha_k} \left|\left|\vec{x} - \vec{x}^{(k)}\right|\right|_2^2 \right)$$

$$= \nabla f\left(\vec{x}^{(k)}\right) + \frac{\vec{x} - \vec{x}^{(k)}}{\alpha_k}$$

Setting the gradient to zero

$$\vec{x}^{(k+1)} = \arg\min_{\vec{x}} f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(\vec{x} - \vec{x}^{(k)}\right) + \frac{1}{2\,\alpha_k} \left|\left|\vec{x} - \vec{x}^{(k)}\right|\right|_2^2$$

Motivation

Linear approximation+ ℓ_2 term is convex

$$\nabla \left(f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(\vec{x} - \vec{x}^{(k)}\right) + \frac{1}{2\,\alpha_k} \left|\left|\vec{x} - \vec{x}^{(k)}\right|\right|_2^2 \right)$$
$$= \nabla f\left(\vec{x}^{(k)}\right) + \frac{\vec{x} - \vec{x}^{(k)}}{\alpha_k}$$

Setting the gradient to zero

$$\vec{x}^{(k+1)} = \arg\min_{\vec{x}} f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(\vec{x} - \vec{x}^{(k)}\right) + \frac{1}{2\,\alpha_k} \left|\left|\vec{x} - \vec{x}^{(k)}\right|\right|_2^2$$
$$= \vec{x}^{(k)} - \alpha_k \nabla f\left(\vec{x}^{(k)}\right)$$

Proximal gradient method

Idea: Minimize local first-order approximation + h

$$\begin{aligned}\vec{x}^{(k+1)} &= \arg\min_{\vec{x}} f\left(\vec{x}^{(k)}\right) + \nabla f\left(\vec{x}^{(k)}\right)^T \left(\vec{x} - \vec{x}^{(k)}\right) + \frac{1}{2\,\alpha_k}\left|\left|\vec{x} - \vec{x}^{(k)}\right|\right|_2^2 \\ &\quad + h\left(\vec{x}\right) \\ &= \arg\min_{\vec{x}} \frac{1}{2}\left|\left|x - \left(\vec{x}^{(k)} - \alpha_k\,\nabla f\left(\vec{x}^{(k)}\right)\right)\right|\right|_2^2 + \alpha_k\, h\left(\vec{x}\right) \\ &= \mathrm{prox}_{\alpha_k\, h}\left(\vec{x}^{(k)} - \alpha_k\,\nabla f\left(\vec{x}^{(k)}\right)\right)\end{aligned}$$

Proximal operator:

$$\mathrm{prox}_h\left(y\right) := \arg\min_{\vec{x}} h\left(\vec{x}\right) + \frac{1}{2}\left|\left|y - x\right|\right|_2^2$$

Proximal gradient method

Method to solve the optimization problem

$$\text{minimize} \quad f\left(\vec{x}\right) + h\left(\vec{x}\right),$$

where f is differentiable and prox_h is tractable

Proximal-gradient iteration:

$$\vec{x}^{(0)} = \text{arbitrary initialization}$$

$$\vec{x}^{(k+1)} = \text{prox}_{\alpha_k h}\left(\vec{x}^{(k)} - \alpha_k \nabla f\left(\vec{x}^{(k)}\right)\right)$$

Interpretation as a fixed-point method

A vector $\vec{x}^*$ is a solution to

$$\text{minimize} \quad f\left(\vec{x}\right) + h\left(\vec{x}\right),$$

if and only if it is a fixed point of the proximal-gradient iteration for any $\alpha > 0$

$$\vec{x}^* = \text{prox}_{\alpha\, h}\left(\vec{x}^* - \alpha\, \nabla f\left(\vec{x}^*\right)\right)$$

Proof

$\vec{x}^*$ is the solution to

$$\min_{\vec{x}} \quad \alpha\, h\left(\vec{x}\right) + \frac{1}{2}\left|\left|\vec{x}^* - \alpha\, \nabla f\left(\vec{x}^*\right) - \vec{x}\right|\right|_2^2$$

if and only if there is a subgradient $\vec{g}$ of h at $\vec{x}^*$ such that

Proof

$\vec{x}^*$ is the solution to

$$\min_{\vec{x}} \quad \alpha\, h\left(\vec{x}\right) + \frac{1}{2}\left|\left|\vec{x}^* - \alpha\, \nabla f\left(\vec{x}^*\right) - \vec{x}\right|\right|_2^2$$

if and only if there is a subgradient $\vec{g}$ of h at $\vec{x}^*$ such that $\alpha \nabla f\left(\vec{x}^*\right) + \alpha \vec{g} = \vec{0}$

Proof

$\vec{x}^*$ is the solution to

$$\min_{\vec{x}} \quad \alpha\, h\left(\vec{x}\right) + \frac{1}{2}\left|\left|\vec{x}^* - \alpha\, \nabla f\left(\vec{x}^*\right) - \vec{x}\right|\right|_2^2$$

if and only if there is a subgradient $\vec{g}$ of h at $\vec{x}^*$ such that
$\alpha \nabla f\left(\vec{x}^*\right) + \alpha \vec{g} = \vec{0}$

$\vec{x}^*$ minimizes $f + h$ if and only if there is a subgradient $\vec{g}$ of h at $\vec{x}^*$ such that

Proof

$\vec{x}^*$ is the solution to

$$\min_{\vec{x}} \quad \alpha\, h(\vec{x}) + \frac{1}{2}\,||\vec{x}^* - \alpha\,\nabla f(\vec{x}^*) - \vec{x}||_2^2$$

if and only if there is a subgradient $\vec{g}$ of h at $\vec{x}^*$ such that $\alpha\nabla f(\vec{x}^*) + \alpha\vec{g} = \vec{0}$

$\vec{x}^*$ minimizes $f + h$ if and only if there is a subgradient $\vec{g}$ of h at $\vec{x}^*$ such that $\nabla f(\vec{x}^*) + \vec{g} = \vec{0}$

Proximal operator of ℓ_1 norm

The proximal operator of the ℓ_1 norm is the soft-thresholding operator

$$\text{prox}_{\alpha \|\cdot\|_1}(y) = \mathcal{S}_\alpha(y)$$

where $\alpha > 0$ and

$$\mathcal{S}_\alpha(y)_i := \begin{cases} y_i - \text{sign}(y_i)\,\alpha & \text{if } |y_i| \geq \alpha \\ 0 & \text{otherwise} \end{cases}$$

Iterative Shrinkage-Thresholding Algorithm (ISTA)

The proximal gradient method for the problem

$$\text{minimize} \quad \frac{1}{2}\,||A\vec{x} - \vec{y}||_2^2 + \lambda\,||\vec{x}||_1$$

is called ISTA

ISTA iteration:

$$\vec{x}^{(0)} = \text{arbitrary initialization}$$
$$\vec{x}^{(k+1)} = \mathcal{S}_{\alpha_k\,\lambda}\left(\vec{x}^{(k)} - \alpha_k\,A^T\left(A\vec{x}^{(k)} - \vec{y}\right)\right)$$

Proximal operator of nuclear norm

The solution X to

$$\min_{X \in \mathbb{R}^{m \times n}} \frac{1}{2} ||Y - X||_F^2 + \tau ||X||_*$$

is obtained by soft-thresholding the SVD of Y

$$X_{\text{prox}} = \mathcal{D}_\tau (Y)$$

$$\mathcal{D}_\tau (M) := U \mathcal{S}_\tau (S) V^T \qquad \text{where } M = U S V^T$$

$$\mathcal{S}_\tau (S)_{ii} := \begin{cases} S_{ii} - \tau & \text{if } S_{ii} > \tau \\ 0 & \text{otherwise} \end{cases}$$

Proximal gradient method

Proximal gradient method for the problem

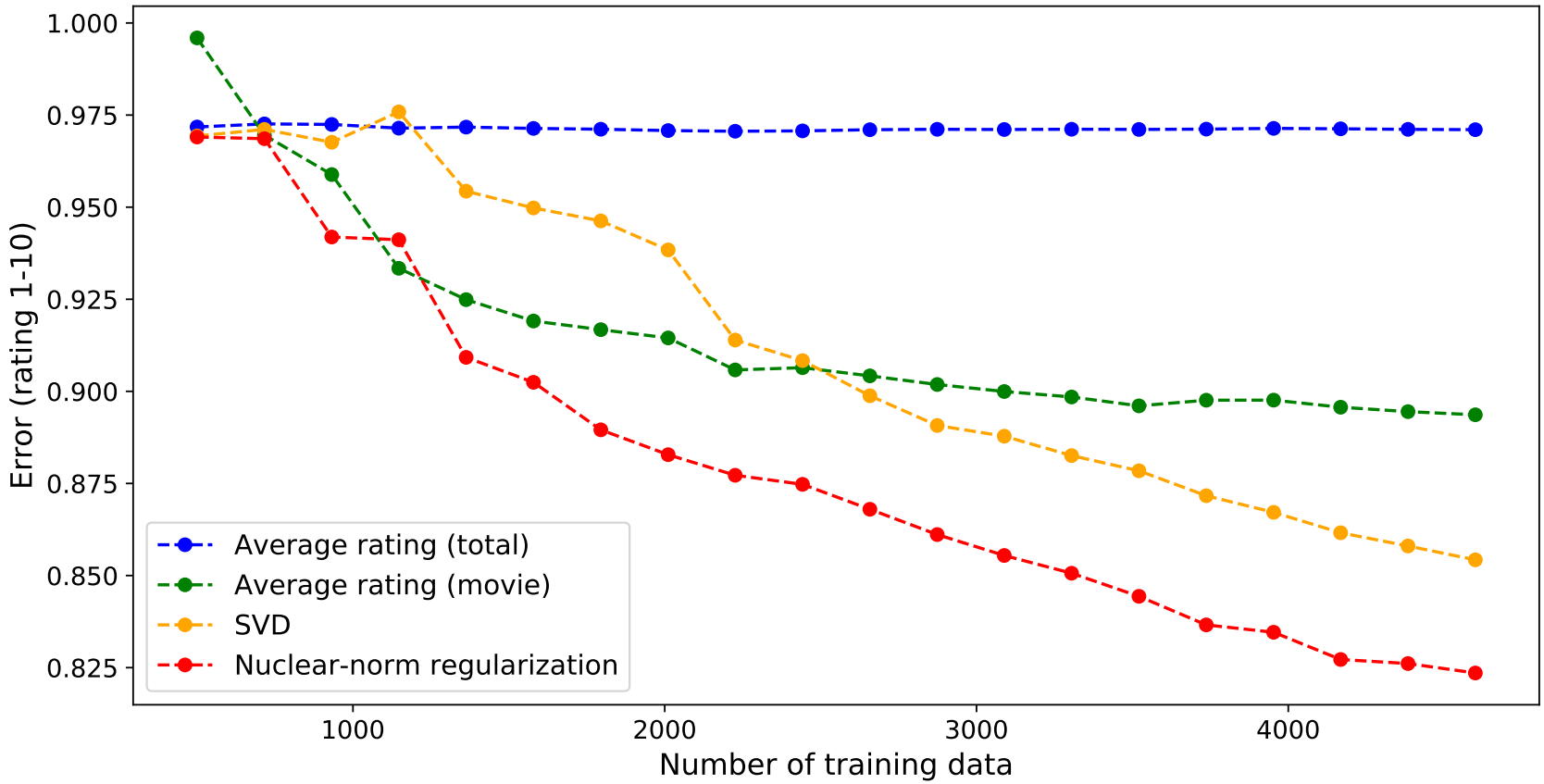
$$\min_{X \in \mathbb{R}^{m \times n}} ||X_\Omega - \vec{y}||_2^2 + \lambda ||X||_*$$

$$X^{(0)} = \text{arbitrary initialization}$$

$$M^{(k)} = X^{(k)} - \alpha_k \left(X_\Omega^{(k)} - \vec{y} \right)$$

$$X^{(k+1)} = \mathcal{D}_{\alpha_k \lambda} \left(M^{(k)} \right)$$

Results


1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
Error (rating 1-10)
1000
2000
3000
4000
Number of training data
Average rating (total)
Average rating (movie)
SVD
Nuclear-norm regularization