Sequential and Regularized Least Squares, Static Nonlinear Models and Gradient Descent

Roland Hostettler

September 13, 2018

Recap

- The general linear model is given by

$$\boldsymbol{y} = \boldsymbol{G}\boldsymbol{\theta} + \boldsymbol{r},\ \mathrm{E}\{\boldsymbol{r}\} = 0,\ \mathrm{Cov}\{\boldsymbol{r}\} = \boldsymbol{R}$$

- The weighted linear least squares estimator is

$$\hat{\boldsymbol{\theta}}_{WLS} = (\boldsymbol{G}^{\mathsf{T}}\boldsymbol{R}^{-1}\boldsymbol{G})^{-1}\boldsymbol{G}^{\mathsf{T}}\boldsymbol{R}^{-1}\boldsymbol{y}$$

  with $\mathrm{E}\{\hat{\boldsymbol{\theta}}_{WLS}\} = \boldsymbol{\theta}$ and $\mathrm{Cov}\{\hat{\boldsymbol{\theta}}_{WLS}\} = (\boldsymbol{G}^{\mathsf{T}}\boldsymbol{R}^{-1}\boldsymbol{G})^{-1}$.
- The LS estimator is a special case of the WLS squares estimator with $\boldsymbol{R} \propto \boldsymbol{I}$
- It holds that

$$\mathrm{Cov}\{\hat{\boldsymbol{\theta}}_{\mathrm{WLS}}\} \leq \mathrm{Cov}\{\hat{\boldsymbol{\theta}}_{\mathrm{LS}}\}$$

Intended Learning Outcomes

After this lecture, you will be able to:

- Describe and identify sequential and regularized linear least squares estimators;
- distinguish (weighted) linear least squares and regularized linear least squares estimators and their properties;
- identify the challenges encountered in nonlinear sensor models;
- restate the gradient descent algorithm for nonlinear least squares.

Sequential Linear Least Squares: Formulation

- Sometimes data arrives sequentially; how can we update an existing estimate with the new data?
- Given the data $\boldsymbol{y}_{1:n-1} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{n-1}\}$, we have

$$\hat{\boldsymbol{\theta}}_{n-1} = (\boldsymbol{G}^{\mathsf{T}}\boldsymbol{R}^{-1}\boldsymbol{G})^{-1}\boldsymbol{G}^{\mathsf{T}}\boldsymbol{R}^{-1}\boldsymbol{y}$$
$$\mathrm{Cov}\{\hat{\boldsymbol{\theta}}_{n-1}\} = (\boldsymbol{G}^{\mathsf{T}}\boldsymbol{R}^{-1}\boldsymbol{G})^{-1} \triangleq \boldsymbol{P}_{n-1}$$

- Updated cost function when $\boldsymbol{y}_n$ arrives:

$$J_{\mathsf{SLS}}(\boldsymbol{\theta}) = (\boldsymbol{y} - \boldsymbol{G}\boldsymbol{\theta})^{\mathsf{T}}\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{G}\boldsymbol{\theta}) + (\boldsymbol{y}_n - \boldsymbol{C}_n\boldsymbol{\theta})^{\mathsf{T}}\boldsymbol{R}_n^{-1}(\boldsymbol{y}_n - \boldsymbol{C}_n\boldsymbol{\theta}),$$

Sequential Linear Least Squares: Derivation (1/2)

- Updated cost function when $\boldsymbol{y}_n$ arrives:

$$J_{\text{SLS}}(\boldsymbol{\theta}) = (\boldsymbol{y} - \boldsymbol{G}\boldsymbol{\theta})^{\mathsf{T}} \boldsymbol{R}^{-1} (\boldsymbol{y} - \boldsymbol{G}\boldsymbol{\theta}) + (\boldsymbol{y}_n - \boldsymbol{C}_n \boldsymbol{\theta})^{\mathsf{T}} \boldsymbol{R}_n^{-1} (\boldsymbol{y}_n - \boldsymbol{C}_n \boldsymbol{\theta}),$$

- Gradient:

$$\frac{\partial J_{\text{SLS}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -2(\boldsymbol{G}^{\mathsf{T}} \boldsymbol{R}^{-1} \boldsymbol{y} + \boldsymbol{C}_n^{\mathsf{T}} \boldsymbol{R}_n^{-1} \boldsymbol{y}_n) + 2(\boldsymbol{G}^{\mathsf{T}} \boldsymbol{R}^{-1} \boldsymbol{G} + \boldsymbol{C}_n^{\mathsf{T}} \boldsymbol{R}_n^{-1} \boldsymbol{C}_n)\boldsymbol{\theta}$$

- Updated least squares estimator:

$$\hat{\boldsymbol{\theta}}_n = (\boldsymbol{P}_{n-1}^{-1} + \boldsymbol{C}_n^{\mathsf{T}} \boldsymbol{R}_n^{-1} \boldsymbol{C}_n)^{-1} (\boldsymbol{G}^{\mathsf{T}} \boldsymbol{R}^{-1} \boldsymbol{y} + \boldsymbol{C}_n^{\mathsf{T}} \boldsymbol{R}_n^{-1} \boldsymbol{y}_n).$$

Sequential Linear Least Squares: Derivation (2/2)

- Updated least squares estimator:

$$\hat{\boldsymbol{\theta}}_n = (\boldsymbol{P}_{n-1}^{-1} + \boldsymbol{C}_n^{\mathsf{T}} \boldsymbol{R}_n^{-1} \boldsymbol{C}_n)^{-1} (\boldsymbol{G}^{\mathsf{T}} \boldsymbol{R}^{-1} \boldsymbol{y} + \boldsymbol{C}_n^{\mathsf{T}} \boldsymbol{R}_n^{-1} \boldsymbol{y}_n).$$

  $\Rightarrow$ Problem: Does not re-use $\hat{\boldsymbol{\theta}}_{n-1}$ and $\boldsymbol{P}_{n-1}$ well

- Alternative form:

$$\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_{n-1} + \boldsymbol{L}_n (\boldsymbol{y}_n - \boldsymbol{C}_n \hat{\boldsymbol{\theta}}_{n-1})$$

  with $\boldsymbol{L}_n = \boldsymbol{P}_{n-1} \boldsymbol{C}_n^{\mathsf{T}} (\boldsymbol{C}_n \boldsymbol{P}_{n-1} \boldsymbol{C}_n^{\mathsf{T}} + \boldsymbol{R}_n)^{-1}$.

Sequential Linear Least Squares: Properties

- Sequential linear least squares:

$$\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_{n-1} + \boldsymbol{L}_n(\boldsymbol{y}_n - \boldsymbol{C}_n\hat{\boldsymbol{\theta}}_{n-1})$$
$$\boldsymbol{L}_n = \boldsymbol{P}_{n-1}\boldsymbol{C}_n^{\mathsf{T}}(\boldsymbol{C}_n\boldsymbol{P}_{n-1}\boldsymbol{C}_n^{\mathsf{T}} + \boldsymbol{R}_n)^{-1}$$

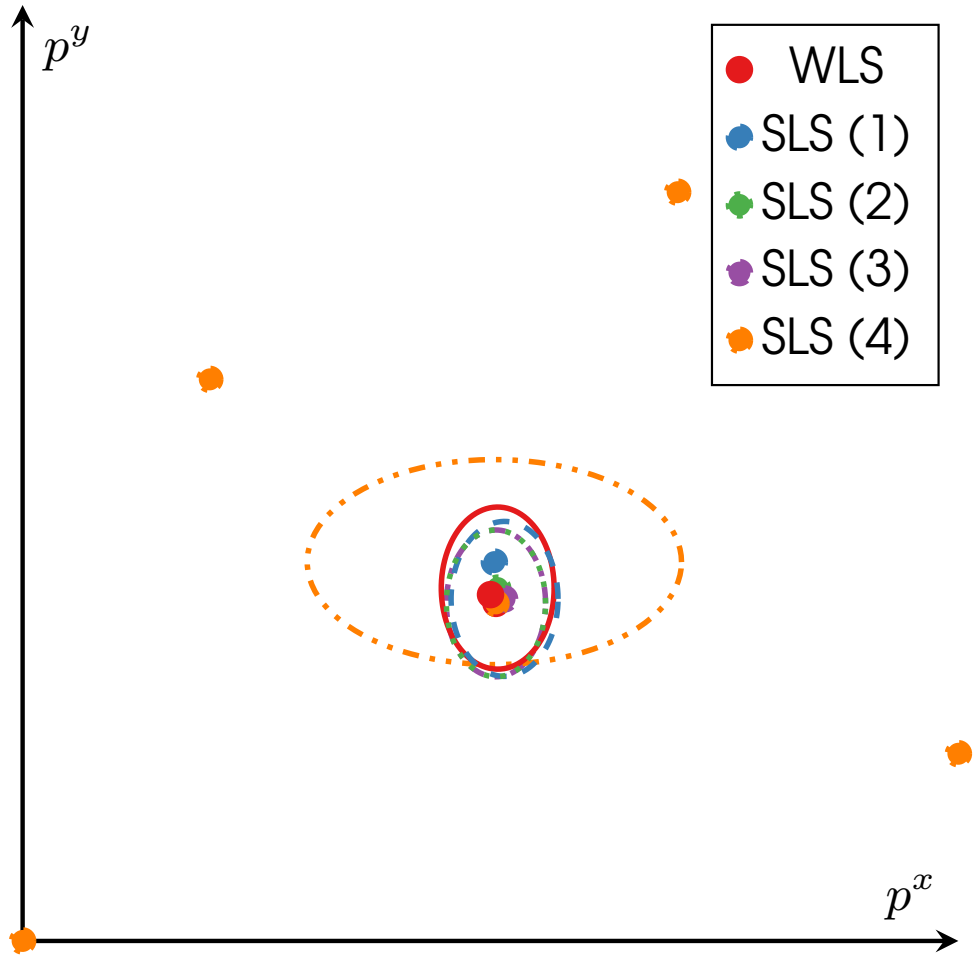
- Mean:

$$\begin{aligned}\mathrm{E}\{\hat{\boldsymbol{\theta}}_n\} &= \mathrm{E}\{\hat{\boldsymbol{\theta}}_{n-1}\} + \boldsymbol{L}_n(\mathrm{E}\{\boldsymbol{y}_n\} - \boldsymbol{C}_n\,\mathrm{E}\{\hat{\boldsymbol{\theta}}_{n-1}\}) \\ &= \boldsymbol{\theta} + \boldsymbol{L}_n(\boldsymbol{C}_n\boldsymbol{\theta} - \boldsymbol{C}_n\boldsymbol{\theta}) \\ &= \boldsymbol{\theta},\end{aligned}$$

- Covariance:

$$\mathrm{Cov}\{\hat{\boldsymbol{\theta}}_n\} = \boldsymbol{P}_{n-1} - \boldsymbol{L}_n(\boldsymbol{C}_n\boldsymbol{P}_{n-1}\boldsymbol{C}^{\mathsf{T}} + \boldsymbol{R}_n)\boldsymbol{L}_n^{\mathsf{T}}.$$

Example: Localizing a Target (1)



Regularized Linear Least Squares: Formulation

- Sometimes we have prior knowledge about $\boldsymbol{\theta}$
- For example, we could assume that

$$\mathrm{E}\{\boldsymbol{\theta}\} = \boldsymbol{m} \text{ and } \mathrm{Cov}\{\boldsymbol{\theta}\} = \boldsymbol{P}$$

- Cost function with regularization term:

$$J_{\text{ReLS}}(\boldsymbol{\theta}) = \underbrace{(\boldsymbol{y} - \boldsymbol{G}\boldsymbol{\theta})^{\mathsf{T}} \boldsymbol{R}^{-1} (\boldsymbol{y} - \boldsymbol{G}\boldsymbol{\theta})}_{\text{Data}} + \underbrace{(\boldsymbol{\theta} - \boldsymbol{m})^{\mathsf{T}} \boldsymbol{P}^{-1} (\boldsymbol{\theta} - \boldsymbol{m})}_{\text{Prior Knowledge}}$$

- This corresponds to a *Bayesian linear model*

Regularized Linear Least Squares: Derivation

- Gradient of the cost function:

$$\begin{aligned}\frac{\partial J_{\text{ReLS}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{\partial}{\partial \boldsymbol{\theta}}(\boldsymbol{y}-\boldsymbol{G}\boldsymbol{\theta})^{\mathsf{T}}\boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{G}\boldsymbol{\theta}) \\ &\quad + \frac{\partial}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}-\boldsymbol{m})^{\mathsf{T}}\boldsymbol{P}^{-1}(\boldsymbol{\theta}-\boldsymbol{m}) \\ &= -2\boldsymbol{G}^{\mathsf{T}}\boldsymbol{R}^{-1}\boldsymbol{y} + 2\boldsymbol{G}^{\mathsf{T}}\boldsymbol{R}^{-1}\boldsymbol{G}\boldsymbol{\theta} - 2\boldsymbol{P}^{-1}\boldsymbol{m} + 2\boldsymbol{P}^{-1}\boldsymbol{\theta} \\ &= -2(\boldsymbol{G}^{\mathsf{T}}\boldsymbol{R}^{-1}\boldsymbol{y} + \boldsymbol{P}^{-1}\boldsymbol{m}) + 2(\boldsymbol{G}^{\mathsf{T}}\boldsymbol{R}^{-1}\boldsymbol{G} + \boldsymbol{P}^{-1})\boldsymbol{\theta}\end{aligned}$$

- Setting to zero and solving for $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}}_{\text{ReLS}} = (\boldsymbol{G}^{\mathsf{T}}\boldsymbol{R}^{-1}\boldsymbol{G} + \boldsymbol{P}^{-1})^{-1}(\boldsymbol{G}^{\mathsf{T}}\boldsymbol{R}^{-1}\boldsymbol{y} + \boldsymbol{P}^{-1}\boldsymbol{m})$$

- Alternative formulation (using matrix inversion lemma):

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\text{ReLS}} &= \boldsymbol{m} + \boldsymbol{K}(\boldsymbol{y}-\boldsymbol{G}\boldsymbol{m}) \\ \boldsymbol{K} &= \boldsymbol{P}\boldsymbol{G}^{\mathsf{T}}(\boldsymbol{G}\boldsymbol{P}\boldsymbol{G}^{\mathsf{T}} + \boldsymbol{R})^{-1}\end{aligned}$$

Regularized Linear Least Squares: Properties

- Regularized linear least squares:

$$\hat{\boldsymbol{\theta}}_{\text{ReLS}} = \boldsymbol{m} + \boldsymbol{K}(\boldsymbol{y} - \boldsymbol{G}\boldsymbol{m})$$
$$\boldsymbol{K} = \boldsymbol{P}\boldsymbol{G}^{\mathsf{T}}(\boldsymbol{G}\boldsymbol{P}\boldsymbol{G}^{\mathsf{T}} + \boldsymbol{R})^{-1}$$

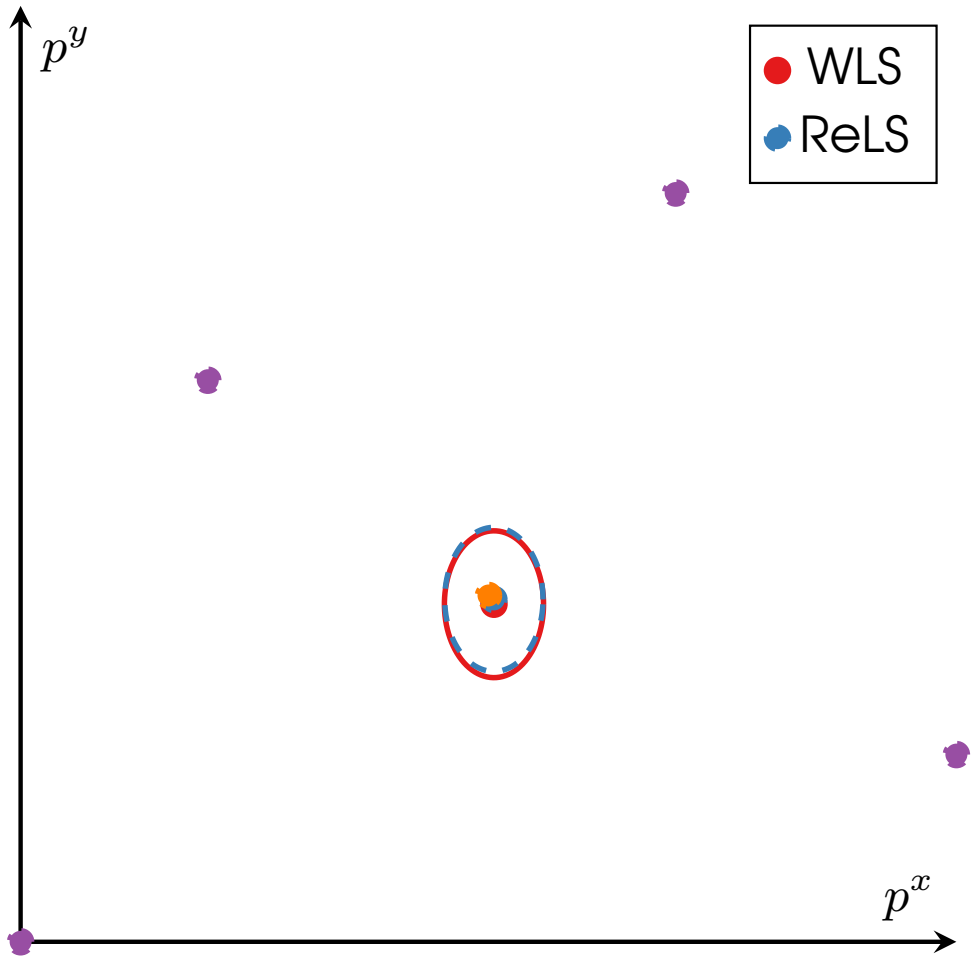
- Mean:

$$\begin{aligned}\mathrm{E}\{\hat{\boldsymbol{\theta}}_{\text{ReLS}}\} &= \mathrm{E}\{\boldsymbol{m} + \boldsymbol{K}(\boldsymbol{y} - \boldsymbol{G}\boldsymbol{m})\} \\ &= \boldsymbol{m} + \boldsymbol{K}(\boldsymbol{G}\,\mathrm{E}\{\boldsymbol{\theta}\} - \boldsymbol{G}\boldsymbol{m}).\end{aligned}$$

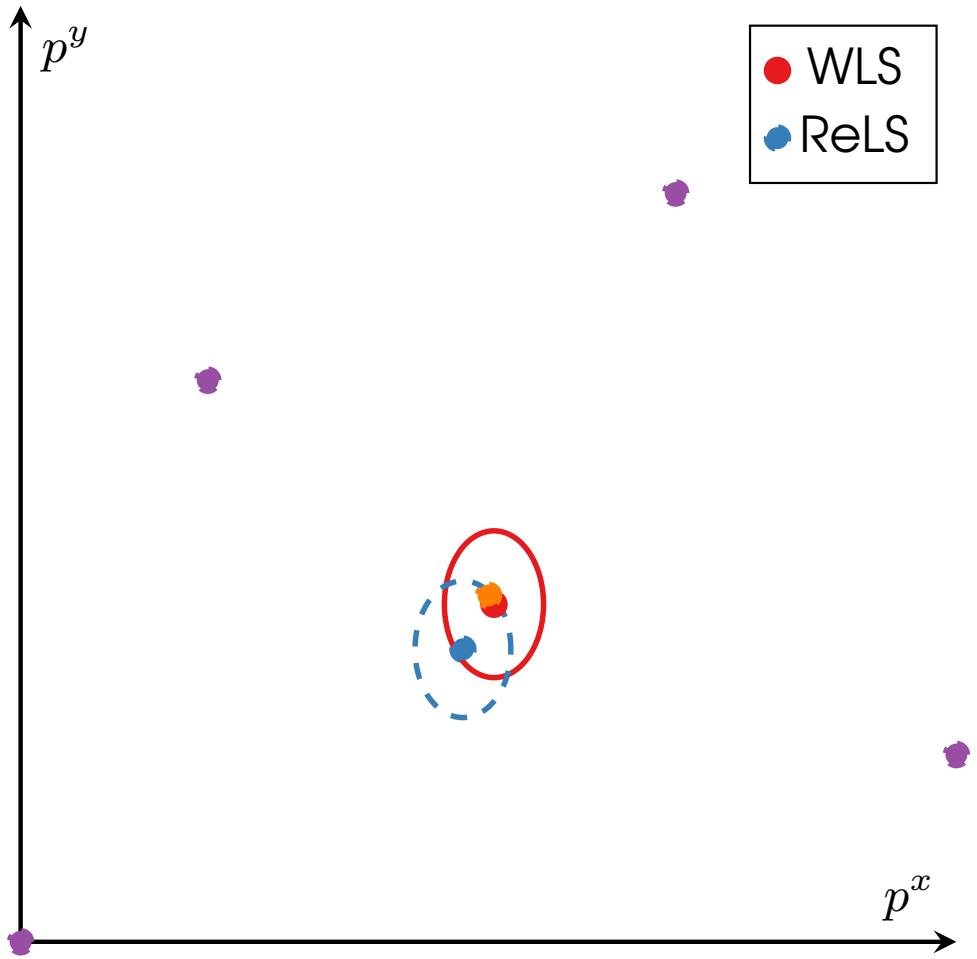
- Covariance:

$$\mathrm{Cov}\{\hat{\boldsymbol{\theta}}_{\text{ReLS}}\} = \boldsymbol{P} - \boldsymbol{K}(\boldsymbol{G}\boldsymbol{P}\boldsymbol{G}^{\mathsf{T}} + \boldsymbol{R})\boldsymbol{K}^{\mathsf{T}}.$$

Example: Localizing a Target (2)



Example: Localizing a Target (3)



Static Nonlinear Models

- Linear models have closed form solutions, but are limited in many cases
- General nonlinear model:

$$\boldsymbol{y} = g(\boldsymbol{\theta}) + \boldsymbol{r},$$

- General cost function:

$$J_{\text{WLS}}(\boldsymbol{\theta}) = (\boldsymbol{y} - g(\boldsymbol{\theta}))^{\mathsf{T}} \boldsymbol{R}^{-1} (\boldsymbol{y} - g(\boldsymbol{\theta})).$$

- For some models, closed form solutions do exist, ...
- ...but for most they do not

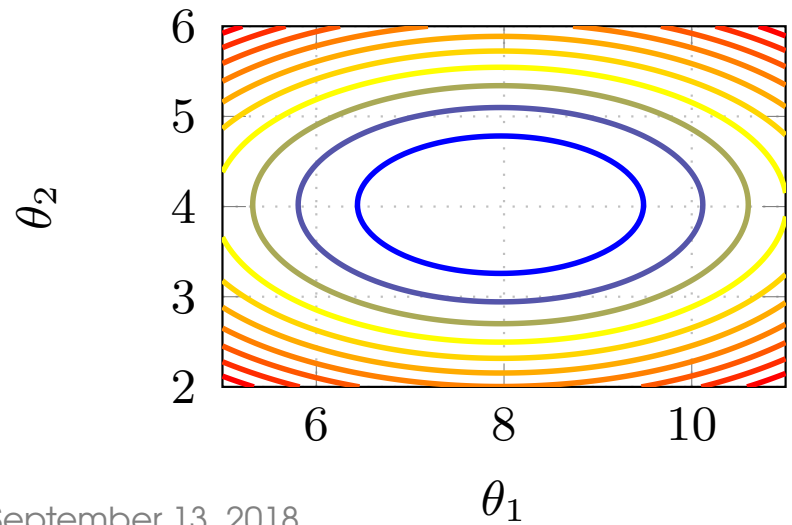
Numerical Optimization

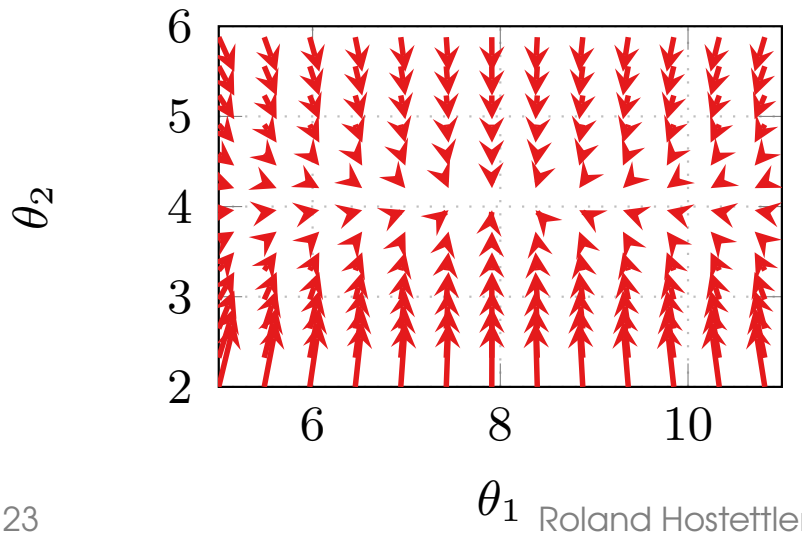
- Iterative algorithms to find the stationary points or roots of a function
- Generally find local minima $\Rightarrow$ Require good initialization
- Two predominant approaches:
  1. Methods that find the minima of a function, that is, methods that solve
  $$x = \underset{x}{\operatorname{argmin}}\, f(x)$$
  2. Methods that find the zeros of the gradient, that is, methods that solve
  $$\frac{\partial f(x)}{\partial x} = 0$$

Gradient Descent: Formulation

- The *gradient* of $f(x)$ w.r.t. x points to the direction where $f(x)$ increases as a function of x
- Changing x in the opposite direction of the gradient decreases $f(x)$
- If the function to minimize is $J_{\text{WLS}}(\boldsymbol{\theta})$, the cost is decreased by the iteration

$$\hat{\boldsymbol{\theta}}^{(i+1)} = \hat{\boldsymbol{\theta}}^{(i)} - \gamma \nabla_{\boldsymbol{\theta}} J_{\text{WLS}}(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(i)}},$$





Gradient Descent: Derivation (1/3)

- Scalar cost function

$$J_{\text{LS}}(\boldsymbol{\theta}) = \sum_{n=1}^{N} (y_n - g_n(\boldsymbol{\theta}))^2$$

- Gradient

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} J_{\text{LS}}(\boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}} \sum_{n=1}^{N} (y_n - g_n(\boldsymbol{\theta}))^2 \\ &= \sum_{n=1}^{N} -2\nabla_{\boldsymbol{\theta}} g_n(\boldsymbol{\theta})(y_n - g_n(\boldsymbol{\theta}))\end{aligned}$$

Gradient Descent: Derivation (2/3)

- Gradient

$$\nabla_{\boldsymbol{\theta}} J_{\mathsf{LS}}(\boldsymbol{\theta}) = \sum_{n=1}^{N} -2 \nabla_{\boldsymbol{\theta}} g_n(\boldsymbol{\theta}) (y_n - g_n(\boldsymbol{\theta}))$$

- Vector form:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} J_{\mathsf{LS}}(\boldsymbol{\theta}) &= -2 \begin{bmatrix} \frac{\partial g_1(\boldsymbol{\theta})}{\partial \theta_1} & \frac{\partial g_2(\boldsymbol{\theta})}{\partial \theta_1} & \cdots & \frac{\partial g_N(\boldsymbol{\theta})}{\partial \theta_1} \\ \frac{\partial g_1(\boldsymbol{\theta})}{\partial \theta_2} & \frac{\partial g_2(\boldsymbol{\theta})}{\partial \theta_2} & & \vdots \\ \vdots & & \ddots & \frac{\partial g_N(\boldsymbol{\theta})}{\partial \theta_{K-1}} \\ \frac{\partial g_1(\boldsymbol{\theta})}{\partial \theta_K} & \cdots & \frac{\partial g_{N-1}(\boldsymbol{\theta})}{\partial \theta_K} & \frac{\partial g_N(\boldsymbol{\theta})}{\partial \theta_K} \end{bmatrix} (\boldsymbol{y} - g(\boldsymbol{\theta})) \\ &= \boldsymbol{G}_{\boldsymbol{\theta}}^{\mathsf{T}} (\boldsymbol{y} - g(\boldsymbol{\theta}))\end{aligned}$$

- $\boldsymbol{G}_{\boldsymbol{\theta}}$ is the Jacobian matrix of $g(\boldsymbol{\theta})$

Gradient Descent: Derivation (3/3)

- Generalization to WLS cost function:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} J_{\text{WLS}}(\boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}} (\boldsymbol{y} - g(\boldsymbol{\theta}))^{\mathsf{T}} \boldsymbol{R}^{-1} (\boldsymbol{y} - g(\boldsymbol{\theta})) \\ &= -2 \boldsymbol{G}_{\boldsymbol{\theta}}^{\mathsf{T}} \boldsymbol{R}^{-1} (\boldsymbol{y} - g(\boldsymbol{\theta}))\end{aligned}$$

- The direction of the negative gradient is

$$-\nabla_{\boldsymbol{\theta}} J_{\text{WLS}}(\boldsymbol{\theta}) = -\boldsymbol{G}_{\boldsymbol{\theta}}^{\mathsf{T}} \boldsymbol{R}^{-1} (\boldsymbol{y} - g(\boldsymbol{\theta}))$$

- The parameter update becomes:

$$\begin{aligned}\hat{\boldsymbol{\theta}}^{(i+1)} &= \hat{\boldsymbol{\theta}}^{(i)} + \gamma \Delta \boldsymbol{\theta}^{(i+1)} \\ \Delta \boldsymbol{\theta}^{(i+1)} &= \boldsymbol{G}_{\boldsymbol{\theta}}^{\mathsf{T}} \boldsymbol{R}^{-1} (\boldsymbol{y} - g(\hat{\boldsymbol{\theta}}^{(i)}))\end{aligned}$$

- The Jacobian matrix $\boldsymbol{G}_{\boldsymbol{\theta}}$ is evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(i)}$

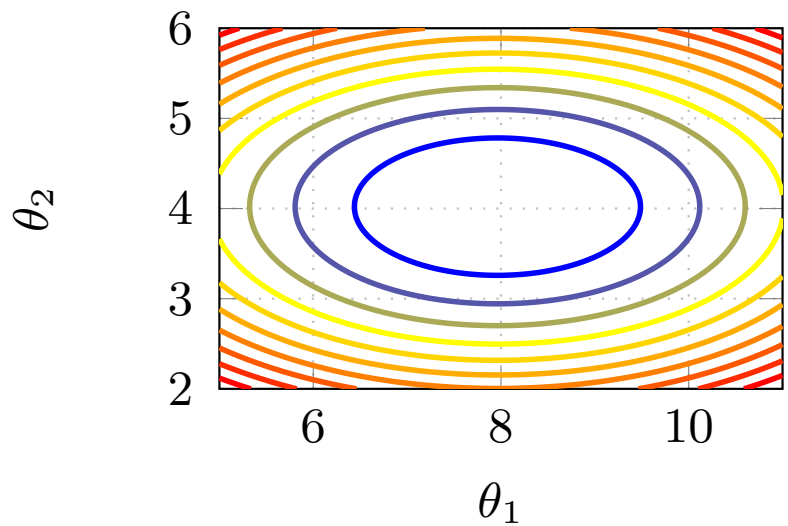
Gradient Descent: Line Search (1/2)

- Parameter update:

$$\hat{\boldsymbol{\theta}}^{(i+1)} = \hat{\boldsymbol{\theta}}^{(i)} + \gamma \Delta \boldsymbol{\theta}^{(i+1)}$$

$$\Delta \boldsymbol{\theta}^{(i+1)} = \boldsymbol{G}_{\boldsymbol{\theta}}^{\mathsf{T}} \boldsymbol{R}^{-1} (\boldsymbol{y} - g(\hat{\boldsymbol{\theta}}^{(i)}))$$

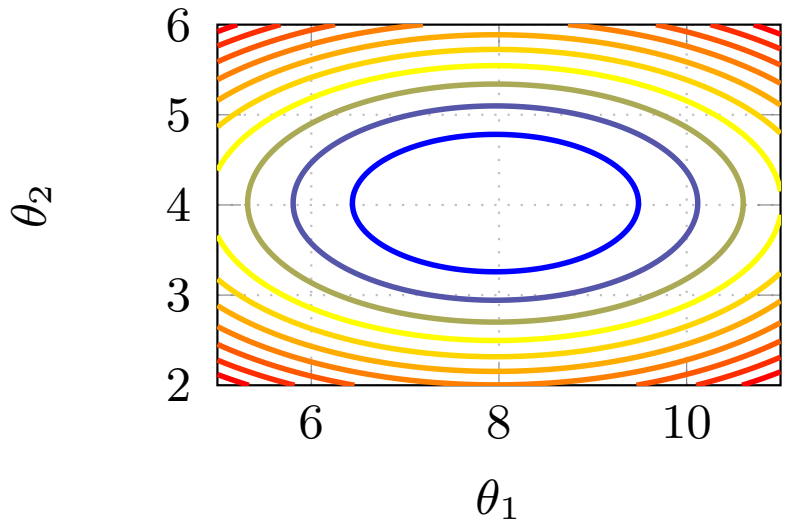
- How long should γ be?
  - Large γ: Might overshoot
  - Small γ: Too slow progress



Gradient Descent: Line Search (2/2)

Adaptive step length selection strategy:

1. Choose $\gamma = 1$
2. Calculate $\hat{\boldsymbol{\theta}}^{(i+1)}$ and the corresponding cost
3. If the cost is lowered, accept $\hat{\boldsymbol{\theta}}^{(i+1)}$, otherwise set $\gamma \leftarrow \gamma/2$ and return to step 2



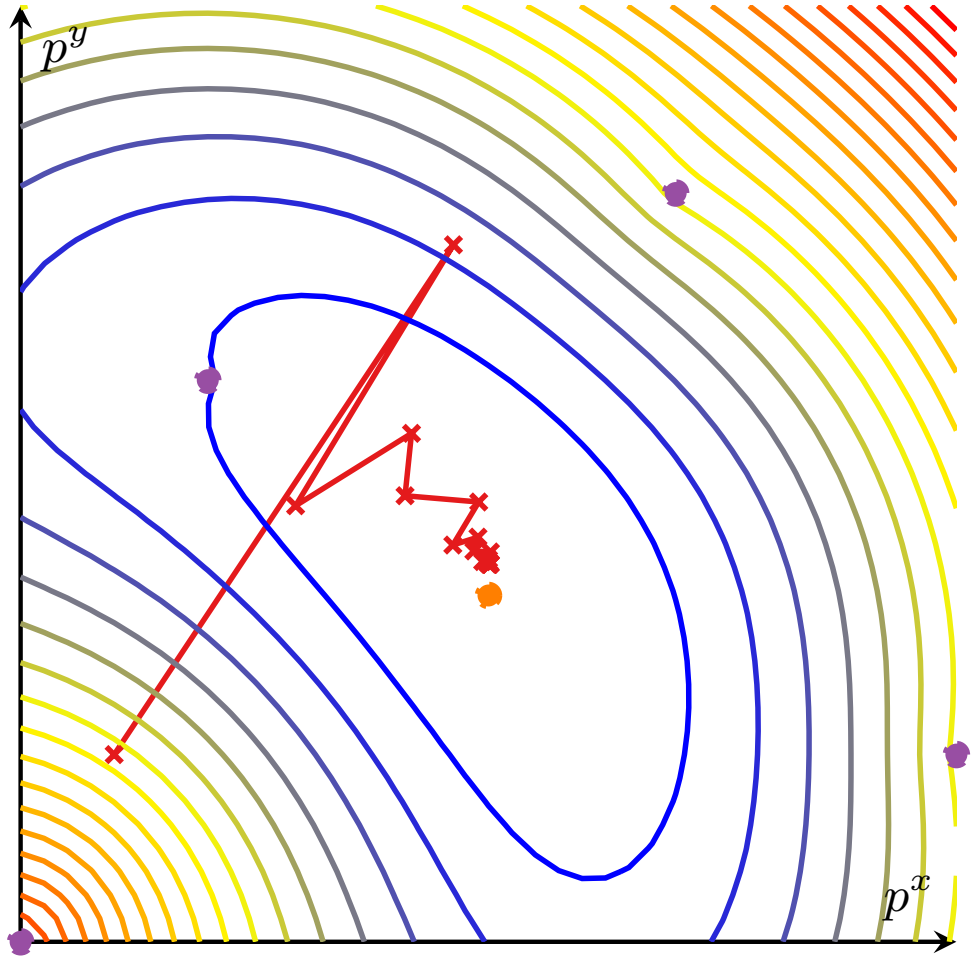
Gradient Descent: Algorithm

Algorithm 1 Gradient Descent with Line Search

Require: Initial guess $\hat{\boldsymbol{\theta}}^{(0)}$, data $\boldsymbol{y}$, function $g(\boldsymbol{\theta})$, Jacobian $\boldsymbol{G}_{\boldsymbol{\theta}}$
Ensure: Parameter estimate $\hat{\boldsymbol{\theta}}_{\text{WLS}}$

Set $i \leftarrow 0$
repeat
  Calculate $\Delta\boldsymbol{\theta}^{(i+1)} = \boldsymbol{G}_{\boldsymbol{\theta}}^{\mathsf{T}} \boldsymbol{R}^{-1} (\boldsymbol{y} - g(\hat{\boldsymbol{\theta}}^{(i)}))$
  Set $\gamma^{(i+1)} \leftarrow 1$
  repeat ▷ Line Search
    Calculate $\hat{\boldsymbol{\theta}}^{(i+1)} = \hat{\boldsymbol{\theta}}^{(i)} + \gamma^{(i+1)} \Delta\boldsymbol{\theta}^{(i+1)}$
    Set $\gamma^{(i+1)} \leftarrow \gamma^{(i+1)}/2$
  until $J_{\text{WLS}}(\hat{\boldsymbol{\theta}}^{(i+1)}) < J_{\text{WLS}}(\hat{\boldsymbol{\theta}}^{(i)})$
  Set $i \leftarrow i + 1$
until Converged

Example: Localizing a Target (4)



Summary

- The sequential linear least squares estimator is

$$\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_{n-1} + \boldsymbol{L}_n(\boldsymbol{y}_n - \boldsymbol{C}_n\hat{\boldsymbol{\theta}}_{n-1})$$
$$\boldsymbol{L}_n = \boldsymbol{P}_{n-1}\boldsymbol{C}_n^{\mathsf{T}}(\boldsymbol{C}_n\boldsymbol{P}_{n-1}\boldsymbol{C}_n^{\mathsf{T}} + \boldsymbol{R}_n)^{-1}$$

  with $\mathrm{E}\{\hat{\boldsymbol{\theta}}_n\} = \boldsymbol{\theta}$ and
  $\boldsymbol{P}_n = \mathrm{Cov}\{\hat{\boldsymbol{\theta}}_n\} = \boldsymbol{P}_{n-1} - \boldsymbol{L}_n(\boldsymbol{C}_n\boldsymbol{P}_{n-1}\boldsymbol{C}^{\mathsf{T}} + \boldsymbol{R}_n)\boldsymbol{L}_n^{\mathsf{T}}$
- The regularized linear least squares estimator is

$$\hat{\boldsymbol{\theta}}_{\text{ReLS}} = \boldsymbol{m} + \boldsymbol{K}(\boldsymbol{y} - \boldsymbol{G}\boldsymbol{m})$$
$$\boldsymbol{K} = \boldsymbol{P}\boldsymbol{G}^{\mathsf{T}}(\boldsymbol{G}\boldsymbol{P}\boldsymbol{G}^{\mathsf{T}} + \boldsymbol{R})^{-1}$$

  with $\mathrm{E}\{\hat{\boldsymbol{\theta}}_{\text{ReLS}}\} = \boldsymbol{m} + \boldsymbol{K}(\boldsymbol{G}\,\mathrm{E}\{\boldsymbol{\theta}\} - \boldsymbol{G}\boldsymbol{m})$ and
  $\mathrm{Cov}\{\hat{\boldsymbol{\theta}}_{\text{ReLS}}\} = \boldsymbol{P} - \boldsymbol{K}(\boldsymbol{G}\boldsymbol{P}\boldsymbol{G}^{\mathsf{T}} + \boldsymbol{R})\boldsymbol{K}^{\mathsf{T}}$
- The gradient descent algorithm is a numerical method for solving nonlinear least squares problems