

# Introduction to optimization

---

Julia Wrobel

# Overview

Today, we cover:

- Intro to optimization
- Rates of convergence
- Beginning of gradient methods
  - Steepest descent
  - Newton's Method

Announcements

- HW2 posted and due 2/11 at 10:00AM

Readings:

- Peng Chapter 2 (rates of convergence)
- Peng Chapter 3 (general optimization)

# Optimization terminology

We will consider the following general optimization problem:

$$\begin{aligned} & \text{minimize}_x \quad f(\mathbf{x}) \\ & \text{subject to} \quad g_j(\mathbf{x}) \leq 0, \quad j = 1, 2, \dots, m; \\ & \quad h_k(\mathbf{x}) = 0, \quad k = 1, 2, \dots, l. \end{aligned}$$

- $\mathbf{x} \in \mathbf{R}^p$ : **optimization variable** (in this class, could be a scalar, vector or a matrix)
- $f(\mathbf{x}) : \mathbf{R}^p \rightarrow \mathbf{R}$ : **objective function**
- $g_j : R^p \rightarrow R$  and  $g_j(\mathbf{x}) \leq 0$ : **inequality constraints**
- $h_k : R^p \rightarrow R$  and  $h_k(\mathbf{x}) = 0$ : **equality constraints**
- If no constraints: **unconstrained problem**

# Optimization terminology

We will consider the following general optimization problem:

$$\text{minimize}_x \quad f(\mathbf{x})$$

$$\text{subject to} \quad g_j(\mathbf{x}) \leq 0, \quad j = 1, 2, \dots, m;$$

$$h_k(\mathbf{x}) = 0, \quad k = 1, 2, \dots, l.$$

# Least squares linear regression

Given data  $X \in R^{n \times p}$  and  $Y \in R^n$  with  $\text{rank}(X) = p$

$$\underset{\beta}{\text{minimize}} \|Y - X\beta\|_2^2$$

- **unconstrained** optimization problem
- any  $\beta \in R^p$  is **feasible**
- the optimal value  $f^* = \|Y - X(X^T X)^{-1} X^T Y\|_2^2$
- the globally optimal  $\beta^* = (X^T X)^{-1} X^T Y$ 
  - also locally optimal, the only locally optimal point

# Unconstrained optimization problem

Consider minimizing differentiable function  $f$

$$\text{minimize}_x f(x)$$

A point  $x^*$  is called **stationary** if

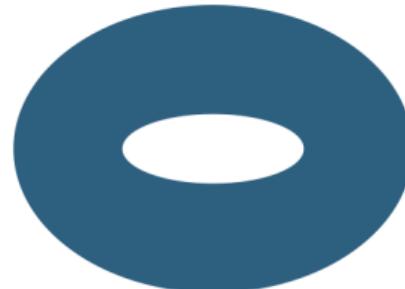
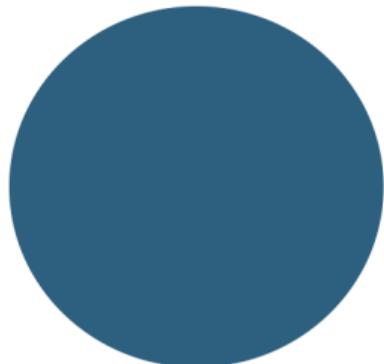
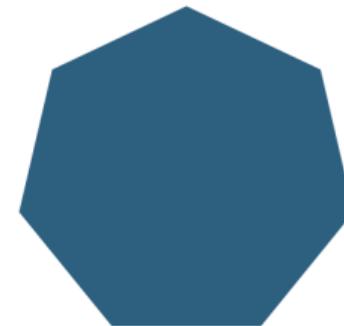
$$\nabla f(x^*) = 0.$$

All local optimal points are **stationary** points.

Globally optimal  $x^*$  satisfies  $f(x^*) = 0$ , but locally optimal and stationary points also satisfy it.

For **convex**  $f$ , any solution to  $f(x^*) = 0$  is globally optimal.

# Convex sets



# Convex optimization problems

- Very common in statistics, *easier* to solve, generally have nice algorithms

**Definition:** A function  $f : R^p \rightarrow R$  is **convex** if for all  $x_1, x_2 \in R^p$  and all  $\alpha \in [0, 1]$ ,

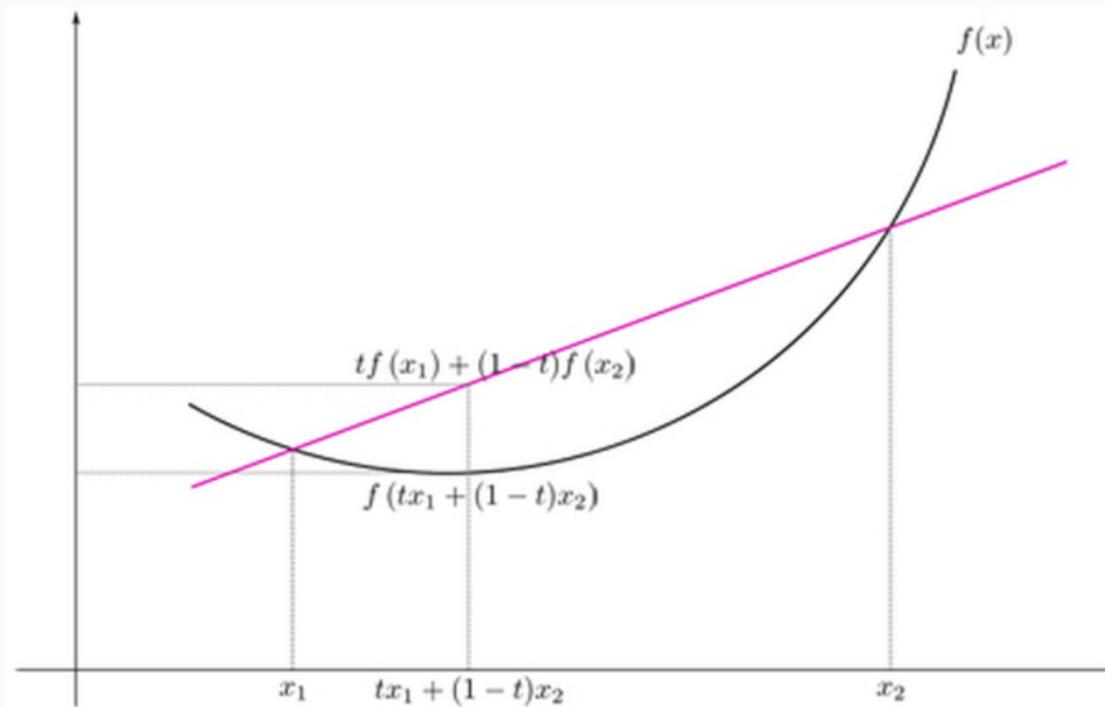
$$f(\alpha x_1 + (1 - \alpha)x_2) \leq \alpha f(x_1) + (1 - \alpha)f(x_2)$$

and is **strictly convex** if for all  $x_1, x_2 \in R^p$ ,  $x_1 \neq x_2$ , and all  $\alpha \in (0, 1)$

$$f(\alpha x_1 + (1 - \alpha)x_2) < \alpha f(x_1) + (1 - \alpha)f(x_2)$$

- **Interpretation:** The chord between two points is always above the function

# Convex optimization problems



**Figure 1:** Convex function, basic definition

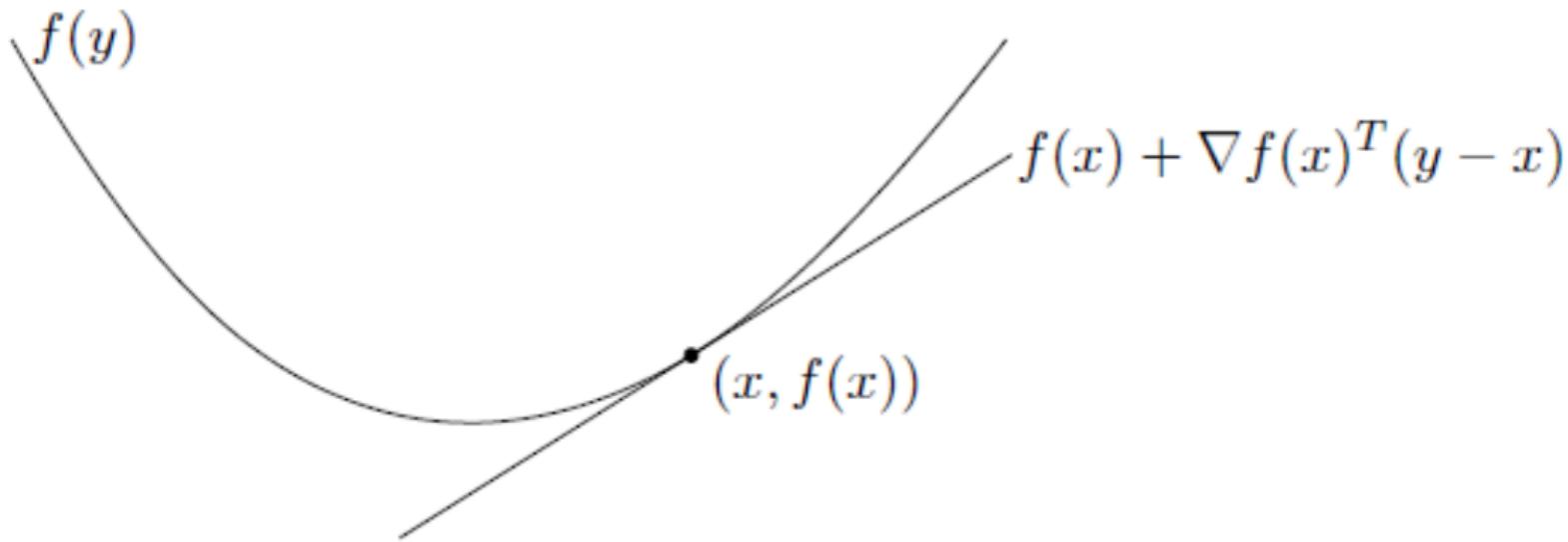
# Convex optimization problems

## Theorem

*First order conditions (for differentiable  $f$ )*

- $f$  is convex  $\iff f(y) \geq f(x) + \nabla f(x)^\top (y - x), \forall x, y \in \mathbb{R}^p.$
- $f$  is strictly convex  $\iff f(y) > f(x) + \nabla f(x)^\top (y - x), \forall x, y \in \mathbb{R}^p$  and  $x \neq y.$
- **Interpretation:** function lies above its tangent

# Convex optimization problems



**Figure 2:** Convex function, first order condition

# Convex optimization problems

## Theorem

*Second order conditions (for twice differentiable  $f$ )*

- $f$  is convex  $\iff$  Hessian  $\nabla^2 f(x) \succeq 0, \forall x \in \mathbb{R}^p$ . (pos.semi-def)
- $f$  is strictly convex  $\iff$  Hessian  $\nabla^2 f(x) \succ 0, \forall x \in \mathbb{R}^p$  (strictly pos.semi-def)
- Often easiest to check in practice, i.e.

$$f(x) = x^2, \quad \nabla^2 f(x) = 2 > 0.$$

$$f(x) = \|x\|_2^2, \quad \nabla f(x) = 2x, \quad \nabla^2 f(x) = 2I \succ 0.$$

# Examples of convex functions

- $-\log(x)$
- $e^x$
- $|x|^p, p \geq 1$
- Any norm on  $\mathbb{R}^p$
- $-\log(\det(\Sigma))$ , where  $\Sigma$  is positive definite

# Operations that preserve convexity

- Non-negative weighted sum:  $\sum_{i=1}^k w_i f_i$ , where  $w_i \geq 0$  and  $f_i, i = 1, \dots, k$  are convex functions.
- If  $f$  is convex, and  $g(x) = f(Ax + b)$ , then  $g$  is convex.
- If  $f_1, \dots, f_k$  are convex functions, then  $\max(f_1, \dots, f_k)$  is also convex.
- ... not exhaustive list

# Example

Least squares loss function is convex

$$f(\beta) = \|Y - X\beta\|_2^2$$

Why?

- The hessian is  $\nabla^2 f(\beta) = 2X^T X \succeq 0$  (semi positive definite)
- $f(\beta) = g(Y - X\beta)$ , where  $g(x) = \|x\|_2^2$  is convex as a norm squared

- A point  $x \in \mathbb{R}^p$  is **feasible** if it satisfies all the constraints.  
Otherwise, it's **infeasible**.
- The **optimal value**  $f^*$  is the minimal value of  $f$  over the set  
of feasible points

# Optimization terminology

We will consider the following general optimization problem:

$$\begin{aligned} & \text{minimize}_x \quad f(\mathbf{x}) \\ & \text{subject to} \quad g_j(\mathbf{x}) \leq 0, \quad j = 1, 2, \dots, m; \\ & \quad h_k(\mathbf{x}) = 0, \quad k = 1, 2, \dots, l. \end{aligned}$$

- The **optimal value**  $f^*$  is the minimal value of  $f$  over the set of feasible points
- $x^*$  is **globally optimal** if  $x$  is *feasible* and  $f(x^*) = f^*$
- $x^*$  is **locally optimal** if  $x$  is *feasible* and for each feasible  $x$  in the neighborhood  $\|x - x^*\|_2 \leq R$  for some  $R > 0$ ,  $f(x^*) \leq f(x)$ .

# Recall unconstrained optimization problem

Consider minimizing differentiable function  $f$

$$\text{minimize}_x f(x)$$

A point  $x^*$  is called **stationary** if

$$\nabla f(x^*) = 0.$$

All local optimal points are **stationary** points.

Globally optimal  $x^*$  satisfies  $f(x^*) = 0$ , but locally optimal and stationary points also satisfy it.

# Unconstrained convex optimization problem

Consider

$$\text{minimize}_x f(x)$$

This is a **convex** optimization problem if  $f(x)$  is **convex**

**Important property 1:** any locally optimal point of a convex problem is globally optimal

**Important property 2:** If  $f$  is differentiable,  $x^*$  is optimal if and only if

$$\nabla f(x)|_{x=x^*} = 0.$$

## Example: Least Squares

Least squares solves

$$\underset{\beta}{\text{minimize}} \|Y - X\beta\|_2^2$$

This is a convex unconstrained optimization problem, so the solution must satisfy

$$\begin{aligned}\nabla \|Y - X\beta\|_2^2 &= \nabla (\|Y\|_2^2 - 2Y^T X\beta + \beta^T X^T X\beta) \\ &= -2X^T Y + 2X^T X\beta = 0\end{aligned}$$

## Example: Least Squares

This is equivalent to

$$X^T X \beta = X^T Y$$

If  $X^T X$  is **invertible**, global solution is  $\beta^* = (X^T X)^{-1} X^T Y$ .

■

# Example: Maximum Likelihood Estimation

Observations  $x_i, i = 1, \dots, n$ , independent samples from distribution with density  $f(x; \theta)$  with some parameter  $\theta \in \mathbb{R}^d$

## Maximum Likelihood Estimator (MLE)

$$\hat{\theta} = \arg \max_{\theta} \prod_{i=1}^n f(x_i; \theta)$$

Typically, we maximize **log-likelihood** which is equivalent to

$$\hat{\theta} = \arg \min_{\theta} \left\{ - \sum_{i=1}^n \log f(x_i; \theta) \right\}$$

This is **convex optimization** if  $-\log(f)$  is convex.

# MLE example

Normal likelihood with known variance  $\sigma^2$

$$f(x; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \theta)^2}{2\sigma^2}\right)$$

Here  $\theta$  is the unknown mean.

Log-likelihood

$$\log f(x; \theta) = -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x - \theta)^2 = C - \frac{1}{2\sigma^2}(x - \theta)^2$$

**MLE** optimization problem

$$\hat{\theta} = \arg \min_{\theta} \left\{ -\sum_{i=1}^n -\frac{1}{2\sigma^2}(x_i - \theta)^2 \right\} = \arg \min_{\theta} \sum_{i=1}^n (x_i - \theta)^2$$

# MLE example

MLE optimization problem

$$\hat{\theta} = \arg \min_{\theta} \left\{ -\sum_{i=1}^n -\frac{1}{2\sigma^2} (x_i - \theta)^2 \right\} = \arg \min_{\theta} \sum_{i=1}^n (x_i - \theta)^2$$

This is **convex optimization problem**. Why?

The optimality conditions

$$-2 \sum_{i=1}^n x_i + 2n\theta = 0.$$

The optimal  $\hat{\theta} = n^{-1} \sum_{i=1}^n x_i = \bar{x}$  - sample mean.

# Summary

Unconstrained optimization problem with differentiable  $f$ :

$$\text{minimize}_x f(x).$$

To find global optimum, need to solve optimality conditions

$$\nabla f(x) = 0.$$

For **convex**  $f$  - any solution to above is **globally optimal**.

Least squares problem has closed form solution.

What if exact solution is not tractable? Need **numerical methods**

## Example 1

$$f(x) = (x - 50)^2 + e^x / 50$$

We will come back to this example shortly.

# Rates of convergence

One of the ways algorithms can be compared is via their rates of convergence to some limiting value.

- Typically we have an iterative algorithm that is trying to find the max/min of an objective function  $f$ 
  - Want to estimate how long it will take to reach that optimal value
- Three rates of convergence we will focus on:
  - **linear** (slowest)
  - **superlinear** (faster)
  - **quadratic** (fastest)

Algorithms that require more information about  $f$  (such as its derivative) tend to converge more quickly.

# Linear convergence

Suppose we have a sequence  $\{x_n\}$  such that  $x_n \rightarrow x_\infty \in \mathcal{R}^k$ .  
Convergence is **linear** if there exists  $r \in (0, 1)$  such that:

$$\frac{\|x_{n+1} - x_\infty\|}{\|x_n - x_\infty\|} \leq r$$

for all sufficiently large  $n$ .

# Superlinear convergence

We say a sequence  $\{x_n\}$  converges to  $x_\infty$  **superlinearly** if we have

$$\lim_{n \rightarrow \infty} \frac{\|x_{n+1} - x_\infty\|}{\|x_n - x_\infty\|} = 0$$

for all sufficiently large  $n$ .

# Quadratic convergence

Quadratic convergence is the fastest form of convergence discussed here. We say a sequence  $\{x_n\}$  converges to  $x_\infty$  at a **quadratic** rate if there exists some constant  $0 < M < \infty$  such that

$$\frac{\|x_{n+1} - x_\infty\|}{\|x_n - x_\infty\|^2} \leq M$$

for all sufficiently large  $n$ .

# Gradient methods: steepest (gradient) descent

- Choose a step size  $\alpha > 0$ 
  - Sometimes called **learning rate** or **learning step**
- Start with an initial guess  $x_0$
- At each iteration  $t$ , compute  $x_{t+1} = x_t - \alpha \nabla f(x_t)$
- Continue until some **convergence criterion** is met
  - i.e.  $f(x_{t+1}) \approx f(x_t)$

**Idea:** Move  $\alpha$  units in the direction of *steepest descent*, which is the direction that is orthogonal to the contours of  $f$  at the point  $x_n$

- This attempts to find solution to  $\nabla f(x) = 0$

# Steepest descent

In practice, can require many steps (iterations) to reach the minimum when parameters are highly correlated.

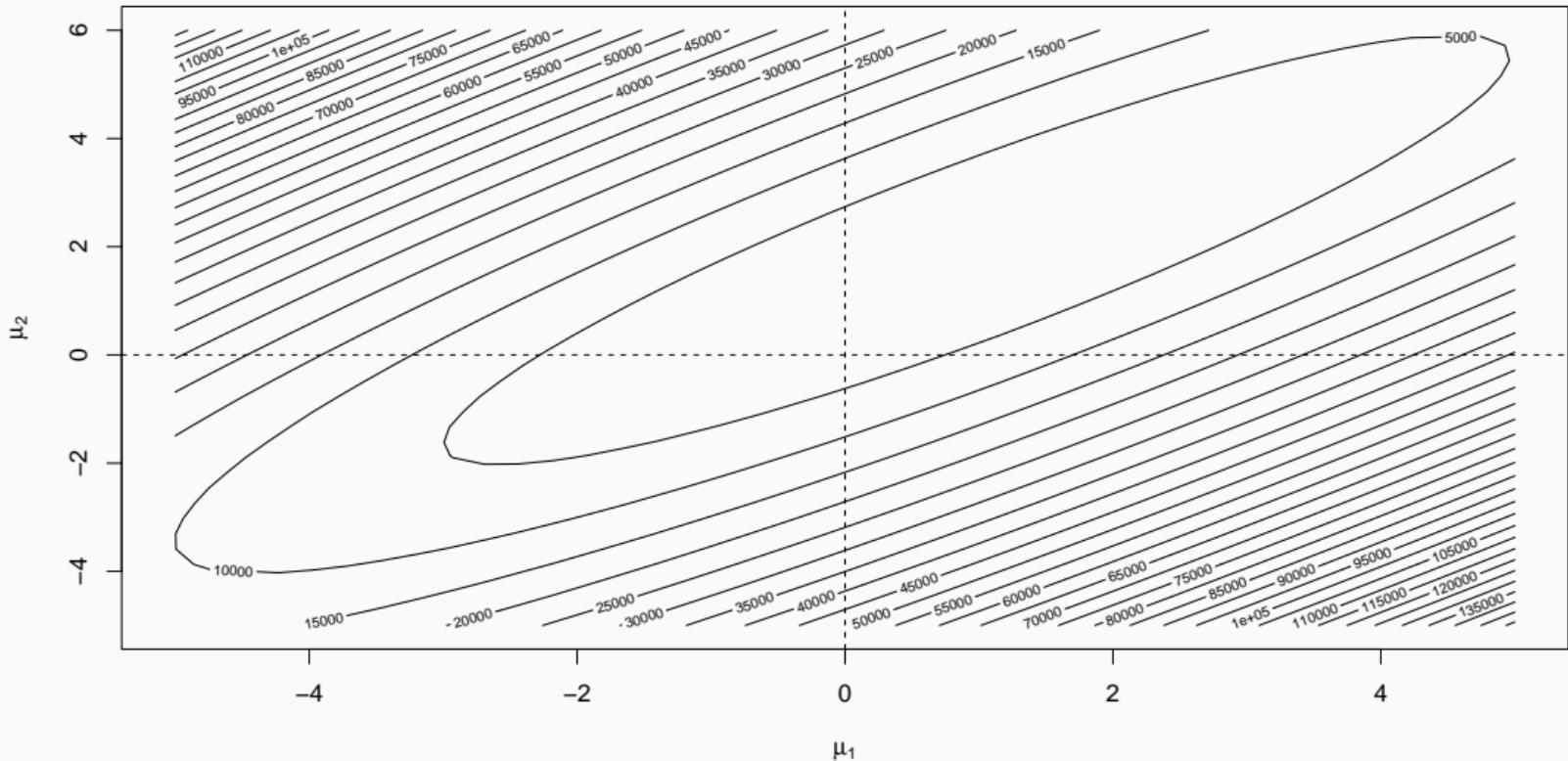
**Example:** Bivariate Normal.

- Can use steepest descent to estimate the MLE of the mean
- True  $\mu = (1, 2)$
- True  $\Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$

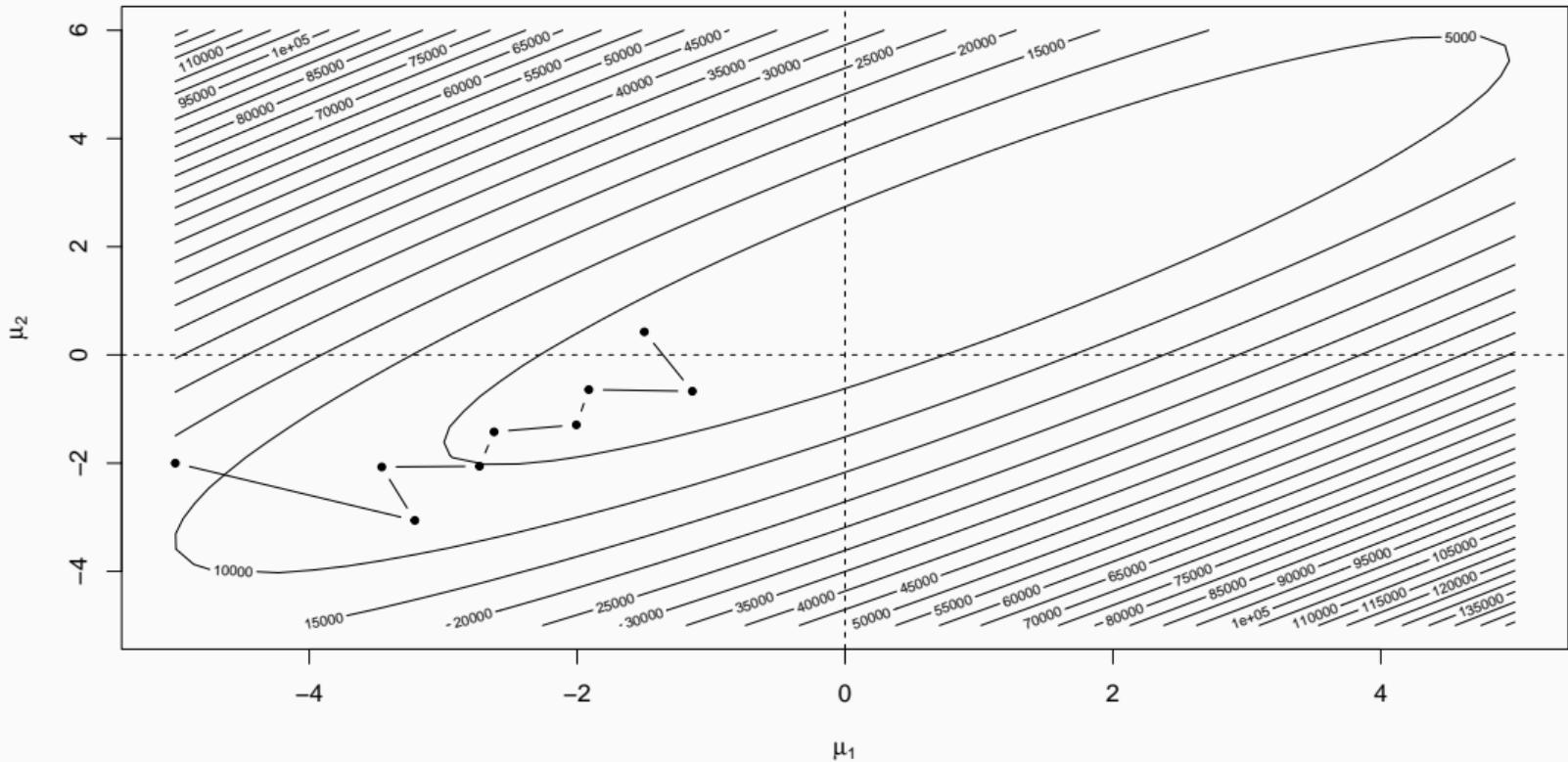
Parameters are highly correlated!

- Try starting value  $\mu_0 = (-5, -2)$

# Steepest descent



# Steepest descent



# Steepest descent - example

- For simplicity, focus on one-dimensional case first

$$f(x) = (x - 50)^2 + e^x / 50, \quad \nabla f(x) = 2x - 100 + e^x / 50 = 0$$

The choice of step size is very important!!

- Too small**  $\alpha$  - very small difference between updates, larger number of iterations
- Too large**  $\alpha$  - oscillations, may not converge

# Lab exercise

Use **Exercise 1** in lab to check different values of  $\alpha$  on the given function.

How did we monitor convergence in this code? Why?

# Steepest descent in practice

- Very simple
- Only requires the first derivative
- Used in many machine learning methods, i.e. in neural nets  
(with additional stochastic updates)

# Newton's method

Goal is to find solution  $x^*$  to

$$\nabla f(x) = 0$$

By Taylor expansion, can approximate  $\nabla f(x^*)$  around a given point  $x$ :

$$\nabla f(x^*) = \nabla f(x) + \nabla^2 f(x)(x^* - x) + \text{higher order terms.}$$

# Newton's method

Since  $\nabla f(x^*) = 0$ , must have  $\nabla f(x) + \nabla^2 f(x)(x^* - x) \approx 0$ , leading to

$$x^* \approx x - \{\nabla^2 f(x)\}^{-1} \nabla f(x).$$

One dimensional case update, Newton's method

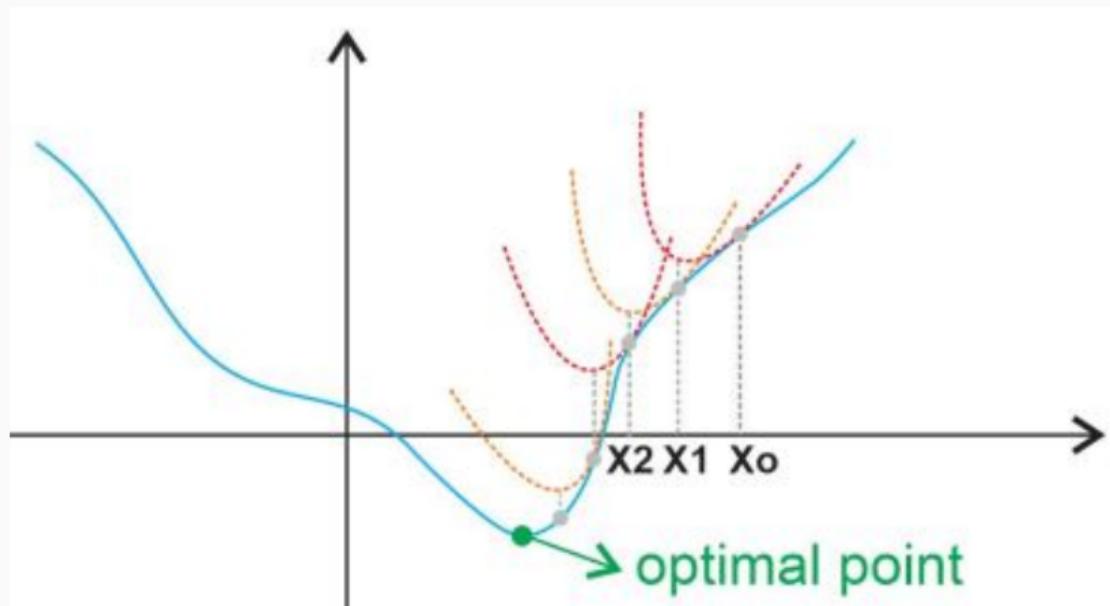
$$x_{t+1} = x_t - \frac{f'(x_t)}{f''(x_t)}$$

Steepest descent (one dimensional)

$$x_{t+1} = x_t - \alpha f'(x_t)$$

# Newton's method - illustration

- The closer is  $x_0$  to the optimal value  $x^*$ , the faster is the convergence



## Newton's method - example

$$f(x) = (x - 50)^2 + e^x/50, \quad \nabla f(x) = 2x - 100 + e^x/50 = 0$$

$$\nabla f'(x) = 2 + e^x/50$$

See **Exercise 2** in lab to implement this example.

- How does it compare to the steepest descent approach in number of iterations?
- Computation time?

# Recall

Convex optimization problem:

$$\text{minimize}_x f(x), \quad f - \text{convex function}.$$

To find global optimum, need to solve optimality conditions

$$\nabla f(x) = 0.$$

Steepest descent algorithm and Newton's method aim to find any solution to the above, so may be applied with nonconvex problems as well **but** - only guaranteed to converge to a **global** optimum if convex - solution may be a local min, local max, or saddle point

# Resources

- old paper on convexity in GLMs
- Peng, Advanced Statistical Computing, chapters 2 and 3