# Scaled Descent

CPSC 406 – Computational Optimization

# Scaled descent

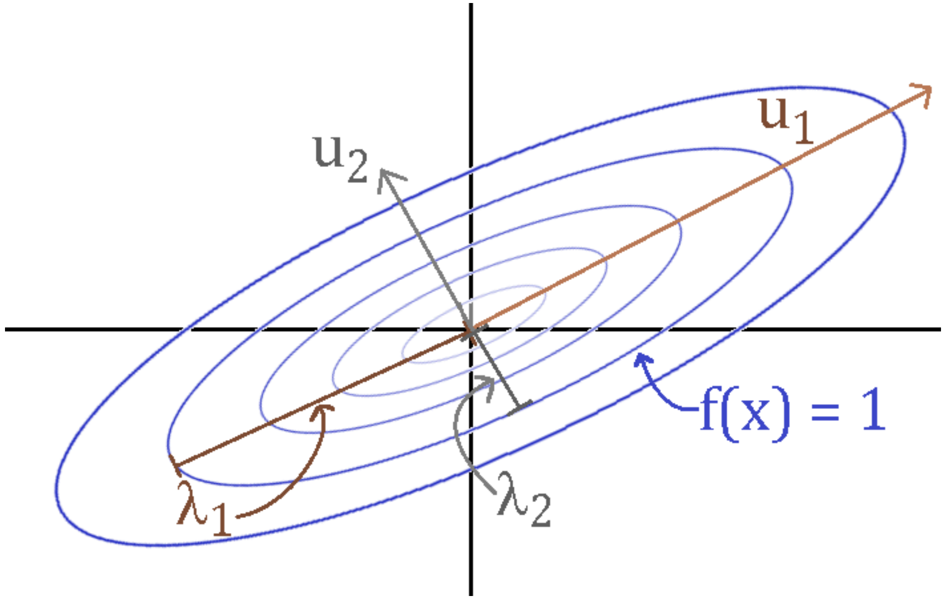
- conditioning
- scaled gradient direction
- Gauss Newton

# Zig-zagging

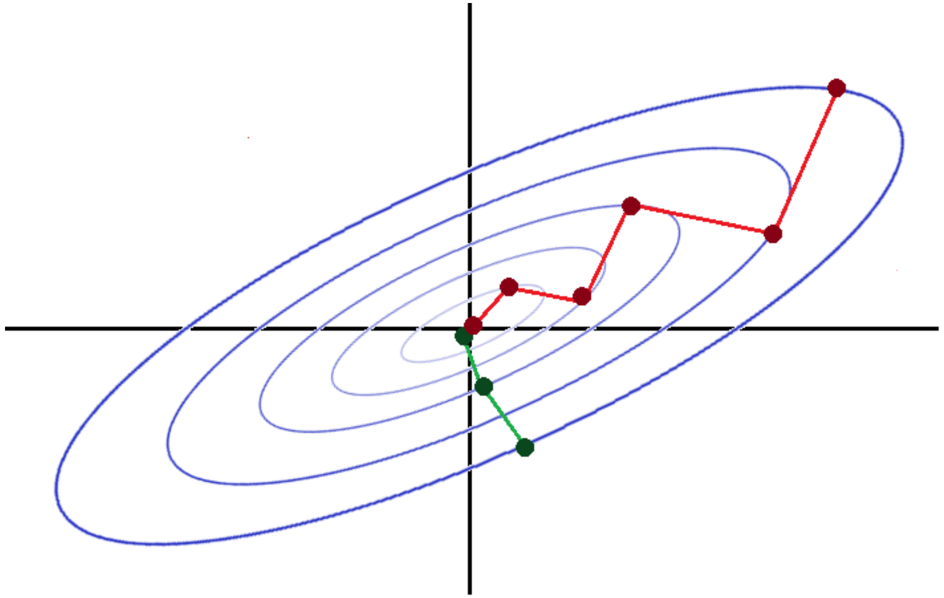
Consider the quadratic function with $H$ symmetric and positive definite

$$f(x) = \frac{1}{2} x^\top H x, \qquad H = U \Lambda U^\top$$

**level sets** are ellipsoids:

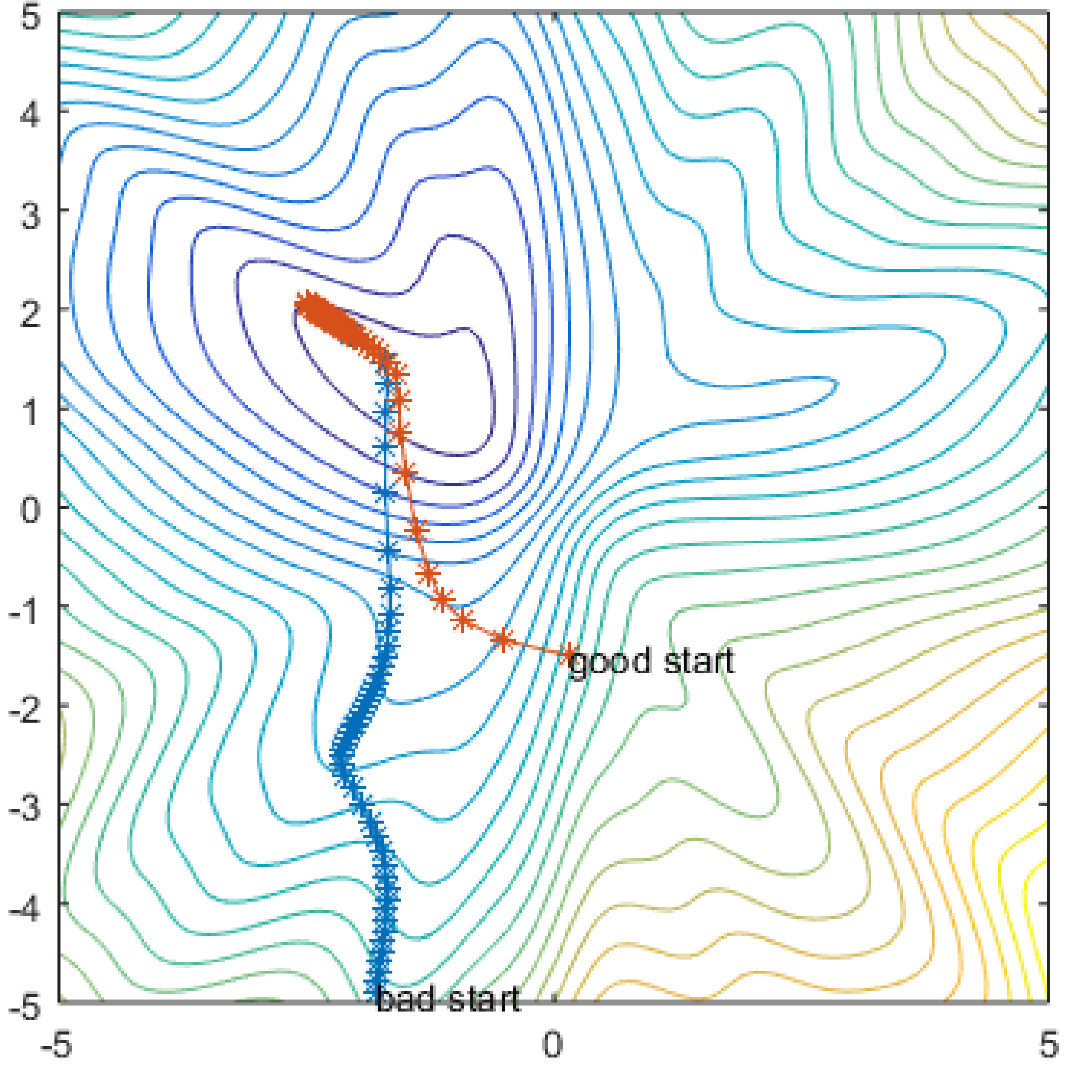



gradient descent from two starting points:



- eigenvectors of $H$ are principal axes
- eigenvalues are the lengths of the “unit ellipse” axes

# Gradient descent



# Gradient descent zig-zags

Let $x^1, x^2, \ldots$ be the iterates generated by gradient descent with **exact** linesearch. Then

$$(x^{k+1} - x^k)^T (x^{k+2} - x^{k+1}) = 0$$

Proof: exact steplength satisfies

$$\alpha^k = \operatorname*{argmin}_{\alpha > 0} \phi(\alpha) := f(x^k + \alpha d^k), \quad d^k = -\nabla f(x^k)$$

- optimality of step $\alpha = \alpha^k$

$$0 = \phi'(\alpha^k) = \frac{d}{d\alpha} f(\underbrace{x^k + \alpha^k d^k}_{= x^{k+1}}) = (d^k)^T \nabla f(x^{k+1}) = -\nabla f(x^k)^T \nabla f(x^{k+1})$$

- because $x^{k+1} - x^k = \alpha^k d^k$ and $x^{k+2} - x^{k+1} = \alpha^{k+1} d^{k+1}$

$$\nabla f(x^k)^T \nabla f(x^{k+1}) = 0 \quad \iff \quad (x^{k+1} - x^k)^T (x^{k+2} - x^{k+1}) = 0$$

# Condition number

The **condition number** of an $n \times n$ positive definite matrix $H$ is

$$\kappa(H) = \frac{\lambda_{\max}(H)}{\lambda_{\min}(H)} \geq 1$$

- ill-conditioned if $\kappa(H) \gg 1$
- condition number of Hessian influences speed of convergence of gradient descent
  - $\kappa(H) = 1$: gradient descent converges in one step
  - $\kappa(H) \gg 1$: gradient descent zig-zags
- if $f$ is twice continuously differentiable, define the **condition number** of $f$ at solution $x^*$ as

$$\kappa(f) = \kappa(\nabla^2 f(x^*))$$

# Scaled gradient method

$$\min_x f(x) \qquad f : \mathbb{R}^n \to \mathbb{R}$$

- make a linear change of variables: $x = Sy$ where $S$ is nonsingular to get rescaled problem

$$\min_y \; g(y) := f(Sy)$$

- apply gradient descent to scaled problem

$$y^{k+1} = y^k - \alpha^k \nabla g(y^k) \quad \text{with} \quad \nabla g(y) = S^\intercal \nabla f(Sy)$$

- multiply on left by $S$ to get $x$-update

$$x^{k+1} = Sy^{k+1} = S(y^k - \alpha^k \nabla g(y^k)) = x^k - \alpha^k SS^T \nabla f(x^k)$$

**scaled gradient** method

$$x^{k+1} = x^k + \alpha^k d^k, \qquad d^k = -\underbrace{SS^T}_{\succ 0} \nabla f(x^k)$$

# Scaled descent

- If $\nabla f(x) \neq 0$, the scaled negative gradient $d = -SS^T \nabla f(x)$ is a descent direction

$$f'(x; d) = d^T \nabla f(x) = -\nabla f(x)^T (SS^T) \nabla f(x) < 0$$

  because $D := SS^T \succ 0$

- Recall: a matrix $D$ is **positive definite** if and only if
  - $D = U \Lambda U^\intercal$ with $\Lambda \succ 0$ diagonal and $U$ nonsingular
  - $D = SS^\intercal$ with $S$ nonsingular

**scaled gradient method**

- for $k = 0, 1, 2, \ldots$
  - choose scaling matrix $D_k \succ 0$
  - compute $d^k = -D \nabla f(x^k)$
  - choose stepsize $\alpha^k > 0$ via linesearch on $\phi(\alpha) = f(x^k + \alpha d^k)$
  - update $x^{k+1} = x^k + \alpha^k d^k$

# Choosing the scaling matrix

Observe relationship between optimizing $f$ and optimizing its scaling $g$

$$\min_y g(y) = f(Sy) \quad \text{with} \quad x \equiv Sy$$

**condition number** of $\nabla^2 f(x)$ governs convergence of gradient descent

$$\nabla^2 g(y) = S^\intercal \nabla^2 f(Sy) S$$

- choose $S$ such that $\nabla^2 g$ is well-conditioned, ie, $\kappa(\nabla^2 g) \approx 1$

**Example (quadratic)**

$$f(x) = \tfrac{1}{2} x^T H x + b^\intercal x + \gamma, \quad \nabla^2 f(x) = H = U \Lambda U^T \succ 0$$

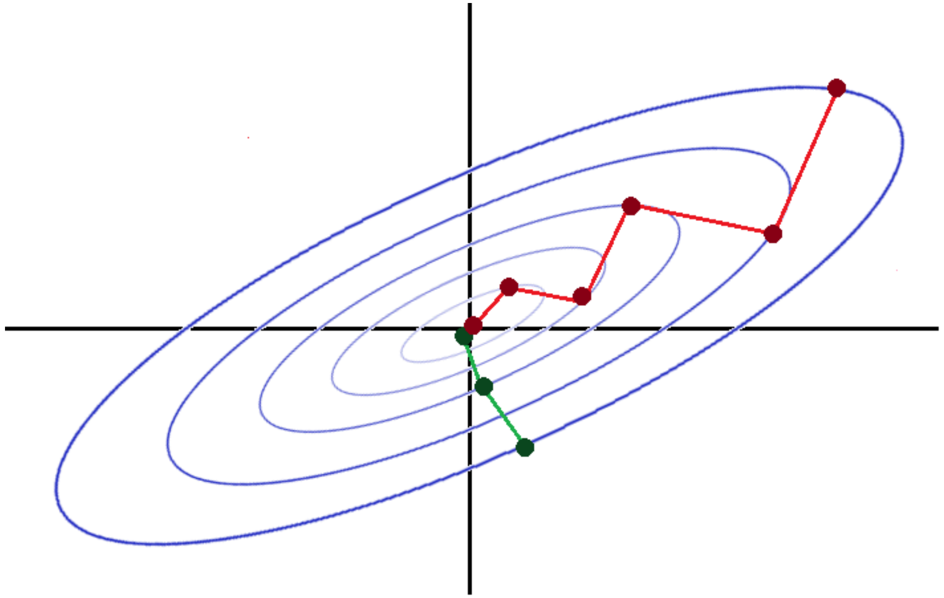
- pick $S$ such that $S^T H S = I$, ie, $S = H^{-1/2} := U \Lambda^{-1/2} U^T$
- gives perfectly conditioned $\nabla^2 g$

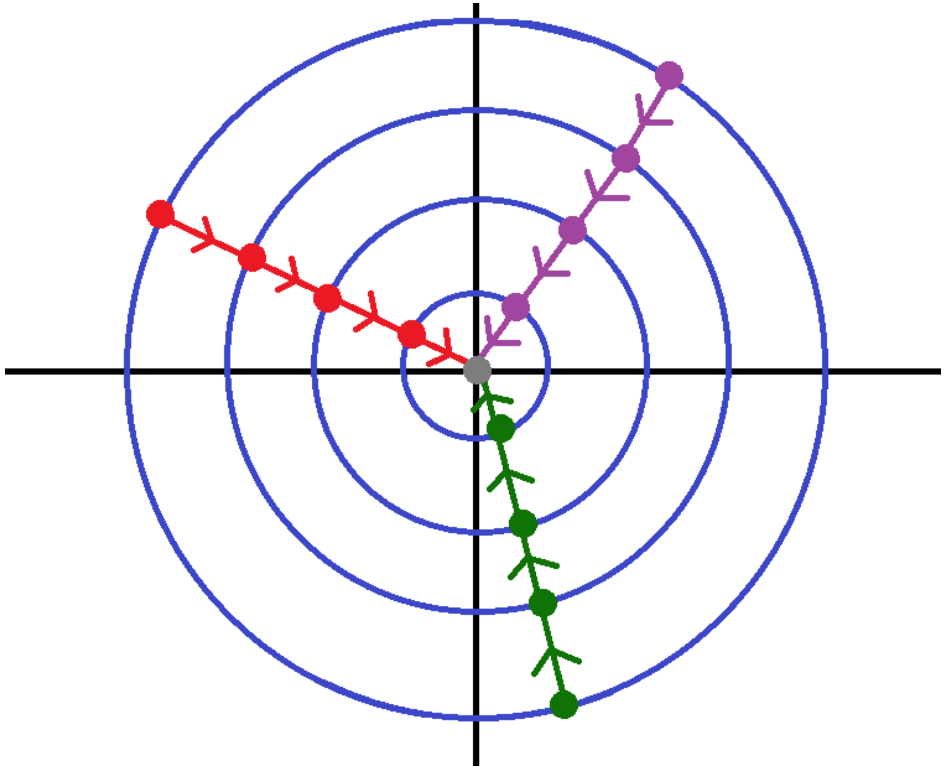
$$\kappa(S^T H S) = \kappa(H^{-1/2} H H^{-1/2}) = \kappa(I) = 1$$

# Level sets of scaled and unscaled problems

Close to solution $x^*$, levels sets of

- $f$ are ellipsoids and $\kappa(f) > 1$
- $g$ are circles for ideal $S$ because $\kappa(g) \approx 1$





# Question

Consider the change of variables $x = Sy$ to the quadratic function

$$f(x) = \frac{1}{2} x^T H x,$$

to obtain the scaled function

$$g(y) = f(Sy).$$

Which choice of the nonsingular scaling matrix $S$ will transform the level sets of $g(y)$ into circles (i.e., result in a perfectly conditioned Hessian for $g$)?

a. $S = I$ (the identity matrix)

b. $S = H$

c. $S = H^{-1/2}$

d. $S = \mathrm{diag}(H)$ (the diagonal part of $H$)

# Common scalings

Make $S^{(k)}\nabla^2 f(x^{(k)})S^{(k)}$ as well conditioned as possible

$$S^{(k)}(S^{(k)})^T = \begin{cases} (\nabla f(x^{(k)}))^{-1} & \text{Newton } (\kappa = 1) \\ (\nabla f(x^{(k)}) + \lambda I)^{-1} & \text{damped Newton} \\ \mathbf{Diag}\left(\frac{\partial^2 f(x^{(k)})}{\partial x_i^2}\right)^{-1} & \text{diagonal scaling} \end{cases}$$

# Gauss Newton
# Nonlinear Least Squares

# Nonlinear least squares

- NLLS (nonlinear least-squares) problem

$$\min_{x \in \mathbb{R}^n} \quad f(x) := \frac{1}{2}\|r(x)\|_2^2, \quad r : \mathbb{R}^n \to \mathbb{R}^m \quad (\text{typically}, m > n).$$

- gradient and residual vector (Jacobian $J(x)$)

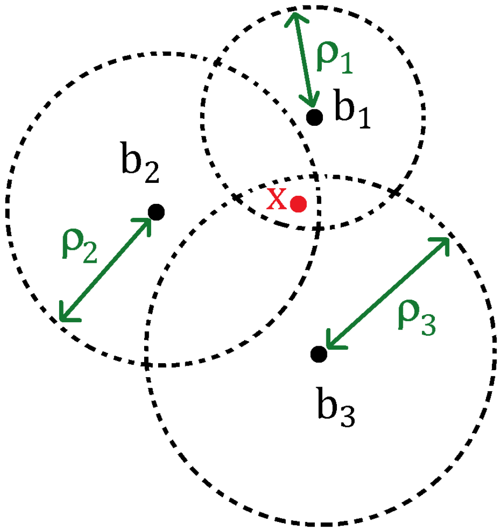
$$r(x) = \begin{bmatrix} r_1(x) \\ r_2(x) \\ \vdots \\ r_m(x) \end{bmatrix}, \quad \nabla f(x) = J(x)^T r(x), \quad J(x) = \begin{bmatrix} \nabla r_1(x)^T \\ \nabla r_2(x)^T \\ \vdots \\ \nabla r_m(x)^T \end{bmatrix}$$

- reduces to linear least-squares when $r$ is **affine**

$$r(x) = Ax - b$$

# Example – localication problem

- estimate $x \in \mathbb{R}^2$ from approximate distances to known fixed beacons




**data**

- $m$ beacons at known locations $b_1, \ldots, b_m$
- approximate distances

$$d_i = \|x - b_i\|_2 + \epsilon_i$$

where $\epsilon_i$ is measurement error

- NLLS position estimate solves

$$\min_x \quad \frac{1}{2}\sum_{i=1}^{m} r_i(x), \quad r_i(x) = \|x - b_i\|_2 - d_i$$

- must settle for locally optimal solution

# Linearization of residual

- **linearize** $r(x)$ about $\bar{x}$

$$r(x) = \begin{bmatrix} r_1(x) \\ r_2(x) \\ \vdots \\ r_m(x) \end{bmatrix} = \begin{bmatrix} r_1(\bar{x}) + \nabla r_1(\bar{x})^T(x - \bar{x}) \\ r_2(\bar{x}) + \nabla r_2(\bar{x})^T(x - \bar{x}) \\ \vdots \\ r_m(\bar{x}) + \nabla r_m(\bar{x})^T(x - \bar{x}) \end{bmatrix} + o(\|x - \bar{x}\|)$$

$$= J(\bar{x})(x - \bar{x}) + r(\bar{x}) + o(\|x - \bar{x}\|)$$

$$= J(\bar{x})x - \underbrace{(J(\bar{x})\bar{x} - r(\bar{x}))}_{=:b(\bar{x})} + o(\|x - \bar{x}\|)$$

- **pure Gauss Newton** iteration: use linearized least-squares problem used to determine $x^{(k+1)}$

$$x^{(k+1)} = \underset{x}{\operatorname{argmin}} \ \tfrac{1}{2}\|J(x^k)x - b(x^k)\|_2^2 \quad \text{or} \quad x^{(k+1)} = J(x^k)\backslash b(x^k)$$

# Gauss Newton as scaled descent

- expand the least squares subproblem (set $J_k := J(x^k)$ and $b_k := b(x^k)$). If $J_k$ full rank,

$$\begin{aligned} x^{(k+1)} &= \underset{x}{\operatorname{argmin}} \ \|J_k x - b_k\|^2 \\ &= (J_k^T J_k)^{-1} J_k^\intercal b_k \\ &= (J_k^T J_k)^{-1} J_k^T (J_k x^k - r_k) \\ &= x^k - (J_k^T J_k)^{-1} J_k^T r_k \end{aligned}$$

- interpret at **scaled** gradient descent

$$x^{k+1} = x^k + d^k, \qquad d^k := \underbrace{(J_k^T J_k)^{-1}}_{=D_k \succ 0} \underbrace{(-J_k^T r_k)}_{=-\nabla f(x^k)}$$

- Hessian of objective $f(x) = \frac{1}{2}\|r(x)\|^2$

$$\nabla^2 f(x) = J(x)^T J(x) + \sum_{i=1}^{m} \nabla^2 r_i(x)$$

# Gauss Newton for NLLS

$$\min_x \; f(x) = \tfrac{1}{2}\|r(x)\|_2^2, \quad r : \mathbb{R}^n \to \mathbb{R}^m$$

- linesearch on nonlinear objective $f(x) = \frac{1}{2}\|r(x)\|^2$ required to ensure convergence

$$x^{k+1} = x^k + \alpha^k d^k, \qquad d^k = \operatorname*{argmin}_d \; \|J_k d - r_k\|^2$$

## Gauss Newton for NLLS

- given starting point $x^0$ and stopping tolerance $\epsilon > 0$
- for $k = 0, 1, 2, \ldots$
  1. compute residual $r_k = r(x^k)$ and Jacobian $J_k = J(x^k)$
  2. compute step $d^k = \operatorname{argmin}_d \; \|J_k d + r_k\|^2$, ie, $d^k = -J_k \backslash r_k$
  3. choose stepsize $\alpha^k \in (0, 1]$ via linesearch on $f(x)$
  4. update $x^{k+1} = x^k + \alpha^k d^k$
  5. stop if $\|r(x^{k+1})\| < \epsilon$ or $\|\nabla f(x^k)\| = \|J_k^T r_k\| < \epsilon$