DTU

January 17-18, 2023

02946 Scientific Computing for X-Ray Computed Tomography

Optimization Methods for Tomography

About me

Martin S. Andersen

Associate Professor
Section for Scientific Computing
DTU Compute

Head of Studies, MSc Eng MMC

PhD from UCLA (EE)
MSc from Aalborg University (EE)

Research in optimization and its applications

June 2023: 02953 Convex Optimization (5 ECTS)

Tentative schedule

Chapter 13 in textbook

Tuesday, January 17	Unconstrained optimization Lipschitz continuity Majorization minimization Convexity Step size rules & stopping criteria Power iteration
Wednesday, January 18	Constrained optimization Convex sets Proximal gradient method Optimality conditions Accelerated proximal gradient method Smoothing techniques

Optimization for tomography

Maximum likelihood (ML) estimation

$$\hat{\boldsymbol{x}} = \operatorname*{argmax}_{\boldsymbol{x}} \{\pi(\boldsymbol{b} \,|\, \boldsymbol{x})\} = \operatorname*{argmin}_{\boldsymbol{x}} \{-\ln \pi(\boldsymbol{b} \,|\, \boldsymbol{x})\}$$

Maximum a posteriori (MAP) estimation

$$\begin{aligned} \hat{\boldsymbol{x}} &= \operatorname*{argmax}_{\boldsymbol{x}} \{\pi(\boldsymbol{b} \,|\, \boldsymbol{x})\pi(\boldsymbol{x})\} \\ &= \operatorname*{argmin}_{\boldsymbol{x}} \{-\ln \pi(\boldsymbol{b} \,|\, \boldsymbol{x}) - \ln \pi(\boldsymbol{x})\} \end{aligned}$$

Example

$$\text{minimize} \quad \tfrac{1}{2}\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \gamma\, R(\boldsymbol{x}) + \text{const.}$$

Unconstrained optimization

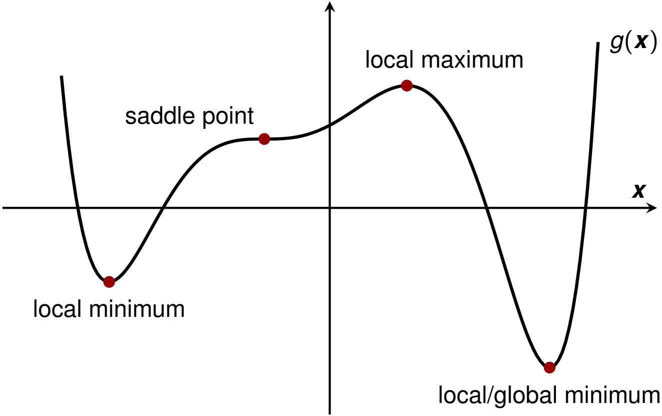
$$\text{minimize} \quad g(\boldsymbol{x})$$

- variable $\boldsymbol{x} \in \mathbb{R}^n$
- objective function $g : \mathbb{R}^n \to \mathbb{R}$ continuously differentiable
- global minimum at $\boldsymbol{x}^\star$ if $g(\boldsymbol{y}) \geq g(\boldsymbol{x}^\star)$ for all $\boldsymbol{y} \in \mathbb{R}^n$
- $\boldsymbol{x}$ is a stationary point of g if

$$\nabla g(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial g(\boldsymbol{x})}{\partial x_1} \\ \vdots \\ \frac{\partial g(\boldsymbol{x})}{\partial x_n} \end{bmatrix} = \mathbf{0}$$

- stationarity is a necessary condition for global optimality

Stationary points



Gradient method

Iterative update of image

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - t_k \nabla g(\boldsymbol{x}^{(k)}), \quad k = 0, 1, 2, \ldots$$

- step size $t_k > 0$
- directional derivative of g at $\boldsymbol{x}^{(k)}$ in the direction $-\nabla g(\boldsymbol{x}^{(k)})$ is

$$-\nabla g(\boldsymbol{x}^{(k)})^T \nabla g(\boldsymbol{x}^{(k)}) = -\|\nabla g(\boldsymbol{x}^{(k)})\|_2^2$$

- directional derivative is negative unless $\boldsymbol{x}^{(k)}$ is a stationary point
- implies that $-\nabla g(\boldsymbol{x}^{(k)})$ is a descent direction if $\boldsymbol{x}^{(k)}$ is not stationary
- descent it guaranteed if we choose t_k such that $g(\boldsymbol{x}^{(k+1)}) < g(\boldsymbol{x}^{(k)})$

Exact line search

Cauchy's step size rule: minimize g along the current search direction

$$t_k = \operatorname*{argmin}_{t>0} \left\{ g(\boldsymbol{x}^{(k)} - t\nabla g(\boldsymbol{x}^{(k)})) \right\}$$

- "greedy" heuristic
- may be as expensive to solve as original problem

Example: least-squares objective

$$g(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2^2$$

- gradient $\nabla g(\boldsymbol{x}) = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})$
- exact line search

$$t_k = \underset{t>0}{\operatorname{argmin}} \left\{ g(\boldsymbol{x}^{(k)} - t\nabla g(\boldsymbol{x}^{(k)})) \right\} = \frac{\|\nabla g(\boldsymbol{x}^{(k)})\|_2^2}{\|\boldsymbol{A}\nabla g(\boldsymbol{x}^{(k)})\|_2^2} = \frac{\|\boldsymbol{A}^T \varrho^{(k)}\|_2^2}{\|\boldsymbol{A}\boldsymbol{A}^T \varrho^{(k)}\|_2^2}$$

which follows from $\varrho^{(k)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(k)}$ and

$$\frac{d}{dt} g(\boldsymbol{x}^{(k)} - t\nabla g(\boldsymbol{x}^{(k)})) = t\|\boldsymbol{A}\nabla g(\boldsymbol{x}^{(k)})\|_2^2 - \|\nabla g(\boldsymbol{x}^{(k)})\|_2^2 = 0$$

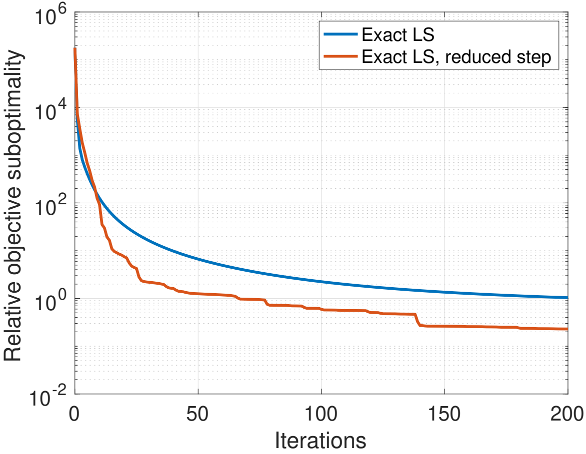
Example: least-squares objective (cont.)

Relative suboptimality

$$\frac{|g(\boldsymbol{x}^{(k)}) - g(\boldsymbol{x}^{\star})|}{|g(\boldsymbol{x}^{\star})|}$$

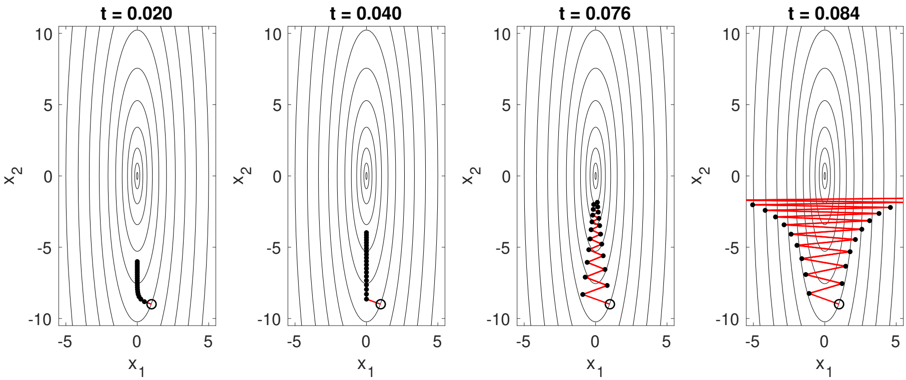
Adjusted step size

$$\gamma t_k, \quad \gamma = 0.9$$



Example: gradient method with fixed step size (first 20 iterations)

$$g(\boldsymbol{x}) = \frac{1}{2}(25x_1^2 + x_2^2)$$



Lipschitz continuity

Gradient ∇g is Lipschitz continuous if there exists a constant L such that

$$\|\nabla g(\boldsymbol{x}) - \nabla g(\boldsymbol{y})\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2, \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$$

Quadratic upper bound

If ∇g is Lipschitz continuous with constant L, then

$$g(\boldsymbol{y}) \leq g(\boldsymbol{x}) + \nabla g(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2, \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$$

Quadratic upper bound: derivation

- Define restriction of g to line through $\boldsymbol{x}$ and $\boldsymbol{y}$: $\phi(\tau) = g(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x}))$
- Newton–Leibniz integral rule: $\phi(1) - \phi(0) = \int_0^1 \phi'(\tau)\, d\tau$

$$\begin{aligned}
g(\boldsymbol{y}) - g(\boldsymbol{x}) &= \int_0^1 \nabla g(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x}))^T(\boldsymbol{y} - \boldsymbol{x})\, d\tau \\
&= \nabla g(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \int_0^1 (\nabla g(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})) - \nabla g(\boldsymbol{x}))^T(\boldsymbol{y} - \boldsymbol{x})\, d\tau \\
&\leq \nabla g(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \int_0^1 \|\nabla g(\boldsymbol{x} + \tau(\boldsymbol{y} - \boldsymbol{x})) - \nabla g(\boldsymbol{x})\|_2 \|\boldsymbol{y} - \boldsymbol{x}\|_2\, d\tau \\
&\leq \nabla g(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \int_0^1 \tau L \|\boldsymbol{y} - \boldsymbol{x}\|_2^2\, d\tau
\end{aligned}$$

Example

Least-squares objective

$$g(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2^2$$

with gradient $\nabla g(\boldsymbol{x}) = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})$

Implies that

$$\|\nabla g(\boldsymbol{y}) - \nabla g(\boldsymbol{x})\|_2 = \|\boldsymbol{A}^T\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{x})\|_2 \leq \|\boldsymbol{A}^T\boldsymbol{A}\|_2\|\boldsymbol{y} - \boldsymbol{x}\|_2$$

and hence ∇g is Lipschitz continuous with constant $L = \|\boldsymbol{A}\|_2^2$

Twice continuously differentiable functions

Suppose g is twice continuously differentiable with Hessian matrix

$$\nabla^2 g(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial^2 g(\boldsymbol{x})}{\partial x_1^2} & \frac{\partial^2 g(\boldsymbol{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 g(\boldsymbol{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 g(\boldsymbol{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 g(\boldsymbol{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 g(\boldsymbol{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 g(\boldsymbol{x})}{\partial x_n \partial x_1} & \frac{\partial^2 g(\boldsymbol{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 g(\boldsymbol{x})}{\partial x_n^2} \end{bmatrix}$$

Bounded Hessian

∇g is Lipschitz continuous with constant L iff $\|\nabla^2 g(\boldsymbol{x})\|_2 \leq L$ for all $\boldsymbol{x}$

Exercise 13.1: Step size rules for least-squares problems

Consider the gradient method applied to the least-squares objective function $g(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2^2$, i.e.,

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - t_k \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x}^{(k)} - \boldsymbol{b}), \quad k = 0, 1, 2 \ldots$$

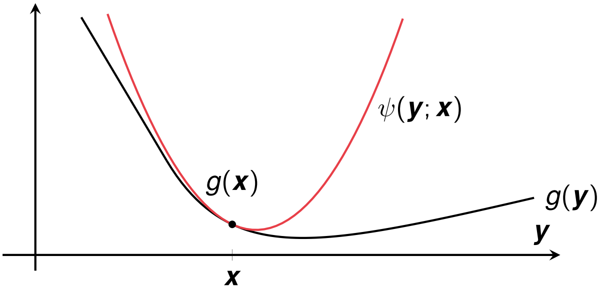
where $\boldsymbol{x}^{(0)}$ is an initial guess. For each of the following step size rules, show that the gradient iteration can be implemented such that each iteration only requires a single matrix-vector multiplication with $\boldsymbol{A}$ and one with $\boldsymbol{A}^T$.

1. The step size t_k is constant, i.e., $t_k = t > 0$ for all k.
2. The step size t_k is found by means of the exact line search.

Majorization

A function $\psi(\boldsymbol{y}; \boldsymbol{x})$ is said to be a majorization of g at $\boldsymbol{x}$ if

$$\psi(\boldsymbol{x}; \boldsymbol{x}) = g(\boldsymbol{x}) \quad \text{and} \quad \psi(\boldsymbol{y}; \boldsymbol{x}) \geq g(\boldsymbol{y}), \quad \text{for all } \boldsymbol{y}$$



Majorization minimization

Iterative update based on majorization

$$\boldsymbol{x}^{(k+1)} = \underset{\boldsymbol{y}}{\operatorname{argmin}} \left\{ \psi(\boldsymbol{y}; \boldsymbol{x}^{(k)}) \right\}, \quad k = 0, 1, 2, \ldots$$

- $\boldsymbol{x}^{(k+1)}$ minimizes the majorization $\psi(\boldsymbol{y}; \boldsymbol{x}^{(k)})$ so

$$\psi(\boldsymbol{x}^{(k+1)}; \boldsymbol{x}^{(k)}) \leq \psi(\boldsymbol{x}^{(k)}; \boldsymbol{x}^{(k)})$$

- properties of majorization imply that

$$g(\boldsymbol{x}^{(k+1)}) \leq \psi(\boldsymbol{x}^{(k+1)}; \boldsymbol{x}^{(k)}) \leq \psi(\boldsymbol{x}^{(k)}; \boldsymbol{x}^{(k)}) = g(\boldsymbol{x}^{(k)})$$

Majorization minimization: quadratic majorization

Use quadratic upper bound to construct majorization

$$\psi(\boldsymbol{y};\boldsymbol{x}) = g(\boldsymbol{x}) + \nabla g(\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2$$

with ∇g Lipschitz continuous with constant L

Gradient method with constant step size

$$\boldsymbol{x}^{(k+1)} = \underset{\boldsymbol{y}}{\operatorname{argmin}} \left\{\psi(\boldsymbol{y};\boldsymbol{x}^{(k)})\right\} = \boldsymbol{x}^{(k)} - t_k \nabla g(\boldsymbol{x}^{(k)}), \quad t_k = \frac{1}{L}$$

Analysis of gradient method with constant step size

Majorization property $g(\mathbf{x}^{(k+1)}) \leq \psi(\mathbf{x}^{(k+1)}; \mathbf{x}^{(k)})$ implies that

$$g(\mathbf{x}^{(k+1)}) \leq g(\mathbf{x}^{(k)}) - \frac{1}{2L}\|\nabla g(\mathbf{x}^{(k)})\|_2^2$$

- summing inequality for $k = 0, \ldots, N$,

$$\frac{1}{2L}\sum_{k=0}^{N}\|\nabla g(\mathbf{x}^{(k)})\|_2^2 \leq g(\mathbf{x}^{(0)}) - g(\mathbf{x}^{(N+1)}) \leq g(\mathbf{x}^{(0)}) - g(\mathbf{x}^{\star})$$

- converges to stationary point if $g(\mathbf{x}^{(0)}) - g(\mathbf{x}^{\star})$ is finite
- step size $t_k \in (0, 2/L)$ yields a descent unless $\nabla g(\mathbf{x}^{(k)}) = 0$

Exercise 13.2: Lipschitz continuous gradients

Suppose $g_1 : \mathbb{R}^n \to \mathbb{R}$ and $g_2 : \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable.

Show that if ∇g_1 and ∇g_2 are Lipschitz continuous with constants L_1 and L_2, respectively, then $\nabla g(\boldsymbol{x}) = \nabla g_1(\boldsymbol{x}) + \nabla g_2(\boldsymbol{x})$ is Lipschitz continuous with constant $L = L_1 + L_2$.

SIRT-like methods

Recall the SIRT method

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \lambda_k \boldsymbol{D}\boldsymbol{A}^T \boldsymbol{M}(\boldsymbol{A}\boldsymbol{x}^{(k)} - \boldsymbol{b})$$

where $\lambda_k \in (0, 2)$ and $\boldsymbol{M}$ and $\boldsymbol{D}$ are positive diagonal matrices

May be viewed as *scaled* gradient method for minimizing

$$g(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})^T \boldsymbol{M}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_{\boldsymbol{M}}^2$$

with gradient $\nabla g(\boldsymbol{x}) = \boldsymbol{A}^T \boldsymbol{M}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})$

SIRT-like methods (cont.)

Gradient satisfies

$$\begin{aligned}\|\nabla g(\boldsymbol{y}) - \nabla g(\boldsymbol{x})\|_{\boldsymbol{D}} &= \|\boldsymbol{D}^{1/2}\boldsymbol{A}^T\boldsymbol{M}\boldsymbol{A}\boldsymbol{D}^{1/2}\boldsymbol{D}^{-1/2}(\boldsymbol{y}-\boldsymbol{x})\|_2 \\ &\leq \|\boldsymbol{M}^{1/2}\boldsymbol{A}\boldsymbol{D}^{1/2}\|_2^2\|\boldsymbol{D}^{-1/2}(\boldsymbol{y}-\boldsymbol{x})\|_2\end{aligned}$$

Assuming that $\boldsymbol{D}$ and $\boldsymbol{M}$ satisfy $\|\boldsymbol{M}^{1/2}\boldsymbol{A}\boldsymbol{D}^{1/2}\|_2 \leq 1$,

$$\|\nabla g(\boldsymbol{y}) - \nabla g(\boldsymbol{x})\|_{\boldsymbol{D}} \leq \|\boldsymbol{y}-\boldsymbol{x}\|_{\boldsymbol{D}^{-1}}$$

SIRT-like methods (cont.)

Condition $\|\nabla g(\boldsymbol{y}) - \nabla g(\boldsymbol{x})\|_{\boldsymbol{D}} \leq \|\boldsymbol{y} - \boldsymbol{x}\|_{\boldsymbol{D}^{-1}}$ implies quadratic upper bound

$$g(\boldsymbol{y}) \leq g(\boldsymbol{x}) + \nabla g(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^T \boldsymbol{D}^{-1}(\boldsymbol{y} - \boldsymbol{x})$$

Majorization minimization method for minimizing g

$$\begin{aligned} \boldsymbol{x}^{(k+1)} &= \boldsymbol{x}^{(k)} - \lambda_k \boldsymbol{D} \nabla g(\boldsymbol{x}^{(k)}) \\ &= \boldsymbol{x}^{(k)} - \lambda_k \boldsymbol{D} \boldsymbol{A}^T \boldsymbol{M}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) \end{aligned}$$

with $\lambda_k \in (0, 2)$

SIRT-like methods (cont.)

$\|\boldsymbol{M}^{1/2}\boldsymbol{A}\boldsymbol{D}^{1/2}\|_2 \leq 1$ is satisfied for $\boldsymbol{D}$ and $\boldsymbol{M}$ defined as

$$\boldsymbol{D}_{jj}^{-1} = \sum_{i=1}^{m} |\boldsymbol{A}_{ij}|^{\alpha}, \qquad \boldsymbol{M}_{ii}^{-1} = \sum_{j=1}^{n} |\boldsymbol{A}_{ij}|^{2-\alpha}, \quad \alpha \in [0,2]$$

- we define $|\boldsymbol{A}_{ij}|^0 = 1$ when $\boldsymbol{A}_{ij} = 0$
- objective function $g(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_{\boldsymbol{M}}^2$ depends on α
- Cimmino's method: $\alpha = 0$
- SIRT: $\alpha = 1$
- "parallel" coordinate descent: $\alpha = 2$

Exercise 13.3: SIRT-like methods

Recall that the SIRT iteration solves a weighted least-squares problem of the form

$$\text{minimize} \quad \frac{1}{2}\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_{\boldsymbol{M}}^2, \qquad \boldsymbol{M} \text{ diag. positive definite.}$$

1. Show that $\|\boldsymbol{M}^{1/2}\boldsymbol{A}\boldsymbol{D}^{1/2}\|_2 \leq 1$ if $\boldsymbol{M}$ and $\boldsymbol{D}$ are diagonal matrices and

$$\boldsymbol{D}_{jj}^{-1} = \sum_{i=1}^{m} |\boldsymbol{A}_{ij}|^{\alpha}, \qquad \boldsymbol{M}_{ii}^{-1} = \sum_{j=1}^{n} |\boldsymbol{A}_{ij}|^{2-\alpha}, \qquad \alpha \in [0, 2].$$

   Hint: Show that $\|\boldsymbol{M}^{1/2}\boldsymbol{A}\boldsymbol{D}^{1/2}\boldsymbol{x}\|_2^2 \leq \|\boldsymbol{x}\|_2^2$ when $\alpha \in [0, 2]$.

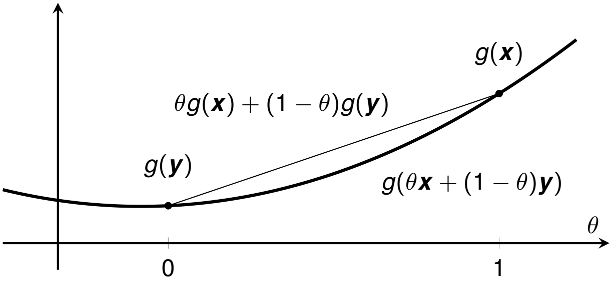
2. Implement the SIRT iteration in MATLAB with α as an input parameter.
3. Compute reconstructions for different α (see textbook for details).

Convexity

$g : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if

$$g(\theta \boldsymbol{x} + (1-\theta)\boldsymbol{y}) \leq \theta g(\boldsymbol{x}) + (1-\theta) g(\boldsymbol{y}), \quad \theta \in [0,1]$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ (g is concave if $-g$ is convex)

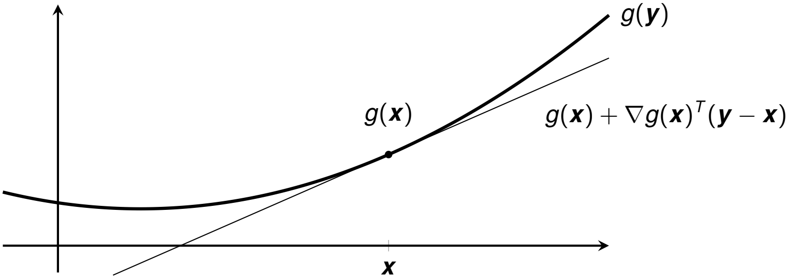


Convexity: first-order condition

Continuously differentiable g is convex if and only if

$$g(\boldsymbol{y}) \geq g(\boldsymbol{x}) + \nabla g(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x})$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$



Convexity: implications

- stationary points are global minimizers

$$g(\boldsymbol{y}) \geq g(\boldsymbol{x}^\star) + \nabla g(\boldsymbol{x}^\star)^T(\boldsymbol{y} - \boldsymbol{x}^\star) = g(\boldsymbol{x}^\star), \quad \text{for all } \boldsymbol{y} \in \mathbb{R}^n$$

- gradient method with step size $t_k = \gamma/L$ and $\gamma \in (0, 2)$ satisfies

$$g(\boldsymbol{x}^{(k)}) - g(\boldsymbol{x}^\star) \leq \frac{2L\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^\star\|_2^2}{4 + \gamma(2-\gamma)k}.$$

  if ∇g is Lipschitz continuous with constant L
- suboptimality satisfies $g(\boldsymbol{x}^{(k)}) - g^\star = O(1/k)$
- at most $O(1/\epsilon)$ iterations required before $g(\boldsymbol{x}^{(k)}) - g^\star \leq \epsilon$

Strong convexity

g is strongly convex with parameter $\mu > 0$ if

$$g(\theta \boldsymbol{x} + (1-\theta)\boldsymbol{y}) \leq \theta g(\boldsymbol{x}) + (1-\theta) g(\boldsymbol{y}) - \frac{\theta(1-\theta)\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

for all $\theta \in [0, 1]$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

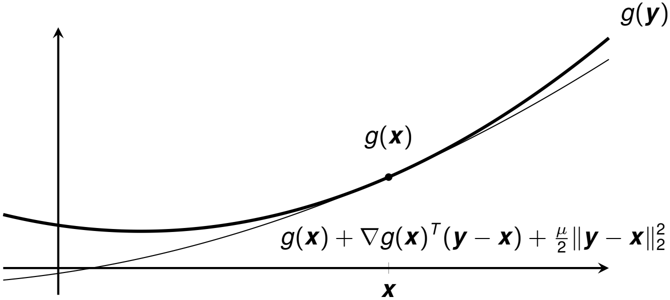
Interpretation: $\tilde{g}(\boldsymbol{x}) = g(\boldsymbol{x}) - \frac{\mu}{2}\|\boldsymbol{x}\|_2^2$ is convex

Strong convexity: first-order condition

Continuously differentiable g: strongly convex with parameter $\mu > 0$ if

$$g(\boldsymbol{y}) \geq g(\boldsymbol{x}) + \nabla g(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2, \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$$

Implies that minimizer is unique



Strong convexity: implications

- minimizing quadratic lower bound wrt. $\boldsymbol{y}$ yields

$$g(\boldsymbol{y}) \geq g(\boldsymbol{x}) - \frac{1}{2\mu}\|\nabla g(\boldsymbol{x})\|_2^2 \implies g(\boldsymbol{x}) - g(\boldsymbol{x}^\star) \leq \frac{1}{2\mu}\|\nabla g(\boldsymbol{x})\|_2^2$$

- substitute $\boldsymbol{x}^\star$ for $\boldsymbol{y}$ in first-order condition

$$\begin{aligned} g(\boldsymbol{x}^\star) &\geq g(\boldsymbol{x}) + \nabla g(\boldsymbol{x})^T(\boldsymbol{x}^\star - \boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{x}^\star - \boldsymbol{x}\|_2^2 \\ &\geq g(\boldsymbol{x}) - \|\nabla g(\boldsymbol{x})\|_2\|\boldsymbol{x}^\star - \boldsymbol{x}\|_2 + \frac{\mu}{2}\|\boldsymbol{x}^\star - \boldsymbol{x}\|_2^2 \end{aligned}$$

$g(\boldsymbol{x}^\star) \leq g(\boldsymbol{x})$ implies that

$$\|\boldsymbol{x}^\star - \boldsymbol{x}\|_2 \leq \frac{2}{\mu}\|\nabla g(\boldsymbol{x})\|_2$$

Strong convexity: implications (cont.)

- gradient method with step size $t_k = 2/(L+\mu)$ with satisfies

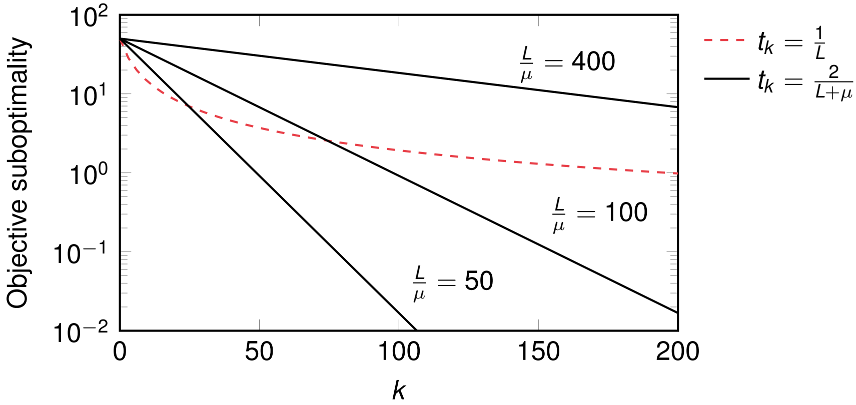
$$g(\boldsymbol{x}^{(k)}) - g(\boldsymbol{x}^\star) \leq \frac{L}{2}\left(\frac{L-\mu}{L+\mu}\right)^{2k} \|\boldsymbol{x}^{(0)} - \boldsymbol{x}^\star\|_2^2,$$

  and

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^\star\|_2 \leq \left(\frac{L-\mu}{L+\mu}\right)^{k} \|\boldsymbol{x}^{(0)} - \boldsymbol{x}^\star\|_2$$

  if g is μ-strongly convex with L-Lipschitz gradient
- implies that $\boldsymbol{x}^{(k)}$ converges linearly to $\boldsymbol{x}^\star$

Comparison of worst-case suboptimality bounds



Example: least-squares problem

$$g(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2^2$$

Linearization of gradient around $\boldsymbol{x}^\star$ yields

$$\nabla g(\boldsymbol{x}^{(k)}) = \nabla g(\boldsymbol{x}^\star) + \nabla^2 g(\boldsymbol{x}^\star)(\boldsymbol{x}^{(k)} - \boldsymbol{x}^\star)$$

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - t\boldsymbol{A}^T\boldsymbol{A}(\boldsymbol{x}^{(k)} - \boldsymbol{x}^\star),$$

Subtract $\boldsymbol{x}^\star$ from both sides, take norm, and use $\|\boldsymbol{M}\boldsymbol{x}\|_2 \leq \|\boldsymbol{M}\|_2\|\boldsymbol{x}\|_2$

$$\begin{aligned}\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^\star &= (\boldsymbol{I} - t\boldsymbol{A}^T\boldsymbol{A})(\boldsymbol{x}^{(k)} - \boldsymbol{x}^\star) \\ &= (\boldsymbol{I} - t\boldsymbol{A}^T\boldsymbol{A})^{k+1}(\boldsymbol{x}^{(0)} - \boldsymbol{x}^\star)\end{aligned}$$

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^\star\|_2 \leq \|\boldsymbol{I} - t\boldsymbol{A}^T\boldsymbol{A}\|_2^k\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^\star\|_2$$

Example: least-squares problem (cont.)

Suppose eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}$ belong to the interval $[\mu, L]$ where $L = \|\boldsymbol{A}\|_2^2$

Choose t such that it minimizes the spectral radius of $\boldsymbol{I} - t\boldsymbol{A}^T\boldsymbol{A}$

$$\begin{aligned} t^\star &= \operatorname*{argmin}_t \{\|\boldsymbol{I} - t\boldsymbol{A}^T\boldsymbol{A}\|_2\} \\ &= \operatorname*{argmin}_t \{\max_{\lambda \in [\mu, L]} |1 - t\lambda|\} \\ &= \operatorname*{argmin}_t \{\max\{1 - t\mu, 1 - tL, t\mu - 1, tL - 1\}\} \\ &= \frac{2}{L + \mu} \end{aligned}$$

Spectral radius of $\boldsymbol{I} - t^\star\boldsymbol{A}^T\boldsymbol{A}$ is $(L - \mu)/(L + \mu)$

Exercises 13.4 and 13.5

13.4 **Strong convexity.** Suppose g is a twice continuously differentiable and strongly convex function with strong convexity parameter μ.

1. Show that the smallest eigenvalue of $\nabla^2 g(\boldsymbol{x})$ is bounded below by μ.
2. Consider the regularized least-squares objective function

$$g(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \frac{\delta}{2}\|\boldsymbol{x}\|_2^2, \qquad \delta > 0.$$

   Derive the Lipschitz constant for ∇g and a lower bound on μ.

13.5 **Poisson measurements.** The negative log-likelihood function is

$$g(\boldsymbol{x}) = \mathbf{1}^T \exp(-\boldsymbol{A}\boldsymbol{x}) + \exp(-\boldsymbol{b})^T \boldsymbol{A}\boldsymbol{x} + \text{const.}$$

where $\boldsymbol{b} = -\log(\boldsymbol{I}/\boldsymbol{I}_0)$ and $\boldsymbol{I}$ is assumed to be positive.
(Refer to textbook for questions.)

Power iteration for matrix norm estimation

$$\|\boldsymbol{H}\|_2 = \sup_{\boldsymbol{x} \neq 0} \left\{ \frac{\boldsymbol{x}^T \boldsymbol{H} \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} \right\} = \lambda_{\max}(\boldsymbol{H}) \qquad \boldsymbol{H} \text{ symmetric}$$

Power iteration

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{H}\boldsymbol{x}^{(k)} / \|\boldsymbol{H}\boldsymbol{x}^{(k)}\|_2, \quad k = 0, 1, 2, \ldots, \quad \text{with } \boldsymbol{x}^{(0)} \text{ random}$$

$$\hat{\lambda}^{(k)} = \|\boldsymbol{H}\boldsymbol{x}^{(k)}\|_2 \xrightarrow{\text{a.s.}} \lambda_{\max}(\boldsymbol{H}) \quad \text{as} \quad k \to \infty$$

Why it works: let $\boldsymbol{H} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T$ and $\boldsymbol{\alpha} = \boldsymbol{V}^T\boldsymbol{x}^{(0)}$

$$\boldsymbol{x}^{(k)} = \boldsymbol{H}^k\boldsymbol{x}^{(0)} / \|\boldsymbol{H}^k\boldsymbol{x}^{(0)}\|_2, \quad k = 1, 2, \ldots$$

$$\boldsymbol{H}^k\boldsymbol{x}^{(0)} = \boldsymbol{V}\boldsymbol{\Lambda}^k\boldsymbol{V}^T\boldsymbol{x}^{(0)} = \sum_{i=1}^{n} \alpha_i \lambda_i^k \boldsymbol{v}_i, \quad k = 1, 2, \ldots$$

Power iteration for matrix norm estimation

$$\|\boldsymbol{H}\|_2 = \sup_{\boldsymbol{x}\neq 0}\left\{\frac{\boldsymbol{x}^T\boldsymbol{H}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}}\right\} = \lambda_{\max}(\boldsymbol{H}) \qquad \boldsymbol{H} \text{ symmetric}$$

Power iteration

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{H}\boldsymbol{x}^{(k)}/\|\boldsymbol{H}\boldsymbol{x}^{(k)}\|_2, \quad k = 0, 1, 2, \ldots, \quad \text{with } \boldsymbol{x}^{(0)} \text{ random}$$

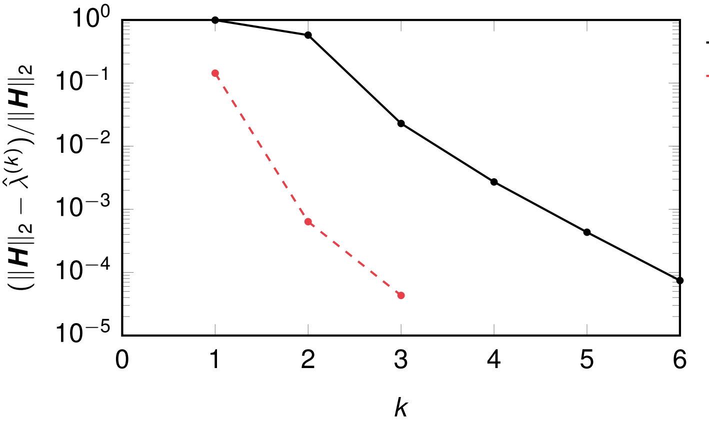
$$\hat{\lambda}^{(k)} = \|\boldsymbol{H}\boldsymbol{x}^{(k)}\|_2 \xrightarrow{\text{a.s.}} \lambda_{\max}(\boldsymbol{H}) \quad \text{as} \quad k \to \infty$$

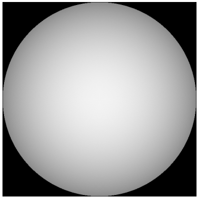
Why it works: let $\boldsymbol{H} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T$ and $\boldsymbol{\alpha} = \boldsymbol{V}^T\boldsymbol{x}^{(0)}$

$$\boldsymbol{x}^{(k)} = \boldsymbol{H}^k\boldsymbol{x}^{(0)}/\|\boldsymbol{H}^k\boldsymbol{x}^{(0)}\|_2, \quad k = 1, 2, \ldots$$

$$\lambda_1^{-k}\boldsymbol{H}^k\boldsymbol{x}^{(0)} = \alpha_1\boldsymbol{v}_1 + \alpha_2\left(\frac{\lambda_2}{\lambda_1}\right)^k\boldsymbol{v}_2 + \cdots + \alpha_n\left(\frac{\lambda_n}{\lambda_1}\right)^k\boldsymbol{v}_n$$

Example: $\boldsymbol{H} = \boldsymbol{A}^T\boldsymbol{A}$





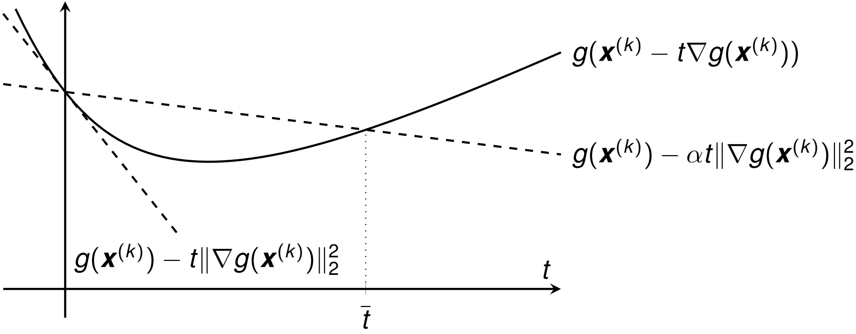
$\hat{\boldsymbol{x}}$

Remarks: (i) avoid forming $\boldsymbol{H}$, (ii) similar to MATLAB's `normest()`

Backtracking line search

Armijo condition for gradient method

$$g(\boldsymbol{x}^{(k)} - t\nabla g(\boldsymbol{x}^{(k)})) \leq g(\boldsymbol{x}^{(k)}) - \alpha t\|\nabla g(\boldsymbol{x})\|_2^2$$



Backtracking line search (cont.)

Backtracking line search

Require: $\alpha \in (0, \frac{1}{2})$, $\beta \in (0, 1)$, and $t = t_0 > 0$

while $g(\boldsymbol{x}^{(k)} - t\nabla g(\boldsymbol{x}^{(k)})) > g(\boldsymbol{x}^{(k)}) - \alpha t \|\nabla g(\boldsymbol{x}^{(k)})\|_2^2$ **do**

$t \leftarrow t\beta$

end while

- α controls a trade-off between max. step length and required decrease
- β controls backtracking "aggresiveness"
- typical values are $\alpha = 10^{-2}$ and $\beta = 0.7$

Barzilai–Borwein step size rules

Quadratic approximation

$$g(\boldsymbol{y}) \approx g(\boldsymbol{x}) + \nabla g(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{\alpha}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2$$

$$\nabla g(\boldsymbol{y}) - \nabla g(\boldsymbol{x}) \approx \alpha(\boldsymbol{y} - \boldsymbol{x})$$

Define $\Delta \boldsymbol{y} = \nabla g(\boldsymbol{x}^{(k)}) - \nabla g(\boldsymbol{x}^{(k-1)})$ and $\Delta \boldsymbol{s} = \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}$ $(k \geq 1)$

$$t_k^{\mathrm{BB1}} = \alpha_k^{-1}, \quad \alpha_k = \operatorname*{argmin}_{\alpha} \left\{\|\Delta \boldsymbol{y} - \alpha \Delta \boldsymbol{s}\|_2^2\right\} = \frac{\Delta \boldsymbol{s}^T \Delta \boldsymbol{y}}{\|\Delta \boldsymbol{s}\|_2^2}$$

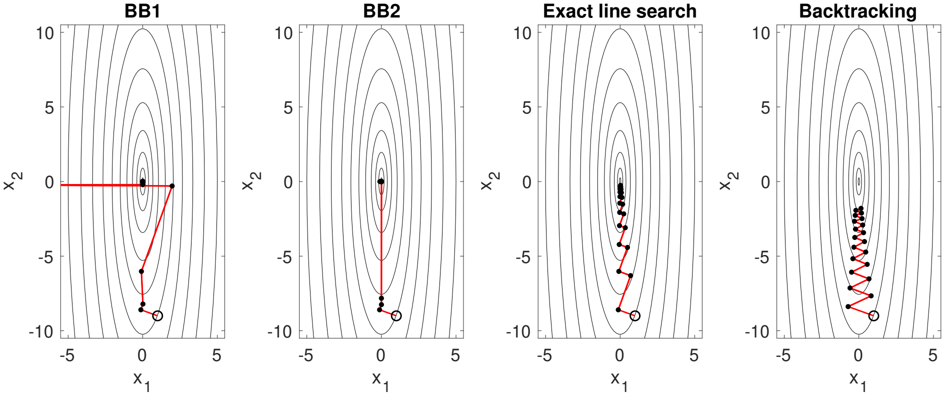
$$t_k^{\mathrm{BB2}} = \operatorname*{argmin}_{\beta} \left\{\|\beta \Delta \boldsymbol{y} - \Delta \boldsymbol{s}\|_2^2\right\} = \frac{\Delta \boldsymbol{s}^T \Delta \boldsymbol{y}}{\|\Delta \boldsymbol{y}\|_2^2}$$

Barzilai–Borwein step size rules (cont.)

- first step size ($k = 0$) must be chosen using another method
- not a descent method ($g(\boldsymbol{x}^{(k+1)}) \leq g(\boldsymbol{x}^{(k)})$ not guaranteed)
- convergence guaranteed if g is strongly convex and quadratic
- safe-guarding is generally required to ensure convergence

Example: least-squares problem

$$t_k^{\mathrm{BB1}} = \frac{\|\nabla g(\boldsymbol{x}^{(k-1)})\|_2^2}{\|\boldsymbol{A}\nabla g(\boldsymbol{x}^{(k-1)})\|_2^2}, \qquad t_k^{\mathrm{BB2}} = \frac{\|\boldsymbol{A}\nabla g(\boldsymbol{x}^{(k-1)})\|_2^2}{\|\boldsymbol{A}^T\boldsymbol{A}\nabla g(\boldsymbol{x}^{(k-1)})\|_2^2}$$



Stopping criteria

Approximate stationarity conditions

$$\|\nabla g(\boldsymbol{x}^{(k)})\|_2 \leq \epsilon, \qquad \|\underbrace{\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}}_{-t_k \nabla g(\boldsymbol{x}^{(k)})}\|_2 \leq \epsilon \|\boldsymbol{x}^{(k)}\|_2$$

not scale invariant; change of variables $\tilde{g}(\boldsymbol{y}) = g(\boldsymbol{C}\boldsymbol{y})$ yields

$$\|\nabla \tilde{g}(\boldsymbol{y}^{(k)})\|_2 = \|\boldsymbol{C}^T \nabla g(\boldsymbol{x}^{(k)})\|_2 \leq \epsilon, \qquad \nabla \tilde{g}(\boldsymbol{y}) = \boldsymbol{C}^T \nabla g(\boldsymbol{C}\boldsymbol{y})$$

Strongly convex objective

$$\|\nabla g(\boldsymbol{x}^{(k)})\|_2 \leq \sqrt{2\mu\epsilon_{\mathrm{obj}}}, \qquad \|\nabla g(\boldsymbol{x}^{(k)})\|_2 \leq \frac{\mu\epsilon_{\mathrm{dist}}}{2},$$

imply that $g(\boldsymbol{x}^{(k)}) - g(\boldsymbol{x}^\star) \leq \epsilon_{\mathrm{obj}}$ and $\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^\star\|_2 \leq \epsilon_{\mathrm{dist}}$

Example: Tikhonov regularized least-squares

$$\text{minimize} \quad g(\boldsymbol{x}) \equiv \tfrac{1}{2}\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \tfrac{\alpha}{2}\|\boldsymbol{x}\|_2^2$$

$\alpha > 0$ is a lower bound on strong convexity parameter (Exercise 13.4)

Stopping criteria

$$\|\nabla g(\boldsymbol{x}^{(k)})\|_2 = \|\alpha\, \boldsymbol{x}^{(k)} - \boldsymbol{A}^T \varrho^{(k)}\|_2 \leq \frac{\alpha\, \epsilon_{\text{dist}}}{2}, \qquad \varrho^{(k)} = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^{(k)}$$

ensures that $\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^\star\|_2 \leq \epsilon_{\text{dist}}$

Exercise 13.6: Step sizes

Apply the gradient method to the problem of minimizing

$$g(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2^2$$

where $\boldsymbol{A}$ and $\boldsymbol{b}$ are generated as follows:

```
>> I0 = 1e6; n = 128;
>> A = paralleltomo(n)*(2/n);
>> x = reshape(phantomgallery('grains',n),[],1);
>> I = poissrnd(I0*exp(-A*x));
>> b = -log(I/I0);
```

Plot (semi-log. y-axis) the obj. value for the first 200 iterations using:

1. Exact line search
2. Backtracking line search
3. BB1 step size
4. BB2 step size

Constrained optimization

$$\text{minimize} \quad g(\boldsymbol{x}) + h(\boldsymbol{x})$$

- $g : \mathbb{R}^n \to \mathbb{R}$ convex and differentiable
- $h : \mathbb{R}^n \to \mathbb{R} \cup \{-\infty, \infty\}$ closed convex
- h not necessarily continuously differentiable but "simple"

Special case

$$\begin{array}{ll} \text{minimize} & g(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in \mathcal{C}, \end{array}$$

corresponds to $h(\boldsymbol{x}) = I_{\mathcal{C}}(\boldsymbol{x}) = \begin{cases} 0, & \boldsymbol{x} \in \mathcal{C} \\ \infty, & \boldsymbol{x} \notin \mathcal{C} \end{cases}$

Convex sets

$\mathcal{C} \subseteq \mathbb{R}^n$ is a convex set if and only if

$$\theta \boldsymbol{x} + (1-\theta)\boldsymbol{y} \in \mathcal{C}, \quad \theta \in [0,1], \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$$

convex set | nonconvex set

Majorization minimization

Suppose ∇g is Lipschitz continuous with constant L

- majorization of $g + h$ at $\boldsymbol{x}$

$$\psi(\boldsymbol{y};\boldsymbol{x}) = g(\boldsymbol{x}) + \nabla g(\boldsymbol{x})^T(\boldsymbol{y}-\boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 + h(\boldsymbol{y})$$

- majorization minimization

$$\begin{aligned}\boldsymbol{x}^{(k+1)} &= \operatorname*{argmin}_{\boldsymbol{y}} \left\{ g(\boldsymbol{x}^{(k)})^T\boldsymbol{y} + h(\boldsymbol{y}) + \frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}^{(k)}\|_2^2 \right\} \\ &= \operatorname*{argmin}_{\boldsymbol{y}} \left\{ h(\boldsymbol{y}) + \frac{L}{2}\|\boldsymbol{y}-(\boldsymbol{x}^{(k)} - (1/L)\nabla g(\boldsymbol{x}^{(k)}))\|_2^2 \right\}\end{aligned}$$

Proximal gradient method

Proximal operator associated with h and $t > 0$

$$\mathrm{prox}_{th}(\boldsymbol{x}) = \underset{\boldsymbol{y}}{\mathrm{argmin}} \left\{ h(\boldsymbol{y}) + \frac{1}{2t} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2 \right\}$$

- easy to evaluate if h is "simple"
- strong convexity implies that $\mathrm{prox}_{th}(\boldsymbol{x})$ is unique

Proximal gradient method

$$\boldsymbol{x}^{(k+1)} = \mathrm{prox}_{th}(\boldsymbol{x}^{(k)} - t\nabla g(\boldsymbol{x}^{(k)})), \quad t = \frac{1}{L}, \quad k = 0, 1, 2, \ldots$$

Examples

- $h(\boldsymbol{x}) = I_{\mathcal{C}}(\boldsymbol{x})$ where $\mathcal{C}$ is a closed, convex set

$$\operatorname{prox}_{th}(\boldsymbol{x}) = P_{\mathcal{C}}(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{y}\in\mathcal{C}}\{\|\boldsymbol{y} - \boldsymbol{x}\|_2^2\}$$

- $h(\boldsymbol{x}) = I_{\mathcal{C}}(\boldsymbol{x})$ with $\mathcal{C} = \{\boldsymbol{x} \mid l_i \le x_i \le u_i,\ i = 1, \ldots, n\}$

$$\operatorname{prox}_{th}(\boldsymbol{x}) = \max(\boldsymbol{l}, \min(\boldsymbol{u}, \boldsymbol{x}))$$

- $h(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$

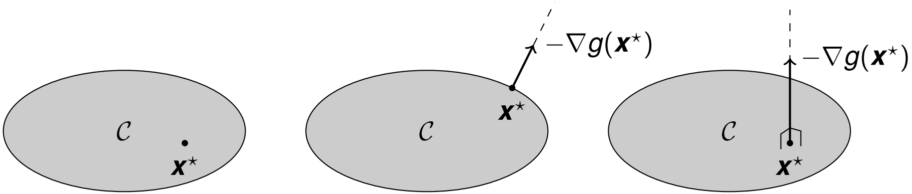
$$\operatorname{prox}_{th}(\boldsymbol{x}) = \operatorname{diag}(\operatorname{sgn}(\boldsymbol{x}))\max(\operatorname{abs}(\boldsymbol{x}) - t\mathbb{1}, \boldsymbol{0})$$

Optimality condition

$x^\star$ is a minimizer of $g + h$ if and only if

$$x^\star = \text{prox}_{th}(x^\star - t\nabla g(x^\star)), \quad t > 0$$

Special case: $h(x) = I_{\mathcal{C}}(x)$ where $\mathcal{C}$ is closed, convex



Example: nonnegativity constraints

$$\begin{array}{ll} \text{minimize} & g(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in \mathcal{C} \end{array}$$

with $\mathcal{C} = \{\boldsymbol{x} \mid x_i \geq 0, i = 1, \ldots, n\}$

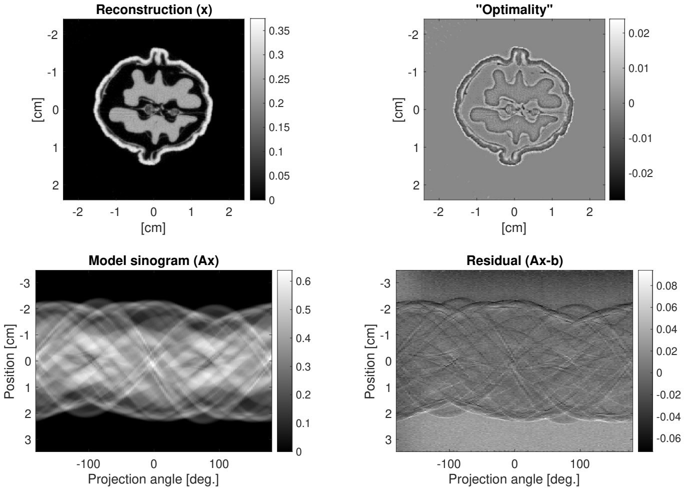
Optimality condition

$$\boldsymbol{x}^\star = \max(\boldsymbol{0}, \boldsymbol{x}^\star - t\nabla g(\boldsymbol{x}^\star)), \quad t > 0$$

or equivalently, for $i = 1, \ldots, n$

$$(x_i^\star = 0 \ \wedge \ [\nabla g(\boldsymbol{x}^\star)]_i \geq 0) \quad \vee \quad (x_i^\star > 0 \ \wedge \ [\nabla g(\boldsymbol{x}^\star)]_i = 0)$$

Example: reconstruction with nonnegativity constraints



Accelerated proximal gradient method

Accelerated proximal gradient method

Require: initial vector $\boldsymbol{x}^{(0)}$, $\boldsymbol{y} = \boldsymbol{x}^{(0)}$, $t_0 = 1$

for $k = 0, 1, 2, \ldots$ **do**

$$\boldsymbol{x}^{(k+1)} = \mathrm{prox}_{(1/L)h}(\boldsymbol{y} - (1/L)\nabla g(\boldsymbol{y}))$$

$$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$$

$$\boldsymbol{y} = \boldsymbol{x}^{(k+1)} + \frac{t_k - 1}{t_{k+1}}(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)})$$

end for

- improved worst-case bound: $f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^\star) = O(1/k^2)$ where $f = g + h$
- not a descent method

Example

$$\text{minimize} \quad g(\boldsymbol{x}) + h(\boldsymbol{x})$$

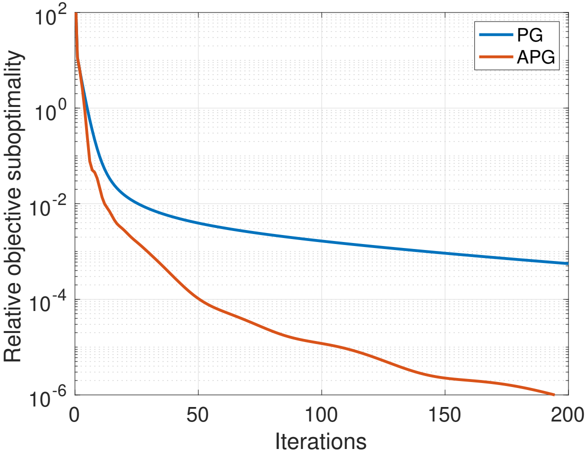
where

$$g(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2^2$$

$$h(\boldsymbol{x}) = \frac{\gamma}{2}\|\boldsymbol{x}\|_2 + I_{\mathcal{C}}(\boldsymbol{x})$$

$$\mathcal{C} = \{\boldsymbol{x} \mid x_i \geq 0,\ i = 1, \ldots, n\}$$

(several ways to "split" objective)



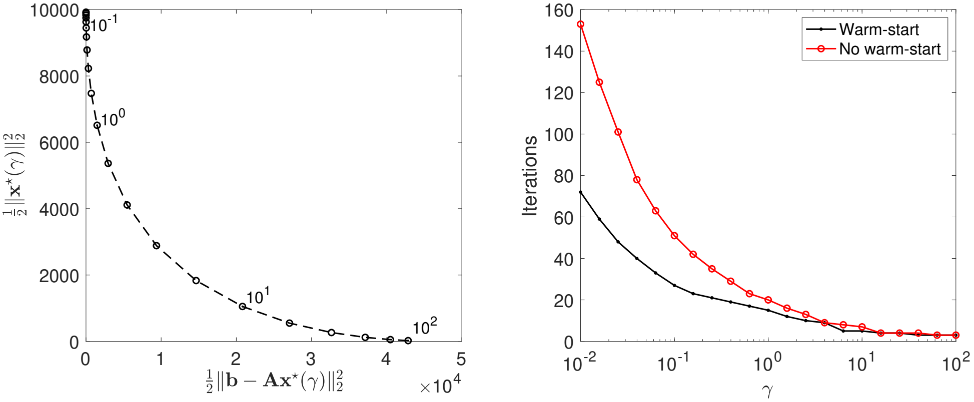
Regularized least-squares

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \alpha\, R(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in \mathcal{C} \end{array}$$

- special cases: Tikhonov, generalized Tikhonov, and TV regularization
- trade-off between two objectives
- often of interest to solve problem with different values of
- trade-off curve (aka L-curve), parameterized by γ

$$\left(\frac{1}{2}\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^\star(\gamma)\|_2^2, R(\boldsymbol{x}^\star(\gamma))\right)$$

Tracing the trade-off curve: Tikhonov regularization



Warm-start: use $\boldsymbol{x}^\star(\gamma)$ as initial guess when solving for $\boldsymbol{x}^\star(\gamma')$

Total variation regularized least-squares

$$\mathrm{TV_a}(\boldsymbol{x}) = \|\boldsymbol{D}\boldsymbol{x}\|_1 = \sum_{i=1}^{2n} |\boldsymbol{d}_i^T \boldsymbol{x}|, \qquad \boldsymbol{D} = \begin{bmatrix} \boldsymbol{I}_N \otimes \boldsymbol{D}_M \\ \boldsymbol{D}_N \otimes \boldsymbol{I}_M \end{bmatrix}$$

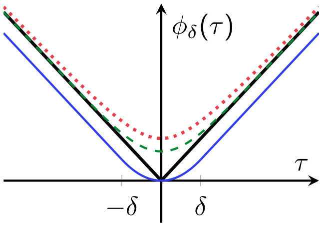
- $\mathrm{TV_a}(\boldsymbol{x})$ is convex but not everywhere differentiable
- $\mathrm{TV_a}(\boldsymbol{x})$ is not “simple” (proximal operator is not cheap to eval.)
- smooth approximation

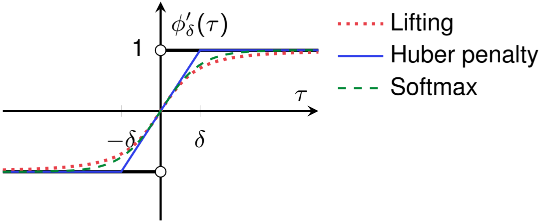
$$\mathrm{TV_a^\delta}(\boldsymbol{x}) = \sum_{i=1}^{2n} \phi_\delta(\boldsymbol{d}_i^T \boldsymbol{x}), \qquad \nabla \mathrm{TV_a^\delta}(\boldsymbol{x}) = \sum_{i=1}^{2n} \boldsymbol{d}_i \nabla \phi_\delta(\boldsymbol{d}_i^T \boldsymbol{x})$$

- more advanced methods exist (splitting methods, etc.)

Smooth approximation to absolute value function

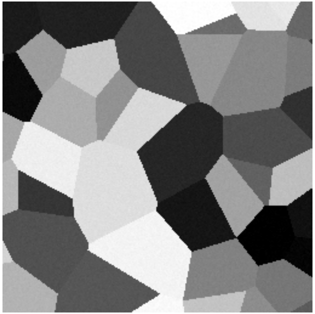
- Lifting: $\phi_\delta(\tau) = \left\| \begin{bmatrix} \tau \\ \delta \end{bmatrix} \right\|_2 = \sqrt{\tau^2 + \delta^2}$
- Huber penalty (scaled): $\phi_\delta(\tau) = \begin{cases} \frac{\tau^2}{2\delta}, & |\tau| \leq \delta \\ |\tau| - \frac{\delta}{2}, & |\tau| > \delta \end{cases}$
- Softmax: $\phi_\delta(\tau) = \delta \log(e^{\tau/\delta} + e^{-\tau/\delta})$

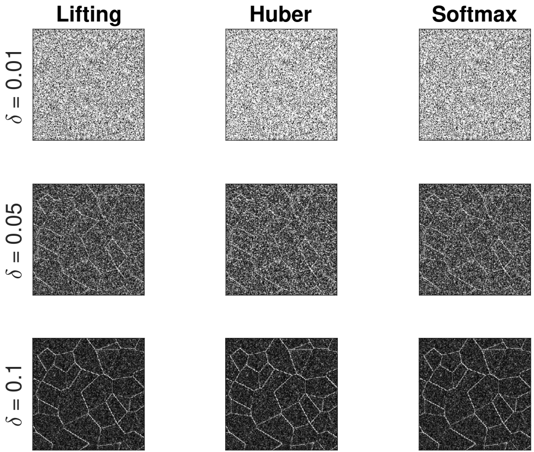




Example: gradient of $\mathrm{TV}_{\mathrm{a}}^{\delta}(\boldsymbol{x})$

```
>> N = 256;
>> X = phantomgallery('grains',N) ...
      + 1e-2*randn(N,N);
```





Extension to isotropic TV

$$\mathrm{TV}_{\mathrm{i}}^{\delta}(\boldsymbol{x}) = \sum_{i=1}^{n} \phi_\delta(\boldsymbol{D}_i \boldsymbol{x}), \qquad \boldsymbol{D}_i = \begin{bmatrix} \boldsymbol{i}_i^T(\boldsymbol{I}_N \otimes \boldsymbol{D}_M) \\ \boldsymbol{i}_i^T(\boldsymbol{D}_N \otimes \boldsymbol{I}_M) \end{bmatrix}$$

$\phi_\delta : \mathbb{R}^2 \to \mathbb{R}$ approximates 2-norm of vector in $\mathbb{R}^2$

Approximation	$\phi_\delta(\boldsymbol{y})$	$\nabla \phi_\delta(\boldsymbol{y})$
Lifting	$\left\\| \begin{bmatrix} \boldsymbol{y} \\ \delta \end{bmatrix} \right\\|_2$	$\left\\| \begin{bmatrix} \boldsymbol{y} \\ \delta \end{bmatrix} \right\\|_2^{-1} \boldsymbol{y}$
Huber	$\begin{cases} \frac{\boldsymbol{y}^T\boldsymbol{y}}{2\delta}, & \\|\boldsymbol{y}\\|_2 \leq \delta \\ \\|\boldsymbol{y}\\|_2 - \frac{\delta}{2}, & \\|\boldsymbol{y}\\|_2 > \delta \end{cases}$	$\frac{1}{\max(\delta, \\|\boldsymbol{y}\\|_2)} \boldsymbol{y}$
Softmax	$\delta \log(e^{\\|\boldsymbol{y}\\|_2/\delta} + e^{-\\|\boldsymbol{y}\\|_2/\delta})$	$\begin{cases} \frac{\tanh(\\|\boldsymbol{y}\\|_2/\delta)}{\\|\boldsymbol{y}\\|_2} \boldsymbol{y}, & \boldsymbol{y} \neq \boldsymbol{0} \\ \boldsymbol{0}, & \boldsymbol{y} = \boldsymbol{0} \end{cases}$

Exercise 13.7: Smooth approximation of total variation penalty

Show that the three smooth approximations of the absolute value function all have a Lipschitz continuous derivative with Lipschitz constant $L = 1/\delta$.

- Lifting: $\phi_\delta(\tau) = \left\| \begin{bmatrix} \tau \\ \delta \end{bmatrix} \right\|_2 = \sqrt{\tau^2 + \delta^2}$
- Huber penalty (scaled): $\phi_\delta(\tau) = \begin{cases} \frac{\tau^2}{2\delta}, & |\tau| \leq \delta \\ |\tau| - \frac{\delta}{2}, & |\tau| > \delta \end{cases}$
- Softmax: $\phi_\delta(\tau) = \delta \log(e^{\tau/\delta} + e^{-\tau/\delta})$

Exercise 13.8: Regularized weighted least-squares problems

Consider the following weighted least-squares problems with two different regularization terms: (i) generalized Tikhonov regularization

$$\boldsymbol{x}_{\mathrm{GTik}} = \operatorname*{argmin}_{\boldsymbol{x}} \left\{ \frac{1}{2}\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_{\boldsymbol{W}}^2 + \frac{\gamma}{2}\|\boldsymbol{D}\boldsymbol{x}\|_2^2 \right\} \qquad (1)$$

and (ii) total variation regularization

$$\boldsymbol{x}_{\mathrm{TV}} = \operatorname*{argmin}_{\boldsymbol{x}} \left\{ \frac{1}{2}\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_{\boldsymbol{W}}^2 + \gamma\|\boldsymbol{D}\boldsymbol{x}\|_1 \right\}. \qquad (2)$$

The variable $\boldsymbol{x} \in \mathbb{R}^n$ represents an image of size $N \times N$ (i.e., $n = N^2$).
(Refer to textbook for questions.)