References

Convex Optimization , Chapter 9, S. Boyd & L. Vandenberghe, Cambridge University Press 2004.

Problem: Unconstrained minimization I

Consider the function $f : \mathbf{x} \in \mathbb{R}^d \rightarrow f(\mathbf{x}) \in \mathbb{R}$. We want to find $\hat{\mathbf{x}}$ such that:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} f(\mathbf{x})$$

Assuming f is differentiable (and convex), a necessary (and sufficient) condition for $\hat{\mathbf{x}}$ to be optimal is:

$$\nabla f(\hat{\mathbf{x}}) = 0$$

Sometimes this minimum cannot be found analytically, so instead we aim at designing an algorithm that generates a sequence of points $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(k)}$ that converges towards $\hat{\mathbf{x}}$.

Problem: Unconstrained minimization II

Examples:

- Quadratic minimization can be solve analytically:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \left\{ f(\mathbf{x}) = \|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2 \right\}$$

- Unconstrained Geometric Programming:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \left\{ f(\mathbf{x}) = \log\left(\sum_{i=1}^{m} \exp(\mathbf{a}_i^T \mathbf{x} + b_i) \right) \right\}$$

The optimal condition is:

$$\nabla f(\hat{\mathbf{x}}) = \frac{1}{\sum_{i=1}^{m} \exp(\mathbf{a}_i^T \mathbf{x} + b_i)} \sum_{i=1}^{m} \exp(\mathbf{a}_i^T \mathbf{x} + b_i)\, \mathbf{a}_i$$

Descent methods

Theorem (General descent method)

Given *a starting point* $\mathbf{x}^{(0)}$

Repeat

1. *Determine a descent direction* $\Delta\mathbf{x}^{(k)}$
2. *Line search: Choose a step size* $t^{(k)} > 0$
3. *Update:* $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t^{(k)}\ \Delta\mathbf{x}^{(k)}$

Until *stopping criterion is satisfied.*

Choice of a starting point $\mathbf{x}^{(0)}$

Choice of the step size $t^{(k)}$ I

When updating the current position:

- small steps: inefficient
- large step: potentially bad results

1. Exact line search: t is chosen to minimize f along the ray $\{\mathbf{x} + t\Delta\mathbf{x} | t \geq 0\}$ or:

$$t = \arg\min_{s \geq 0} f(\mathbf{x} + s\Delta\mathbf{x})$$

Exercise: Show that $\nabla f(\mathbf{x}^{(k)} + t\Delta\mathbf{x}^{(k)})^T \Delta\mathbf{x}^{(k)} = 0$ when t is fixed by exact line search.

Choice of the step size $t^{(k)}$ II

Remark: most of the time you need numerical methods to find the root of $\phi'(s)$ with

$$\phi(s) = f(\mathbf{x} + s\Delta\mathbf{x})$$

1. Dichotomous and Golden search
2. Bisection
3. Newton's method

Choice of the step size $t^{(k)}$ III

2. Backtracking line search: Since most line search are inexact in practice, the step length is usually chosen to approximatly minimize f or reduce f enough. Backtracking line search is one of them:

Theorem (Backtracking line search)

Given a descent direction $\Delta\mathbf{x}$ for f at $\mathbf{x}$, and 2 constants $\alpha \in [0; 0.5]$ $\beta \in]0; 1[$,
Given $t := 1$
while $f(\mathbf{x} + t\Delta\mathbf{x}) > f(\mathbf{x}) + \alpha t \nabla f(\mathbf{x})^T \Delta\mathbf{x}$, $t := \beta t$

Choice of a descent direction $\Delta\mathbf{x}$ I

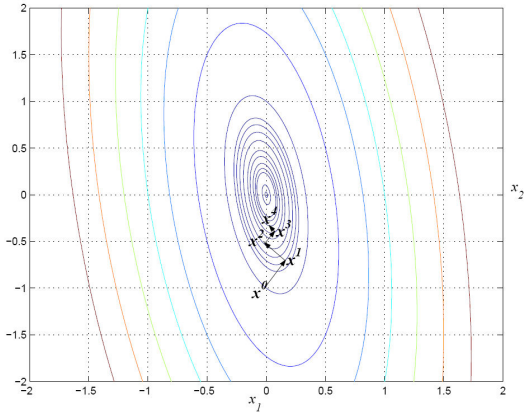
A natural choice for the search direction is the negative gradient $\Delta\mathbf{x} = -\nabla f(\mathbf{x})$, and the resulting algorithm is then called gradient descent algorithm.

Definition (Gradient)

With $\mathbf{x} \in \mathbb{R}^d$ or $\mathbf{x} = (x_1, x_2, \cdots, x_d)^T$, the gradient is defined by:

$$\nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_d} \end{pmatrix}$$

Choice of a descent direction $\Delta\mathbf{x}$ II



Choice stopping criterion is satisfied

$$\|\nabla f(x^{(k+1)})\| < \epsilon$$

Newton's method I

Definition (Hessian matrix)

$$H_f(\mathbf{x}) = \nabla^2 f(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \vdots & & \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_d \partial x_d} \end{pmatrix}$$

Theorem (Newton's method)

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - [\nabla^2 f(\mathbf{x}^{(k)})]^{-1} \nabla f(\mathbf{x}^{(k)})$$

or rewritten using the Hessian matrix

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - [H_f(\mathbf{x}^{(k)})]^{-1} \nabla f(\mathbf{x}^{(k)})$$

Newton's method II

Exercise: Prove the Newton method.

Alternatives:

- Quasi-Newton methods
- ...

Applications

Finding minimum or maximum as many applications: we cite here just one the Mean-shift approach.