Chapter 4: Unconstrained Nonlinear Optimization

Edoardo Amaldi

DEIB – Politecnico di Milano
edoardo.amaldi@polimi.it

Website: http://home.deib.polimi.it/amaldi/OPT-18-19.shtml

Academic year 2018-19

4.1 Examples

1) Statistical estimation turns out to be a nonlinear optimization problem

Random variable X with density $f(x, \underline{\theta})$, where $\underline{\theta} \in \mathbb{R}^m$ is parameter vector, and independent observations $x_1, \ldots, x_n$.

example: for Gaussian, \mu and \Sigma

Maximum likelihood: Estimates $\hat{\underline{\theta}}$ of $\underline{\theta}$ are derived by maximizing

$$L(\underline{\theta}) = f(x_1, \underline{\theta})\ f(x_2, \underline{\theta}) \ldots f(x_n, \underline{\theta})$$

Likelihood: product of densities evaluated in all the observations. NOTE: densities can be nonlinear (ex. Gaussian), thus L is nonlinear —> it is better to maximize a sum rather than a product, thus we consider the logarithm

Assumption: $\exists\, \underline{\theta}$ for which all factors are positive.

Since ln(.) is monotonically increasing, $\hat{\underline{\theta}}$ also maximizes

$$\ln(L(\underline{\theta})) = \sum_{j=1}^{n} \ln(f(x_j, \underline{\theta}))$$

If f is differentiable w.r.t. $\underline{\theta}$ at $\hat{\underline{\theta}}$, necessary optimality conditions:

$$\sum_{j=1}^{n} \frac{\nabla_{\underline{\theta}} f(x_j, \hat{\underline{\theta}})}{f(x_j, \hat{\underline{\theta}})} = \underline{0}$$

For Guassian density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp - \frac{(x-\mu)^2}{2\sigma^2}$$

and $\underline{\theta} = (\mu, \sigma)$, we obtain

$$\ln(L(\underline{\theta})) = \ln((\frac{1}{\sigma\sqrt{2\pi}})^n \prod_{j=1}^{n} \exp - \frac{(x_j-\mu)^2}{2\sigma^2}) = -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2$$

Minimum is achieved in a stationary point:

$$\frac{\partial[\ln(L(\underline{\theta}))]}{\partial\mu} = \frac{1}{\sigma^2}\sum_{j=1}^{n}(x_j-\mu) = 0$$

and

$$\frac{\partial[\ln(L(\underline{\theta}))]}{\partial\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{j=1}^{n}(x_j-\mu)^2 = 0$$

Thus

$$\hat{\mu} = \frac{1}{n}\sum_{j=1}^{n} x_j \qquad\qquad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_j-\hat{\mu})^2}$$

2) Training multilayer neural networks turns out to be a nonlinear optimization problem

Supervised learning:

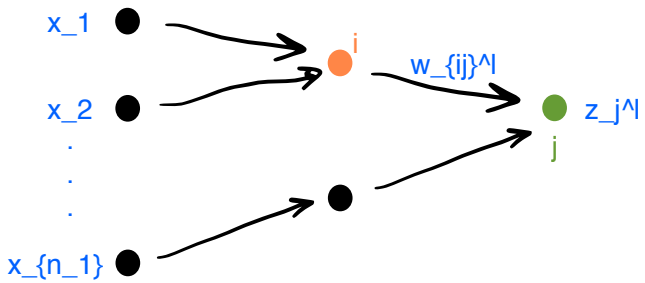
Given a training set $T = \{(\underline{x}^1, \underline{y}^1), \ldots, (\underline{x}^p, \underline{y}^p)\}$ where $\underline{y}^k \in [0,1]^{n_{out}}$ desiderd output for $\underline{x}^k \in \mathbb{R}^{n_{in}}$, construct a model that maps $\underline{x}^k$'s into $\underline{y}^k$'s as well as possible.

Multilayer networks: (one possible ML model to perform supervised learning)

L layers with n_l units in layer l, $n_1 = n_{in}$ and $n_L = n_{out}$.

First layer of inputs $x_1, \ldots, x_{n_1}$, other layers with activation units.

Illustration:



Output of unit j of layer l:

$$z_j^l = \phi\left(\sum_{i=1}^{n_{l-1}} w_{ij}^l z_i^{l-1} - w_{0j}^l\right)$$

where weights w_{ij} are to be determined and $\phi : \mathbb{R} \to \mathbb{R}$ is sigmoid $\phi(t) = \frac{1}{1+e^{-t}}$.

Multilayer network defnines a mapping $h(\underline{w}, .)$ from $\mathbb{R}^{n_1}$ to $\mathbb{R}^{n_L}$ parametrized by $\underline{w} = \{w_{ij}^l : l = 1, \ldots, L; i = 1, \ldots, n_{l-1}; j = 1, \ldots, n_l\}$.

NOTE: the topology of the network (# of layers and # of unit per layer) should be designed a-priori

Training problem: Given $T = \{(\underline{x}^1, \underline{y}^1), \ldots, (\underline{x}^p, \underline{y}^p)\}$, select appropriate values of $\underline{w}$ to approximate as well as possible the mapping underlying T.

In general one minimizes sum of squared errors

$$\frac{1}{2} \sum_{k=1}^{p} (\|\underline{y}^k - h(\underline{w}, \underline{x}^k)\|)^2$$

Quite challenging, typically nonconvex with multiple local minima.

Illustration

4.2 Optimality conditions

Generic optimization problem:

$$\min_{\underline{x} \in S} \; f(\underline{x})$$

where $S \subseteq \mathbb{R}^n$, $f: \; S \to \mathbb{R}$ and $f \in \mathcal{C}^1$ or $\mathcal{C}^2$.

it is not a completely unconstrained case, actually, since we consider a subset…but we consider only very simple constraints (ex. nonnegativity)

Unconstrained case: $S = \mathbb{R}^n$

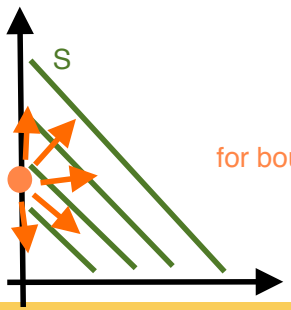
Extension of the necessary and sufficient optimality conditions (1st and 2nd order).

I can move along d while remaining in the feasible region S

Definition: $\underline{d} \in \mathbb{R}^n$ is a *feasible direction* at $\underline{\overline{x}}$ if

$$\exists \; \overline{\alpha} > 0 \;\; \text{such that } \underline{\overline{x}} + \alpha \underline{d} \in S \qquad \forall \alpha \in [0, \overline{\alpha}] \tag{1}$$

Illustrations:



for boundary points, the feasible directions draw a cone

At any interior point all directions (all $\underline{d} \in \mathbb{R}^n$) are feasible.

REMEMBER: we are always MINIMIZING the objective function!

First order necessary local optimality conditions:

If $f \in \mathcal{C}^1$ on S and $\overline{x}$ is a *local minimum* of f over S, then for any feasible direction $\underline{d} \in \mathbb{R}^n$ at $\overline{x}$

$$\nabla^t f(\overline{x})\underline{d} \geq 0, \quad (@)$$

namely all feasible directions are ascent directions.

Proof:

Let \bar{x} be a point s.t. there exists a feasible direction d and the angle between d and -\nabla f(x) is less than 90 degrees. Then, by moving along d, I decrease the value of the objective function (i.e. the value is better) —> if this is the case, \bar{x} cannot be a local minimum. It must be -\nabla f(x)^T d \leq 0, thus (@)

According to (1), consider

I consider the function f only along the feasible direction d

$$\phi\colon [0, \overline{\alpha}] \to \mathbb{R} \quad \text{such that} \quad \phi(\alpha) = f(\overline{x} + \alpha\underline{d})$$

Since $\overline{x}$ is a local minimum of f over S, $\alpha = 0$ is a local minimum of $\phi(\alpha)$.

Taylor series of ϕ at $\alpha = 0$

$$\phi(\alpha) = \phi(0) + \alpha\phi'(0) + o(\alpha)$$

N.B.: $u(\alpha) = o(\alpha)$ if $u(\alpha)$ tends to 0 faster than α when $\alpha \to 0$.

\phi(\alpha) = \phi(0) + \alpha \phi'(0) —> \phi(\alpha) - \phi(0) = \alpha \phi'(0) < 0 because \alpha > 0

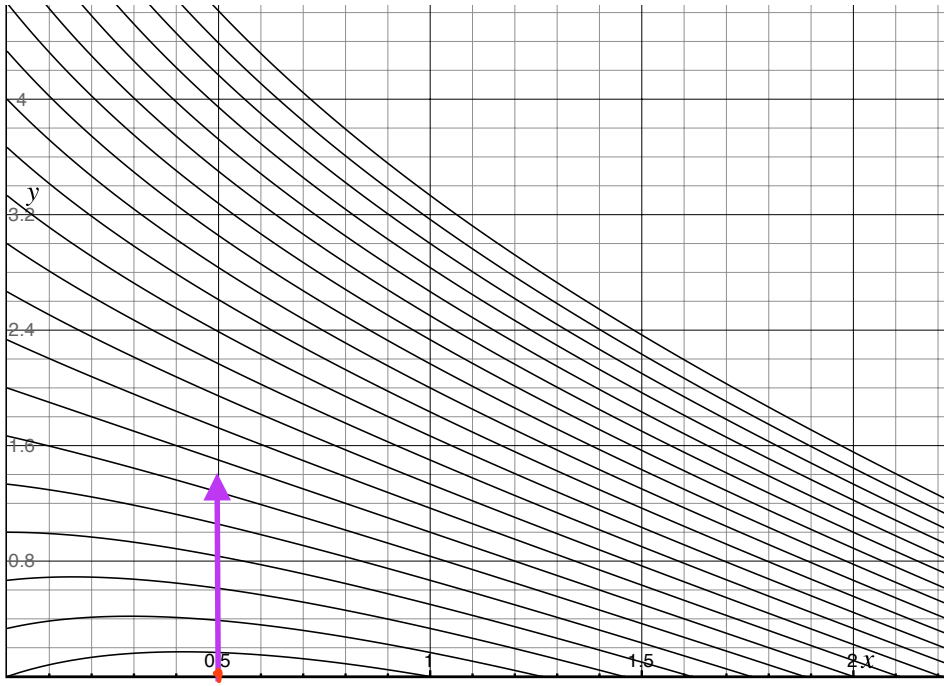
Suppose $\phi'(0) < 0$: if $\alpha \to 0_+$ we can neglect $o(\alpha)$ and we have $\phi(\alpha) - \phi(0) < 0$, then 0 would not be a local minimum.

because \phi(\alpha) - \phi(0) < 0 means that there exists \alpha s.t. the value of \phi at \alpha is less than the value of \phi at 0

Therefore $\phi'(0) \geq 0$ and, since $\phi'(\alpha) = \nabla^t f(\overline{x} + \alpha\underline{d})\underline{d}$, we have $\nabla^t f(\overline{x})\underline{d} \geq 0$.

□

Example: $\min_{x_1,\ x_2 \geq 0}\ f(x_1, x_2) = x_1^2 - x_1 + x_2 + x_1 x_2$



$\underline{x}^* = (\frac{1}{2}\ 0)^t$ is a global minimum because $\nabla^t f(\underline{x}^*)\underline{d} \geq \underline{0}$ for all feasible directions $\underline{d}$ in $\underline{x}^*$ (all those with $d_2 \geq 0$), even if $\nabla^t f(\underline{x}^*) = (0\ \frac{3}{2}) \neq \underline{0}$.

Second order necessary local optimality conditions:

we need to be able to compute the Hessian

If $f \in \mathcal{C}^2$ on S and $\overline{x}$ is a *local minimum* of f over S then

i) $\nabla^t f(\overline{x})\underline{d} \geq 0$ for every $\underline{d} \in \mathbb{R}^n$ feasible direction at $\overline{x}$, (same as 1^{st} order)

ii) if $\nabla^t f(\overline{x})\underline{d} = 0$ then $\underline{d}^t \nabla^2 f(\overline{x})\underline{d} \geq 0$. (this does not mean that the Hessian is positive semidefinite, since we are not considering all possible d, but only the ones orthogonal to the gradient)

we will have as condition "positive semidef. Hessian" in the unconstrained case

Proof:

Similarly for (ii).

Suppose $\nabla^t f(\overline{x})\underline{d} = 0$, then

see slide 7

$$\phi(\alpha) = \phi(0) + \alpha \underbrace{\phi'(0)}_{0} + \frac{1}{2}\alpha^2 \phi''(0) + o(\alpha^2).$$

If $\phi''(0) = \underline{d}^t \nabla^2 f(\overline{x})\underline{d} < 0$, for sufficiently small values of α we would have

$$\phi(\alpha) - \phi(0) \leq \frac{1}{2}\alpha^2 \phi''(0) < 0,$$

namely 0 would not be a local minimum of $\phi(\alpha)$.

Hence $\phi''(0) = \underline{d}^t \nabla^2 f(\overline{x})\underline{d} \geq 0$.

□

Corollary: (Unconstrained case)

If $f \in \mathcal{C}^2$ on S and $\overline{x} \in int(S)$ is a *local minimum* of F over S, then

1. $\nabla f(\overline{x}) = \underline{0}$ (stationarity condition)
2. $\nabla^2 f(\overline{x})$ is positive semidefinite.

unconstrained:

S = \mathbb{R}^n

or

we are considering an interior point

Proof:

Since $\overline{x} \in int(S)$, all $\underline{d} \in \mathbb{R}^n$ are feasible directions at $\overline{x}$.

Then $\nabla^t f(\overline{x})\underline{d} \geq 0$ for every $\underline{d}$ and $-\underline{d}$ imply (1).

(2) is an immediate consequence of $\underline{d}^t \nabla^2 f(\overline{x})\underline{d} \geq 0$ for all $\underline{d} \in \mathbb{R}^n$.

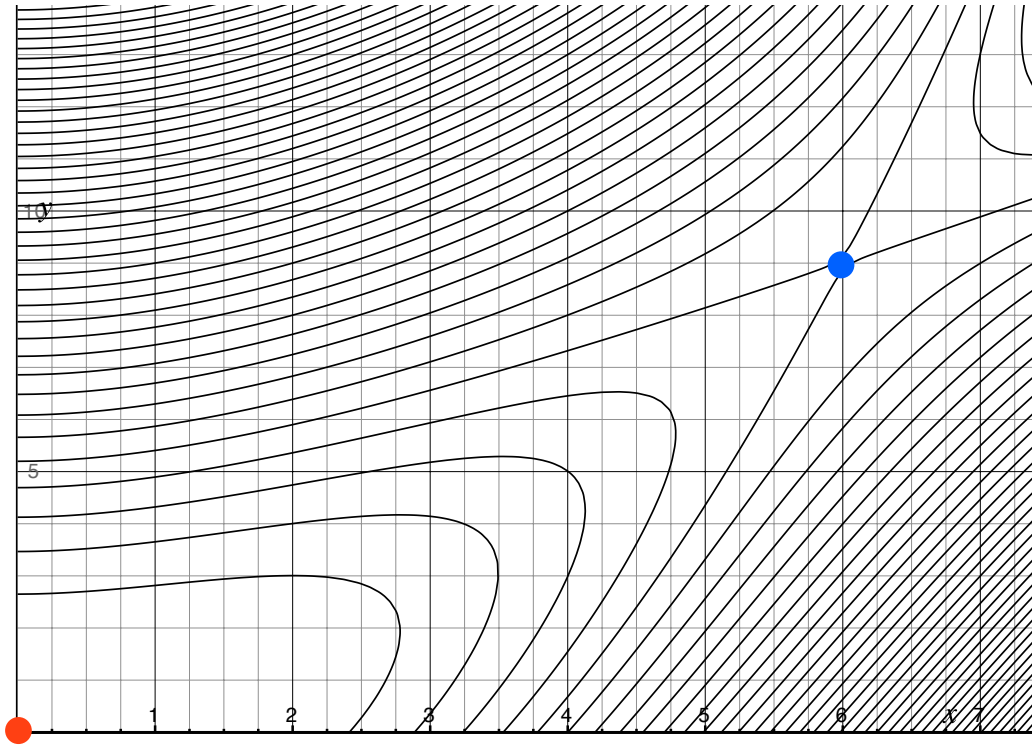
□

Types of candidate points: local minima, local maxima and saddle points.

These optimality conditions are not sufficient.

E.g. $f(x) = x^3$ with $f'(0) = 0$ and $f''(0) = 0$ but $x = 0$ is not a local minimum.

Example: $min_{x_1,\ x_2 \geq 0}\ f(x_1, x_2) = x_1^3 - x_1^2 x_2 + 2x_2^2$



Candidate points: (0 0) and (6 9). (0 0) is on the boundary and (6 9) is not a local minimum even though, for $x_1 = 6$, $x_2 = 9$ it is a local minimum w.r.t. x_2 and, for $x_2 = 9$, $x_1 = 6$ it is a local minimum w.r.t. x_1.

NOTE: the Hessian matrix of this function at (6,9) is not positive definite

Sufficient local optimality conditions:

I need to strengthen the condition (ii) of 2^{nd} order necessary cond. iot get a sufficient cond.

If $f \in \mathcal{C}^2$ on S and $\overline{x} \in int(S)$ such that $\nabla f(\overline{x}) = \underline{0}$ and $\nabla^2 f(\overline{x})$ is positive definite, then $\overline{x}$ is a *strict local minimum* of f over S, namely

$$f(\underline{x}) > f(\overline{x}) \qquad \forall \underline{x} \in \mathcal{N}_\epsilon(\overline{x}) \cap S.$$

Proof:

Let $\underline{d} \in \mathcal{B}_\epsilon(\underline{0})$ be any feasible direction such that $\overline{x} + \underline{d} \in S \cap \mathcal{B}_\epsilon(\overline{x})$.

Then

$$f(\overline{x} + \underline{d}) = f(\overline{x}) + \nabla^t f(\overline{x})\underline{d} + \frac{1}{2}\underline{d}^t \nabla^2 f(\overline{x})\underline{d} + o(\|\underline{d}\|^2)$$

with $\nabla f(\overline{x}) = \underline{0}$.

Since $\nabla^2 f(\overline{x})$ is positive definite, $\exists a > 0$ such that $\underline{d}^t \nabla^2 f(\overline{x})\underline{d} \geq a\|\underline{d}\|^2$ where a is a smallest eigenvalue of $\nabla^2 f(\overline{x})$.

Thus for $\|\underline{d}\|$ sufficiently small

$$f(\overline{x} + \underline{d}) - f(\overline{x}) \geq \frac{a}{2}\|\underline{d}\|^2 > 0$$

which implies $f(\overline{x} + \underline{d}) > f(\overline{x})$, namely $\overline{x}$ is a strict local minimum along $\underline{d}$.

□

Since this holds $\forall \underline{d} \in \mathbb{R}^n$ such that $\overline{x} + \underline{d} \in S \cap \mathcal{B}_\epsilon(\overline{x})$, f is locally strictly convex.

Convex problems

$$\min_{x \in C \subseteq \mathbb{R}^n} f(\underline{x}) \qquad \text{where } C \subseteq \mathbb{R}^n \text{ convex and } f: C \to \mathbb{R} \text{ convex}$$

We know: if f convex, every local minimum is a global minimum.

Necessary and sufficient (NS) conditions for global optimality:

Let f be convex of class $\mathcal{C}^1$ on $C \subseteq \mathbb{R}^n$ convex. $\underline{x}^*$ is a *global minimum* of f on C if and only if

$$\nabla^t f(\underline{x}^*)(\underline{y} - \underline{x}^*) \geq 0 \qquad \forall \underline{y} \in C.$$

Proof:

NC: if $f \in \mathcal{C}^1$ and $\underline{x}^*$ is a local minimum (also global minimum due to convexity) then $\nabla^t f(\underline{x}^*)\underline{d} \geq 0 \quad \forall \underline{d}$ feasible directions at $\underline{x}^*$, namely $\forall \underline{d} = \underline{y} - \underline{x}^*$ with $\underline{y} \in C$.

by 1st order necessary condition

any feasible direction can be expressed in this way

SC: f is convex if and only if

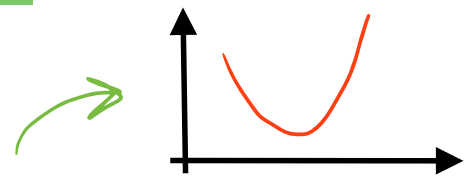
$$f(\underline{y}) \geq f(\underline{x}^*) + \nabla^t f(\underline{x}^*)(\underline{y} - \underline{x}^*) \qquad \forall \underline{y} \in C.$$

characterization of convex functions involving supporting hyperplanes: a convex function lives always above its linear approximation

Then $\nabla f(\underline{x}^*)(\underline{y} - \underline{x}^*) \geq 0$ implies that $f(\underline{y}) \geq f(\underline{x}^*)$ for every $\underline{y} \in C$. □

i.e. x* is a global minimum

Recall: Let $C \subseteq R^n$ be convex. Then $\underline{x} \in C$ is an *extreme point* of C if it cannot be expressed as a convex combination of two different points of C.



Property: (maximization of convex functions)

Let f be a convex function defined on a convex bounded closed set C. If f has a (finite) *maximum* over C, then $\exists$ an *optimal extreme point* of C.

Proof:

we are somehow generalising what we said in the LP case, where we had linear functions and we found that the optimal solution is an extreme point —> also in this case, we only have to focus on the boundary of C: on its extreme points

Suppose that $\underline{x}^*$ is a global maximum of f over C, but not an extreme point.

Then I can express x* as convex combination of two points in C…for example of two points on the boundary (y1 and y2), that exist for sure because C is bounded.

1) Verify that the maximum is achieved at a point on the boundary ∂C.

Since C is convex bounded and closed, for any $\underline{x}^* \in int(C)$ there exist $\underline{y}_1, \underline{y}_2 \in \partial C$ and $\alpha \in [0,1]$ such that $\underline{x}^* = \alpha \underline{y}_1 + (1-\alpha)\underline{y}_2$.

Due to convexity of f, we have

$$f(\underline{x}^*) \leq \alpha f(\underline{y}_1) + (1-\alpha) f(\underline{y}_2) \leq \min\{f(\underline{y}_1), f(\underline{y}_2)\}.$$

Thus also $\underline{y}_1$ and $\underline{y}_2$ are global maxima.

2) Suppose $\underline{x}^* \in \partial C$ is not an extreme point.

Consider $T_1 = C \cap H$, where H is a supporting hyperplane at $\underline{x}^* \in \partial C$.

Clearly $dim(T_1) \leq n - 1$.

Since T_1 is compact, $\exists$ a global optimum $\underline{x}_1$ of f over T_1 such that

$$\max_{\underline{x} \in T_1} f(\underline{x}) = f(\underline{x}_1) = f(\underline{x}^*)$$

and, as previously, we have $\underline{x}_1 \in \partial T_1$.

Claim: If $\underline{x}_1$ is an extreme point of T_1, $\underline{x}_1$ is also an extreme point of C.

If $\underline{x}_1$ is not an extreme point of T_1, we similarly define $T_2, \ldots$

In the worst case $dim(T_n) = 0$. Such an isolated $\underline{x}_n$ is clearly an extreme point. Since an extreme point of T_i is also an extreme point of T_{i-1}, $\underline{x}_n$ must be an extreme point of C.

□

Illustrations:

Special case: Linear programming (convex & concave, finite number of extreme points)