OSM Boot Camp **Math Notes**

Harrison Beard

Summer 2018

Mon, 23 Jul. 2018

- **Topic**. Nonlinear Optimization.
  –

$$\cancel{\begin{array}{ll}\max\limits_{\mathbf{x}} & \mathbf{c}^{\mathsf{T}}\mathbf{x} \\ \text{s.t.} & \mathbf{A}\mathbf{x} \preceq \mathbf{b} \\ & \mathbf{x} \succeq \mathbf{0}\end{array}}$$

$$\Downarrow$$

$$\text{nonlinear } \min_{\mathbf{x}} f : \mathbb{R}^n \to \mathbb{R}$$

Start with a guess $\mathbf{x}_0$. This yields through the algorithm an $\mathbf{x}_1$, then $\mathbf{x}_1 \mapsto \mathbf{x}_2$, and $\mathbf{x}_2 \mapsto \mathbf{x}_3$, and so forth. Eventually we get convergence. This is all according to the rule

$$\mathbf{x}_{i+1} = f(\mathbf{x}_i).$$

Typically, f does one of two things: It could move in a direction that decreases the objective function (**descent function**) or it could approximate the objective function near $\mathbf{x}_i$ with some simpler function, and then that function itself is then optimized (**local approximation methods**).

- **Topic**. Convergence.
  – What does it look like?

(*i*). $\|\mathbf{x}_{i+1} - \mathbf{x}_i\| < \varepsilon$

(*ii*). $\frac{\|\mathbf{x}_{i+1} - \mathbf{x}_i\|}{\|\mathbf{x}_i\|} < \varepsilon$

(*iii*). $\|\mathbf{D}f(\mathbf{x}_i)\| < \varepsilon$, by the FONC

(*iv*). $|f(\mathbf{x}_{i+1}) - f(\mathbf{x}_i)| < \varepsilon$

  – Quadratic Optimization: f is minimized where g is minimized where

$$g(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathsf{T}}\mathbf{Q}\mathbf{x} - \mathbf{b}^{\mathsf{T}}\mathbf{x} + c,$$

where $\mathbf{Q} = \mathbf{A}^{\mathsf{T}} + \mathbf{A}$. A minimizer exists only if $\mathbf{Q} > 0$. The minimizer is the solution to $\mathbf{Q}\mathbf{x} = \mathbf{b}$, and $\mathbf{0} = \mathbf{D}g(\mathbf{x}) = \mathbf{Q}\mathbf{x} - \mathbf{b}$.

  – In general, we find a solution to the linear system of equations of n equations with n unknowns:

(*i*). LU-Decomposition

(*ii*). QR-Decomposition

(*iii*). Cholesky

  – All the above algorithms are $\mathcal{O}(n^3)$ in time.

- **Topic**. Standard Least Squares.

- For $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{A} \in \mathrm{M}_{m\times n}(\mathbb{R})$, the problem of finding an $\mathbf{x}^* \in \mathbb{R}^n$ to minimize $\|\mathbf{A}\mathbf{x} = \mathbf{b}\|_2$ is the same as minimizing

$$\mathbf{x}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x} - 2\mathbf{A}\mathbf{x}\mathbf{b}.$$

Note

$$\langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle = \mathbf{x}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x} - 2\mathbf{A}\mathbf{x}\mathbf{b} + \mathbf{b}^\mathsf{T}\mathbf{b}.$$

We also have

$$\mathbf{A}^\mathsf{T}\mathbf{A} \ = \mathbf{A}^\mathsf{T}\mathbf{b}.$$

The solution is the same as minimizing $g(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x} - 2\mathbf{A}\mathbf{x}\mathbf{b}$:

$$\begin{aligned} \mathbf{0} &= \mathbf{D}g(\mathbf{x}) \\ &= \mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{x} \\ &= \mathbf{A}^\mathsf{T}\mathbf{b}. \end{aligned}$$

- **Topic**. Gradient Descent.
  - Move in the direction of $-\mathbf{D}f^\mathsf{T}(\mathbf{x}_i)$, the direction of steepest descent. The new approximation:

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \alpha\mathbf{D}f^\mathsf{T}(\mathbf{x}_i),$$

for some value of α. To choose α_i, choose

$$\alpha_i^* = \arg\min_{\alpha_i} f\left(\mathbf{x}_i - \alpha\mathbf{D}f^\mathsf{T}(\mathbf{x}_i)\right),$$

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \alpha_i^*\mathbf{D}f^\mathsf{T}(\mathbf{x}_i).$$

This policy of proceeding down the surface is called **steepest descent.**

- **Topic**. Newton's Method: multivariate version. Note that the Hessian $\mathbf{D}^2 f$ has to be positive definite.

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \left(\mathbf{D}^2 f(\mathbf{x}_i)\right)^{-1}\mathbf{D}f^\mathsf{T}(\mathbf{x}_i).$$

Converges quadratically.
  - Problems with Newton:

(*i*). If $\mathbf{x}_0$ is too far from $\mathbf{x}^*$.

(*ii*). When $\mathbf{D}^2 f(\mathbf{x}_i)$ is not positive definite $\left(\mathbf{D}^2 f(\mathbf{x}_i) \not> 0\right)$.

(*iii*). When $\left(\mathbf{D}^2 f(\mathbf{x}_i)\right)^{-1}\mathbf{D}f^\mathsf{T}(\mathbf{x}_i)$ is too expensive to compute or unstable, or impossible.

Wed, 25 Jul. 18

Conjugate Gradient Methods.

- Different than Quasi Newton Method in that they don't store the $n \times n$ Hess (or approximations)
- Most useful when obj. fn. is of form:
$$\frac{1}{2}\mathbf{x}^\mathsf{T}\mathbf{Q}\mathbf{x} - \mathbf{b}^\mathsf{T}\mathbf{x} + c,$$
where $\mathbf{Q}$ is symmetric, $\mathbf{Q} > 0$, and $\mathbf{Q}$ is sparse (most of the entries are zero).
- Each step of Conj. Grad. has temporal and spatial complexity $\mathcal{O}(m)$, where m is the number of nonzero entries.

Nonlinear Least Squares.

- Of the form
$$f = \mathbf{r}^\mathsf{T}\mathbf{r}.$$

1. If the dimension is not too big:
   (a) if $\mathbf{x}_0$ is close to $\mathbf{x}^*$:
      i. If computing $\left(\mathbf{D}^2 f(\mathbf{x})\right)^{-1}\mathbf{D}f^\mathsf{T}(\mathbf{x})$ is cheap and feasible, then use Newton's.
      ii. Else,
         - If $f = \mathbf{r}^\mathsf{T}\mathbf{r}$, use Gauss-Newton.
         - Use BFGS.
   (b) Else, use a gradient descent until you get a better "$\mathbf{x}_0$".
   (c) If all other methods are not converging rapidly, then try conjugate gradient.
2. If dimension large and Hess sparse, use conj. grad.

Gradient Methods.

Key Concept 1: Proposition 9.2.1.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function that is differentiable at $\mathbf{x} \in \mathbb{R}^n$. Among all unit vectors in $\mathbb{R}^n$, the unit vector $\mathbf{u} \in \mathbb{R}^n$ has the gradient directional derivative $\mathbf{D}_\mathbf{u} f(\mathbf{x})$ at $\mathbf{x}$ and has the normalized gradient
$$\mathbf{u} = \mathbf{D}f(\mathbf{x})^\mathsf{T} / \left\|\mathbf{D}f(\mathbf{x})^\mathsf{T}\right\|.$$

□ **Proof.** By C-S, for $\mathbf{u} \in \mathbb{R}^n$, we have

$$\begin{aligned} |\mathbf{D}f_\mathbf{u}(\mathbf{x})| &= |\mathbf{D}f(\mathbf{x})\mathbf{u}| \\ &= \left|\left\langle \mathbf{D}f(\mathbf{x})^\mathsf{T}, \mathbf{u}\right\rangle\right| \\ &\leq \left\|\mathbf{D}f(\mathbf{x})^\mathsf{T}\right\|. \end{aligned}$$

But if we let $\mathbf{u} = \mathbf{D}f(\mathbf{x})^\mathsf{T} / \left\|\mathbf{D}f(\mathbf{x})^\mathsf{T}\right\|$ we have

$$\begin{aligned} \mathbf{D}f_\mathbf{u}(\mathbf{x}) &= \left\langle \mathbf{D}f(\mathbf{x})^\mathsf{T}, \mathbf{D}f(\mathbf{x})^\mathsf{T}\right\rangle / \left\|\mathbf{D}f(\mathbf{x})^\mathsf{T}\right\| \\ &= \left\|\mathbf{D}f(\mathbf{x})^\mathsf{T}\right\|, \end{aligned}$$

so the normalized gradient $\mathbf{u} = \mathbf{D}f(\mathbf{x})^\mathsf{T} / \left\|\mathbf{D}f(\mathbf{x})^\mathsf{T}\right\|$ maximizes the directional derivative. ■

- **Gradient Descent Methods** are of the form
$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{D}f(\mathbf{x}_k)^\mathsf{T}.$$
(*i*). You could choose $\alpha_k = 1$. If descent, keep $\alpha_k = 1$; (*ii*). else, let $\alpha_k = \frac{1}{2} \cdot \alpha_k$. (*iii*). Then let $\alpha_{k+1} = 1$ and return 1. Then return to (*i*).
- Line searching:
$$\alpha_k = \arg\min_{\alpha \in (0,\infty)} f(\mathbf{x}_{k+1}).$$
This method is called **Steepest Descent**.

Gradient Methods.

Key Concept 2: Proposition.

Let $f : \mathbb{R} \to \mathbb{R}$ be $\mathcal{C}^1$. If $\mathbf{d}_k = -\mathbf{D}f(\mathbf{x}_k)^\mathsf{T} \neq \mathbf{0}$ and α_k is chosen with line search. Then setting
$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{D}f(\mathbf{x}_k)^\mathsf{T}$$
yields
$$f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k).$$

□ **Proof.** $\phi_k(\alpha_k) \leq \phi_k(\alpha)$ for $\alpha \geq 0$; by chain rule,
$$\begin{aligned} \phi_k'(0) &= -\mathbf{D}f(\mathbf{x}_k)\mathbf{D}f(\mathbf{x}_k)^\mathsf{T} \\ &= -\left\|\mathbf{D}f(\mathbf{x}_k)^\mathsf{T}\right\|^2 \\ &< 0. \end{aligned}$$
Since $f \in \mathcal{C}^1$, the function $\phi(\alpha) \in \mathcal{C}^1$, which means $\phi'(\alpha)$ is negative on some open nhbd of 0. Then $\phi(\alpha)$ is decreasing on that nhbd. i.e., $\exists \bar{\alpha} > 0 : \phi(\alpha) < \phi(0) \forall \alpha \in (0, \bar{\alpha}]$, so
$$\begin{aligned} f(\mathbf{x}_{k+1}) &= \phi_k(\alpha_k) \\ &\leq \phi_k(\bar{\alpha}) \\ &< \phi_k(0) \\ &= f(\mathbf{x}_k). \end{aligned}$$
■

Steepest Descent.

◊ **Example.** For a quadratic $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\mathsf{T}\mathbf{Q}\mathbf{x} - \mathbf{b}^\mathsf{T}\mathbf{x} + c$, $\mathbf{Q} > 0$, we can find an explicit formula for α_k in steepest descent method. *Note.* $\mathbf{D}f(\mathbf{x}_k)^\mathsf{T} = \mathbf{Q}\mathbf{x}_k - \mathbf{b}$. α_k minimizes $\phi(\alpha) = f\left(\mathbf{x}_k - \alpha \mathbf{D}f(\mathbf{x}_k)^\mathsf{T}\right)$.
$$\phi'(\alpha_k) = 0.$$

$$\begin{aligned} 0 &= \phi'(\alpha_k) \\ &= -\mathbf{D}f\left(\mathbf{x}_k - \alpha_k \mathbf{D}f(\mathbf{x}_k)^\mathsf{T}\mathbf{D}f(\mathbf{x}_k)^\mathsf{T}\right) \\ &= \left(\left(\mathbf{x}_k - \alpha_k \mathbf{D}f(\mathbf{x}_k)^\mathsf{T}\right)^\mathsf{T}\mathbf{Q} - \mathbf{b}^\mathsf{T}\right)\mathbf{D}f(\mathbf{x}_k)^\mathsf{T} \\ &= -\left(\mathbf{x}_k - \alpha_k \mathbf{D}f(\mathbf{x}_k)^\mathsf{T}\right)^\mathsf{T}\mathbf{Q}\mathbf{D}f(\mathbf{x}_k)^\mathsf{T} + \mathbf{b}^\mathsf{T}\mathbf{D}f(\mathbf{x}_k)^\mathsf{T}. \end{aligned}$$

This implies that

$$\begin{aligned}\left(\alpha_k \mathbf{D}f\left(\mathbf{x}_k\right)^{\mathsf{T}}\right)^{\mathsf{T}} \mathbf{Q}\mathbf{D}f\left(\mathbf{x}_k\right)^{\mathsf{T}} &= \left(\mathbf{x}_k^{\mathsf{T}}\mathbf{Q} - \mathbf{b}^{\mathsf{T}}\right)\mathbf{D}f\left(\mathbf{x}_k\right)^{\mathsf{T}} \\ &= \mathbf{D}f\left(\mathbf{x}_k\right)\mathbf{D}f\left(\mathbf{x}_k\right)^{\mathsf{T}},\end{aligned}$$

so

$$\alpha_k = \frac{\mathbf{D}f\left(\mathbf{x}_k\right)\mathbf{D}f\left(\mathbf{x}_k\right)^{\mathsf{T}}}{\mathbf{D}f\left(\mathbf{x}_k\right)\mathbf{Q}\mathbf{D}f\left(\mathbf{x}_k\right)^{\mathsf{T}}}.$$

♦

ⓘ **Note.** The important thing to note about steepest descent is that
- The next direction is orthogonal to the last direction.
- Each step stops at a point tangent to the level set.

Let $(\lambda_1, \ldots, \lambda_n)$ be the eigenvalues of $\mathbf{Q}$. If the eigenvalues are all equal, then we have circle level sets. If they are all very disparate, we get ellipsoid level sets.

Recall that the definition of gradient in the univariate case is

$$\frac{\partial f}{\partial x} = \lim_{h\to 0}\frac{f(x+h)-f(x)}{h}.$$

In the multivariate case, the analog is

$$\mathbf{D}_i f\left(\mathbf{x}\right) \approx \lim_{h\to 0}\frac{f\left(\mathbf{x}+h\mathbf{e}_i\right)-f\left(\mathbf{x}\right)}{h},$$

where $\mathbf{e}_i$ is the ith basis vector in the domain of f.

How do we select h? The general rule of thumb is to select $\boxed{h \approx 2\sqrt{\mathrm{Rerr}_f}}$.

Key Concept 3: Theorem.

Let $f \in \mathcal{C}^2\left(\mathbb{R}^n, \mathbb{R}\right)$ be computed as $\tilde{f}$ with a Rerr_f near $\mathbf{x}_0$, as assume that $\left|f\left(\mathbf{x}\right)\right| < M$ and $\left\|\mathbf{D}^2 f\left(\mathbf{x}\right)\right\| < L$ near $\mathbf{x}_0$. Assume that $h > 0$. For $i \in \{1, \ldots, n\}$, let

$$\widetilde{\mathbf{D}_i} f\left(\mathbf{x}_0\right) = \left(\tilde{f}\left(\mathbf{x}_0 \oplus h\mathbf{e}_i\right) \ominus \tilde{f}\left(\mathbf{x}_0\right)\right) \oslash h,$$

where $\oplus, \ominus, \oslash$ are computer operators for $+, -, /$. Then, we have

$$\left\|\mathbf{D}f_i\left(\mathbf{x}_0\right) - \widetilde{\mathbf{D}_i} f\left(\mathbf{x}_0\right)\right\| \leq \frac{1}{2}hL + \frac{2M\mathrm{Rerr}_f + \varepsilon_{\text{machine}}}{h} + \varepsilon_{\text{machine}}.$$

ⓘ **Note.**

$$\mathbf{D}^2_{ij} f\left(\mathbf{x}_0\right) \approx \frac{f\left(\mathbf{x}_0 + h\mathbf{e}_i + h\mathbf{e}_j\right) - f\left(\mathbf{x}_0 + h\mathbf{e}_i\right) - f\left(\mathbf{x}_0 + h\mathbf{e}_j\right) - f\left(\mathbf{x}_0\right)}{h^2}.$$

Newton's Method.

- Let $f : \mathbb{R}^n \to \mathbb{R}$. It is $\mathcal{C}^2$ and $\mathbf{x}^* \in \mathbb{R}^n$ is a local minimizer of f satisfying $\mathbf{D}^2 f(\mathbf{x}^*) > 0$. Let

$$q(\mathbf{x}) = f(\mathbf{x}_k) + \mathbf{D}f(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_k)^{\mathsf{T}} \mathbf{D}^2 f(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k),$$

  and $\mathbf{x}_{k+1}$ is defined to be the minimizer of $q(\mathbf{x})$. Then

$$\begin{aligned} \mathbf{D}q(\mathbf{x}) &= \mathbf{D}f(\mathbf{x}_k) + (\mathbf{x}^* - \mathbf{x}_k)^{\mathsf{T}} \mathbf{D}^2 f(\mathbf{x}_k) \\ &= 0 \end{aligned}$$

  implies that

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{D}^2 f(\mathbf{x}_k)^{-1} \mathbf{D}f(\mathbf{x}_k)^{-1}.$$

- $$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{D}^2 f(\mathbf{x}_k + \mu_k \mathbf{I})^{-1} \mathbf{D}f(\mathbf{x}_k)^{\mathsf{T}}.$$

F, 27 Jul. 18

Broyden's Method.

- A simpler method than the BFGS method.
- In order to minimize $f(\mathbf{x})$, we can minimize g_k, where

$$g_k(\mathbf{x}) = f(\mathbf{x}_k) + \mathbf{D}f(\mathbf{x}_k)(\mathbf{x}-\mathbf{x}_k) + \frac{1}{2}(\mathbf{x}-\mathbf{x}_k)\mathbf{A}_k(\mathbf{x}_k)(\mathbf{x}-\mathbf{x}_k).$$

Let $\mathbf{A}_0 = \mathbf{D}^2 f(\mathbf{x}_0)$, and

$$\mathbf{D}g_{k+1}(\mathbf{x}) = \mathbf{D}f(\mathbf{x}_k) + (\mathbf{x}-\mathbf{x}_{k+1})^{\mathsf{T}}\mathbf{A}_{k+1},$$

and

$$\mathbf{D}f(\mathbf{x}_{k+1}) - \mathbf{D}f(\mathbf{x}_k) = (\mathbf{x}_{k+1}-\mathbf{x}_k)^{\mathsf{T}}\mathbf{A}_{k+1}. \tag{9.12}$$

- Now let

$$\mathbf{y}_k \coloneqq \mathbf{D}f(\mathbf{x}_{k+1}) - \mathbf{D}f(\mathbf{x}_k)$$

and

$$\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k,$$

so (9.12) becomes $\mathbf{y}_k^{\mathsf{T}} = \mathbf{s}_k^{\mathsf{T}}\mathbf{A}_{k+1}$, and we have that

$$\mathbf{A}_{k+1} = \mathbf{A}_k + \frac{\mathbf{y}_k - \mathbf{A}_k\mathbf{s}_k}{\|\mathbf{s}_k\|^2}\mathbf{s}_k^{\mathsf{T}},$$

which is the best approximation of $\mathbf{A}_{k+1}$ given $\mathbf{A}_k$ because it minimizes the normed difference between $\mathbf{A}$ and $\mathbf{A}_k$.

- So, **Broyden's Method** essentially boils down to the following Quasi-Newton method.

$$\begin{cases} \mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{A}_k^{-1}\mathbf{D}f(\mathbf{x}_k)^{\mathsf{T}} \\ \mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k \\ \mathbf{y}_k = \mathbf{D}f(\mathbf{x}_{k+1})^{\mathsf{T}} - \mathbf{D}f(\mathbf{x}_k)^{\mathsf{T}} \\ \mathbf{A}_{k+1} = \mathbf{A}_k + \frac{\mathbf{y}_k - \mathbf{A}_k\mathbf{s}_k}{\|\mathbf{s}_k\|^2}\mathbf{s}_k^{\mathsf{T}}. \end{cases}$$

The main idea is that we are using an *approximation* for the Hessian each step, instead of solving for the Hessian in more expensive ways. For big n, it makes a big difference.

Key Concept 4: Sherman-Morrison-Woodbury (SMW) Proposition.

For nonsingular $n \times n$ matrix $\mathbf{A}$, $n \times \ell$ matrix $\mathbf{B}$, nonsingular $\ell \times \ell$ matrix $\mathbf{C}$, and $\ell \times n$ matrix $\mathbf{D}$, we have

$$(\mathbf{A}+\mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{D}\mathbf{A}^{-1}. \tag{9.13}$$

□ **Proof.** (*If $\mathbf{A}^{-1}$ is known, the cost of finding $(\mathbf{A}+\mathbf{BCD})^{-1}$ via SMW is $\mathcal{O}\left(\ell^3 + n\ell\right)$*). Start with We have If , then

$$\boxed{\mathbf{x}_{k-1} = \mathbf{x}_k - \mathbf{A}_k^{-1}\mathbf{D}f(\mathbf{x}_k)^{\mathsf{T}}}.$$

■

BFGS Method.

- The evolution of $\mathbf{x}$ is as follows:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{B}_k \mathbf{D} f\left(\mathbf{x}_k\right)^{\mathsf{T}}.$$

Applying Taylor's theorem, we have

$$f\left(\mathbf{x}_{k+1}\right) = f\left(\mathbf{x}_k\right) + \mathbf{D} f\left(\mathbf{x}_k\right)\left(\mathbf{x}_{k+1} - \mathbf{x}_k\right) + o\left(\left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\|\right),$$

so if $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|$ is sufficiently small, and if $\mathbf{B}_k > 0$ (positive definite), then $f\left(\mathbf{x}_{k+1}\right) < f\left(\mathbf{x}_k\right)$.