Algorithms for tensor approximations

Lek-Heng Lim

MSRI Summer Graduate Workshop

July 7–18, 2008

Synopsis

- **Naïve:** the Gauss-Seidel heuristic.
- **Harmonic analysis:** pursuits algorithms.
- **Real algebraic geometry:** semi-definite programming.
- **Riemannian geometry:** Grassman-Newton method.

Recap: best low rank approximation of a hypermatrix

- **Outer product rank:** $\mathcal{A} \in \mathbb{R}^{l \times m \times n}$. Want $\mathbf{u}_i \in \mathbb{R}^l$, $\mathbf{v}_i \in \mathbb{R}^m$, $\mathbf{w}_i \in \mathbb{R}^n$ unit vectors, $\sigma_i \in \mathbb{R}$, that minimize
$$\left\|\mathcal{A} - \sum\nolimits_{i=1}^{r} \sigma_i \mathbf{u}_i \otimes \mathbf{v}_i \otimes \mathbf{w}_i\right\|.$$
- **Symmetric outer product rank:** $\mathcal{A} \in \mathsf{S}^3(\mathbb{R}^n)$. Want $\mathbf{v}_i$ unit vector, $\lambda_i \in \mathbb{R}$, that minimize
$$\left\|\mathcal{A} - \sum\nolimits_{i=1}^{r} \lambda_i \mathbf{v}_i \otimes \mathbf{v}_i \otimes \mathbf{v}_i\right\|.$$
- **Nonnegative outer product rank:** $\mathcal{A} \in \mathbb{R}_+^{l \times m \times n}$. Want $\mathbf{x}_i \in \mathbb{R}_+^l$, $\mathbf{y}_i \in \mathbb{R}_+^m$, $\mathbf{z}_i \in \mathbb{R}_+^n$ unit vectors, $\delta_i \in \mathbb{R}_+$, that minimize
$$\left\|\mathcal{A} - \sum\nolimits_{i=1}^{r} \delta_i \mathbf{x}_i \otimes \mathbf{y}_i \otimes \mathbf{z}_i\right\|.$$

Recap: best low rank approximation of a hypermatrix

- **Multilinear rank:** $\mathcal{A} \in \mathbb{R}^{l \times m \times n}$. Want $U \in \mathbb{R}^{l \times r_1}$, $V \in \mathbb{R}^{m \times r_2}$, $W \in \mathbb{R}^{n \times r_3}$ matrices with orthonormal columns, $\mathcal{C} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$, that minimize
$$\|\mathcal{A} - (U, V, W) \cdot \mathcal{C}\|.$$
- **Hybrid:** $\mathcal{A} \in \mathbb{R}^{l \times m \times n}$. Want $\mathcal{B}_1, \dots, \mathcal{B}_r \in \mathbb{R}^{l \times m \times n}$ with
$$\operatorname{rank}_{\boxplus}(\mathcal{B}_i) \leq (r_1, r_2, r_3), \quad \|\mathcal{B}_i\| = 1,$$
that minimize
$$\left\|\mathcal{A} - \sum\nolimits_{i=1}^{r} \sigma_i \mathcal{B}_i\right\|.$$

Gauss-Seidel method

- Optimal solution $\mathcal{B}_*$ to $\operatorname{argmin}_{\operatorname{rank}_\otimes(\mathcal{B})\le r}\|\mathcal{A}-\mathcal{B}\|_F$ not easy to compute since the objective function is non-convex.
- A widely used strategy is a nonlinear **Gauss-Seidel** algorithm, better known as the **Alternating Least Squares** algorithm:

Algorithm: ALS for optimal rank-r approximation

initialize $X^{(0)} \in \mathbb{R}^{l\times r}, Y^{(0)} \in \mathbb{R}^{m\times r}, Z^{(0)} \in \mathbb{R}^{n\times r}$;
initialize $s^{(0)}, \varepsilon > 0, k = 0$;
while $\rho^{(k+1)}/\rho^{(k)} > \varepsilon$;
$\quad X^{(k+1)} \leftarrow \operatorname{argmin}_{\bar{X}\in\mathbb{R}^{l\times r}}\|T - \sum_{\alpha=1}^{r}\bar{\mathbf{x}}_\alpha^{(k+1)}\otimes\mathbf{y}_\alpha^{(k)}\otimes\mathbf{z}_\alpha^{(k)}\|_F^2$;
$\quad Y^{(k+1)} \leftarrow \operatorname{argmin}_{\bar{Y}\in\mathbb{R}^{m\times r}}\|T - \sum_{\alpha=1}^{r}\mathbf{x}_\alpha^{(k+1)}\otimes\bar{\mathbf{y}}_\alpha^{(k+1)}\otimes\mathbf{z}_\alpha^{(k)}\|_F^2$;
$\quad Z^{(k+1)} \leftarrow \operatorname{argmin}_{\bar{Z}\in\mathbb{R}^{n\times r}}\|T - \sum_{\alpha=1}^{r}\mathbf{x}_\alpha^{(k+1)}\otimes\mathbf{y}_\alpha^{(k+1)}\otimes\bar{\mathbf{z}}_\alpha^{(k+1)}\|_F^2$;
$\quad \rho^{(k+1)} \leftarrow \|\sum_{\alpha=1}^{r}[\mathbf{x}_a^{(k+1)}\otimes\mathbf{y}_\alpha^{(k+1)}\otimes\mathbf{z}_\alpha^{(k+1)} - \mathbf{x}_\alpha^{(k)}\otimes\mathbf{y}_\alpha^{(k)}\otimes\mathbf{z}_\alpha^{(k)}]\|_F^2$;
$\quad k \leftarrow k+1$;

- Coordinate cycling heuristic. May not converge.

Best r-term approximation

$$f \approx \alpha_1 f_1 + \alpha_2 f_2 + \cdots + \alpha_r f_r.$$

- **Target function** $f \in \mathcal{H}$ vector space, cone, etc.
- $f_1, \dots, f_r \in \mathscr{D} \subset \mathcal{H}$ **dictionary**.
- $\alpha_1, \dots, \alpha_r \in \mathbb{R}$ or $\mathbb{C}$ (linear), $\mathbb{R}_+$ (convex), $\mathbb{R} \cup \{-\infty\}$ (tropical).
- $\approx$ with respect to $\varphi : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$, some measure of 'nearness' between pairs of points (e.g. norms, metric, volumes, expectation, entropy, Brègman divergences, etc), want

$$\operatorname{argmin}\{\varphi(f, \alpha_1 f_1 + \dots \alpha_r f_r) \mid f_i \in \mathscr{D}\}.$$

- For concreteness, $\mathcal{H}$ separable Hilbert space; measure of nearness is a norm, but not necessarily the one induced by its inner product.
- Reference: various papers by A. Cohen, R. DeVore, V. Temlyakov.

Recap: dictionaries

- Discrete cosine:

$$\mathscr{D} = \left\{ \sqrt{\tfrac{2}{N}} \cos(k + \tfrac{1}{2})(n + \tfrac{1}{2})\tfrac{\pi}{N} \;\middle|\; k \in [N-1] \right\} \subseteq \mathbb{C}^N.$$

- Taylor:

$$\mathscr{D} = \{x^n \mid n \in \mathbb{N} \cup \{0\}\} \subseteq C^\omega(\mathbb{R}).$$

- Fourier:

$$\mathscr{D} = \{\cos(nx), \sin(nx) \mid n \in \mathbb{Z}\} \subseteq L^2(-\pi, \pi).$$

- Peter-Weyl:

$$\mathscr{D} = \{\langle \pi(x)\mathbf{e}_i, \mathbf{e}_j \rangle \mid \pi \in \widehat{G}, i, j \in [d_\pi]\} \subseteq L^2(G).$$

Recap: dictionaries

- Paley-Wiener:

$$\mathscr{D} = \{\operatorname{sinc}(x - n) \mid n \in \mathbb{Z}\} \subseteq H^2(\mathbb{R}).$$

- Gabor:

$$\mathscr{D} = \{e^{i\alpha nx} e^{-(x-m\beta)^2/2} \mid (m, n) \in \mathbb{Z} \times \mathbb{Z}\} \subseteq L^2(\mathbb{R}).$$

- Wavelet:

$$\mathscr{D} = \{2^{n/2}\psi(2^n x - m) \mid (m, n) \in \mathbb{Z} \times \mathbb{Z}\} \subseteq L^2(\mathbb{R}).$$

- Friends of wavelets: $\mathscr{D} \subseteq L^2(\mathbb{R}^2)$ beamlets, brushlets, curvelets, ridgelets, wedgelets, multiwavelets.

Approximants

Definition

Dictionary $\mathscr{D} \subset \mathcal{H}$. For $r \in \mathbb{N}$, the set of **r-term approximants** is

$$\Sigma_r(\mathscr{D}) := \Big\{ \sum\nolimits_{i=1}^{r} \alpha_i f_i \in \mathcal{H} \;\Big|\; \alpha_i \in \mathbb{C}, f_i \in \mathscr{D} \Big\}.$$

Let $f \in \mathcal{H}$. The **error of r-term approximation** is

$$\sigma_n(f) := \inf\nolimits_{g \in \Sigma_r(\mathscr{D})} \|f - g\|.$$

- Linear combination of two r-term approximants may have more than r non-zero terms.
- $\Sigma_r(\mathscr{D})$ not a subspace of $\mathcal{H}$. Hence **nonlinear approximation**.
- In contrast with usual (linear) approximation, ie.

$$\inf\nolimits_{g \in \operatorname{span}(\mathscr{D})} \|f - g\|.$$

Small is beautiful

$$f \approx \sum_{i \in \mathscr{I} \subseteq \mathscr{D}} \alpha_i f_i$$

- Want good approximation, ie. $\|f - \sum_{i \in \mathscr{I} \subseteq \mathscr{D}} \alpha_i f_i\|$ small.
- Want sparse/concentrated representation, ie. $|\mathscr{I}|$ small.
- Sparsity depends on choice of $\mathscr{D}$.
  - $\mathscr{D}_{10} = \{10^n \mid n \in \mathbb{Z}\}, \mathscr{D}_3 = \{3^n \mid n \in \mathbb{Z}\} \subseteq \mathbb{R}$,
  $$\begin{aligned} \tfrac{1}{3} &= [0.33333\cdots]_{10} = \textstyle\sum_{n=1}^{\infty} 3 \cdot 10^{-n} \\ &= [0.1]_3 = 1 \cdot 3^{-1}. \end{aligned}$$
  - $\mathscr{D}_{\text{fourier}} = \{\cos(nx), \sin(nx) \mid n \in \mathbb{Z}\}$,
  $$\tfrac{1}{2}x = \sin(x) - \tfrac{1}{2}\sin(2x) + \tfrac{1}{3}\sin(3x) - \cdots .$$
  - $\mathscr{D}_{\text{taylor}} = \{x^n \mid n \in \mathbb{N} \cup \{0\}\}$,
  $$\sin(x) = x - \tfrac{1}{6}x^3 + \tfrac{1}{120}x^5 - \cdots .$$

Bigger is better

- **Union of dictionaries:** allows for efficient (sparse) representation of different features
  - $\mathscr{D} = \mathscr{D}_{\text{fourier}} \cup \mathscr{D}_{\text{wavelets}}$,
  - $\mathscr{D} = \mathscr{D}_{\text{spikes}} \cup \mathscr{D}_{\text{sinusoids}} \cup \mathscr{D}_{\text{splines}}$,
  - $\mathscr{D} = \mathscr{D}_{\text{wavelets}} \cup \mathscr{D}_{\text{curvelets}} \cup \mathscr{D}_{\text{beamlets}} \cup \mathscr{D}_{\text{ridgelets}}$.
- $\mathscr{D}$ **overcomplete** or **redundant** dictionary. Trade off: computational complexity.
- **Rule of thumb:** the larger and more diverse the dictionary, the more efficient/sparser the representation.
- **Observation:** $\mathscr{D}$ above all zero dimensional (at most countably infinite).
- **Question:** What about dictionaries with a continuously varying families of functions?

Dictionaries of positive dimensions

- Neural networks:

$$\mathscr{D} = \{\sigma(\mathbf{w}^\top \mathbf{x} + w_0) \in L^2(\mathbb{R}^n) \mid (w_0, \mathbf{w}) \in \mathbb{R} \times \mathbb{R}^n\}$$

  where $\sigma : \mathbb{R} \to \mathbb{R}$ sigmoid function, eg. $\sigma(x) = [1 + \exp(-x)]^{-1}$.

- Exponential:

$$\mathscr{D} = \{e^{-tx} \mid t \in \mathbb{R}_+\} \qquad \text{or} \qquad \mathscr{D} = \{e^{\tau x} \mid \tau \in \mathbb{C}\}.$$

- Separable:

$$\mathscr{D} = \{g \in L^2(\mathbb{R}^3) \mid g(x, y, z) = \vartheta(x)\varphi(y)\psi(z)\}$$

  where $\vartheta, \varphi, \psi : \mathbb{R} \to \mathbb{R}$.

- Symmetric separable:

$$\mathscr{D} = \{g \in L^2(\mathbb{R}^3) \mid g(x, y, z) = \varphi(x)\varphi(y)\varphi(z)\}$$

  where $\varphi : \mathbb{R} \to \mathbb{R}$.

Same thing different names

- rth secant (quasiprojective) variety of the Segre variety is the set of r term approximants.
- If $\mathscr{D} = \operatorname{Seg}(\mathbb{R}^l, \mathbb{R}^m, \mathbb{R}^n)$, then

$$\Sigma_r(\mathscr{D}) = \{\mathcal{A} \in \mathbb{R}^{l \times m \times n} \mid \operatorname{rank}_{\otimes}(\mathcal{A}) \leq r\}.$$

- Outer product decomposition:

$$\begin{aligned}\mathscr{D} &= \{\mathbf{u} \otimes \mathbf{v} \otimes \mathbf{w} \mid (\mathbf{u}, \mathbf{v}, \mathbf{w}) \in \mathbb{R}^l \times \mathbb{R}^m \times \mathbb{R}^n\} \\ &= \{\mathcal{A} \in \mathbb{R}^{l \times m \times n} \mid \operatorname{rank}_{\otimes}(\mathcal{A}) \leq 1\}.\end{aligned}$$

- Symmetric outer product decomposition:

$$\mathscr{D} = \{\mathbf{v} \otimes \mathbf{v} \otimes \mathbf{v} \mid \mathbf{v} \in \mathbb{R}^n\} = \{\mathcal{A} \in \mathsf{S}^3(\mathbb{R}^n) \mid \operatorname{rank}_{\mathsf{S}}(\mathcal{A}) \leq 1\}.$$

- Nonnegative outer product decomposition:

$$\begin{aligned}\mathscr{D} &= \{\mathbf{x} \otimes \mathbf{y} \otimes \mathbf{z} \mid (\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \mathbb{R}^l_+ \times \mathbb{R}^m_+ \times \mathbb{R}^n_+\} \\ &= \{\mathcal{A} \in \mathbb{R}^{l \times m \times n}_+ \mid \operatorname{rank}_+(\mathcal{A}) \leq 1\}.\end{aligned}$$

Pursuit algorithms

- Stepwise projection:

$$g_k = \operatorname{argmin}_{g \in \mathscr{D}} \{ \|f - h\| \mid h \in \operatorname{span}\{g_1, \dots, g_{k-1}, g\} \},$$
$$f_k = \operatorname{proj}_{\operatorname{span}\{g_1, \dots, g_k\}}(f).$$

- Orthonormal matching pursuit:

$$g_k = \operatorname{argmax}_{g \in \mathscr{D}} |\langle f - f_{k-1}, g \rangle|,$$
$$f_k = \operatorname{proj}_{\operatorname{span}\{g_1, \dots, g_k\}}(f).$$

- Pure greedy:

$$g_k = \operatorname{argmax}_{g \in \mathscr{D}} |\langle f - f_{k-1}, g \rangle|,$$
$$f_k = f_{k-1} + \langle f - f_{k-1}, g_k \rangle g_k.$$

- Relaxed greedy:

$$g_k = \operatorname{argmin}_{g \in \mathscr{D}} \{ \|f - h\| \mid h \in \operatorname{span}\{f_{k-1}, g\} \},$$
$$f_k = \alpha_k f_{k-1} + \beta_k g_k.$$

Recap: hypermatrices are functions on finite sets

Totally ordered finite sets: $[n] = \{1 < 2 < \cdots < n\}$, $n \in \mathbb{N}$.

- Hypermatrix (order 3)

$$f : [l] \times [m] \times [n] \to \mathbb{R}.$$

- If $f(i,j,k) = a_{ijk}$, then f is represented by $\mathcal{A} = [\![a_{ijk}]\!]_{i,j,k=1}^{l,m,n} \in \mathbb{R}^{l \times m \times n}$.
- $\ell^2([l] \times [m] \times [n]) = \ell^2([l]) \otimes \ell^2([m]) \otimes \ell^2([n])$: $\mathcal{A}, \mathcal{B} \in \mathbb{R}^{l \times m \times n}$,

$$\langle \mathcal{A}, \mathcal{B} \rangle = \sum\nolimits_{i,j,k=1}^{l,m,n} a_{ijk} b_{ijk}.$$

- Frobenius norm

$$\|\mathcal{A}\|_F^2 = \sum\nolimits_{i,j,k=1}^{l,m,n} a_{ijk}^2.$$

Pursuit algorithms for tensor approximations

- Tensor approximation.
  - Target function
    $$f : [l] \times [m] \times [n] \to \mathbb{R}.$$
  - Dictionary of **separable functions**,
    $$\mathscr{D}_{\otimes} = \{g : [l] \times [m] \times [n] \to \mathbb{R} \mid g(i,j,k) = \vartheta(i)\varphi(j)\psi(k)\},$$
    where $\vartheta : [l] \to \mathbb{R}$, $\varphi : [m] \to \mathbb{R}$, $\psi : [n] \to \mathbb{R}$.
- Symmetric tensor approximation.
  - Target function:
    $$f : [n] \times [n] \times [n] \to \mathbb{R}$$
    with $f(i,j,k) = f(j,i,k) = \cdots = f(k,j,i)$.
  - Dictionary of symmetric separable functions:
    $$\mathscr{D}_{\mathsf{S}} = \{g : [n] \times [n] \times [n] \to \mathbb{R} \mid g(i,j,k) = \vartheta(i)\vartheta(j)\vartheta(k)\},$$
    where $\vartheta : [l] \to \mathbb{R}$.

Pursuit algorithms for tensor approximations

- Nonnegative tensor approximation.
  - Target function
$$f : [l] \times [m] \times [n] \to \mathbb{R}_+.$$
  - Dictionary of nonnegative separable functions,
$$\mathscr{D}_+ = \{g : [l] \times [m] \times [n] \to \mathbb{R}_+ \mid g(i, j, k) = \vartheta(i)\varphi(j)\psi(k)\},$$
where $\vartheta : [l] \to \mathbb{R}_+$, $\varphi : [m] \to \mathbb{R}_+$, $\psi : [n] \to \mathbb{R}_+$.

Some history

- f polynomial in variables $\mathbf{x} = (x_1, \dots, x_N)$. Suppose $f : \mathbb{R}^N \to \mathbb{R}$ non-negative valued, ie. $f(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^N$.
- **Question:** Can we write f as a sum of squares of polynomials,

$$f(\mathbf{x}) = \sum\nolimits_{j=1}^{M} p_j(\mathbf{x})^2 \quad ?$$

- **Answer (Hilbert):** Not in general, eg.
$f(w, x, y, z) = w^4 + x^2y^2 + y^2z^2 + z^2x^2 - 4xyzw$.
- **Hilbert's 17th Problem:** Can we write f as a sum of squares of rational functions,

$$f(\mathbf{x}) = \sum\nolimits_{j=1}^{M} \left(\frac{p_j(\mathbf{x})}{q_j(\mathbf{x})}\right)^2 \quad ?$$

- **Answer (Artin):** Yes!

SDP based algorithms

- **Observation 1:**

$$F(x_{11}, \dots, z_{nr}) = \|A - \textstyle\sum_{\alpha=1}^{r} \mathbf{x}_\alpha \otimes \mathbf{y}_\alpha \otimes \mathbf{z}_\alpha\|_F^2$$
$$= \textstyle\sum_{i,j,k=1}^{l,m,n} \left(a_{ijk} - \sum_{\alpha=1}^{r} x_{i\alpha} y_{j\alpha} z_{k\alpha}\right)^2$$

  is a polynomial of total degree 6 (resp. $2k$ for order k-tensors) in variables $x_{11}, \dots, z_{nr}$.

- **Multivariate polynomial optimization:** non-convex problem

$$\operatorname{argmin} F(x_{11}, \dots, z_{nr})$$

  may be relaxed to a convex problem (thus global optima is guranteed) which can in turn be solved using semidefinite programming (SDP).

- [Lasserre; 2001], [Parrilo; 2003], [Parrilo, Sturmfels; 2003].

How it works

- **Observation 2:** If $F - \lambda$ can be expressed as a sum of squares of polynomials

$$F(x_{11}, \dots, z_{nr}) - \lambda = \sum_{i=1}^{n} P_i(x_{11}, \dots, z_{nr})^2,$$

then λ is a global lower bound for F, ie.

$$F(x_{11}, \dots, z_{nr}) \geq \lambda$$

for all $x_{11}, \dots, z_{nr} \in \mathbb{R}$.

- **Simple strategy:** Find the largest λ^* such that $F - \lambda^*$ is a sum of squares. Then λ^* is often $\min F(x_{11}, \dots, z_{nr})$.

Sketch

- Write $\mathbf{v} = (1, x_{11}, \dots, z_{nr}, \dots, x_{l1}y_{m1}z_{n1}, \dots, z_{nr}^6)^\top$, the D-tuple of monomials of total degree ≤ 6, where

$$D := \binom{r(l+m+n)+3}{3}.$$

- Write $F(x_{11}, \dots, z_{nr}) = \boldsymbol{\alpha}^\top \mathbf{v}$ where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_D) \in \mathbb{R}^D$ are the coefficients of the respective monomials.
- Since $\deg(F)$ is even, F may also be written as

$$F(x_{11}, \dots, z_{nr}) = \mathbf{v}^\top M \mathbf{v}$$

  for some $M \in \mathbb{R}^{D \times D}$.
- So

$$F(x_{11}, \dots, z_{nr}) - \lambda = \mathbf{v}^\top (M - \lambda E_{11}) \mathbf{v}$$

  where $E_{11} = \mathbf{e}_1 \mathbf{e}_1^\top \in \mathbb{R}^{D \times D}$.

Sketch

- **Observation 3:** The RHS is a sum of squares iff $M - \lambda E_{11}$ is positive semidefinite (since $M - \lambda E_{11} = B^\top B$). Hence we have

$$\begin{array}{rl} \text{minimize} & -\lambda \\ \text{subjected to} & \mathbf{v}^\top(S + \lambda E_{11})\mathbf{v} = F, \\ & S \succeq 0. \end{array}$$

- This is an SDP problem

$$\begin{array}{rll} \text{minimize} & 0 \circ S - \lambda & \\ \text{subjected to} & S \circ B_1 + \lambda = \alpha_1, & \\ & S \circ B_k = \alpha_k, & k = 2, \dots, D \\ & S \succeq 0, & \lambda \in \mathbb{R}. \end{array}$$

Properties

- May be solved in polynomial time.
- Like all SDP-based algorithms, duality produces a certificate that tells us whether we have arrived at a globally optimal solution.
- The *duality gap*, ie. difference between the values of the primal and dual objective functions, is 0 at a global minima.
- **Complexity:** For rank-r approximations to order-k tensors $A \in \mathbb{R}^{d_1 \times \cdots \times d_k}$,
$$D = \binom{r(d_1 + \cdots + d_k) + k}{k}$$
is large even for moderate d_i, r and k.
- **Sparsity to the rescue:** The polynomials that we are interested in are always sparse (eg. for $k = 3$, only terms of the form xyz or $x^2y^2z^2$ or $uvwxyz$ appear).

Newton polytope

Newton polytope of a polynomial f is the convex hull of the powers of the monomials in f.

Example

Newton polytope of $f(x, y) = 3.67x^4y^10 + -2.03x^3y^3 + 5.74x^3 - 20.1y^2 - 7.23$ is the convex hull of the points $(4, 10), (3, 3), (3, 0), (2, 0), (0, 0)$ in $\mathbb{R}^2$. Newton polytope of $g(x, y, z) = 1.7x^4y^6z^2 + 7.4x^3z^5 - 3.0y^4 + 0.1yz^2$ is the convex hull of the points $(4, 6, 2), (3, 0, 5), (0, 4, 0), (0, 1, 2)$ in $\mathbb{R}^3$.

Theorem (Reznick)

If $f(\mathbf{x}) = \sum_{i=1}^m p_i(\mathbf{x})^2$, then the powers of the monomials in p_i must lie in $\frac{1}{2}$ Newton(f).

Multilinear polynomial

- The Newton polytope for a polynomial of the form

$$f(x_{11}, \dots, z_{nr}) = -\lambda + \sum_{i,j,k=1}^{l,m,n} \left(a_{ijk} - \sum_{\alpha=1}^{r} x_{i\alpha} y_{j\alpha} z_{k\alpha} \right)^2$$

  is spanned by 1 and monomials of the form $x_{i\alpha}^2 y_{j\alpha}^2 z_{k\alpha}^2$ (ie. monomials of the form $x_{i\alpha} y_{j\alpha} z_{k\alpha}$ and $x_{i\alpha} y_{j\alpha} z_{k\alpha} x_{i\beta} y_{j\beta} z_{k\beta}$ may all be dropped).
- So if $f(x_{11}, \dots, z_{nr}) = \sum_{j=1}^{N} p_j(x_{11}, \dots, z_{nr})^2$, then only 1 and monomials of the form $x_{i\alpha} y_{j\alpha} z_{k\alpha}$ may occur in $p_1, \dots, p_N$.
- In other words, we have reduced the size of the problem from $\binom{r(l+m+n)+3}{3}$ to $rlmn + 1$.

Global convergence

- If polynomials of the form

$$-\lambda + \sum_{i,j,k=1}^{l,m,n} \left(a_{ijk} - \sum_{\alpha=1}^{r} x_{i\alpha} y_{j\alpha} z_{k\alpha} \right)^2$$

  can *always* be written as a sum of polynomials (we don't know), then the SDP algorithm for optimal low-rank tensor approximation will *always* converge globally.

- Numerical experiments performed by Parrilo on general polynomials yield $\lambda^* = \min F$ in all cases.

Best multilinear rank approximation

- Given $\mathcal{A} \in \mathbb{R}^{l \times m \times n}$, want $\operatorname{rank}_{\boxplus}(\mathcal{B}) = (r_1, r_2, r_3)$ with

$$\min \|\mathcal{A} - \mathcal{B}\|_F = \min \|\mathcal{A} - (X, Y, Z) \cdot \mathcal{C}\|_F$$

$\mathcal{C} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$, $X \in \mathbb{R}^{l \times r_1}$, $Y \in \mathbb{R}^{m \times r_2}$, $Z \in \mathbb{R}^{n \times r_3}$ orthonormal.

- Problem overparameterized and equivalent to

$$\max \left\|(X^\top, Y^\top, Z^\top) \cdot \mathcal{A}\right\|_F = \max \|\mathcal{A} \cdot (X, Y, Z)\|_F \,,$$

$X^\top X = I, Y^\top Y = I, Z^\top Z = I$.

- Problem defined on a product of Grassmann manifolds since

$$\|\mathcal{A} \cdot (X, Y, Z)\|_F = \|\mathcal{A} \cdot (XQ_1, YQ_2, ZQ_3)\|_F,$$

for any $(Q_1, Q_2, Q_3) \in O(l) \times O(m) \times O(n)$. Only the subspaces spanned by X, Y, Z matters.

- Problem reformulated as

$$\max_{(X,Y,Z) \in \operatorname{Gr}(l,r_1) \times \operatorname{Gr}(m,r_2) \times \operatorname{Gr}(n,r_3)} \Phi(X, Y, Z).$$

Newton and Quasi-Newton algorithms on manifolds

- $\mathbf{T}_X$ tangent space at $X \in \mathrm{Gr}(n, r)$

$$\mathbb{R}^{n\times r} \ni \Delta \in \mathbf{T}_X \qquad \Longleftrightarrow \qquad \Delta^\top X = 0$$

1. Compute Grassmann gradient $\nabla\Phi \in \mathbf{T}_{(X,Y,Z)}$.
2. Compute Hessian or update Hessian approximation

$$H : \Delta \in \mathbf{T}_{(X,Y,Z)} \to H\Delta \in \mathbf{T}_{(X,Y,Z)}.$$

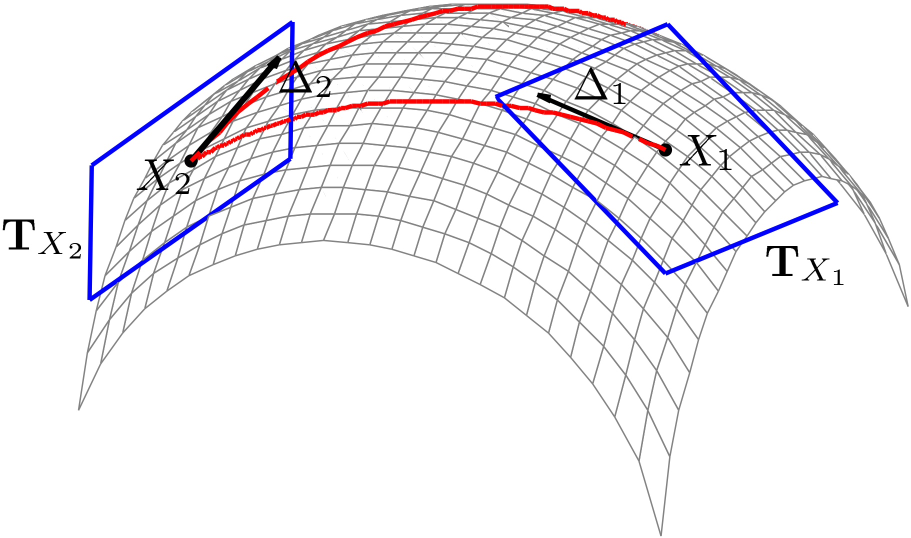
3. At $(X, Y, Z) \in \mathrm{Gr}(l, r_1) \times \mathrm{Gr}(m, r_2) \times \mathrm{Gr}(n, r_3)$, solve

$$H\Delta = -\nabla\Phi$$

   for search direction Δ.
4. Update iterate (X, Y, Z): Move along geodesic from (X, Y, Z) in the direction given by Δ.

- Optimize over a product of three (or more) Grassmannians.
- [Gabay, 1982], [Arias, Edelman, Smith; 1999], [Eldén, Savas; 2008].

Picture



Quasi-Newton and BFGS update

The BFGS update

$$H_{k+1} = H_k - \frac{H_k \mathbf{s}_k \mathbf{s}_k^\top H_k}{\mathbf{s}_k^\top H_k \mathbf{s}_k} + \frac{\mathbf{y}_k \mathbf{y}_k^\top}{\mathbf{y}_k^\top \mathbf{y}_k}$$

where

$$\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k = t_k \mathbf{p}_k,$$
$$\mathbf{y}_k = \nabla f_{k+1} - \nabla f_k.$$

On Grassmann manifold the vectors are defined on different points belonging to different tangent spaces.

Different ways of parallel transporting vectors

$X \in \mathrm{Gr}(n, r)$, $\Delta_1, \Delta_2 \in \mathbf{T}_X$ and $X(t)$ geodesic path along Δ_1

- Parallel transport using global coordinates

$$\Delta_2(t) = T_{\Delta_1}(t)\Delta_2$$

  we have also

$$\Delta_1 = X_\perp D_1 \quad \text{and} \quad \Delta_2 = X_\perp D_2$$

  where $X_\perp$ basis for $\mathbf{T}_X$. Let $X(t)_\perp$ be basis for $\mathbf{T}_{X(t)}$.

- Parallel transport using local coordinates

$$\Delta_2(t) = X(t)_\perp D_2.$$

Parallel transport in local coordinates

All transported tangent vectors have the same coordinate representation in the basis $X(t)_\perp$ at all points on the path $X(t)$.

Plus No need to transport the gradient or the Hessian.

Minus Need to compute $X(t)_\perp$.

In global coordinate we compute

- $\mathbf{T}_{k+1} \ni \mathbf{s}_k = t_k T_{\Delta_k}(t_k)\mathbf{p}_k$
- $\mathbf{T}_{k+1} \ni \mathbf{y}_k = \nabla f_{k+1} - T_{\Delta_k}(t_k)\nabla f_k$
- $T_{\Delta_k}(t_k) H_k T_{\Delta_k}^{-1}(t_k) : \mathbf{T}_{k+1} \longrightarrow \mathbf{T}_{k+1}$

$$H_{k+1} = H_k - \frac{H_k \mathbf{s}_k \mathbf{s}_k^\top H_k}{\mathbf{s}_k^\top H_k \mathbf{s}_k} + \frac{\mathbf{y}_k \mathbf{y}_k^\top}{\mathbf{y}_k^\top \mathbf{y}_k}$$

Limited memory BFGS

Compact representation of BFGS in Euclidean space:

$$H_k = H_0 + \begin{bmatrix} S_k & H_0 Y_k \end{bmatrix} \begin{bmatrix} R_k^{-\top}(D_k + Y_k^\top H_0 Y_k) R_k^{-1} & -R_k^{-\top} \\ -R_k^{-1} & 0 \end{bmatrix} \begin{bmatrix} S_k^\top \\ Y_k^\top H_0 \end{bmatrix}$$

where

$$\begin{aligned} S_k &= [\mathbf{s}_0, \dots, \mathbf{s}_{k-1}], \\ Y_k &= [\mathbf{y}_0, \dots, \mathbf{y}_{k-1}], \\ D_k &= \operatorname{diag}\left[\mathbf{s}_0^\top \mathbf{y}_0, \dots, \mathbf{s}_{k-1}^\top \mathbf{y}_{k-1}\right], \\ R_k &= \begin{bmatrix} \mathbf{s}_0^\top \mathbf{y}_0 & \mathbf{s}_0^\top \mathbf{y}_1 & \cdots & \mathbf{s}_0^\top \mathbf{y}_{k-1} \\ 0 & \mathbf{s}_1^\top \mathbf{y}_1 & \cdots & \mathbf{s}_1^\top \mathbf{y}_{k-1} \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \mathbf{s}_{k-1}^\top \mathbf{y}_{k-1} \end{bmatrix}. \end{aligned}$$

Limited memory BFGS

Limited memory BFGS [Byrd et al; 1994]. Replace H_0 by $\gamma_k I$ and keep the m most resent $\mathbf{s}_j$ and $\mathbf{y}_j$,

$$H_k = \gamma_k I + \begin{bmatrix} S_k & \gamma_k Y_k \end{bmatrix} \begin{bmatrix} R_k^{-\top}(D_k + \gamma_k Y_k^\top Y_k)R_k^{-1} & -R_k^{-\top} \\ -R_k^{-1} & 0 \end{bmatrix} \begin{bmatrix} S_k^\top \\ \gamma_k Y_k^\top \end{bmatrix}$$

where

$$\begin{aligned} S_k &= [\mathbf{s}_{k-m}, \dots, \mathbf{s}_{k-1}], \\ Y_k &= [\mathbf{y}_{k-m}, \dots, \mathbf{y}_{k-1}], \\ D_k &= \operatorname{diag}\left[\mathbf{s}_{k-m}^\top \mathbf{y}_{k-m}, \dots, \mathbf{s}_{k-1}^\top \mathbf{y}_{k-1}\right], \\ R_k &= \begin{bmatrix} \mathbf{s}_{k-m}^\top \mathbf{y}_{k-m} & \mathbf{s}_{k-m}^\top \mathbf{y}_{k-m+1} & \cdots & \mathbf{s}_{k-m}^\top \mathbf{y}_{k-1} \\ 0 & \mathbf{s}_{k-m+1}^\top \mathbf{y}_{k-m+1} & \cdots & \mathbf{s}_{k-m+1}^\top \mathbf{y}_{k-1} \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \mathbf{s}_{k-1}^\top \mathbf{y}_{k-1} \end{bmatrix}. \end{aligned}$$

L-BFGS on the Grassmann manifold

- In each iteration, parallel transport vectors in S_k and Y_k to $\mathbf{T}_k$, ie. perform
$$\bar{S}_k = TS_k, \qquad \bar{Y}_k = TY_k$$
where T is the transport matrix.
- No need to modify R_k or D_k
$$\langle \mathbf{u}, \mathbf{v} \rangle = \langle T\mathbf{u}, T\mathbf{v} \rangle$$
where $\mathbf{u}, \mathbf{v} \in \mathbf{T}_k$ and $T\mathbf{u}, T\mathbf{v} \in \mathbf{T}_{k+1}$.
- H_k nonsingular, Hessian is singular. No problem $\mathbf{T}_k$ at $\mathbf{x}_k$ is invariant subspace of H_k, ie. if $\mathbf{v} \in \mathbf{T}_k$ then $H_k\mathbf{v} \in \mathbf{T}_k$.
- [Savas, L.; 2008]