ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Prof. D. Kressner
M. Steinlechner

1 ► ISTA: A projected proximal gradient method

Last week, we have derived the explicit form of the proximal operator for the ℓ_1-norm: Let $g(\mathbf{x}) = \|x\|_1$, then

$$\Big(\operatorname{prox}_{\gamma g}(x)\Big)_i = \begin{cases} x_i - \gamma & \text{if } x_i > \gamma, \\ x_i + \gamma & \text{if } x_i < -\gamma, \\ 0 & \text{if } -\gamma \leq x_i \leq \gamma. \end{cases},$$

which can be written in short form as $\Big(\operatorname{prox}_{\gamma g}(x)\Big)_i = \max\{|x_i| - \gamma, 0\} \operatorname{sign}(x_i)$.

A very common application is the solution of ℓ_1-regularized least squares problems, the so-called *lasso* model:

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax - b\|^2 + \gamma\|x\|_1 \tag{1}$$

We now want to solve this problem using a simple projected proximal gradient scheme as defined by equation (5.49) in the script. This simple method is also known as ISTA.

a) Create the following test example of an overdetermined least squares problem,

```
rng(7);
m = 500;
n = 2500;
A = randn( m, n );
% make the columns have unit l2 norm
A = A ./ repmat( sqrt(sum(A.^2)), m, 1 );
% create a sparse exact solution
x_star = zeros(n, 1);
idx = randperm( n, 100 );
x_star(idx) = randn( 100, 1 );
% create corresponding right hand side
noise = 10^(-2) * randn( m, 1 );
b = A*x_star + noise;
```

Using this data set, implement ISTA in Matlab to solve the optimization problem (1).

For simplicity, use the constant step size $\alpha = 0.1$ and regularization parameter $\gamma = 0.1\|A^T b\|_\infty$. Stop after 500 iterations to obtain a solution $\mathbf{x}_{\text{end}}$. Plot the convergence of the cost function, where you use the last iterate as a reference value. How many nonzero elements does your solution $\mathbf{x}_{\text{end}}$ have?

2 ► Nonnegative Matrix Factorization (NMF)

The nonnegative rank-k approximation of a nonnegative matrix $A \in \mathbb{R}^{m\times n}$ is a statsistical tool to extract *features* from A. Typical applications are pattern recognition, recommendation systems, or spectral analysis.

The task reads

$$\min_{W\in\mathbb{R}^{m\times k},\ H\in\mathbb{R}^{k\times n}} \frac{1}{2}\|A - WH\|_F^2, \quad W \geq 0,\ H \geq 0,$$

where $\|M\|_F = (\sum_{ij} m_{ij}^2)^{1/2}$ is the Frobenius norm of a matrix. Typically, $k \ll \min(m, n)$.

(a) (i) Let

$$f\colon \mathbb{R}^{m\times k} \times \mathbb{R}^{k\times n} \to \mathbb{R},\ (W, H) \mapsto \frac{1}{2}\|A - WH\|_F^2.$$

At a point (W_0, H_0) calculate the partial gradients $\nabla_W f(W_0, H_0) \in R^{m\times k}$ and $\nabla_H f(W_0, H_0) \in \mathbb{R}^{k\times n}$.

Hint. The Frobenius norm is induced by the inner product $\langle A, B\rangle_F = \operatorname{trace}(A^T B)$. In particular, $\|A - B\|_F^2 = \|A\|^2 - 2\langle A, B\rangle_F + \|B\|_F^2$. The gradient is always defined by $\langle \nabla f(x), h\rangle = f'(x)h$.

(ii) Calculate the exact values $\alpha_W \geq 0$ and $\alpha_H \geq 0$ which minimize

$$\alpha \mapsto f(W_0 - \alpha\nabla_W f(W_0, H_0), H_0), \quad \text{and} \quad \alpha \mapsto f(W_0, H_0 - \alpha\nabla_H f(W_0, H_0)),$$

respectively.

(iii) Given arbitrary W and H, how do the orthogonal projections $P_1(W)$ and $P_2(H)$ onto the feasible set read?

*(b) Prove that the NMF problem always has at least one global solution.

Hint. Restrict to the case $\|WH\|_F \leq \|A\|_F$. Show that one can assume that each non-zero row of H has norm one. Prove that one can restrict to bounded W then.

3 ► Solving NMF with alternating projected gradients

(a) Implement in MATLAB the following function for NMF:

Require: matrix $A \geq 0$, $k \in [1, \min(m, n)]$, starting guesses $W_0, H_0 \geq 0$, maxiter
  $i \leftarrow 1$, $W \leftarrow W_0$, $H \leftarrow H_0$
  while $\|P(\nabla f(W, H))\|_F > 10^{-5}\|\nabla f(W_0, H_0)\|_F$ and $i <$ maxiter **do**
    $W \leftarrow P_1(W - \alpha_W \nabla_W f(W, H))$
    $H \leftarrow P_2(H - \alpha_H \nabla_H f(W, H))$
    $i \leftarrow i + 1$
  end while

The syntax has to be
`function [W,H] = nmf_projgrad(A,W0,H0,maxiter)`.

(b) Download the zip file from the homepage and extract it into a folder. Put your m-file into the same folder and run the script.

A nonnegative matix A is generated whose 38 columns are vectorized greyscale images taken from the Extended Yale Face Database B.[1] The script runs the NMF algorithm with $k = 8$ (be patient) and shows 8 "basis faces" (the columns of W) found by NMF. Every face in $\tilde{A} = WH$ is a superposition of these "faces". Another figure shows one such decomposition. Please experiment with the values of k!

[1] Georghiades et al., From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose, IEEE Trans. Pattern Anal. Mach. Intelligence, 23 (2001); Lee et al., Acquiring Linear Subspaces for Face Recognition under Variable Lighting, IEEE Trans. Pattern Anal. Mach. Intelligence, 27 (2005).