The Ordered Subsets Mirror Descent Optimization Method with Applications to Tomography

by

Aharon Ben-Tal, Tamar Margalit and Arkadi Nemirovski [*]
MINERVA Optimization Center,
Faculty of Industrial Engineering and Management
Technion – Israel Institute of Technology
Haifa, Israel.

November 20, 2004

Abstract. We describe an optimization problem arising in reconstructing 3D medical images from Positron Emission Tomography (PET). A mathematical model of the problem, based on the Maximum Likelihood principle is posed as a problem of minimizing a convex function of several millions variables over the standard simplex. To solve a problem of these characteristics, we develop and implement a new algorithm, Ordered Subsets Mirror Descent, and demonstrate, theoretically and computationally, that it is well suited for solving the PET reconstruction problem.

Key words: positron emission tomography, maximum likelihood, image reconstruction, convex optimization, mirror descent.

1 Introduction

The goal of this paper is to develop a *practical* algorithm for an extremely large-scale convex optimization problem arising in Nuclear Medicine - that of reconstructing images from data acquired by Positron Emission Tomography (PET).

The PET technique is described in Section 2, and the corresponding mathematical optimization problem is given in Section 3. The specific characteristics of the problem rules out most advanced optimization methods, and as a result we focus on gradient-type methods. Specifically, we develop an accelerated version of the Mirror Descent (MD) method ([Nem78]). The acceleration is based on the *Incremental Gradient* idea ([Ber95], [Ber96], [Ber97], [Luo91], [Luo94], [Tse98]), also known as the *Ordered Subsets* (OS) technique in the Medical Imaging Literature ([Hud94], [Man95], [Kam98]). The MD method is described in Section 4. The accelerated version, OSMD, is studied in Section 5 in particular for a specific setup of OSMD, suitable for the PET reconstruction problem. In Section 6 we report the results of testing the OSMD algorithm on several realistic cases, and also to the classical Subgradient Descent method. Our conclusion from these tests is that OSMD is a reliable and efficient algorithm for PET reconstruction, which compares favourably with the best currently commercially used methods.

[*] The research is part of the PARAPET Project supported by the EUC Esprit Programme.

2 Positron Emission Tomography

Positron Emission Tomography (PET) is a powerful, non-invasive, medical diagnostic imaging technique for measuring the metabolic activity of cells in the human body. It has been in clinical use since the early 1990s. PET imaging is unique in that it shows the **chemical functioning** of organs and tissues, while other imaging techniques - such as X-ray, computerized tomography (CT) and magnetic resonance imaging (MRI) - show **anatomic structures**. PET is the only method that can detect and display metabolic changes in tissue, distinguish normal tissue from those that are diseased, such as in cancer, differentiate viable from dead or dying tissue, show regional blood flow, and determine the distribution and fate of drugs in the body. It is useful clinically in patients with certain conditions affecting the brain and the heart as well as in patients with certain types of cancer. Because of its accuracy, effectiveness, and cost efficiency, PET is becoming indispensable for the diagnosis of disease and treatment of patients.

2.1 The physical principles of PET

A PET scan involves the use of a small amount of a radioactive material which has the property of emitting positrons (positively charged electrons). Such a substance is referred to as *positron emitter*. One of the prime reasons for the importance of PET in medical research and practice is the existence of positron-emitting isotopes of elements such as carbon, oxygen and fluorine. These isotopes can be attached or tagged to biochemical compounds such as glucose, ammonia, water etc. to form radioactive tracers that will mimic their stable counterparts biologically (i.e., the radio-tracer element does not modify the biochemical behavior of the molecule). The choice of the biochemical compound and the radioactive tracer depends on the particular medical information being sought. When these radioactive drugs (or "radio-pharmaceuticals") are administered to a patient, either by injection or inhalation of gas, they distribute within the body according to the physiologic pathways associated with their stable counterparts.

The scan begins after a delay ranging from seconds to minutes to allow for the radio-tracer transport to the organ of interest. Then, the radio-isotope decays to a more stable atom by emitting a positron from its nucleus. The emitted positron loses most of its kinetic energy after traveling only a few millimeters in living tissue. It is then highly susceptible to interaction with an electron, an event that annihilates both particles. The mass of the two particles is converted into 1.02 million electron volts of energy, divided equally between two gamma rays.

The two gamma rays fly off the point of annihilation in nearly opposite directions along a line with a completely random orientation (i.e., uniformly distributed in space). They penetrate the surrounding tissue and are recordered outside the patient by a PET scanner consisting of circular arrays (*rings*) of gamma radiation detectors.

Since the two gamma rays are emitted simultaneously and travel in almost exactly opposite directions, their source can be established with high accuracy. This is achieved by grouping the radiation detectors in pairs. Two opposing detectors register a signal only if both sense high-energy photons within a short ($\sim 10^{-8}$sec) timing window. Detection of two events at the same time is referred to as *coincidence event*. Each detector is in coincidence with a number of detectors opposite so as to cover a *field of view* (FOV) about half as large in diameter as the diameter of the detector array.

A coincidence event is assigned to a *line of response* (LOR) connecting the two relevant detectors. In the two-dimensional case, an LOR is identified by the angle ϕ and the distance

s from the scanner axis (the center of the FOV). A certain pair of detectors is identified by the LOR joining their centers, and is sometimes referred to as a *bin*. The total number of coincidence events detected by a specific pair of detectors, approximates the line integral of the radio-tracer concentration along the relevant LOR. Considering the total number of coincidence events detected by all pairs of detectors with the same angle ϕ, we get a parallel set of such line integrals, known as a *parallel projection* set or shortly, as a projection.

The measured data set is the collection of numbers of coincidences counted by different pairs of detectors, or equivalently, the number of counts in all bins that intersect the FOV. Based on the measured data, a mathematical algorithm, applied by a computer, tries to reconstruct the spatial distribution of the radioactivity within the body. The principle of image reconstruction by computerized tomography is that an object can be reproduced from a set of its projections taken at different angles. The validity of such a reconstruction depends, of course, on the number of counts collected. The number of projections is a parameter of the scanner, and it determines the size of the mathematical reconstruction problem.

Note that there are several factors affecting quantitative accuracy of the measured data (e.g., detector efficiency, attenuation, scatter, random events, etc.). Therefore, the total number of counts is typically much smaller than the total number of emissions.

The final result of the scan study is usually presented as a set of two-dimensional images (known as *slices*), which together compose the three-dimensional mapping of the tracer distribution within the body.

3 The optimization problem

For consistent data, i.e. free of noise and measurement errors, there is a unique analytic solution of the two-dimensional inversion problem of recovering a 2D image from the set of its one-dimensional projections. This solution derived by Radon in 1917 and becomes later the basis for computerized tomography. The method, named *Filtered Back-projection* (FBP), was first applied for 2D PET image reconstruction by Shepp and Logan in 1974 ([She74]).

The images obtained by the FBP method as well as other analytical methods, which are based on inverse transforms, tend to be "streaky" and noisy. To address the problem of noise, the study of statistical (iterative) reconstruction techniques has received much attention in the past few years. Iterative methods allow to incorporate naturally physical constraints and a priori knowledge not contained in the measured projections e.g., the Poisson nature of the emission process.

The formulation of the PET reconstruction problem as a maximum likelihood (ML) problem rather than as an inverse problem was initially suggested by Rockmore and Mackovski in 1976 ([Roc76]). It became feasible when Shepp and Vardi in 1982 ([She82]) and Vardi, Shepp and Kaufman in 1985 ([Var85]) showed how the Expectation Maximization (EM) algorithm could be used for the ML computation.

Mathematical model and the Maximum-Likelihood problem The goal of ML estimation, as applied to emission tomography, is to find the expected number of annihilations by maximizing the probability of the set of observations, i.e. the detected coincidence events.

The mathematical model is based on the realistic assumption that photon counts follow a Poisson process. To simplify the computations, we form a finite parameter space by imposing a grid of boxes (*voxels*) over the emitting object. Let $X(j)$ denote the number of radioactive events emitted from voxel j. It is assumed that $X(j), j = 1, \ldots, n$ are independent Poisson-distributed random variables with unknown means λ_j,

$$X(j) \sim \text{Poisson}(\lambda_j).$$

Let p_{ij} be the probability that an emission from voxel j will be detected in bin i. Note that p_{ij} defines a transition matrix (likelihood matrix) assumed to be known from the geometry of the detector array. The probability to detect an event emitted from voxel j is:

$$p_j = \sum_{i=1}^{m} p_{ij}. \tag{1}$$

The number of events emitted from voxel j and detected in bin i is defined by $X(i,j) = p_{ij}X(j)$. By a Bernoulli thinning process with the probabilities p_{ij}, for different j and i, $\{X(i,j)\}$ are also independent Poisson random variables. Let $Y(i)$ denote the total number of events detected by bin i, i.e.,

$$Y(i) = \sum_j p_{ij} X(i,j), \tag{2}$$

then, $Y(i)$ is also a Poisson random variable, with the mean

$$\mu_i = \sum_j p_{ij} \lambda_j, \tag{3}$$

and $Y(i)$'s are independent of each other. A more accurate model of the observations would be

$$\mu_i = \sum_j m_i p_{ij} \lambda_j + r_i + s_i,$$

where, r_i and s_i are known values for random and scatter coincidences, and m_i are known attenuation coefficients, but we will use the simplified model. We denote by y_i the observations, namely the realizations of the random variables $Y(i)$.

The problem of PET image reconstruction can be formulated in the context of an incomplete data problem: the *complete data* (but unobserved) are the number of counts emitted from each voxel ($X(j)$); the *incomplete data* (observed) are counts of photons collected in various bins (y_i); and the parameter to be estimated is the expected number of counts emitted from each voxel (λ_j). Thus, the reconstruction problem is equivalent to a parameter estimation problem, and a *maximum likelihood* function can be formulated. In general, the likelihood function can be defined as the joint probability density of the measured data known up to the unobservable parameters to be estimated. Maximizing this likelihood function with respect to the unobservable parameters yields the parameters with which the data are most consistent.

According to (2) and (3) the vector of observed data $y = (y_1, \ldots, y_m)^T$ has the following likelihood function:

$$\begin{aligned} L(\lambda) = p(Y = y|\lambda) &= \prod_{i=1}^{m} e^{-\mu_i} \frac{\mu_i^{y_i}}{y_i!} \\ &= \prod_{i=1}^{m} (exp[-\textstyle\sum_{j=1}^{n} \lambda_j p_{ij}] \frac{[\sum_{j=1}^{n} \lambda_j p_{ij}]^{y_i}}{y_i!}) \end{aligned} \tag{4}$$

The maximum likelihood estimate of λ is the vector $\overline{\lambda}$ maximizing $L(\lambda)$ or equivalently its logarithm:

$$\ln L(\lambda) = -\sum_{j=1}^{n} \lambda_j p_j + \sum_{i=1}^{m} y_i \ln(\sum_{j=1}^{n} \lambda_j p_{ij}) - \text{constant}. \tag{5}$$

Note that the function $\ln L(\lambda)$ is concave ([She82]). Therefore, we can write the following convex minimization problem with non-negativity constraints:

$$F(\lambda) \equiv \sum_{j=1}^{n} p_j \lambda_j - \sum_{i=1}^{m} y_i \ln \left(\sum_{j=1}^{n} p_{ij} \lambda_j \right) \to \min \mid \lambda \geq 0, \tag{6}$$

The optimal solution to the problem is the Maximum Likelihood estimate of the (discretized) density of the tracer.

Problem (6) is an extremely large-scale convex optimization program: the design dimension n (the number of voxels) normally is $128^3 = 2,097,152$, while the number m of bins (i.e., the number of log-terms in the objective) can vary from 6,000,000 to 20,000,000, depending on the type of the tomograph. On a 450 MHz Pentium III computer with 200 Mb RAM, a single computation of the value and the gradient of the objective (i.e., multiplication of given vectors once by the matrix $P = ||p_{ij}||$ and once by P^T) takes from 15 to 45 minutes, depending on m.

The huge sizes of the PET Image Reconstruction problem impose severe restrictions on the type of optimization techniques which could be used to solve (6):

A. With the design dimension of order of $n = 10^6$, the only option is to use methods whose computational effort per iteration is linear in n. Even with this complexity per iteration, the overall number of iterations should be at most few tens – otherwise the running time of the method will be too large for actual clinical applications.

B. The objective in (6) is not defined on the whole $\mathbf{R}^n$ and may blow up to ∞ as λ approaches a "bad" boundary point of the nonnegative orthant (e.g. the origin); moreover, (6) is a constrained problem, however simple the constraint might look.

Observation A rules out basically all advanced optimization methods, like Interior Point ones (or other Newton-based optimization techniques): in spite of the fast convergence in terms of iteration counts, these techniques (at least in their "theoretically valid" forms) will "never" finish even the first iteration... In principle, it could be possible to use quasi-Newton techniques. Such an approach, however, would require resolving difficulties coming from B, without a clear reward for the effort: to the best of our knowledge, in the case when the number of iterations is restricted to only a small fraction of the design dimension (see A), there is no theoretical or computational evidence in favor of quasi-Newton methods.

Consequently, in our case, the most promising methods seem to be simple gradient-descent type methods aimed at solving convex problems with simple constraints. For these methods, the complexity per iteration is linear in n. Moreover, in favorable circumstances, the rate of convergence of gradient-type methods, although poor, is independent (or nearly so) of the design dimension. As a result, with a gradient-type method one usually reaches the first one or two digits of the optimal value in a small number of iterations, and then the method "dies", i.e., in many subsequent iterations no more progress in accuracy is obtained. Note that this "convergence pattern" is, essentially, what is needed in the PET Image Reconstruction problem. Indeed, this is an inverse (and as such – an ill-posed) optimization problem; practice demonstrates that when solving it to high accuracy, in terms of the optimal value, (which is possible in the 2D case), the quality of the image first improves and then tends to deteriorate, resulting eventually in a highly noisy image. Thus, in the case in question we in fact are not interested in high-accuracy solutions, which makes gradient descent techniques an appropriate choice.

4 The Mirror Descent Scheme and minimization over a simplex

4.1 The general Mirror Descent scheme

The general Mirror Descent (gMD) scheme is aimed at solving a convex optimization problem

$$f(x) \to \min \mid x \in X \subset \mathbf{R}^n, \tag{7}$$

where X is a convex compact set in $\mathbf{R}^n$ and f is a Lipschitz continuous convex function on X.

> Note that the PET Image Reconstruction problem with $p_{ij} > 0$ can be easily converted to (7). Indeed, from the KKT conditions for (6) we deduce the complementarity equations
>
> $$\left(p_j - \sum_i y_i \frac{p_{ij}}{\sum\limits_\ell p_{i\ell}\lambda_\ell} \right) \lambda_j = 0, \;\; j = 1, ..., n.$$
>
> Summing up these equations, we see that any optimal solution λ to problem (6) must satisfy the equation
>
> $$\sum_j p_j \lambda_j = B \equiv \sum_i y_i.$$
>
> Thus, we loose nothing by adding to problem (6) the equality constraint $\sum\limits_j p_j\lambda_j = B$. If we further introduce the change of variables
>
> $$x_j = \frac{p_j \lambda_j}{B},$$
>
> we end up with the optimization program
>
> $$f(x) \equiv -\sum_{i=1}^m y_i \ln(\sum_j r_{ij} x_j) \to \min \mid x \in \Delta_n \equiv \{x \in \mathbf{R}^n : x \geq 0, \sum_i x_i = 1\}, \tag{8}$$
>
> where
>
> $$r_{ij} = B\frac{p_{ij}}{p_j}$$
>
> which is equivalent to (6). The new formulation (8) is of the form (7), with the standard simplex Δ_n playing the role of X. Besides this, the resulting objective f is convex and Lipschitz continuous on $X = \Delta_n$, provided that $p_{ij} > 0$.

The *setup* for the gMD method is given by the following entities:

1. A *compact convex* set $Y \supset X$;
2. A *norm* $\|\cdot\|$ on $\mathbf{R}^n$ and its associated projector of Y onto X

   $$\pi(y) \in \operatorname*{Argmin}_{x \in X} \|y - x\|,$$

   along with the corresponding *separator*

   $$\eta(y) \in \mathbf{R}^n : \quad \|\eta(y)\|_* \leq 1, \quad \eta^T(y)(y - x) \geq \|y - \pi(y)\| \;\; \forall x \in X, \tag{9}$$

   where

   $$\|\xi\|_* = \max\{\xi^T x \mid \|x\| \leq 1\}$$

   is the norm on $\mathbf{R}^n$ conjugate to $\|\cdot\|$;

3. A positive real α and a continuously differentiable convex function $w : Y \to \mathbf{R}$ which is α-strongly convex on Y w.r.t. the norm $\|\cdot\|$, i.e.

$$(w'(x) - w'(y))^T(x - y) \geq \alpha\|y - x\|^2 \quad \forall x, y \in Y \qquad [w' \equiv \nabla w]$$

It is assumed that we can compute efficiently

- The projector $\pi(y)$ and the separator $\eta(y)$, $y \in Y$;
- The Legendre transformation

$$W(\xi) = \max_{y \in Y}\left[\xi^T y - w(y)\right]$$

of $w(\cdot)$, $\xi \in \mathbf{R}^n$.

Note that α-strong convexity of w on Y implies, via the standard duality relations ([RW98], Proposition 12.54), that W is continuously differentiable on the entire $\mathbf{R}^n$ with Lipschitz continuous gradient:

$$\|W'(\xi) - W'(\eta)\| \leq \frac{1}{\alpha}\|\xi - \eta\|_* \quad \forall \xi, \eta \in \mathbf{R}^n. \tag{10}$$

Moreover, the mapping $\xi \mapsto W'(\xi) = \operatorname{argmax}_{x \in Y}[\xi^T x - w(x)]$ is a parameterization of Y.

The gMD method for solving (7) generates sequences $\xi_t \in \mathbf{R}^n$, $\widehat{x}_t \in Y$, $x_t \in X$ as follows:

- *Initialization:* Choose (arbitrarily) $x_0 \in X$ and set $\xi_1 = w'(x_0)$;
- *Step t,$t = 1, 2, ...$:*

S.1) Set

$$\widehat{x}_t = W'(\xi_t); \quad x_t = \pi(\widehat{x}_t); \quad \eta_t = \eta(\widehat{x}_t).$$

S.2) Compute the value $f(x_t)$ and a subgradient $f'(x_t)$ of f at x_t. If $f'(x_t) = 0$, then x_t is the exact minimizer of f on X, and we terminate. If $f'(x_t) \neq 0$, we set

$$\xi_{t+1} = w'(\widehat{x}_t) - \gamma_t[f'(x_t) + \|f'(x_t)\|_* \eta_t], \tag{11}$$

where $\gamma_t > 0$ is a stepsize, and pass to step $t + 1$.

- *Approximate solution* x^t generated in course of the first t steps of the method is the best (with the smallest value of f) of the points $x_1, ..., x_t$: $x^t \in \operatorname{Argmin}_{x \in \{x_1, ..., x_t\}} f(x)$.

The convergence properties of the Mirror Descent method are summarized in the following

Theorem 4.1 *Assume that f is convex and Lipschitz continuous on X, with Lipschitz constant, w.r.t. $\|\cdot\|$, equal to $L_{\|\cdot\|}(f)$, and that the subgradients $f'(x_t)$ used in the gMD satisfy the condition*

$$\|f'(x_t)\|_* \leq L_{\|\cdot\|}(f).$$

Then for every $t \geq 1$ one has

$$f(x^t) - \min_{x \in X} f(x) \leq \min_{1 \leq s \leq r \leq t} \frac{\Gamma(w) + \frac{2}{\alpha}\sum\limits_{\tau=s}^{r} \gamma_\tau^2 \|f'(x_\tau)\|_*^2}{\sum\limits_{\tau=s}^{r} \gamma_\tau}, \tag{12}$$

where

$$\Gamma(w)=\max_{x,y\in Y}[w(x)-w(y)-(x-y)^T w'(y)].$$

In particular, whenever $\gamma_t\to+0$ as $t\to\infty$ and $\sum\limits_\tau \gamma_\tau=\infty$, one has $f(x^t)-\min\limits_{x\in X} f(x)\to 0$ as $t\to\infty$. Moreover, with the stepsizes chosen as

$$\gamma_\tau=\frac{C(\alpha\Gamma(w))^{1/2}}{\|f'(x_\tau)\|_*\sqrt{t}}, \tag{13}$$

one has

$$f(x^t)-\min_{x\in X} f(x)\le \widehat{C}(C)L_{\|\cdot\|}(f)\sqrt{\frac{\Gamma(w)}{\alpha}}t^{-1/2},\ \ t=1,2,... \tag{14}$$

with certain universal function $\widehat{C}(\cdot)$.

The theorem, in a slightly modified setting, is proved in [Nem78]. Here it will be derived as a straightforward simplification of the proof of Theorem 5.1 below.

4.2 $\|\cdot\|_p$-Mirror Descent and minimization over the standard simplex

As we have seen, the PET Image Reconstruction problem can be converted to the form of (8), i.e., posed as the problem of minimizing a convex function f over the standard simplex Δ_n. Therefore we focus on the gMD scheme as applied to the particular case of $X=\Delta_n$.

Let us choose somehow $p\in(1,2]$ and consider the following setup for gMD:

$$Y=\{x\mid \|x\|_p\le 1\}\ [\supset \Delta_n];\quad \|\cdot\|=\|\cdot\|_p;\quad w(x)=\frac{1}{2}\|x\|_p^2. \tag{15}$$

This setup defines a family $\{\mathrm{MD}_p\}_{1<p\le 2}$ of *ℓ_p-Mirror Descent methods* for minimizing convex functions over the standard simplex Δ_n (in fact, MD_p can be used to minimize a convex function over a convex subset of the unit $\|\cdot\|_p$-ball). A natural question is *which one of these methods is best suited for minimization over Δ_n ?* To answer this question, note first that for setup (15), a straightforward calculation yields that

$$W(\xi)=\begin{cases}\frac{1}{2}\|\xi\|_q^2, & \|\xi\|_q\le 1\\ \|\xi\|_q-\frac{1}{2}, & \|\xi\|_q>1\end{cases},\quad q=\frac{p}{p-1}. \tag{16}$$

Moreover, it is known (to be self-contained, we reproduce the proof in Appendix 1) that the parameter α of strong convexity of w w.r.t. the $\|\cdot\|_p$-norm satisfies the relation

$$\alpha\equiv\alpha_p(n)\ge O(1)(p-1), \tag{17}$$

and the quantity $\Gamma(w)$ defined in (12) is

$$\Gamma(w)=O(1)$$

(here and in what follows, $O(1)$ are appropriate positive absolute constants). Consequently, the efficiency estimate (14) becomes

$$f(x^t)-\min_{x\in X} f(x)\le \widehat{C}(C)\frac{L_{\|\cdot\|_p}(f)}{\sqrt{p-1}}t^{-1/2},\ \ t=1,2,... \tag{18}$$

Recalling that for every $x \in \mathbf{R}^n$ one clearly has $\|x\|_p \leq \|x\|_1 \leq \|x\|_p n^{\frac{p-1}{p}}$, and therefore

$$L_{\|\cdot\|_1}(f) \leq L_{\|\cdot\|_p}(f) \leq L_{\|\cdot\|_1} n^{\frac{p-1}{p}},$$

we derive from (18) that

$$f(x^t) - \min_{x \in X} f(x) \leq \widehat{C}(C) \frac{n^{\frac{p-1}{p}}}{\sqrt{p-1}} L_{\|\cdot\|_1}(f) t^{-1/2}, \;\; t = 1, 2, ... \tag{19}$$

Assuming $n > 1$ and minimizing the right hand side over $p \in (1, 2]$, we see that a good choice of p is

$$p = p(n) = 1 + \frac{O(1)}{\ln n}. \tag{20}$$

With this choice of p, the efficiency estimate (19) becomes

$$f(x^t) - \min_{x \in X} f(x) \leq \widehat{C}(C) \frac{\sqrt{\ln n} L_{\|\cdot\|_1}(f)}{\sqrt{t}}, \;\; t = 1, 2, ..., \tag{21}$$

while the underlying stepsizes are

$$\gamma_t = \frac{C}{\|f'(x_\tau)\|_* \sqrt{\ln n} \sqrt{t}} \qquad [C > 0]. \tag{22}$$

In what follows, we refer to the Mirror Descent method with the setup given by (15), (20), (22) (where $C = O(1)$) as to $\|\cdot\|_1$-*Mirror Descent method* MD_1.

Discussion. In the family $\{\mathrm{MD}_p\}_{1 \leq p \leq 2}$ of Mirror Descent methods, the special case MD_2 is well known – it is a kind of the standard Subgradient Descent method originating from [Sho67] and [Pol67] and studied in numerous papers (for the "latest news" on SD, see [KLP99] and references therein). The only modification needed to get from the MD scheme not a "kind of" the Subgradient Descent, but *exactly* the standard SD method

$$x_{t+1} = \pi_X(x_t - \gamma_t f'(x_t)), \quad \pi_X(x) = \operatorname*{argmin}_{y \in X} \|x - y\|_2, \tag{23}$$

for minimizing a convex function over a convex subset X of the unit Euclidean ball, is to set in (15) $p = 2$ and $Y = X$ rather than $p = 2$ and $Y = \{x \mid \|x\|_2 \leq 1\}$. Our analysis demonstrates, however, that when minimizing over the standard simplex, the "non-Euclidean" Mirror Descent MD_1 is preferable to the usual SD. Indeed, the best efficiency estimate known so far for SD as applied to minimizing a convex Lipschitz continuous function f over the standard simplex Δ_n is

$$f(x^t) - \min_{x \in \Delta_n} f(x) \leq O(1) \frac{L_{\|\cdot\|_2}(f)}{\sqrt{t}},$$

while the efficiency bound for MD_1 is

$$f(x^t) - \min_{x \in \Delta_n} f(x) \leq O(1) \frac{\sqrt{\ln n} L_{\|\cdot\|_1}(f)}{\sqrt{t}};$$

the ratio of these efficiency estimates is

$$R = O(1) \frac{L_{\|\cdot\|_2}(f)}{\sqrt{\ln n} L_{\|\cdot\|_1}(f)}.$$

Now, the ratio $L_{\|\cdot\|_2}(f)/L_{\|\cdot\|_1}(f)$ is always ≥ 1 and can be as large as $O(1)\sqrt{n}$ (in the case where all partial derivatives of f are of order of 1, and their sum is identically zero). It follows that for the problem of minimization over the standard simplex, *as far as the efficiency estimates are concerned*, the "non-Euclidean" Mirror Descent MD_1 can outperform the standard Subgradient Descent by a factor of order of $(n/\ln n)^{1/2}$, which, for large n, can make a huge difference.

4.3 $\mathbf{MD}_1$ and complexity of large-scale convex minimization over a simplex

We next show that the efficiency estimate of MD_1 as applied to minimization of Lipschitz continuous functions over an n-dimensional simplex cannot be improved by more than an $O(\ln n)$-factor, provided that n is large. Thus, MD_1 is a "nearly optimal" method, in the sense of Information-based Complexity theory, for large-scale convex minimization over the standard simplex.

Consider the family $\mathcal{F} \equiv \mathcal{F}(L, n)$ of all problems

$$f(x) \to \min \mid x \in \Delta_n \equiv \{x \in \mathbf{R}^n_+ : \sum_{i=1}^n x_i = 1\}$$

associated with convex functions $f : \Delta_n \to \mathbf{R}$ which are Lipschitz continuous and whose Lipschitz constant (taken w.r.t. $\|\cdot\|_1$) does not exceed a given positive L. The *Information-based Complexity* $\mathrm{Compl}(\varepsilon)$ *of the family* $\mathcal{F}$ is defined as follows. Let $\mathcal{B}$ be a routine which, as applied to a problem f from the family $\mathcal{F}$, successively generates *search points* $x_t = x_t(\mathcal{B}, f) \in \mathbf{R}^n$ and *approximate solutions* $x^t = x^t(\mathcal{B}, f)$; the only restriction on the mechanism of generating the search points and the approximate solutions is that both x_t and x^t should be deterministic functions of the values $f(x_\tau)$ and the subdifferentials $\partial f(x_\tau)$ of the objective taken at the previous search points x_τ, $\tau < t$, so that x_1, x^1 are independent of f, x_2, x^2 depend only on $f(x_1), \partial f(x_1)$, and so on. We define the *complexity of* $\mathcal{F}$ *w.r.t.* $\mathcal{B}$ as the function

$$\mathrm{Compl}_{\mathcal{B}}(\varepsilon) = \inf\{T : f(x^t(\mathcal{B}, f)) - \min_{\Delta_n} f \leq \varepsilon \quad \forall (t \geq T, f \in \mathcal{F}\},$$

i.e., as smallest number of steps after which the inaccuracy of approximate solutions generated by $\mathcal{B}$ is at most ε, whatever is $f \in \mathcal{F}$. The *complexity of the family* $\mathcal{F}$ is defined as

$$\mathrm{Compl}(\varepsilon) = \min_{\mathcal{B}} \mathrm{Compl}_{\mathcal{B}}(\varepsilon),$$

where the minimum is taken over all aforementioned "solution methods" $\mathcal{B}$. Note that the efficiency bound (21) says that

$$\mathrm{Compl}_{\mathrm{MD}_1}(\varepsilon) \leq O(1)\left[\frac{L^2 \ln n}{\varepsilon^2} + 1\right], \quad \varepsilon > 0. \tag{24}$$

On the other hand, the following statement takes place (for the proof, see Appendix 2):

Proposition 4.1 *The Information-based complexity of the family $\mathcal{F}(L, n)$ is at least* $O(1) \min\left[\frac{L^2}{\varepsilon^2}, n\right]$.

Comparing (24) with the lower complexity bound given by Proposition 4.1, we see that in the case of $\varepsilon \geq Ln^{-1/2}$ the accuracy guarantees given by MD_1 as applied to optimization problems from $\mathcal{F}$ cannot be improved by more than factor $O(\ln n)$.

5 Incremental Gradient version of the Mirror Descent scheme - the OSMD method

The objective function in the PET Image Reconstruction problem is a sum of a huge number m of simple convex functions. A natural way to exploit this fact in order to reduce the computational effort per iteration is offered by the *Incremental Gradient* technique (see e.g, [Ber95]), which in the medical imaging literature is known as the *Ordered Subsets* (OS) scheme (see [Hud94]).

The idea of the OS scheme is very simple: when solving problem (7) with the objective of the form

$$f(x) = \sum_{\ell=1}^{k} f_\ell(x), \tag{25}$$

one replaces at iteration t the "true" gradient $f'(x_t)$ with "partial gradient" $f'_{\ell(t)}(x_t)$, with $\ell(t)$ running, in the cyclic order, through the set $1, ..., k$ of indices of the components $f_1, ..., f_k$. With this approach, one reduces the computational effort required to compute f', and thus – reduces the complexity of an iteration. Computational practice in many cases demonstrates that such a modification does not affect much the quality of approximate solutions generated after a given number of iterations, provided that k is not too large.

Below, we present the Ordered Subsets version of the general Mirror Descent scheme and demonstrate that its convergence properties are similar to those of the original scheme.

The Ordered Subsets Mirror Descent scheme for solving problem (7) with objective of the form (25) (where all components f_ℓ are convex and Lipschitz continuous on X) has the same setup $(Y, X, \|\cdot\|, w, W)$ as the original gMD scheme and is as follows:

- *Initialization:* Choose $x_0 \in X$ and set $\xi_1 = w'(x_0)$;
- *Outer iteration t,$t = 1, 2, ...$:*

O.1) Given ξ_t, run a k-iteration *inner loop* as follows:

- Initialization: Set $\xi_t^1 = \xi_t$;
- Inner iteration ℓ, $\ell = 1, ..., k$:

  I.1) Given ξ_t^ℓ, compute

  $$\widehat{x}_t^\ell = W'(\xi_t^\ell); \quad x_t^\ell = \pi(\widehat{x}_t^\ell); \quad \eta_t^\ell = \eta(\widehat{x}_t^\ell)$$

  (cf. step S.1 in the original MD scheme).
  I.2) Compute the value $f_\ell(x_t^\ell)$ and a subgradient $f'_\ell(x_t^\ell)$ of f_ℓ at the point x_t^ℓ and set

  $$\xi_t^{\ell+1} = \xi_t^\ell - \gamma_t[f'_\ell(x_t^\ell) + \|f'_\ell(x_t^\ell)\|_* \eta_t^\ell],$$

  where $\gamma_t > 0$ is a stepsize.

O.2) Set

$$\xi_{t+1} = w'(W'(\xi_t^{m+1}))$$

and pass to Outer iteration $t + 1$.

- *Approximate solution* x^t generated in course of t steps of the method is the point $x^1_{\tau(t)}$, where

$$\tau(t) \in \operatorname*{Argmin}_{t/2 \le \tau \le t} \widetilde{f}_\tau, \quad \widetilde{f}_\tau = \sum_{\ell=1}^{k} f_\ell(x_\tau^\ell)$$

(note that $\widetilde{f}_\tau$ is a natural estimate of $f(x^1_\tau)$).

The main theoretical result of our paper summarizes the convergence properties of the Ordered Subsets version of the Mirror Descent scheme in the following

Theorem 5.1 *Assume that f_ℓ, $\ell = 1, ..., m$, are convex and Lipschitz continuous on X, with Lipschitz constants w.r.t. $\|\cdot\|$ not exceeding $L_{\|\cdot\|}(f)$, and that the subgradients $f'_\ell(x^\ell_t)$ used in the Mirror Descent method satisfy the condition*

$$\|f'_\ell(x^\ell_t)\|_* \leq L_{\|\cdot\|}(f).$$

Assume, in addition, that the $\|\cdot\|$-projector $\pi(\cdot)$ is Lipschitz continuous on Y, with a Lipschitz constant β w.r.t. $\|\cdot\|$, i.e.,

$$\|\pi_X(x) - \pi_X(x')\| \leq \beta\|x - x'\|, \quad \forall x, x' \in Y.$$

Then for every $t \geq 1$ one has

$$f(x^t) - \min_{x\in X} f(x) \leq \frac{\Gamma(w) + 2k(k+1)\beta\alpha^{-1}L^2_{\|\cdot\|}(f) \sum\limits_{t/2\leq\tau\leq t} \gamma^2_\tau}{\sum\limits_{t/2\leq\tau\leq t} \gamma_\tau} + 4k^2\beta\alpha^{-1}L^2_{\|\cdot\|}(f) \max_{t/2\leq\tau\leq t} \gamma_\tau. \quad (26)$$

In particular, whenever $\gamma_t \to +0$ and $\sum\limits_{t/2\leq\tau\leq t} \gamma_\tau \to \infty$ as $t \to \infty$, one has $f(x^t) - \min\limits_{x\in X} f(x) \to 0$ as $t \to \infty$. Moreover, with the stepsizes chose as

$$\gamma_t = \frac{(\alpha\beta^{-1}\Gamma(w))^{1/2}}{kL_t\sqrt{t}}, \quad (27)$$

where L_t are any numbers satisfying

$$0 < L_{\min} \leq L_t \leq L_{\max} < \infty,$$

one has

$$f(x^t) - \min_{x\in X} f(x) \leq O(1)k\sqrt{\frac{\beta\Gamma(w)}{\alpha}}\left(L_{\max} + \frac{L^2_{\|\cdot\|}(f)}{L_{\min}}\right)t^{-1/2}, \quad t = 1, 2, ... \quad (28)$$

Proof. 1^0. Let x_* be a minimizer of f on X, let $W_*(\xi) = W(\xi) - \xi^T x_*$, and let

$$g^\ell_\tau = f'_\ell(x^\ell_\tau), \quad h^\ell_\tau = g^\ell_\tau + \|g^\ell_\tau\|_*\eta^\ell_\tau.$$

Observe, first, that from $\|\eta^\ell_\tau\|_* \leq 1$ and $\|f'_\ell(x^\ell_\tau)\|_* \leq L_{\|\cdot\|}(f)$ it follows that

$$\|h^\ell_\tau\|_* \leq 2\|g^\ell_\tau\|_* \leq 2L, \quad L = L_{\|\cdot\|}(f), \quad (29)$$

whence

$$\|\xi^\ell_\tau - \xi^{\ell+1}_\tau\|_* \leq 2\gamma_\tau L.$$

Besides this, by (10) and by assumptions on $\pi(\cdot)$ and f_ℓ we have

$$\|W'(\xi) - W'(\eta)\| \leq \tfrac{1}{\alpha}\|\xi - \eta\|_* \quad \forall \xi, \eta \in \mathbf{R}^n,$$

$$\|\pi(x) - \pi(y)\| \leq \beta\|x - y\| \quad \forall x, y \in Y,$$

$$|f_\ell(x) - f_\ell(y)| \leq L\|x - y\| \quad \forall x, y \in X.$$

Combining these relations and taking into account the description of the method, we get

$$\begin{array}{ll} (a) & \|\widehat{x}^\ell_\tau - \widehat{x}^1_\tau\| \leq 2k\alpha^{-1}\gamma_\tau L,\ \ell = 1,...,k; \\ (b) & \|x^\ell_\tau - x^1_\tau\| \leq 2k\beta\alpha^{-1}\gamma_\tau L,\ \ell = 1,...,k; \\ (c) & |f_\ell(x^\ell_\tau) - f_\ell(x^1_\tau)| \leq 2k\beta\alpha^{-1}\gamma_\tau L^2,\ \ell = 1,...,k. \end{array} \tag{30}$$

2^0. Since W_* differs from W by a linear function, relation (10) holds true for W_* as well, whence

$$W_*(\xi+\eta) = W_*(\xi)+\eta^T W_*'(\xi)+\int\limits_0^1 [W_*'(\xi+t\eta)-W_*'(\xi)]^T\eta dt \leq W_*(\xi)+\eta^T W_*'(\xi)+\frac{1}{2\alpha}\|\eta\|_*^2. \tag{31}$$

Besides this, whenever $\xi \in \mathbf{R}^n$, we have

$$W'(\xi) = \mathrm{argmax}_{x\in Y}[\xi^T x - w(x)],$$

and since w is continuously differentiable on Y, it follows that

$$[\xi - w'(W'(\xi))]^T(W'(\xi) - y) \geq 0 \quad \forall y \in Y.$$

It follows that

$$\begin{array}{rcl} W_*(\xi) & = & W(\xi) - \xi^T x_* = \xi^T W'(\xi) - w(W'(\xi)) - \xi^T x_* \\ & = & [w'(W'(\xi))]^T W'(\xi) - w(W'(\xi)) + [\xi - w'(W'(\xi))]^T(W'(\xi) - x_*) - [w'(W'(\xi))]^T x_* \\ & \geq & [w'(W'(\xi))]^T W'(\xi) - w(W'(\xi)) - [w'(W'(\xi))]^T x_* \\ & = & W_*(w'(W'(\xi))). \end{array} \tag{32}$$

We now have

$$\begin{array}{rll} & W_*(\xi^{\ell+1}_\tau) = W_*(\xi^\ell_\tau - \gamma_\tau h^\ell_\tau) & \\ \leq & W_*(\xi^\ell_\tau) - \gamma_\tau[h^\ell_\tau]^T W_*'(\xi^\ell_\tau) + \frac{1}{2\alpha}\gamma_\tau^2\|h^\ell_\tau\|_*^2 & \text{[by (31)]} \\ \leq & W_*(\xi^\ell_\tau) - \gamma_\tau[h^\ell_\tau]^T W_*'(\xi^\ell_\tau) + \frac{2}{\alpha}\gamma_\tau^2 L^2 & \text{[by (29)]} \\ = & W_*(\xi^\ell_\tau) - \gamma_\tau[h^\ell_\tau]^T[\widehat{x}^\ell_\tau - x_*] + \frac{2}{\alpha}\gamma_\tau^2 L^2 & \\ = & W_*(\xi^\ell_\tau) + \frac{2}{\alpha}\gamma_\tau^2 L^2 + \gamma_\tau[h^\ell_\tau]^T[x_* - \widehat{x}^\ell_\tau] & \\ = & W_*(\xi^\ell_\tau) + \frac{2}{\alpha}\gamma_\tau^2 L^2 + \gamma_\tau[f_\ell'(x^\ell_\tau)]^T[x_* - \widehat{x}^\ell_\tau] + \gamma_t\|f_\ell'(x^\ell_\tau)\|_*[\eta^\ell_\tau]^T[x_* - \widehat{x}^\ell_\tau]. & \end{array}$$

The last term here is $\leq -\|\widehat{x}_\tau^\ell - x_\tau^\ell\|$ by (9), so that

$$\begin{array}{rcl} W_*(\xi_\tau^{\ell+1}) & \leq & W_*(\xi_\tau^\ell) + \frac{2}{\alpha}\gamma_\tau^2 L^2 - \gamma_t \|f'_\ell(x_\tau^\ell)\|_* \|\widehat{x}_\tau^\ell - x_\tau^\ell\| + \gamma_\tau [f'_\ell(x_\tau^\ell)]^T [x_* - \widehat{x}_\tau^\ell] \\ & = & W_*(\xi_\tau^\ell) + \frac{2}{\alpha}\gamma_\tau^2 L^2 - \gamma_t \|f'_\ell(x_\tau^\ell)\|_* \|\widehat{x}_\tau^\ell - x_\tau^\ell\| \\ & & + \gamma_\tau [f'_\ell(x_\tau^\ell)]^T [x_* - x_\tau^\ell] + \gamma_\tau [f'_\ell(x_\tau^\ell)]^T [x_\tau^\ell - \widehat{x}_\tau^\ell] \end{array}$$

The last term here is $\leq \|f'_\ell(x_\tau^\ell)\|_* \|\widehat{x}_\tau^\ell - x_\tau^\ell\|$, whence

$$\begin{array}{rcll} W_*(\xi_\tau^{\ell+1}) & \leq & W_*(\xi_\tau^\ell) + \frac{2}{\alpha}\gamma_\tau^2 L^2 + \gamma_\tau [f'_\ell(x_\tau^\ell)]^T [x_* - x_\tau^\ell] & \\ & \leq & W_*(\xi_\tau^\ell) + \frac{2}{\alpha}\gamma_\tau^2 L^2 + \gamma_\tau [f_\ell(x_*) - f_\ell(x_\tau^\ell)] & \text{[convexity of } f_\ell] \\ & \leq & W_*(\xi_\tau^\ell) + \frac{2}{\alpha}\gamma_\tau^2 L^2 + \gamma_\tau [f_\ell(x_*) - f_\ell(x_\tau^1)] + \gamma_\tau [f_\ell(x_\tau^1) - f_\ell(x_\tau^\ell)]. & \end{array}$$

Since the last term is $\leq 2k\beta\alpha^{-1}\gamma_\tau L^2$ by (30.c), we come to

$$W_*(\xi_\tau^{\ell+1}) \leq W(\xi_\tau^\ell) - \gamma_\tau [f_\ell(x_\tau^1) - f_\ell(x_*)] + 2(k+1)\beta\alpha^{-1}\gamma_\tau^2 L^2.$$

Adding up these inequalities for $\ell = 1, ..., k$, we conclude that

$$W_*(\xi_\tau^{k+1}) \leq W(\xi_\tau^1) - \gamma_\tau [f(x_\tau^1) - f(x_*)] + 2k(k+1)\beta\alpha^{-1}\gamma_\tau^2 L^2.$$

Since $\xi_\tau^1 = \xi_\tau$ and $\xi_{\tau+1} = w'(W'(\xi_\tau^{k+1}))$, the latter inequality, by (32), implies that

$$W_*(\xi_{\tau+1}) \leq W_*(\xi_\tau) - \gamma_\tau [f(x_\tau^1) - f(x_*)] + 2k(k+1)\beta\alpha^{-1}\gamma_\tau^2 L^2. \tag{33}$$

Summing up the resulting inequalities over τ, $t/2 \leq \tau \leq t$, and denoting by $\bar{t}$ the smallest value of τ in this range, we get

$$\begin{array}{rcl} \left[\min\limits_{\bar{t} \leq \tau \leq t} f(x_\tau^1) - f(x_*)\right] \sum\limits_{\tau=\bar{t}}^{t} \gamma_\tau & \leq & \sum\limits_{\tau=\bar{t}}^{t} \gamma_\tau [f(x_\tau^1) - f(x_*)] \\ & \leq & W_*(\xi_{\bar{t}}) - W_*(\xi_{t+1}) + 2k(k+1)\beta\alpha^{-1}L^2 \sum\limits_{\tau=\bar{t}}^{t} \gamma_\tau^2. \end{array} \tag{34}$$

Now, since W is the Legendre transformation of $w\Big|_Y$ and $x_* \in X \subset Y$, we have $W_*(\xi_{t+1}) = W(\xi_{t+1}) - \xi_{t+1}^T x_* \geq -w(x_*)$, while, by construction, $\xi_{\bar{t}} = w'(y_{\bar{t}})$ for certain $y_{\bar{t}} \in Y$. It follows that $W_*(\xi_{\bar{t}}) = [w'(y_{\bar{t}})]^T y_{\bar{t}} - w(y_{\bar{t}}) - [w'(y_{\bar{t}})]^T x_*$, whence

$$W_*(\xi_{\bar{t}}) - W_*(\xi_{t+1}) \leq w(x_*) - \Big[w(y_{\bar{t}}) + [w'(y_{\bar{t}})]^T (x_* - y_{\bar{t}})\Big] \leq \Gamma(w).$$

Thus, (34) implies that

$$\min_{\bar{t} \leq \tau \leq t} f(x_\tau^1) - f(x_*) \leq \frac{\Gamma(w) + 2k(k+1)\beta\alpha^{-1}L^2 \sum\limits_{\tau=\bar{t}}^{t} \gamma_\tau^2}{\sum\limits_{\tau=\bar{t}}^{t} \gamma_\tau}. \tag{35}$$

At the same time, from (30.c) it follows that whenever $t \geq \tau \geq t/2$, one has

$$|\widetilde{f}_\tau - f(x^1_\tau)| \leq 2k^2\beta\alpha^{-1}L^2 \max_{t/2\leq\tau\leq t} \gamma_\tau \qquad [\widetilde{f}_\tau = \sum_{\ell=1}^m f_\ell(x^\ell_\tau))]$$

Taking into account the latter inequality, the inequality (30) and the rule for generating x^t, we come to (26).

The remaining statements of Theorem 5.1 are straightforward consequences of (26). ■

Remark 5.1 The theoretical efficiency estimate of OSMD stated by Theorem 5.1 is not better (in fact, it is larger, by a factor $O(k\beta^{1/2})$) than the estimate stated in Theorem 4.1 for gMD. The advantage of the Ordered Subsets techniques is a matter of practical experience in several difficult application areas (e.g., Training of Neural Nets ([Ber97]) and Tomography ([Hud94])). In this regard, the role of Theorem 5.1 is to make the approach legitimate theoretically.

5.1 Ordered Subsets implementation of MD$_1$

From now on, we focus on problem (7) with objective of the form (25), and assume that X is the standard n-dimensional simplex Δ_n. Our current goal is to complete the description of the associated Ordered Subsets version of MD$_1$. The only elements which still are missing are the calculation of the projector

$$\pi(x) \equiv \pi_p(x) = \operatorname*{argmin}_{y\in\Delta_n} \|x-y\|_p,$$

of the separator $\eta(x)$ and an explicit upper bound on the Lipschitz constant of this projector w.r.t. $\|\cdot\|_p$-norm, i.e., on the quantity

$$\beta(p) = \sup_{x,x'\in\mathbf{R}^n, x\neq x'} \frac{\|\pi_p(x)-\pi_p(x')\|_p}{\|x-x'\|_p}.$$

The required information is provided by the following result:

Proposition 5.1 *Let $1 < p < \infty$. Then*

(i) *The projector $\pi_p(x)$ is independent of p, and is given component-wise by*

$$(\pi_p(x))_j = (x_j + \lambda(x))_+, \ j = 1, ..., n, \qquad [a_+ = \max[0,a]] \tag{36}$$

where $\lambda(x)$ is the unique root of the equation

$$\sum_{j=1}^n (x_j + \lambda)_+ = 1. \tag{37}$$

In particular, $\pi_p(x)$, for every $p > 1$, is also a $\|\cdot\|_1$-projector of $\mathbf{R}^n$ onto Δ_n:

$$\pi_p(x) \in \operatorname{Argmin}\{\|x-y\|_1 : y \in \Delta_n\}.$$

The separator $\eta_p(x)$:

$$\|\eta_p(x)\|_q \leq 1, \ \eta_p^T(x)(x-y) \geq \|x-\pi_p(x)\|_p \quad \forall y \in X \qquad \left[q = \tfrac{p}{p-1}\right]$$

is readily given by $\pi_p(x)$*:*

$$\begin{array}{lll} x \in X & \Rightarrow & \eta_p(x) = 0; \\ x \notin X & \Rightarrow & \eta_p(x) = [\nabla \|z\|_p]_{z=x-\pi_p(x)} = \left\{ \frac{|\delta_i|^{p-1}\,\mathrm{sign}(\delta_i)}{\|\delta\|_p^{p-1}} \right\}_{i=1}^n, \ \delta = x - \pi_p(x). \end{array} \tag{38}$$

(ii) $\beta(p) \le 2$.

Proof. 0^0. Relation (38) is evident, since $\|\cdot\|_p$ is continuously differentiable outside of the origin for $p > 1$.

1^0. Let us verify first that $\pi_p(x)$ is indeed given by (36) and thus is independent of p. There is nothing to prove when $x \in \Delta_n$ (in this case the unique root of (37) is $\lambda(x) = 0$, and (36) says correctly that $\pi_p(x) = x$). Now let $x \notin \Delta_n$. It is immediately seen that $\lambda(x)$ is well-defined; let y be the vector with the coordinates given by the right hand side of (36). This vector clearly belongs to Δ_n, and the vector $d = y - x$ is as follows: there exists a nonempty subset J of the index set $\{1, ..., n\}$ such that $d_j = \lambda(x)$ for $j \in J$ and $d_j < \lambda(x)$ and $y_j = 0$ for $j \notin J$. In order to verify that y is the $\|\cdot\|_p$-projection of x onto Δ_n, it suffices to prove that if $\delta = \left.\frac{\partial \|z\|_p}{\partial z}\right|_{z=d}$, then the linear form $\delta^T u$ attains its minimum over $u \in \Delta_n$ at the point y. We have

$$\delta_j = \theta |d_j|^{p-1}\,\mathrm{sign}(d_j), \ j = 1, ..., n \quad [\theta > 0],$$

i.e., same as for the vector d itself, for certain μ it holds $\delta_j = \mu$, $j \in J$ and $\delta_j < \mu$, $y_j = 0$ for $j \notin J$, so that the linear form $\delta^T u$ indeed attains its minimum over $u \in \Delta_n$ at the point y.

2^0. Now let us prove that $\beta(p) \le 2$. Observe that $\pi_p(x)$ is Lipschitz continuous (since $\pi_p(\cdot)$ is independent of p, and the $\|\cdot\|_2$-projector onto a closed convex set is Lipschitz continuous, with constant 1, w.r.t. $\|\cdot\|_2$).

2^0.1. Let $J(x) = \{j \mid x_j + \lambda(x) \ge 0\}$, and let $k(x)$ be the cardinality of $J(x)$. Since $\lambda(x)$ solves (37), we have $k(x) \ge 1$ and $\lambda(x) = \frac{1}{k(x)}\left[1 - \sum\limits_{j \in J(x)} x_j\right]$. Denoting by $e(x)$ the characteristic vector of the set $J(x)$ and by $E(x)$ the matrix $\mathrm{Diag}(e(x))$, we therefore get

$$\pi_p(x) = E(x)x + \frac{1}{k(x)} e(x) - \frac{1}{k(x)} e(x) e^T(x) x. \tag{39}$$

Let $\mathcal{J}$ be the set of all nonempty subsets of the index set $\{1, ..., n\}$, and let $X[J] = \{x \mid J(x) = J\}$ for $J \in \mathcal{J}$. From (39) it follows that for every $J \in \mathcal{J}$ we have

$$\begin{array}{rcl} & & x, y \in X[J] \Rightarrow \\ \|\pi_p(x) - \pi_p(y)\|_p & \le & \|E(x)(x-y)\|_p + \frac{1}{k(x)} \|e(x)e^T(x)(x-y)\|_p \\ & \le & \|x - y\|_p + \frac{1}{k(x)} \|e(x)\|_p \|e(x)\|_{\frac{p}{p-1}} \|x - y\|_p \\ & = & 2\|x - y\|_p. \end{array} \tag{40}$$

2^0.2. Let

$$\mathcal{X} = \{x \in \mathbf{R}^n : \exists (J \in \mathcal{J}, j \le n) : \mathrm{Card}(J) x_j = \sum_{j' \in J} x_{j'} - 1\}.$$

Note that $\mathcal{X}$ is the union of finitely many hyperplanes. We claim that if $x, y \in \mathbf{R}^n$ are such that the segment $[x, y]$ does not intersect $\mathcal{X}$, then $J(x) = J(y)$ and, consequently (see (40)),

$$\|\pi_p(x) - \pi_p(y)\|_p \leq 2\|x - y\|_p. \tag{41}$$

Indeed, assume that $J(x) \neq J(y)$, or, which is the same, the sets $\{j : x_j \geq -\lambda(x)\}$ and $\{j : y_j \geq -\lambda(y)\}$ are distinct from each other. Since $\lambda(\cdot)$ clearly is continuous, it follows that on the segment $[x, y]$ there exists a point $\bar{x}$ such that one of the coordinates of the point equals to $-\lambda(\bar{x})$, i.e., to $\frac{1}{k(\bar{x})}\left[\sum\limits_{j' \in J(\bar{x})} \bar{x}_{j'} - 1\right]$. In other words, $\bar{x} \in \mathcal{X}$, which contradicts the assumption.

2^0.3. Now let $y, y' \in \mathbf{R}^n \backslash \mathcal{X}$. Since $\mathcal{X}$ is a union of finitely many hyperplanes, the segment $[y, y']$ can be partitioned into subsequent segments $[y, y_1]$, $[y_1, y_2]$,..., $[y_s, y']$ in such a way that the interior of every segment of the partition does not intersect $\mathcal{X}$. By the result of 2^0.2, $\pi_p(\cdot)$ is Lipschitz continuous with constant 2 w.r.t. $\|\cdot\|_p$ on the interiors of the above segments. Since $\pi_p(\cdot)$, as we just have mentioned, is continuous, it follows that

$$\|\pi_p(y) - \pi_p(y')\|_p \leq 2\|y - y'\|_p.$$

The latter relation holds true for all pairs $y, y' \in \mathbf{R}^n \backslash \mathcal{X}$, i.e., for all pairs from a set which is dense in $\mathbf{R}^n$; since $\pi_p(\cdot)$ is continuous, this relation in fact holds for all y, y'. ∎

Remark 5.2 *The upper bound 2 on $\beta(p)$ cannot be improved, unless one restricts the range of values of p and/or values of n. Indeed, the $\|\cdot\|_p$-distance from the origin to a vertex of Δ_n is 1, while the $\|\cdot\|_p$-distance between the $\|\cdot\|_p$-projections of these points onto Δ_n, i.e., the $\|\cdot\|_p$-distance from a vertex to the barycenter of Δ_n, is $\left(\frac{n-1}{n^p} + \left(\frac{n-1}{n}\right)^p\right)^{1/p}$; when n is large and p is close to 1, the latter quantity is close to 2.*

We see that to project onto Δ_n is easy: computation of $\pi(x)$ requires, basically, the same effort as ordering the coordinates of x, which can be done in time $O(n \ln n)$.

6 Implementation and testing

In this section, we present results of the Mirror Descent method as applied to the PET image reconstruction problem based on several sets of simulated and real clinical data. We compare the results obtained by OSMD$_1$ and MD$_1$. In addition, we compare the results of MD to those of the usual Subgradient Descent method.

6.1 Implementation of the algorithms

In our experiments, we have worked with several sets of tomography data. Each data set gives rise to a particular optimization problem of the form of (8) which was solved by the Mirror Descent scheme (in both the usual and the Ordered Subset versions). The setup for Mirror Descent was

$$Y = \{x \mid \|x\|_p \leq 1\} \; [\supset \Delta_n], \quad \|\cdot\| = \|\cdot\|_1, \quad w(x) = \frac{1}{2}\|x\|_p^2, \quad p = p(n) = 1 + \frac{1}{\ln n}.$$

This setup differs from (15) – (20) by setting $\|\cdot\| = \|\cdot\|_1$ instead of $\|\cdot\| = \|\cdot\|_{p(n)}$; with the above $p(n)$, this modification does not affect the theoretical efficiency estimate of the algorithm.

The indicated setup defines the algorithm up to the stepsize policy. The latter for the "No Ordered Subsets" version MD of the method was chosen as (cf. (22)):

$$\gamma_t = \frac{C}{\|f'(x_T)\|_\infty \sqrt{\ln n}\sqrt{t}}$$

with $C = 0.03$ (this value of the stepsize factor C was found the best one in our preliminary experiments and never was changed afterwards).

The Ordered Subsets version OSMD of the method uses 24-component representation (25) of the objective, the components being partial sums of the terms in the sum (8), with $m/24$ subsequent terms in every one of the partial sums. The stepsizes here were chosen according to the rule (cf. (27))

$$\gamma_t = \frac{C}{24L_t\sqrt{t}\sqrt{\ln n}},$$

where L_t is a current guess for the $\|\cdot\|_1$-Lipschitz constant of the objective; in our implementation, this guess, starting with the second outer iteration, was defined as $\sum\limits_{1\leq\ell\leq 24} \|f'_\ell(x^\ell_{t-1})\|_\infty$. The stepsize factor C in OSMD was set to 0.3.

In our experiments we have used, as a "reference point", the standard Subgradient Descent method (23) (in the usual "No subsets") version with the "theoretical" stepsize policy

$$\gamma_t = \frac{C}{\|f'(x_t)\|_2\sqrt{t}}.$$

The stepsize factor C was tuned to get as good reconstruction as possible; the resulting "optimal value" of it turned out to be 0.006.

The starting point x_0 in all our runs was the barycenter of the simplex Δ_n.

Measuring quality of reconstructions. In Medical Imaging, the standard way to evaluate the quality of a reconstruction algorithm is to apply the algorithm to simulated data and to check how the resulting pictures reproduce important for a particular application elements of the true image (in tomography, these elements could be, e.g., small areas with high density of the tracer mimicking tumors). In what follows we combine this, basically qualitative, way of evaluation with a quantitative one, where the quality of the approximate solution x^t to (8) yielded after t steps of the method is measured by the quantity $\varepsilon_t = f(x^t) - \min\limits_{\Delta_n} f$. Note that this quantity is not "observable" (since the true optimal value $f_* = \min\limits_{\Delta_n} f$ is unknown). We can, however, easily compute a *lower bound* on f_*. Assume, e.g., that we have run a "no subset" version of the method, and in course of computations have computed the values $f(x_t)$ and subgradients $f'(x_t)$ of the objective at N search points x_t, $1 \leq t \leq N$. Then we can build the standard piecewise-linear minorant $f^N(\cdot)$ of our objective:

$$f^N(x) = \max_{1\leq t\leq N} \left[[f(x_t) - x_t^T f'(x_t)] + x^T f'(x_t)\right] \leq f(x).$$

The quantity $f_*^N \equiv \min\limits_{x\in\Delta_n} f^N(x)$ clearly is a lower bound on f_*, so that the "observable" quantities

$$\widehat{\varepsilon}_t = f(x_t) - f_*^N,\ 1 \leq t \leq N$$

are upper bounds on the actual inaccuracies ε_t. In our experiments, the bound f_*^N was computed at the post-optimization phase according to the relation

$$f_*^N \equiv \min_{x\in\Delta_n}\max_{t\le N}\left[[f(x_t)-x_t^Tf'(x_t)]+x^Tf'(x_t)\right] = \max_{\lambda\in\Delta_N}\phi(\lambda),$$
$$\phi(\lambda)\equiv \min_{x\in\Delta_n}\sum_{t=1}^N \lambda_t\left[\underbrace{[f(x_t)-x_t^Tf'(x_t)]}_{d_t}+x^Tf'(x_t)\right] = \sum_{t=1}^N\lambda_t d_t + \min_{j\le n}[\sum_{t=1}^N \lambda_t f'(x_t)]_j,$$

which reduces the computation of f_*^N to maximizing a convex function $\phi(\lambda)$ of N variables. In our experiments, the total number of iterations N was just 10, and there was no difficulty in minimizing ϕ.

In the Ordered Subsets version of the method, the policy for bounding f_* from below was similar: here after N outer iterations we know the values and the subgradients of the components f_ℓ, $\ell=1,...,k$, in decomposition (25) along the points x_t^ℓ, $t=1,...,N$. This allows to build a piecewise minorant

$$f^N(x)=\sum_{\ell=1}^k \max_{t=1,...,N}\left[[f_\ell(x_t^\ell)-[x_t^\ell]^Tf'_\ell(x_t^\ell)]+x^Tf'_\ell(x_t^\ell)\right]$$

of the objective and to use, as the lower bound on f_*, the quantity

$$f_*^N \equiv \min_{x\in\Delta_n} f^N(x) = \max_{\mu}\left\{\psi(\mu):\mu=\{\mu_{t\ell}\}\ge 0, \sum_t \mu_{t\ell}=1,\ \ell=1,...,k\right\},$$
$$\psi(\mu)\equiv \min_{x\in\Delta_n}\sum_{t,\ell}\mu_{t\ell}\left[\underbrace{[f_\ell(x_t^\ell)-[x_t^\ell]^Tf'_\ell(x_t^\ell)]}_{d_{t\ell}}+x^Tf'_\ell(x_t^\ell)\right] = \sum_{t,\ell}\mu_{t\ell}d_{t\ell}+\min_j\left[\sum_{t,\ell}\mu_{t\ell}f'_\ell(x_t^\ell)\right]_j.$$

6.2 Results

We tested the algorithms on 5 sets of tomography data; the first four are simulated scans of *phantoms* (artificial bodies), obtained from the Eidolon simulator ([Zai98], [Zai99]) of `PRT-1` PET-scanner. The phantoms (`Cylinder`, `Utah`, `Spheres`, `Jaszczak`) are 3D cylinders with piecewise constant density of the tracer; they are commonly used in Tomography to test the effectiveness of scanners and reconstruction methods (for more details, see [Thi99]). The fifth data set `Brain` is obtained from the `GE Advance` PET-scanner in an actual brain study.

All experiments were carried out on the INTEL Marlinspike Windows NT Workstation (500 MHz 1Mb Cache INTEL Pentium III Xeon processor, 2GB RAM). A single outer iteration of OSMD takes nearly the same time as a single iteration of MD, namely, appr. 2 min in each of the four "phantom" tests ($n=515,871, m=3,170,304$), and appr. 90 min in the `Brain` test ($n=2,763,635, m\approx 25,000,000$). About 95% of the running time is used to compute the value and the gradient of the objective.

Our numerical results are summarized in Table 1.

Note that in OSMD there is no necessity to compute the true values of the objective along the iterates x_t^ℓ, and an attempt to compute these values would increase the execution time by factor k. For the sake of this paper we, however, did compute the values $f(x_t^1)$.

A more detailed description of the data and the results is as follows.

Table 1: Objective values along iterations (for OSMD, $x_t = x_t^1$)

Itr#	`Cylinder` $f(x_t) \times 10^{-8}$		`Utah` $f(x_t) \times 10^{-8}$		`Spheres` $f(x_t) \times 10^{-7}$		`Jazszak` $f(x_t) \times 10^{-7}$		`Brain` $f(x_t) \times 10^{-9}$	
	MD	OSMD	MD	OSMD	MD	OSMD	MD	OSMD	MD	OSMD
1	-2.382	-2.382	-2.549	-2.549	-4.295	-4.295	-5.021	-5.021	-1.463	-1.463
2	-2.648	-2.725	-2.807	-2.902	-4.767	-5.132	-5.643	-5.908	-1.725	-1.848
3	-2.708	-2.732	-2.890	-2.926	-5.079	-5.191	-5.867	-5.968	-1.867	-2.001
4	-2.732	-2.732	-2.929	-2.939	-5.189	-5.200	-5.970	-6.000	-1.951	-2.012
5	-2.723	-2.734	-2.917	-2.938	-5.168	-5.212	-5.950	-5.988	-1.987	-2.015
6	-2.738	-2.738	-2.943	-2.937	-5.230	-5.216	-6.001	-6.005	-1.978	-2.015
7	-2.727	-2.740	-2.923	-2.936	-5.181	-5.205	-5.967	-5.991	-1.997	-2.016
8	-2.740	-2.742	-2.942	-2.936	-5.227	-5.218	-6.007	-6.005	-2.008	-2.016
9	-2.731	-2.737	-2.925	-2.937	-5.189	-5.212	-5.974	-5.994	-1.999	-2.016
10	-2.741	-2.741	-2.941	-2.937	-5.225	-5.205	-6.030	-6.002	-2.009	-2.016
Lower bound	-2.754		-2.966		-5.283		-6.093		-2.050	

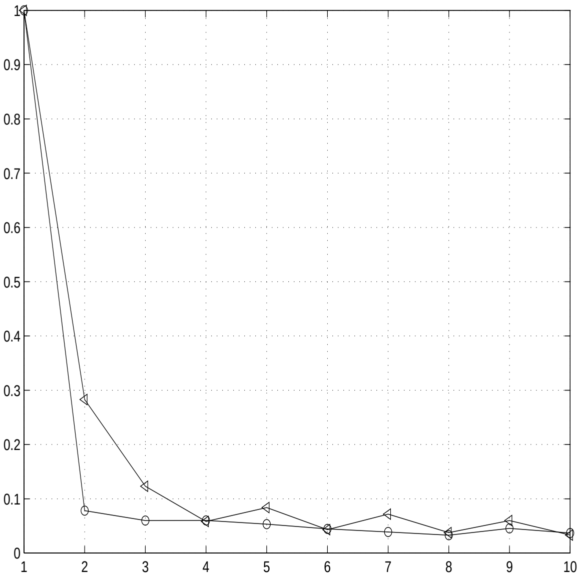

Figure 1. `Cylinder`, progress in accuracy: plot of $\theta(t) = \frac{f(x_t)-f_*^{10}}{f(x_1)-f_*^{10}}$ $\left[\geq \frac{f(x_t)-f_*}{f(x_1)-f_*}\right]$

$\triangle$ – MD; o – OSMD

"Cylinder" $(n = 515,871, m = 3,170,304)$**.** This phantom is a cylinder with a uniform density of the tracer. Fig. 1 displays the "progress in accuracy" in the experiment.

"Utah" $(n = 515,871, m = 3,170,304)$**.** This phantom (see Fig. 2) is a pair of co-axial cylinders with 2 vertical tubes in the inner cylinder, and the density of the tracer is high between the cylinders and in one of the tubes, low in the other tube and is moderate within the inner cylinder outside the tubes. The phantom allows to test the ability of an algorithm to reconstruct the borders between areas with different densities of the tracer and the ratios of these densities. Fig. 3 displays the "progress in accuracy".

In clinical applications, the yield of a reconstruction algorithm is a collection of *slices* – pictures of different 2D cross-sections of the resulting 3D image. To give an idea of the quality of our reconstructions, Fig. 4 represents their slices (the cross-sections of the outer cylinder by a plane orthogonal to its axis); in all our pictures, white corresponds to high, and black – to low density of the tracer.

"Spheres" $(n = 515,871, m = 3,170,304)$**.** This phantom is a cylinder containing 6 spheres of different radii centered at the mid-slice of the cylinder. The density of the tracer is high within the spheres and low outside of them. The mid-slice of the phantom is shown on Fig. 5. The phantom is used to test tumor detection capability, mainly for torso studies.

Fig. 6 displays the "progress in accuracy". The mid-slices of our 3D reconstructions are shown on Fig. 7. The `Spheres` experiment clearly demonstrates the advantages of the $\|\cdot\|_1$-Mirror Descent as compared to the usual Subgradient Descent. The best progress in accuracy we were able to get with SD was to reduce in 10 iterations the initial residual in the objective by factor 5.26, which is 3.5 times worse than the similar factor (18.51) for MD. What is much more

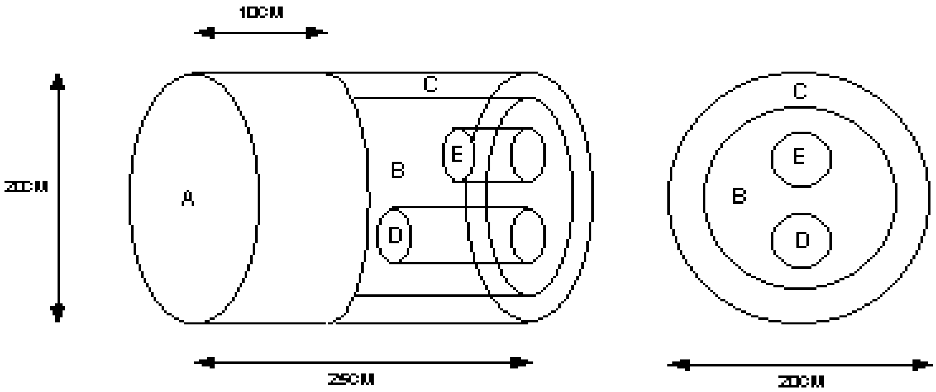

Figure 2. The `Utah` phantom

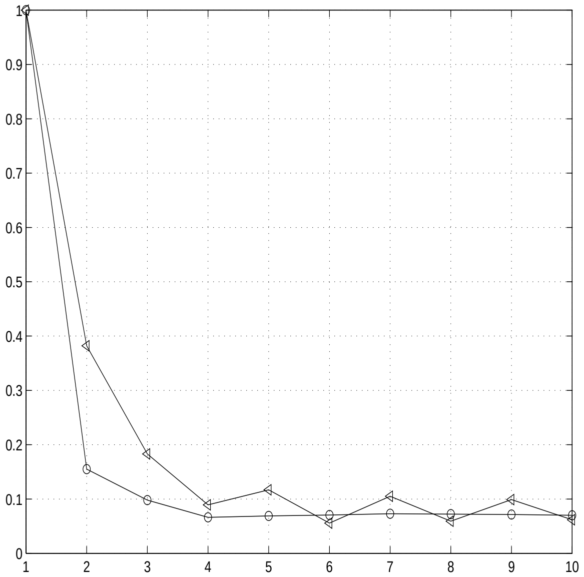

Figure 3. `Utah`, progress in accuracy: plot of $\theta(t) = \frac{f(x_t)-f_*^{10}}{f(x_1)-f_*^{10}}$ $\left[\geq \frac{f(x_t)-f_*}{f(x_1)-f_*}\right]$

$\triangle$ – MD; o – OSMD

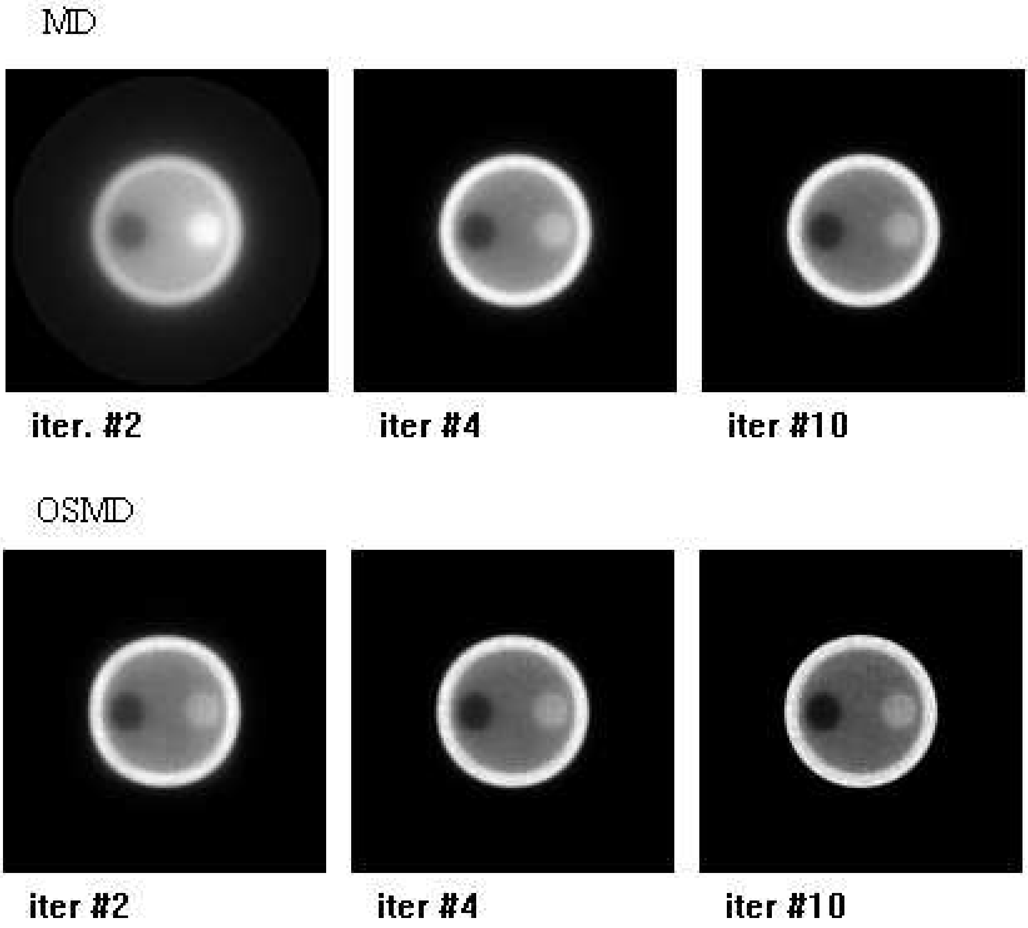

Figure 4. `Utah`, near-top slice of the reconstruction.

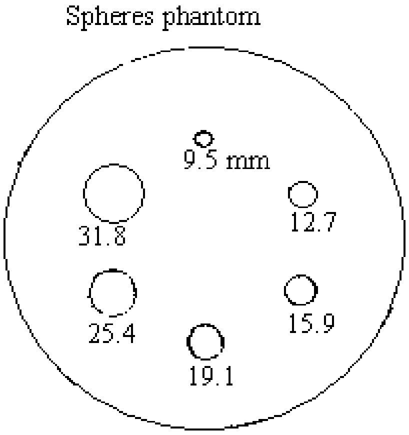

Figure 5. Mid-slice of the `Spheres` phantom

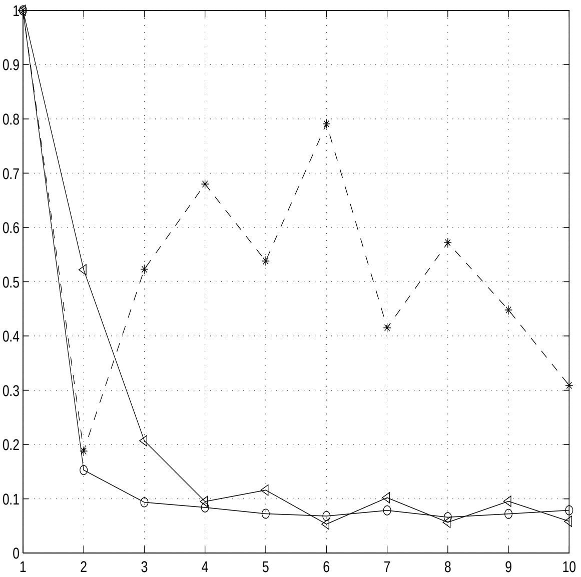

Figure 6. `Spheres`, progress in accuracy: plot of $\theta(t) = \frac{f(x_t)-f_*^{10}}{f(x_1)-f_*^{10}}$ $\left[\geq \frac{f(x_t)-f_*}{f(x_1)-f_*}\right]$

Δ – MD; o – OSMD; * – Subgradient Descent

dangerous from the clinical viewpoint, is that the reconstructions given by SD can be heavily affected by artifacts, as can be seen from Fig. 8.

“Jaszczak” ($n = 515,871, m = 3,170,304$). This phantom is a cylinder containing a number of vertical tubes of different cross-sections. The density of the tracer is high outside of the tubes and is zero inside them. The mid-slice of the phantom is shown on Fig. 9. The number and the sizes of tubes “recognized” by a reconstruction algorithm allow to quantify the resolution of the algorithm.

Fig. 10 displays the “progress in accuracy”. The mid-slices of our 3D reconstructions are shown on Fig. 11. The `Jaszczak` experiment clearly demonstrates the advantages of OSMD as compared to MD. We see that the quality of the image after just two outer iterations of OSMD is at leas as good as the one obtained after four iterations of MD. Likewise, four iterations of OSMD result in an image comparable to the one obtained by MD in ten iterations.

“Brain” ($n = 2,763,635, m \approx 25,000,000$). This data is an actual clinical brain study of a patient with the Alzheimer disease.

Fig. 12 displays the “progress in accuracy”. The mid-slices of our 3D reconstructions are shown on Fig. 13. The `Brain` experiment again demonstrates the advantages of OSMD as compared to MD. Indeed, OSMD produced in 4 iterations an image which is as good the one produced after 10 iterations of MD.

The quality of our reconstructions compares favourably with the one given by the commercially used algorithms (based of filtered back-projection). As compared to the “golden standard” of the new generation of 3D imaging algorithms – the so-called OSEM (Ordered Subset Expectation Maximization) algorithm, OSMD is highly competitive both in image quality and

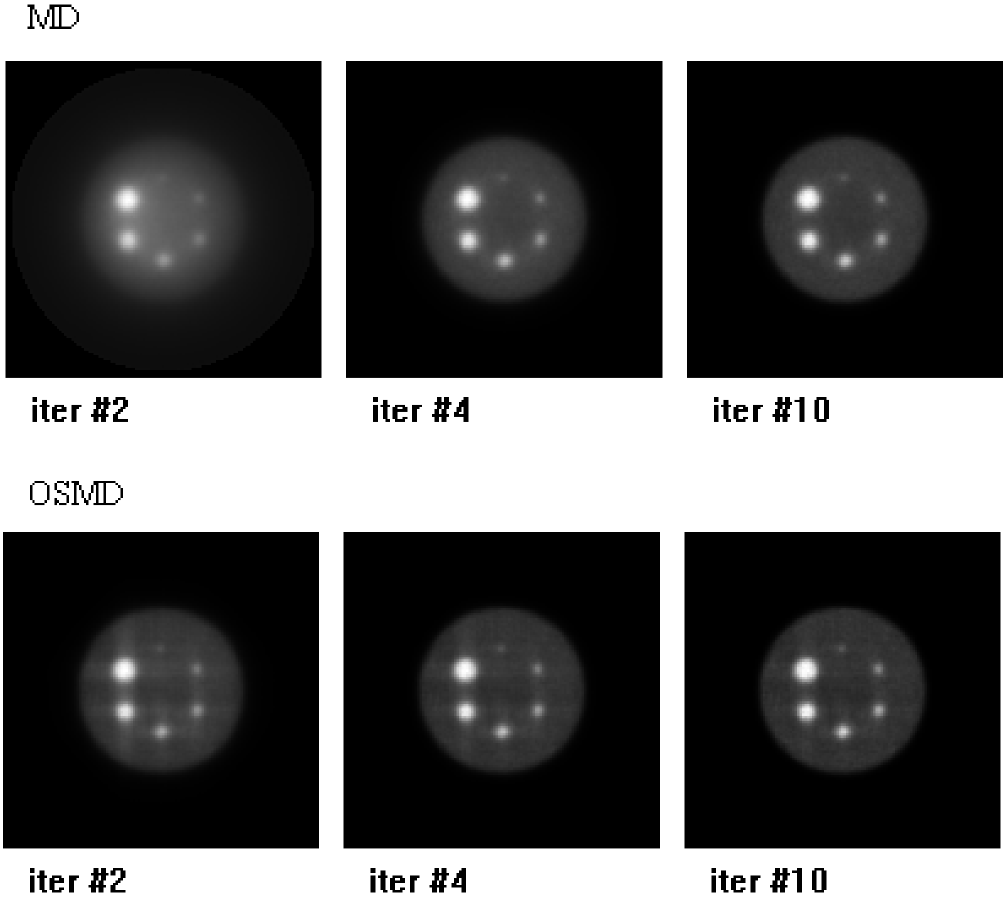

Figure 7. `Spheres`, mid-slice of the reconstructions.

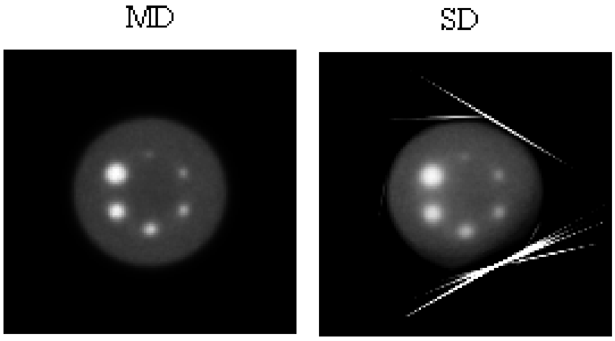

Figure 8. `Spheres`, mid-slice of the SD reconstruction after 10 iterations

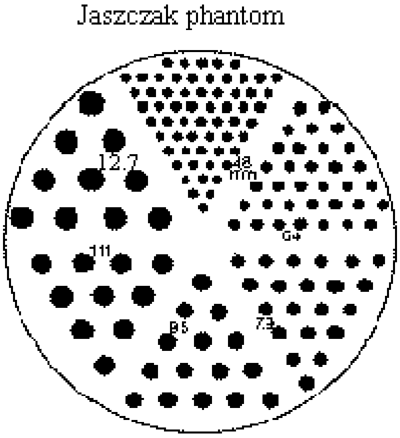

Figure 9. Mid-slice of the `Jaszczak` phantom

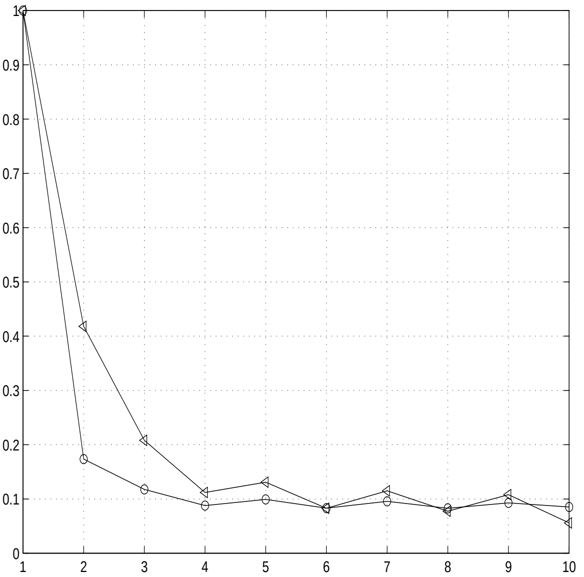

Figure 10. `Jaszczak`, progress in accuracy: plot of $\theta(t) = \frac{f(x_t)-f_*^{10}}{f(x_1)-f_*^{10}} \quad \left[\geq \frac{f(x_t)-f_*}{f(x_1)-f_*}\right]$

$\triangle$ – MD; o – OSMD

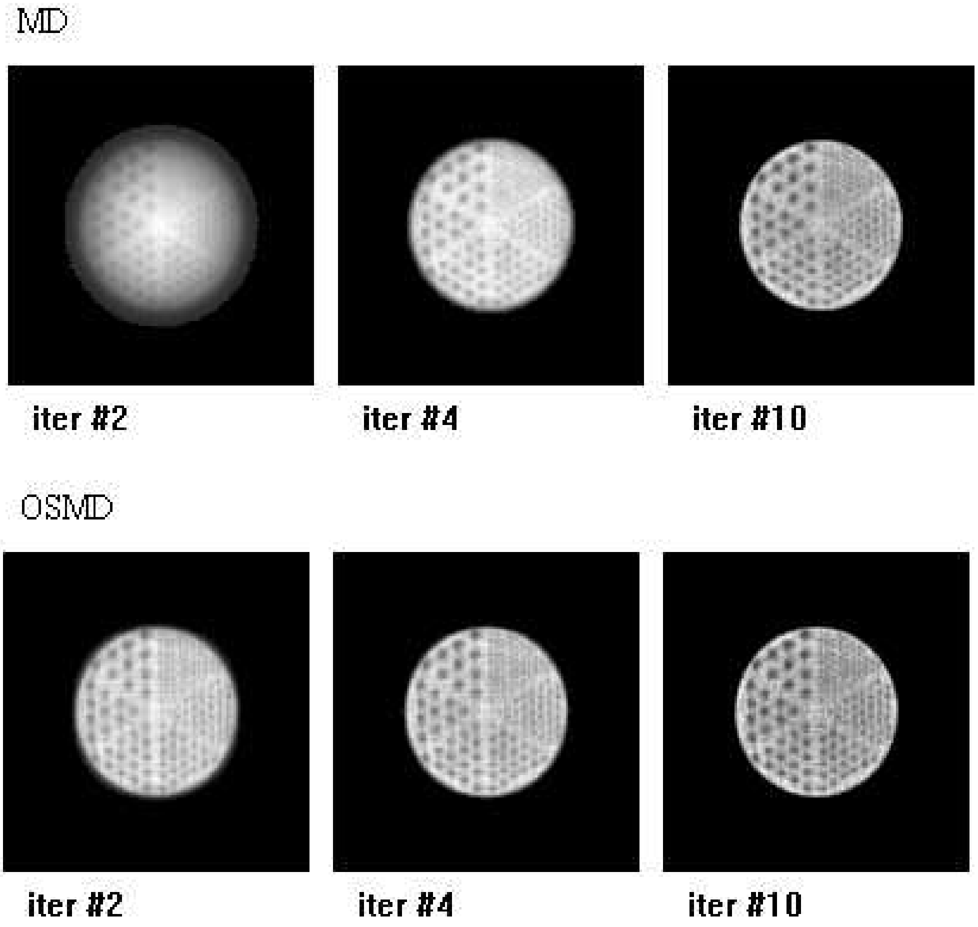

Figure 11. Jaszczak, mid-slice of the reconstructions.

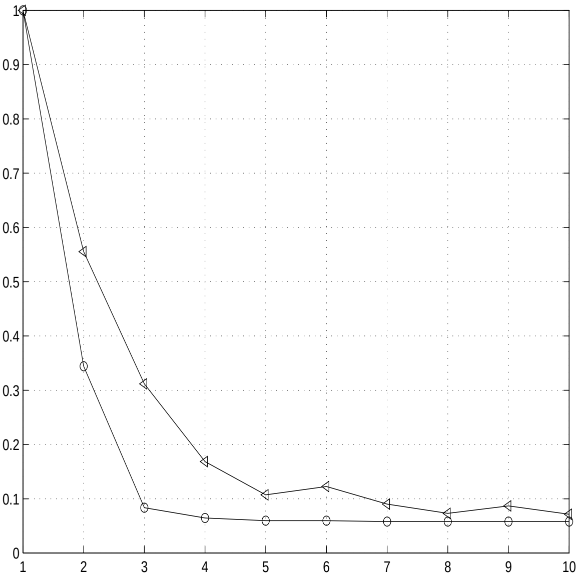

Figure 12. `Brain`, progress in accuracy: plot of $\theta(t) = \frac{f(x_t)-f_*^{10}}{f(x_1)-f_*^{10}} \quad \left[\geq \frac{f(x_t)-f_*}{f(x_1)-f_*}\right]$

$\triangle$ – MD; o – OSMD

computational effort. Moreover, the OSMD algorithm possesses a solid theoretical background (guaranteed efficiency estimates), which is not the case for OSEM.

7 Conclusions

The outlined results of our research suggest the following conclusions:

1. Simple gradient-descent type optimization techniques, which seem to be the only option when solving really large-scale (hundreds thousands and millions of variables) convex optimization problems, can be quite successful and can yield a solution of a satisfactory quality in few iterations.

2. When implementing gradient-type optimization techniques, one should try to adjust the method to the "geometry" of the problem. For such an adjustment, the general Mirror Descent scheme can be used.

3. Implementing gradient-descent type techniques in an "incremental gradient" fashion can accelerate significantly the solution process.

Acknowledgments. We gratefully acknowledge the help of members of the PARAPET consortium, especially Matthew Jacobson and Dr. Ron Levkovitz. We are greatly indebted to anonymous referees for their suggestions aimed at improving the structure of the paper.

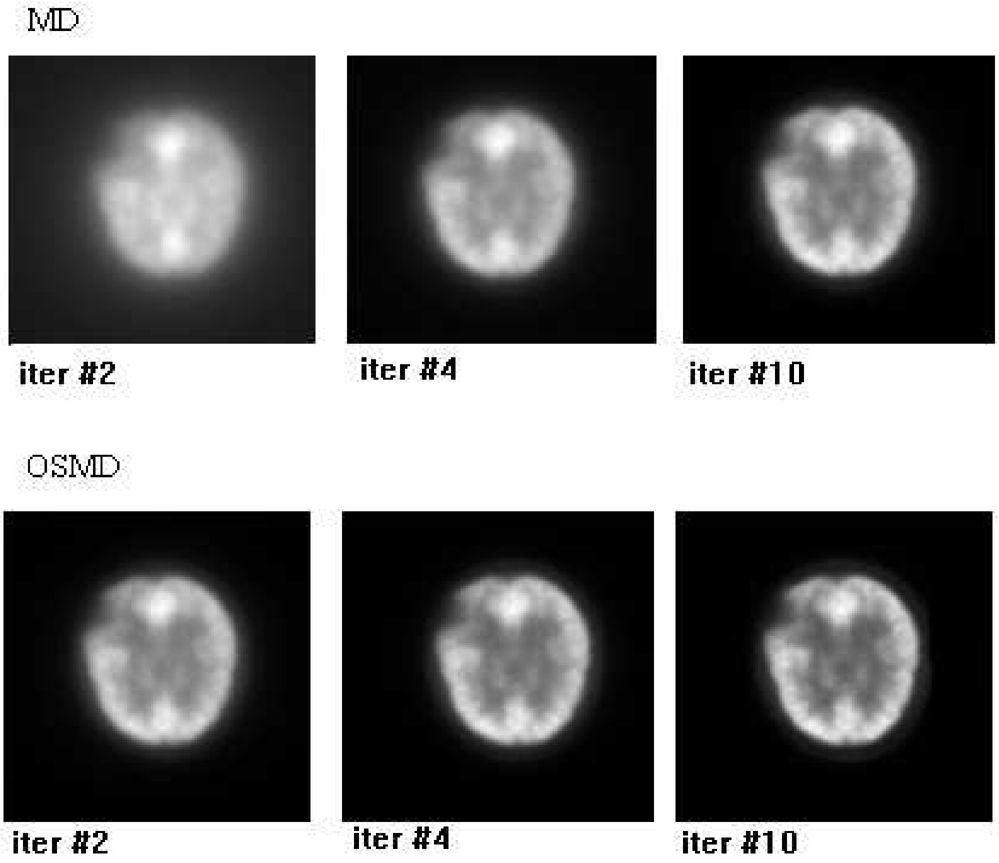

Figure 13. `Brain`, near-mid slice of the reconstructions.

[the top-left missing part is the area affected by the Alzheimer disease]

References

[Ber97] Bertsekas, D.P. (1997), "A new class of incremental gradient methods for least squares problems", *SIAM J. on Optimization* Vol. 7, No. 4, pp 913-926.

[Ber95] Bertsekas, D.P. (1995), *Nonlinear Programming*, Athena Scientific, Belmont, Massachusetts.

[Ber96] Bertsekas, D.P. (1996), "Incremental least squares methods and the extended Kalman filter", *SIAM J. on Optimization* Vol. 6, pp. 807-822.

[Hud94] Hudson, H.M. and Larkin, R.S. (1994), "Accelerated image reconstruction using ordered subsets of projection data", *IEEE Trans. on Medical Imaging* Vol. 13, No. 4, pp. 601-609.

[Kam98] Kamphuis, C. and Beekman, F.J. (1998), "Accelerated iterative transmission CT reconstruction using an ordered subsets convex algorithm", *IEEE Trans. on Medical Imaging* Vol. 17, No. 6, pp. 1101-1105.

[KLP99] Kiwiel, K.C., Larson, T., and Lindberg, P.O. (1999), "The efficiency of ballstep subgradient level methods for convex optimization", *Mathematics of Operations Research* Vol. 24, 237-254.

[Lan84] Lange, K. and Carson, R. (1984), "EM reconstruction algorithms for emission and transmission tomography", *J. Comp. Assist. Tomogr.* Vol. 8, pp. 306-316.

[Luo91] Luo, Z.Q. (1991), "On the convergence of the LMS algorithm with adaptive learning rate for linear feedforward networks", *Neural Computation* Vol. 3, pp. 226-245.

[Luo94] Luo, Z.Q. and Tseng P. (1994), "Analysis of an approximate gradient projection method with applications to the backpropagation algorithm", *Optimization Methods and Software* Vol. 4, pp. 85-101.

[Man95] Manglos, S.H., Gagne, G.M., Krol, A., Thomas, F.D. and Narayanaswamy, R. (1995), "Transmission maximum-likelihood reconstruction with ordered subsets for cone bean CT", *Phys. Med. Biol.* Vol. 40, No. 7, pp. 1225-1241.

[Nem78] Nemirovski, A. and Yudin, D. (1978) *Problem complexity and method efficiency in optimization* – Nauka Publishers, Moscow, 1978 (in Russian); English translation: John Wiley & Sons, 1983.

[Pol67] Polyak, B.T. (1967) "A general method for solving extremal problems", *Soviet Math. Doklady* Vol. 174, No.1, pp. 33-36.

[She74] Shepp, L.A. and Logan, B.F. (1974), "The Fourier reconstruction of a head section", *IEEE Trans. Nucl. Sci.* Vol.32, pp. 21-43.

[RW98] Rockafellar, R.T., and Wets, R.J-B. (1998), *Variational Analysis*, Springer, 1998.

[Roc76] Rockmore, A. and Makobvski, A. (1976), "A maximum likelihood approach to emission image reconstruction from projections", *IEEE Trans. Nuc. Sci.* Vol. 23, pp. 1428-1432.

[She82] Shepp, L.A. and Vardi, Y. (1982), "Maximum likelihood reconstruction for emission tomography", *IEEE Trans. on Medical Imaging* Vol. MI-1, No. 2, pp. 113-122.

[Sho67] Shor, N.Z. (1967), "Generalized gradient descent with application to block programming", *Kibernetika*, No. 3, 1967 (in Russian).

[Thi99] Thielemans, K. (1999), "A data library for the PARAPET project", PARAPET ESPRIT consortium deliveravle 1.2.

[Tse98] Tseng, P. (1998), "An incremental gradient (-projection) method with momentum term and adaptive stepsize rule", *SIAM J. on Optimization* Vol. 8, No. 2, pp. 506-531.

[Var85] Vardi, Y., Shepp, L.A. and Kaufman, L. (1985), "A Statistical Model for Positron Emission Tomography", *J. Amer. Statist. Assoc.* Vol. 80, pp. 8-37.

[Zai98] Zaidi, H., Labbe, C. and Morel, C. (1998), "Implementation of a Monte Carlo simulation environment for fully 3D PET on a high performance parallel platform", *Parallel Computing* Vol. 24, pp. 1523-1536.

[Zai99] Zaidi, H., Hermann Scheurer, A.K. and Morel, C. (1999), "An Object-Oriented Monte Carlo Simulator for 3D Cylindrical PET Tomographs", *Computer Methods and Programs in Biomedicine* Vol. 58, pp. 133-145.

8 Appendix 1: Strong convexity of $\frac{1}{2}\|\cdot\|_p^2$

Here we reproduce the proof of the following known fact (see, e.g., [Nem78]):

Lemma 8.1 *Let $1 < p \le 2$, and let $w(x) = \frac{1}{2}\|x\|_p^2 : \mathbf{R}^n \to \mathbf{R}$. Then the function w is α-strongly convex w.r.t. the norm $\|\cdot\|_p$, with*

$$\alpha = p - 1. \tag{42}$$

Proof. It is known ([RW98], Propositions 12.54, 12.60) that the fact that a continuously differentiable convex function $v : \mathbf{R}^n \to \mathbf{R}$ is α-strongly convex on $\mathbf{R}^n$ w.r.t. a norm $\|\cdot\|$ is equivalent to the fact that the Legendre transformation

$$V(\xi) = \max_{x\in\mathbf{R}^n} [\xi^T x - v(x)]$$

of v is continuously differentiable and satisfies the relation

$$V(\xi + \eta) \le V(\xi) + \eta^T \nabla V(\xi) + \frac{1}{2\alpha}\|\eta\|_*^2 \quad \forall \xi, \eta \in \mathbf{R}^n, \tag{43}$$

where $\|\cdot\|_*$ is the norm conjugate to $\|\cdot\|$. In our case, $\|\cdot\| = \|\cdot\|_p$ and $V(\xi) = \frac{1}{2}\|\xi\|_q^2$, $q = p/(p-1) \ge 2$, so that V is twice continuously differentiable outside of the origin (and, of course, is convex); therefore, in order to verify that (43) is satisfied with $\alpha = p - 1$, it suffices to prove that

$$\eta^T \nabla^2 V(\xi)\eta \le \frac{1}{p-1}\|\eta\|_q^2 \tag{44}$$

for every $\xi \neq 0$. By homogeneity, $\nabla^2 V(t\xi) = \nabla^2 V(\xi)$, $t > 0$, so that when proving (44), we may assume that $\|\xi\|_q = 1$. We now have

$$\begin{array}{rcll}
\eta^T \nabla V(\xi) & = & \|\xi\|_q^{2-q} \sum\limits_{i=1}^n |\xi|_i^{q-1} \operatorname{sign}(\xi_i)\eta_i, & \\
\eta^T \nabla^2 V(\xi)\eta & = & (2-q)\|\xi\|_q^{2-2q} \left(\sum\limits_{i=1}^n |\xi|_i^{q-1} \operatorname{sign}(\xi_i)\eta_i\right)^2 & \\
& & +(q-1)\|\xi\|_q^{2-q} \sum\limits_{i=1}^n |\xi_i|^{q-2}\eta_i^2 & \\
& \leq & (q-1)\sum\limits_{i=1}^n |\xi_i|^{q-2}\eta_i^2 & [\text{since } q \geq 2, \|\xi\|_q = 1] \\
& \leq & (q-1)\left(\sum\limits_i |\xi_i|^q\right)^{\frac{q-2}{q}} \left(\sum\limits_i |\eta|_i^q\right)^{\frac{2}{q}} & [\text{Hölder's inequality}] \\
& \leq & (q-1)\|\eta\|_q^2, &
\end{array}$$

so that (44) is satisfied, due to $q - 1 = \frac{1}{p-1}$. ∎

9 Appendix 2: Proof of Proposition 4.1

1^0. W.l.o.g., we can assume that n is a power of 2: $n = 2^k$. It is known that there exists an orthogonal basis $u^1, ..., u^m$ in $\mathbf{R}^m$, $m = 2^{k-1}$, such that $|u_j^\ell| = 1$ for all $\ell, j = 1, ..., m$. Let $e^\ell = \begin{pmatrix} u^\ell \\ -u^\ell \end{pmatrix} \in \mathbf{R}^{2m} = \mathbf{R}^n$, $\ell = 1, ..., m$. Note that

1^0.A. $\|e^\ell\|_2^2 = n$, $\ell = 1, ..., m$;

1^0.B. $[e^\ell]^T e^{\ell'} = 0$, $1 \leq \ell < \ell' \leq m$.

1^0.C. $\sum\limits_{t=1}^n e_t^\ell = 0$, $\ell = 1, ..., m$.

1^0.D. For every linear combination $e[\lambda] = \sum\limits_{\ell=1}^m \lambda_\ell e^\ell$ one has $e_i[\lambda] = -e_{m+i}[\lambda]$, $i = 1, ..., m$, whence

$$\|e[\lambda]\|_\infty = \max_{i \leq n} e_i[\lambda] = -\min_{i \leq n} e_i[\lambda].$$

2^0. Let $\delta > 0$, $1 < k \leq m$, and let $\mathcal{B}$ be a method for solving problems from $\mathcal{F} = \mathcal{F}(L, n)$. Let us set

$$\varepsilon(\mathcal{B}, k) = \sup_{f \in \mathcal{F}} \left[f(x^{k-1}(\mathcal{B}, f)) - \min_{\Delta_n} f \right].$$

We are about to prove that

$$\varepsilon(\mathcal{B}, k) \geq \frac{L}{\sqrt{k}}. \tag{45}$$

Note that this inequality immediately implies the desired lower bound on the information-based complexity of $\mathcal{F}$.

From the viewpoint of the behaviour of $\mathcal{B}$ at the first $k-1$ steps (which is the only issue we are interested in when proving (45)), we change nothing when assuming that $\mathcal{B}$, as applied to a problem from $\mathcal{F}$, performs exactly k steps; the search points generated by the method at the first $k-1$ steps are as given by the search rules specifying the method, and the last search point x_k is the kth approximate solution generated by $\mathcal{B}$ as applied to the problem. Thus, from now on we assume that the point $x^{k-1}(\mathcal{B}, f)$ in (45) is the k-st search point generated by $\mathcal{B}$ as applied to f.

3^0. To prove (45), we intend to construct a "difficult" for $\mathcal{B}$ objective f as the pointwise maximum of k linear functions with orthogonal descent directions chosen from the set $\{\pm\ell^1, ..., \pm\ell^m\}$. These linear functions will be successively constructed according to the adversary principle. i.e., when $\mathcal{B}$ requires evaluation at search point x_i, the ith linear function is defined such that little progress is achieved while consistency with previous information is maintained. The construction is as follows. Let x_1 be the first search point of the method (this point is problem-independent), let

$$\ell_1 \in \mathop{\mathrm{Argmax}}_{1\leq\ell\leq k} |x_1^T e^\ell|, \quad \sigma_1 = \mathrm{sign}(x_1^T e^{\ell_1}) \ \left[\mathrm{sign}(s) = \begin{cases} 1, & s \geq 0 \\ -1, & s < 0 \end{cases}\right],\ f^1(x) = L\sigma_1 x^T e^{\ell_1} - \delta.$$

Suppose we have defined already $x_1, ..., x_p$, $\ell_1, ..., \ell_p$, $f^1(\cdot), ..., f^p(\cdot)$, $\sigma_1, ..., \sigma_p \in \{-1; 1\}$ in such a way that

(a_p) $1 \leq \ell_i \leq k$ and the indices $\ell_1, ..., \ell_p$ are distinct from each other;

(b_p) $f^i(x) = \max\limits_{j=1,...,i} [L\sigma_j x^T e^{\ell_j} - j\delta]$, $i = 1, ..., p$;

(c_p) $x_1, ..., x_i$ is the initial i-element segment of the *trajectory* (the sequence of search points) of $\mathcal{B}$ as applied to $f^i(\cdot)$;

(d_p) $\sigma_i x_i^T e^{\ell_i} = \max\{|x_i^T e^\ell| \mid \ell \in \{1, ..., k\}\backslash\{\ell_1, ..., \ell_{i-1}\}\}$, $i = 1, ..., p$.

Note that with our initialization conditions (a_1) – (d_1) do hold.

In the case of $p < k$, let us extend the collection we have built to a similar collection of $(p+1)$-element tuples; to this end we define x_{p+1} as the $(p+1)$th search point of $\mathcal{B}$ as applied to $f^p(\cdot)$, ℓ_{p+1} as the index from the set $I^p = \{1, ..., k\}\backslash\{\ell_1, ..., \ell_p\}$ which maximizes the quantities $x_{p+1}^T e^\ell$ over $\ell \in I^p$, and σ_{p+1} as $\mathrm{sign}(x_{p+1}^T e^{\ell_{p+1}})$, and finally set

$$f^{p+1}(x) = \max\{f^p(x), L\sigma_{p+1} x^T e^{\ell_{p+1}} - (p+1)\delta\}.$$

It is easily seen that when $1 \leq i \leq j \leq p+1$, one has $f^j(x) = f^i(x)$ in a neighbourhood of x_i; with this observation, (a_{p+1}) – (d_{p+1}) immediately follow from (a_p) – (d_p) and our construction.

After k steps of the aforementioned construction, we get a function

$$f(x) \equiv f^k(x) = \max_{1\leq i\leq k} [L\sigma_i x^T e^{\ell_i} - i\delta]$$

such that the trajectory of $\mathcal{B}$ on f is $x_1, .., x_k$, so that x_k is the result of $\mathcal{B}$ as applied to f. Observe that $f \in \mathcal{F}(L, n)$, due to $\|e^\ell\|_\infty = 1$. In view of (d_p), we have

$$f(x_k) \geq -k\delta. \tag{46}$$

On the other hand, let us bound from above the minimum value of f over Δ_n. We have

$$f(x) = \max_{i=1,...,k} [L\sigma_i x^T e^{\ell_i} - i\delta] \leq g(x) \equiv \max_{i=1,...,k} L\sigma_i x^T e^{\ell_i}$$

and therefore

$$\begin{array}{rcl}
\min\limits_{x\in\Delta_n} f(x) & \leq & \min\limits_{x\in\Delta_n} g(x) = L\min\limits_{x\in\Delta_n}\max\limits_{i\leq k} x^T[\sigma_i e^{\ell_i}] \\
& = & L\min\limits_{x\in\Delta_n}\max\limits_{\lambda\in\Delta_k} x^T \underbrace{\left[\sum_{i=1}^k \lambda_i\sigma_i e^{\ell_i}\right]}_{\widetilde{e}[\lambda]} \\
& = & L\max\limits_{\lambda\in\Delta_k}\min\limits_{x\in\Delta_n} x^T\widetilde{e}[\lambda] = L\max\limits_{\lambda\in\Delta_k}\min\limits_{i=1,\ldots,n}\widetilde{e}_i[\lambda] \\
& = & L\max\limits_{\lambda\in\Delta_k}[-\|\widetilde{e}_i[\lambda]\|_\infty] \qquad\qquad \text{[see } 1^0\text{.D]} \\
& = & -L\min\limits_{\lambda\in\Delta_k}\|\widetilde{e}[\lambda]\|_\infty \leq -Ln^{-1/2}\min\limits_{\lambda\in\Delta_k}\|\widetilde{e}[\lambda]\|_2 \\
& = & -Ln^{-1/2}\min\limits_{\lambda\in\Delta_k}\sqrt{\sum\limits_{i=1}^k \lambda_i^2\sigma_i^2\|e^{\ell_i}\|_2^2} \qquad \text{[see } 1^0\text{.B]} \\
& = & -Ln^{-1/2}\min\limits_{\lambda\in\Delta_k}\sqrt{\sum\limits_{i=1}^k \lambda_i^2 n} \qquad \text{[see } 1^0\text{.A]} \\
& \leq & -Lk^{-1/2},
\end{array}$$

We see that $\min\limits_{x\in\Delta_n} f(x) \leq -Lk^{-1/2}$, which combines with (46) to yield that

$$f(x_k) - \min_{x\in\Delta_n} f \geq Lk^{-1/2} - k\delta.$$

Since $f \in \mathcal{F}$, $x_k = x^{k-1}(\mathcal{B}, f)$ and $\delta > 0$ is arbitrary, (45) follows. ■