From greedy approximation to greedy optimization

Vladimir Temlyakov

July, 2014

Toy example

Let $\Psi := \{\psi\}_{k=1}^{\infty}$ be an orthonormal basis for a Hilbert space H. For any $f \in H$ there is a convergent (in H) orthogonal expansion

$$f = \sum_{k=1}^{\infty} \langle f, \psi_k \rangle \psi_k.$$

A classical way of approximation of f is to take a partial sum

$$S_n(f, \Psi) := \sum_{k=1}^{n} \langle f, \psi_k \rangle \psi_k.$$

For the error we have

$$\|f - S_n(f, \Psi)\|^2 = \sum_{k=n+1}^{\infty} |\langle f, \psi_k \rangle|^2.$$

m-term approximation

In nonlinear approximation we use the m-term approximation

$$\sum_{k\in\Lambda}\langle f,\psi_k\rangle\psi_k, \quad |\Lambda| = m.$$

It is clear that the optimal (from the point of view of the error) choice of Λ is the set of m biggest in absolute value coefficients $\langle f,\psi_k\rangle$. We can realize this choice by picking the biggest coefficients one by one. This results in the reordering (greedy reordering) of the orthogonal expansion:

$$f = \sum_{i=1}^{\infty}\langle f,\psi_{k_i}\rangle\psi_{k_i}, \quad |\langle f,\psi_{k_1}\rangle| \ge |\langle f,\psi_{k_2}\rangle| \ge \dots.$$

Major questions of greedy approximation

1. Let instead of an orthonormal basis Ψ we have a redundant system $\mathcal{D}$. How to approximate with regard to $\mathcal{D}$?

Major questions of greedy approximation

1. Let instead of an orthonormal basis Ψ we have a redundant system $\mathcal{D}$. How to approximate with regard to $\mathcal{D}$?
2. How to work in a Banach space X instead of a Hilbert space H?

Notations

We begin with the case where approximation takes place in a Banach space X equipped with a norm $\|\cdot\| := \|\cdot\|_X$. We formulate our approximation problem in the following general way.

Definition (Dictionary)

We say a set of functions $\mathcal{D}$ from X is a dictionary if each $g \in X$ has norm one ($\|g\|_X = 1$) and the closure of Span $\mathcal{D}$ coincides with X.

We let $\Sigma_m(\mathcal{D})$ denote the collection of all functions (elements) in X which can be expressed as a linear combination of at most m elements of $\mathcal{D}$.

m-sparse elements

Thus each function $s \in \Sigma_m(\mathcal{D})$ can be written in the form

$$s = \sum_{g \in \Lambda} c_g g, \quad \Lambda \subset \mathcal{D}, \quad \#\Lambda \le m,$$

where the c_g are real numbers. In some cases, it may be possible to write an element from $\Sigma_m(\mathcal{D})$ in this form in more than one way. The space $\Sigma_m(\mathcal{D})$ is not linear: the sum of two functions from $\Sigma_m(\mathcal{D})$ is generally not in $\Sigma_m(\mathcal{D})$.

Examples

Perhaps the first example of approximation involving dictionaries was considered by E. Schmidt in 1907, who considered the approximation of functions $f(x,y)$ of two variables in $L_2([0,1]^2)$ by functions of the form

$$B_m(x,y) = \sum_{j=1}^{m} c_j u_j(x) v_j(y).$$

This approximation problem can be seen as an m-term approximation with regard to the dictionary

$$\Pi = \{g \, : g(x,y) = u(x)v(y);$$

$$u, v \in L_2([0,1]), \|u\|_{L_2} = \|v\|_{L_2} = 1\}.$$

One more example

Another approximation problem of this type which is well known in statistics is the projection pursuit regression problem. The problem is to approximate in L_2 a given multivariate function $f \in L_2$ by a sum of ridge functions, i.e. by

$$W_m(x) = \sum_{j=1}^{m} r_j(\langle \omega_j, x \rangle), \qquad x, \omega_j \in \mathbb{R}^d, \quad j = 1, \dots, m,$$

where r_j, $j = 1, \dots, m$, are univariate functions.

More examples

Another example, from signal processing, uses the Gabor functions

$$g_{a,b}(x) := e^{iax}e^{-bx^2}$$

and approximates a univariate function by linear combinations of the elements

$$\{g_{a,b}(x-c) : a, c \in \mathbb{R}, b > 0\}.$$

Best m-term approximation

For a function $f \in X$ we define its best m-term approximation error

$$\sigma_m(f, \mathcal{D})_X := \inf_{s \in \Sigma_m(\mathcal{D})} \|f - s\|_X.$$

We concentrate on an important problem of finding good methods of m-term approximation in the case of general dictionary $\mathcal{D}$ and on studying their efficiency. Let us begin this discussion in the special case of a Hilbert space with the inner product $\langle \cdot, \cdot \rangle$. We define first the Weak Greedy Algorithm (WGA) in Hilbert space H. We describe this algorithm for a general dictionary $\mathcal{D}$.

WGA

Let a sequence $\tau = \{t_k\}_{k=1}^{\infty}$, $0 \le t_k \le 1$, be given.
WGA We define $f_0^\tau := f$. Then for each $m \ge 1$, we inductively define:

1. $\varphi_m^\tau \in \mathcal{D}$ is any satisfying

$$|\langle f_{m-1}^\tau, \varphi_m^\tau \rangle| \ge t_m \sup_{g \in \mathcal{D}} |\langle f_{m-1}^\tau, g \rangle|;$$

WGA

Let a sequence $\tau = \{t_k\}_{k=1}^{\infty}$, $0 \le t_k \le 1$, be given.
WGA We define $f_0^\tau := f$. Then for each $m \ge 1$, we inductively define:

1. $\varphi_m^\tau \in \mathcal{D}$ is any satisfying

$$|\langle f_{m-1}^\tau, \varphi_m^\tau \rangle| \ge t_m \sup_{g \in \mathcal{D}} |\langle f_{m-1}^\tau, g \rangle|;$$

2. $f_m^\tau := f_{m-1}^\tau - \langle f_{m-1}^\tau, \varphi_m^\tau \rangle \varphi_m^\tau$;

WGA

Let a sequence $\tau = \{t_k\}_{k=1}^{\infty}$, $0 \le t_k \le 1$, be given.
WGA We define $f_0^\tau := f$. Then for each $m \ge 1$, we inductively define:

1. $\varphi_m^\tau \in \mathcal{D}$ is any satisfying

$$|\langle f_{m-1}^\tau, \varphi_m^\tau \rangle| \ge t_m \sup_{g \in \mathcal{D}} |\langle f_{m-1}^\tau, g \rangle|;$$

2. $f_m^\tau := f_{m-1}^\tau - \langle f_{m-1}^\tau, \varphi_m^\tau \rangle \varphi_m^\tau$;
3. $G_m^\tau(f, \mathcal{D}) := \sum_{j=1}^m \langle f_{j-1}^\tau, \varphi_j^\tau \rangle \varphi_j^\tau$.

Historical comment

In the case $t_k = 1$, $k = 1, \dots$ the WGA is called Pure Greedy Algorithm (PGA). The PGA was proposed by J.H. Friedman and W. Stuetzle in 1981 for the ridge dictionary. We note that in a particular case $t_k = t$, $k = 1, 2, \dots$, the WGA was considered by L. Jones (1987) (also for the ridge dictionary). The WGA provides for each $f \in H$ an expansion into a series (greedy expansion)

$$f \sim \sum_{j=1}^{\infty} c_j(f)\varphi_j^\tau, \quad c_j(f) := \langle f_{j-1}^\tau, \varphi_j^\tau \rangle.$$

In general it is not an expansion into an orthogonal series but it has some similar properties.

Parseval's formula

The coefficients $c_j(f)$ of an expansion are obtained by the Fourier formulas with f replaced by the residuals f_{j-1}^{τ}. It is easy to see that

$$\|f_m^{\tau}\|^2 = \|f_{m-1}^{\tau}\|^2 - |c_m(f)|^2.$$

There are convergence results for the greedy expansion and, therefore, from the above equality we get for this expansion an analog of the Parseval formula for orthogonal expansions:

$$\|f\|^2 = \sum_{j=1}^{\infty} |c_j(f)|^2.$$

Rate of convergence

For a general dictionary $\mathcal{D}$ we define the class of functions

$$\mathcal{A}_1^o(\mathcal{D}, M) := \{f \in H : f = \sum_{k\in\Lambda} c_k w_k, \quad w_k \in \mathcal{D}, \ \#\Lambda < \infty$$

$$\sum_{k\in\Lambda} |c_k| \le M\}$$

and we define $\mathcal{A}_1(\mathcal{D}, M)$ as the closure (in H) of $\mathcal{A}_1^o(\mathcal{D}, M)$. Furthermore, we define $\mathcal{A}_1(\mathcal{D})$ as the union of the classes $\mathcal{A}_1(\mathcal{D}, M)$ over all $M > 0$. For $f \in \mathcal{A}_1(\mathcal{D})$, we define the norm

$$|f|_{\mathcal{A}_1(\mathcal{D})}$$

as the smallest M such that $f \in \mathcal{A}_1(\mathcal{D}, M)$.

First results

It was proved in [DeVore,T., 1996] that for a general dictionary $\mathcal{D}$ the Pure Greedy Algorithm provides the following estimate

$$\|f - G_m(f,\mathcal{D})\| \le |f|_{\mathcal{A}_1(\mathcal{D})} m^{-1/6}. \tag{1}$$

(In this and similar estimates we consider that the inequality holds for all possible choices of $\{G_m\}$.) That paper contains also an example of a dictionary $\mathcal{D}$ and an element f such that

$$\|f - G_m(f,\mathcal{D})\| > \frac{1}{2}|f|_{\mathcal{A}_1(\mathcal{D})} m^{-1/2}, \quad m \ge 4.$$

Further results

We proved in [Konyagin,T., 1999] an estimate

$$\|f - G_m(f, \mathcal{D})\| \le 4|f|_{\mathcal{A}_1(\mathcal{D})} m^{-11/62}$$

which improves a little the original one (see (1)).
E. Livshitz and T. (2002) proved the following lower estimate.
There exist a dictionary $\mathcal{D}$ and an element $f \in H$, $f \neq 0$, such that

$$\|f - G_m(f, \mathcal{D})\| \ge C m^{-0.27} |f|_{\mathcal{A}_1(\mathcal{D})}$$

with a positive constant C.
A. Sil'nichenko improved the exponent $11/62$ to 0.182 in the upper estimate and E. Livshitz improved the exponent 0.27 to 0.1898 in the lower estimate.

Open Problem

Find the right order of the sequence

$$\sup_{f,H,\mathcal{D}} \|f - G_m(f,\mathcal{D})\|/|f|_{\mathcal{A}_1(\mathcal{D})}.$$

WOGA

Let a sequence $\tau = \{t_k\}_{k=1}^{\infty}$, $0 \le t_k \le 1$, be given. We define the Weak Orthogonal Greedy Algorithm (WOGA).

WOGA We define $f_0^{o,\tau} := f$. Then for each $m \ge 1$ we inductively define:

1. $\varphi_m^{o,\tau} \in \mathcal{D}$ is any element satisfying

$$|\langle f_{m-1}^{o,\tau}, \varphi_m^{o,\tau} \rangle| \ge t_m \sup_{g \in \mathcal{D}} |\langle f_{m-1}^{o,\tau}, g \rangle|;$$

WOGA

Let a sequence $\tau = \{t_k\}_{k=1}^{\infty}$, $0 \le t_k \le 1$, be given. We define the Weak Orthogonal Greedy Algorithm (WOGA).

WOGA We define $f_0^{o,\tau} := f$. Then for each $m \ge 1$ we inductively define:

1. $\varphi_m^{o,\tau} \in \mathcal{D}$ is any element satisfying

$$|\langle f_{m-1}^{o,\tau}, \varphi_m^{o,\tau} \rangle| \ge t_m \sup_{g \in \mathcal{D}} |\langle f_{m-1}^{o,\tau}, g \rangle|;$$

2.

$$G_m^{o,\tau}(f, \mathcal{D}) := P_{H_m^\tau}(f), \quad \text{where} \quad H_m^\tau := \operatorname{Span}(\varphi_1^{o,\tau}, \dots, \varphi_m^{o,\tau});$$

WOGA

Let a sequence $\tau = \{t_k\}_{k=1}^{\infty}$, $0 \le t_k \le 1$, be given. We define the Weak Orthogonal Greedy Algorithm (WOGA).

WOGA We define $f_0^{o,\tau} := f$. Then for each $m \ge 1$ we inductively define:

1. $\varphi_m^{o,\tau} \in \mathcal{D}$ is any element satisfying

$$|\langle f_{m-1}^{o,\tau}, \varphi_m^{o,\tau} \rangle| \ge t_m \sup_{g \in \mathcal{D}} |\langle f_{m-1}^{o,\tau}, g \rangle|;$$

2. $$G_m^{o,\tau}(f, \mathcal{D}) := P_{H_m^\tau}(f), \quad \text{where} \quad H_m^\tau := \mathrm{Span}(\varphi_1^{o,\tau}, \dots, \varphi_m^{o,\tau});$$

3. $f_m^{o,\tau} := f - G_m^{o,\tau}(f, \mathcal{D})$.

Rate of convergence

Theorem (T., 2000)

Let $\mathcal{D}$ be an arbitrary dictionary in H. Then for each $f \in \mathcal{A}_1(\mathcal{D}, M)$ we have

$$\|f - G_m^{o,\tau}(f, \mathcal{D})\| \le M(1 + \sum_{k=1}^{m} t_k^2)^{-1/2}.$$

Notations

Let X be a Banach space with norm $\|\cdot\|$.

Definition

We say that a set of elements (functions) $\mathcal{D}$ from X is a symmetric dictionary if each $g \in \mathcal{D}$ has norm equal to one ($\|g\| = 1$),

$$g \in \mathcal{D} \quad \text{implies} \quad -g \in \mathcal{D},$$

and closure of Span $\mathcal{D} = X$.

For an element $f \in X$ we denote by F_f a norming (peak) functional for f:

$$\|F_f\| = 1, \qquad F_f(f) = \|f\|.$$

Two forms

The greedy step (the first step) of the PGA can be interpreted in two ways.

- First, we look at the mth step for an element $\varphi_m \in \mathcal{D}$ and a number λ_m satisfying

$$\|f_{m-1} - \lambda_m \varphi_m\|_H = \inf_{g \in \mathcal{D}, \lambda} \|f_{m-1} - \lambda g\|_H. \qquad (2)$$

Two forms

The greedy step (the first step) of the PGA can be interpreted in two ways.

- First, we look at the mth step for an element $\varphi_m \in \mathcal{D}$ and a number λ_m satisfying

$$\|f_{m-1} - \lambda_m \varphi_m\|_H = \inf_{g \in \mathcal{D}, \lambda} \|f_{m-1} - \lambda g\|_H. \tag{2}$$

- Second, we look for an element $\varphi_m \in \mathcal{D}$ such that

$$\langle f_{m-1}, \varphi_m \rangle = \sup_{g \in \mathcal{D}} \langle f_{m-1}, g \rangle. \tag{3}$$

Two forms

The greedy step (the first step) of the PGA can be interpreted in two ways.

- First, we look at the mth step for an element $\varphi_m \in \mathcal{D}$ and a number λ_m satisfying

$$\|f_{m-1} - \lambda_m \varphi_m\|_H = \inf_{g \in \mathcal{D}, \lambda} \|f_{m-1} - \lambda g\|_H. \tag{2}$$

- Second, we look for an element $\varphi_m \in \mathcal{D}$ such that

$$\langle f_{m-1}, \varphi_m \rangle = \sup_{g \in \mathcal{D}} \langle f_{m-1}, g \rangle. \tag{3}$$

In a Hilbert space both versions (2) and (3) result in the same PGA. In a general Banach space the corresponding versions of (2) and (3) lead to different greedy algorithms.

XGA

The Banach space version of (2) is straightforward: instead of the Hilbert norm $\|\cdot\|_H$ in (2) we use the Banach norm $\|\cdot\|_X$. This results in the following greedy algorithm.

X-Greedy Algorithm (XGA) We define $f_0 := f$, $G_0 := 0$. Then, for each $m \geq 1$, we inductively define

XGA

The Banach space version of (2) is straightforward: instead of the Hilbert norm $\|\cdot\|_H$ in (2) we use the Banach norm $\|\cdot\|_X$. This results in the following greedy algorithm.

X-Greedy Algorithm (XGA) We define $f_0 := f$, $G_0 := 0$. Then, for each $m \geq 1$, we inductively define

1. $\varphi_m \in \mathcal{D}$, $\lambda_m \in \mathbb{R}$ are such that (we assume existence)

$$\|f_{m-1} - \lambda_m \varphi_m\|_X = \inf_{g \in \mathcal{D}, \lambda} \|f_{m-1} - \lambda g\|_X. \tag{4}$$

XGA

The Banach space version of (2) is straightforward: instead of the Hilbert norm $\|\cdot\|_H$ in (2) we use the Banach norm $\|\cdot\|_X$. This results in the following greedy algorithm.

X-Greedy Algorithm (XGA) We define $f_0 := f$, $G_0 := 0$. Then, for each $m \geq 1$, we inductively define

1. $\varphi_m \in \mathcal{D}$, $\lambda_m \in \mathbb{R}$ are such that (we assume existence)

$$\|f_{m-1} - \lambda_m \varphi_m\|_X = \inf_{g \in \mathcal{D}, \lambda} \|f_{m-1} - \lambda g\|_X. \tag{4}$$

2. Denote

$$f_m := f_{m-1} - \lambda_m \varphi_m, \qquad G_m := G_{m-1} + \lambda_m \varphi_m.$$

Dual greedy algorithm

The second version of the PGA in a Banach space is based on the concept of a norming (peak) functional. We note that in a Hilbert space a norming functional F_f acts as follows

$$F_f(g) = \langle f/\|f\|, g\rangle.$$

Therefore, (3) can be rewritten in terms of the norming functional $F_{f_{m-1}}$ as

$$F_{f_{m-1}}(\varphi_m) = \sup_{g\in\mathcal{D}} F_{f_{m-1}}(g). \tag{5}$$

This observation leads to the class of dual greedy algorithms. We define the Weak Dual Greedy Algorithm with weakness $\tau := \{t_k\}_{k=1}^{\infty}$ (WDGA(τ)).

WDGA

Weak Dual Greedy Algorithm (WDGA(τ)) Let $\tau := \{t_m\}_{m=1}^{\infty}$, $t_m \in [0,1]$, be a weakness sequence. We define $f_0 := f$. Then, for each $m \geq 1$, we inductively define

1. $\varphi_m \in \mathcal{D}$ is any satisfying

$$F_{f_{m-1}}(\varphi_m) \geq t_m \|F_{f_{m-1}}\|_{\mathcal{D}}. \tag{6}$$

WDGA

Weak Dual Greedy Algorithm (WDGA(τ)) Let $\tau := \{t_m\}_{m=1}^{\infty}$, $t_m \in [0,1]$, be a weakness sequence. We define $f_0 := f$. Then, for each $m \geq 1$, we inductively define

1. $\varphi_m \in \mathcal{D}$ is any satisfying

$$F_{f_{m-1}}(\varphi_m) \geq t_m \|F_{f_{m-1}}\|_{\mathcal{D}}. \tag{6}$$

2. Define a_m as

$$\|f_{m-1} - a_m \varphi_m\| = \min_{a \in \mathbb{R}} \|f_{m-1} - a\varphi_m\|.$$

WDGA

Weak Dual Greedy Algorithm (WDGA(τ)) Let $\tau := \{t_m\}_{m=1}^{\infty}$, $t_m \in [0,1]$, be a weakness sequence. We define $f_0 := f$. Then, for each $m \geq 1$, we inductively define

1. $\varphi_m \in \mathcal{D}$ is any satisfying

$$F_{f_{m-1}}(\varphi_m) \geq t_m \|F_{f_{m-1}}\|_{\mathcal{D}}. \tag{6}$$

2. Define a_m as

$$\|f_{m-1} - a_m\varphi_m\| = \min_{a\in\mathbb{R}} \|f_{m-1} - a\varphi_m\|.$$

3. Denote

$$f_m := f_{m-1} - a_m\varphi_m.$$

Remark

First results on greedy approximation in Banach spaces were obtained by M. Donahue, L. Gurvits, C. Darken, and E. Sontag, 1997.

Let $\tau := \{t_k\}_{k=1}^{\infty}$ be a given sequence of nonnegative numbers $t_k \le 1$, $k = 1, \dots$. We define first the Weak Chebyshev Greedy Algorithm (WCGA) that is a generalization for Banach spaces of Weak Orthogonal Greedy Algorithm defined for Hilbert spaces.

WCGA

WCGA We define $f_0^c := f_0^{c,\tau} := f$. Then for each $m \geq 1$ we inductively define

1. $\varphi_m^c := \varphi_m^{c,\tau} \in \mathcal{D}$ is any satisfying

$$F_{f_{m-1}^c}(\varphi_m^c) \geq t_m \sup_{g \in \mathcal{D}} F_{f_{m-1}^c}(g).$$

WCGA

WCGA We define $f_0^c := f_0^{c,\tau} := f$. Then for each $m \geq 1$ we inductively define

1. $\varphi_m^c := \varphi_m^{c,\tau} \in \mathcal{D}$ is any satisfying

$$F_{f_{m-1}^c}(\varphi_m^c) \geq t_m \sup_{g \in \mathcal{D}} F_{f_{m-1}^c}(g).$$

2. Define

$$\Phi_m := \Phi_m^\tau := \mathrm{Span}\{\varphi_j^c\}_{j=1}^m,$$

and $G_m^c := G_m^{c,\tau}$ to be the best approximant to f from Φ_m.

WCGA

WCGA We define $f_0^c := f_0^{c,\tau} := f$. Then for each $m \geq 1$ we inductively define

1. $\varphi_m^c := \varphi_m^{c,\tau} \in \mathcal{D}$ is any satisfying

$$F_{f_{m-1}^c}(\varphi_m^c) \geq t_m \sup_{g \in \mathcal{D}} F_{f_{m-1}^c}(g).$$

2. Define

$$\Phi_m := \Phi_m^\tau := \mathrm{Span}\{\varphi_j^c\}_{j=1}^m,$$

and $G_m^c := G_m^{c,\tau}$ to be the best approximant to f from Φ_m.

3. Denote $f_m^c := f_m^{c,\tau} := f - G_m^c$.

Modulus of smoothness

We consider here approximation in uniformly smooth Banach spaces.

Definition

For a Banach space X we define the modulus of smoothness

$$\rho(u) := \sup_{\|x\|=\|y\|=1} \left(\frac{1}{2}(\|x+uy\| + \|x-uy\|) - 1\right).$$

The uniformly smooth Banach space is the one with the property

$$\lim_{u\to 0} \rho(u)/u = 0.$$

Rate of convergence

We denote the closure of the convex hull of $\mathcal{D}$ by $A_1(\mathcal{D})$.

Theorem (T., 2001)

Let X be a uniformly smooth Banach space with the modulus of smoothness $\rho(u) \le \gamma u^q$, $1 < q \le 2$. Then for a sequence $\tau := \{t_k\}_{k=1}^{\infty}$, $t_k \le 1$, $k = 1, 2, \dots$, we have for any $f \in A_1(\mathcal{D})$ that

$$\|f_m^{c,\tau}\| \le C(q,\gamma)(1 + \sum_{k=1}^{m} t_k^p)^{-1/p}, \quad p := \frac{q}{q-1},$$

with a constant $C(q,\gamma)$ which may depend only on q and γ.

WGAFR

Weak Greedy Algorithm with Free Relaxation (WGAFR). Let $\tau := \{t_m\}_{m=1}^{\infty}$, $t_m \in [0,1]$, be a weakness sequence. We define $f_0 := f$ and $G_0 := 0$. Then for each $m \geq 1$ we define:

1. $\varphi_m \in \mathcal{D}$ is any element satisfying

$$F_{f_{m-1}}(\varphi_m) \geq t_m \sup_{g \in \mathcal{D}} F_{f_{m-1}}(g).$$

WGAFR

Weak Greedy Algorithm with Free Relaxation (WGAFR). Let $\tau := \{t_m\}_{m=1}^{\infty}$, $t_m \in [0,1]$, be a weakness sequence. We define $f_0 := f$ and $G_0 := 0$. Then for each $m \geq 1$ we define:

1. $\varphi_m \in \mathcal{D}$ is any element satisfying

$$F_{f_{m-1}}(\varphi_m) \geq t_m \sup_{g \in \mathcal{D}} F_{f_{m-1}}(g).$$

2. Find w_m and λ_m such that

$$\|f - ((1-w_m)G_{m-1} + \lambda_m \varphi_m)\| = \inf_{\lambda, w} \|f - ((1-w)G_{m-1} + \lambda \varphi_m)\|$$

and define $G_m := (1-w_m)G_{m-1} + \lambda_m \varphi_m$.

WGAFR

Weak Greedy Algorithm with Free Relaxation (WGAFR). Let $\tau := \{t_m\}_{m=1}^{\infty}$, $t_m \in [0,1]$, be a weakness sequence. We define $f_0 := f$ and $G_0 := 0$. Then for each $m \geq 1$ we define:

1. $\varphi_m \in \mathcal{D}$ is any element satisfying

$$F_{f_{m-1}}(\varphi_m) \geq t_m \sup_{g \in \mathcal{D}} F_{f_{m-1}}(g).$$

2. Find w_m and λ_m such that

$$\|f - ((1-w_m)G_{m-1} + \lambda_m \varphi_m)\| = \inf_{\lambda, w} \|f - ((1-w)G_{m-1} + \lambda \varphi_m)\|$$

and define $G_m := (1-w_m)G_{m-1} + \lambda_m \varphi_m$.

3. Let $f_m := f - G_m$.

Rate of convergence

Theorem (T., 2008)

Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u) \le \gamma u^q$, $1 < q \le 2$. Take a number $\epsilon \ge 0$ and two elements f, f^ϵ from X such that

$$\|f - f^\epsilon\| \le \epsilon, \quad f^\epsilon / B \in A_1(\mathcal{D}),$$

with some number $B = C(f, \epsilon, \mathcal{D}, X) > 0$. Then, for both algorithms WCGA and WGAFR we have ($p := q/(q-1)$)

$$\|f_m\| \le \max\left(2\epsilon, C(q,\gamma)(B+\epsilon)(1+\sum_{k=1}^{m} t_k^p)^{-1/p}\right).$$

Modulus of smoothness

We assume that the set

$$D := \{x : E(x) \le E(0)\}$$

is bounded. For a bounded set D define the modulus of smoothness of E on D as follows

$$\rho(E,u) := \frac{1}{2} \sup_{x \in D, \|y\|=1} |E(x+uy) + E(x-uy) - 2E(x)|. \quad (7)$$

Modulus of smoothness

We assume that the set

$$D := \{x : E(x) \le E(0)\}$$

is bounded. For a bounded set D define the modulus of smoothness of E on D as follows

$$\rho(E,u) := \frac{1}{2} \sup_{x \in D, \|y\|=1} |E(x+uy) + E(x-uy) - 2E(x)|. \quad (7)$$

A typical assumption in convex optimization is of the form ($\|y\| = 1$)

$$|E(x+uy) - E(x) - \langle E'(x), uy \rangle| \le Cu^2$$

which corresponds to the case $\rho(E,u)$ of order u^2. We assume that E is Fréchet differentiable.

The Frank-Wolfe-type algorithm

Let $\tau := \{t_k\}_{k=1}^{\infty}$ be a given weakness sequence of numbers $t_k \in [0,1]$, $k = 1, \dots$.

Weak Relaxed Greedy Algorithm (WRGA(co)). We define $G_0 := G_0^{r,\tau} := 0$. Then, for each $m \geq 1$ we define:

1. $\varphi_m := \varphi_m^{r,\tau} \in \mathcal{D}$ is any element satisfying

$$\langle -E'(G_{m-1}), \varphi_m - G_{m-1} \rangle \geq t_m \sup_{g \in \mathcal{D}} \langle -E'(G_{m-1}), g - G_{m-1} \rangle.$$

The Frank-Wolfe-type algorithm

Let $\tau := \{t_k\}_{k=1}^{\infty}$ be a given weakness sequence of numbers $t_k \in [0,1]$, $k = 1, \dots$.

Weak Relaxed Greedy Algorithm (WRGA(co)). We define $G_0 := G_0^{r,\tau} := 0$. Then, for each $m \geq 1$ we define:

1. $\varphi_m := \varphi_m^{r,\tau} \in \mathcal{D}$ is any element satisfying

$$\langle -E'(G_{m-1}), \varphi_m - G_{m-1} \rangle \geq t_m \sup_{g \in \mathcal{D}} \langle -E'(G_{m-1}), g - G_{m-1} \rangle.$$

2. Find $0 \leq \lambda_m \leq 1$ such that

$$E((1-\lambda_m)G_{m-1} + \lambda_m \varphi_m) = \inf_{0 \leq \lambda \leq 1} E((1-\lambda)G_{m-1} + \lambda \varphi_m)$$

and define $G_m := G_m^{r,\tau} := (1-\lambda_m)G_{m-1} + \lambda_m \varphi_m$.

Rate of approximation

Theorem (T., 2012)

Let E be a uniformly smooth convex function with modulus of smoothness $\rho(E,u) \le \gamma u^q$, $1 < q \le 2$. Then, for a sequence $\tau := \{t_k\}_{k=1}^{\infty}$, $t_k \le 1$, $k = 1, 2, \dots$, we have for any $f \in A_1(\mathcal{D})$ that

$$E(G_m) - E(f) \le \left(C_1(q,\gamma) + C_2(q,\gamma) \sum_{k=1}^{m} t_k^p \right)^{1-q}, \quad p := \frac{q}{q-1},$$

with positive constants $C_1(q,\gamma)$, $C_2(q,\gamma)$ which may depend only on q and γ.

WGAFR(co)

Weak Greedy Algorithm with Free Relaxation (WGAFR(co)). Let $\tau := \{t_m\}_{m=1}^{\infty}$, $t_m \in [0,1]$, be a weakness sequence. We define $G_0 := 0$. Then for each $m \geq 1$ we have:

1. $\varphi_m \in \mathcal{D}$ is any element satisfying

$$\langle -E'(G_{m-1}), \varphi_m \rangle \geq t_m \sup_{g \in \mathcal{D}} \langle -E'(G_{m-1}), g \rangle.$$

WGAFR(co)

Weak Greedy Algorithm with Free Relaxation (WGAFR(co)). Let $\tau := \{t_m\}_{m=1}^{\infty}$, $t_m \in [0,1]$, be a weakness sequence. We define $G_0 := 0$. Then for each $m \geq 1$ we have:

1. $\varphi_m \in \mathcal{D}$ is any element satisfying

$$\langle -E'(G_{m-1}), \varphi_m \rangle \geq t_m \sup_{g \in \mathcal{D}} \langle -E'(G_{m-1}), g \rangle.$$

2. Find w_m and λ_m such that

$$E((1-w_m)G_{m-1} + \lambda_m \varphi_m) = \inf_{\lambda, w} E((1-w)G_{m-1} + \lambda \varphi_m)$$

and define $G_m := (1-w_m)G_{m-1} + \lambda_m \varphi_m$.

Rate of convergence for WGAFR(co)

Theorem (T, 2012)

Let E be a uniformly smooth convex function with modulus of smoothness $\rho(E,u) \le \gamma u^q$, $1 < q \le 2$. Take a number $\epsilon \ge 0$ and an element f^ϵ from D such that

$$E(f^\epsilon) \le \inf_{x \in D} E(x) + \epsilon, \quad f^\epsilon / B \in A_1(\mathcal{D}),$$

with some number $B = C(E, \epsilon, \mathcal{D}) \ge 1$. Then we have ($p := q/(q-1)$)

$$E(G_m) - \inf_{x \in D} E(x) \le$$

$$\max\left(2\epsilon, C_1(E,q,\gamma) B^q \left(C_2(E,q,\gamma) + \sum_{k=1}^m t_k^p\right)^{1-q}\right).$$

Gradient type algorithms

The most difficult part of an algorithm is to find an element $\varphi_m \in \mathcal{D}$ to be used in approximation process. We consider greedy methods for finding $\varphi_m \in \mathcal{D}$. We have two types of greedy steps to find $\varphi_m \in \mathcal{D}$.

I. Gradient greedy step. At this step we look for an element $\varphi_m \in \mathcal{D}$ such that

$$\langle -E'(G_{m-1}), \varphi_m \rangle \geq t_m \sup_{g \in \mathcal{D}} \langle -E'(G_{m-1}), g \rangle.$$

Algorithms that use the first derivative of the objective function E are called *first order* optimization algorithms.

Zero order algorithms

II. E-greedy step. At this step we look for an element $\varphi_m \in \mathcal{D}$ which satisfies (we assume existence):

$$\inf_{c\in\mathbb{R}} E(G_{m-1} + c\varphi_m) = \inf_{g\in\mathcal{D},c\in\mathbb{R}} E(G_{m-1} + cg).$$

Algorithms that only use the values of the objective function E are called *zero order* optimization algorithms.

Approximation step

After we found $\varphi_m \in \mathcal{D}$ we can proceed in different ways. We now list some typical steps that are motivated by the corresponding steps in greedy approximation theory. These steps or their variants are used in optimization algorithms like *gradient method*, *reduced gradient method*, *conjugate gradients*, *gradient pursuits*.

(A) Best step in the direction $\varphi_m \in \mathcal{D}$. We choose c_m such that

$$E(G_{m-1} + c_m\varphi_m) = \inf_{c\in\mathbb{R}} E(G_{m-1} + c\varphi_m)$$

and define

$$G_m := G_{m-1} + c_m\varphi_m.$$

Other approximation steps

(B) Shortened best step in the direction $\varphi_m \in \mathcal{D}$. We choose c_m as in (A) and for a given parameter $b > 0$ define

$$G_m^b := G_{m-1}^b + bc_m\varphi_m.$$

Usually, $b \in (0,1)$. This is why we call it *shortened*.

(C) Chebyshev-type (fully corrective) methods. We choose $G_m \in \operatorname{span}(\varphi_1, \dots, \varphi_m)$ which satisfies

$$E(G_m) = \inf_{c_j, j=1,\dots,m} E(c_1\varphi_1 + \cdots + c_m\varphi_m).$$

(D) Fixed relaxation. For a given sequence $\{r_k\}_{k=1}^{\infty}$ of relaxation parameters $r_k \in [0,1)$ we choose $G_m := (1-r_m)G_{m-1} + c_m\varphi_m$ with c_m from

$$E((1-r_m)G_{m-1} + c_m\varphi_m) = \inf_{c\in\mathbb{R}} E((1-r_m)G_{m-1} + c\varphi_m).$$

More approximation steps

(F) Free relaxation. We choose $G_m \in \operatorname{span}(G_{m-1}, \varphi_m)$ which satisfies

$$E(G_m) = \inf_{c_1, c_2} E(c_1 G_{m-1} + c_2 \varphi_m).$$

(G) Prescribed coefficients. For a given sequence $\{c_k\}_{k=1}^{\infty}$ of positive coefficients in the case of greedy step **I** we define

$$G_m := G_{m-1} + c_m \varphi_m. \tag{8}$$

In the case of greedy step **II** we define G_m by formula (8) with the greedy step **II** modified as follows: $\varphi_m \in \mathcal{D}$ is an element satisfying

$$E(G_{m-1} + c_m \varphi_m) = \inf_{g \in \mathcal{D}} E(G_{m-1} + c_m g).$$

Problem

We are interested in the following fundamental problem of sparse approximation.

Problem

How to design a practical algorithm that builds sparse approximations comparable to best m-term approximations?

Problem

We are interested in the following fundamental problem of sparse approximation.

Problem

How to design a practical algorithm that builds sparse approximations comparable to best m-term approximations?

In other words: How to choose elements from the dictionary for good m-term approximation?

Haar basis

Remark

In the case $X = L_p$, $1 < p < \infty$, $\mathcal{D} = \mathcal{H}_p$, *the recipe is very simple: for a given* $f \in L_p$,

$$f = \sum_I c_I(f) H_{I,p},$$

choose those $H_{I,p}$ *for which the* $|c_I(f)|$ *are the largest.*

Haar basis

Remark

In the case $X = L_p$, $1 < p < \infty$, $\mathcal{D} = \mathcal{H}_p$, *the recipe is very simple: for a given* $f \in L_p$,

$$f = \sum_I c_I(f) H_{I,p},$$

choose those $H_{I,p}$ *for which the* $|c_I(f)|$ *are the largest.*

This discovery led to the actively developing theory of greedy-type bases. The above recipe gives us the Thresholding Greedy Algorithm (TGA).

Lebesgue-type inequality for the TGA

Theorem (T., 1998)

For each $f \in L_p(\mathbb{T}^d)$ *we have*

$$\|f - G_m(f,\mathcal{T})\|_p \le (1 + 3m^{h(p)})\sigma_m(f,\mathcal{T})_p, \quad 1 \le p \le \infty,$$

where $h(p) := |1/2 - 1/p|$.

Lebesgue-type inequality for the TGA

Theorem (T., 1998)

For each $f \in L_p(\mathbb{T}^d)$ *we have*

$$\|f - G_m(f,\mathcal{T})\|_p \le (1 + 3m^{h(p)})\sigma_m(f,\mathcal{T})_p, \quad 1 \le p \le \infty,$$

where $h(p) := |1/2 - 1/p|$.

Remark

There is a positive absolute constant C *such that for each* m *and* $1 \le p \le \infty$ *there exists a function* $f \ne 0$ *with the property*

$$\|G_m(f,\mathcal{T})\|_p \ge Cm^{h(p)}\|f\|_p.$$

Problem for the trigonometric system

Thus the recipe that works well for the Haar basis does not work well for the trigonometric system.

Problem

How to choose harmonics (frequencies) for good m-term trigonometric approximation?

Problem for the trigonometric system

Thus the recipe that works well for the Haar basis does not work well for the trigonometric system.

Problem

How to choose harmonics (frequencies) for good m-term trigonometric approximation?

Remark

It turns out that the following recipe works well for $2 \le p < \infty$. We describe the mth iteration for approximating f. Suppose f_{m-1} is the residual after $m-1$ iterations. Then we look for the largest Fourier coefficient of the function $f_{m-1}|f_{m-1}|^{p-2}$ and choose the corresponding harmonic.

Lebesgue-type inequality for the WCGA

Theorem (T.,2013)

Let $\mathcal{D}$ be the normalized in L_p, $2 \le p < \infty$, real d-variate trigonometric system. Then for any $f \in L_p$ the WCGA with weakness parameter t gives

$$\|f_{C(t,p,d)m\ln(m+1)}\|_p \le C\sigma_m(f,\mathcal{D})_p. \tag{9}$$

Lebesgue-type inequality for the WCGA

Theorem (T.,2013)

Let $\mathcal{D}$ be the normalized in L_p, $2 \le p < \infty$, real d-variate trigonometric system. Then for any $f \in L_p$ the WCGA with weakness parameter t gives

$$\|f_{C(t,p,d)m\ln(m+1)}\|_p \le C\sigma_m(f,\mathcal{D})_p. \tag{9}$$

The Open Problem 7.1 (p. 91) from [Temlyakov, 2003] asks if (9) holds without an extra $\ln(m+1)$ factor. The above theorem is the first result on the Lebesgue-type inequalities for the WCGA with respect to the trigonometric system. It provides a progress in solving the above mentioned open problem, but the problem is still open.

THANK YOU!