From greedy approximation to greedy optimization

Vladimir Temlyakov

December 10, 2013
1. Introduction

2. Greedy approximation in Hilbert spaces

3. Greedy approximation in Banach spaces

4. Greedy algorithms for convex optimization
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

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

$$f = \sum_{i=1}^{\infty} \langle f, \psi_{k_i} \rangle \psi_{k_i}, \quad |\langle f, \psi_{k_1} \rangle| \geq |\langle f, \psi_{k_2} \rangle| \geq \ldots.$$
Major questions of greedy approximation

1. Let instead of an orthonormal basis $\Psi$ 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 $\Psi$ we have a redundant system $D$. How to approximate with regard to $D$?
2. How to work in a Banach space $X$ instead of a Hilbert space $H$?
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 $\text{Span } \mathcal{D}$ 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 $\mathcal{D}$. 
Thus each function $s \in \Sigma_m(\mathcal{D})$ can be written in the form

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

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

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

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

$$W_m(x) = \sum_{j=1}^{m} r_j(\langle \omega_j, x \rangle),$$

where $r_j, j = 1, \ldots, m$, are univariate functions.
Another example, from signal processing, uses the Gabor functions
\[ g_{a,b}(x) := e^{i a x} e^{-b x^2} \]
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

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

We concentrate on an important problem of finding good methods of $m$-term approximation in the case of general dictionary $D$ 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 $D$. 
Let a sequence \( \tau = \{t_k\}_{k=1}^{\infty}, 0 \leq t_k \leq 1 \), be given. \[ \text{WGA}\] We define \( f^\tau_0 := f \). Then for each \( m \geq 1 \), we inductively define:

1. \( \varphi^\tau_m \in D \) is any satisfying

\[
|\langle f^\tau_{m-1}, \varphi^\tau_m \rangle| \geq t_m \sup_{g \in D} |\langle f^\tau_{m-1}, g \rangle|;
\]
Let a sequence $\tau = \{t_k\}_{k=1}^{\infty}$, $0 \leq t_k \leq 1$, be given. 

**WGA** We define $f_{\tau}^0 := f$. Then for each $m \geq 1$, we inductively define:

1. $\phi_{\tau}^m \in D$ is any satisfying

   $$\left| \langle f_{m-1}^\tau, \phi_{\tau}^m \rangle \right| \geq t_m \sup_{g \in D} \left| \langle f_{m-1}^\tau, g \rangle \right|;$$

2. $f_{\tau}^m := f_{m-1}^\tau - \langle f_{m-1}^\tau, \phi_{\tau}^m \rangle \phi_{\tau}^m;$
Let a sequence $\tau = \{ t_k \}_{k=1}^{\infty}$, $0 \leq t_k \leq 1$, be given.

**WGA** We define $f_0^\tau := f$. Then for each $m \geq 1$, we inductively define:

1. $\varphi_m^\tau \in \mathcal{D}$ is any satisfying
   \[
   |\langle f_{m-1}^\tau, \varphi_m^\tau \rangle| \geq t_m \sup_{g \in \mathcal{D}} |\langle f_{m-1}^\tau, g \rangle|;
   \]

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

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

Vladimir Temlyakov
From greedy approximation to greedy optimization
In the case $t_k = 1$, $k = 1, \ldots$ 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 $t_k = t$, $k = 1, 2, \ldots$, 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)

$$f \sim \sum_{j=1}^{\infty} c_j(f) \varphi_j^T,$$

$$c_j(f) := \langle f_{j-1}^T, \varphi_j^T \rangle.$$

In general it is not an expansion into an orthogonal series but it has some similar properties.
The coefficients $c_j(f)$ of an expansion are obtained by the Fourier formulas with $f$ replaced by the residuals $f_{j-1}^\tau$. It is easy to see that

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

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

$$
A^o_1(\mathcal{D}, M) := \{f \in H : f = \sum_{k \in \Lambda} c_k w_k, \; w_k \in \mathcal{D}, \; \#\Lambda < \infty \}
$$

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

and we define $A_1(\mathcal{D}, M)$ as the closure (in $H$) of $A^o_1(\mathcal{D}, M)$. Furthermore, we define $A_1(\mathcal{D})$ as the union of the classes $A_1(\mathcal{D}, M)$ over all $M > 0$. For $f \in A_1(\mathcal{D})$, we define the norm

$$
|f|_{A_1(\mathcal{D})}
$$

as the smallest $M$ such that $f \in A_1(\mathcal{D}, M)$.
It was proved in [DeVore, T., 1996] that for a general dictionary $\mathcal{D}$ the Pure Greedy Algorithm provides the following estimate

$$\|f - G_m(f, \mathcal{D})\| \leq |f|_{A_1(\mathcal{D})} m^{-1/6}. \quad (1)$$

(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

$$\|f - G_m(f, \mathcal{D})\| > \frac{1}{2} |f|_{A_1(\mathcal{D})} m^{-1/2}, \quad m \geq 4.$$
We proved in [Konyagin, T., 1999] an estimate
\[ \| f - G_m(f, \mathcal{D}) \| \leq 4 |f|_{A_1(\mathcal{D})} m^{-11/62} \]
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
\[ \| f - G_m(f, \mathcal{D}) \| \geq C m^{-0.27} |f|_{A_1(\mathcal{D})} \]
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.
Find the right order of the sequence

$$\sup_{H,D} \frac{\|f - G_m(f, D)\|}{\|f\|_{A_1(D)}}.$$
Let a sequence \( \tau = \{ t_k \}_{k=1}^{\infty}, 0 \leq t_k \leq 1 \), be given. We define the Weak Orthogonal Greedy Algorithm (WOGA).

WOGA We define \( f_{0,\tau} := f \). Then for each \( m \geq 1 \) we inductively define:

1. \( \varphi_{m,\tau} \in D \) is any element satisfying

\[
|\langle f_{m-1,\tau}, \varphi_{m,\tau} \rangle| \geq t_m \sup_{g \in D} |\langle f_{m-1,\tau}, g \rangle|;
\]
Let a sequence $\tau = \{t_k\}_{k=1}^{\infty}$, $0 \leq t_k \leq 1$, be given. We define the Weak Orthogonal Greedy Algorithm (WOGA).

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

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

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

2. $G_m^{o,\tau}(f, \mathcal{D}) := P_{H_m^\tau}(f)$, where $H_m^\tau := \text{Span}(\varphi_1^{o,\tau}, \ldots, \varphi_m^{o,\tau})$;
Let a sequence \( \tau = \{t_k\}_{k=1}^{\infty}, 0 \leq t_k \leq 1 \), be given. We define the Weak Orthogonal Greedy Algorithm (WOGA).

**WOGA** We define \( f_{0,\tau}^o := f \). Then for each \( m \geq 1 \) we inductively define:

1. \( \varphi_{m,\tau}^o \in D \) is any element satisfying
   \[
   |\langle f_{m-1,\tau}^o, \varphi_{m,\tau}^o \rangle| \geq t_m \sup_{g \in D} |\langle f_{m-1,\tau}^o, g \rangle|;
   \]

2. \( G_{m,\tau}^o (f, D) := P_{H_{m,\tau}^o} (f) \), where \( H_{m,\tau}^o := \text{Span}(\varphi_{1,\tau}^o, \ldots, \varphi_{m,\tau}^o) \);

3. \( f_{m,\tau}^o := f - G_{m,\tau}^o (f, D) \).
Rate of convergence

Theorem (T., 2000)

Let $\mathcal{D}$ be an arbitrary dictionary in $H$. Then for each $f \in A_1(\mathcal{D}, M)$ we have

$$\|f - G^o,\tau_m(f, \mathcal{D})\| \leq M(1 + \sum_{k=1}^{m} t_k^2)^{-1/2}.$$
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 ($\|g\| = 1$), $g \in \mathcal{D}$ implies $-g \in \mathcal{D}$, and closure of $\text{Span} \mathcal{D} = X$.

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

$$\|F_f\| = 1, \quad F_f(f) = \|f\|.$$
Two forms

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

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

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

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

$$\langle f_{m-1}, \varphi_m \rangle = \sup_{g \in \mathcal{D}} \langle f_{m-1}, g \rangle.$$  \hfill (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.
Two forms

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

1. First, we look at the $m$th step for an element $\varphi_m \in D$ and a number $\lambda_m$ satisfying

$$\|f_m - 1 - \lambda_m \varphi_m\|_H = \inf_{g \in D, \lambda} \|f_m - 1 - \lambda g\|_H. \quad (2)$$

2. Second, we look for an element $\varphi_m \in D$ such that

$$\langle f_m - 1, \varphi_m \rangle = \sup_{g \in D} \langle f_m - 1, g \rangle. \quad (3)$$

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

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

\textbf{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

\[
f_m := f_{m-1} - \lambda_m \varphi_m, \quad G_m := G_{m-1} + \lambda_m \varphi_m.
\]
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

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

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($\tau$)).
Weak Dual Greedy Algorithm (WDGA(\(\tau\))) 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

\[ \varphi_m \in \mathcal{D} \text{ is any satisfying} \]

\[ F_{f_{m-1}}(\varphi_m) \geq t_m \| F_{f_{m-1}} \| \mathcal{D}. \]  

(6)
Weak Dual Greedy Algorithm (WDGA(\(\tau\))) 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

\[F_{f_{m-1}}(\varphi_m) \geq t_m \|F_{f_{m-1}}\|_\mathcal{D} .\]  

2. Define \(a_m\) as

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

\[
F_{f_{m-1}}(\varphi_m) \geq t_m \| F_{f_{m-1}} \| \mathcal{D}.
\] (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

\[
f_m := f_{m-1} - a_m \varphi_m.
\]
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 $t_k \leq 1$, $k = 1, \ldots$. 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** We define $f_0^c := f_{0,\tau}^c := f$. Then for each $m \geq 1$ we inductively define

1. $\varphi_m^c := \varphi_{m,\tau}^c \in \mathcal{D}$ 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** 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

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

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

Vladimir Temlyakov
From greedy approximation to greedy optimization
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

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

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

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

$$\lim_{{u \to 0}} \rho(u)/u = 0.$$
We denote the closure of the convex hull of \( D \) by \( A_1(D) \).

**Theorem (T., 2001)**

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

\[
\|f^{c,\tau}_m\| \leq C(q, \gamma)(1 + \sum_{k=1}^m t_k^p)^{-1/p}, \quad p := \frac{q}{q - 1},
\]

with a constant \( C(q, \gamma) \) which may depend only on \( q \) and \( \gamma \).
Weak Greedy Algorithm with Free Relaxation (WGAFR). Let \( \tau := \{t_m\}_{m=1}^{\infty}, \quad t_m \in [0, 1], \) 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

\[
F_{f_{m-1}}(\varphi_m) \geq t_m \sup_{g \in \mathcal{D}} F_{f_{m-1}}(g).
\]
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. \( \varphi_m \in D \) is any element satisfying
   \[
   F_{f_{m-1}}(\varphi_m) \geq t_m \sup_{g \in D} F_{f_{m-1}}(g).
   \]

2. Find \( w_m \) and \( \lambda_m \) 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 \( G_m := (1 - w_m)G_{m-1} + \lambda_m\varphi_m. \)
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 \( \lambda_m \) 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 \( f_m := f - G_m. \)
Theorem (T., 2008)

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

$$\|f - f^\epsilon\| \leq \epsilon, \quad f^\epsilon/B \in A_1(D),$$

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

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

We assume that the set

\[ D := \{ x : E(x) \leq E(0) \} \]

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

\[
\rho(E, u) := \frac{1}{2} \sup_{x \in D, \|y\| = 1} \left| E(x + uy) + E(x - uy) - 2E(x) \right|.
\] (7)
We assume that the set
\[ D := \{ x : E(x) \leq E(0) \} \]
is bounded. For a bounded set \( D \) define the modulus of smoothness of \( E \) on \( D \) as follows
\[
\rho(E, u) := \frac{1}{2} \sup_{x \in D, \|y\| = 1} |E(x + uy) + E(x - uy) - 2E(x)|. \tag{7}
\]
A typical assumption in convex optimization is of the form \( (\|y\| = 1) \)
\[
|E(x + uy) - E(x) - \langle E'(x), uy \rangle| \leq Cu^2
\]
which corresponds to the case \( \rho(E, u) \) of order \( u^2 \). We assume that \( E \) is Fréchet differentiable.
Let \( \tau := \{t_k\}_{k=1}^\infty \) be a given weakness sequence of numbers \( t_k \in [0, 1], \ k = 1, \ldots. \)

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

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

\[
\langle -E'(G_{m-1}), \varphi_m - G_{m-1} \rangle \geq t_m \sup_{g \in D} \langle -E'(G_{m-1}), g - G_{m-1} \rangle.
\]
The Frank-Wolfe-type algorithm

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

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

1. \( \varphi_m := \varphi_{m}^{r, \tau} \in D \) is any element satisfying

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

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 
\( G_m := G_{m}^{r, \tau} := (1 - \lambda_m)G_{m-1} + \lambda_m \varphi_m. \)
Theorem (T., 2012)

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

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

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

Weak Greedy Algorithm with Free Relaxation (WGAFR(co)). Let $	au := \{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

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


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 \( \lambda_m \) 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
\[
G_m := (1 - w_m)G_{m-1} + \lambda_m \varphi_m.
\]
Theorem (T, 2012)

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

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

with some number $B = C(E, \epsilon, D) \geq 1$. Then we have

$$(p := q/(q-1))$$

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

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

\[
G^b_m := G^b_{m-1} + bc_m \varphi_m.
\]

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

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

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

(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 \( G_m \in \text{span}(G_{m-1}, \varphi_m) \) 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

\[
E(G_{m-1} + c_m \varphi_m) = \inf_{g \in \mathcal{D}} E(G_{m-1} + c_m g).
\]
THANK YOU!