Some remarks on convex optimization on Banach spaces

Vladimir Temlyakov

Innopolis, October 11, 2024

Setting

A typical problem of convex optimization is to find an approximate solution to the problem

$$\inf_{x \in D} E(x) \tag{1}$$

under assumption that E is a convex function. We consider a convex function E defined on a Banach space X. Let X be a Banach space with norm $\|\cdot\|$.

We say that a set of elements (functions) $\mathcal{D}$ from X is a dictionary if each $g \in \mathcal{D}$ has norm bounded by one ($\|g\| \le 1$) and the closure of $\operatorname{span} \mathcal{D}$ is X.

We denote the closure (in X) of the convex hull of $\mathcal{D}^{\pm} := \{\pm g, g \in \mathcal{D}\}$ by $A_1(\mathcal{D})$.

Typical constraints

Sparsity Constraints: The set $\Sigma_n(\mathcal{D})$ of functions

$$g = \sum_{g \in \Lambda} c_g g, \quad |\Lambda| = n,$$

is called the set of *sparse* functions of order n with respect to the dictionary $\mathcal{D}$. One common assumption is to minimize E on $D = \Sigma_n(\mathcal{D})$, i.e. to look for an n sparse minimizer of (1).

Typical constraints

Sparsity Constraints: The set $\Sigma_n(\mathcal{D})$ of functions

$$g = \sum_{g \in \Lambda} c_g g, \quad |\Lambda| = n,$$

is called the set of *sparse* functions of order n with respect to the dictionary $\mathcal{D}$. One common assumption is to minimize E on $D = \Sigma_n(\mathcal{D})$, i.e. to look for an n sparse minimizer of (1).
ℓ_1 constraints: Minimize E over $A_1(\mathcal{D})$. A slightly more general setting is to minimize E over one of the sets

$$\mathcal{L}_M := \{g \in X : \ g/M \in A_1(\mathcal{D})\}.$$

Example

There has been considerable interest in solving the convex unconstrained optimization problem

$$\min_x \frac{1}{2}\|y - \Phi x\|_2^2 + \lambda \|x\|_1 \tag{2}$$

where $x \in \mathbb{R}^n$, $y \in \mathbb{R}^k$, Φ is an $k \times n$ matrix, λ is a nonnegative parameter, $\|v\|_2$ denotes the Euclidian norm of v, and $\|v\|_1$ is the ℓ_1 norm of v.

Example

There has been considerable interest in solving the convex unconstrained optimization problem

$$\min_x \frac{1}{2}\|y - \Phi x\|_2^2 + \lambda \|x\|_1 \tag{2}$$

where $x \in \mathbb{R}^n$, $y \in \mathbb{R}^k$, Φ is an $k \times n$ matrix, λ is a nonnegative parameter, $\|v\|_2$ denotes the Euclidian norm of v, and $\|v\|_1$ is the ℓ_1 norm of v.
Problems of the form (2) have become familiar over the past three decades, particularly in statistical and signal processing contexts.
Problem (2) is closely related to the following convex constrained optimization problem

$$\min_x \frac{1}{2}\|y - \Phi x\|_2^2 \quad \text{subject to} \quad \|x\|_1 \leq A. \tag{3}$$

Optimization in $\mathbb{R}^n$

The problem (3) is a convex optimization problem of the energy function $E(x, \Phi) := \frac{1}{2}\|y - \Phi x\|_2^2$ on the octahedron $\{x : \|x\|_1 \leq A\}$ in $\mathbb{R}^n$.

Optimization in $\mathbb{R}^n$

The problem (3) is a convex optimization problem of the energy function $E(x, \Phi) := \frac{1}{2}\|y - \Phi x\|_2^2$ on the octahedron $\{x : \|x\|_1 \le A\}$ in $\mathbb{R}^n$.

- The domain of optimization is simple and all dependence on the matrix Φ is in the energy function $E(x, \Phi)$, which makes the problem difficult.

Optimization in $\mathbb{R}^n$

The problem (3) is a convex optimization problem of the energy function $E(x, \Phi) := \frac{1}{2}\|y - \Phi x\|_2^2$ on the octahedron $\{x : \|x\|_1 \leq A\}$ in $\mathbb{R}^n$.

- The domain of optimization is simple and all dependence on the matrix Φ is in the energy function $E(x, \Phi)$, which makes the problem difficult.
- Also, the domain is in the high dimensional space $\mathbb{R}^n$.

Recast of the above example

In typical applications, for instance in compressed sensing, k is much smaller than n. We recast the above problem as an optimization problem with $E(z) := \frac{1}{2}\|y - z\|_2^2$ over the domain $A_1(\mathcal{D})$.

Recast of the above example

In typical applications, for instance in compressed sensing, k is much smaller than n. We recast the above problem as an optimization problem with $E(z) := \frac{1}{2}\|y - z\|_2^2$ over the domain $A_1(\mathcal{D})$.

- We optimize with respect to a dictionary $\mathcal{D} := \{\varphi_i\}_{i=1}^n$, which is associated with a $k \times n$ matrix $\Phi = [\varphi_1 \dots \varphi_n]$ with $\varphi_j \in \mathbb{R}^k$ being the column vectors of Φ.

Recast of the above example

In typical applications, for instance in compressed sensing, k is much smaller than n. We recast the above problem as an optimization problem with $E(z) := \frac{1}{2}\|y - z\|_2^2$ over the domain $A_1(\mathcal{D})$.

- We optimize with respect to a dictionary $\mathcal{D} := \{\varphi_i\}_{i=1}^n$, which is associated with a $k \times n$ matrix $\Phi = [\varphi_1 \dots \varphi_n]$ with $\varphi_j \in \mathbb{R}^k$ being the column vectors of Φ.
- In this formulation the energy function $E(z)$ is very simple and all dependence on Φ is in the form of the domain $A_1(\mathcal{D})$.

Recast of the above example

In typical applications, for instance in compressed sensing, k is much smaller than n. We recast the above problem as an optimization problem with $E(z) := \frac{1}{2}\|y - z\|_2^2$ over the domain $A_1(\mathcal{D})$.

- We optimize with respect to a dictionary $\mathcal{D} := \{\varphi_i\}_{i=1}^n$, which is associated with a $k \times n$ matrix $\Phi = [\varphi_1 \dots \varphi_n]$ with $\varphi_j \in \mathbb{R}^k$ being the column vectors of Φ.
- In this formulation the energy function $E(z)$ is very simple and all dependence on Φ is in the form of the domain $A_1(\mathcal{D})$.
- Other important feature of the new formulation is that optimization takes place in the $\mathbb{R}^k$ with relatively small k.

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

$$\rho(E,u) := \frac{1}{2} \sup_{x \in D, \|y\|=1} |E(x+uy) + E(x-uy) - 2E(x)|. \quad (4)$$

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

$$\rho(E,u) := \frac{1}{2} \sup_{x \in D, \|y\|=1} |E(x+uy) + E(x-uy) - 2E(x)|. \quad (4)$$

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 $t \in (0, 1]$ be a given weakness parameter.

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

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

$$\langle -E'(G_{m-1}), \varphi_m - G_{m-1} \rangle \geq t \sup_{g \in \mathcal{D}} \langle -E'(G_{m-1}), g - G_{m-1} \rangle.$$

The Frank-Wolfe-type algorithm

Let $t \in (0,1]$ be a given weakness parameter.
Weak Relaxed Greedy Algorithm (WRGA(co)). We define $G_0 := 0$.
Then, for each $m \geq 1$ we define:

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

$$\langle -E'(G_{m-1}), \varphi_m - G_{m-1} \rangle \geq t \sup_{g \in \mathcal{D}} \langle -E'(G_{m-1}), g - G_{m-1} \rangle.$$

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

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 parameter $t \in (0,1]$ we have

$$E(G_m) - \inf_{f \in A_1(\mathcal{D})} E(f) \le C(t,q,\gamma) m^{1-q}.$$

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 parameter $t \in (0,1]$ we have

$$E(G_m) - \inf_{f \in A_1(\mathcal{D})} E(f) \le C(t,q,\gamma) m^{1-q}.$$

In the case $q = 2$ it goes back to Frank and Wolfe, 1956 (special case) and to Tewari, Ravikumar, and Dhillon, 2012.

Chebyshev algorithm

Weak Chebyshev Greedy Algorithm (WCGA(co)). Let $t \in (0,1]$ be a weakness parameter. We define $G_0 := 0$. Then for each $m \geq 1$ we have the following inductive definition.

- (1) $\varphi_m \in \mathcal{D}$ is any element satisfying

$$|\langle -E'(G_{m-1}), \varphi_m \rangle| \geq t \sup_{g \in \mathcal{D}} |\langle -E'(G_{m-1}), g \rangle|.$$

Chebyshev algorithm

Weak Chebyshev Greedy Algorithm (WCGA(co)). Let $t \in (0,1]$ be a weakness parameter. We define $G_0 := 0$. Then for each $m \geq 1$ we have the following inductive definition.

- (1) $\varphi_m \in \mathcal{D}$ is any element satisfying

$$|\langle -E'(G_{m-1}), \varphi_m \rangle| \geq t \sup_{g \in \mathcal{D}} |\langle -E'(G_{m-1}), g \rangle|.$$

- (2) Define $\Phi_m := \operatorname{span}\{\varphi_j\}_{j=1}^m$, and define G_m to be the point from Φ_m at which E attains the minimum: $E(G_m) = \inf_{x \in \Phi_m} E(x)$.

Chebyshev algorithm with function evaluations

E-Greedy Chebyshev Algorithm (EGCA(co)). We define $G_0 := 0$. Then for each $m \geq 1$ we have the following inductive definition.

- (1) $\varphi_m \in \mathcal{D}$ is any element satisfying (assume existence)

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

Chebyshev algorithm with function evaluations

E-Greedy Chebyshev Algorithm (EGCA(co)). We define $G_0 := 0$. Then for each $m \geq 1$ we have the following inductive definition.

- (1) $\varphi_m \in \mathcal{D}$ is any element satisfying (assume existence)

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

- (2) Define $\Phi_m := \operatorname{span}\{\varphi_j\}_{j=1}^m$, and define G_m to be the point from Φ_m at which E attains the minimum: $E(G_m) = \inf_{x \in \Phi_m} E(x)$.

WGAFR(co)

Weak Greedy Algorithm with Free Relaxation (WGAFR(co)). Let $t \in (0,1]$, be a weakness parameter. We define $G_0 := 0$. Then for each $m \geq 1$ we have:

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

$$|\langle -E'(G_{m-1}), \varphi_m \rangle| \geq t \sup_{g \in \mathcal{D}} |\langle -E'(G_{m-1}), g \rangle|.$$

WGAFR(co)

Weak Greedy Algorithm with Free Relaxation (WGAFR(co)). Let $t \in (0,1]$, be a weakness parameter. We define $G_0 := 0$. Then for each $m \geq 1$ we have:

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

$$|\langle -E'(G_{m-1}), \varphi_m \rangle| \geq t \sup_{g \in \mathcal{D}} |\langle -E'(G_{m-1}), g \rangle|.$$

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

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 for the WCGA(co), ECGA(co), and WGAFR(co)

$$E(G_m) - \inf_{x \in D} E(x) \le \max\left(2\epsilon, C_1(t,E,q,\gamma) B^q m^{1-q}\right).$$

Extra conditions on E

E1. Smoothness. We assume that E is a convex function with $\rho(E, u) \leq \gamma u^2$.

Extra conditions on E

E1. Smoothness. We assume that E is a convex function with $\rho(E,u) \le \gamma u^2$.

E2. Restricted strong convexity. We assume that for any S-sparse element f we have

$$E(f) - E(f_0) \ge \beta \|f - f_0\|^2. \tag{5}$$

Incoherence property of a dictionary

Definition (A)

A. We say that $f = \sum_{i \in T} x_i g_i$ has ℓ_1 incoherence property with parameters S, V, and r if for any $A \subset T$ and any Λ such that $A \cap \Lambda = \emptyset$, $|A| + |\Lambda| \le S$ we have for any $\{c_i\}$

$$\sum_{i \in A} |x_i| \le V |A|^r \| f_A - \sum_{i \in \Lambda} c_i g_i \|. \tag{6}$$

Incoherence property of a dictionary

Definition (A)

A. We say that $f = \sum_{i \in T} x_i g_i$ has ℓ_1 incoherence property with parameters S, V, and r if for any $A \subset T$ and any Λ such that $A \cap \Lambda = \emptyset$, $|A| + |\Lambda| \le S$ we have for any $\{c_i\}$

$$\sum_{i \in A} |x_i| \le V |A|^r \| f_A - \sum_{i \in \Lambda} c_i g_i \|. \tag{6}$$

A dictionary $\mathcal{D}$ has ℓ_1 incoherence property with parameters K, S, V, and r if for any $A \subset B$, $|A| \le K$, $|B| \le S$ we have for any $\{c_i\}_{i \in B}$

$$\sum_{i \in A} |c_i| \le V |A|^r \| \sum_{i \in B} c_i g_i \|.$$

Exponential rate of convergence

Theorem (T., 2013)

Let E satisfy assumptions **E1** *and* **E2**. *Suppose for a point of minimum f_0 we have $\|f_0 - f^\epsilon\| \le \epsilon$ with K-sparse $f := f^\epsilon$ satisfying property* **A**. *Then for the WCGA(co) with weakness parameter t and for the EGCA(co) we have for $K + m \le S$*

$$E(G_m) - E(f_0)$$

$$\le \max\left((E(0) - E(f_0))\exp\left(-\frac{c_1 m}{K^{2r}}\right), 8(\gamma^2/\beta)\epsilon^2\right) + 2\gamma\epsilon^2,$$

where $c_1 := \frac{\beta t^2}{64\gamma V^2}$.

The End

Thank you!