Blended Matching Pursuit

A sparse unregularized method for convex minimization

Cyrille W. Combettes

Joint work with Sebastian Pokutta

Georgia Tech H. Milton Stewart School of Industrial and Systems Engineering

Second Conference on Discrete Optimization and Machine Learning
RIKEN Center for Advanced Intelligence Project, Tokyo, Japan

July 30th, 2019

Table of contents

Problem

- Let $f : \mathbb{R}^n \to \mathbb{R}$ be a smooth and convex function and $\mathcal{D} \subset \mathbb{R}^n$ be a normalized and symmetrical basis, possibly overcomplete
- Find a sparse (relative to $\mathcal{D}$) ϵ-approximate solution to

$$\min_{x \in \mathbb{R}^n} f(x)$$

Problem

- Let $f : \mathbb{R}^n \to \mathbb{R}$ be a smooth and convex function and $\mathcal{D} \subset \mathbb{R}^n$ be a normalized and symmetrical basis, possibly overcomplete
- Find a sparse (relative to $\mathcal{D}$) ϵ-approximate solution to

$$\min_{x \in \mathbb{R}^n} f(x)$$

- Build a point $x = \sum_{i=1}^{m} \lambda_i v_i$ ($v_i \in \mathcal{D}$) such that m is small and

$$f\left(\sum_{i=1}^{m} \lambda_i v_i\right) \leqslant \min_{\mathbb{R}^n} f + \epsilon$$

Goals

- Provide an alternative to the constrained/regularized methods which require tuning of s or λ:

$$\begin{aligned} \min\ & f(x) \\ \text{s.t.}\ & m \leqslant s \end{aligned}$$

(general $\mathcal{D}$)

$$\begin{aligned} \min\ & f(x) \\ \text{s.t.}\ & \|x\|_0 \leqslant s \end{aligned}$$

($\mathcal{D} = \{\pm e_1, \ldots, \pm e_n\}$)

$$\begin{aligned} \min\ & f(x) \\ \text{s.t.}\ & \|x\|_1 \leqslant s \end{aligned}$$

($\mathcal{D} = \{\pm e_1, \ldots, \pm e_n\}$
and ℓ_1-convex relaxation)

$$\begin{aligned} \min\ & f(x) + \lambda \|x\|_1 \\ \text{s.t.}\ & \xcancel{\|x\|_1 \leqslant s} \end{aligned}$$

($\mathcal{D} = \{\pm e_1, \ldots, \pm e_n\}$
and lasso regularization)

Goals

- Provide an alternative to the constrained/regularized methods which require tuning of s or λ:

$$\begin{aligned} \min\ & f(x) \\ \text{s.t.}\ & m \leqslant s \end{aligned}$$
(general $\mathcal{D}$)

$$\begin{aligned} \min\ & f(x) \\ \text{s.t.}\ & \|x\|_0 \leqslant s \end{aligned}$$
$(\mathcal{D} = \{\pm e_1, \ldots, \pm e_n\})$

$$\begin{aligned} \min\ & f(x) \\ \text{s.t.}\ & \|x\|_1 \leqslant s \end{aligned}$$
$(\mathcal{D} = \{\pm e_1, \ldots, \pm e_n\}$ and ℓ_1-convex relaxation)

$$\begin{aligned} \min\ & f(x) + \lambda \|x\|_1 \\ \text{s.t.}\ & \cancel{\|x\|_1 \leqslant s} \end{aligned}$$
$(\mathcal{D} = \{\pm e_1, \ldots, \pm e_n\}$ and lasso regularization)

Also: possibility of faster convergence by allowing iterates to go sometimes outside the feasible region?

Goals

- Provide an alternative to the constrained/regularized methods which require tuning of s or λ:

$$\min\ f(x) \quad \text{s.t.}\ m \leqslant s \quad \text{(general } \mathcal{D}\text{)}$$

$$\min\ f(x) \quad \text{s.t.}\ \|x\|_0 \leqslant s \quad (\mathcal{D} = \{\pm e_1, \dots, \pm e_n\})$$

$$\min\ f(x) \quad \text{s.t.}\ \|x\|_1 \leqslant s \quad (\mathcal{D} = \{\pm e_1, \dots, \pm e_n\} \text{ and } \ell_1\text{-convex relaxation})$$

$$\min\ f(x) + \lambda\|x\|_1 \quad \text{s.t.}\ \cancel{\|x\|_1 \leqslant s} \quad (\mathcal{D} = \{\pm e_1, \dots, \pm e_n\} \text{ and lasso regularization})$$

  Also: possibility of faster convergence by allowing iterates to go sometimes outside the feasible region?

- Provide an unconstrained method that keeps **each iterate** sparse, to avoid expensive reoptimizations

Generalized Matching Pursuit (GMP)

Locatello et al. [2017]

- **Gradient descent**: optimal descent direction but potentially poor sparsity

$$x_{t+1} \leftarrow x_t - \gamma_t \nabla f(x_t)$$

The update term $-\nabla f(x_t)$ may be a combination of many atoms

Generalized Matching Pursuit (GMP)

Locatello et al. [2017]

- **Gradient descent**: optimal descent direction but potentially poor sparsity

$$x_{t+1} \leftarrow x_t - \gamma_t \nabla f(x_t)$$

  The update term $-\nabla f(x_t)$ may be a combination of many atoms

- **GMP**: to preserve sparsity, move in the direction of an atom $v_t \in \mathcal{D}$

$$x_{t+1} \leftarrow x_t + \gamma_t v_t$$

Generalized Matching Pursuit (GMP)

Locatello et al. [2017]

- **Gradient descent**: optimal descent direction but potentially poor sparsity

$$x_{t+1} \leftarrow x_t - \gamma_t \nabla f(x_t)$$

The update term $-\nabla f(x_t)$ may be a combination of many atoms

- **GMP**: to preserve sparsity, move in the direction of an atom $v_t \in \mathcal{D}$

$$x_{t+1} \leftarrow x_t + \gamma_t v_t$$

  - Select the v_t that best approximates the direction $-\nabla f(x_t)$:

$$v_t \leftarrow \arg\max_{v \in \mathcal{D}} \langle -\nabla f(x_t), v \rangle$$

Generalized Matching Pursuit (GMP)

Locatello et al. [2017]

- **Gradient descent**: optimal descent direction but potentially poor sparsity

$$x_{t+1} \leftarrow x_t - \gamma_t \nabla f(x_t)$$

  The update term $-\nabla f(x_t)$ may be a combination of many atoms

- **GMP**: to preserve sparsity, move in the direction of an atom $v_t \in \mathcal{D}$

$$x_{t+1} \leftarrow x_t + \gamma_t v_t$$

  - Select the v_t that best approximates the direction $-\nabla f(x_t)$:

$$v_t \leftarrow \arg\min_{v \in \mathcal{D}} \langle \nabla f(x_t), v \rangle$$

Generalized Matching Pursuit (GMP)

Locatello et al. [2017]

- **Gradient descent**: optimal descent direction but potentially poor sparsity

$$x_{t+1} \leftarrow x_t - \gamma_t \nabla f(x_t)$$

The update term $-\nabla f(x_t)$ may be a combination of many atoms

- **GMP**: to preserve sparsity, move in the direction of an atom $v_t \in \mathcal{D}$

$$x_{t+1} \leftarrow x_t + \gamma_t v_t$$

  - Select the v_t that best approximates the direction $-\nabla f(x_t)$:

$$v_t \leftarrow \arg\min_{v \in \mathcal{D}} \langle \nabla f(x_t), v \rangle$$

  - The progress in function value is at least:

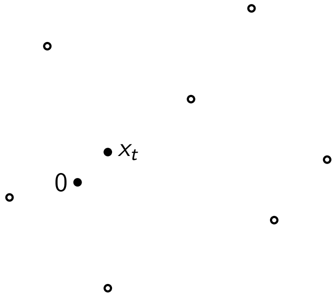
$$f(x_t) - f(x_{t+1}) \geqslant \frac{\langle \nabla f(x_t), v_t \rangle^2}{2L}$$

Generalized Matching Pursuit

Locatello et al. [2017]

Algorithm GMP

$\mathcal{S}_0 \leftarrow \{x_0\}$ where $x_0 \in \mathcal{D}$
for $t = 0$ **to** $T-1$ **do**
  $v_t \leftarrow \arg\min_{v \in \mathcal{D}} \langle \nabla f(x_t), v \rangle$
  $\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{v_t\}$
  $x_{t+1} \leftarrow \arg\min_{x_t + \mathbb{R} v_t} f$
end for

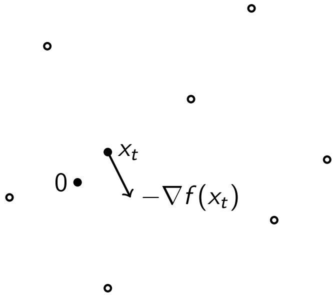


Generalized Matching Pursuit

Locatello et al. [2017]

Algorithm GMP

1: $\mathcal{S}_0 \leftarrow \{x_0\}$ where $x_0 \in \mathcal{D}$
2: **for** $t = 0$ **to** $T-1$ **do**
3: $\quad v_t \leftarrow \arg\min_{v \in \mathcal{D}} \langle \nabla f(x_t), v \rangle$
4: $\quad \mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{v_t\}$
5: $\quad x_{t+1} \leftarrow \arg\min_{x_t + \mathbb{R} v_t} f$
6: **end for**



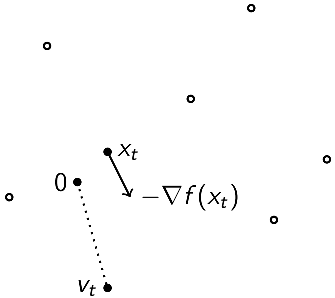
Generalized Matching Pursuit

Locatello et al. [2017]

Algorithm GMP

$\mathcal{S}_0 \leftarrow \{x_0\}$ where $x_0 \in \mathcal{D}$
for $t = 0$ **to** $T - 1$ **do**
  $v_t \leftarrow \arg\min_{v \in \mathcal{D}} \langle \nabla f(x_t), v \rangle$
  $\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{v_t\}$
  $x_{t+1} \leftarrow \arg\min_{x_t + \mathbb{R} v_t} f$
end for

- v_t approximates the direction $-\nabla f(x_t)$

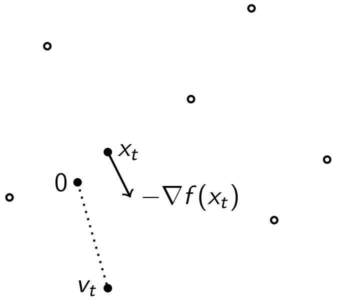


Generalized Matching Pursuit

Locatello et al. [2017]

Algorithm GMP

$\mathcal{S}_0 \leftarrow \{x_0\}$ where $x_0 \in \mathcal{D}$
for $t = 0$ **to** $T-1$ **do**
  $v_t \leftarrow \arg\min_{v \in \mathcal{D}} \langle \nabla f(x_t), v \rangle$
  $\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{v_t\}$
  $x_{t+1} \leftarrow \arg\min_{x_t + \mathbb{R} v_t} f$
end for



- v_t approximates the direction $-\nabla f(x_t)$
- Sparsity of x_{t+1} is $|\mathcal{S}_{t+1}| \leqslant |\mathcal{S}_t| + 1$

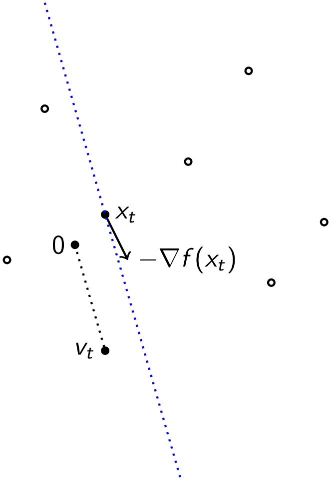
Generalized Matching Pursuit

Locatello et al. [2017]

Algorithm GMP

1: $\mathcal{S}_0 \leftarrow \{x_0\}$ where $x_0 \in \mathcal{D}$
2: **for** $t = 0$ **to** $T-1$ **do**
3: $\quad v_t \leftarrow \arg\min_{v \in \mathcal{D}} \langle \nabla f(x_t), v \rangle$
4: $\quad \mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{v_t\}$
5: $\quad x_{t+1} \leftarrow \arg\min_{x_t + \mathbb{R} v_t} f$
6: **end for**

- v_t approximates the direction $-\nabla f(x_t)$
- Sparsity of x_{t+1} is $|\mathcal{S}_{t+1}| \leqslant |\mathcal{S}_t| + 1$
- x_t moves in the direction of v_t



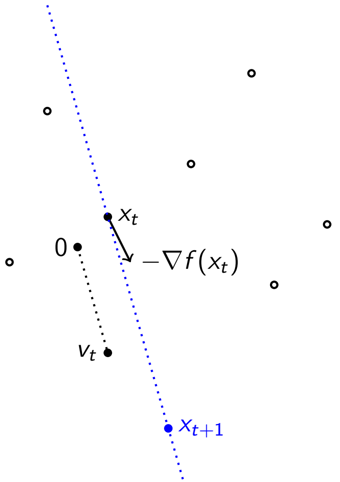
Generalized Matching Pursuit

Locatello et al. [2017]

Algorithm GMP

1: $\mathcal{S}_0 \leftarrow \{x_0\}$ where $x_0 \in \mathcal{D}$
2: **for** $t = 0$ **to** $T - 1$ **do**
3: $\quad v_t \leftarrow \arg\min_{v \in \mathcal{D}} \langle \nabla f(x_t), v \rangle$
4: $\quad \mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{v_t\}$
5: $\quad x_{t+1} \leftarrow \arg\min_{x_t + \mathbb{R} v_t} f$
6: **end for**

- v_t approximates the direction $-\nabla f(x_t)$
- Sparsity of x_{t+1} is $|\mathcal{S}_{t+1}| \leqslant |\mathcal{S}_t| + 1$
- x_t moves in the direction of v_t



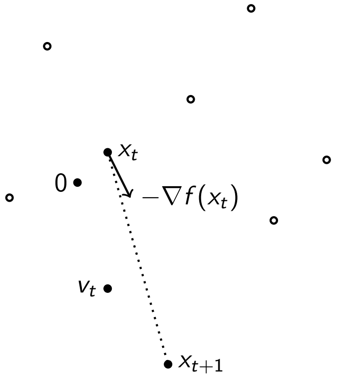
Generalized Matching Pursuit

Locatello et al. [2017]

Algorithm GMP

$\mathcal{S}_0 \leftarrow \{x_0\}$ where $x_0 \in \mathcal{D}$
for $t = 0$ **to** $T-1$ **do**
  $v_t \leftarrow \arg\min_{v \in \mathcal{D}} \langle \nabla f(x_t), v \rangle$
  $\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{v_t\}$
  $x_{t+1} \leftarrow \arg\min_{x_t + \mathbb{R} v_t} f$
end for

- v_t approximates the direction $-\nabla f(x_t)$
- Sparsity of x_{t+1} is $|\mathcal{S}_{t+1}| \leqslant |\mathcal{S}_t| + 1$
- x_t moves in the direction of v_t
- $x_{t+1} = x_t + \gamma_t v_t = x_0 + \sum_{s=0}^{t} \gamma_s v_s$

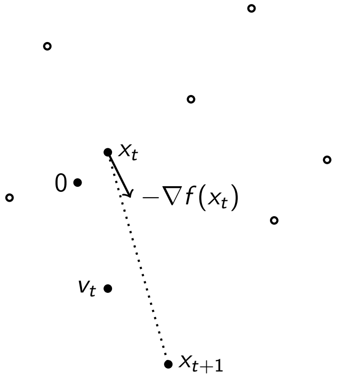


Generalized Matching Pursuit

Locatello et al. [2017]

Algorithm GMP

$\mathcal{S}_0 \leftarrow \{x_0\}$ where $x_0 \in \mathcal{D}$
for $t = 0$ **to** $T - 1$ **do**
    $v_t \leftarrow \arg\min_{v \in \mathcal{D}} \langle \nabla f(x_t), v \rangle$
    $\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{v_t\}$
    $x_{t+1} \leftarrow \arg\min_{x_t + \mathbb{R} v_t} f$
end for



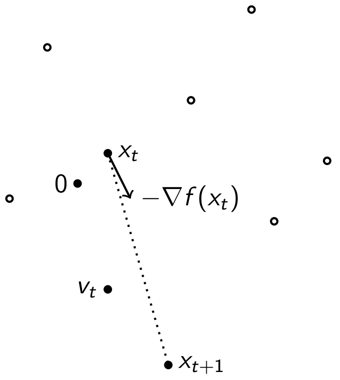
- v_t approximates the direction $-\nabla f(x_t)$
- Sparsity of x_{t+1} is $|\mathcal{S}_{t+1}| \leqslant |\mathcal{S}_t| + 1$
- x_t moves in the direction of v_t
- $x_{t+1} = x_t + \gamma_t v_t = x_0 + \sum_{s=0}^{t} \gamma_s v_s$

Generalized Matching Pursuit

Locatello et al. [2017]

Algorithm GMP

1: $\mathcal{S}_0 \leftarrow \{x_0\}$ where $x_0 \in \mathcal{D}$
2: **for** $t = 0$ **to** $T - 1$ **do**
3: $\quad v_t \leftarrow \arg\min_{v \in \mathcal{D}} \langle \nabla f(x_t), v \rangle$
4: $\quad \mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{v_t\}$
5: $\quad x_{t+1} \leftarrow \arg\min_{x_t + \mathbb{R} v_t} f$
6: **end for**



- v_t approximates the direction $-\nabla f(x_t)$
- Sparsity of x_{t+1} is $|\mathcal{S}_{t+1}| \leqslant |\mathcal{S}_t| + 1$
- x_t moves in the direction of v_t
- $x_{t+1} = x_t + \gamma_t v_t = x_0 + \sum_{s=0}^{t} \gamma_s v_s$

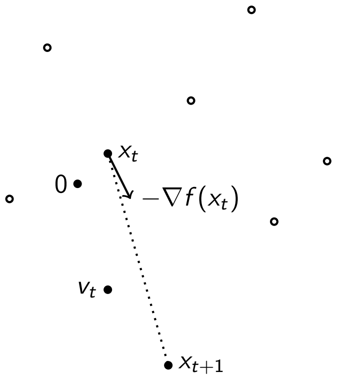
Generalized Matching Pursuit

Locatello et al. [2017]

Algorithm GMP

$\mathcal{S}_0 \leftarrow \{x_0\}$ where $x_0 \in \mathcal{D}$
for $t = 0$ **to** $T-1$ **do**
  $v_t \leftarrow \arg\min_{v \in \mathcal{D}} \langle \nabla f(x_t), v \rangle$
  $\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{v_t\}$
  $x_{t+1} \leftarrow \arg\min_{\mathrm{span}(\mathcal{S}_{t+1})} f$
end for

- v_t approximates the direction $-\nabla f(x_t)$
- Sparsity of x_{t+1} is $|\mathcal{S}_{t+1}| \leqslant |\mathcal{S}_t| + 1$
- x_t moves in the direction of v_t
- $x_{t+1} = x_t + \gamma_t v_t = x_0 + \sum_{s=0}^{t} \gamma_s v_s$



Generalized/Orthogonal Matching Pursuit

GMP: $x_{t+1} \leftarrow \arg\min_{x_t + \mathbb{R} v_t} f$

OMP: $x_{t+1} \leftarrow \arg\min_{\text{span}(\mathcal{S}_{t+1})} f$

(sequence of projected gradient (PG) steps)

Generalized/Orthogonal Matching Pursuit

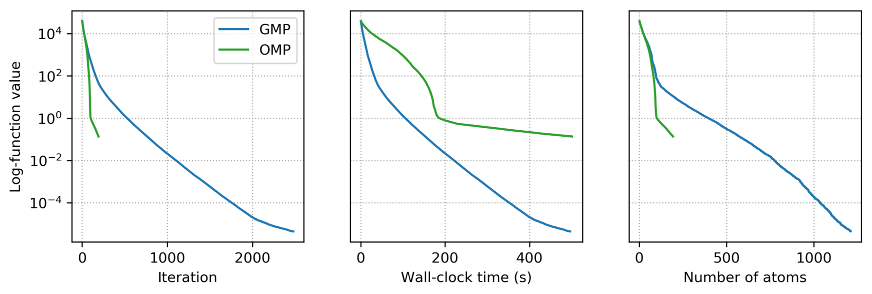
GMP: $x_{t+1} \leftarrow \arg \min_{x_t + \mathbb{R} v_t} f$

OMP: $x_{t+1} \leftarrow \arg \min_{\text{span}(\mathcal{S}_{t+1})} f$

(sequence of projected gradient (PG) steps)

Example (Sparse recovery/machine learning)

- Measured signal/observed data: $y = Ax^* + \mathcal{N}(0, \sigma^2 I_m)$ where $\|x^*\|_0 \ll n$
- Goal: recover/learn x^*
- Problem: minimize $f(x) = \|y - Ax\|_2^2$ with $\mathcal{D} = \{\pm e_1, \dots, \pm e_n\}$



Generalized/Orthogonal Matching Pursuit

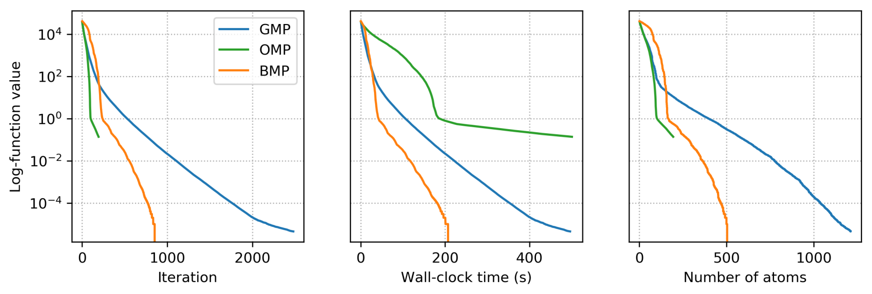
GMP: $x_{t+1} \leftarrow \arg\min_{x_t + \mathbb{R}v_t} f$

OMP: $x_{t+1} \leftarrow \arg\min_{\mathrm{span}(\mathcal{S}_{t+1})} f$

(sequence of projected gradient (PG) steps)

Example (Sparse recovery/machine learning)

- Measured signal/observed data: $y = Ax^* + \mathcal{N}(0, \sigma^2 I_m)$ where $\|x^*\|_0 \ll n$
- Goal: recover/learn x^*
- Problem: minimize $f(x) = \|y - Ax\|_2^2$ with $\mathcal{D} = \{\pm e_1, \ldots, \pm e_n\}$



Blended Matching Pursuit
Ideas

- Unify the best of GMP (speed) and OMP (sparsity)

Blended Matching Pursuit

Ideas

- Unify the best of GMP (speed) and OMP (sparsity)

GMP step

$$v_t \leftarrow \arg\min_{v \in \mathcal{D}} \langle \nabla f(x_t), v \rangle$$

$$x_{t+1} \leftarrow \arg\min_{x_t + \mathbb{R} v_t} f$$

$$\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{v_t\}$$

PG step

$$\widetilde{\nabla} f(x_t) \leftarrow \mathrm{proj}_{\mathrm{span}(\mathcal{S}_t)}(\nabla f(x_t))$$

$$x_{t+1} \leftarrow \arg\min_{x_t + \mathbb{R}\widetilde{\nabla} f(x_t)} f$$

$$\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t$$

Blended Matching Pursuit

Ideas

- Unify the best of GMP (speed) and OMP (sparsity)

GMP step

$$v_t \leftarrow \arg\min_{v \in \mathcal{D}} \langle \nabla f(x_t), v \rangle$$
$$x_{t+1} \leftarrow \arg\min_{x_t + \mathbb{R} v_t} f$$
$$\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{v_t\}$$

Potentially more progress
but decreases the sparsity level

PG step

$$\widetilde{\nabla} f(x_t) \leftarrow \mathrm{proj}_{\mathrm{span}(\mathcal{S}_t)}(\nabla f(x_t))$$
$$x_{t+1} \leftarrow \arg\min_{x_t + \mathbb{R} \widetilde{\nabla} f(x_t)} f$$
$$\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t$$

Progress only over $\mathrm{span}(\mathcal{S}_t)$
but keeps the sparsity level intact

Blended Matching Pursuit

Ideas

- Unify the best of GMP (speed) and OMP (sparsity)

GMP step

$$v_t \leftarrow \arg\min_{v \in \mathcal{D}} \langle \nabla f(x_t), v \rangle$$

$$x_{t+1} \leftarrow \arg\min_{x_t + \mathbb{R} v_t} f$$

$$\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{v_t\}$$

Potentially more progress
but decreases the sparsity level

PG step

$$\widetilde{\nabla} f(x_t) \leftarrow \mathrm{proj}_{\mathrm{span}(\mathcal{S}_t)}(\nabla f(x_t))$$

$$x_{t+1} \leftarrow \arg\min_{x_t + \mathbb{R}\widetilde{\nabla} f(x_t)} f$$

$$\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t$$

Progress only over $\mathrm{span}(\mathcal{S}_t)$
but keeps the sparsity level intact

- Blend GMP steps with PG steps:

$$\textbf{GMP}, \underbrace{\textbf{PG}, \ldots, \textbf{PG}}_{\substack{\text{partially optimize} \\ \text{over } \mathrm{span}(\mathcal{S}_t)}}, \underbrace{\textbf{GMP}}_{\substack{\text{add 1 atom and} \\ \text{enter new space} \\ \mathrm{span}(\mathcal{S}_t \cup \{v_t\})}}, \underbrace{\textbf{PG}, \ldots, \textbf{PG}}_{\substack{\text{partially optimize} \\ \text{over } \mathrm{span}(\mathcal{S}_t \cup \{v_t\})}}, \textbf{GMP}, \ldots$$

Blended Matching Pursuit

Ideas

- Unify the best of GMP (speed) and OMP (sparsity)

GMP step

$$v_t \leftarrow \arg\min_{v \in \mathcal{D}} \langle \nabla f(x_t), v \rangle$$
$$x_{t+1} \leftarrow \arg\min_{x_t + \mathbb{R} v_t} f$$
$$\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{v_t\}$$

Potentially more progress
but decreases the sparsity level

PG step

$$\widetilde{\nabla} f(x_t) \leftarrow \mathrm{proj}_{\mathrm{span}(\mathcal{S}_t)}(\nabla f(x_t))$$
$$x_{t+1} \leftarrow \arg\min_{x_t + \mathbb{R}\widetilde{\nabla} f(x_t)} f$$
$$\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t$$

Progress only over $\mathrm{span}(\mathcal{S}_t)$
but keeps the sparsity level intact

- Blend GMP steps with PG steps:

$$\textbf{GMP}, \underbrace{\textbf{PG}, \ldots, \textbf{PG}}_{\substack{\text{partially optimize} \\ \text{over span}(\mathcal{S}_t)}}, \underbrace{\textbf{GMP}}_{\substack{\text{add 1 atom and} \\ \text{enter new space} \\ \text{span}(\mathcal{S}_t \cup \{v_t\})}}, \underbrace{\textbf{PG}, \ldots, \textbf{PG}}_{\substack{\text{partially optimize} \\ \text{over span}(\mathcal{S}_t \cup \{v_t\})}}, \textbf{GMP}, \ldots$$

- Additional speed-up: lazify the linear minimization oracle with a *weak-separation* oracle $\mathrm{LPsep}_{\mathcal{D}}(\nabla f(x_t), \phi_t)$ [Braun et al., 2017]

Find $v_t \in \mathcal{D}$ such that $\langle \nabla f(x_t), v_t \rangle \leqslant \phi_t/2$

Blended Matching Pursuit

Algorithm design

- How to decide which step to perform? Check progress:

$$f(x_t) - f(x_{t+1}) \geqslant \begin{cases} \dfrac{\min_{v \in \mathcal{D}} \langle \nabla f(x_t), v \rangle^2}{2L} & \text{if } \textbf{GMP step} \\ \geqslant \dfrac{\|\widetilde{\nabla} f(x_t)\|^2}{2L} \geqslant \dfrac{\min_{v \in \mathcal{S}_t} \langle \nabla f(x_t), v \rangle^2}{2L} & \text{if } \textbf{PG step} \end{cases}$$

- How to decide which step to perform? Check progress:

$$f(x_t) - f(x_{t+1}) \geqslant \begin{cases} \dfrac{\min_{v\in\mathcal{D}}\langle\nabla f(x_t), v\rangle^2}{2L} & \text{if } \textbf{GMP step} \\ \geqslant \dfrac{\|\widetilde{\nabla} f(x_t)\|^2}{2L} \geqslant \dfrac{\min_{v\in\mathcal{S}_t}\langle\nabla f(x_t), v\rangle^2}{2L} & \text{if } \textbf{PG step} \end{cases}$$

- Introduce a quantity $\phi_t \leqslant 0$ to monitor BMP

Blended Matching Pursuit

Algorithm design

- How to decide which step to perform? Check progress:

$$f(x_t) - f(x_{t+1}) \geqslant \begin{cases} \dfrac{\min_{v \in \mathcal{D}} \langle \nabla f(x_t), v \rangle^2}{2L} & \text{if } \textbf{GMP step} \\ \geqslant \dfrac{\|\widetilde{\nabla} f(x_t)\|^2}{2L} \geqslant \dfrac{\min_{v \in \mathcal{S}_t} \langle \nabla f(x_t), v \rangle^2}{2L} & \text{if } \textbf{PG step} \end{cases}$$

- Introduce a quantity $\phi_t \leqslant 0$ to monitor BMP
- Before adding a new atom (**GMP step**), check if *enough* progress can be made with a **PG step**:

Blended Matching Pursuit

Algorithm design

- How to decide which step to perform? Check progress:

$$f(x_t) - f(x_{t+1}) \geqslant \begin{cases} \dfrac{\min_{v \in \mathcal{D}} \langle \nabla f(x_t), v \rangle^2}{2L} & \text{if } \textbf{GMP step} \\ \geqslant \dfrac{\|\widetilde{\nabla} f(x_t)\|^2}{2L} \geqslant \dfrac{\min_{v \in \mathcal{S}_t} \langle \nabla f(x_t), v \rangle^2}{2L} & \text{if } \textbf{PG step} \end{cases}$$

- Introduce a quantity $\phi_t \leqslant 0$ to monitor BMP
- Before adding a new atom (**GMP step**), check if *enough* progress can be made with a **PG step**:
  - **If** $\min_{v \in \mathcal{S}_t} \langle \nabla f(x_t), v \rangle \leqslant \phi_t / 2$ **then** take a **PG step**
  - **Else**: check the guarantee on progress for a **GMP step**

Blended Matching Pursuit

Algorithm design

- How to decide which step to perform? Check progress:

$$f(x_t) - f(x_{t+1}) \geqslant \begin{cases} \dfrac{\min_{v \in \mathcal{D}} \langle \nabla f(x_t), v \rangle^2}{2L} & \text{if } \textbf{GMP step} \\ \geqslant \dfrac{\|\widetilde{\nabla} f(x_t)\|^2}{2L} \geqslant \dfrac{\min_{v \in \mathcal{S}_t} \langle \nabla f(x_t), v \rangle^2}{2L} & \text{if } \textbf{PG step} \end{cases}$$

- Introduce a quantity $\phi_t \leqslant 0$ to monitor BMP
- Before adding a new atom (**GMP step**), check if *enough* progress can be made with a **PG step**:
  - **If** $\min_{v \in \mathcal{S}_t} \langle \nabla f(x_t), v \rangle \leqslant \phi_t/2$ **then** take a **PG step**
  - **Else**: check the guarantee on progress for a **GMP step**
    - **If** $\text{LPsep}_{\mathcal{D}}(\nabla f(x_t), \phi_t)$ finds a $v_t \in \mathcal{D}$ s.t. $\langle \nabla f(x_t), v_t \rangle \leqslant \phi_t/2$ **then** take a **GMP step**
    - **Else**: $0 \geqslant \min_{v \in \mathcal{D}} \langle \nabla f(x_t), v \rangle > \phi_t/2$ so ϕ_t is too low and needs to be rescaled:

Perform a **dual step**: $\phi_{t+1} \leftarrow \phi_t/2$

Blended Matching Pursuit

Pseudocode

Algorithm BMP

$\mathcal{S}_0,\ \phi_0 \leftarrow \{x_0\},\ \min_{v\in\mathcal{D}}\langle\nabla f(x_0), v\rangle/2$ where $x_0 \in \mathcal{D}$

for $t = 0$ **to** $T-1$ **do**

  if $\min_{v\in\mathcal{S}_t}\langle\nabla f(x_t), v\rangle \leqslant \phi_t/2$ **then**

    $x_{t+1} \leftarrow \arg\min_{x_t+\mathbb{R}\widetilde{\nabla} f(x_t)} f$ {PG step}

    $\mathcal{S}_{t+1}, \phi_{t+1} \leftarrow \mathcal{S}_t, \phi_t$

  else

    $v_t \leftarrow \mathrm{LPsep}_{\mathcal{D}}(\nabla f(x_t), \phi_t)$

    if $v_t =$ **false then**

      $x_{t+1} \leftarrow x_t$ {dual step}

      $\mathcal{S}_{t+1}, \phi_{t+1} \leftarrow \mathcal{S}_t, \phi_t/2$

    else

      $x_{t+1} \leftarrow \arg\min_{x_t+\mathbb{R}v_t} f$ {GMP step}

      $\mathcal{S}_{t+1}, \phi_{t+1} \leftarrow \mathcal{S}_t \cup \{v_t\}, \phi_t$

    end if

  end if

end for

Definitions

- f is **L-smooth of order $\ell > 1$** if $L > 0$ and

$$\forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n, \ f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leqslant \frac{L}{\ell} \|y - x\|^{\ell}$$

Definitions

- f is ***L*-smooth of order ℓ** > 1 if $L > 0$ and

$$\forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n, \; f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leqslant \frac{L}{\ell} \|y - x\|^{\ell}$$

- f is ***S*-strongly convex of order *s*** > 1 if $S > 0$ and

$$\forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n, \; f(y) - f(x) - \langle \nabla f(x), y - x \rangle \geqslant \frac{S}{s} \|y - x\|^{s}$$

Definitions

- f is **L-smooth of order ℓ** > 1 if $L > 0$ and

$$\forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n, \quad f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leqslant \frac{L}{\ell} \|y - x\|^{\ell}$$

- f is **S-strongly convex of order s** > 1 if $S > 0$ and

$$\forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n, \quad f(y) - f(x) - \langle \nabla f(x), y - x \rangle \geqslant \frac{S}{s} \|y - x\|^{s}$$

- f is **C-sharp of order θ** $\in]0, 1[$ if $C > 0$ and

$$\operatorname{dist}\left(x, \arg\min_{\mathbb{R}^n} f\right) \leqslant C \left(f(x) - \min_{\mathbb{R}^n} f\right)^{\theta}$$

holds around $\arg\min_{\mathbb{R}^n} f$.

Definitions

On sharpness and smoothness

Fact

If f is smooth of order $\ell > 1$ and sharp of order $\theta \in]0,1[$, then $\ell\theta \leqslant 1$.

Definitions

On sharpness and smoothness

Fact

If f is smooth of order $\ell > 1$ and sharp of order $\theta \in]0,1[$, then $\ell\theta \leqslant 1$.

We have

$$\begin{cases} f(x) - f(x^*) \leqslant \dfrac{L}{\ell}\|x - x^*\|^{\ell} & \text{by smoothness} \\ \|x - x^*\| \leqslant C(f(x) - f(x^*))^{\theta} & \text{by sharpness} \end{cases}$$

Definitions

On sharpness and smoothness

Fact

If f is smooth of order $\ell > 1$ and sharp of order $\theta \in]0,1[$, then $\ell\theta \leqslant 1$.

We have

$$\begin{cases} f(x) - f(x^*) \leqslant \dfrac{L}{\ell}\|x - x^*\|^{\ell} & \text{by smoothness} \\ \|x - x^*\| \leqslant C(f(x) - f(x^*))^{\theta} & \text{by sharpness} \end{cases}$$

so

$$\frac{1}{C}\left(\frac{\ell}{L}\right)^{\theta} \leqslant \|x - x^*\|^{\ell\theta - 1}.$$

Fact

If f is smooth of order $\ell > 1$ and sharp of order $\theta \in]0,1[$, then $\ell\theta \leqslant 1$.

We have

$$\begin{cases} f(x) - f(x^*) \leqslant \frac{L}{\ell}\|x - x^*\|^\ell & \text{by smoothness} \\ \|x - x^*\| \leqslant C(f(x) - f(x^*))^\theta & \text{by sharpness} \end{cases}$$

so

$$\frac{1}{C}\left(\frac{\ell}{L}\right)^\theta \leqslant \|x - x^*\|^{\ell\theta - 1}.$$

Therefore,

$$\ell\theta \leqslant 1.$$

- Strong convexity $\Rightarrow$ sharpness:

$$\frac{S}{s}\|x-x^*\|^s \leqslant f(x)-f(x^*) \quad \Rightarrow \quad \|x-x^*\| \leqslant \left(\frac{s}{S}\right)^{1/s}(f(x)-f(x^*))^{1/s}$$

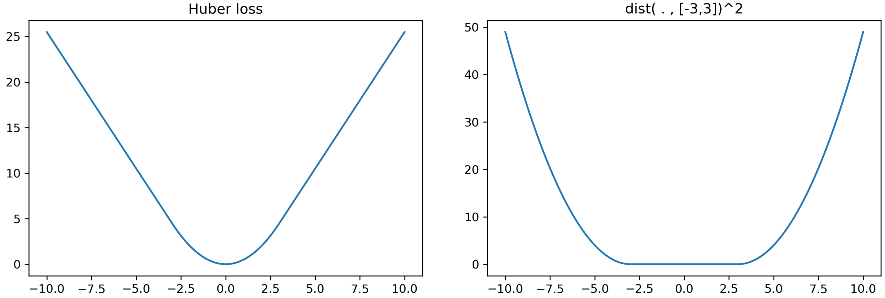
Definitions

On sharpness and strong convexity

- Strong convexity $\Rightarrow$ sharpness:

$$\frac{S}{s}\|x - x^*\|^s \leqslant f(x) - f(x^*) \quad \Rightarrow \quad \|x - x^*\| \leqslant \left(\frac{s}{S}\right)^{1/s} (f(x) - f(x^*))^{1/s}$$

- Convexity+sharpness $\not\Rightarrow$ strong convexity



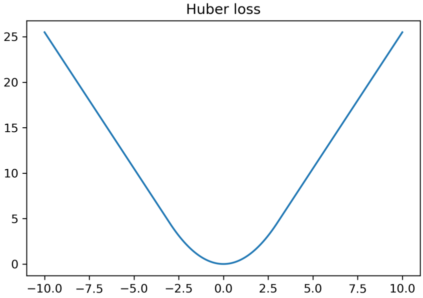
Definitions

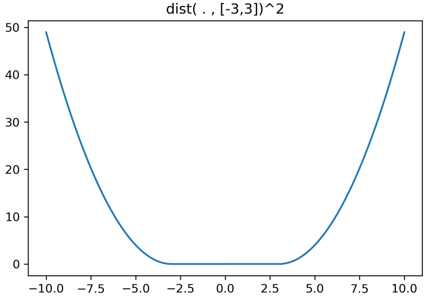
On sharpness and strong convexity

- Strong convexity $\Rightarrow$ sharpness:

$$\frac{S}{s}\|x - x^*\|^s \leqslant f(x) - f(x^*) \quad \Rightarrow \quad \|x - x^*\| \leqslant \left(\frac{s}{S}\right)^{1/s} (f(x) - f(x^*))^{1/s}$$

- Convexity+sharpness $\not\Rightarrow$ strong convexity





- Sharpness holds for all *well-behaved* convex functions [Bolte et al., 2007]

Blended Matching Pursuit

Convergence analysis

Properties of f	BMP convergence rate	Lower bound on complexity[1]
Smooth convex	$T(\epsilon) = \mathcal{O}\left(\frac{1}{\epsilon^{1/(\ell-1)}}\right)$	$T(\epsilon) = \Omega\left(\frac{1}{\epsilon^{1/(1.5\ell-1)}}\right)$
Smooth convex sharp with $\ell\theta = 1$	$T(\epsilon) = \mathcal{O}\left(\ln\left(\frac{1}{\epsilon}\right)\right)$	$T(\epsilon) = \Omega\left(\ln\left(\frac{1}{\epsilon}\right)\right)$
Smooth convex sharp with $\ell\theta < 1$	$T(\epsilon) = \mathcal{O}\left(\frac{1}{\epsilon^{(1-\ell\theta)/(\ell-1)}}\right)$	$T(\epsilon) = \Omega\left(\frac{1}{\epsilon^{(1-\ell\theta)/(1.5\ell-1)}}\right)$

[1] Nemirovskii and Nesterov [1985].

Blended Matching Pursuit

Convergence analysis

Properties of f	BMP convergence rate	Lower bound on complexity[1]
Smooth convex	$T(\epsilon) = \mathcal{O}\left(\frac{1}{\epsilon^{1/(\ell-1)}}\right)$	$T(\epsilon) = \Omega\left(\frac{1}{\epsilon^{1/(1.5\ell-1)}}\right)$
Smooth convex sharp with $\ell\theta = 1$	$T(\epsilon) = \mathcal{O}\left(\ln\left(\frac{1}{\epsilon}\right)\right)$	$T(\epsilon) = \Omega\left(\ln\left(\frac{1}{\epsilon}\right)\right)$
Smooth convex sharp with $\ell\theta < 1$	$T(\epsilon) = \mathcal{O}\left(\frac{1}{\epsilon^{(1-\ell\theta)/(\ell-1)}}\right)$	$T(\epsilon) = \Omega\left(\frac{1}{\epsilon^{(1-\ell\theta)/(1.5\ell-1)}}\right)$

- Open question: can we close the gap using acceleration?

[1] Nemirovskii and Nesterov [1985].

Computational experiments

BMP vs. GMP, OMP, BCG [Braun et al., 2019], CoGEnT [Rao et al., 2015]

- Measured signal/observed data: $y = Ax^* + \mathcal{N}(0, \sigma^2 I_m)$ where $\|x^*\|_0 \ll n$
- Goal: recover/learn x^* using $\mathcal{D} = \{\pm e_1, \ldots, \pm e_n\}$

Computational experiments

BMP vs. GMP, OMP, BCG [Braun et al., 2019], CoGEnT [Rao et al., 2015]

- Measured signal/observed data: $y = Ax^* + \mathcal{N}(0, \sigma^2 I_m)$ where $\|x^*\|_0 \ll n$
- Goal: recover/learn x^* using $\mathcal{D} = \{\pm e_1, \ldots, \pm e_n\}$
- Different methods:

BMP, GMP, and OMP solve

$$\begin{aligned} \min \quad & \|y - Ax\|_2^2 \\ \text{s.t.} \quad & x \in \mathbb{R}^n \end{aligned}$$

BCG and CoGEnT solve

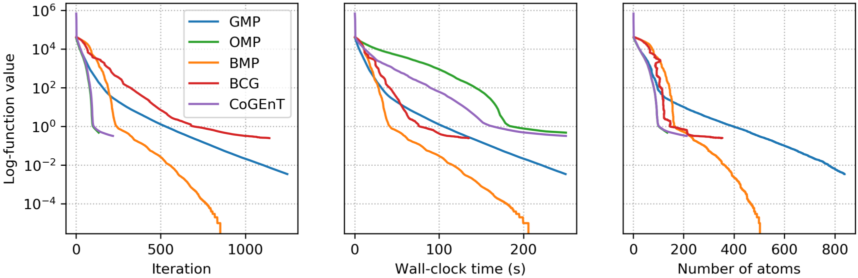
$$\begin{aligned} \min \quad & \|y - Ax\|_2^2 \\ \text{s.t.} \quad & \|x\|_1 \leqslant \|x^*\|_1 \end{aligned}$$

where $\|x^*\|_1$ is favorably given

Computational experiments

BMP vs. GMP, OMP, BCG [Braun et al., 2019], CoGEnT [Rao et al., 2015]

Let $f : x \in \mathbb{R}^{2000} \mapsto \|y - Ax\|_2^2$, $A \in \mathbb{R}^{500\times 2000}$, and $\mathcal{D} = \{\pm e_1, \ldots, \pm e_{2000}\}$

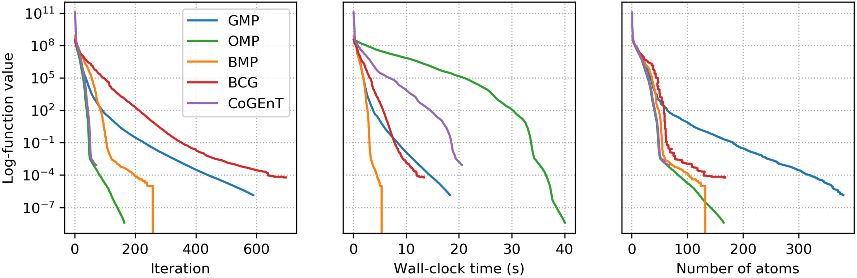


- BMP converges faster (in time) than the other methods
- BMP has close-to-optimal sparsity while having no explicit sparsity information (which the constrained methods BCG and CoGEnT have)

Computational experiments

BMP vs. GMP, OMP, BCG [Braun et al., 2019], CoGEnT [Rao et al., 2015]

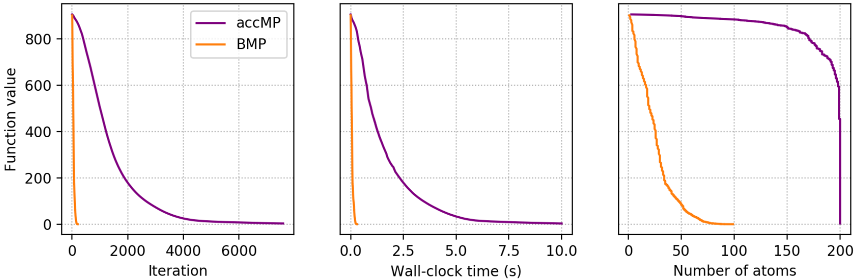
Let $f : x \in \mathbb{R}^{1000} \mapsto \|y - Ax\|_3^5$, $A \in \mathbb{R}^{250\times 1000}$, and $\mathcal{D} = \{\pm e_1, \ldots, \pm e_{1000}\}$



Computational experiments

BMP vs. accMP [Locatello et al., 2018]

Let $f : x \in \mathbb{R}^{100} \mapsto \frac{1}{2}\|x - b\|_2^2$ and $\mathcal{D}$ be a set of 200 random points



Thank you!

References

[1] J. Bolte, A. Daniilidis, and A. Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 2007.

[2] G. Braun, S. Pokutta, and D. Zink. Lazifying conditional gradient algorithms. *ICML*, 2017.

[3] G. Braun, S. Pokutta, D. Tu, and S. Wright. Blended conditional gradients. *ICML*, 2019.

[4] C. W. Combettes and S. Pokutta. Blended matching pursuit. *arXiv preprint arXiv:1904.12335*, 2019.

[5] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 1956.

[6] E. S. Levitin and B. T. Polyak. Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics*, 1966.

[7] F. Locatello, R. Khanna, M. Tschannen, and M. Jaggi. A unified optimization view on generalized matching pursuit and Frank-Wolfe. *AISTATS*, 2017.

[8] F. Locatello, A. Raj, S. P. Karimireddy, G. Rätsch, B. Schölkopf, S. U. Stich, and M. Jaggi. On matching pursuit and coordinate descent. *ICML*, 2018.

[9] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 1993.

[10] A. S. Nemirovskii and Y. E. Nesterov. Optimal methods of smooth convex minimization. *USSR Computational Mathematics and Mathematical Physics*, 1985.

[11] N. Rao, S. Shah, and S. Wright. Forward-backward greedy algorithms for atomic norm regularization. *IEEE Transactions on Signal Processing*, 2015.