Blended Matching Pursuit

Cyrille W. Combettes (Georgia Institute of Technology)
Sebastian Pokutta (Zuse Institute Berlin and TU Berlin)

Problem

Let $\mathcal{D} \subset \mathcal{H}$ be a dictionary and $f : \mathcal{H} \to \mathbb{R}$ be a smooth, convex, and coercive function. Solve, without sparsity-inducing constraints:

Problem. For any $\epsilon > 0$, find $x \in \mathcal{H}$ satisfying $f(x) - \min_{\mathcal{H}} f \leqslant \epsilon$ and which is sparse relative to $\mathcal{D}$, i.e., $x = \sum_{i=1}^{m} \lambda_i v_i$ where $v_1, \ldots, v_m \in \mathcal{D}$ and m is *small*.

Preliminaries

Let $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ be a separable Hilbert space with induced norm $\| \cdot \|$. A set $\mathcal{D} \subset \mathcal{H}$ of normalized vectors is a *dictionary* if it is at most countable and $\mathrm{cl}(\mathrm{span}(\mathcal{D})) = \mathcal{H}$, and in this case its elements are referred to as *atoms*. For any set $\mathcal{S} \subseteq \mathcal{H}$, let $\mathcal{S}' \coloneqq \mathcal{S} \cup -\mathcal{S}$ denote the *symmetrization of* $\mathcal{S}$. Let $f : \mathcal{H} \to \mathbb{R}$ be a Fréchet differentiable function. We say that f is:

(i) *L-smooth of order $\ell > 1$* if $L > 0$ and for all $x, y \in \mathcal{H}$,

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leqslant \frac{L}{\ell} \|y - x\|^{\ell},$$

(ii) *S-strongly convex of order $s > 1$* if $S > 0$ and for all $x, y \in \mathcal{H}$,

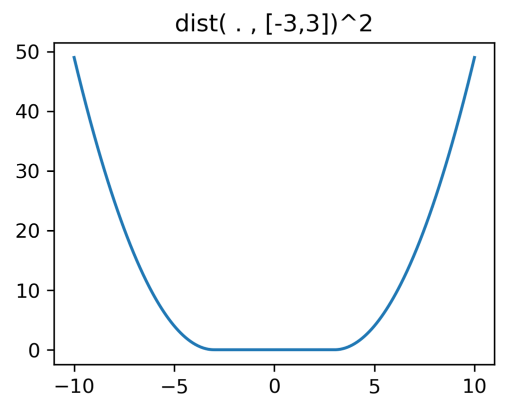
$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \geqslant \frac{S}{s} \|y - x\|^{s},$$

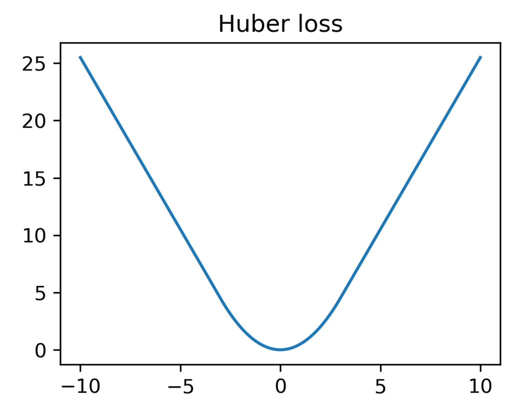
(iii) *C-sharp of order $\theta \in \,]0,1[$ on $\mathcal{K}$* if $\mathcal{K} \subset \mathcal{H}$ is a bounded set, $\varnothing \neq \arg\min_{\mathcal{H}} f \subset \mathrm{int}(\mathcal{K})$, and for all $x \in \mathcal{K}$,

$$\mathrm{dist}\left(x, \arg\min_{\mathcal{H}} f\right) \leqslant C \left(f(x) - \min_{\mathcal{H}} f\right)^{\theta}.$$

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

Fact 2. A strongly convex function is sharp but a (convex and) sharp function is not necessarily strongly convex:





Lemma [1]**.** Sharpness holds for all *well-behaved* convex functions in $\mathbb{R}^n$.

Generalized/Orthogonal Matching Pursuit

Gradient descent follows the optimal descent direction but produces poor sparsity as $-\nabla f(x_t)$ may be a combination of many atoms. To preserve sparsity, **GMP** moves in the direction of an atom $v_t \in \mathcal{D}'$, keeping track of the *active set* $\mathcal{S}_{t+1} = \mathcal{S}_t \cup \{v_t\}$. OMP reoptimizes f over $\mathrm{span}(\mathcal{S}_{t+1})$ and each iteration is typically a sequence of projected gradient steps (**PG steps**). OMP achieves higher sparsity than **GMP** but each iteration is expensive: the sequence of **PG steps** is overkill and can be truncated.

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

Lazification. BMP speeds-up the linear oracle with a *weak-separation* oracle $\mathbf{LPsep}_{\mathcal{D}'}(\nabla f(x_t), \phi_t, \kappa)$ [2]: Find $v_t \in \mathcal{D}'$ such that $\langle \nabla f(x_t), v_t \rangle \leqslant \phi_t / \kappa$.

Blending. BMP blends **GMP steps** with **PG steps**:

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

The idea is to promote **PG steps** as long as the progress offered is *comparable* to that of a **GMP step**. To this end, we want to compare $\min_{v \in \mathcal{S}_t'} \langle \nabla f(x_t), v \rangle$ to $\min_{v \in \mathcal{D}'} \langle \nabla f(x_t), v \rangle$, which quantity is not available because of the lazification.

Dual gap estimates. Hence, we introduce *dual gap estimates* $|\phi_t|$. This designation comes from $\min_{v \in \mathcal{D}'} \langle \nabla f(x_t), v \rangle$ being a dual gap in our setting. Indeed, there exists $\rho > 0$ such that for all $t \in [\![0, T]\!]$ and $x^* \in \arg\min_{\mathcal{H}} f$,

$$\epsilon_t \leqslant \langle \nabla f(x_t), x_t - x^* \rangle \leqslant \max_{u, v \in \rho\,\mathrm{conv}(\mathcal{D}')} \langle \nabla f(x_t), u - v \rangle = -2\rho \min_{v \in \mathcal{D}'} \langle \nabla f(x_t), v \rangle \qquad (1)$$

where $\epsilon_t = f(x_t) - \min_{\mathcal{H}} f$. We initialize $\phi_0 \leftarrow \min_{v \in \mathcal{D}'} \langle \nabla f(x_0), v \rangle / \tau$ so $\epsilon_0 \leqslant 2\tau\rho|\phi_0|$. Then the **criterion** Line 3 measures the progress offered by a **PG step**. If it is not satisfactory, then $\mathbf{LPsep}_{\mathcal{D}'}$ is called to evaluate if there exists a **GMP step** with satisfactory progress. Else, it shows that $\min_{v \in \mathcal{D}'} \langle \nabla f(x_t), v \rangle > \phi_t$ and by (1), $\epsilon_t \leqslant 2\rho|\phi_t|$ so we have detected an improved dual gap estimate. A **dual step** updates $\phi_{t+1} \leftarrow \phi_t / \tau$ and gives $\epsilon_{t+1} \leqslant 2\tau\rho|\phi_{t+1}|$; only **dual steps** update ϕ_t.

Algorithm Blended Matching Pursuit (BMP)

Input: Start atom $x_0 \in \mathcal{D}$, parameter $\eta > 0$ balancing speed vs. sparsity, $\kappa \geqslant 1$, $\tau > 1$.

1: $\mathcal{S}_0, \phi_0 \leftarrow \{x_0\}, \min_{v \in \mathcal{D}'} \langle \nabla f(x_0), v \rangle / \tau$
2: **for** $t = 0$ **to** $T - 1$ **do**
3:   **if** $\min_{v \in \mathcal{S}_t'} \langle \nabla f(x_t), v \rangle \leqslant \phi_t / \eta$ **then**
4:     $\widetilde{\nabla} f(x_t) \leftarrow \mathrm{proj}_{\mathrm{span}(\mathcal{S}_t)}(\nabla f(x_t))$
5:     $x_{t+1} \leftarrow \arg\min_{x_t + \mathbb{R}\widetilde{\nabla} f(x_t)} f$ {**PG step**}
6:     $\mathcal{S}_{t+1}, \phi_{t+1} \leftarrow \mathcal{S}_t, \phi_t$
7:   **else**
8:     $v_t \leftarrow \mathbf{LPsep}_{\mathcal{D}'}(\nabla f(x_t), \phi_t, \kappa)$
9:     **if** $v_t =$ false **then**
10:       $x_{t+1} \leftarrow x_t$ {**dual step**}
11:       $\mathcal{S}_{t+1}, \phi_{t+1} \leftarrow \mathcal{S}_t, \phi_t / \tau$
12:     **else**
13:       $x_{t+1} \leftarrow \arg\min_{x_t + \mathbb{R} v_t} f$ {**GMP step**}
14:       $\mathcal{S}_{t+1}, \phi_{t+1} \leftarrow \mathcal{S}_t \cup \{v_t\}, \phi_t$
15:     **end if**
16:   **end if**
17: **end for**

Convergence results

Properties of f	BMP rate	Complexity lower bound [3]
Smooth convex	$\mathcal{O}\left(\frac{1}{\epsilon^{1/(\ell-1)}}\right)$	$\Omega\left(\frac{1}{\epsilon^{1/(1.5\ell-1)}}\right)$
Smooth convex sharp $\ell\theta = 1$	$\mathcal{O}\left(\ln\left(\frac{1}{\epsilon}\right)\right)$	$\Omega\left(\ln\left(\frac{1}{\epsilon}\right)\right)$
Smooth convex sharp $\ell\theta < 1$	$\mathcal{O}\left(\frac{1}{\epsilon^{(1-\ell\theta)/(\ell-1)}}\right)$	$\Omega\left(\frac{1}{\epsilon^{(1-\ell\theta)/(1.5\ell-1)}}\right)$

Computational experiments

We measure a signal/observe data $y = Ax^* + \mathcal{N}(0, \sigma^2 I_m)$ where $\|x^*\|_0 \ll n$ and we want to recover/learn x^* from the dictionary $\mathcal{D} = \{\pm e_1, \ldots, \pm e_n\}$.

BMP, GMP, and OMP solve

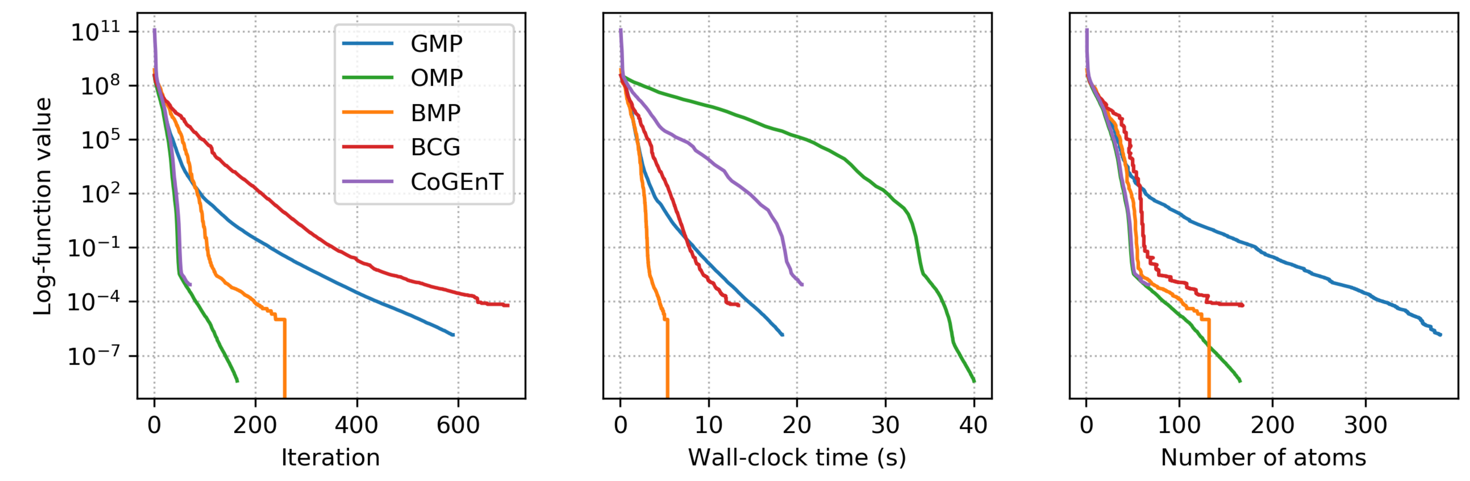
$$\min \; \|y - Ax\|_2^2 \quad \text{s.t.} \; x \in \mathbb{R}^n$$

BCG and CoGEnT solve

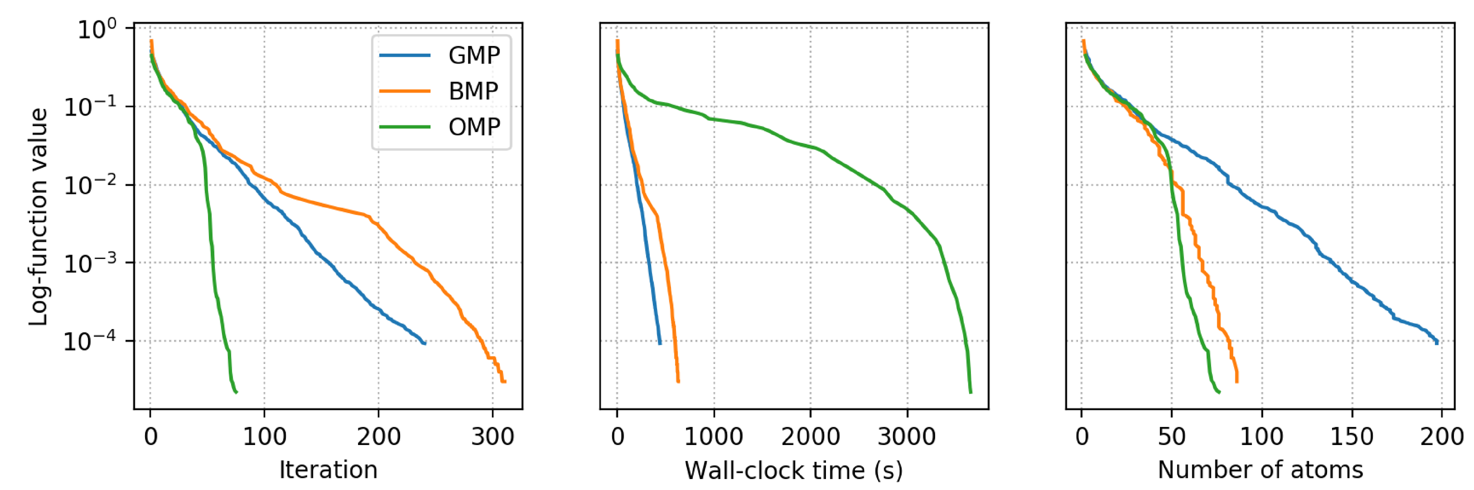
$$\min \; \|y - Ax\|_2^2 \quad \text{s.t.} \; \|x\|_1 \leqslant \|x^*\|_1$$

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

(i) $A \in \mathbb{R}^{250 \times 1000}$ and $f : x \in \mathbb{R}^{1000} \mapsto \|y - Ax\|_3^5$



(ii) Gisette dataset: $f : x \in \mathbb{R}^{5000} \mapsto (1/1000) \sum_{i=1}^{1000} \ln(1 + e^{-y_i a_i^\top x})$



References

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

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

[3] A. Nemirovskii and Y. Nesterov. Optimal methods of smooth convex minimization. *Comput. Math. Math. Phys.*, 1985.