First of all, we would like to thank the reviewers for their valuable feedback. Below we address each reviewer's comments and place in **General** a discussion on the *dual gap estimates* $|\phi_t|$, as it was asked by multiple reviewers.

[General] The *duality gap* defined in the constrained setting (see, e.g., Jaggi [2013]) provides a good quality bound on the primal gap and yields important insights for convergence analyses and algorithm design. In our setting, we can derive a similar inequality as follows. Since $\mathcal{D}'$ is symmetric and $0 \in \operatorname{int}(\operatorname{conv}(\mathcal{D}'))$, there exists (an unknown) $\rho > 0$ such that $\{x^*, x_0, \ldots, x_T\} \subset \rho \operatorname{conv}(\mathcal{D}')$. Define $v_t^{\text{FW}} := \operatorname{argmin}_{v \in \mathcal{D}'} \langle \nabla f(x_t), v \rangle$; note that $\langle \nabla f(x_t), v_t^{\text{FW}} \rangle \leq 0$ since $\mathcal{D}' = -\mathcal{D}'$. Then $\epsilon_t := f(x_t) - f(x^*) \leq \langle \nabla f(x_t), x_t - x^* \rangle \leq \max_{(u,v) \in (\rho \operatorname{conv}(\mathcal{D}'))^2} \langle \nabla f(x_t), u - v \rangle = -2\rho \langle \nabla f(x_t), v_t^{\text{FW}} \rangle$ **(1)**, which is our desired inequality. The design of BMP strongly relies on this last quantity, essential to the blending of full steps and constrained steps (see **(2)**) but unfortunately, it involves the unknown ρ so it is not functional here. We remedy this by working with *estimations* of these quantities: these are the *dual gap estimates* $|\phi_t|$. We set $\phi_0 \leftarrow \langle \nabla f(x_0), v_0^{\text{FW}} \rangle / \tau \leq 0$ (L2 in Alg. 3) so $f(x_0) - f(x^*) \leq 2\tau\rho|\phi_0|$ by **(1)**. Suppose t' is the first iteration where a *dual step* is taken. For all $t \in [\![0, t']\!]$, we have $\phi_t = \phi_0$ and $\epsilon_t \leq \epsilon_0$, so $\epsilon_t \leq 2\tau\rho|\phi_t|$. However, this inequality becomes looser and looser, as the primal gaps ϵ_t decrease while the $|\phi_t|$ stay the same. Thus, when we detect an improved dual bound via a negative weak-separation oracle call, i.e., $\langle \nabla f(x_t), v_t^{\text{FW}} \rangle > \phi_t$ (see L11-12 in Alg. 3 and the negative call in Oracle 2), the *dual step* updates the estimate $|\phi_t|$ accordingly: by **(1)**, this implies that $\epsilon_t \leq 2\rho|\phi_t|$ so by setting $\phi_{t+1} \leftarrow \phi_t/\tau$ and $x_{t+1} \leftarrow x_t$, we obtain $\epsilon_{t+1} \leq 2\tau\rho|\phi_{t+1}|$. Therefore, a dual step updates $|\phi_t|$ and tightens the bound on the primal gap (since $\tau > 1$). Furthermore, dividing by τ provides a geometric, hence fast, rescaling of the dual gap estimates, which can also be seen in the proofs: the number of required dual steps is $\mathcal{O}(\ln(1/\epsilon))$. In addition, note that ϕ_t is also an estimate for primal progress, roughly $\mathcal{O}(\phi_t^2)$ by smoothness; see **(2)**.

(2): Assume WLOG that f is L-smooth of order $\ell = 2$ and that the atoms have unit norm. It can be shown (see proofs) that the progress $f(x_t) - f(x_{t+1})$ is at least $\langle \nabla f(x_t), v_t \rangle^2 / 2L$ for a full step, $\langle \nabla f(x_t), v_t^{\text{FW-}\mathcal{S}} \rangle^2 / 2L$ for a constrained step, and $\langle \nabla f(x_t), v_t^{\text{FW}} \rangle^2 / 2L$ in GMP. Since ϕ_t is a (scaled) estimation of $\langle \nabla f(x_t), v_t^{\text{FW}} \rangle$, the criterion L5 in Alg. 3 tests whether a constrained step would yield progress within a multiplicative factor to that of a GMP step, and this without adding a new atom (hence preserving sparsity). The second test at L11-12 for deciding between a full step (which is a GMP step with a weak-separation oracle) and a dual step is discussed above.

[Reviewer #1] 1. `for` loops: Indeed these are the typical example of poor Python performance. However, we compare all methods using the same code framework and we suspect the results to differ only by a constant factor from those using `numba`. We will provide evidence of this and modify the code to adapt to `numba`.
2. Data generation: We reproduced the CoGEnT experiments for comparison purposes, where the data is generated using i.i.d. Gaussians and the design can be near orthogonal indeed. However we are in the process of conducting additional experiments with more complex structures. These will be added in the revised version once all is finalized.
3. L15: Sparsity of the solution is indeed what is important however, in some applications, it is not necessarily the last iterate: an earlier iterate might be selected as solution as it might provide better test error and avoid overfitting (early stopping). Note that the lasso regularization or the ℓ_1-ball constrained method also maintain sparsity at each iterate.
4. L35: Agreed, we mean that BMP does not require RIP or incoherence properties for its convergence to be analyzed.
5. Fact 2.1: Yes the cited paper [Nemirovskii and Nesterov, 1985] mentions such a result, we will provide the reference.
6. L83: Indeed, we will specify that we assume the function to be smooth.
7. About ϕ_t: Please see **General**.
8. Code not working: We used underscores in the case where some values can be ignored when unpacking, e.g., `a, _, c = (1, 2, 3)`, and this raised an error on your machine. We prepared an online version at https://colab.research.google.com/drive/1VpVET2_lw6kEXttRabIfpHdFiwo4Hp1V. Apart from Figure 2, each experiment takes an average total ∼15 minutes to run. The time limit can be reduced by setting `time_tol` (in seconds) to a lower value; `time_tol` is set at the beginning of each experiment.
9. L97: Indeed the correct spelling is Łojasiewicz. L152: Yes we could use λ_{i_j} since the correction option L22 is never used in our experiments. L66: We will remove the $\mathcal{H}^*$ notation and simply use $\operatorname{argmin}_{\mathcal{H}} f$.

[Reviewer #2] 1. The coordinate-like steps correspond to the *full steps* $x_{t+1} \leftarrow x_t + \gamma_t v_t$ (L17), where a step is taken in the direction of an atom. This reduces to a coordinate step if the atoms are the canonical vectors (in finite dimension) $\pm e_1, \ldots, \pm e_n$. The gradient descent steps are the *constrained steps* $x_{t+1} \leftarrow x_t - \gamma_t \widetilde{\nabla} f(x_t)$ (L7).
2. Actually $\operatorname{span}(\mathcal{D}) = \mathcal{H}$ by definition of $\mathcal{D}$. There is no need to address the case $\operatorname{span}(\mathcal{D}) \subsetneq \mathcal{H}$ as a simple reduction $\mathcal{H} \leftarrow \operatorname{span}(\mathcal{D})$ would yield the same analyses.
3. Yes we will remove the $\mathcal{H}^*$ notation and simply use $\operatorname{argmin}_{\mathcal{H}} f$.
4. We have no evidence of this and it is an open problem we are interested in.
5. Indeed we could comment, e.g., that the rate derived is the same as that of the sharp case, up to a constant factor. Note however that the sharp case subsumes the strongly convex case as mentioned L83-85, L95-96, and L221-223.
6. The term "lazified" comes from the cited paper [Braun et al., 2017].

[Reviewer #3] Please see **General**.