We thank the reviewers for their feedback.

On the role of $x^\star_{min}$ (R1, R3)

We expect that the assumption $x^\star_{min} \geq \Omega(\|\mathbf{x}^\star\|_2/\sqrt{k})$ (where $x^\star_{min} = \min_{i:x^\star_i \neq 0} |x^\star_i|$) is likely an artifact of our proof. However, we expect a proof without this assumption (if feasible) to be more complicated, possibly distracting from the main ideas of the (already lengthy) proof. On the experimental side, both our setting and the setting considered in [53] uses Gaussian signals without any restriction on $x^\star_{min}$, which might indicate that the assumption on $x^\star_{min}$ is in fact not necessary for mirror descent to reconstruct the signal $\mathbf{x}^\star$.

On the theoretical side, the assumption on $x^\star_{min}$ appears in two places in our analysis. On a high level, the inner product $\mathbf{X}(t)^T\mathbf{x}^\star$ is a key quantity in showing the convergence of mirror descent. Using the fact that we have no mismatched signs (Lemma 5 equation (20)), we have the simple lower bound $|\mathbf{X}(t)^T\mathbf{x}^\star| = |\sum_{i=1}^n X_i(t)x^\star_i| \geq \|\mathbf{X}(t)_{\mathcal{S}}\|_1 x^\star_{min}$, where $\mathcal{S} = \{i : x^\star_i \neq 0\}$ denotes the support of the signal. Our technical lemmas then guarantee that the discrepancy between the empirical gradient ∇F and the population gradient ∇f is sufficiently small compared to $\mathbf{X}(t)^T\mathbf{x}^\star$. We believe that it might be possible to control the inner product via a more refined analysis of the trajectory of mirror descent instead of assuming $x^\star_{min} \geq \Omega(\|\mathbf{x}^\star\|_2/\sqrt{k})$, however it is likely to require different techniques from the ones used in our analysis. Second, the assumption on $x^\star_{min}$ allows us to separate *all* support coordinates from off-support coordinates at the end of the initial warm-up stage, in the sense that $|X_i(t)| \gg |X_j(t)|$ for all $i \in \mathcal{S}$, $j \notin \mathcal{S}$. Intuitively, we neither expect nor need $|X_i(t)| > |X_j(t)|$ if $|x^\star_i|$ is very small. Rather, it should suffice if above inequality holds for $i \in \mathcal{S}$ corresponding to "large" coordinates. We anticipate that it might be possible to make this intuition rigorous, potentially utilising similar tools as the ones used in [10] to eliminate the need for an assumption on $x^\star_{min}$.

Experiment in a setting with $k^2 < m \ll n$ (R3)

Following the reviewer's suggestion, we repeated the experiment of Section 5 in various settings with $k^2 < m \ll n$. We present an example in Figure 2, where we increased the dimension of the signal to $n = 50000$ and kept everything else as described in Section 5. We observe the same qualitative behaviour as in Figure 1 (we only include the relative ℓ_2 error, as the Bregman divergence also shows the same behaviour as in Figure 1).

On noise in the measurement model (R1)

We ran discrete-time mirror descent in a measurement model with additive white Gaussian noise, $Y_j = (\mathbf{A}_j^T\mathbf{x}^\star)^2 + \varepsilon_j$ where $\varepsilon_j \sim \mathcal{N}(0, \sigma^2)$ i.i.d. for some $\sigma^2 > 0$. We show the results of an experiment with $n = 50000$, $m = 1000$, $k = 10$ and $\sigma^2 = 0.1$ below. Figure 3 suggests that mirror descent can also reconstruct sparse signals in the model with noise, and the parameter β seems to affect convergence in a similar way as in the noiseless case. The precision up to which we have linear convergence barely improves as we decrease β from 10^{-10} to 10^{-14}, which we suspect is because the presence of noise in the measurement model limits the attainable accuracy. In all our experiments, the relative ℓ_2 error increases after reaching a minimum, which suggests that additional techniques such as early stopping might be needed. We leave this to future work, as the analysis of the noisy model is likely to involve novel ideas.

On the improved sample complexity of HWF (R2)

Theorem 2 requires the number of measurements m to be of order k^2 (ignoring logarithmic factors). The empirical results in [53] suggest that HWF is able to reconstruct sparse signals from far fewer measurements ($m < k^2$) if the signal contains one large entry. There are two obvious candidate explanations for this discrepancy: it could be the case that 1) our proof is suboptimal and the sample complexity in Theorem 2 is overly pessimistic, or that 2) the statement of Theorem 2 does not hold if $m < k^2$, regardless of the maximum magnitude entry of the signal (Theorem 2 not only guarantees convergence towards the underlying signal, but also characterizes the speed of convergence). To investigate which of the two explanations seems more likely, we consider an experiment with $n = 1000$, $m = 500$, $k = 100$ and one entry of the signal set to 0.7. In this setting we have $m < k^2$, and the assumptions of Theorem 2 are not satisfied. Figure 4 suggests that explanation 2) seems more likely: while we have convergence towards the underlying signal, the convergence behaviour is not as described by Theorem 2. In particular, we do not observe linear convergence up to a precision depending on β.

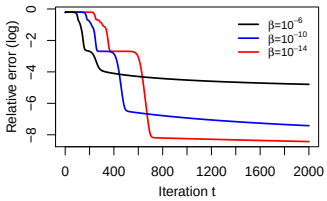

Figure 2: Relative ℓ_2 error (log-scale) of HWF for $n = 50000$, $m = 1000$ and $k = 10$ (no noise).

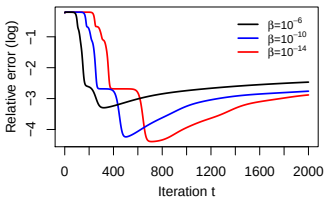

Figure 3: Relative ℓ_2 error (log-scale) of HWF for $n = 50000$, $m = 1000$, $k = 10$ and $\sigma^2 = 0.1$.

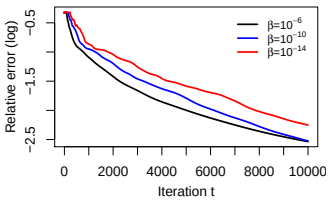

Figure 4: Relative ℓ_2 error (log-scale) of HWF for $n = 1000$, $m = 500$, $k = 100$ and one entry of $\mathbf{x}^\star$ set to 0.7 (no noise).