Auxiliary Material
Efficient Learning using Forward-Backward Splitting

A Batch Convergence Proofs and Corollaries

Before we begin the proofs, we provide a technical lemma.

Lemma 5 (Bounding Step Differences)**.** *Assume that the norms of the subgradients of the functions f and r are bounded as in Eq. (5):*

$$\|\partial f(\boldsymbol{w})\|^2 \leq Af(\boldsymbol{w}) + G^2, \quad \|\partial r(\boldsymbol{w})\|^2 \leq Ar(\boldsymbol{w}) + G^2 \ .$$

Let $\eta_{t+1} \leq \eta_{t+\frac{1}{2}} \leq \eta_t$ and suppose that $\eta_t \leq 2\eta_{t+\frac{1}{2}}$. If we use the FOBOS *update of Eqs. (2) and (3), then for a constant $c \leq 4$ and any vector $\boldsymbol{w}^\star$,*

$$\begin{aligned} &2\eta_t(1 - cA\eta_t)f(\boldsymbol{w}_t) + 2\eta_{t+\frac{1}{2}}(1 - cA\eta_{t+\frac{1}{2}})r(\boldsymbol{w}_{t+1}) \\ &\quad \leq \ 2\eta_t f(\boldsymbol{w}^\star) + 2\eta_{t+\frac{1}{2}} r(\boldsymbol{w}^\star) + \|\boldsymbol{w}_t - \boldsymbol{w}^\star\|^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{w}^\star\|^2 + 7\eta_t\eta_{t+\frac{1}{2}}G^2 \ . \end{aligned} \tag{12}$$

Proof of Lemma 5 We begin with a few simple consequences of the forward-looking subgradient steps before proceeding with the core of the proof. Note first that for some $\boldsymbol{g}_t^f \in \partial f(\boldsymbol{w}_t)$ and $\boldsymbol{g}_{t+1}^r \in \partial r(\boldsymbol{w}_{t+1})$, we have as in Eq. (4)

$$\boldsymbol{w}_{t+1} - \boldsymbol{w}_t = -\eta_t \boldsymbol{g}_t^f - \eta_{t+\frac{1}{2}} \boldsymbol{g}_{t+1}^r \ . \tag{13}$$

The definition of a subgradient implies that for any $\boldsymbol{g}_{t+1}^r \in \partial r(\boldsymbol{w}_{t+1})$ (and similarly for any $\boldsymbol{g}_t^f \in \partial f(\boldsymbol{w}_t)$ with $f(\boldsymbol{w}_t)$ and $f(\boldsymbol{w}^\star)$) implies that

$$r(\boldsymbol{w}^\star) \geq r(\boldsymbol{w}_{t+1}) + \left\langle \boldsymbol{g}_{t+1}^r, \boldsymbol{w}^\star - \boldsymbol{w}_{t+1} \right\rangle \quad \Rightarrow \quad -\left\langle \boldsymbol{g}_{t+1}^r, \boldsymbol{w}_{t+1} - \boldsymbol{w}^\star \right\rangle \leq r(\boldsymbol{w}^\star) - r(\boldsymbol{w}_{t+1}). \tag{14}$$

From the Cauchy-Shwartz Inequality and Eq. (13), we obtain

$$\begin{aligned} \left\langle \boldsymbol{g}_{t+1}^r, (\boldsymbol{w}_{t+1} - \boldsymbol{w}_t) \right\rangle \ &= \ \left\langle \boldsymbol{g}_{t+1}^r, (-\eta_t \boldsymbol{g}_t^f - \eta_{t+\frac{1}{2}} \boldsymbol{g}_{t+1}^r) \right\rangle \\ &\leq \ \|\boldsymbol{g}_{t+1}^r\| \|\eta_{t+\frac{1}{2}} \boldsymbol{g}_{t+1}^r + \eta_t \boldsymbol{g}_t^f\| \ \leq \ \eta_{t+\frac{1}{2}} \|\boldsymbol{g}_{t+1}^r\|^2 + \eta_t \|\boldsymbol{g}_{t+1}^r\| \|\boldsymbol{g}_t^f\| \\ &\leq \ \eta_{t+\frac{1}{2}} \left(Ar(\boldsymbol{w}_{t+1}) + G^2\right) + \eta_t \max\left\{Af(\boldsymbol{w}_t) + G^2, Ar(\boldsymbol{w}_{t+1}) + G^2\right\} \ . \end{aligned} \tag{15}$$

We now proceed to bound the difference between $\boldsymbol{w}^\star$ and $\boldsymbol{w}_{t+1}$, and using a telescoping sum we will eventually bound $f(\boldsymbol{w}_t) + r(\boldsymbol{w}_t) - f(\boldsymbol{w}^\star) - r(\boldsymbol{w}^\star)$. First, we expand norm squared of the difference between $\boldsymbol{w}_t$ and $\boldsymbol{w}_{t+1}$,

$$\begin{aligned} \|\boldsymbol{w}_{t+1} - \boldsymbol{w}^\star\|^2 \ &= \ \|\boldsymbol{w}_t - (\eta_t \boldsymbol{g}_t^f + \eta_{t+\frac{1}{2}} \boldsymbol{g}_{t+1}^r) - \boldsymbol{w}^\star\|^2 \\ &= \ \|\boldsymbol{w}_t - \boldsymbol{w}^\star\|^2 - 2\left[\eta_t \left\langle \boldsymbol{g}_t^f, \boldsymbol{w}_t - \boldsymbol{w}^\star \right\rangle + \eta_{t+\frac{1}{2}} \left\langle \boldsymbol{g}_{t+1}^r, \boldsymbol{w}_t - \boldsymbol{w}^\star \right\rangle\right] + \|\eta_t \boldsymbol{g}_t^f + \eta_{t+\frac{1}{2}} \boldsymbol{g}_{t+1}^r\|^2 \\ &= \ \|\boldsymbol{w}_t - \boldsymbol{w}^\star\|^2 - 2\eta_t \left\langle \boldsymbol{g}_t^f, \boldsymbol{w}_t - \boldsymbol{w}^\star \right\rangle + \|\eta_t \boldsymbol{g}_t^f + \eta_{t+\frac{1}{2}} \boldsymbol{g}_{t+1}^r\|^2 \\ &\quad -2\eta_{t+\frac{1}{2}} \left[\left\langle \boldsymbol{g}_{t+1}^r, \boldsymbol{w}_{t+1} - \boldsymbol{w}^\star \right\rangle - \left\langle \boldsymbol{g}_{t+1}^r, \boldsymbol{w}_{t+1} - \boldsymbol{w}_t \right\rangle\right] \ . \end{aligned} \tag{16}$$

We can bound the third term above by noting that

$$\begin{aligned} &\|\eta_t \boldsymbol{g}_t^f + \eta_{t+\frac{1}{2}} \boldsymbol{g}_{t+1}^r\|^2 \\ &\quad = \ \eta_t^2 \|\boldsymbol{g}_t^f\|^2 + 2\eta_t\eta_{t+\frac{1}{2}} \left\langle \boldsymbol{g}_t^f, \boldsymbol{g}_{t+1}^r \right\rangle + \eta_{t+\frac{1}{2}}^2 \|\boldsymbol{g}_{t+1}^r\|^2 \\ &\quad \leq \ \eta_t^2 Af(\boldsymbol{w}_t) + 2A\eta_t\eta_{t+\frac{1}{2}} \max\left\{f(\boldsymbol{w}_t), r(\boldsymbol{w}_{t+1})\right\} + \eta_{t+\frac{1}{2}}^2 Ar(\boldsymbol{w}_{t+1}) + 4\eta_t^2 G^2 \ . \end{aligned}$$

We now use Eq. (15) to bound the last term of Eq. (16) and the above bound on $\eta_t \boldsymbol{g}_t^f + \eta_{t+\frac{1}{2}} \boldsymbol{g}_{t+1}^r$ to get that

$$\begin{aligned}
&\|\boldsymbol{w}_{t+1} - \boldsymbol{w}^\star\|^2 \\
&\leq \|\boldsymbol{w}_t - \boldsymbol{w}^\star\|^2 - 2\eta_t \left\langle \boldsymbol{g}_t^f, \boldsymbol{w}_t - \boldsymbol{w}^\star \right\rangle - 2\eta_{t+\frac{1}{2}} \left\langle \boldsymbol{g}_{t+1}^r, \boldsymbol{w}_{t+1} - \boldsymbol{w}^\star \right\rangle + \|\eta_t \boldsymbol{g}_t^f + \eta_{t+\frac{1}{2}} \boldsymbol{g}_{t+1}^r\|^2 \\
&\quad + 2\eta_{t+\frac{1}{2}} \left(\eta_{t+\frac{1}{2}} A r(\boldsymbol{w}_{t+1}) + 2A\eta_t \max\{f(\boldsymbol{w}_t), r(\boldsymbol{w}_{t+1})\} + 2\eta_t G^2 \right) \\
&\leq \|\boldsymbol{w}_t - \boldsymbol{w}^\star\|^2 + 2\eta_t \left(f(\boldsymbol{w}^\star) - f(\boldsymbol{w}_t)\right) + 2\eta_{t+\frac{1}{2}} \left(r(\boldsymbol{w}^\star) - r(\boldsymbol{w}_t)\right) + 7\eta_t^2 G^2 \\
&\quad + \eta_t^2 A f(\boldsymbol{w}_t) + 3A\eta_t \eta_{t+\frac{1}{2}} \max\left\{f(\boldsymbol{w}_t), r(\boldsymbol{w}_t)\right\} + 2\eta_{t+\frac{1}{2}}^2 A r(\boldsymbol{w}_{t+1}) \qquad (17) \\
&\leq \|\boldsymbol{w}_t - \boldsymbol{w}^\star\|^2 + 7\eta_t^2 G^2 \\
&\quad + 2\eta_t \left(f(\boldsymbol{w}^\star) - (1 - cA\eta_t) f(\boldsymbol{w}_t)\right) + 2\eta_{t+\frac{1}{2}} \left(r(\boldsymbol{w}^\star) - (1 - cA\eta_{t+\frac{1}{2}}) r(\boldsymbol{w}_{t+1}) \right) \ . \qquad (18)
\end{aligned}$$

To obtain Eq. (17) we used the standard convexity bounds established earlier in Eq. (14). The final bound given by Eq. (18) is due to the fact that $3A\eta_t\eta_{t+\frac{1}{2}} \leq 6A\eta_t^2$ and that for $a, b \geq 0$, $\max\{a, b\} \leq a + b$. Moving the $f(\cdot)$ and $r(\cdot)$ terms to the left hand side of the $\leq$ gives the desired inequality. □

Using the above lemma, the analysis for FOBOS in a batch setting is straightforward. In this setting we set $\eta_{t+\frac{1}{2}} = \eta_{t+1}$ and update $\boldsymbol{w}_t$ to $\boldsymbol{w}_{t+1}$ as prescribed by Eq. (2) and Eq. (3).

Proof of Theorem 1 Rearranging the $f(\boldsymbol{w}^\star)$ and $r(\boldsymbol{w}^\star)$ terms from the bound in lemma 5, we sum the loss terms over t from 1 through T and get a canceling telescoping sum:

$$\begin{aligned}
&\sum_{t=1}^{T} \left[\eta_t \left((1 - cA\eta_t) f(\boldsymbol{w}_t) - f(\boldsymbol{w}^\star)\right) + \eta_{t+1} \left((1 - cA\eta_{t+1}) r(\boldsymbol{w}_{t+1}) - r(\boldsymbol{w}^\star)\right)\right] \\
&\quad \leq \|\boldsymbol{w}_1 - \boldsymbol{w}^\star\|^2 - \|\boldsymbol{w}_{T+1} - \boldsymbol{w}^\star\|^2 + 7G^2 \sum_{t=1}^{T} \eta_t^2 \ \leq \ \|\boldsymbol{w}_1 - \boldsymbol{w}^\star\|^2 + 7G^2 \sum_{t=1}^{T} \eta_t^2 \ . \qquad (19)
\end{aligned}$$

Now we bound the one-off $r(\boldsymbol{w}_{t+1})$ terms by noting that

$$\begin{aligned}
&\sum_{t=1}^{T} \eta_{t+1} \left((1 - cA\eta_{t+1}) r(\boldsymbol{w}_{t+1}) - r(\boldsymbol{w}^\star)\right) \\
&= \sum_{t=1}^{T} \eta_t \left((1 - cA\eta_t) r(\boldsymbol{w}_t) - r(\boldsymbol{w}^\star)\right) + \eta_{T+1} \left((1 - cA\eta_{T+1}) r(\boldsymbol{w}_{t+1}) - r(\boldsymbol{w}^\star)\right) + \eta_1 r(\boldsymbol{w}^\star) \\
&\geq \sum_{t=1}^{T} \eta_t \left((1 - cA\eta_t) r(\boldsymbol{w}_{t+1}) - r(\boldsymbol{w}^\star)\right) + r(\boldsymbol{w}^\star)(\eta_1 - \eta_{T+1}) \\
&\geq \sum_{t=1}^{T} \eta_t \left((1 - cA\eta_t) r(\boldsymbol{w}_t) - r(\boldsymbol{w}^\star)\right) \ . \qquad (20)
\end{aligned}$$

Using the fact that $\|\boldsymbol{w}_1 - \boldsymbol{w}^\star\| = \|\boldsymbol{w}^\star\| \leq D$, we combine Eq. (19)) with Eq. (20) to get the desired bound. □

Corollary 6 (Convergence of decreasing step sizes)**.** *Assume that the conditions of Thm. 1 hold and the step sizes η_t are such that $\eta_t \to 0$, and that $\sum_{t=1}^{\infty} \eta_t = \infty$. Then*

$$\liminf_{t \to \infty} f(\boldsymbol{w}_t) + r(\boldsymbol{w}_t) - \left(f(\boldsymbol{w}^\star) + r(\boldsymbol{w}^\star)\right) = 0 \ .$$

We can give tighter convergence results when we assume that f and r are Lipschitz or when we can guarantee that the $\|\partial f\|$ and $\|\partial r\|$ are bounded. In this case, we have

Corollary 7. *In addition to the conditions of Thm. 1, assume that the norm of any subgradient from ∂f and the norm of any subgradient from ∂r are bounded by G. Then*

$$\min_{t\in\{1,\ldots,T\}} (f(\boldsymbol{w}_t) + r(\boldsymbol{w}_t)) - (f(\boldsymbol{w}^\star) + r(\boldsymbol{w}^\star)) \leq \frac{D^2 + 7G^2\sum_{t=1}^T \eta_t^2}{2\sum_{t=1}^T \eta_t}. \tag{21}$$

Corollary 8 (Optimal fixed step rate)**.** *Assume that the conditions of Cor. 7 hold and that we run* FOBOS *for a* predefined *T iterations with $\eta_t = \frac{D}{\sqrt{7T}G}$. Then the following bound holds.*

$$\min_{i\in\{1,\ldots,T\}} f(\boldsymbol{w}_t) + r(\boldsymbol{w}_t) - (f(\boldsymbol{w}^\star) + r(\boldsymbol{w}^\star)) \leq \frac{3DG}{\sqrt{T}}.$$

In corollaries 7 and 8, our assumption on bounded subgradients of the functions f and r is in practice not restrictive. If we know that an optimal $\boldsymbol{w}^\star$ lies in some closed and bounded set Ω and that $\Omega \subseteq \mathrm{dom}(f+r)$, then [17, Theorem 24.7] guarantees that ∂f and ∂r are bounded for $\boldsymbol{w} \in \Omega$. The lingering question is thus whether we can guarantee that such a set Ω exists and that our iterates $\boldsymbol{w}_t$ remain in Ω. We now describe a simple setting to show that ∂f and ∂r are indeed often bounded. If $r(\boldsymbol{w})$ is a norm and f is lower bounded by 0, then we know that $r(\boldsymbol{w}^\star) \leq f(\boldsymbol{w}^\star) + r(\boldsymbol{w}^\star) \leq f(\boldsymbol{w}_1) + r(\boldsymbol{w}_1)$. Using standard bounds on norms, we get that for some $\gamma > 0$

$$\|\boldsymbol{w}^\star\|_\infty \leq \gamma r(\boldsymbol{w}^\star) \leq \gamma(f(\boldsymbol{w}_1) + r(\boldsymbol{w}_1)) = \gamma f(\boldsymbol{w}_1) \ ,$$

where for the last inequality we used the assumption that $r(\boldsymbol{w}_1) = 0$. Thus, we obtain that $\boldsymbol{w}^\star$ lies in a hypercube. We can easily project onto this box by truncating elements of $\boldsymbol{w}_t$ lying outside it at any iteration without affecting the bounds in Eq. (21). Concretely, this follows since Euclidean projection Π_Ω to a convex set Ω with $\boldsymbol{w}^\star \in \Omega$ satisfies $\|\Pi_\Omega(\boldsymbol{w}_{t+1}) - \boldsymbol{w}^\star\| \leq \|\boldsymbol{w}_{t+1} - \boldsymbol{w}^\star\|$. Further, so long as Ω is a norm ball, we know that

$$r(\Pi_\Omega(\boldsymbol{w}_{t+1})) \leq r(\boldsymbol{w}_{t+1}) \ . \tag{22}$$

Thus, looking at Eq. (17) in our proof of Theorem 1 we notice that $r(\boldsymbol{w}^\star) - r(\boldsymbol{w}_{t+1}) \leq r(\boldsymbol{w}^\star) - r(\Pi_\Omega(\boldsymbol{w}_{t+1}))$ and the series of inequalities through Eq. (18) still hold (with $A = 0$). In general, so long as Eq. (22) holds and $\boldsymbol{w}^\star \in \Omega$, we can project $\boldsymbol{w}_{t+1}$ into Ω without affecting convergence guarantees.

B Online Regret Proofs and Corollaries

Proof of Theorem 3 Looking at lemma 5, we immediately see that if $\|\partial f\|$ and $\|\partial r\|$ are bounded by G,

$$f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{w}^\star) + r(\boldsymbol{w}_{t+1}) - r(\boldsymbol{w}^\star) \leq \frac{1}{2\eta_t}\left(\|\boldsymbol{w}_t - \boldsymbol{w}^\star\|^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{w}^\star\|^2\right) + \frac{7}{2}G^2\eta_t. \tag{23}$$

Now we use Eq. (23) to obtain that

$$\begin{aligned} R_{f+r}(T) &= \sum_{t=1}^T (f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{w}^\star) + r(\boldsymbol{w}_t) - r(\boldsymbol{w}^\star)) + r(\boldsymbol{w}_{T+1}) - r(\boldsymbol{w}^\star) - r(\boldsymbol{w}_1) + r(\boldsymbol{w}^\star) \\ &\leq GD + \sum_{t=1}^T \frac{1}{2\eta_t}\left(\|\boldsymbol{w}_t - \boldsymbol{w}^\star\|^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{w}^\star\|^2\right) + \frac{7G^2}{2}\sum_{t=1}^T \eta_t \end{aligned}$$

since $r(\boldsymbol{w}) \leq r(\boldsymbol{0}) + G\|\boldsymbol{w}\| \leq GD$. We can rewrite the above bound and see

$$\begin{aligned} R_{f+r}(T) &\leq GD + \frac{1}{2\eta_1}\|\boldsymbol{w}_1 - \boldsymbol{w}^\star\|^2 + \frac{1}{2}\sum_{t=2}^T \|\boldsymbol{w}_t - \boldsymbol{w}^\star\|^2\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \frac{7G^2}{2}\sum_{t=1}^T \eta_t \\ &\leq GD + \frac{D^2}{2\eta_1} + \frac{D^2}{2}\sum_{t=2}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \frac{7G^2}{2}\sum_{t=1}^T \eta_t \ , \end{aligned}$$

where we used again the bound on the distance of each $\boldsymbol{w}_t$ to $\boldsymbol{w}^\star$ for the last inequality. Lastly, we use the fact that the sum $\frac{1}{\eta_1} + \sum_{t=2}^{T}(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}})$ telescopes and get that

$$R_{f+r}(T) \;\leq\; GD + \frac{D^2}{2\eta_T} + \frac{7G^2}{2}\sum_{t=1}^{T}\eta_t \;\;.$$

Setting $\eta_t = c/\sqrt{t}$ and recognizing as in [7] that $\sum_{t=1}^{T}\eta_t \leq 2c\sqrt{T}$ concludes the proof. □

We assume as in Sec. 3 that we are minimizing $f(\boldsymbol{w}) + r(\boldsymbol{w})$. Suppose that on each step of FOBOS, we choose instead of some $\boldsymbol{g}_t^f \in \partial f(\boldsymbol{w}_t)$ a stochastic estimate of the gradient $\tilde{\boldsymbol{g}}_t^f$ where $E[\tilde{\boldsymbol{g}}_t^f] \in \partial f(\boldsymbol{w}_t)$. We assume that we still use the true r (which is generally easy, as it is simply the regularization function). It is straightforward to use Theorem 3 above as in the derivation of Theorems 2 and 3 from [14] to derive the following corollary on the convergence rate of stochastic FOBOS.

Corollary 9. *Assume that the conditions on ∂f, ∂r, and $\boldsymbol{w}^\star$ hold as in the previous theorems and let* FOBOS *be run for T iterations. Let s be an integer chosen uniformly at random from $\{1, \ldots, T\}$. If $\eta_t = \frac{D}{4G\sqrt{t}}$, then*

$$E_s[f(\boldsymbol{w}_s) + r(\boldsymbol{w}_s)] \leq f(\boldsymbol{w}^\star) + r(\boldsymbol{w}^\star) + \frac{2GD + 4GD\sqrt{T}}{T}.$$

With probability at least $1-\delta$, $f(\boldsymbol{w}_s) + r(\boldsymbol{w}_s) \leq f(\boldsymbol{w}^\star) + r(\boldsymbol{w}^\star) + \frac{2GD+4GD\sqrt{T}}{\delta T}$.

C High-dimensional Efficiency

Proof of Proposition 4 It suffices to show that the proposition is correct for $T = 2$ and then use an inductive argument, because the proposition trivially holds for $T = 1$. We provide here a direct proof for each norm separately by examining the updates we derived in Sec. 4 and showing that $\boldsymbol{w}_2 = \boldsymbol{w}^\star$.

Note that the objective functions are separable for $q = 1$. Therefore, for ℓ_1-regularization it suffices to prove the proposition for any component of the vector $\boldsymbol{w}$. We omit the index of the component and denote by $w_0, w_1, w_2, w_3, \ldots$ one coordinate of $\boldsymbol{w}$ along the iterations of $\mathcal{P}.1$ and by $w^\star$ the result for the same component when solving $\mathcal{P}.2$. We need to show that $w^\star = w_2$. Expanding the ℓ_1-update of Eq. (6) over two iterations we get the following:

$$\begin{aligned} w_2 \;&=\; \mathrm{sign}(w_1)\,[|w_1| - \lambda_2]_+ \;=\; \mathrm{sign}(w_1)\,\big[\big|\mathrm{sign}(w_0)\,[|w_0| - \lambda_1]_+\big| - \lambda_2\big]_+ \\ &=\; \mathrm{sign}(w_0)\,[|w_0| - \lambda_1 - \lambda_2]_+ \;\;, \end{aligned}$$

where we used the positivity of $|\cdot|$. Examining $\mathcal{P}.2$ and using Eq. (6) again we get

$$w^\star = \mathrm{sign}(w_0)\,[|w_0| - \lambda_1 - \lambda_2]_+ \;\;.$$

Therefore, $w^\star = w_2$ as claimed.

Next we prove the proposition for ℓ_2, returning to using the entire vector for the proof. Using the explicit ℓ_2-update from Eq. (7), we can expand the norm of the vector $\boldsymbol{w}_1$ due to the program $\mathcal{P}.1$ as follows,

$$\|\boldsymbol{w}_1\| \;=\; \left[1 - \frac{\lambda_1}{\|\boldsymbol{w}_0\|}\right]_+ \|\boldsymbol{w}_0\| \;=\; [\,\|\boldsymbol{w}_0\| - \lambda_1]_+ \;\;.$$

Similarly, we get that $\|\boldsymbol{w}_2\| = [\|\boldsymbol{w}_1\| - \lambda_2]_+$. Combining the norm equalities we see that the norm of $\boldsymbol{w}_2$ due to the succession of the two updates is

$$\|\boldsymbol{w}_2\| \;=\; \big[[\|\boldsymbol{w}_0\| - \lambda_1]_+ - \lambda_2\big]_+ \;=\; [\|\boldsymbol{w}_0\| - \lambda_1 - \lambda_2]_+ \;\;.$$

Computing directly the norm of $\boldsymbol{w}^\star$ due to the update given by Eq. (7) yields

$$\|\boldsymbol{w}^\star\| \;=\; \left[1 - \frac{\lambda_1 + \lambda_2}{\|\boldsymbol{w}_0\|}\right]_+ \|\boldsymbol{w}_0\| \;=\; [\|\boldsymbol{w}_0\| - \lambda_1 - \lambda_2]_+ \;\;.$$

Thus, $\boldsymbol{w}^\star$ and $\boldsymbol{w}_2$ have the same norm. Since the update itself retains the direction of the original vector $\boldsymbol{w}_0$, we get that $\boldsymbol{w}^\star = \boldsymbol{w}_2$ as needed.

We now turn to the most complicated update and proof of the three norms, the ℓ_∞ norm. We start by recapping the programs $\mathcal{P}.1$ and $\mathcal{P}.2$ for $T = 2$ and $q = \infty$,

$$\mathcal{P}.1: \quad \boldsymbol{w}_1 = \underset{\boldsymbol{w}}{\operatorname{argmin}} \left\{ \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}_0\|^2 + \lambda_1 \|\boldsymbol{w}\|_\infty \right\} \tag{24}$$

$$\boldsymbol{w}_2 = \underset{\boldsymbol{w}}{\operatorname{argmin}} \left\{ \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{w}_1\|^2 + \lambda_2 \|\boldsymbol{w}\|_\infty \right\} \ , \tag{25}$$

$$\mathcal{P}.2: \quad \boldsymbol{w}^\star = \underset{\boldsymbol{w}}{\operatorname{argmin}} \left\{ \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{v}\|_2^2 + (\lambda_1 + \lambda_2) \|\boldsymbol{w}\|_\infty \right\} \ . \tag{26}$$

We prove the equivalence of the two programs in two stages. First, we examine the case $\|\boldsymbol{w}_0\|_1 > \lambda_1 + \lambda_2$, and then consider the complement case $\|\boldsymbol{w}_0\|_1 \leq \lambda_1 + \lambda_2$. For concreteness and simplicity, we assume that $\boldsymbol{w}_0 \succeq \boldsymbol{0}$, since, clearly, the objective is symmetric in $\boldsymbol{w}_0$ and $-\boldsymbol{w}_0$. We thus can assume that all entries of $\boldsymbol{w}_0$ are non-negative. In the proof we use the following operators: $[\boldsymbol{v}]_+$ now denotes the positive component of each entry of $\boldsymbol{v}$, $\min\{\boldsymbol{v}, \theta\}$ denotes the component-wise minimum between the elements of $\boldsymbol{v}$ and θ, and likewise $\max\{\boldsymbol{v}, \theta\}$ is the component-wise maximum. Starting with the case $\|\boldsymbol{w}_0\|_1 > \lambda_1 + \lambda_2$, we examine Eq. (24). From Lagrange duality we know that that $\boldsymbol{w}_1 = \boldsymbol{w}_0 - \boldsymbol{\alpha}_1$, where $\boldsymbol{\alpha}_1$ is the solution of

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\boldsymbol{\alpha} - \boldsymbol{w}_0\|_2^2 \ \text{ s.t. } \ \|\boldsymbol{\alpha}\|_1 \leq \lambda_1 \ .$$

As described by [6] and reviewed above in Sec. 4, $\boldsymbol{\alpha}_1 = [\boldsymbol{w}_0 - \theta_1]_+$ for some $\theta_1 \in \mathbb{R}_+$. The form of $\boldsymbol{\alpha}_1$ readily translates to the following form for $\boldsymbol{w}_1$: $\boldsymbol{w}_1 = \boldsymbol{w}_0 - \boldsymbol{\alpha}_1 = \min(\boldsymbol{w}_0, \theta_1)$. Applying similar reasoning to the second step of $\mathcal{P}.1$ yields $\boldsymbol{w}_2 = \boldsymbol{w}_1 - \boldsymbol{\alpha}_2 = \boldsymbol{w}_0 - \boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2$, where $\boldsymbol{\alpha}_2$ is the minimizer of

$$\frac{1}{2}\|\boldsymbol{\alpha} - \boldsymbol{w}_1\|_2^2 = \frac{1}{2}\|\boldsymbol{\alpha} - (\boldsymbol{w}_0 - \boldsymbol{\alpha}_1)\|_2^2 \quad \text{s.t. } \|\boldsymbol{\alpha}\|_1 \leq \lambda_2 \ .$$

Again, we have $\boldsymbol{\alpha}_2 = [\boldsymbol{w}_1 - \theta_2]_+ = [\boldsymbol{w}_0 - \boldsymbol{\alpha}_1 - \theta_2]_+$ for some $\theta_2 \in \mathbb{R}_+$. The successive steps then imply that

$$\boldsymbol{w}_2 = \min\{\boldsymbol{w}_1, \theta_2\} = \min\left\{\min\{\boldsymbol{w}_0, \theta_1\}, \theta_2\right\} \ .$$

We next show that regardless of the ℓ_1-norm of $\boldsymbol{w}_0$, $\theta_2 \leq \theta_1$. Intuitively, if $\theta_2 > \theta_1$, the second minimization step of $\mathcal{P}.1$ would perform no shrinkage of $\boldsymbol{w}_1$ to get $\boldsymbol{w}_2$. Formally, assume for the sake of contradiction that $\theta_2 > \theta_1$. Under this assumption, we would have that $\boldsymbol{w}_2 = \min\{\min\{\boldsymbol{w}_0, \theta_1\}, \theta_2\} = \min\{\boldsymbol{w}_0, \theta_1\} = \boldsymbol{w}_1$. In turn, we obtain that $\boldsymbol{0}$ belongs to the subgradient set of Eq. (25) when evaluated at $\boldsymbol{w} = \boldsymbol{w}_1$, thus,

$$\boldsymbol{0} \in \boldsymbol{w}_1 - \boldsymbol{w}_1 + \lambda_2 \partial \|\boldsymbol{w}_1\|_\infty = \lambda_2 \partial \|\boldsymbol{w}_1\|_\infty \ .$$

Clearly, the set $\partial \|\boldsymbol{w}_1\|_\infty$ can contain $\boldsymbol{0}$ only when $\boldsymbol{w}_1 = \boldsymbol{0}$. Since we assumed that $\lambda_1 < \|\boldsymbol{w}_0\|_1$, and hence that $\boldsymbol{\alpha}_1 \preceq \boldsymbol{w}_0$ and $\boldsymbol{\alpha}_1 \neq \boldsymbol{w}_0$, we have that $\boldsymbol{w}_1 = \boldsymbol{w}_0 - \boldsymbol{\alpha}_1 \neq \boldsymbol{0}$. This contradiction implies that $\theta_2 \leq \theta_1$.

We now examine the solution vectors to the dual problems of $\mathcal{P}.1$, $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$. We know that $\|\boldsymbol{\alpha}_1\|_1 = \lambda_1$ so that $\|\boldsymbol{w}_0 - \boldsymbol{\alpha}_1\|_1 > \lambda_2$ and hence $\boldsymbol{\alpha}_2$ is at the boundary $\|\boldsymbol{\alpha}_2\|_1 = \lambda_2$ (see again [6]). Furthermore, the sum of the these vectors is

$$\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2 = [\boldsymbol{w}_0 - \theta_1]_+ + \left[\boldsymbol{w}_0 - [\boldsymbol{w}_0 - \theta_1]_+ - \theta_2\right]_+ \ . \tag{27}$$

Let v denote a component of $\boldsymbol{w}_0$ greater than θ_1. For any such component the right hand side of Eq. (27) amounts to

$$[v - (v - \theta_1) - \theta_2]_+ + [v - \theta_1]_+ = [\theta_1 - \theta_2]_+ + v - \theta_1 = v - \theta_1 = [v - \theta_1]_+ \ ,$$

where we used the fact that $\theta_2 \leq \theta_1$ to eliminate the term $[\theta_1 - \theta_2]_+$. Next, let u denote a component of $\boldsymbol{w}_0$ smaller than θ_1. In this case, the right hand side of Eq. (27) amounts to $[u - 0 - \theta_2]_+ + 0 =$

$[u-\theta_2]_+$. Recapping, the end result is that the vector sum $\boldsymbol{\alpha}_1+\boldsymbol{\alpha}_2$ equals $[\boldsymbol{w}_0-\theta_2]_+$. Moreover, $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ are in $\mathbb{R}^n_+$ as we assumed that $\boldsymbol{w}_0 \succeq \mathbf{0}$, and thus

$$\| [\boldsymbol{w}_0-\theta_2]_+ \|_1 = \|\boldsymbol{\alpha}_1+\boldsymbol{\alpha}_2\|_1 = \lambda_1+\lambda_2 \ . \tag{28}$$

We now show that $\mathcal{P}.2$ has the same dual solution as the sequential updates above. The dual of $\mathcal{P}.2$ is

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\boldsymbol{\alpha}-\boldsymbol{w}_0\|_2^2 \ \text{ s.t. } \ \|\boldsymbol{\alpha}\|_1 \leq \lambda_1+\lambda_2.$$

Denoting by $\boldsymbol{\alpha}_0$ the solution of the above dual problem, we have $\boldsymbol{w}^\star = \boldsymbol{w}_0-\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_0 = [\boldsymbol{w}_0-\theta]_+$ for some $\theta \in \mathbb{R}_+$. Examining the norm of $\boldsymbol{\alpha}_0$ we obtain that

$$\|\boldsymbol{\alpha}_0\|_1 = \left\|[\boldsymbol{w}_0-\theta]_+\right\|_1 = \lambda_1+\lambda_2 \tag{29}$$

because we assumed that $\|\boldsymbol{w}_0\|_1 > \lambda_1+\lambda_2$. We can view the terms $\left\|[\boldsymbol{w}_0-\theta_2]_+\right\|_1$ from Eq. (28) and $\left\|[\boldsymbol{w}_0-\theta]_+\right\|_1$ from Eq. (29) as functions of θ_2 and θ, respectively. The functions are strictly decreasing functions of θ and θ_2 over the interval $[0, \|\boldsymbol{w}_0\|_\infty]$. Therefore, they are invertible for $0 < \lambda_1+\lambda_2 < \|\boldsymbol{w}_0\|_1$. Since $\left\|[\boldsymbol{w}_0-\theta]_+\right\|_1 = \left\|[\boldsymbol{w}_0-\theta_2]_+\right\|_1$, we must have $\theta_2=\theta$. Recall that the solution of Eq. (26) is $\boldsymbol{w}^\star = \min\{\boldsymbol{w}_0,\theta\}$, and the solution of the sequential update induced by Eq. (24) and Eq. (25) is $\min\{\min\{\boldsymbol{w}_0,\theta_1\},\theta_2\} = \min\{\boldsymbol{w}_0,\theta_2\}$. The programs $\mathcal{P}.1$ and $\mathcal{P}.2$ therefore result in the same vector $\min\{\boldsymbol{w}_0,\theta_2\} = \min\{\boldsymbol{w}_0,\theta\}$ and their induced updates are equivalent.

We now examine the case when $\|\boldsymbol{w}_0\|_1 \leq \lambda_1+\lambda_2$. If the 1-norm of $\boldsymbol{w}_0$ is also smaller than λ_1, $\|\boldsymbol{w}_0\|_1 \leq \lambda_1$, then the dual solution for the first step of $\mathcal{P}.1$ is $\boldsymbol{\alpha}_1 = \boldsymbol{w}_0$, which makes $\boldsymbol{w}_1 = \boldsymbol{w}_0-\boldsymbol{\alpha}_1 = \mathbf{0}$ and hence $\boldsymbol{w}_2 = \mathbf{0}$. The dual solution for the combined problem is clearly $\boldsymbol{\alpha}_0 = \boldsymbol{w}_0$; again, $\boldsymbol{w}^\star = \boldsymbol{w}_0-\boldsymbol{\alpha}_0 = \mathbf{0}$. We are thus left with the case $\lambda_1 < \|\boldsymbol{w}_0\|_1 \leq \lambda_1+\lambda_2$. We straightforwardly get that the solution to Eq. (26) is $\boldsymbol{w}^\star = \mathbf{0}$. We now prove that the iterated solution obtained by $\mathcal{P}.1$ results in the zero vector as well. First, consider the dual solution $\boldsymbol{\alpha}_1$, which is the minimizer of $\|\boldsymbol{\alpha}-\boldsymbol{w}_0\|^2$ subject to $\|\boldsymbol{\alpha}\|_1 \leq \lambda_1$. Since $\boldsymbol{\alpha}_1 = [\boldsymbol{w}_0-\theta_1]_+$ for some $\theta_1 \geq 0$, we know that each component of $\boldsymbol{\alpha}_1$ is between zero and its corresponding component in $\boldsymbol{w}_0$, therefore, $\|\boldsymbol{w}_0-\boldsymbol{\alpha}_1\|_1 = \|\boldsymbol{w}_0\|_1 - \|\boldsymbol{\alpha}\|_1 = \|\boldsymbol{w}_0\|_1-\lambda_1 \leq \lambda_2$. The dual of the second step of $\mathcal{P}.1$ distills to the minimization $\frac{1}{2}\|\boldsymbol{\alpha}-(\boldsymbol{w}_0-\boldsymbol{\alpha}_1)\|^2$ subject to $\|\boldsymbol{\alpha}\|_1 \leq \lambda_2$. Since we showed that $\|\boldsymbol{w}_0-\boldsymbol{\alpha}\|_1 \leq \lambda_2$, we get $\boldsymbol{\alpha}_2 = \boldsymbol{w}_0-\boldsymbol{\alpha}_1$. This means that $\theta_2 = 0$. Recall that the solution of $\mathcal{P}.1$ is $\min\{\boldsymbol{w}_0,\theta_2\}$, which amounts to the zero vector when $\theta_2=0$. We have thus showed that both optimization problems result in the zero vector. This proves the equaivalence of $\mathcal{P}.1$ and $\mathcal{P}.2$ for $q=\infty$. □