We organize the appendices as follows:

- **Appendix A:** Proof of Theorem 2.1
- **Appendix B:** Proof of Corollary 2.2
- **Appendix C:** Alternative Proof and Generalzed Version of Corollary 2.3
- **Appendix D:** Proof of Theorem 3.3
- **Appendix E:** Derivation of Dual Problem for ℓ_1-Regularized Loss Minimization
- **Appendix F:** Examples of Computing $\operatorname{dist}(h_j, \mathbf{y})$
- **Appendix G:** Remarks on Computing α

A. Proof of Theorem 2.1

Theorem 2.1 (Convergence Progress at Iteration t)**.** *Let Δ_t and Δ_{t+1} be the optimality gaps after iterations t and $t+1$ of Algorithm 2. Then for all $t \geq 1$ if the algorithm does not converge at iteration $t+1$, we have*

$$\Delta_{t+1} \leq \Delta_t - \left(\tfrac{\gamma}{2}\tau_t^2\Delta_t^2\right)^{1/3} . \tag{6}$$

Proof. Note: throughout this proof, we use α to refer to α_{t+1} in order to simplify notation.

When $\alpha = 1$, we have

$$\begin{aligned}\Delta_{t+1} &= f(\mathbf{y}_{t+1}) - f(\mathbf{x}_{t+1}) &(24)\\ &= f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) &(25)\\ &\leq 0\,. &(26)\end{aligned}$$

This is because $\mathcal{C}_{t+1}$ includes all constraints active at $\mathbf{x}_t$, ensuring $f(\mathbf{x}_{t+1}) \geq f(\mathbf{x}_t)$. Thus, when $\alpha = 1$, the algorithm converges at iteration $t+1$, and the theorem holds.

To consider the case $\alpha < 1$, we begin by writing

$$\begin{aligned}\Delta_{t+1} &= f(\mathbf{y}_{t+1}) - f(\mathbf{x}_{t+1}) &(27)\\ &= \left[f(\mathbf{y}_{t+1}) - f(\mathbf{x}_t)\right] + \left[f(\mathbf{x}_t) - f(\mathbf{x}_{t+1})\right] . &(28)\end{aligned}$$

Our approach is to bound these terms as functions of Δ_t, τ_t, and α. We will eliminate α from this result by bounding over all $\alpha \in [0, 1]$.

Bounding First Term in (28): Because f is strongly convex with parameter γ, we can write

$$\begin{aligned}f(\mathbf{y}_{t+1}) &= f(\alpha\mathbf{x}_t + (1-\alpha)\mathbf{y}_t) &(29)\\ &\leq \alpha f(\mathbf{x}_t) + (1-\alpha)f(\mathbf{y}_t) - \alpha(1-\alpha)\tfrac{\gamma}{2}\left\|\mathbf{x}_t - \mathbf{y}_t\right\|_2^2 . &(30)\end{aligned}$$

This implies

$$\begin{aligned}f(\mathbf{y}_{t+1}) - f(\mathbf{x}_t) &\leq (1-\alpha)\left[f(\mathbf{y}_t) - f(\mathbf{x}_t)\right] - \alpha(1-\alpha)\tfrac{\gamma}{2}\left\|\mathbf{x}_t - \mathbf{y}_t\right\|_2^2 &(31)\\ &= (1-\alpha)\Delta_t - \alpha(1-\alpha)\tfrac{\gamma}{2}\left\|\mathbf{x}_t - \mathbf{y}_t\right\|_2^2 . &(32)\end{aligned}$$

Furthermore, since $\mathbf{y}_{t+1} = \alpha\mathbf{x}_t + (1-\alpha)\mathbf{y}_t$, we have

$$\begin{aligned}\alpha(1-\alpha)\tfrac{\gamma}{2}\left\|\mathbf{x}_t - \mathbf{y}_t\right\|_2^2 &= \alpha(1-\alpha)\tfrac{\gamma}{2}\left[\frac{\left\|\mathbf{y}_{t+1} - \mathbf{y}_t\right\|_2^2}{\alpha^2}\right] &(33)\\ &\geq \frac{(1-\alpha)}{\alpha}\tfrac{\gamma}{2}\tau_t^2\,. &(34)\end{aligned}$$

Above, the inequality is true because $\alpha < 1$ implies $\|\mathbf{y}_{t+1} - \mathbf{y}_t\|_2 \geq \tau_t$; there exists an $h_j \notin \mathcal{C}_t$ such that $h_j(\mathbf{y}_{t+1}) = 0$, but since $h_j \notin \mathcal{C}_t$, no point on the boundary of h_j—$\mathbf{y}_{t+1}$ included—may be within a radius τ_t of $\mathbf{y}_t$.

Combining (32) and (34), we have

$$f(\mathbf{y}_{t+1}) - f(\mathbf{x}_t) \leq (1-\alpha)\Delta_t - \frac{(1-\alpha)}{\alpha}\tfrac{\gamma}{2}\tau_t^2 \,. \tag{35}$$

Bounding the Second Term in (28): To bound the second term for the case that $\alpha < 1$, let h_j be the (possibly non-unique) constraint such that $h_j(\mathbf{y}_{t+1}) = 0$ and $h_j(\mathbf{x}_t) > 0$, and recall the definition

$$\operatorname{dist}(h_j, \mathbf{x}_t) = \inf_{\mathbf{z}\,:\,h_j(\mathbf{z})=0} \|\mathbf{z} - \mathbf{x}_t\|_2 \,. \tag{36}$$

Because $h(\mathbf{y}_{t+1}) = 0$, the set $\{\mathbf{z} \,:\, h_j(\mathbf{z}) = 0\}$ is non-empty, and we can define $\mathbf{z}_t$ as a value of $\mathbf{z}$ that minimizes $\|\mathbf{z} - \mathbf{x}_t\|_2$ over this set. We have

$$\begin{aligned}
\operatorname{dist}(h_j, \mathbf{x}_t) &= \|\mathbf{z}_t - \mathbf{x}_t\|_2 && (37)\\
&= \left\|\mathbf{z}_t - \tfrac{1}{\alpha}\left(\mathbf{y}_{t+1} - (1-\alpha)\mathbf{y}_t\right)\right\|_2 && (38)\\
&= \frac{1-\alpha}{\alpha}\left\|\tfrac{-\alpha}{1-\alpha}\mathbf{z}_t + \tfrac{1}{1-\alpha}\mathbf{y}_{t+1} - \mathbf{y}_t\right\|_2 && (39)\\
&\geq \frac{1-\alpha}{\alpha}\tau_t \,. && (40)
\end{aligned}$$

The last step is due to the convexity of h_j and the fact that $h_j \notin \mathcal{C}_t$ (otherwise we could not have $h_j(\mathbf{x}_t) > 0$ since $\mathbf{x}_t$ is feasible for all constraints in $\mathcal{C}_t$). Applying convexity of h_j, we have

$$h_j\left(\tfrac{-\alpha}{1-\alpha}\mathbf{z}_t + \tfrac{1}{1-\alpha}\mathbf{y}_{t+1}\right) \geq 0 \,, \tag{41}$$

since $h_j(\mathbf{z}_t) = h_j(\mathbf{y}_{t+1}) = 0$, and $\frac{-\alpha}{1-\alpha} + \frac{1}{1-\alpha} = 1$ with the first term being negative. The fact that this affine combination of $\mathbf{z}_t$ and $\mathbf{y}_{t+1}$ violates (or is tight at) h_j while $\mathbf{y}_t$ is feasible for h_j implies (40) since $\mathbf{y}_t$ is at least a distance τ_t from the boundary $\{\mathbf{z} \,:\, h_j(\mathbf{z}) = 0\}$ (since $h_j \notin \mathcal{C}_t$).

We can use our bound on $\operatorname{dist}(h_j, \mathbf{x}_t)$ to bound $f(\mathbf{x}_t) - f(\mathbf{x}_{t+1})$.

$$\begin{aligned}
f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) &\leq -(\mathbf{x}_{t+1} - \mathbf{x}_t)^T \nabla f(\mathbf{x}_t) - \tfrac{\gamma}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 && (42)\\
&\leq -\tfrac{\gamma}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 && (43)\\
&\leq -\tfrac{\gamma}{2}\operatorname{dist}(h_j, \mathbf{x}_t)^2 && (44)\\
&\leq -\tfrac{\gamma}{2}\frac{(1-\alpha)^2}{\alpha^2}\tau_t^2 \,. && (45)
\end{aligned}$$

Above the first inequality results from strong convexity. The second inequality requires an optimality conditions argument. In particular $\mathbf{x}_t$ minimizes f subject to constraints $\{h_j \,:\, h_j(\mathbf{x}_t) = 0\}$, while $\mathbf{x}_{t+1}$ minimizes f subject to a superset of these constraints. This means $\mathbf{x}_{t+1}$ is feasible for the first problem and $(\mathbf{x}_{t+1} - \mathbf{x}_t)^T \nabla f(\mathbf{x}_t) \geq 0$. Finally, the third inequality results from the fact that $\mathbf{x}_{t+1}$ cannot violate h_j.

Completing the Proof: Adding (35) and (45), we have

$$\begin{aligned}
\Delta_{t+1} &\leq (1-\alpha)\,\Delta_t - \frac{(1-\alpha)}{\alpha}\tfrac{\gamma}{2}\tau_t^2 - \frac{(1-\alpha)^2}{\alpha^2}\tfrac{\gamma}{2}\tau_t^2 && (46)\\
&= (1-\alpha)\,\Delta_t - \frac{1-\alpha}{\alpha^2}\tfrac{\gamma}{2}\tau_t^2 && (47)\\
&= (1-\alpha)\left(\Delta_t - \frac{1}{\alpha^2}\tfrac{\gamma}{2}\tau_t^2\right) \,. && (48)
\end{aligned}$$

It is worth noting that this partial result formalizes the main intuition for BLITZ. When α is close to 1, $\mathbf{y}_t$ becomes close to $\mathbf{x}_{t-1}$ and the resulting suboptimality gap becomes small (via the left part of (48)). At the same time, if α is close to 0, there must exist an h_j that is substantially violated by $\mathbf{x}_{t-1}$. As a result, $f(\mathbf{x}_t)$ improves significantly from $f(\mathbf{x}_{t-1})$ and the resulting suboptimality gap again becomes small (this time via the right side of (48)).

We complete our proof by bounding (48) over all $\alpha \in [0, 1]$. A relatively simple bound is the following:

$$\Delta_{t+1} \leq \Delta_t - \left(\tfrac{\gamma}{2}\tau_t^2 \Delta_t^2\right)^{1/3} . \tag{49}$$

This can be obtained by solving for α in

$$\alpha \Delta_t = \frac{1}{\alpha^2} \tfrac{\gamma}{2} \tau_t^2 \tag{50}$$

and then writing $\Delta_{t+1} \leq (1 - \alpha')\Delta_t$ where α' is the solution from above.

□

B. Proof of Corollary 2.2

Corollary 2.2 (Linear Convergence)**.** *For $t \geq 1$, define*

$$\Delta'_t = f(\mathbf{y}_t) - f(\mathbf{x}_{t-1}) \,, \tag{7}$$

and suppose we run Algorithm 2 choosing τ_t as

$$\tau_t = \sqrt{\tfrac{2}{\gamma}(1 - r)^3 \Delta'_t} \tag{8}$$

for some $r \in [0, 1)$. Then for $t \geq 1$, we have

$$f(\mathbf{y}_t) - f(\mathbf{x}^\star) \leq r^{t-1} \Delta_0 \,. \tag{9}$$

Proof. The proof is a direct application of Theorem 2.1. However, since Δ_t is not known when selecting τ_t, we instead use Δ'_t to upper-bound Δ_t. (To see that Δ'_t upper-bounds Δ_t, note that since all constraints that are tight at $\mathbf{x}_{t-1}$ are included in the working set at iteration t, we have $f(\mathbf{x}_t) \geq f(\mathbf{x}_{t-1})$. Plugging into the definitions of Δ_t and Δ'_t, we have $\Delta'_t \geq \Delta_t$.)

Applying Theorem 2.1 while choosing τ_t as in (8), we have

$$\begin{aligned}
\Delta_t &\leq \Delta_{t-1} - \left(\tfrac{1}{2}\gamma \tau_{t-1}^2 \Delta_{t-1}^2\right)^{1/3} && (51)\\
&= \Delta_{t-1} - \left((1 - r)^3 \Delta'_{t-1} \Delta_{t-1}^2\right)^{1/3} && (52)\\
&\leq \Delta_{t-1} - (1 - r)\Delta_{t-1} && (53)\\
&= r\Delta_{t-1} \ . && (54)
\end{aligned}$$

This completes our proof. □

C. Alternative Proof and Generalized Version of Corollary 2.3

Corollary 2.3 immediately follows from Theorem 2.1. In this appendix, we present a simpler alternative proof. Furthermore, this proof leads to a more general constraint elimination rule. In particular, while Corollary 2.3 is assumed to be used with the BLITZ algorithm (and subproblems are assumed to be solved exactly), the more general rule can be applied with *any* feasible point $\mathbf{y}$ and suboptimality gap Δ.

Recall Corollary 2.3:

Corollary 2.3 (Constraint Elimination)**.** *For $t \geq 1$, define Δ'_t as in (7). If*

$$\operatorname{dist}(h_j, \mathbf{y}_t) > \sqrt{\tfrac{2}{\gamma}\Delta'_t} \,, \tag{10}$$

then $h_j(\mathbf{x}^\star) < 0$, and h_j may be eliminated from (P1).

Here we prove the following:

Theorem C.1 (FLEX Constraint Elimination). *For (P1), let* $\mathbf{y}$ *be any feasible point and let* Δ *be a suboptimality gap such that* $f(\mathbf{y}) - f(\mathbf{x}^\star) \leq \Delta$. *If*

$$\operatorname{dist}(h_j, \mathbf{y}) > \sqrt{\tfrac{2}{\gamma}\Delta}\,, \tag{55}$$

then $h_j(\mathbf{x}^\star) < 0$*, and* h_j *may be eliminated from (P1).*

Proof. By optimality conditions of (P1), we know

$$\langle \nabla f(\mathbf{x}^\star), \mathbf{y} - \mathbf{x}^\star \rangle \geq 0\,. \tag{56}$$

By strong convexity of f, we have

$$\begin{aligned} f(\mathbf{y}) &\geq f(\mathbf{x}^\star) + \langle \nabla f(\mathbf{x}^\star), \mathbf{y} - \mathbf{x}^\star \rangle + \tfrac{\gamma}{2} \left\| \mathbf{y} - \mathbf{x}^\star \right\|_2^2 && (57) \\ &\geq f(\mathbf{x}^\star) + \tfrac{\gamma}{2} \left\| \mathbf{y} - \mathbf{x}^\star \right\|_2^2 \,. && (58) \end{aligned}$$

Assume $\operatorname{dist}(h_j, \mathbf{y}) > \sqrt{\tfrac{2}{\gamma}\Delta}$. This implies

$$\begin{aligned} \left\| \mathbf{y} - \mathbf{x}^\star \right\|_2^2 &\leq \tfrac{2}{\gamma} \left[f(\mathbf{y}) - f(\mathbf{x}^\star) \right] && (59) \\ &\leq \tfrac{2}{\gamma}\Delta && (60) \\ &< \operatorname{dist}(h_j, \mathbf{y})^2 \,. && (61) \end{aligned}$$

We have shown $\operatorname{dist}(h_j, \mathbf{y}) > \left\| \mathbf{y} - \mathbf{x}^\star \right\|_2$. By definition of $\operatorname{dist}(h_j, \mathbf{y})$, we must have $h_j(\mathbf{x}^\star) < 0$. Therefore, h_j is not active at the solution. □

We note that in our experience, such screening/constraint elimination rules are rather conservative in general. This means that for many problems, few constraints are eliminated unless the problem is somehow easy to begin with (in our case, if the feasible point $\mathbf{y}$ is already close to the solution $\mathbf{x}^\star$; in the ℓ_1-regularized learning case, screening rules perform best when the regularization λ is large).

As a result, for hard problems, we find it much more efficient to be aggressive eliminating constraints and then periodically reconsider constraints later. When reconsidering constraint h_j in BLITZ, we compute $\operatorname{dist}(h_j, \mathbf{y})$ to determine whether h_j should be added to $\mathcal{C}$. With $\operatorname{dist}(h_j, \mathbf{y})$ already computed, applying Theorem C.1 requires negligible additional computation.

D. Proof of Theorem 3.3

Theorem 3.3 (Progress for ℓ_1 with Approximate Solver). *For (P5), define* Δ_t *as in (16), and assume* $\mathbf{x}_t$ *and* $\mathbf{w}_t$ *satisfy (17). If* $\alpha_{t+1} = 1$*, assume* $g(\mathbf{w}_{t+1}) \geq g(\mathbf{w}_t)$*. If* $\alpha_{t+1} < 1$*, let* h_j *be the (possibly non-unique) constraint such that* $h_j(\mathbf{x}_t) > 0$ *and* $h_j(\mathbf{y}_{t+1}) = 0$ *and assume* $g(\mathbf{w}_{t+1}) \geq \max_\delta g(\mathbf{w}_t + \delta \mathbf{e}_j)$*. Then for* $t \geq 1$*, we have*

$$\Delta_{t+1} \leq \max\left\{ \Delta_t - \left(\tfrac{1}{2L}(1-\epsilon_t)^2 \tau_t^2 \Delta_t^2 \right)^{1/3}, \epsilon_t \Delta_t \right\} \,. \tag{18}$$

This proof is similar to the proof of Theorem 2.1. The main addition is the incorporation of partial subproblem solutions. The relation between $\mathbf{x}_t$ and $\mathbf{w}_t$ as defined in (15) is important, and for this reason, our proof applies only to the ℓ_1-regularized loss minimization problem and not the general setting of Theorem 2.1.

Like in the proof of Theorem 2.1, we use α to refer to α_{t+1}. Note that when $\alpha = 1$, we have

$$\begin{aligned} \Delta_{t+1} &= f(\mathbf{y}_{t+1}) - g(\mathbf{w}_{t+1}) && (62) \\ &= f(\mathbf{x}_t) - g(\mathbf{w}_{t+1}) && (63) \\ &\leq f(\mathbf{x}_t) - g(\mathbf{w}_t) && (64) \\ &\leq \epsilon_t (f(\mathbf{x}_t) - f(\mathbf{y}_t)) && (65) \\ &= \epsilon_t \Delta_t \,. && (66) \end{aligned}$$

Thus, when $\alpha = 1$, the theorem holds. For the remainder of the proof, we consider the case $\alpha < 1$. We write

$$\begin{aligned}
\Delta_{t+1} &= f(\mathbf{y}_{t+1}) - g(\mathbf{w}_{t+1}) && (67)\\
&= f((1-\alpha)\mathbf{y}_t + \alpha\mathbf{x}_t) - g(\mathbf{w}_{t+1}) && (68)\\
&\leq (1-\alpha)f(\mathbf{y}_t) + \alpha f(\mathbf{x}_t) - \tfrac{1}{2L}\alpha(1-\alpha)\left\|\mathbf{x}_t - \mathbf{y}_t\right\|_2^2 && (69)\\
&= (1-\alpha)\left[f(\mathbf{y}_t) - g(\mathbf{w}_t)\right] + \alpha\left[f(\mathbf{x}_t) - g(\mathbf{w}_t)\right] + \left[g(\mathbf{w}_t) - g(\mathbf{w}_{t+1})\right] - \tfrac{1}{2L}\alpha(1-\alpha)\left\|\mathbf{x}_t - \mathbf{y}_t\right\|_2^2 && (70)\\
&\leq (1-\alpha)\Delta_t + \alpha\epsilon_t\Delta_t + \left[g(\mathbf{w}_t) - g(\mathbf{w}_{t+1})\right] - \tfrac{1}{2L}\alpha(1-\alpha)\left\|\mathbf{x}_t - \mathbf{y}_t\right\|_2^2 && (71)\\
&= (1-\alpha(1-\epsilon_t))\Delta_t - \tfrac{1}{2L}\alpha(1-\alpha)\left\|\mathbf{x}_t - \mathbf{y}_t\right\|_2^2 + \left[g(\mathbf{w}_t) - g(\mathbf{w}_{t+1})\right]\,. && (72)
\end{aligned}$$

The remaining steps of the proof bound the second and third terms of (72) as functions of α and τ_t. We then achieve the final result by bounding over all $\alpha \in [0, 1]$.

For the second term of (72), we have

$$\begin{aligned}
\tfrac{1}{2L}\alpha(1-\alpha)\left\|\mathbf{x}_t - \mathbf{y}_t\right\|_2^2 &= \tfrac{1}{2L}\alpha(1-\alpha)\left[\frac{\left\|\mathbf{y}_{t+1} - \mathbf{y}_t\right\|_2^2}{\alpha^2}\right] && (73)\\
&\geq \frac{1-\alpha}{\alpha}\tfrac{1}{2L}\tau_t^2\,. && (74)
\end{aligned}$$

The inequality above results from the definition of α, the condition $\alpha < 1$, and the definition of τ_t ($\left\|\mathbf{y}_{t+1} - \mathbf{y}\right\|_2$ must be at least τ_t, otherwise α must be 1).

Now let us consider the third term of (72). Recall that $\mathbf{x}_t = \xi_t \cdot p(\mathbf{A}\mathbf{w}_t, \mathbf{b})$, where p maps dual variables $\mathbf{w}_t$ to the primal variables $\mathbf{x}_t$ and $\xi_t \in [0, 1]$ scales this result toward $\mathbf{0}$ so that $\mathbf{x}_t$ satisfies all constraints in $\mathcal{C}_t$. Since $\alpha < 1$, there must be an h_j such that $h_j(\mathbf{x}_t) > 0$, $h_j(\mathbf{y}_{t+1}) = 0$, and $h_j(\mathbf{x}_{t+1}) \leq 0$. For this h_j, we have

$$\begin{aligned}
& h_j(\mathbf{y}_{t+1}) = 0 && (75)\\
\Rightarrow\quad & \left|\mathbf{A}_j^T\mathbf{y}_{t+1}\right| - \lambda = 0 && (76)\\
\Rightarrow\quad & \left|\mathbf{A}_j^T\left[\alpha\mathbf{x}_t + (1-\alpha)\mathbf{y}_t\right]\right| - \lambda = 0 && (77)\\
\Rightarrow\quad & \alpha\left|\mathbf{A}_j^T\mathbf{x}_t\right| + (1-\alpha)\left|\mathbf{A}_j^T\mathbf{y}_t\right| - \lambda \geq 0 && (78)\\
\Rightarrow\quad & \left|\mathbf{A}_j^T\mathbf{x}_t\right| - \lambda \geq \frac{(1-\alpha)}{\alpha}\left(\lambda - \left|\mathbf{A}_j^T\mathbf{y}_t\right|\right) && (79)\\
\Rightarrow\quad & \frac{\left|\mathbf{A}_j^T\mathbf{x}_t\right| - \lambda}{\left\|\mathbf{A}_j\right\|_2} \geq \frac{(1-\alpha)}{\alpha}\frac{\lambda - \left|\mathbf{A}_j^T\mathbf{y}_t\right|}{\left\|\mathbf{A}_j\right\|_2} && (80)\\
\Rightarrow\quad & \frac{\left|\mathbf{A}_j^T\mathbf{x}_t\right| - \lambda}{\left\|\mathbf{A}_j\right\|_2} \geq \frac{(1-\alpha)}{\alpha}\tau_t\,. && (81)
\end{aligned}$$

Above we have used the fact that $\operatorname{dist}(h_j, \mathbf{y}_t) = \frac{\lambda - \left|\mathbf{A}_j^T\mathbf{y}_t\right|}{\left\|\mathbf{A}_j\right\|_2} \geq \tau_t$. Otherwise, h_j would have been included in $\mathcal{C}_t$, making $h_j(\mathbf{x}_t) \leq 0$. Since $\xi_t \in [0, 1]$ we have

$$\frac{\left|\mathbf{A}_j^T p(\mathbf{A}\mathbf{w}_t, \mathbf{b})\right| - \lambda}{\left\|\mathbf{A}_j\right\|_2} \geq \frac{(1-\alpha)}{\alpha}\tau_t\,. \tag{82}$$

However, $\mathbf{A}_j^T p(\mathbf{A}\mathbf{w}_t, \mathbf{b})$ is also the derivative of the loss $\sum_i \phi_i(\mathbf{a}_i^T\mathbf{w})$ with respect to w_j. Using standard coordinate descent analysis, if we consider minimizing $-g(\mathbf{w})$ with an update at coordinate j, we have

$$g(\mathbf{w}_t) - g(\mathbf{w}_{t+1}) \leq \min_{\delta_j} g(\mathbf{w}_t) - g(\mathbf{w}_t + \mathbf{e}_j \delta) \tag{83}$$

$$\leq \min_{\delta} \tfrac{L}{2} \|\mathbf{A}_j\|_2^2 \delta^2 + [\mathbf{A}_j^T p(\mathbf{A}\mathbf{w}_t, \mathbf{b})]\delta + \lambda |\delta| \tag{84}$$

$$\leq -\tfrac{1}{2L} \left[\frac{\lambda - \left|\mathbf{A}_j^T p(\mathbf{A}\mathbf{w}_t, \mathbf{b})\right|}{\|\mathbf{A}_j\|_2} \right]^2 \tag{85}$$

$$\leq -\tfrac{1}{2L} \frac{(1-\alpha)^2}{\alpha^2} \tau_t^2 \,. \tag{86}$$

Above, the second inequality comes from our assumption that ϕ_i is smooth. Combining (86) and (74) with (72), we have

$$\Delta_{t+1} \leq (1 - \alpha(1-\epsilon_t))\, \Delta_t - \frac{(1-\alpha)}{\alpha} \tfrac{1}{2L} \tau_t^2 - \frac{(1-\alpha)^2}{\alpha^2} \tfrac{1}{2L} \tau_t^2 \tag{87}$$

$$= (1 - \alpha(1-\epsilon_t))\, \Delta_t - \frac{1-\alpha}{\alpha^2} \tfrac{1}{2L} \tau_t^2 \tag{88}$$

$$= \epsilon_t \Delta_t + (1-\alpha) \left((1-\epsilon_t)\Delta_t - \frac{1}{\alpha^2} \tfrac{1}{2L} \tau_t^2 \right) . \tag{89}$$

We complete our proof by bounding over all $\alpha \in [0, 1]$. A relatively simple bound is the following:

$$\Delta_{t+1} \leq \epsilon_t \Delta_t + \max_{\alpha \in [0,1]} \min \left\{ (1-\alpha)(1-\epsilon_t)\Delta_t, \left((1-\epsilon_t)\Delta_t - \frac{1}{\alpha^2} \tfrac{1}{2L} \tau_t^2 \right)_+ \right\} \tag{90}$$

$$= \max \left\{ \epsilon_t \Delta_t, \Delta_t - \left(\tfrac{1}{2L} (1-\epsilon_t)^2 \tau_t^2 \Delta_t^2 \right)^{1/3} \right\} . \tag{91}$$

E. Derivation of Dual Problem for ℓ_1-Regularized Loss Minimization

In this appendix, we derive the dual of the ℓ_1-regularized learning problem from Section 3.

$$\min_{\mathbf{w}} \sum_{i=1}^n \phi_i(\mathbf{a}_i^T \mathbf{w}) + \lambda \|\mathbf{w}\|_1 = \min_{\mathbf{w}} \sum_{i=1}^n \phi_i^{**}(\mathbf{a}_i^T \mathbf{w}) + \lambda \|\mathbf{w}\|_1 \tag{92}$$

$$= \min_{\mathbf{w}} \sum_{i=1}^n \max_{x_i} \left[(\mathbf{a}_i^T \mathbf{w}) x_i - \phi_i^*(x_i) \right] + \lambda \|\mathbf{w}\|_1 \tag{93}$$

$$= \min_{\mathbf{w}} \max_{\mathbf{x}} - \sum_{i=1}^n \phi_i^*(x_i) + \langle \mathbf{A}\mathbf{w}, \mathbf{x} \rangle + \lambda \|\mathbf{w}\|_1 \tag{94}$$

$$= \max_{\mathbf{x}} \min_{\mathbf{w}} - \sum_{i=1}^n \phi_i^*(x_i) + \langle \mathbf{A}\mathbf{w}, \mathbf{x} \rangle + \lambda \|\mathbf{w}\|_1 \tag{95}$$

$$= \max_{\mathbf{x}} - \sum_{i=1}^n \phi_i^*(x_i) + \min_{\mathbf{w}} \langle \mathbf{A}\mathbf{w}, \mathbf{x} \rangle + \lambda \|\mathbf{w}\|_1 \tag{96}$$

$$= \max_{\mathbf{x} \,:\, \|\mathbf{A}^T \mathbf{x}\|_\infty \leq \lambda} \sum_{i=1}^n -\phi_i^*(x_i) \,. \tag{97}$$

Note that ϕ_i^* refers to the conjugate function of ϕ_i:

$$\phi_i^*(x_i) = \max_v \langle v, x_i \rangle - f(v) \,. \tag{98}$$

We now derive this function for squared and logistic loss.

E.1. Conjugate Function for Squared Loss

$$\max_v \langle v, x_i\rangle - \tfrac{1}{2}(v-b_i)^2 = -\tfrac{1}{2}b_i^2 + \max_v (x_i+b_i)v - \tfrac{1}{2}v^2 \tag{99}$$

$$= -\tfrac{1}{2}b_i^2 + \tfrac{1}{2}(b_i+x_i)^2 \tag{100}$$

by setting $v = x_i + b_i$.

E.2. Conjugate Function for Logistic Loss

We are looking to solve

$$\max_v \ \langle v, x_i\rangle - \log(1+\exp(-b_i v)) \,. \tag{101}$$

Differentiating, we have

$$x_i + \frac{b_i \exp(-b_i v)}{1+\exp(-b_i v)} = 0 \tag{102}$$

$$\Rightarrow x_i = \frac{-b_i \exp(-b_i v)}{1+\exp(-b_i v)} \tag{103}$$

$$\Rightarrow v = -\frac{1}{b_i}\log\left(\frac{-x_i}{x_i+b_i}\right) \,. \tag{104}$$

We can substitute this into (101) to obtain

$$\phi^*(x_i) = -\frac{x_i}{b_i}\log\left(\frac{-x_i}{x_i+b_i}\right) - \log\left(1-\frac{x_i}{x_i+b_i}\right) \tag{105}$$

$$= -\frac{x_i}{b_i}\log\left(-\frac{x_i}{b_i}\right) + \frac{x_i}{b_i}\log\left(1+\frac{x_i}{b_i}\right) - \log\left(1-\frac{x_i}{x_i+b_i}\right) \tag{106}$$

$$= -\frac{x_i}{b_i}\log\left(-\frac{x_i}{b_i}\right) + \left(1+\frac{x_i}{b_i}\right)\log\left(1+\frac{x_i}{b_i}\right) \,. \tag{107}$$

F. Examples of Computing $\mathrm{dist}(h_j, \mathbf{y})$

In this appendix, we briefly include examples for evaluating $\mathrm{dist}(h_j, \mathbf{y})$.

F.1. Linear Constraints

The most common scenario is that h_j is linear. For some vector $\mathbf{a}$ and scaler b, let

$$h_j(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + b \,. \tag{108}$$

In this case,

$$\mathrm{dist}(h_j, \mathbf{y}) = \inf_{\mathbf{z}\,:\,h_j(\mathbf{z})=0} \|\mathbf{z}-\mathbf{y}\|_2 \tag{109}$$

$$= \|(\mathbf{y}+\mu\mathbf{a}) - \mathbf{y}\|_2 \tag{110}$$

$$= |\mu|\,\|\mathbf{a}\|_2 \,, \tag{111}$$

where the scaler μ is such that

$$h_j(\mathbf{y}+\mu\mathbf{a}) = \langle \mathbf{a}, \mathbf{y}\rangle + \mu\|\mathbf{a}\|_2^2 + b = 0 \,. \tag{112}$$

This leaves us with

$$\mathrm{dist}(h_j, \mathbf{y}) = \frac{|\langle \mathbf{a}, \mathbf{y}\rangle + b|}{\|\mathbf{a}\|_2} \,. \tag{113}$$

F.2. Constraints for ℓ_1-Regularized Loss Minimization

When $h_j(\mathbf{x}) = \left|\mathbf{A}_j^T \mathbf{x}\right| - \lambda$, the constraint h_j can be viewed as the combination of two linear constraints:

$$h_j^+(\mathbf{x}) = \mathbf{A}_j^T \mathbf{x} - \lambda\,, \quad \text{and} \tag{114}$$

$$h_j^-(\mathbf{x}) = -\mathbf{A}_j^T \mathbf{x} - \lambda\,. \tag{115}$$

In the BLITZ algorithm, the fact $\mathbf{y}$ is feasible implies $\left|\mathbf{A}_j^T \mathbf{x}\right| \leq \lambda$ and we have

$$\operatorname{dist}(h_j, \mathbf{y}) = \frac{\lambda - \left|\mathbf{A}_j^T \mathbf{y}\right|}{\|\mathbf{A}_j\|_2}\,. \tag{116}$$

F.3. Spherical Constraints

$\operatorname{dist}(h_j, \mathbf{y})$ is also easy to compute when $\{\mathbf{x} : h_j(\mathbf{x}) = 0\}$ is a sphere. Specifically, let

$$h_j(\mathbf{x}) = a\,\|\mathbf{x} - \mathbf{b}\|_2^2 - c\,. \tag{117}$$

Assume $a > 0$ and also assume that $c \geq 0$ since $h_j(\mathbf{x}) \leq 0$ could never be satisfied otherwise. The minimizer of $\|\mathbf{z} - \mathbf{y}\|_2$ subject to $h_j(\mathbf{z}) = 0$ is given by

$$\mathbf{z}^\star = \mathbf{b} + \mu(\mathbf{y} - \mathbf{b})\,, \tag{118}$$

where $\mu \geq 1$ is chosen such that $h_j(\mathbf{z}^\star) = 0$. More specifically, we have

$$a\mu^2\,\|\mathbf{y} - \mathbf{b}\|_2^2 - c = 0 \tag{119}$$

$$\Rightarrow \quad \mu = \sqrt{\frac{c}{a}\frac{1}{\|\mathbf{y} - \mathbf{b}\|_2^2}}\,. \tag{120}$$

$$\tag{121}$$

This implies

$$\|\mathbf{z}^\star - \mathbf{y}\|_2 = \|(\mu - 1)(\mathbf{b} - \mathbf{y})\|_2 \tag{122}$$

$$= (\mu - 1)\,\|\mathbf{y} - \mathbf{b}\|_2 \tag{123}$$

$$= \sqrt{\frac{c}{a}} - \|\mathbf{y} - \mathbf{b}\|_2\,. \tag{124}$$

F.4. Smooth Constraints

For arbitrary h_j, evaluating $\operatorname{dist}(h_j, \mathbf{y})$ is potentially difficult. Despite h_j being convex, minimizing $\|\mathbf{z} - \mathbf{y}\|_2$ subject to $h_j(\mathbf{z}) = 0$ is a not a convex problem in general due to the domain $\{\mathbf{z} : h_j(\mathbf{z}) = 0\}$.

The guarantees of BLITZ still hold, however, if we use a lower bound of $\operatorname{dist}(h_j, \mathbf{y})$ when determining the working set. If the gradient of h_j exists and is Lipschitz continuous with constant L, then obtaining a lower bound is straightforward. We can define

$$h_j'(\mathbf{x}) = h_j(\mathbf{y}) + (\mathbf{x} - \mathbf{y})^T \nabla h_j(\mathbf{y}) + \tfrac{L}{2}\,\|\mathbf{x} - \mathbf{y}\|_2^2\,. \tag{125}$$

$h_j'(\mathbf{x})$ upper-bounds $h_j(\mathbf{x})$ for all $\mathbf{x}$. As a result, the set $\{\mathbf{x} : h_j'(\mathbf{x}) \leq 0\}$ is a subset of $\{\mathbf{x} : h_j(\mathbf{x}) \leq 0\}$, and we have

$$\operatorname{dist}(h_j', \mathbf{y}) \leq \operatorname{dist}(h_j, \mathbf{y})\,. \tag{126}$$

Evaluating $\operatorname{dist}(h_j', \mathbf{y})$ is straightforward since $\{\mathbf{x} : h_j'(\mathbf{x}) = 0\}$ is a sphere.

G. Remarks on Computing α

Here we briefly discuss how to compute α. Recall that

$$\alpha = \max\left\{\alpha' \in [0,1] \,:\, \alpha'\mathbf{x} + (1-\alpha')\mathbf{y} \in \mathcal{D}\right\} . \tag{127}$$

That is, α is chosen such that $\mathbf{y} = \alpha\mathbf{x} + (1-\alpha)\mathbf{y}$ is the closest feasible point to $\mathbf{x}$ on the line segment $[\mathbf{y}, \mathbf{x}]$.

One way to find α is to define an α_j for each constraint h_j as

$$\alpha_j = \max\left\{\alpha' \in [0,1] \,:\, h_j(\alpha'\mathbf{x} + (1-\alpha')\mathbf{y}) \leq 0\right\} . \tag{128}$$

Then we simply set

$$\alpha = \min_j \alpha_j . \tag{129}$$

If $h_j(\mathbf{x}) \leq 0$, then clearly $\alpha_j = 1$. Otherwise, for general h_j, evaluating (128) can be accomplished in logarithmic time using the bisection algorithm. For the common case that h_j is linear, α_j can be computed in closed form:

$$h_j(\alpha_j\mathbf{x} + (1-\alpha_j)\mathbf{y}) = 0 \tag{130}$$

$$\Rightarrow \quad \alpha_j h_j(\mathbf{x}) + (1-\alpha_j)h_j(\mathbf{y}) = 0 \tag{131}$$

$$\Rightarrow \quad \alpha_j = \frac{-h_j(\mathbf{y})}{h_j(\mathbf{x}) - h_j(\mathbf{y})} . \tag{132}$$

Note that in BLITZ, $h_j(\mathbf{y}) \leq 0$, and since any constraint for which $h_j(\mathbf{y}) = 0$ is included in $\mathcal{C}$, it is always the case that $\alpha_j > 0$.