A. Proof of Theorem 2.1

**Theorem 2.1** (Safeness of StingyCD). *In Algorithm 2, every skipped update would, if computed, result in $\delta = 0$. That is, if $q^{(t-1)} \leq \tau_i$ and $x_i^{(t-1)} = 0$, then

$$
\max \left\{ -x_i^{(t-1)}, \frac{\langle A_i, b - Ax^{(t-1)} \rangle - \lambda}{\| A_i \|^2} \right\} = 0.
$$

**Proof.** Since $x_i^{(t-1)} = 0$, we need to prove that if $q^{(t-1)} \leq \tau_i$, then

$$
\langle A_i, r^{(t-1)} \rangle - \lambda \leq 0,
$$

where we have used the definition $r^{(t-1)} = b - Ax^{(t-1)}$.

We show by induction that $q^{(t)} = \| rr - r^{(t)} \|^2$. The base case is that $q^{(t-1)} = 0$ whenever StingyCD performs the update $rr \leftarrow r^{(t-1)}$. The inductive step is that

$$
q^{(t)} = q^{(t-1)} - 2\delta \langle A_i, r^{(t-1)} - rr \rangle + \delta^2 \| A_i \|^2,
$$

and

$$
= \| r^{(t-1)} - rr \|^2 - 2\delta \langle A_i, r^{(t-1)} - rr \rangle + \delta^2 \| A_i \|^2.
$$

Recall the definition $\tau_i = \text{sign} (g_i) \frac{g_i^2}{\| A_i \|^2}$, where $g_i = -\langle A_i, rr \rangle + \lambda$. It follows that

$$
q^{(t-1)} \leq \tau_i \implies \| r^{(t-1)} - rr \|^2 \leq \text{sign} (g_i) \frac{g_i^2}{\| A_i \|^2},
$$

and

$$
\| r^{(t-1)} - rr \| \leq \frac{g_i}{\| A_i \|},
$$

$$
\| A_i \| \| r^{(t-1)} - rr \| \leq -\langle A_i, rr \rangle + \lambda,
$$

$$
\| A_i \| \| r^{(t-1)} - rr \| - \lambda \leq 0,
$$

$$
\langle A_i, r^{(t-1)} \rangle - \lambda \leq 0.
$$

\hfill \Box

B. Proof of Theorem 2.2

**Theorem 2.2** (Per iteration time complexity of StingyCD). *Algorithm 2 can be implemented so that iteration $t$ requires

- Less time than an identical iteration of Algorithm 1 if $q^{(t-1)} \leq \tau_i$ and $x_i^{(t-1)} = 0$ (the update is skipped) and $rr$ is not updated. Specifically, StingyCD requires $O(1)$ time, while CD requires $O(\text{NNZ} (A_i))$ time.
- The same amount of time (up to an $O(1)$ term) as a CD iteration if the update is not skipped and $rr$ is not updated. In particular, both algorithms require the same number of $O(\text{NNZ} (A_i))$ operations.
- More time than a CD iteration if $rr$ is updated. In this case, StingyCD requires $O(\text{NNZ} (A_i))$ time.

**Proof.** Note that at each iteration, CD computes a dot product of length $\text{NNZ} (A_i)$ to compute $\delta$. If $\delta \neq 0$, an additional $O(\text{NNZ} (A_i))$ operation is required to update $r^{(t)}$. 

Case 1: the update is skipped and \( \text{rr} \) is not updated In this case, the only computation StingyCD performs during this iteration is (i.) deciding not to update the reference vector, (ii.) choosing a coordinate to update, and (iii.) checking whether \( q^{(t-1)} \leq \tau_i \) and \( x_i^{(t-1)} = 0 \). Steps (i.) and (ii.) can be easily be defined so that they require \( \mathcal{O}(1) \) time, and checking the conditions for (iii.) also requires constant time.

Case 2: the update is not skipped and \( \text{rr} \) is not updated In this case, the only additional operation that we have not already considered is the update to \( q^{(t)} \). This update can be performed in constant time by caching previous computations of \( \langle A_i, \text{rr} \rangle, \langle A_i, r^{(t-1)} \rangle \), and \( \| A_i \|^2 \). The value of \( \langle A_i, \text{rr} \rangle \) was computed when computing the threshold \( \tau_i \), and \( \langle A_i, r^{(t-1)} \rangle \) and \( \| A_i \|^2 \) are necessary to compute \( \delta \).

Case 3: \( \text{rr} \) is updated In this case, computing \( \tau_i \) for all \( i \) requires computing \( \langle A_i, \text{rr} \rangle \) for all columns in \( A \). This is a matrix-vector multiply that requires \( \mathcal{O}(\text{NNZ}(A_i)) \) operations.

C. Proof of Theorem 3.2

Theorem 3.2 (Equation for \( P(U^{(t)}) \)). Assume \( x_i^{(t-1)} = 0 \) and \( \tau_i \in (-q^{(t-1)}, q^{(t-1)}) \). Then Assumption 3.1 implies

\[
P(U^{(t)}) = \begin{cases} 
\frac{1}{2} I_{(1 - \tau_i/q^{(t-1)})}(\frac{n-1}{2}, \frac{1}{2}) & \text{if } \tau_i \geq 0, \\
1 - \frac{1}{2} I_{(1 + \tau_i/q^{(t-1)})}(\frac{n-1}{2}, \frac{1}{2}) & \text{otherwise},
\end{cases}
\]

where \( I_x(a, b) \) is the regularized incomplete beta function.

Proof. Recall the illustration form Figure 2:

Because we assume \( r^{(t-1)} \) is uniformly distributed on the boundary of \( S^{(t)} \), the probability that \( r^{(t-1)} \in A_i \) is given by dividing the area of \( A_i \cap \text{bd}(S^{(t)}) \) by the area of \( \text{bd}(S^{(t)}) \). The region \( A_i \cap \text{bd}(S^{(t)}) \) is a hyperspherical cap. In the case that \( \text{rr} \notin A_i \), we know from (Li, 2011) that the area of \( A_i \cap \text{bd}(S^{(t)}) \) is given by

\[
\frac{1}{2} \text{area}(S^{(t)}) I_{\sin(\theta)}(\frac{n-1}{2}, \frac{1}{2}),
\]

where \( \text{area}(S^{(t)}) \) is the surface area of \( S^{(t)} \) and \( \theta \) is the angle indicated in the diagram.

When \( \tau_i \geq 0 \), note that by definition of \( \tau_i \), we have \( \text{rr} \notin A_i \). It follows then that when \( \tau_i \geq 0 \), we have

\[
P(U_i) = \frac{\frac{1}{2} \text{area}(S^{(t)}) I_{\sin(\theta)}(\frac{n-1}{2}, \frac{1}{2})}{\text{area}(S^{(t)})}
= \frac{1}{2} I_{(1 - \cos(\theta))^2}(\frac{n-1}{2}, \frac{1}{2})
(14)
\]

\[
= \frac{1}{2} I_{(1 + \tau_i/q^{(t-1)})}(\frac{n-1}{2}, \frac{1}{2}).
(15)
\]

In the case that \( \tau_i < 0 \), we have \( \text{rr} \in A_i \), and we can use symmetry to see that

\[
P(U_i) = 1 - \frac{1}{2} I_{(1 + \tau_i/q^{(t-1)})}(\frac{n-1}{2}, \frac{1}{2}).
(17)
\]

\( \square \)
D. Details of estimating $P(U^{(t)})$ in StingyCD+

In §3.1, we defined the probability $P(U^{(t)})$. Assuming $\tau_i \in (-q^{(t-1)}, q^{(t-1)})$, we have

$$
P(U^{(t)}) = \begin{cases} 
\frac{1}{2} I_1(1-\tau_i/q^{(t-1)})(\frac{n-1}{2}, \frac{1}{2}) & \text{if } \tau_i \geq 0, \\
1 - \frac{1}{2} I_1(1+\tau_i/q^{(t-1)})(\frac{n-1}{2}, \frac{1}{2}) & \text{otherwise}, 
\end{cases}
$$

(18)

where $I_a(a, b)$ is the regularized incomplete beta function.

In our implementation of StingyCD+, we compute $P(U^{(t)})$ approximately using a lookup table. First, we make use of the approximation

$$
\frac{1}{2} I_1(1-\tau_i/q^{(t-1)})(\frac{n-1}{2}, \frac{1}{2}) \approx 1 - \Phi \left( \sqrt{\tau_i(n-1)/q^{(t-1)}} \right).
$$

(19)

Above, $\Phi$ is the standard normal CDF.

Using (19) is not strictly necessary. Using (19) leads to a simpler implementation, however, since we no longer need to compute the regularized incomplete beta function. Instead we only need to define a lookup table for the standard normal CDF. We expect this approximation has negligible effect on StingyCD+, since (19) is a very close approximation for moderately large $n$.

Using (19), our StingyCD+ implementation uses a lookup table of 128 values for $1 - \Phi(\sqrt{x})$. Values of $x$ are spaced uniformly between 0 and 32 inclusive, meaning the table stores the values $1 - \Phi(0), 1 - \Phi(\sqrt{0.25}), 1 - \Phi(\sqrt{0.5}), \ldots, 1 - \Phi(\sqrt{32})$.

To estimate $P(U^{(t)})$ during each iteration, StingyCD+ first computes $\tau_i(n-1)/q^{(t-1)}$ and then reads the closest value from the table that results in an upper bound for $P(U^{(t)})$. For example, if $\tau_i(n-1)/q^{(t-1)} = 0.2$, our approximation of $P(U^{(t)})$ is $1 - \Phi(\sqrt{0.25}) = 0.308 \ldots$. If $\tau_i(n-1)/q^{(t-1)} = -0.2$, then our approximation of $P(U^{(t)})$ is $\Phi(\sqrt{0.5}) = 0.760 \ldots$.

E. Proof of Theorem 3.3

**Theorem 3.3 (StingyCD+ converges to a solution of (P)).** In StingyCD+, assume $\xi^{(t)} \leq \text{NNZ} (x^{(t-1)})$ for all $t > 0$. Also, for each $i \in [m]$, assume the largest number of consecutive iterations during which get_next_coordinate() does not return $i$ is bounded as $t \to \infty$. Then

$$
\lim_{t \to \infty} f(x^{(t)}) = f(x^*).
$$

Before proving the theorem, we introduce and prove a few lemmas.

**Lemma E.1.** Given the assumptions of Theorem 3.3, let $M$ be a number larger than the maximum number of consecutive iterations get_next_coordinate() does not return coordinate $i$ for all $i \in [m]$ as $t \to \infty$. Consider any iteration $t > 0$ of StingyCD+ and any $i \in [m]$ such that $x_i^{(t-1)} \neq 0$. Then there exists an iteration $t' \geq t$ during which StingyCD+ computes an update to coordinate $i$. Furthermore, we have $t' \leq t + mM$.

**Proof.** Define $C_i^{(t-1)}$ as the set of coordinates that correspond to nonzero entries in $x_i^{(t-1)}$:

$$
C_i^{(t-1)} = \{ i : x_i^{(t-1)} \neq 0 \}.
$$

(20)

Let $i_{\text{delayed}}$ denote the unique coordinate in $C_i^{(t-1)}$ such that the delay $D_i^{(t)}$ is largest:

$$
i_{\text{delayed}} = \arg\max_{i \in C_i^{(t-1)}} D_i^{(t)}.
$$

(21)

This coordinate is unique because $t_{\text{last}}$ differs for all $i \in C_i^{(t-1)}$—StingyCD+ updates at most one coordinate during each iteration.

We must have $D_i^{(t)} \geq \text{NNZ} (x_i^{(t-1)})$, since the NNZ $(x_i^{(t-1)}) - 1$ coordinates in $C_i^{(t-1)}$ not equal to $i_{\text{delayed}}$ were updated before $i_{\text{delayed}}$ was last updated (otherwise (21) would not hold). Thus, counting these updates, as well as the update to weight $i_{\text{delayed}}$ during iteration $t_{\text{last}}$, we must have $D_i^{(t)} \geq \text{NNZ} (x_i^{(t-1)})$. 


Now let \( k \geq 0 \) be the smallest such \( k \) for which \( \text{get\_next\_coordinate()} \) returns \( i_{\text{delayed}} \) during iteration \( t + k \). Note that \( k < M \). We must have \( D_{i_{\text{delayed}}}^{(t+k)} \geq \text{NNZ} (x^{(t+k-1)}) \), since (i) until an update for coordinate \( i \) is computed, \( D_i^{(t)} \) is non-decreasing with \( t \) for all \( i \), (ii) we have \( D_{i_{\text{delayed}}}^{(t)} \geq \text{NNZ} (x^{(t-1)}) \), and (iii) whenever \( \text{NNZ} (x^{(t)}) = \text{NNZ} (x^{(t-1)}) + 1 \) for \( t' \in \{t, t + 1, \ldots, t + k - 1\} \), we must also have \( D_{i_{\text{delayed}}}^{(t+1)} = D_{i_{\text{delayed}}}^{(t)} + 1 \)—an update to a zero entry of \( x \) increases the delay for all coordinates by 1.

In addition, since \( i_{\text{delayed}} \in C^{(t-1)} \) and \( i_{\text{delayed}} \) has not been updated since before iteration \( t \), we must have \( x_{i_{\text{delayed}}}^{(t+k-1)} \neq 0 \). Thus, by definition of \( P(U^{(t+k)}) \), we must have \( P(U^{(t+k)}) = 1 \). Applying the assumption that \( \xi^{(t+k)} \leq \text{NNZ} (x^{(t+k-1)}) \), it follows that

\[
P(U^{(t+k)}) D_{i_{\text{delayed}}}^{(t+k)} = D_{i_{\text{delayed}}}^{(t+k)} \geq \text{NNZ} (x^{(t+k-1)}) + 1 \geq \xi^{(t+k)}. \tag{22}
\]

Thus, the condition for skipping update \( t + k \) in StingyCD+ is not satisfied. That is, during iteration \( t + k \), StingyCD+ computes an update to coordinate \( i_{\text{delayed}} \). It follows that \( D_{i_{\text{delayed}}}^{(t+k+1)} = 1 \). That is, \( i_{\text{delayed}} \) now corresponds to the weight with smallest delay among nonzero weights.

Now consider any \( i \) such that \( x_i^{(t-1)} \neq 0 \). This coordinate was last updated during iteration \( t_i^{\text{last}} \). It follows that if coordinate \( i \) is not updated by iteration \( t_i^{\text{last}} + (m - 1)M \), then \( i \) corresponds to the weight with largest delay among nonzero weights. This is because we have shown that the nonzero weight with maximum delay is updated within \( M \) iterations, after which it becomes the nonzero weight with smallest delay. Thus, before coordinate \( i \) is updated again, at most \( (m - 1) \) other coordinates correspond to the nonzero weight with largest delay, each of which requires at most \( M \) iterations to update. It follows that after an additional \( M \) iterations—that is, by iteration \( t_i^{\text{last}} + mM \)—coordinate \( i \) must be updated.

**Lemma E.2.** Given the assumptions of Theorem 3.3, then for some set \( F \), StingyCD+ converges to a solution of the problem

\[
\min_{\mathbf{x} \in \mathbb{R}^n} \quad f(\mathbf{x}) := \frac{1}{2} \| \mathbf{A} \mathbf{x} - \mathbf{b} \|^2 + \lambda \langle \mathbf{1}, \mathbf{x} \rangle
\]

s.t. \( \mathbf{x} \geq 0 \)

\( x_i = 0 \ \forall i \in F \). \tag{P’}

**Proof.** First note that \( f(x^{(t)}) \) is nonincreasing with \( t \). This is because whenever \( x^{(t)} \neq x^{(t-1)} \), we can write

\[
x^{(t)} = x^{(t-1)} + \delta \mathbf{e}_i \tag{23}
\]

for some coordinate \( i \), where

\[
\delta = \arg \min_{\delta' : x_i^{(t-1)} + \delta' \geq 0} f(x^{(t-1)} + \delta' \mathbf{e}_i) = \max \left\{ -x_i^{(t-1)}, \frac{\langle \mathbf{A}_i, \mathbf{b} - \mathbf{A} x^{(t-1)} \rangle - \lambda}{\| \mathbf{A}_i \|^2} \right\}. \tag{24}
\]

Second, note that for all \( t \), \( x_i^{(t)} \geq 0 \). From the definition of \( f \), it follows that \( f(x^{(t)}) \geq 0 \) for all \( t \).

Thus, \( f(x^{(t)}) \) is a bounded monotone sequence, which implies that \( \lim_{t \to \infty} f(x^{(t)}) \) exists.

Now let us assume that \( x^{(t)} \) does not converge to a solution of (P’) for some set \( F \). Then there exists a value \( \nu > 0 \) for which the following holds: for all \( t' > 0 \), there exists an iteration \( t > t' \) such that for some \( i \) where \( x_i^{(t-1)} \neq 0 \), we have

\[
|\delta| = \max \left\{ -x_i^{(t-1)}, \frac{\langle \mathbf{A}_i, x^{(t-1)} \rangle - \lambda}{\| \mathbf{A}_i \|^2} \right\} \geq \nu. \tag{25}
\]

In other words, if StingyCD+ updated coordinate \( i \) (corresponding to a nonzero weight) during iteration \( t \), the magnitude of the update would be at least positive value \( \nu \).

Also, note that after any update \( \delta \) to a coordinate \( i \) during iteration \( t \) of StingyCD+, we have (by Taylor expansion)

\[
f(x^{(t)}) - f(x^{(t-1)}) = \left( \lambda - \langle \mathbf{A}_i, x^{(t-1)} \rangle \right) \delta + \frac{1}{2} \| \mathbf{A}_i \|^2 \delta^2 \leq -\frac{1}{2} \| \mathbf{A}_i \|^2 \delta^2. \tag{26}
\]
Now define \( \hat{f} = \lim_{t \to \infty} f(x^{(t)}) \). Consider an iteration \( t' \) such that \( f(x^{(t')}) \leq \hat{f} + \epsilon \), where we define \( \epsilon > 0 \) later.

According to (25), there exists an iteration \( t > t' \) such that for some \( i \) for which \( x_i^{(t-1)} > 0 \), we have

\[
\left| \max \left\{ -x_i^{(t-1)}, \frac{\langle A_i, r^{(t-1)} \rangle - \lambda}{\|A_i\|^2} \right\} \right| \geq \nu.
\]  

(28)

According to Lemma E.1, StingyCD+ will compute at least one update to coordinate \( i \) between iterations \( t \) and \( t + mM \). During each of the iterations between iteration \( t \) and \( t + mM \), suppose that coordinate \( i' \) is updated by an amount \( \delta' \). It must be the case then that

\[
\delta' \leq \sqrt{2\epsilon} \|A_i\|.
\]  

(29)

Otherwise the fact that \( \hat{f} = \lim_{t \to \infty} f(x^{(t)}) \) would be violated due to (27).

Now let \( T \) denote the iteration during which coordinate \( i \) is next updated. From the triangle inequality and (29), it follows that

\[
\| r^{(t-1)} - r^{(T-1)} \| \leq mM \sqrt{2\epsilon}.
\]  

(30)

This implies that

\[
\frac{\langle A_i, r^{(T-1)} \rangle}{\|A_i\|^2} - \frac{\langle A_i, r^{(t-1)} \rangle}{\|A_i\|^2} \in \left[ -mM \sqrt{2\epsilon}, +mM \sqrt{2\epsilon} \right].
\]  

(31)

Now let \( \delta \) be the update to coordinate \( i \) during iteration \( T \). It follows that

\[
|\delta| = \left| \max \left\{ x_i^{(t-1)}, \frac{\langle A_i, r^{(t-1)} \rangle - \lambda}{\|A_i\|^2} \right\} \right|
\]  

(32)

\[
\geq \max \left\{ x_i^{(t-1)}, \frac{\langle A_i, r^{(t-1)} \rangle - \lambda}{\|A_i\|^2} \right\} - \frac{mM \sqrt{2\epsilon}}{\|A_i\|}
\]  

(33)

\[
\geq \nu - \frac{mM \sqrt{2\epsilon}}{\|A_i\|}.
\]  

(34)

Now let us define \( s = \min_{i' : \|A_i'\| > 0} \|A_i'\| \).

\[
\epsilon = \frac{1}{2} \left( \frac{\nu s}{mM} \right)^2
\]  

(35)

Then it follows that

\[
|\delta| > \frac{1}{2} \nu.
\]  

(36)

From (27), it follows that

\[
f(x^{(T)}) \leq f(x^{(T-1)}) - \frac{1}{2} \|A_i\|^2 \delta^2 \leq \hat{f} + \epsilon - \frac{1}{2} s^2 \nu^2 < \hat{f},
\]  

(37)

which violates the assumption that \( \lim_{t \to \infty} f(x^{(t)}) = \hat{f} \).

Thus, StingyCD+ must converge to a solution of \((P')\) for some set \( F \).

\[ \square \]

Proof of Theorem 3.3. Suppose that StingyCD+ does not converge to a solution to \((P)\).

Now define \( \tilde{f} = \lim_{t \to \infty} f(x^{(t)}) \). Also define \( \tilde{r} = \lim_{t \to \infty} r^{(t)} \) and \( \tilde{x} = \lim_{t \to \infty} x^{(t)} \).
Lemma E.2 guarantees that the algorithm at least converges to a solution of $(P')$ for some set $\mathcal{F}$. Using this assumption, if StingyCD+ does not converge to $(P')$’s solution then there exists a $\nu > 0$ such that for some $i$ such that $\tilde{x}_i \neq 0$, we have
\[
\langle A_i, \tilde{x} \rangle - \lambda \geq \nu. \tag{38}
\]

Consider an iteration $t'$ such that $f(x^{(t'-1)}) \leq \hat{f} + \epsilon$, where we define $\epsilon > 0$ later. By Taylor expansion, we have for any $t \geq t'$,
\[
f(x^{(t)}) = f(\bar{x}) + \langle \nabla f(\bar{x}), x^{(t)} - \bar{x} \rangle + \frac{1}{2} \|Ax^{(t)} - \bar{x}\|^2 \tag{39}
\]
\[
\geq \hat{f} + \frac{1}{2} \|\tilde{r} - r^{(t-1)}\|^2. \tag{40}
\]

This implies that for any $t \geq t'$, we have
\[
\|\tilde{r} - r^{(t-1)}\| \leq \sqrt{2\epsilon}. \tag{41}
\]

Define $\epsilon = \min_{i' : \|A_{i'}\| \neq 0} \frac{\nu^2}{8\|A_{i'}\|^2}$. It follows then that for all $t \geq t'$,
\[
\langle A_{i'}, r^{(t-1)} \rangle - \lambda \geq \langle A_{i'}, \tilde{x} \rangle - \|A_{i'}\| \sqrt{2\epsilon} - \lambda \geq \nu - \|A_{i'}\| \sqrt{2\epsilon} \geq \frac{1}{2} \nu. \tag{42}
\]

Also, if we assume $-\langle A_{i'}, r \rangle + \lambda > 0$, we must have
\[
\tau_i = \frac{(\langle A_{i'}, r \rangle + \lambda)^2}{\|A_{i'}\|^2} \tag{43}
\]
\[
\leq \frac{(\langle A_{i'}, r^{(t-1)} \rangle + \lambda + \|A_{i'}\| \|r^{(t-1)} - r\|)^2}{\|A_{i'}\|^2} \tag{44}
\]
\[
\leq (q^{(t-1)} - \frac{1}{2} \nu)^2 \tag{45}
\]
\[
< q^{(t-1)}. \tag{46}
\]

Otherwise, we must have $-\langle A_{i'}, r \rangle + \lambda < 0$, which ensures $\tau_i \leq 0 \leq q^{(t-1)}$. In addition, $q^{(t-1)}$ is bounded as $t \to \infty$ due to (41). As a result, whenever $i$ is returned by get_next_coordinate() during an iteration $t > t'$, then $P(U^{(t)})$ is bounded away from zero. As $t \to \infty$, the delay $D_i^{(t)}$ increases as, at a minimum, nonzero-valued coordinates are updated. Thus, for an eventual iteration $T$, we have
\[
P(U^{(t)}) D_i^{(t)} \geq \xi^{(t)}. \tag{47}
\]

At this point, an update to coordinate $i$ is computed. From (42), it follows that
\[
\delta \geq \frac{1}{2} \frac{\nu}{\|A_{i'}\|^2}. \tag{48}
\]

which ensures that
\[
f(x^{(T)}) \leq f(x^{(T-1)}) - \frac{1}{2} \|A_{i'}\|^2 \delta^2 \tag{49}
\]
\[
\leq f(\bar{x}) + \epsilon - \frac{1}{2} \frac{\nu^2}{\|A_{i'}\|^2} \tag{50}
\]
\[
\leq f(\bar{x}) - \frac{3}{8} \frac{\nu^2}{\|A_{i'}\|^2}. \tag{51}
\]

This contradicts the definition of $\bar{x}$. Thus, our assumption that $x^{(t)}$ does not converge to a solution of $(P)$ is incorrect.
\[
\square
\]
F. Generalizing StingyCD to Linear SVMs

In this section, we briefly describe how to apply StingyCD to the problem

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \|Mx\|^2 - \langle 1, x \rangle \\
\text{s.t.} & \quad x \in [0, C]^n.
\end{align*}
\]

(PSVM)

We note that (PSVM) is very similar to (P). If not for the constraint that \(x \leq C1\), in fact, (PSVM) would be an instance of (P)—we could solve (PSVM) by defining \(A = M\), \(b = 0\), and \(\lambda = -1\) and then running Algorithm 2.

To incorporate the new constraint, our CD update becomes

\[
\delta_{\text{SVM}} = \min \left\{ C - x_i^{(t-1)}, \delta \right\}.
\]

In this case, StingyCD’s same rule applies for guaranteeing coordinate \(i\) remains 0 during iteration \(t\). With a minor change, we can also check if \(x_i^{(t-1)}\) is guaranteed to remain \(C\) during iteration \(t\). Specifically, if \(x_i^{(t-1)} = C\) and \(q^{(t-1)} \leq -\tau_i\), then it is guaranteed that \(\delta_{\text{SVM}} = 0\).
G. Additional comparisons for Lasso problems

This section contains results using additional values of $\lambda$ for the experiments in §6.1. In general, we find the results to be quite consistent, regardless of $\lambda$. Only “CD + Safe Screening” seems to be greatly affected by this parameter.

G.1. Full results for finance dataset

Number of examples: $1.6 \times 10^4$. Number of features: $5.5 \times 10^5$. 

![Graphs showing relative suboptimality, support set precision, and support set recall for different values of $\lambda$ in the finance dataset.](image)
G.2. Full results for allstate dataset

Number of examples: $2.5 \times 10^5$. Number of features: $1.5 \times 10^4$.

allstate, $\lambda = 0.1\lambda_{\text{max}}$, $\|x^*\|_0 = 176$

allstate, $\lambda = 0.05\lambda_{\text{max}}$, $\|x^*\|_0 = 1404$

allstate, $\lambda = 0.02\lambda_{\text{max}}$, $\|x^*\|_0 = 4821$

allstate, $\lambda = 0.01\lambda_{\text{max}}$, $\|x^*\|_0 = 6828$
H. Additional comparisons for logistic regression problems

H.1. Full results for lending_club dataset

Number of examples: $1.1 \times 10^5$. Number of features: $3.1 \times 10^4$.

- **lending_club, $\lambda = 0.05 \lambda_{\max}, \|x^*\|_0 = 272**

- **lending_club, $\lambda = 0.02 \lambda_{\max}, \|x^*\|_0 = 878**

- **lending_club, $\lambda = 0.01 \lambda_{\max}, \|x^*\|_0 = 1937**

- **lending_club, $\lambda = 0.005 \lambda_{\max}, \|x^*\|_0 = 3780**
H.2. Full results for kdda dataset

Number of examples: $8.4 \times 10^6$. Number of features: $2.2 \times 10^6$.

StingyCD+ ProxNewton with Working Sets
CD ProxNewton with Working Sets StingyCD+ ProxNewton CD ProxNewton

$\lambda = 0.02 \lambda_{\max}$, $\|x^*\|_0 = 195$

$\lambda = 0.01 \lambda_{\max}$, $\|x^*\|_0 = 383$

$\lambda = 0.005 \lambda_{\max}$, $\|x^*\|_0 = 692$

$\lambda = 0.002 \lambda_{\max}$, $\|x^*\|_0 = 1616$