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* $q^{(t-1)} \le \tau_i$ *and* $x_i^{(t-1)} = 0$*, then*

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

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

$$\left\langle \mathbf{A}_i, \mathbf{r}^{(t-1)} \right\rangle - \lambda \le 0, \tag{3}$$

where we have used the definition $\mathbf{r}^{(t-1)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(t-1)}$.

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

$$\begin{aligned}
q^{(t)} &= q^{(t-1)} - 2\delta \left\langle \mathbf{A}_i, \mathbf{r}^{(t-1)} - \mathbf{rr} \right\rangle + \delta^2 \left\| \mathbf{A}_i \right\|^2 && (4)\\
&= \left\| \mathbf{r}^{(t-1)} - \mathbf{rr} \right\|^2 - 2\delta \left\langle \mathbf{A}_i, \mathbf{r}^{(t-1)} - \mathbf{rr} \right\rangle + \delta^2 \left\| \mathbf{A}_i \right\|^2 && (5)\\
&= \left\| \mathbf{r}^{(t-1)} - \delta \mathbf{A}_i - \mathbf{rr} \right\|^2 && (6)\\
&= \left\| \mathbf{r}^{(t)} - \mathbf{rr} \right\|^2 . && (7)
\end{aligned}$$

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

$$\begin{aligned}
q^{(t-1)} \le \tau_i \quad &\Rightarrow \quad \left\| \mathbf{r}^{(t-1)} - \mathbf{rr} \right\|^2 \le \operatorname{sign}(g_i) \frac{g_i^2}{\left\| \mathbf{A}_i \right\|^2} && (8)\\
&\Rightarrow \quad \left\| \mathbf{r}^{(t-1)} - \mathbf{rr} \right\| \le \frac{g_i}{\left\| \mathbf{A}_i \right\|} && (9)\\
&\Rightarrow \quad \left\| \mathbf{A}_i \right\| \left\| \mathbf{r}^{(t-1)} - \mathbf{rr} \right\| \le -\langle \mathbf{A}_i, \mathbf{rr} \rangle + \lambda && (10)\\
&\Rightarrow \quad \langle \mathbf{A}_i, \mathbf{rr} \rangle + \left\| \mathbf{A}_i \right\| \left\| \mathbf{r}^{(t-1)} - \mathbf{rr} \right\| - \lambda \le 0 && (11)\\
&\Rightarrow \quad \left\langle \mathbf{A}_i, \mathbf{r}^{(t-1)} \right\rangle - \lambda \le 0\,. && (12)
\end{aligned}$$

□

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

Proof. Note that at each iteration, CD computes a dot product of length $\mathrm{NNZ}(\mathbf{A}_i)$ to compute δ. If $\delta \ne 0$, an additional $\mathcal{O}(\mathrm{NNZ}(\mathbf{A}_i))$ operation is required to update $\mathbf{r}^{(t)}$.

Case 1: the update is skipped and $\mathbf{rr}$ 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 $\mathbf{x}_i^{(t-1)} = 0$. 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 $\mathbf{rr}$ 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 $\langle \mathbf{A}_i, \mathbf{rr} \rangle$, $\langle \mathbf{A}_i, \mathbf{r}^{(t-1)} \rangle$, and $\|\mathbf{A}_i\|^2$. The value of $\langle \mathbf{A}_i, \mathbf{rr} \rangle$ was computed when computing the threshold τ_i, and $\langle \mathbf{A}_i, \mathbf{r}^{(t-1)} \rangle$ and $\|\mathbf{A}_i\|^2$ are necessary to compute δ.

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

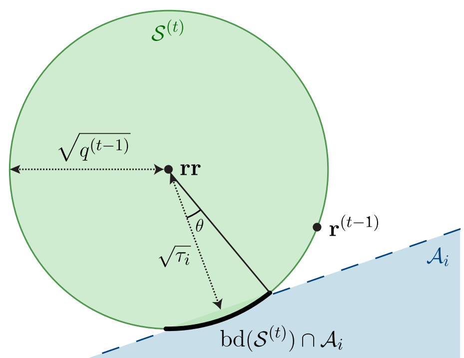
C. Proof of Theorem 3.2

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*

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

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

Proof. Recall the illustration form Figure 2:



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

$$\tfrac{1}{2}\mathrm{area}(\mathcal{S}^{(t)}) I_{\sin(\theta)^2}\left(\tfrac{n-1}{2}, \tfrac{1}{2}\right), \tag{13}$$

where $\mathrm{area}(\mathcal{S}^{(t)})$ is the surface area of $\mathcal{S}^{(t)}$ and θ is the angle indicated in the diagram.

When $\tau_i \geq 0$, note that by definition of τ_i, we have $\mathbf{rr} \notin \mathcal{A}_i$. It follows then that when $\tau_i \geq 0$, we have

$$P(U_t) = \frac{\frac{1}{2}\mathrm{area}(\mathcal{S}^{(t)}) I_{\sin(\theta)^2}\left(\frac{n-1}{2}, \frac{1}{2}\right)}{\mathrm{area}(\mathcal{S}^{(t)})} \tag{14}$$

$$= \tfrac{1}{2} I_{(1-\cos(\theta)^2)}\left(\tfrac{n-1}{2}, \tfrac{1}{2}\right) \tag{15}$$

$$= \tfrac{1}{2} I_{(1-\tau_i/q^{(t-1)})}\left(\tfrac{n-1}{2}, \tfrac{1}{2}\right). \tag{16}$$

In the case that $\tau_i < 0$, we have $\mathbf{rr} \in \mathcal{A}_i$, and we can use symmetry to see that

$$P(U_t) = 1 - \tfrac{1}{2} I_{(1+\tau_i/q^{(t-1)})}\left(\tfrac{n-1}{2}, \tfrac{1}{2}\right). \tag{17}$$

□

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

$$P(U^{(t)}) = \begin{cases} \frac{1}{2} I_{(1-\tau_i/q^{(t-1)})}(\frac{n-1}{2}, \frac{1}{2}) & \text{if } \tau_i \geq 0, \\ 1 - \frac{1}{2} I_{(1+\tau_i/q^{(t-1)})}(\frac{n-1}{2}, \frac{1}{2}) & \text{otherwise,} \end{cases} \tag{18}$$

where $I_x(a, b)$ 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

$$\tfrac{1}{2} I_{(1-\tau_i/q^{(t-1)})}(\tfrac{n-1}{2}, \tfrac{1}{2}) \approx 1 - \Phi\left(\sqrt{\tau_i(n-1)/q^{(t-1)}}\right). \tag{19}$$

Above, Φ 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 $\xi^{(t)} \leq \mathrm{NNZ}\left(\mathbf{x}^{(t-1)}\right)$ 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*

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

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 $\mathcal{C}^{(t-1)}$ as the set of coordinates that correspond to nonzero entries in $\mathbf{x}^{(t-1)}$:

$$\mathcal{C}^{(t-1)} = \{i : x_i^{(t-1)} \neq 0\}. \tag{20}$$

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

$$i_{\text{delayed}} = \operatorname*{argmax}_{i \in \mathcal{C}^{(t-1)}} D_i^{(t)}. \tag{21}$$

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

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

Now let $k \geq 0$ be the smallest such k for which get_next_coordinate() returns i_{delayed} during iteration $t+k$. Note that $k < M$. We must have $D_{i_{\text{delayed}}}^{(t+k)} \geq \text{NNZ}\left(\mathbf{x}^{(t+k-1)}\right)$, since (i) until an update for coordinate i is computed, $D_i^{(t)}$ is nondecreasing with t for all i, (ii) we have $D_{i_{\text{delayed}}}^{(t)} \geq \text{NNZ}\left(\mathbf{x}^{(t-1)}\right)$, and (iii) whenever $\text{NNZ}\left(\mathbf{x}^{(t')}\right) = \text{NNZ}\left(\mathbf{x}^{(t'-1)}\right) + 1$ 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 $\mathbf{x}$ increases the delay for all coordinates by 1.

In addition, since $i_{\text{delayed}} \in \mathcal{C}^{(t-1)}$ and i_{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 $\xi^{(t+k)} \leq \text{NNZ}\left(\mathbf{x}^{(t+k-1)}\right)$, it follows that

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

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_{delayed}. It follows that $D_{i_{\text{delayed}}}^{(t+k+1)} = 1$. That is, i_{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^{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 $\mathcal{F}$, StingyCD+ converges to a solution of the problem*

$$\begin{array}{ll} \underset{\mathbf{x} \in \mathbb{R}^m}{\text{minimize}} & f(\mathbf{x}) := \frac{1}{2}\left\|\mathbf{A}\mathbf{x} - \mathbf{b}\right\|^2 + \lambda \left\langle \mathbf{1}, \mathbf{x} \right\rangle \\ \text{s.t.} & \mathbf{x} \geq 0 \\ & x_i = 0 \ \ \forall i \in \mathcal{F} \end{array} . \tag{P'}$$

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

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

for some coordinate i, where

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

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

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

Now let us assume that $\mathbf{x}^{(t)}$ does not converge to a solution of (P') for some set $\mathcal{F}$. 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

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

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 ν.

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

$$f(\mathbf{x}^{(t)}) - f(\mathbf{x}^{(t-1)}) = \left(\lambda - \left\langle \mathbf{A}_i, \mathbf{r}^{(t-1)} \right\rangle\right) \delta + \frac{1}{2}\left\|\mathbf{A}_i\right\|^2 \delta^2 \tag{26}$$

$$\leq -\frac{1}{2}\left\|\mathbf{A}_i\right\|^2 \delta^2 . \tag{27}$$

Now define $\hat{f} = \lim_{t\to\infty} f(\mathbf{x}^{(t)})$. Consider an iteration t' such that $f(\mathbf{x}^{(t')}) \leq \hat{f} + \epsilon$, 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

$$\left| \max\left\{ -x_i^{(t-1)}, \frac{\langle \mathbf{A}_i, \mathbf{r}^{(t-1)} \rangle - \lambda}{\|\mathbf{A}_i\|^2} \right\} \right| \geq \nu \,. \tag{28}$$

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 δ'. It must be the case then that

$$\delta' \leq \frac{\sqrt{2\epsilon}}{\|\mathbf{A}_{i'}\|} \,. \tag{29}$$

Otherwise the fact that $\hat{f} = \lim_{t\to\infty} f(\mathbf{x}^{(t)})$ 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

$$\left\| \mathbf{r}^{(t-1)} - \mathbf{r}^{(T-1)} \right\| \leq mM\sqrt{2\epsilon} \,. \tag{30}$$

This implies that

$$\frac{\langle \mathbf{A}_i, \mathbf{r}^{(T-1)} \rangle}{\|\mathbf{A}_i\|^2} - \frac{\langle \mathbf{A}_i, \mathbf{r}^{(t-1)} \rangle}{\|\mathbf{A}_i\|^2} \in \left[-\tfrac{mM\sqrt{2\epsilon}}{\|\mathbf{A}_i\|}, +\tfrac{mM\sqrt{2\epsilon}}{\|\mathbf{A}_i\|} \right] \,. \tag{31}$$

Now let δ be the update to coordinate i during iteration T. It follows that

$$|\delta| = \left| \max\left\{ x_i^{(T-1)}, \frac{\langle \mathbf{A}_i, \mathbf{r}^{(T-1)} \rangle - \lambda}{\|\mathbf{A}_i\|^2} \right\} \right| \tag{32}$$

$$\geq \left| \max\left\{ x_i^{(t-1)}, \frac{\langle \mathbf{A}_i, \mathbf{r}^{(t-1)} \rangle - \lambda}{\|\mathbf{A}_i\|^2} \right\} \right| - \frac{mM\sqrt{2\epsilon}}{\|\mathbf{A}_i\|} \tag{33}$$

$$\geq \nu - \frac{mM\sqrt{2\epsilon}}{\|\mathbf{A}_i\|} \,. \tag{34}$$

Now let us define $s = \min_{i' \,:\, \|\mathbf{A}_{i'}\| > 0} \|\mathbf{A}_{i'}\|$.

$$\epsilon = \tfrac{1}{8} \left(\frac{\nu s}{mM} \right)^2 \tag{35}$$

Then it follows that

$$|\delta| > \tfrac{1}{2}\nu \,. \tag{36}$$

From (27), it follows that

$$f(x^{(T)}) \leq f(\mathbf{x}^{(T-1)}) - \tfrac{1}{2} \|\mathbf{A}_i\|^2 \delta^2 \leq \hat{f} + \epsilon - \tfrac{1}{2} s^2 \nu^2 < \hat{f} \,, \tag{37}$$

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

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

□

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

Now define $\hat{f} = \lim_{t\to\infty} f(\mathbf{x}^{(t)})$. Also define $\hat{\mathbf{r}} = \lim_{t\to\infty} \mathbf{r}^{(t)}$ and $\hat{\mathbf{x}} = \lim_{t\to\infty} \mathbf{x}^{(t)}$.

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 $\hat{x}_i \neq 0$, we have

$$\langle \mathbf{A}_i, \hat{\mathbf{r}} \rangle - \lambda \geq \nu \,. \tag{38}$$

Consider an iteration t' such that $f(\mathbf{x}^{(t'-1)}) \leq \hat{f} + \epsilon$, where we define $\epsilon > 0$ later. By Taylor expansion, we have for any $t \geq t'$,

$$f(\mathbf{x}^{(t)}) = f(\hat{\mathbf{x}}) + \left\langle \nabla f(\hat{\mathbf{x}}), \mathbf{x}^{(t)} - \hat{\mathbf{x}} \right\rangle + \tfrac{1}{2} \left\| \mathbf{A}\mathbf{x}^{(t)} - \mathbf{A}\hat{\mathbf{x}} \right\|^2 \tag{39}$$

$$\geq \hat{f} + \tfrac{1}{2} \left\| \hat{\mathbf{r}} - \mathbf{r}^{(t-1)} \right\|^2 \,. \tag{40}$$

This implies that for any $t \geq t'$, we have

$$\left\| \hat{\mathbf{r}} - \mathbf{r}^{(t-1)} \right\| \leq \sqrt{2\epsilon} \,. \tag{41}$$

Define $\epsilon = \min_{i' : \|\mathbf{A}_{i'}\| \neq 0} \frac{\nu^2}{8\|\mathbf{A}_{i'}\|^2}$. It follows then that for all $t \geq t'$,

$$\left\langle \mathbf{A}_i, \mathbf{r}^{(t-1)} \right\rangle - \lambda \geq \langle \mathbf{A}_i, \hat{\mathbf{r}} \rangle - \|\mathbf{A}_i\| \sqrt{2\epsilon} - \lambda \geq \nu - \|\mathbf{A}_i\| \sqrt{2\epsilon} \geq \tfrac{1}{2}\nu \,. \tag{42}$$

Also, if we assume $- \langle \mathbf{A}_i, \mathbf{rr} \rangle + \lambda > 0$, we must have

$$\tau_i = \frac{\left(- \langle \mathbf{A}_i, \mathbf{rr} \rangle + \lambda\right)^2}{\|\mathbf{A}_i\|^2} \tag{43}$$

$$\leq \frac{\left(- \left\langle \mathbf{A}_i, \mathbf{r}^{(t-1)} \right\rangle + \lambda + \|\mathbf{A}_i\| \left\| \mathbf{r}^{(t-1)} - \mathbf{rr} \right\|\right)^2}{\|\mathbf{A}_i\|^2} \tag{44}$$

$$\leq (q^{(t-1)} - \tfrac{1}{2}\nu)^2 \tag{45}$$

$$< q^{(t-1)} \,. \tag{46}$$

Otherwise, we must have $- \langle \mathbf{A}_i, \mathbf{rr} \rangle + \lambda < 0$, 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 coorinates 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

$$\delta \geq \tfrac{1}{2} \frac{\nu}{\|\mathbf{A}_i\|^2} \,, \tag{48}$$

which ensures that

$$f(\mathbf{x}^{(T)}) \leq f(\mathbf{x}^{(T-1)}) - \tfrac{1}{2} \|\mathbf{A}_i\|^2 \delta^2 \tag{49}$$

$$\leq f(\hat{\mathbf{x}}) + \epsilon - \tfrac{1}{2} \tfrac{\nu^2}{\|\mathbf{A}_i\|^2} \tag{50}$$

$$\leq f(\hat{\mathbf{x}}) - \tfrac{3}{8} \tfrac{\nu^2}{\|\mathbf{A}_i\|^2} \,. \tag{51}$$

This contradicts the definition of $\hat{\mathbf{x}}$. Thus, our assumption that $\mathbf{x}^{(t)}$ does not converge to a solution of (P) is incorrect.

□

F. Generalizing StingyCD to Linear SVMs

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

$$\begin{array}{ll} \underset{\mathbf{x} \in \mathbb{R}^n}{\text{minimize}} & \frac{1}{2} \|\mathbf{M}\mathbf{x}\|^2 - \langle \mathbf{1}, \mathbf{x} \rangle \\ \text{s.t.} & \mathbf{x} \in [0, C]^n \end{array} . \tag{PSVM}$$

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

To incorporate the new constraint, our CD update becomes

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

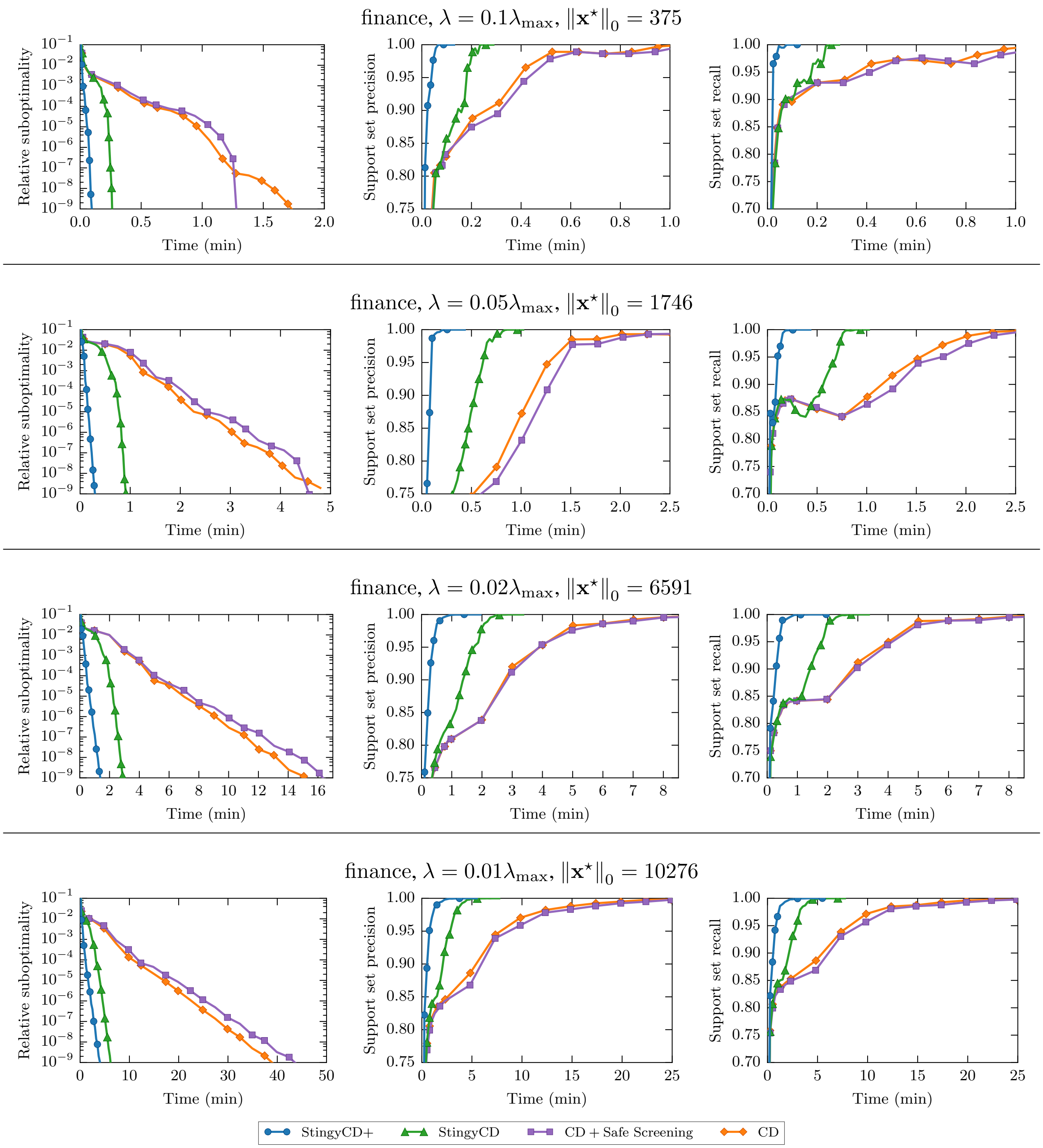
In this case, StingyCD's same rule applies for guaranteeing coordinate i remains 0 during iteration t. With a minor change, we can also check if $x_i^{(t-1)}$ is guaranteed to remain C during iteration t. Specifically, if $x_i^{(t-1)} = C$ and $q^{(t-1)} \leq -\tau_i$, then it is guaranteed that $\delta_{\text{SVM}}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

G. Additional comparisons for Lasso problems

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

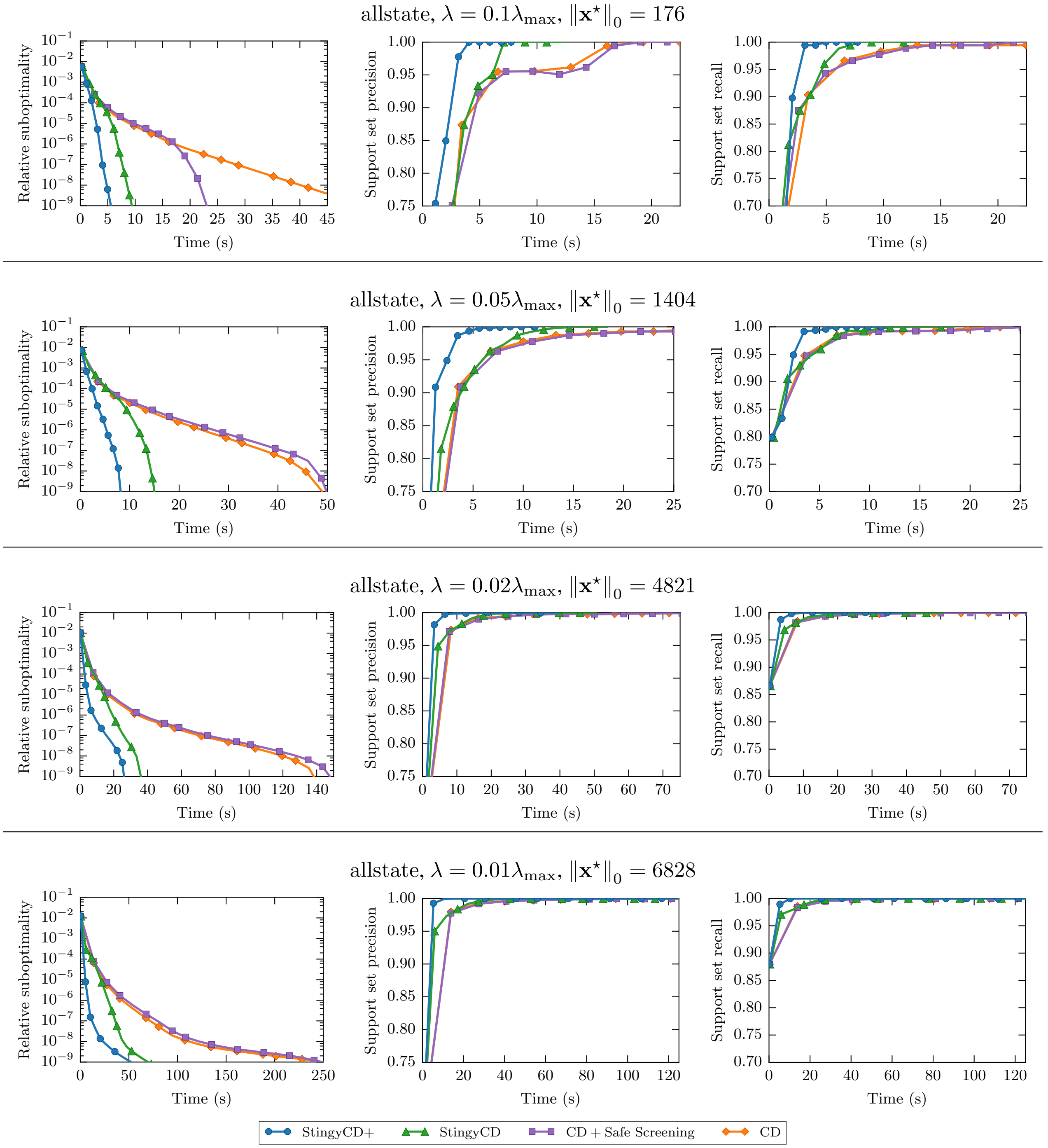
G.1. Full results for finance dataset

Number of examples: 1.6×10^4. Number of features: 5.5×10^5.



G.2. Full results for allstate dataset

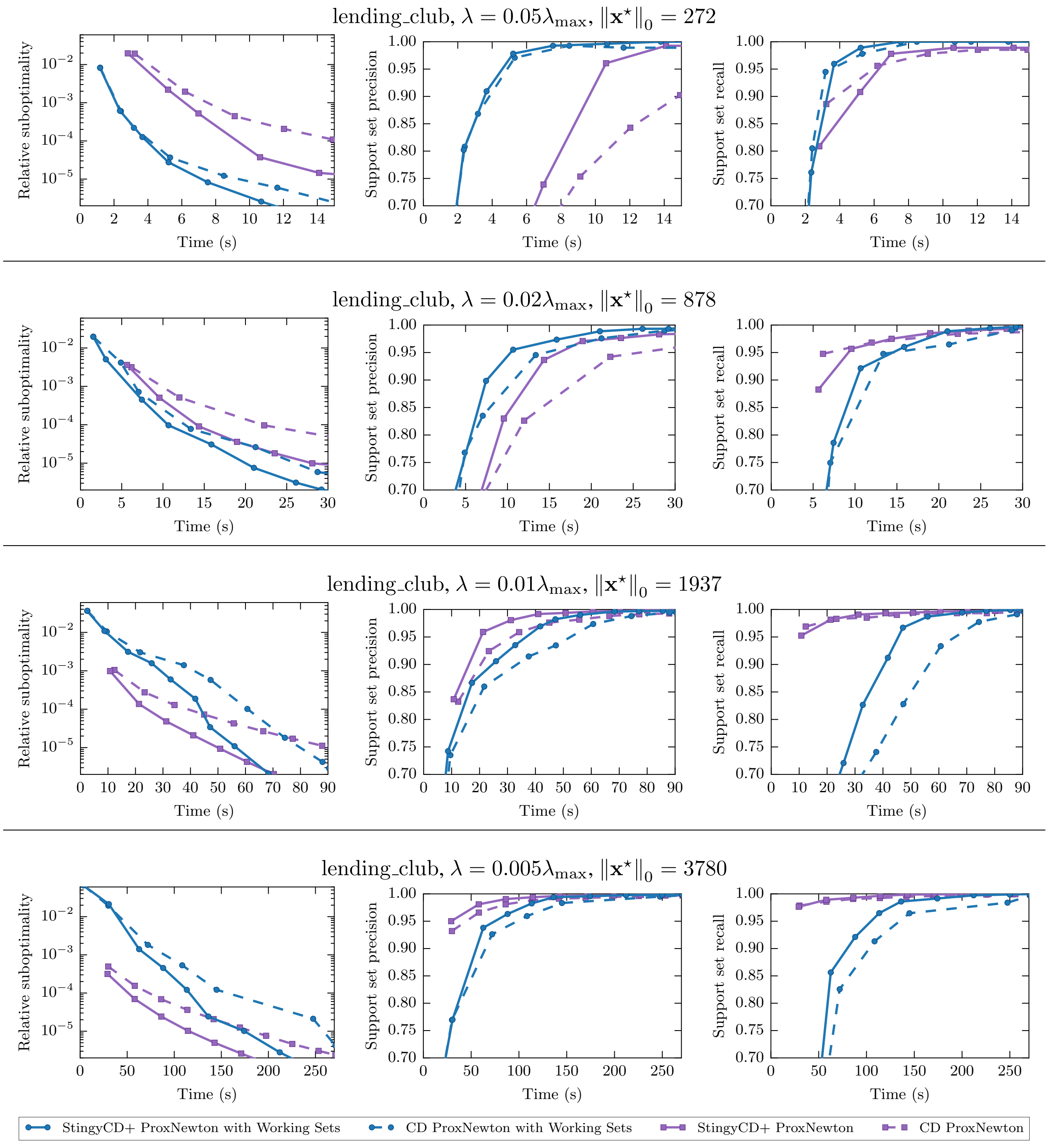
Number of examples: 2.5×10^5. Number of features: 1.5×10^4.



H. Additional comparisons for logistic regression problems

H.1. Full results for lending_club dataset

Number of examples: 1.1×10^5. Number of features: 3.1×10^4.



H.2. Full results for kdda dataset

Number of examples: 8.4×10^6. Number of features: 2.2×10^6.