Parallel Coordinate Descent for L_1-Regularized Loss Minimization: Theory Supplement

Abstract

In this supplementary document, we give detailed proofs of all theoretical results of the main paper.

1. Preliminaries

General form for our optimization problems, modified to use duplicate features and have a twice-differentiable regularization term:

$$\min_{\mathbf{x}\in\mathbb{R}_+^{2d}} \sum_{i=1}^{n} L(\hat{\mathbf{a}}_i^T\mathbf{x}, y_i) + \lambda \sum_{j=1}^{2d} x_j \tag{1}$$

Instantiation of Eq. (1) for Lasso (**?**):

$$F(\mathbf{x}) = \tfrac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 + \lambda\sum_{j=1}^{2d} x_j \tag{2}$$

Instantiation of Eq. (1) for sparse logistic regression:

$$F(\mathbf{x}) = \sum_{i=1}^{n} \log\left(1+\exp\left(-y_i\hat{\mathbf{a}}_i^T\mathbf{x}\right)\right) + \lambda\sum_{j=1}^{2d} x_j \tag{3}$$

Update rule for $x_j \leftarrow x_j + \delta x_j$:

$$\delta x_j = \max\{-x_j, -(\nabla F(\mathbf{x}))_j/\beta\} \tag{4}$$

2. Detailed Proofs: β for Squared Loss and Logistic Loss

Assumptions 2.1 and 3.1 both upper bound the change in objective from updating $\mathbf{x}$ with $\Delta\mathbf{x}$. We show how to do so for Assumption 3.1, which generalizes Assumption 2.1. For both losses, we upper-bound the objective using a second-order Taylor expansion of F around $\mathbf{x}$.

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

2.1. Proof: β for Squared Loss

$$\nabla F(\mathbf{x}) = \mathbf{A}^T\mathbf{A}\mathbf{x} - \mathbf{A}^T\mathbf{y} + \lambda\mathbf{1} \tag{5}$$

$$\nabla^2 F(\mathbf{x}) = \mathbf{A}^T\mathbf{A} \tag{6}$$

where $\mathbf{1}$ is an all-ones vector of the appropriate size. Note that, since derivatives of (2) of order higher than two are zero, the second order Taylor expansion is exact:

$$\begin{aligned} &F(\mathbf{x}+\Delta\mathbf{x}) \\ &= F(\mathbf{x}) + (\Delta\mathbf{x})^T\nabla F(\mathbf{x}) + \tfrac{1}{2}(\Delta\mathbf{x})^T\nabla^2 F(\mathbf{x})\Delta\mathbf{x} \end{aligned} \tag{7}$$

Plugging in the second order derivative gives $\beta = 1$:

$$\begin{aligned} &F(\mathbf{x}+\Delta\mathbf{x}) \\ &= F(\mathbf{x}) + (\Delta\mathbf{x})^T\nabla F(\mathbf{x}) + \tfrac{1}{2}(\Delta\mathbf{x})^T\mathbf{A}^T\mathbf{A}\Delta\mathbf{x}\,. \end{aligned} \tag{8}$$

This bound is exact for squared loss but not for all losses.

2.2. Proof: β for Logistic Loss

Define $p_i = \frac{1}{1+\exp\left(-\mathbf{a}_i^T\mathbf{x}\right)}$, the class conditional probability of y_i given $\mathbf{a}_i$.

$$\tfrac{\partial}{\partial x_j}F(\mathbf{x}) = \lambda + \sum_{i=1}^{n} y_i\mathbf{A}_{ij}\left(p_i - 1\right) \tag{9}$$

$$\tfrac{\partial^2}{\partial x_j \partial x_k}F(\mathbf{x}) = \sum_{i=1}^{n} \mathbf{A}_{ij}\mathbf{A}_{ik}\left(1-p_i\right)p_i \tag{10}$$

Taylor's theorem tells us that there exists an $\hat{\mathbf{x}}$ s.t.

$$\begin{aligned} &F(\mathbf{x}+\Delta\mathbf{x}) \\ &\leq F(\mathbf{x}) + (\Delta\mathbf{x})^T\nabla F(\mathbf{x}) + \tfrac{1}{2}(\Delta\mathbf{x})^T(\nabla^2 F(\hat{\mathbf{x}}))\Delta\mathbf{x} \end{aligned} \tag{11}$$

The second-order term is maximized by setting $p_i = \frac{1}{2}$ in $\frac{\partial^2}{\partial x_j \partial x_k}F(\mathbf{x})$ for each j,k. Plugging this in gives our bound with $\beta = \frac{1}{4}$:

$$\begin{aligned} &F(\mathbf{x}+\Delta\mathbf{x}) \\ &\leq F(\mathbf{x}) + (\Delta\mathbf{x})^T\nabla F(\mathbf{x}) + \tfrac{1}{2}\tfrac{(\Delta\mathbf{x})^T\mathbf{A}^T\mathbf{A}\Delta\mathbf{x}}{4} \end{aligned} \tag{12}$$

3. Duplicated Features

Our work, like Shalev-Shwartz and Tewari (?)'s work, uses duplicated features (with $\mathbf{x} \in \mathbb{R}^{2d}$ and $\mathbf{A} \in \mathbb{R}^{n \times 2d}$), but our actual algorithm does not (so $\mathbf{x} \in \mathbb{R}^{d}_{+}$ and $\mathbf{A} \in \mathbb{R}^{n \times d}$). They point out that the optimization problems with and without duplicated features are equivalent.

To see this, consider the form of $F(\mathbf{x})$ in Eq. (1). x_j only appears in the dot product $\hat{\mathbf{a}}_i^T \mathbf{x}$ via $\mathbf{A}_{i,j} x_j$, and x_{d+j} only appears via $-\mathbf{A}_{i,j} x_{d+j}$, where $\mathbf{A}$ is the original design matrix without duplicated features. Suppose $x_j > 0$ and $x_{d+j} > 0$, and assume w.l.o.g. that $x_j > x_{d+j}$. Then setting $x_j \longleftarrow x_j - x_{d+j}$ and $x_{d+j} \longleftarrow 0$ would give the same value for the loss term $L(\hat{\mathbf{a}}_i^T \mathbf{x}, y_i)$, and it would decrease the regularization penalty by $2\lambda x_{d+j}$. Therefore, at the optimum, at most one of x_j, x_{d+j} will be non-zero, and the objectives with and without duplicated features will be equal.

4. Detailed Proof: Theorem 2.1

Define a potential function, where $\mathbf{x}^*$ is the optimal weight vector:

$$\Psi(\mathbf{x}) = \frac{\beta}{2} \|\mathbf{x} - \mathbf{x}^*\|_2^2 + F(\mathbf{x}) \tag{13}$$

Claim: After updating weight x_j with δx_j,

$$\Psi(\mathbf{x}) - \Psi(\mathbf{x} + \Delta\mathbf{x}) \geq (x_j - x_j^*)(\nabla F(\mathbf{x}))_j \tag{14}$$

To see this:

$$\begin{aligned}
&\Psi(\mathbf{x}) - \Psi(\mathbf{x} + \Delta\mathbf{x}) \\
&\quad = \tfrac{\beta}{2}\left[\|\mathbf{x} - \mathbf{x}^*\|_2^2 - \|\mathbf{x} + \Delta\mathbf{x} - \mathbf{x}^*\|_2^2\right] && (15) \\
&\qquad + F(\mathbf{x}) - F(\mathbf{x} + \Delta\mathbf{x}) \\
&\quad = -\tfrac{\beta}{2}\left[2(\mathbf{x} - \mathbf{x}^*)^T \Delta\mathbf{x} + (\delta x_j)^2\right] && (16) \\
&\qquad + F(\mathbf{x}) - F(\mathbf{x} + \Delta\mathbf{x}) \\
&\quad \geq \beta\left(-\mathbf{x}^T \Delta\mathbf{x} + \mathbf{x}^{*T} \Delta\mathbf{x} - \tfrac{(\delta x_j)^2}{2}\right) && (17) \\
&\qquad - (\Delta\mathbf{x})^T \nabla F(\mathbf{x}) - \tfrac{\beta}{2}(\delta x_j)^2 \\
&\quad = \beta\left(-x_j \delta x_j + x_j^* \delta x_j - (\delta x_j)^2\right) && (18) \\
&\qquad - \delta x_j (\nabla F(\mathbf{x}))_j \\
&\quad \geq \beta\left(-x_j \delta x_j - (\delta x_j)^2\right) - x_j^* (\nabla F(\mathbf{x}))_j && (19) \\
&\qquad - \delta x_j (\nabla F(\mathbf{x}))_j
\end{aligned}$$

Above, Eq. (17) used Assumption 2.1. Eq. (19) used the update rule for choosing δx_j in Eq. (4). Now there are two possible cases stemming from the update rule. Case 1: If $\delta x_j = -x_j$, then Eq. (19) simplifies to

$$\Psi(\mathbf{x}) - \Psi(\mathbf{x} + \Delta\mathbf{x}) \geq (x_j - x_j^*)(\nabla F(\mathbf{x}))_j \tag{20}$$

Case 2: If $\delta x_j = -(\nabla F(\mathbf{x}))_j / \beta$, then Eq. (19) again simplifies to

$$\begin{aligned}
&\Psi(\mathbf{x}) - \Psi(\mathbf{x} + \Delta\mathbf{x}) && (21) \\
&\quad \geq x_j (\nabla F(\mathbf{x}))_j - \beta(\delta x_j)^2 - x_j^* (\nabla F(\mathbf{x}))_j && (22) \\
&\qquad + \beta(\delta x_j)^2 \\
&\quad = (x_j - x_j^*)(\nabla F(\mathbf{x}))_j && (23)
\end{aligned}$$

Having proved our claim, we can now take the expectation of Eq. (14) w.r.t. j, the chosen weight:

$$\begin{aligned}
&\mathbf{E}\left[\Psi(\mathbf{x}) - \Psi(\mathbf{x} + \Delta\mathbf{x})\right] \\
&\quad \geq \mathbf{E}\left[(x_j - x_j^*)(\nabla F(\mathbf{x}))_j\right] && (24) \\
&\quad = \frac{1}{2d} \mathbf{E}\left[(\mathbf{x} - \mathbf{x}^*)^T \nabla F(\mathbf{x})\right] && (25) \\
&\quad \geq \frac{1}{2d} \mathbf{E}\left[F(\mathbf{x}) - F(\mathbf{x}^*)\right] && (26)
\end{aligned}$$

In Eq. (25), we write $\frac{1}{2d}$ instead of $\frac{1}{d}$ (which Shalev-Shwartz and Tewari (?) write), for there are another d duplicates of each of the original d weights. Eq. (26) holds since $F(\mathbf{x})$ is convex.

Summing over $T+1$ iterations gives:

$$\begin{aligned}
&\mathbf{E}\left[\sum_{t=0}^{T} \Psi(\mathbf{x}^{(t)}) - \Psi(\mathbf{x}^{(t+1)})\right] \\
&\quad \geq \frac{1}{2d} \mathbf{E}\left[\sum_{t=0}^{T} F(\mathbf{x}^{(t)})\right] - \frac{T+1}{2d} F(\mathbf{x}^*) && (27) \\
&\quad \geq \frac{T+1}{2d}\left[\mathbf{E}\left[F(\mathbf{x}^{(T)})\right] - F(\mathbf{x}^*)\right] && (28)
\end{aligned}$$

where Eq. (28) used the fact that $F(\mathbf{x}_t)$ decreases monotonically with t. Since $\sum_{t=0}^{T} \Psi(\mathbf{x}^{(t)}) - \Psi(\mathbf{x}^{(t+1)}) = \Psi(\mathbf{x}^{(0)}) - \Psi(\mathbf{x}^{(T+1)})$, rearranging the above inequality gives

$$\begin{aligned}
&\mathbf{E}\left[F(\mathbf{x}_T)\right] - F(\mathbf{x}^*) && (29) \\
&\quad \leq \frac{2d}{T+1} \mathbf{E}\left[\Psi(\mathbf{x}^{(0)}) - \Psi(\mathbf{x}^{(T+1)})\right] && (30) \\
&\quad \leq \frac{2d}{T+1} \mathbf{E}\left[\Psi(\mathbf{x}^{(0)})\right] && (31) \\
&\quad = \frac{2d}{T+1}\left[\frac{\beta}{2}\|\mathbf{x}^*\|_2^2 + F(\mathbf{x}^{(0)})\right] && (32)
\end{aligned}$$

where Eq. (31) used $\Psi(\mathbf{x}) \geq 0$ and Eq. (32) used $\mathbf{x}^{(0)} = 0$.

This bound divides by $T+1$ instead of T (which Shalev-Shwartz and Tewari (?) do). Also, their theorem has an extra factor of $\frac{1}{2}$ on the right-hand side but should not due to the doubled length of $\mathbf{x}$ (though careful analysis without duplicated features could likely re-introduce the $\frac{1}{2}$).

5. Detailed Proof: Theorem 3.1

Start with Eq. (8), and note that the update rule in Eq. (4) implies that $\delta x_j \leq -(\nabla F(\mathbf{x}))_j$ (with $\beta = 1$ for Lasso). This gives us:

$$\begin{aligned} & F(\mathbf{x}+\Delta\mathbf{x}) - F(\mathbf{x}) \\ & \quad \leq -(\Delta\mathbf{x})^T(\Delta\mathbf{x}) + \tfrac{1}{2}(\Delta\mathbf{x})^T \mathbf{A}^T\mathbf{A}\Delta\mathbf{x} \end{aligned} \tag{33}$$

Noting that $\Delta\mathbf{x}$ can only have non-zeros in the indices in $\mathcal{P}_t$, we can rewrite this as

$$\begin{aligned} & F(\mathbf{x}+\Delta\mathbf{x}) - F(\mathbf{x}) \\ & \leq -\textstyle\sum_{j\in\mathcal{P}_t}(\delta x_j)^2 + \tfrac{1}{2}\sum_{i,j\in\mathcal{P}_t}(\mathbf{A}^T\mathbf{A})_{i,j}\delta x_i \delta x_j \end{aligned} \tag{34}$$

Separating out the diagonal terms in the sum over i, j and using $diag(\mathbf{A}^T\mathbf{A}) = \mathbf{1}$ gives the desired result:

$$\begin{aligned} & F(\mathbf{x}+\Delta\mathbf{x}) - F(\mathbf{x}) \\ & \leq -\tfrac{1}{2}\textstyle\sum_{j\in\mathcal{P}_t}(\delta x_j)^2 + \tfrac{1}{2}\sum_{\substack{i,j\in\mathcal{P}_t, \\ i\neq j}}(\mathbf{A}^T\mathbf{A})_{i,j}\delta x_i \delta x_j \end{aligned} \tag{35}$$

6. Detailed Proof: Theorem 3.2

This proof uses the result from Lemma 3.3, which is proven in detail in Sec. 7.

Modify the potential function used for sequential SCD:

$$\Psi(\mathbf{x}) = \frac{\beta}{2}\|\mathbf{x}-\mathbf{x}^*\|_2^2 + \tfrac{1}{1-\epsilon}F(\mathbf{x}), \tag{36}$$

where ϵ is defined as in Eq. (67). Assume that $\mathtt{P}$ is chosen s.t. $\epsilon < 1$.

Write out the change in the potential function from an update $\Delta\mathbf{x}$:

$$\begin{aligned} & \Psi(\mathbf{x}) - \Psi(\mathbf{x}+\Delta\mathbf{x}) \\ & = \tfrac{\beta}{2}\left[\|\mathbf{x}-\mathbf{x}^*\|_2^2 - \|\mathbf{x}+\Delta\mathbf{x}-\mathbf{x}^*\|_2^2\right] \\ & \quad + \tfrac{1}{1-\epsilon}\left[F(\mathbf{x}) - F(\mathbf{x}+\Delta\mathbf{x})\right] & (37) \\ & = \tfrac{\beta}{2}\left[-2\mathbf{x}^T\Delta\mathbf{x} + 2\mathbf{x}^{*T}\Delta\mathbf{x} - (\Delta\mathbf{x})^T(\Delta\mathbf{x})\right] \\ & \quad + \tfrac{1}{1-\epsilon}\left[F(\mathbf{x}) - F(\mathbf{x}+\Delta\mathbf{x})\right] & (38) \\ & = \beta\left[\sum_{j\in\mathcal{P}_t} -x_j\delta x_j + x_j^*\delta x_j - \tfrac{(\delta x_j)^2}{2}\right] \\ & \quad + \tfrac{1}{1-\epsilon}\left[F(\mathbf{x}) - F(\mathbf{x}+\Delta\mathbf{x})\right] & (39) \end{aligned}$$

Take the expectation w.r.t. $\mathcal{P}_t$, and use Lemma 3.3:

$$\begin{aligned} & \mathbf{E}_{\mathcal{P}_t}\left[\Psi(\mathbf{x}) - \Psi(\mathbf{x}+\Delta\mathbf{x})\right] \\ & = \beta\mathtt{P}\mathbf{E}_j\left[-x_j\delta x_j + x_j^*\delta x_j - \tfrac{(\delta x_j)^2}{2}\right] \\ & \quad + \tfrac{1}{1-\epsilon}\mathbf{E}_{\mathcal{P}_t}\left[F(\mathbf{x}) - F(\mathbf{x}+\Delta\mathbf{x})\right] & (40) \\ & \geq \beta\mathtt{P}\mathbf{E}_j\left[-x_j\delta x_j + x_j^*\delta x_j - \tfrac{(\delta x_j)^2}{2}\right] \\ & \quad -\mathtt{P}\tfrac{1}{1-\epsilon}\mathbf{E}_j\left[\delta x_j(\nabla F(\mathbf{x}))_j + \tfrac{\beta}{2}(1+\epsilon)(\delta x_j)^2\right] & (41) \\ & = \beta\mathtt{P}\mathbf{E}_j\left[-x_j\delta x_j + x_j^*\delta x_j - \tfrac{1}{1-\epsilon}(\delta x_j)^2\right] \\ & \quad -\mathtt{P}\tfrac{1}{1-\epsilon}\mathbf{E}_j\left[\delta x_j(\nabla F(\mathbf{x}))_j\right] & (42) \\ & \geq \beta\mathtt{P}\mathbf{E}_j\Big[-x_j\delta x_j - x_j^*(\nabla F(\mathbf{x}))_j/\beta \\ & \qquad -\tfrac{1}{1-\epsilon}(\delta x_j)^2 - \tfrac{1}{1-\epsilon}\delta x_j(\nabla F(\mathbf{x}))_j/\beta\Big] & (43) \end{aligned}$$

where the last inequality used the update rule in Eq. (4), which implies $\delta x_j \geq -(\nabla F(\mathbf{x}))_j/\beta$.

Consider the two cases in the update rule in Eq. (4). Case 1: $\delta x_j = -x_j \geq -(\nabla F(\mathbf{x}))_j/\beta$.

$$\begin{aligned} & \mathbf{E}_{\mathcal{P}_t}\left[\Psi(\mathbf{x}) - \Psi(\mathbf{x}+\Delta\mathbf{x})\right] \\ & \geq \beta\mathtt{P}\mathbf{E}_j\Big[-x_j\delta x_j - x_j^*(\nabla F(\mathbf{x}))_j/\beta \\ & \qquad + \tfrac{1}{1-\epsilon}x_j\delta x_j + \tfrac{1}{1-\epsilon}x_j(\nabla F(\mathbf{x}))_j/\beta\Big] & (44) \\ & = \beta\mathtt{P}\mathbf{E}_j\Big[\tfrac{\epsilon}{1-\epsilon}x_j\delta x_j - x_j^*(\nabla F(\mathbf{x}))_j/\beta \\ & \qquad + \tfrac{1}{1-\epsilon}x_j(\nabla F(\mathbf{x}))_j/\beta\Big] & (45) \\ & \geq \beta\mathtt{P}\mathbf{E}_j\Big[-\tfrac{\epsilon}{1-\epsilon}x_j(\nabla F(\mathbf{x}))_j/\beta - x_j^*(\nabla F(\mathbf{x}))_j/\beta \\ & \qquad + \tfrac{1}{1-\epsilon}x_j(\nabla F(\mathbf{x}))_j/\beta\Big] & (46) \\ & = \mathtt{P}\mathbf{E}_j\Big[(x_j - x_j^*)(\nabla F(\mathbf{x}))_j\Big]. & (47) \end{aligned}$$

Case 2: $\delta x_j = -(\nabla F(\mathbf{x}))_j/\beta \geq -x_j$.

$$\begin{aligned} & \mathbf{E}_{\mathcal{P}_t}\left[\Psi(\mathbf{x}) - \Psi(\mathbf{x}+\Delta\mathbf{x})\right] \\ & \geq \beta\mathtt{P}\mathbf{E}_j\Big[-x_j\delta x_j - x_j^*(\nabla F(\mathbf{x}))_j/\beta\Big] & (48) \\ & \geq \mathtt{P}\mathbf{E}_j\Big[(x_j - x_j^*)(\nabla F(\mathbf{x}))_j\Big]. & (49) \end{aligned}$$

In both cases,

$$\begin{aligned} & \mathbf{E}_{\mathcal{P}_t}\left[\Psi(\mathbf{x}) - \Psi(\mathbf{x}+\Delta\mathbf{x})\right] \\ & \geq \mathtt{P}\mathbf{E}_j\left[(x_j - x_j^*)(\nabla F(\mathbf{x}))_j\right] & (50) \\ & = \tfrac{\mathtt{P}}{2d}(\mathbf{x}-\mathbf{x}^*)^T\nabla F(\mathbf{x}) & (51) \\ & \geq \tfrac{\mathtt{P}}{2d}(F(\mathbf{x}) - F(\mathbf{x}^*)), & (52) \end{aligned}$$

where the last inequality holds since $F(\mathbf{x})$ is convex.

Now sum over $T+1$ iterations (with an expectation over the $\mathcal{P}_t$ from all iterations):

$$\begin{aligned}
&\mathbf{E}\left[\sum_{t=0}^{T} \Psi(\mathbf{x}^{(t)}) - \Psi(\mathbf{x}^{(t+1)})\right] \\
&\quad \geq \tfrac{\mathtt{P}}{2d}\mathbf{E}\left[\sum_{t=0}^{T} F(\mathbf{x}^{(t)}) - F(\mathbf{x}^{*})\right] && (53)\\
&\quad = \tfrac{\mathtt{P}}{2d}\left[\mathbf{E}\left[\sum_{t=0}^{T} F(\mathbf{x}^{(t)})\right] - (T+1)F(\mathbf{x}^{*})\right] && (54)\\
&\quad \geq \tfrac{\mathtt{P}(T+1)}{2d}\left[\mathbf{E}\left[F(\mathbf{x}^{(T)})\right] - F(\mathbf{x}^{*})\right], && (55)
\end{aligned}$$

where Eq. (55) uses the result from Lemma 3.3, which implies that the objective is decreasing in expectation for $\mathtt{P}$ s.t. $\epsilon \leq 1$. (To see why the objective decreases in expectation, plug in the update rule in Eq. (4) into Eq. (68), and note that the right-hand side of Eq. (68) is negative.)

Since $\sum_{t=0}^{T} \Psi(\mathbf{x}^{(t)}) - \Psi(\mathbf{x}^{(t+1)}) = \Psi(\mathbf{x}^{(0)}) - \Psi(\mathbf{x}^{(T+1)})$, rearranging the above inequality gives

$$\begin{aligned}
&\mathbf{E}\left[F(\mathbf{x}^{(T)})\right] - F(\mathbf{x}^{*}) \\
&\quad \leq \tfrac{2d}{\mathtt{P}(T+1)}\mathbf{E}\left[\Psi(\mathbf{x}^{(0)}) - \Psi(\mathbf{x}^{(T+1)})\right] && (56)\\
&\quad \leq \tfrac{2d}{\mathtt{P}(T+1)}\mathbf{E}\left[\Psi(\mathbf{x}^{(0)})\right] && (57)\\
&\quad = \tfrac{2d}{\mathtt{P}(T+1)}\left[\tfrac{\beta}{2}\|\mathbf{x}^{*}\|_2^2 + \tfrac{1}{1-\epsilon}F(\mathbf{x}^{(0)})\right]. && (58)
\end{aligned}$$

7. Detailed Proof: Lemma 3.3

Note: Assume the algorithm chooses a set of $\mathtt{P}$ coordinates, not a multiset.

Starting with Assumption 3.1, we can rearrange terms as follows:

$$\begin{aligned}
&F(\mathbf{x} + \Delta\mathbf{x}) - F(\mathbf{x}) && (59)\\
&\quad \leq (\Delta\mathbf{x})^T \nabla F(\mathbf{x}) + \tfrac{\beta}{2}(\Delta\mathbf{x})^T \mathbf{A}^T\mathbf{A}(\Delta\mathbf{x})
\end{aligned}$$

Take the expectation w.r.t. $\mathcal{P}_t$, the set of updated weights, and use the fact that each set $\mathcal{P}_t$ is equally likely to be chosen.

$$\begin{aligned}
&\mathbf{E}_{\mathcal{P}_t}\left[F(\mathbf{x} + \Delta\mathbf{x}) - F(\mathbf{x})\right] \\
&\quad \leq \mathbf{E}_{\mathcal{P}_t}\left[\sum_{j \in \mathcal{P}_t} \delta x_j (\nabla F(\mathbf{x}))_j\right] && (60)\\
&\qquad + \tfrac{\beta}{2}\mathbf{E}_{\mathcal{P}_t}\left[\sum_{i,j \in \mathcal{P}_t} \delta x_i (\mathbf{A}^T\mathbf{A})_{i,j} \delta x_j\right] \\
&\quad = \mathbf{E}_{\mathcal{P}_t}\left[\sum_{j \in \mathcal{P}_t} \delta x_j (\nabla F(\mathbf{x}))_j + \tfrac{\beta}{2}(\delta x_j)^2\right] && (61)\\
&\qquad + \tfrac{\beta}{2}\mathbf{E}_{\mathcal{P}_t}\left[\sum_{\substack{i,j \in \mathcal{P}_t \\ i \neq j}} \delta x_i (\mathbf{A}^T\mathbf{A})_{i,j} \delta x_j\right] \\
&\quad = \mathtt{P}\mathbf{E}_j\left[\delta x_j (\nabla F(\mathbf{x}))_j + \tfrac{\beta}{2}(\delta x_j)^2\right] && (62)\\
&\qquad + \tfrac{\beta}{2}\mathtt{P}(\mathtt{P}-1)\mathbf{E}_{i,j\,:\,i \neq j}\left[\delta x_i (\mathbf{A}^T\mathbf{A})_{i,j} \delta x_j\right],
\end{aligned}$$

where $\mathbf{E}_j\,[]$ denotes an expectation w.r.t. j chosen uniformly at random from $\{1, \ldots, 2d\}$ and where $\mathbf{E}_{i,j\,:\,i \neq j}\,[]$ denotes an expectation w.r.t. a pair of distinct indices i, j chosen uniformly at random from $\{1, \ldots, 2d\}$.

Since indices in $\mathcal{P}_t$ are chosen uniformly at random, the expectations may be rewritten as

$$\begin{aligned}
&\mathbf{E}_{\mathcal{P}_t}\left[F(\mathbf{x} + \Delta\mathbf{x}) - F(\mathbf{x})\right] && (63)\\
&\quad \leq \tfrac{\mathtt{P}}{2d}\left[(\Delta x)^T(\nabla F(\mathbf{x})) + \tfrac{\beta}{2}(\Delta x)^T(\Delta x)\right] && (64)\\
&\qquad + \tfrac{\beta}{2}\tfrac{\mathtt{P}(\mathtt{P}-1)}{2d(2d-1)}\left[(\Delta x)^T\mathbf{A}^T\mathbf{A}(\Delta x) - (\Delta x)^T(\Delta x)\right],
\end{aligned}$$

where we are overloading the notation Δx: in Eq. (63), Δx only has non-zero entries in elements indexed by $\mathcal{P}_t$; in Eq. (64), Δx can have non-zero entries everywhere (set by the update rule in Eq. (4)).

The spectral radius, i.e., the largest eigenvalue, of a matrix $\mathbf{M}$ may be expressed as $\max_{\mathbf{z}} \frac{\mathbf{z}^T\mathbf{M}\mathbf{z}}{\mathbf{z}^T\mathbf{z}}$; see, e.g., Strang (**?**). Letting ρ be the spectral radius of $\mathbf{A}^T\mathbf{A}$, upper-bound the second-order term in Eq. (64):

$$\begin{aligned}
&\mathbf{E}_{\mathcal{P}_t}\left[F(\mathbf{x} + \Delta\mathbf{x}) - F(\mathbf{x})\right] \\
&\quad \leq \tfrac{\mathtt{P}}{2d}\left[(\Delta x)^T(\nabla F(\mathbf{x})) + \tfrac{\beta}{2}(\Delta x)^T(\Delta x)\right] && (65)\\
&\qquad + \tfrac{\beta}{2}\tfrac{\mathtt{P}(\mathtt{P}-1)}{2d(2d-1)}\left[\rho(\Delta x)^T(\Delta x) - (\Delta x)^T(\Delta x)\right] \\
&\quad = \tfrac{\mathtt{P}}{2d}(\Delta x)^T(\nabla F(\mathbf{x})) && (66)\\
&\qquad + \tfrac{\beta}{2}\tfrac{\mathtt{P}}{2d}\left(1 + \tfrac{(\mathtt{P}-1)(\rho-1)}{2d-1}\right)(\Delta x)^T(\Delta x)
\end{aligned}$$

Letting

$$\epsilon = \tfrac{(\mathtt{P}-1)(\rho-1)}{2d-1}, \qquad (67)$$

we can rewrite the right-hand side in terms of expectations over $j \in \{1, \ldots, 2d\}$ to get the lemma's result:

$$\begin{aligned} &\mathbf{E}_{\mathcal{P}_t}\left[F(\mathbf{x} + \Delta\mathbf{x}) - F(\mathbf{x})\right] \\ &\quad \leq \mathtt{P}\mathbf{E}_j\left[\delta x_j (\nabla F(\mathbf{x}))_j + \frac{\beta}{2}(1+\epsilon)(\delta x_j)^2\right]. \end{aligned} \tag{68}$$

Note: If we let the algorithm choose a multiset, rather than a set, of $\mathtt{P}$ coordinates, then we get $\epsilon = \frac{(\mathtt{P}-1)\rho}{2d}$, which gives worse scaling than the ϵ above. (Compare the two when all features are uncorrelated so that $\rho = 1$. With a set, the ϵ above indicates that we can use $\mathtt{P} = 2d$ and get good scaling; with a multiset, the changed ϵ indicates that we can be hurt by using larger $\mathtt{P}$.)