Appendix

A. Summary of Notation

Topic	Notation	Explanation
Data (sub)Sets and indices	U	Set of N instances in training set
	V	Set of M instances in validation set
	$\mathcal{X}^t$	Subset of instances from U at the t^{th} epoch
	W	A generic reference to both U and V
	$\pi_t^i \in \mathcal{X}$	Assignment of element $i \in W$ to an element of $\mathcal{X}$
Parameters	θ^*	Optimal model parameter (vector)
	θ_t	Updated parameter (vector) at the t^{th} epoch
	$\mathbf{w}^t$	Vector weights associated with each data point in $\mathcal{X}^t$ (at the t^{th} epoch)
Loss Functions	L_T	Training loss which when evaluated on $x_i \in U$ is referred to as L_T^i
	L_V	Validation loss which when evaluated on $x_j \in V$ is referred to as L_V^j
	L	Generic reference to the loss function which when evaluated on x_i is referred to as L^i.
	$\text{Err}(\mathbf{w}^t, \mathcal{X}^t, L, L_T, \theta_t)$	$\left\|\sum_{i\in\mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t) - \nabla_\theta L(\theta_t)\right\|$
	$E(\mathcal{X})$	$\min_{\mathbf{w}} \text{Err}(\mathbf{w}, \mathcal{X}, L, L_T, \theta_t)$
	$F(\mathcal{X})$	$L_{\max} - E(\mathcal{X})$ where $L_{\max}$ is an upperbound on $E(\mathcal{X})$
	$\hat{E}(\mathcal{X})$	The upper bound $\min_{\mathbf{w}} \text{Err}(\mathbf{w}, \mathcal{X}, L, L_T, \theta_t) \leq \hat{E}(\mathcal{X})$ minimized in Section
	$\hat{F}(\mathcal{X})$	The facility location lower bound function $\sum_{i\in W} \max_{j\in\mathcal{X}} \left(L_{\max} - \|\nabla_\theta L^i(\theta_t) - \nabla_\theta L_T^j(\theta_t)\|\right)$ to be maximized
	$E_\lambda(\mathcal{X}, \mathbf{w})$	Regularized version of $E(\mathcal{X})$ defined as $\left\|\sum_{i\in\mathcal{X}} \mathbf{w}\nabla_\theta L_T^i(\theta) - \nabla_\theta L(\theta)\right\|^2 + \lambda\|\mathbf{w}\|^2$. See Section
	$E_\lambda(\mathcal{X})$	$\min_{\mathbf{w}} E_\lambda(\mathcal{X}, \mathbf{w})$
	$F_\lambda(\mathcal{X})$	$L_{\max} - \min_{\mathbf{w}} E_\lambda(\mathcal{X}, \mathbf{w})$ which we prove to be γ-weakly submodular in Section and subsequently maximize
Hyperparameters	σ_T	Upperbound on the gradient of L_T^i
	σ_V	Upperbound on the gradient of L_V^j
	k	Size of selected subset of points
	R	The number of training epochs after which data selection is periodically performed
	α_t	The learning rate schedule at the t^{th} epoch

Table 1. Organization of the notations used througout this paper

B. Proofs of the Technical Results

B.1. Proof of Theorem 1

We begin by first stating and then proving Theorem 1.

Theorem *Any adaptive data selection algorithm (run with full gradient descent), defined via weights* $\mathbf{w}^t$ *and subsets* $\mathcal{X}^t$ *for* $t = 1, \cdots, T$*, enjoys the following guarantees:*

(1). If L_T *is Lipschitz continuous with parameter* σ_T*, optimal model parameters* θ^**, and* $\alpha = \frac{D}{\sigma_T\sqrt{T}}$*, then*

$\min_{t=1:T} L(\theta_t) - L(\theta^*) \leq \frac{D\sigma_T}{\sqrt{T}} + \frac{D}{T}\sum_{t=1}^{T-1} Err(\mathbf{w}^t, \mathcal{X}^t, L, L_T, \theta_t)$.

(2) If L_T is Lipschitz smooth with parameter $\mathcal{L}_T$, optimal model parameters θ^*, and L_T^i satisfies $0 \leq L_T^i(\theta) \leq \beta_T$, $\forall i$, then setting $\alpha = \frac{1}{\mathcal{L}_T}$, we have $\min_{t=1:T} L(\theta_t) - L(\theta^*) \leq \frac{D^2\mathcal{L}_T + 2\beta_T}{2T} + \frac{D}{T}\sum_{t=1}^{T-1} Err(\mathbf{w}^t, \mathcal{X}^t, L, L_T, \theta_t)$.

(3) If L_T is Lipschitz continuous with parameter σ_T, optimal model parameters θ^*, and L is strongly convex with parameter μ, then setting a learning rate $\alpha_t = \frac{2}{\mu(1+t)}$, we have $\min_{t=1:T} L(\theta_t) - L(\theta^*) \leq \frac{2{\sigma_T}^2}{\mu(T+1)} + \sum_{t=1}^{t=T} \frac{2Dt}{T(T+1)} Err(\mathbf{w}^t, \mathcal{X}^t, L, L_T, \theta_t)$.

PROOF Suppose the gradients of validation loss and training loss are sigma bounded by σ_V and σ_T respectively. Let θ_t be the model parameters at epoch t and θ^* be the optimal model parameters.

Let, $L_w(\theta_t) = \sum_{i \in \mathcal{X}^t} w_i^t L_T^i(\theta_t)$ be the weighted subset training loss parameterized by model parameters θ_t at time step t. Let α_t be the learning rate used at epoch t.

From the definition of Gradient Descent, we have:

$$\nabla_\theta L_w(\theta_t)^T(\theta_t - \theta^*) = \frac{1}{\alpha_t}(\theta_t - \theta_{t+1})^T(\theta_t - \theta^*) \tag{6}$$

$$\nabla_\theta L_w(\theta_t)^T(\theta_t - \theta^*) = \frac{1}{2\alpha_t}\left(\|\theta_t - \theta_{t+1}\|^2 + \|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2\right) \tag{7}$$

$$\nabla_\theta L_w(\theta_t)^T(\theta_t - \theta^*) = \frac{1}{2\alpha_t}\left(\left\|\alpha_t \sum_{i \in \mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t)\right\|^2 + \|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2\right) \tag{8}$$

We can rewrite the function $\nabla_\theta L_w(\theta_t)^T(\theta_t - \theta^*)$ as follows:

$$\nabla_\theta L_w(\theta_t)^T(\theta_t - \theta^*) = \nabla_\theta L_w(\theta_t)^T(\theta_t - \theta^*) - \nabla_\theta L(\theta_t)^T(\theta_t - \theta^*) + \nabla_\theta L(\theta_t)^T(\theta_t - \theta^*) \tag{9}$$

Combining the equations (8) and (9) we have,

$$\nabla_\theta L_w(\theta_t)^T(\theta_t - \theta^*) - \nabla_\theta L(\theta_t)^T(\theta_t - \theta^*) + \nabla_\theta L(\theta_t)^T(\theta_t - \theta^*) = \frac{1}{2\alpha_t}\left(\left\|\alpha_t \sum_{i \in \mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t)\right\|^2 + \|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2\right) \tag{10}$$

$$\nabla_\theta L(\theta_t)^T(\theta_t - \theta^*) = \frac{1}{2\alpha_t}\left(\left\|\alpha_t \sum_{i \in \mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t)\right\|^2 + \|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2\right) - (\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t))^T(\theta_t - \theta^*) \tag{11}$$

Summing up equation (11) for different values of $t \in [0, T-1]$ and assuming a constant learning rate of $\alpha_t = \alpha$, we have:

$$\begin{aligned}\sum_{t=0}^{T-1} \nabla_\theta L(\theta_t)^T(\theta_t - \theta^*) =& \frac{1}{2\alpha}\|\theta_0 - \theta^*\|^2 - \|\theta_T - \theta^*\|^2 + \sum_{t=0}^{T-1}\left(\frac{1}{2\alpha}\left\|\alpha \sum_{i \in \mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t)\right\|^2\right) \\ &+ \sum_{t=0}^{T-1}\left((\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t))^T(\theta_t - \theta^*)\right)\end{aligned}$$

Since $\|\theta_T - \theta^*\|^2 \geq 0$, we have:

$$\sum_{t=0}^{T-1} \nabla_\theta L(\theta_t)^T(\theta_t - \theta^*) \leq \frac{1}{2\alpha}\|\theta_0 - \theta^*\|^2 + \sum_{t=0}^{T-1}\left(\frac{1}{2\alpha}\left\|\alpha \sum_{i \in \mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t)\right\|^2\right) + \sum_{t=0}^{T-1}\left((\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t))^T(\theta_t - \theta^*)\right) \tag{12}$$

Case 1 L_T **is lipschitz continuous with parameter** σ_T **and** L **is a convex function**

From convexity of function $L(\theta)$, we know $L(\theta_t) - L(\theta^*) \leq \nabla_\theta L(\theta_t)^T(\theta_t - \theta^*)$. Combining this with Equation 12 we have,

$$\sum_{t=0}^{T-1} L(\theta_t) - L(\theta^*) \leq \frac{1}{2\alpha}\|\theta_0 - \theta^*\|^2 + \sum_{t=0}^{T-1}(\frac{1}{2\alpha}\left\|\alpha \sum_{i\in\mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t)\right\|^2) + \sum_{t=0}^{T-1}\left((\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t))^T(\theta_t - \theta^*)\right) \tag{13}$$

Since, $\|L_T(\theta)\| \leq \sigma_T$, we have $\left\|\alpha \sum_{i\in\mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t)\right\| \leq \sum_{i=1}^{|\mathcal{X}^t|} w_i^t \sigma$. Assuming that the weights at every iteration are normalized such that $\underset{t\in[1,T]}{\forall} \sum_{i=1}^{|\mathcal{X}^t|} w_i^t = 1$ and the training and validation loss gradients are normalized as well, we have $\left\|\alpha \sum_{i\in\mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t)\right\| \leq \sigma_T$. Also assuming that $\|\theta - \theta^*\| \leq D$, we have,

$$\sum_{t=0}^{T-1} L(\theta_t) - L(\theta^*) \leq \frac{D^2}{2\alpha} + \frac{T\alpha\sigma_T^2}{2} + \sum_{t=0}^{T-1} D\left(\|\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t)\|\right) \tag{14}$$

$$\frac{\sum_{t=0}^{T-1} L(\theta_t) - L(\theta^*)}{T} \leq \frac{D^2}{2\alpha T} + \frac{\alpha\sigma_T^2}{2} + \sum_{t=0}^{T-1} \frac{D}{T}\left(\|\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t)\|\right) \tag{15}$$

Since, $\min\left(L(\theta_t) - L(\theta^*)\right) \leq \frac{\sum_{t=0}^{T-1} L(\theta_t) - L(\theta^*)}{T}$, we have:

$$\min\left(L(\theta_t) - L(\theta^*)\right) \leq \frac{D^2}{2\alpha T} + \frac{\alpha\sigma_T^2}{2} + \sum_{t=0}^{T-1} \frac{D}{T}\left(\|\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t)\|\right) \tag{16}$$

Substituting $L_w(\theta_t) = \sum_{i\in\mathcal{X}^t} w_i^t L_T^i(\theta_t)$ in the above equation we have,

$$\min\left(L(\theta_t) - L(\theta^*)\right) \leq \frac{D^2}{2\alpha T} + \frac{\alpha\sigma_T^2}{2} + \sum_{t=0}^{T-1} \frac{D}{T}\left(\left\|\sum_{i\in\mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t) - \nabla_\theta L(\theta_t)\right\|\right) \tag{17}$$

Choosing $\alpha = \frac{D}{\sigma_T\sqrt{T}}$, we have:

$$\min\left(L(\theta_t) - L(\theta^*)\right) \leq \frac{D\sigma_T}{\sqrt{T}} + \sum_{t=0}^{T-1} \frac{D}{T}\left(\left\|\sum_{i\in\mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t) - \nabla_\theta L(\theta_t)\right\|\right) \tag{18}$$

Case 2 L_T **is lipschitz smooth with parameter** $\mathcal{L}_T$**, and** $\forall i$**,** L_T^i **satisfies** $0 \leq L_T^i(\theta) \leq \beta_T$

Since $L_w(\theta_t) = \sum_{i\in\mathcal{X}^t} w_i^t L_T^i(\theta_t)$, from the additive property of lipschitz smooth functions we can say that $L_w(\theta_t)$ is also lipschitz smooth with constant $\sum_{i\in\mathcal{X}^t} w_i^t \mathcal{L}_T$. Assuming that the weights at every iteration are normalized such that $\underset{t\in[0,T]}{\forall} \sum_{i=1}^{|\mathcal{X}^t|} w_i^t = 1$, we can say that $L_w(\theta_t)$ is lipschitz smooth with constant $\mathcal{L}_T$.

From lipschitz smoothness of function $L_w(\theta)$, we have:

$$L_w(\theta_{t+1}) \leq L_w(\theta_t) + \nabla_\theta L_w(\theta_t)^T(\theta_{t+1} - \theta_t) + \frac{L}{2}\|\theta_{t+1} - \theta_t\|^2 \tag{19}$$

Since $\theta_{t+1} - \theta_t = -\alpha\nabla_\theta L_w(\theta_t)$, we have:

$$L_w(\theta_{t+1}) \leq L_w(\theta_t) - \alpha\nabla_\theta L_w(\theta_t)^T\nabla_\theta L_w(\theta_t) + \frac{\mathcal{L}_T}{2}\|\alpha\nabla_\theta L_w(\theta_t)\|^2 \tag{20}$$

$$L_w(\theta_{t+1}) \leq L_w(\theta_t) + \frac{\mathcal{L}_T\alpha^2 - 2\alpha}{2}\|\nabla_\theta L_w(\theta_t)\|^2 \tag{21}$$

Choosing $\alpha = \frac{1}{\mathcal{L}_T}$, we have:

$$L_w(\theta_{t+1}) \leq L_w(\theta_t) - \frac{1}{2\mathcal{L}_T}\|\nabla_\theta L_w(\theta_t)\|^2 \tag{22}$$

Since $\nabla_\theta L_w(\theta_T) = \sum_{i\in\mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t)$, we have:

$$L_w(\theta_{t+1}) \leq L_w(\theta_t) - \frac{1}{2\mathcal{L}_T}\left\|\sum_{i\in\mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t)\right\|^2 \tag{23}$$

$$\frac{1}{2\mathcal{L}_T}\left\|\sum_{i\in\mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t)\right\|^2 \leq L_w(\theta_t) - L_w(\theta_{t+1}) \tag{24}$$

Summing the above equation for different values of t in $[0, T-1]$, we have:

$$\sum_{t=0}^{t=T-1} \frac{1}{2\mathcal{L}_T}\left\|\sum_{i\in\mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t)\right\|^2 \leq \sum_{t=0}^{t=T-1}(L_w(\theta_t) - L_w(\theta_{t+1})) \tag{25}$$

$$\sum_{t=0}^{t=T-1} \frac{1}{2\mathcal{L}_T}\left\|\sum_{i\in\mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t)\right\|^2 \leq L_w(\theta_0) - L_w(\theta_T) \tag{26}$$

Substituting $\alpha = \frac{1}{\mathcal{L}_T}$ in Equation 12, we have:

$$\sum_{t=0}^{T-1} \nabla_\theta L(\theta_t)^T(\theta_t - \theta^*) \leq \frac{\mathcal{L}_T}{2}\|\theta_0 - \theta^*\|^2 + \sum_{t=0}^{T-1}(\frac{1}{2\mathcal{L}_T}\left\|\sum_{i\in\mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t)\right\|^2) + \sum_{t=0}^{T-1}\Big((\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t))^T(\theta_t - \theta^*)\Big) \tag{27}$$

Substituting Equation 26, we have:

$$\sum_{t=0}^{T-1} \nabla_\theta L(\theta_t)^T(\theta_t - \theta^*) \leq \frac{\mathcal{L}_T}{2}\|\theta_0 - \theta^*\|^2 + L_w(\theta_0) - L_w(\theta_T) + \sum_{t=0}^{T-1}\Big((\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t))^T(\theta_t - \theta^*)\Big) \tag{28}$$

$$\sum_{t=0}^{T-1} \nabla_\theta L(\theta_t)^T(\theta_t - \theta^*) \leq \frac{\mathcal{L}_T}{2}\|\theta_0 - \theta^*\|^2 + L_w(\theta_0) + \sum_{t=0}^{T-1}\Big((\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t))^T(\theta_t - \theta^*)\Big) \tag{29}$$

Since $L_T(\theta)$ is bounded by β_T, we have $L_w(\theta) = \sum_{i\in\mathcal{X}^t} w_i^t L_T^i(\theta)$ bounded by β_T as the weights are normalized to 1 every epoch (i.e., $\underset{t\in[0,T]}{\forall}\sum_{i=1}^{|\mathcal{X}^t|} w_i^t = 1$).

$$\sum_{t=0}^{T-1} \nabla_\theta L(\theta_t)^T(\theta_t - \theta^*) \leq \frac{\mathcal{L}_T}{2}\|\theta_0 - \theta^*\|^2 + \beta_T + \sum_{t=0}^{T-1}\Big((\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t))^T(\theta_t - \theta^*)\Big) \tag{30}$$

Dividing the above equation by T, we have:

$$\frac{\sum_{t=0}^{T-1} \nabla_\theta L(\theta_t)^T(\theta_t - \theta^*)}{T} \leq \frac{\mathcal{L}_T}{2T}\|\theta_0 - \theta^*\|^2 + \frac{\beta_T}{T} + \frac{\sum_{t=0}^{T-1}\left((\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t))^T(\theta_t - \theta^*)\right)}{T} \tag{31}$$

Since $\|\theta - \theta^*\| \leq D$, we have,

$$\frac{\sum_{t=0}^{T-1} \nabla_\theta L(\theta_t)^T(\theta_t - \theta^*)}{T} \leq \frac{D^2\mathcal{L}_T}{2T} + \frac{\beta_T}{T} + \frac{D}{T}\sum_{t=0}^{T-1}(\|\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t)\|) \tag{32}$$

From convexity of function $L(\theta)$, we know $L(\theta_t) - L(\theta^*) \leq \nabla_\theta L(\theta_t)^T(\theta_t - \theta^*)$. Combining this with above equation we have,

$$\frac{\sum_{t=0}^{T-1} L(\theta_t) - L(\theta^*)}{T} \leq \frac{D^2\mathcal{L}_T}{2T} + \frac{\beta_T}{T} + \frac{D}{T}\sum_{t=0}^{T-1}(\|\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t)\|) \tag{33}$$

Since, $\min\left(L(\theta_t) - L(\theta^*)\right) \leq \frac{\sum_{t=0}^{T-1} L(\theta_t) - L(\theta^*)}{T}$, we have:

$$\min\left(L(\theta_t) - L(\theta^*)\right) \leq \frac{D^2\mathcal{L}_T}{2T} + \frac{\beta_T}{T} + \frac{D}{T}\sum_{t=0}^{T-1}(\|\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t)\|) \tag{34}$$

Substituting $L_w(\theta_t) = \sum_{i \in \mathcal{X}^t} w_i^t L_T^i(\theta_t)$ in the above equation we have,

$$\min\left(L(\theta_t) - L(\theta^*)\right) \leq \frac{D^2\mathcal{L}_T}{2T} + \frac{\beta_T}{T} + \frac{D}{T}\sum_{t=0}^{T-1}\left(\left\|\sum_{i \in \mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t) - \nabla_\theta L(\theta_t)\right\|\right) \tag{35}$$

Case 3 L_T **is Lipschitz continuous (parameter** σ_T**) and** L **is strongly convex with parameter** μ

Let the learning at time step t be α_t

From Equation 11, we have:

$$\nabla_\theta L(\theta_t)^T(\theta_t - \theta^*) = \frac{1}{2\alpha_t}\left(\left\|\alpha_t \sum_{i \in \mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t)\right\|^2 + \|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2\right) - (\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t))^T(\theta_t - \theta^*) \tag{36}$$

From the strong convexity of loss function L, we have:

$$\nabla_\theta L(\theta_t)^T(\theta_t - \theta^*) \geq L(\theta_t) - L(\theta^*) + \frac{\mu}{2}\|\theta_t - \theta^*\|^2 \tag{37}$$

Combining the above two equations, we have:

$$L(\theta_t) - L(\theta^*) = \frac{1}{2\alpha_t}\left(\left\|\alpha_t \sum_{i \in \mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t)\right\|^2 + \|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2\right) - (\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t))^T(\theta_t - \theta^*) - \frac{\mu}{2}\|\theta_t - \theta^*\|^2 \tag{38}$$

$$L(\theta_t) - L(\theta^*) = \frac{\alpha_t}{2}\left\|\sum_{i \in \mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t)\right\|^2 + \frac{\alpha_t^{-1} - \mu}{2}\|\theta_t - \theta^*\|^2 - \frac{\alpha_t^{-1}}{2}\|\theta_{t+1} - \theta^*\|^2 - (\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t))^T(\theta_t - \theta^*) \tag{39}$$

Setting an learning rate of $\alpha_t = \frac{2}{\mu(t+1)}$ and multiplying by t on both sides, we have:

$$\begin{aligned} t(L(\theta_t) - L(\theta^*)) = \frac{t}{\mu(t+1)} \left\| \sum_{i \in \mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t) \right\|^2 + \frac{\mu t(t-1)}{4} \|\theta_t - \theta^*\|^2 \\ - \frac{\mu t(t+1)}{4} \|\theta_{t+1} - \theta^*\|^2 - t(\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t))^T (\theta_t - \theta^*) \end{aligned} \tag{40}$$

Since, $\|L_T(\theta)\| \leq \sigma_T$, we have $\left\|\alpha \sum_{i \in \mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t)\right\| \leq \sum_{i=1}^{|\mathcal{X}^t|} w_i^t \sigma_T$. Assuming that the weights at every iteration are normalized such that $\underset{t \in [1,T]}{\forall} \sum_{i=1}^{|\mathcal{X}^t|} w_i^t = 1$ and the training and validation loss gradients are normalized as well, we have $\left\|\alpha \sum_{i \in \mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t)\right\| \leq \sigma_T$. Also assuming that $\|\theta - \theta^*\| \leq D$, we have,

$$\begin{aligned} t(L(\theta_t) - L(\theta^*)) = \frac{{\sigma_T}^2 t}{\mu(t+1)} + \frac{\mu t(t-1)}{4} \|\theta_t - \theta^*\|^2 \\ - \frac{\mu t(t+1)}{4} \|\theta_{t+1} - \theta^*\|^2 + Dt \|\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t)\| \end{aligned} \tag{41}$$

Summing the above equation from $t = 1, \cdots, T$, we have:

$$\begin{aligned} \sum_{t=1}^{t=T} t(L(\theta_t) - L(\theta^*)) = \sum_{t=1}^{t=T} \frac{{\sigma_T}^2 t}{\mu(t+1)} + \sum_{t=1}^{t=T} \frac{\mu t(t-1)}{4} \|\theta_t - \theta^*\|^2 \\ - \sum_{t=1}^{t=T} \frac{\mu t(t+1)}{4} \|\theta_{t+1} - \theta^*\|^2 + \sum_{t=1}^{t=T} Dt \|\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t)\| \end{aligned} \tag{42}$$

$$\begin{aligned} \sum_{t=1}^{t=T} t(L(\theta_t) - L(\theta^*)) \leq \sum_{t=1}^{t=T} \frac{{\sigma_T}^2}{\mu} + \sum_{t=1}^{t=T} \frac{\mu t(t-1)}{4} \|\theta_t - \theta^*\|^2 \\ - \sum_{t=1}^{t=T} \frac{\mu t(t+1)}{4} \|\theta_{t+1} - \theta^*\|^2 + \sum_{t=1}^{t=T} Dt \|\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t)\| \end{aligned} \tag{43}$$

$$\begin{aligned} \sum_{t=1}^{t=T} t(L(\theta_t) - L(\theta^*)) \leq \frac{{\sigma_T}^2 T}{\mu} + \frac{\mu}{4}(0 - T(T+1)\|\theta_{T+1} - \theta^*\|^2) \\ + \sum_{t=1}^{t=T} Dt \|\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t)\| \end{aligned} \tag{44}$$

Since $\frac{\mu}{4}(0 - T(T+1)\|\theta_{T+1} - \theta^*\|^2) \leq 0$, we have:

$$\sum_{t=1}^{t=T} t(L(\theta_t) - L(\theta^*)) \leq \frac{{\sigma_T}^2 T}{\mu} + \sum_{t=1}^{t=T} Dt \|\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t)\| \tag{45}$$

Since, $L(\theta_t) - L(\theta^*) \leq \min_{t=1:T} L(\theta_t) - L(\theta^*)$ and multiplying the above equation by $\frac{2}{T(T+1)}$, we have:

$$\frac{2}{T(T+1)} \sum_{t=1}^{t=T} t(\min_{t=1:T} L(\theta_t) - L(\theta^*)) \leq \frac{2}{T(T+1)} \frac{{\sigma_T}^2 T}{\mu} + \frac{2}{T(T+1)} \sum_{t=1}^{t=T} Dt \|\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t)\| \tag{46}$$

This in turn implies:

$$\min_{t=1:T} L(\theta_t) - L(\theta^*) \leq \frac{{\sigma_T}^2 2}{\mu(T+1)} + \sum_{t=1}^{t=T} \frac{2Dt}{T(T+1)} \|\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t)\| \tag{47}$$

B.2. Convergence Analysis with Stochastic Gradient Descent

Theorem *Denote L_V as the validation loss, L_T as the full training loss, and that the parameters satisfy $||\theta||^2 \leq D^2$. Let L denote either the training or validation loss (with gradient bounded by σ). Any adaptive data selection algorithm, defined via weights $\mathbf{w}^t$ and subsets $\mathcal{X}^t$ for $t = 1, \cdots, T$, and run with a learning rate α using stochastic gradient descent enjoys the following convergence bounds:*

- *if L_T is lipschitz continuous and $\alpha = \frac{D}{\sigma_T\sqrt{T}}$, then $\mathbb{E}\left(\min_{t=1:T} L(\theta_t)\right) - L(\theta^*) \leq \frac{D\sigma_T}{\sqrt{T}} + \frac{D}{T}\sum_{t=1}^{T-1}\mathbb{E}(Err(\mathbf{w}^t, \mathcal{X}^t, L, L_T, \theta_t))$*
- *if L_T is Lipschitz continuous, L is strongly convex with a strong convexity parameter μ, then setting a learning rate $\alpha_t = \frac{2}{\mu(1+t)}$, then $\mathbb{E}\left(\min_{t=1:T} L(\theta_t)\right) - L(\theta^*) \leq \frac{{\sigma_T}^2 2}{\mu(T+1)} + \sum_{t=1}^{t=T}\frac{2D}{T(T+1)}\mathbb{E}\left(tErr(\mathbf{w}^t, \mathcal{X}^t, L, L_T, \theta_t)\right)$*

where:
$Err(\mathbf{w}^t, \mathcal{X}^t, L, L_T, \theta_t) = \left\|\sum_{i\in\mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t) - \nabla_\theta L(\theta_t)\right\|$

PROOF Suppose the gradients of validation loss and training loss are sigma bounded by σ_V and σ_T respectively. Let θ_t be the model parameters at epoch t and θ^* be the optimal model parameters. Let α_t is the learning rate at epoch t.

Let $L_w(\theta_t) = \sum_{i\in\mathcal{X}^t} w_i^t L_T^i(\theta_t)$ be the weighted training loss where $L_T^i(\theta_t)$ is the training loss of the i^{th} instance in the subset $\mathcal{X}^t$.

Let $\nabla_\theta L_w^i(\theta_t)$ be the weighted training loss gradient of the i^{th} instance in the subset $\mathcal{X}^t$, we have:

$$\nabla_\theta L_w^i(\theta_t) = w_i^t \nabla_\theta L_T^i(\theta_t)$$

For a particular θ_t, conditional expectation of $\nabla_\theta L_w^i(\theta_t)$ given $\theta = \theta_t$ over the random choice of i (i.e., randomly selecting i^{th} sample from the subset $\mathcal{X}^t$) yields:

$$\begin{aligned}\mathbb{E}(\nabla_\theta L_w^i(\theta_t) \mid \theta = \theta_t) &= \sum_{i\in\mathcal{X}^t} \nabla_\theta w_i^t L_T^i(\theta_t) \\ &= \nabla_\theta L_w(\theta_T)\end{aligned} \tag{48}$$

In the above equation, $\mathbf{w}^t$ can be assumed as weighted probability distribution as $\sum_{i\in\mathcal{X}^t} w_i^t = 1$.

Similarly conditional expectation of $\nabla_\theta L_w^i(\theta_t)^T(\theta_t - \theta^*)$ given $\theta = \theta_t$ is,

$$\mathbb{E}(\nabla_\theta L_w^i(\theta_t)^T(\theta_t - \theta^*) \mid \theta = \theta_t) = \nabla_\theta L_w(\theta_T)^T(\theta_t - \theta^*) \tag{49}$$

Using the fact that $\theta = \theta_t$ can occur for θ in some finite set Θ (i.e., one element for every choice of samples through out all iterations), we have:

$$\begin{aligned}\mathbb{E}(\nabla_\theta L_w^i(\theta_t)^T(\theta_t - \theta^*)) &= \sum_{\theta_t\in\Theta} \mathbb{E}\left(\nabla_\theta L_w^i(\theta_t)^T(\theta_t - \theta^*)\right)\text{prob}(\theta = \theta_t) \\ &= \sum_{\theta_t\in\Theta} \nabla_\theta L_w(\theta_T)^T(\theta_t - \theta^*)\,\text{prob}(\theta = \theta_t) \\ &= \mathbb{E}\left(\nabla_\theta L_w(\theta_T)^T(\theta_t - \theta^*)\right)\end{aligned} \tag{50}$$

From the definition of stochastic gradient descent, we have:

$$\nabla_\theta L_w^i(\theta_t)^T(\theta_t - \theta^*) = \frac{1}{\alpha_t}(\theta_t - \theta_{t+1})^T(\theta_t - \theta^*) \tag{51}$$

$$\nabla_\theta L_w^i(\theta_t)^T(\theta_t - \theta^*) = \frac{1}{2\alpha_t}\left(\|\theta_t - \theta_{t+1}\|^2 + \|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2\right) \tag{52}$$

$$\nabla_\theta L_w^i(\theta_t)^T(\theta_t - \theta^*) = \frac{1}{2\alpha_t}\left(\left\|\alpha_t \nabla_\theta L_w^i(\theta_t)\right\|^2 + \|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2\right) \tag{53}$$

We can rewrite the function $\nabla_\theta L_w^i(\theta_t)^T(\theta_t - \theta^*)$ as follows:

$$\nabla_\theta L_w^i(\theta_t)^T(\theta_t - \theta^*) = \nabla_\theta L_w^i(\theta_t)^T(\theta_t - \theta^*) - \nabla_\theta L(\theta_t)^T(\theta_t - \theta^*) + \nabla_\theta L(\theta_t)^T(\theta_t - \theta^*) \tag{54}$$

Combining the equations Equation 53 ,Equation 54 we have,

$$\nabla_\theta L_w^i(\theta_t)^T(\theta_t - \theta^*) - \nabla_\theta L(\theta_t)^T(\theta_t - \theta^*) + \nabla_\theta L(\theta_t)^T(\theta_t - \theta^*) = \frac{1}{2\alpha_t}\left(\left\|\alpha_t \nabla_\theta L_w^i(\theta_t)\right\|^2 + \|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2\right) \tag{55}$$

$$\nabla_\theta L(\theta_t)^T(\theta_t - \theta^*) = \frac{1}{2\alpha_t}\left(\left\|\alpha_t \nabla_\theta L_w^i(\theta_t)\right\|^2 + \|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2\right) - \left(\nabla_\theta L_w^i(\theta_t) - \nabla_\theta L(\theta_t)\right)^T(\theta_t - \theta^*) \tag{56}$$

Taking expectation on both sides of the above equation, we have:

$$\begin{aligned}\mathbb{E}(\nabla_\theta L(\theta_t)^T(\theta_t - \theta^*)) = &\frac{1}{2\alpha_t}\mathbb{E}\left(\left\|\alpha_t \nabla_\theta L_w^i(\theta_t)\right\|^2\right) + \frac{1}{2\alpha_t}\mathbb{E}\left(\|\theta_t - \theta^*\|^2\right) \\ &- \frac{1}{2\alpha_t}\mathbb{E}\left(\|\theta_{t+1} - \theta^*\|^2\right) - \mathbb{E}\left(\left(\nabla_\theta L_w^i(\theta_t) - \nabla_\theta L(\theta_t)\right)^T(\theta_t - \theta^*)\right)\end{aligned} \tag{57}$$

From Equation 50, we know that $\mathbb{E}(\nabla_\theta L_w^i(\theta_t)^T(\theta_t - \theta^*)) = \mathbb{E}\left(\nabla_\theta L_w(\theta_T)^T(\theta_t - \theta^*)\right)$. Substituting it in the above equation, we have:

$$\begin{aligned}\mathbb{E}(\nabla_\theta L(\theta_t)^T(\theta_t - \theta^*)) = &\frac{1}{2\alpha_t}\mathbb{E}\left(\left\|\alpha_t \nabla_\theta L_w(\theta_t)\right\|^2\right) + \frac{1}{2\alpha_t}\mathbb{E}\left(\|\theta_t - \theta^*\|^2\right) \\ &- \frac{1}{2\alpha_t}\mathbb{E}\left(\|\theta_{t+1} - \theta^*\|^2\right) - \mathbb{E}\left(\left(\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t)\right)^T(\theta_t - \theta^*)\right)\end{aligned} \tag{58}$$

Case 1 L_T **is lipschitz continuous with parameter** σ_T **(i.e.,** $\|\nabla_\theta L_T(\theta)\| \leq \sigma_T$ **)**

From convexity of function $L(\theta)$, we know $L(\theta_t) - L(\theta^*) \leq \nabla_\theta L(\theta_t)^T(\theta_t - \theta^*)$. Combining this with Equation 58 we have,

$$\begin{aligned}\mathbb{E}(L(\theta_t) - L(\theta^*)) \leq &\frac{1}{2\alpha_t}\mathbb{E}\left(\left\|\alpha_t \nabla_\theta L_w(\theta_t)\right\|^2\right) + \frac{1}{2\alpha_t}\mathbb{E}\left(\|\theta_t - \theta^*\|^2\right) \\ &- \frac{1}{2\alpha_t}\mathbb{E}\left(\|\theta_{t+1} - \theta^*\|^2\right) - \mathbb{E}\left(\left(\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t)\right)^T(\theta_t - \theta^*)\right)\end{aligned} \tag{59}$$

Summing up the above equation from $t = 0 \cdots T-1$, we have:

$$\begin{aligned}\sum_{t=0}^{T-1}\mathbb{E}(L(\theta_t) - L(\theta^*)) \leq &\sum_{t=0}^{T-1}\frac{1}{2\alpha_t}\mathbb{E}\left(\left\|\alpha_t \nabla_\theta L_w(\theta_t)\right\|^2\right) + \frac{1}{2\alpha_t}\sum_{t=0}^{T-1}\mathbb{E}\left(\|\theta_t - \theta^*\|^2\right) \\ &- \frac{1}{2\alpha_t}\sum_{t=0}^{T-1}\mathbb{E}\left(\|\theta_{t+1} - \theta^*\|^2\right) - \sum_{t=0}^{T-1}\mathbb{E}\left(\left(\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t)\right)^T(\theta_t - \theta^*)\right)\end{aligned} \tag{60}$$

Since $\mathbb{E}\left(\|\nabla_\theta L_T(\theta)\|\right) \leq \sigma_T$, we have $\mathbb{E}\left(\|\nabla_\theta L_w(\theta)\|\right) \leq \sigma_T$ as the weights are normalized to 1. Substituting it in the

above equation, we have:

$$\sum_{t=0}^{T-1}\mathbb{E}(L(\theta_t)-L(\theta^*)) \leq \sum_{t=0}^{T-1}\frac{\alpha_t\sigma_T^2}{2}+\frac{1}{2\alpha_t}\sum_{t=0}^{T-1}\mathbb{E}\left(\|\theta_t-\theta^*\|^2\right) - \frac{1}{2\alpha_t}\sum_{t=0}^{T-1}\mathbb{E}\left(\|\theta_{t+1}-\theta^*\|^2\right)-\sum_{t=0}^{T-1}\mathbb{E}\left((\nabla_\theta L_w(\theta_t)-\nabla_\theta L(\theta_t))^T(\theta_t-\theta^*)\right) \tag{61}$$

$$\sum_{t=0}^{T-1}\mathbb{E}(L(\theta_t)-L(\theta^*)) \leq \sum_{t=0}^{T-1}\frac{\alpha_t\sigma_T^2}{2}+\frac{1}{2\alpha_t}\mathbb{E}\left(\|\theta_0-\theta^*\|^2\right) - \frac{1}{2\alpha_t}\mathbb{E}\left(\|\theta_T-\theta^*\|^2\right)-\sum_{t=0}^{T-1}\mathbb{E}\left((\nabla_\theta L_w(\theta_t)-\nabla_\theta L(\theta_t))^T(\theta_t-\theta^*)\right) \tag{62}$$

Since $\mathbb{E}\left(\|\theta_T-\theta^*\|^2\right)\geq 0$, we have:

$$\sum_{t=0}^{T-1}\mathbb{E}(L(\theta_t)-L(\theta^*)) \leq \sum_{t=0}^{T-1}\frac{\alpha_t\sigma_T^2}{2}+\frac{1}{2\alpha_t}\mathbb{E}\left(\|\theta_0-\theta^*\|^2\right)-\sum_{t=0}^{T-1}\mathbb{E}\left((\nabla_\theta L_w(\theta_t)-\nabla_\theta L(\theta_t))^T(\theta_t-\theta^*)\right) \tag{63}$$

Also assuming that $\|\theta-\theta^*\|\leq D$, we have,

$$\sum_{t=0}^{T-1}\mathbb{E}(L(\theta_t)-L(\theta^*)) \leq \frac{\alpha_t T\sigma_T^2}{2}+\frac{D^2}{2\alpha_t}-D\sum_{t=0}^{T-1}\mathbb{E}\left(\|\nabla_\theta L_w(\theta_t)-\nabla_\theta L(\theta_t)\|\right) \tag{64}$$

Choosing a constant learning rate of $\alpha_t=\frac{D}{\sigma_T\sqrt{T}}$, we have:

$$\sum_{t=0}^{T-1}\mathbb{E}(L(\theta_t)-L(\theta^*)) \leq \frac{D\sigma_T\sqrt{T}}{2}+\frac{D\sigma_T\sqrt{T}}{2}-D\sum_{t=0}^{T-1}\mathbb{E}\left(\|\nabla_\theta L_w(\theta_t)-\nabla_\theta L(\theta_t)\|\right) \tag{65}$$

Since, $L(\theta_t)-L(\theta^*)\leq \min_{t=1:T}L(\theta_t)-L(\theta^*)$, we have:

$$\sum_{t=0}^{T-1}\mathbb{E}(\min_{t=1:T}L(\theta_t)-L(\theta^*)) \leq \frac{D\sigma_T\sqrt{T}}{2}+\frac{D\sigma_T\sqrt{T}}{2}-D\sum_{t=0}^{T-1}\mathbb{E}\left(\|\nabla_\theta L_w(\theta_t)-\nabla_\theta L(\theta_t)\|\right) \tag{66}$$

Dividing the above equation by T in the both sides, we have:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}(\min_{t=1:T}L(\theta_t)-L(\theta^*)) \leq \sum_{t=0}^{T-1}\frac{D\sigma_T}{\sqrt{T}}-\frac{D}{T}\sum_{t=0}^{T-1}\mathbb{E}\left(\|\nabla_\theta L_w(\theta_t)-\nabla_\theta L(\theta_t)\|\right) \tag{67}$$

$$\mathbb{E}(\min_{t=1:T}L(\theta_t)-L(\theta^*)) \leq \sum_{t=0}^{T-1}\frac{D\sigma_T}{\sqrt{T}}-\frac{D}{T}\sum_{t=0}^{T-1}\mathbb{E}\left(\|\nabla_\theta L_w(\theta_t)-\nabla_\theta L(\theta_t)\|\right) \tag{68}$$

Substituting $\nabla_\theta L_w(\theta_t)=\sum_{i\in\mathcal{X}^t}w_i^t\nabla_\theta L_T^i(\theta_t)$, we have:

$$\mathbb{E}\left(\min_{t=1:T}L(\theta_t)\right)-L(\theta^*) \leq \sum_{t=0}^{T-1}\frac{D\sigma_T}{\sqrt{T}}-\frac{D}{T}\sum_{t=0}^{T-1}\mathbb{E}\left(\left\|\sum_{i\in\mathcal{X}^t}w_i^t\nabla_\theta L_T^i(\theta_t)-\nabla_\theta L(\theta_t)\right\|\right) \tag{69}$$

Case 2 L_T **is Lipschitz continuous,** L **is strongly convex with a strong convexity parameter** μ

From strong convexity of function $L(\theta)$, we have:

$$\mathbb{E}\left(\nabla_\theta L(\theta_t)^T(\theta_t - \theta^*)\right) \geq \mathbb{E}\left(L(\theta_t) - L(\theta^*)\right) + \frac{\mu}{2}\mathbb{E}\left(\|\theta_t - \theta^*\|\right) \tag{70}$$

Combining the above equation with Equation 58, we have:

$$\begin{aligned}\mathbb{E}\left(L(\theta_t) - L(\theta^*)\right) + \frac{\mu}{2}\mathbb{E}\left(\|\theta_t - \theta^*\|\right) \leq \frac{1}{2\alpha_t}\mathbb{E}\left(\|\alpha_t\nabla_\theta L_w(\theta_t)\|^2\right) + \frac{1}{2\alpha_t}\mathbb{E}\left(\|\theta_t - \theta^*\|^2\right) \\ - \frac{1}{2\alpha_t}\mathbb{E}\left(\|\theta_{t+1} - \theta^*\|^2\right) - \mathbb{E}\left((\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t))^T(\theta_t - \theta^*)\right)\end{aligned} \tag{71}$$

$$\begin{aligned}\mathbb{E}\left(L(\theta_t) - L(\theta^*)\right) \leq \frac{1}{2\alpha_t}\mathbb{E}\left(\|\alpha_t\nabla_\theta L_w(\theta_t)\|^2\right) + \frac{1}{2\alpha_t}\mathbb{E}\left(\|\theta_t - \theta^*\|^2\right) \\ - \frac{1}{2\alpha_t}\mathbb{E}\left(\|\theta_{t+1} - \theta^*\|^2\right) - \mathbb{E}\left((\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t))^T(\theta_t - \theta^*)\right) - \frac{\mu}{2}\mathbb{E}\left(\|\theta_t - \theta^*\|\right)\end{aligned} \tag{72}$$

$$\begin{aligned}\mathbb{E}\left(L(\theta_t) - L(\theta^*)\right) \leq \frac{1}{2\alpha_t}\mathbb{E}\left(\|\alpha_t\nabla_\theta L_w(\theta_t)\|^2\right) + \frac{\alpha_t^{-1} - \mu}{2}\mathbb{E}\left(\|\theta_t - \theta^*\|^2\right) \\ - \frac{\alpha_t^{-1}}{2}\mathbb{E}\left(\|\theta_{t+1} - \theta^*\|^2\right) - \mathbb{E}\left((\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t))^T(\theta_t - \theta^*)\right)\end{aligned} \tag{73}$$

Setting an learning rate of $\alpha_t = \frac{2}{\mu(t+1)}$, multiplying by t on both sides and substituting $\nabla_\theta L_w(\theta_t) = \sum_{i\in\mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t)$, we have:

$$\begin{aligned}\mathbb{E}\left(t(L(\theta_t) - L(\theta^*))\right) = \mathbb{E}\left(\frac{t}{\mu(t+1)}\left\|\sum_{i\in\mathcal{X}^t} w_i^t\nabla_\theta L_T^i(\theta_t)\right\|^2\right) + \mathbb{E}\left(\frac{\mu t(t-1)}{4}\|\theta_t - \theta^*\|^2\right) \\ - \mathbb{E}\left(\frac{\mu t(t+1)}{4}\|\theta_{t+1} - \theta^*\|^2\right) - \mathbb{E}\left(t(\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t))^T(\theta_t - \theta^*)\right)\end{aligned} \tag{74}$$

Since, $\|L_T(\theta)\| \leq \sigma_T$, we have $\left\|\alpha\sum_{i\in\mathcal{X}^t} w_i^t\nabla_\theta L_T^i(\theta_t)\right\| \leq \sum_{i=1}^{|\mathcal{X}^t|} w_i^t\sigma_T$. Assuming that the weights at every iteration are normalized such that $\underset{t\in[1,T]}{\forall}\sum_{i=1}^{|\mathcal{X}^t|} w_i^t = 1$ and the training and validation loss gradients are normalized as well, we have $\left\|\alpha\sum_{i\in\mathcal{X}^t} w_i^t\nabla_\theta L_T^i(\theta_t)\right\| \leq \sigma_T$. Also assuming that $\|\theta - \theta^*\| \leq D$, we have,

$$\begin{aligned}\mathbb{E}\left(t(L(\theta_t) - L(\theta^*))\right) = \frac{{\sigma_T}^2 t}{\mu(t+1)} + \frac{\mu t(t-1)}{4}\mathbb{E}\left(\|\theta_t - \theta^*\|^2\right) \\ - \frac{\mu t(t+1)}{4}\mathbb{E}\left(\|\theta_{t+1} - \theta^*\|^2\right) + \mathbb{E}\left(Dt\,\|\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t)\|\right)\end{aligned} \tag{75}$$

Summing the above equation from $t = 1, \cdots, T$, we have:

$$\begin{aligned}\sum_{t=1}^{t=T}\mathbb{E}\left(t(L(\theta_t) - L(\theta^*))\right) = \sum_{t=1}^{t=T}\frac{{\sigma_T}^2 t}{\mu(t+1)} + \sum_{t=1}^{t=T}\frac{\mu t(t-1)}{4}\mathbb{E}\left(\|\theta_t - \theta^*\|^2\right) \\ - \sum_{t=1}^{t=T}\frac{\mu t(t+1)}{4}\mathbb{E}\left(\|\theta_{t+1} - \theta^*\|^2\right) + \sum_{t=1}^{t=T} D\,\mathbb{E}\left(t\,\|\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t)\|\right)\end{aligned} \tag{76}$$

$$\sum_{t=1}^{t=T} \mathbb{E}\left(t(L(\theta_t) - L(\theta^*))\right) \leq \sum_{t=1}^{t=T} \frac{{\sigma_T}^2}{\mu} + \sum_{t=1}^{t=T} \frac{\mu t(t-1)}{4} \mathbb{E}\left(\|\theta_t - \theta^*\|^2\right) - \sum_{t=1}^{t=T} \frac{\mu t(t+1)}{4} \mathbb{E}\left(\|\theta_{t+1} - \theta^*\|^2\right) + \sum_{t=1}^{t=T} D\,\mathbb{E}\left(t\,\|\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t)\|\right) \tag{77}$$

$$\sum_{t=1}^{t=T} \mathbb{E}\left(t(L(\theta_t) - L(\theta^*))\right) \leq \frac{{\sigma_T}^2 T}{\mu} + \frac{\mu}{4}(0 - T(T+1)\,\mathbb{E}\left(\|\theta_{T+1} - \theta^*\|^2\right)) + \sum_{t=1}^{t=T} D\,\mathbb{E}\left(t\,\|\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t)\|\right) \tag{78}$$

Since $\frac{\mu}{4}(0 - T(T+1)\,\mathbb{E}\left(\|\theta_{T+1} - \theta^*\|^2\right)) \leq 0$, we have:

$$\sum_{t=1}^{t=T} \mathbb{E}\left(t(L(\theta_t) - L(\theta^*))\right) \leq \frac{{\sigma_T}^2 T}{\mu} + \sum_{t=1}^{t=T} D\,\mathbb{E}\left(t\,\|\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t)\|\right) \tag{79}$$

Since, $L(\theta_t) - L(\theta^*) \leq \min_{t=1:T} L(\theta_t) - L(\theta^*)$ and multiplying the above equation by $\frac{2}{T(T+1)}$, we have:

$$\frac{2}{T(T+1)}\,\mathbb{E}\left(\sum_{t=1}^{t=T} t(\min_{t=1:T} L(\theta_t) - L(\theta^*))\right) \leq \frac{2}{T(T+1)} \frac{{\sigma_T}^2 T}{\mu} + \frac{2}{T(T+1)} \sum_{t=1}^{t=T} \mathbb{E}\left(Dt\,\|\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t)\|\right) \tag{80}$$

This in turn implies:

$$\mathbb{E}\left(\min_{t=1:T} L(\theta_t)\right) - L(\theta^*) \leq \frac{{\sigma_T}^2 2}{\mu(T+1)} + \sum_{t=1}^{t=T} \frac{2D}{T(T+1)}\,\mathbb{E}\left(t\,\|\nabla_\theta L_w(\theta_t) - \nabla_\theta L(\theta_t)\|\right) \tag{81}$$

B.3. Conditions for adaptive data selection algorithms to reduce the objective value at every iteration

We provide conditions under which the adaptive subset selection strategy reduces the objective value of L (which can either be the training loss L_T or the validation loss L_V):

Theorem 4 *If the Loss Function L is Lipschitz smooth with parameter $\mathcal{L}$, and the gradient of the training loss is bounded by σ_T, the adaptive data selection algorithm will reduce the objective function at every iteration, i.e. $L(\theta_{t+1}) \leq L(\theta_t)$ as long as $(\sum_{i\in\mathcal{X}} w_i \nabla_\theta L_T^i(\theta))^T \nabla_\theta L(\theta) \geq 0$ and the learning rate schedule satisfies $\alpha_t \leq \min_t 2\frac{\|\nabla_\theta L(\theta)\|\cos(\theta_t)}{\mathcal{L}\sigma_T}$, where θ_t is the angle between $\sum_{i\in\mathcal{X}} w_i \nabla_\theta L_T^i(\theta)$ and $\nabla_\theta L(\theta)$.*

Before proving this result, notice that any data selection approach that attempts to minimize the error term $\text{Err}(\mathbf{w}^t, \mathcal{X}^t, L, L_T, \theta_t) = \|\sum_{i\in\mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta_t) - \nabla_\theta L(\theta_t)\|$, will essentially also maximize $\left(\sum_{i\in\mathcal{X}^t} w_i^t \nabla_\theta L_T^i(\theta)\right)^T \nabla_\theta L(\theta)$. Hence we expect the condition above to be satisfied, as long as the learning rate can be selected appropriately.

PROOF Suppose we have a validation set $\mathcal{V}$ and the loss on the validation set or training set is denoted as $L(\theta)$ depending on the usage. Suppose the subset selected by the GRAD-MATCH is denoted by S and the subset training loss is $L_T(\theta, \mathcal{X})$. Since validation or training loss L is lipschitz smooth, we have,

$$L(\theta_{t+1}) \leq L(\theta_t) + \frac{\mathcal{L}\|\Delta\theta\|^2}{2} + \nabla_\theta L(\theta_t)^T \Delta\theta, \quad \text{where, } \Delta\theta = \theta_{t+1} - \theta_t \tag{82}$$

Since, we are using SGD to optimize the subset training loss $L_T(\theta, S)$ model parameters our update equations will be as follows:

$$\theta_{t+1} = \theta_t - \alpha \sum_{i \in \mathcal{X}} w_i \nabla_\theta L_T^i(\theta_t) \tag{83}$$

Plugging our updating rule (Equation 83) in (Equation82):

$$L(\theta_{t+1}) \leq L(\theta_t) - \alpha \nabla_\theta L(\theta_t)^T \sum_{i \in \mathcal{X}} w_i \nabla_\theta L_T^i(\theta)) + \frac{\mathcal{L}}{2} \left\| -\alpha \sum_{i \in \mathcal{X}} w_i \nabla_\theta L_T^i(\theta) \right\|^2 \tag{84}$$

Which gives,

$$L(\theta_{t+1}) - L(\theta_t) \leq -\alpha \nabla_\theta L(\theta_t)^T \sum_{i \in \mathcal{X}} w_i \nabla_\theta L_T^i(\theta) + \frac{\mathcal{L}\alpha^2}{2} \left\| \sum_{i \in \mathcal{X}} w_i \nabla_\theta L_T^i(\theta) \right\|^2 \tag{85}$$

From (Equation 85), note that:

$$L(\theta_{t+1}) \leq L(\theta_l) \;\text{ if }\; \alpha \left(\left(\sum_{i \in \mathcal{X}} w_i \nabla_\theta L_T^i(\theta) \right)^T \nabla_\theta L(\theta_t) - \frac{\mathcal{L}\alpha}{2} \left\| \sum_{i \in \mathcal{X}} w_i \nabla_\theta L_T^i(\theta) \right\|^2 \right) \geq 0 \tag{86}$$

Since we know that $\left\| \sum_{i \in \mathcal{X}} w_i \nabla_\theta L_T^i(\theta) \right\|^2 \geq 0$, we will have the necessary condition:

$$\left(\sum_{i \in \mathcal{X}} w_i \nabla_\theta L_T^i(\theta) \right)^T \nabla_\theta L(\theta_t) \geq 0$$

We can also re-write the condition in (Equation86) as follows:

$$\left(\sum_{i \in \mathcal{X}} w_i \nabla_\theta L_T^i(\theta) \right)^T \nabla_\theta L(\theta_t) \geq \frac{\alpha \mathcal{L}}{2} \left(\sum_{i \in \mathcal{X}} w_i \nabla_\theta L_T^i(\theta) \right)^T \left(\sum_{i \in \mathcal{X}} w_i \nabla_\theta L_T^i(\theta) \right) \tag{87}$$

The Equation 87 gives the necessary condition for learning rate i.e.,

$$\alpha \leq \frac{2}{\mathcal{L}} \frac{\left(\sum_{i \in \mathcal{X}} w_i \nabla_\theta L_T^i(\theta) \right)^T \nabla_\theta L(\theta_t)}{\left(\sum_{i \in \mathcal{X}} w_i \nabla_\theta L_T^i(\theta) \right)^T \left(\sum_{i \in \mathcal{X}} w_i \nabla_\theta L_T^i(\theta) \right)} \tag{88}$$

The above Equation 88 can be written as follows:

$$\begin{aligned} \alpha &\leq \frac{2}{\mathcal{L}} \frac{\|\nabla_\theta L(\theta_t)\| \cos(\Theta_t)}{\left\| \sum_{i \in \mathcal{X}} w_i \nabla_\theta L_T^i(\theta) \right\|} \\ \text{where} \quad \cos\Theta_t &= \frac{\left(\sum_{i \in \mathcal{X}} w_i \nabla_\theta L_T^i(\theta) \right)^T \nabla_\theta L(\theta_t)}{\left\| \sum_{i \in \mathcal{X}} w_i \nabla_\theta L_T^i(\theta) \right\| \|\nabla_\theta L(\theta_t)\|} \end{aligned} \tag{89}$$

Assuming we normalize the subset weights at every iteration i.e., $\forall_t \sum_{i \in [1, |\mathcal{X}|} w_i^t = 1$, we know that the gradient norm $\left\| \sum_{i \in \mathcal{X}} w_i \nabla_\theta L_T^i(\theta) \right\| \leq \sigma_T$, the condition for the learning rate can be written as follows,

$$\alpha \leq \frac{2 \|\nabla_\theta L(\theta_t)\| \cos(\Theta_t)}{\mathcal{L}\sigma_T} \quad \text{where} \quad \cos\Theta_t = \frac{\left(\sum_{i \in \mathcal{X}} w_i \nabla_\theta L_T^i(\theta) \right)^T \nabla_\theta L(\theta_t)}{\left\| \sum_{i \in \mathcal{X}} w_i \nabla_\theta L_T^i(\theta) \right\| \|\nabla_\theta L(\theta_t)\|} \tag{90}$$

Since, the condition mentioned in Equation 90 needs to be true for all values of l, we have the condition for learning rate as follows:

$$\alpha \leq \min_t \frac{2\left\|\nabla_\theta L(\theta_t)\right\| \cos(\Theta_t)}{\mathcal{L}\sigma_T}$$
$$\text{where} \quad \cos\Theta_t = \frac{\left(\sum_{i\in\mathcal{X}} w_i \nabla_\theta L_T^i(\theta)\right)^T \nabla_\theta L(\theta_t)}{\left\|\sum_{i\in\mathcal{X}} w_i \nabla_\theta L_T^i(\theta)\right\| \left\|\nabla_\theta L(\theta_t)\right\|} \tag{91}$$

B.4. Proof of Theorem 2

We first restate Theorem 2

Theorem *If* $|\mathcal{X}| \leq k,\ \max_i ||\nabla_\theta L_T^i(\theta_t)||_2 < \nabla_{\max}$, *then* $F_\lambda(\mathcal{X})$ *is* γ*-weakly submodular, with* $\gamma \geq \frac{\lambda}{\lambda + k\nabla_{\max}^2}$

PROOF We first note that the minimum eigenvalue of $F_\lambda(\mathcal{X})$ is atleast λ. Next, we note that the maximum eigenvalue of $F_\lambda(\mathcal{X})$ is atmost

$$\lambda + \text{Trace}(F(\mathcal{X})) \tag{92}$$
$$= \lambda + \text{Trace}\left(\begin{bmatrix} \nabla L_T^{1\top}(\theta_t) \\ \nabla L_T^{2\top}(\theta_t) \\ \dots \\ \nabla L_T^{k\top}(\theta_t) \end{bmatrix} \begin{bmatrix} \nabla L_T^{1\top}(\theta_t) \\ \nabla L_T^{2\top}(\theta_t) \\ \dots \\ \nabla L_T^{k\top}(\theta_t) \end{bmatrix}^\top\right) \tag{93}$$
$$= \lambda + \sum_{i\in[k]} ||\nabla L_T^i(\theta_t)||^2 \tag{94}$$

which immediately proves the theorem following (Elenberg et al., 2018).

B.5. Proof of Theorem 3

We start this subsection by first restating Theorem 3.

Theorem *If the function* $F_\lambda(\mathcal{X})$ *is* γ*-weakly submodular,* $\mathcal{X}^*$ *is the optimal subset and* $\max_i ||\nabla_\theta L_T^i(\theta_t)||_2 < \nabla_{\max}$, *(both) the greedy algorithm and OMP (Algorithm 2), run with stopping criteria* $E_\lambda(\mathcal{X}) \leq \epsilon$ *achieve sets* $\mathcal{X}$ *such that* $|\mathcal{X}| \leq \frac{|\mathcal{X}^*|}{\gamma}\log(\frac{L_{\max}}{\epsilon})$ *where* $L_{\max}$ *is an upper bound of* F_λ .

PROOF From Theorem 3, we know that $F_\lambda(\mathcal{X})$ is weakly submodular with parameter $\gamma = \frac{\lambda}{\lambda + k\nabla_{\max}^2}$.

We first prove the result using the greedy algorithm, or in particular the submodular set cover algorithm (Wolsey, 1982). Note that an upper bound of F_λ is $L_{\max}$, and consider the stopping criteria of the greedy algorithm to be achieving a subset $\mathcal{X}$ such that $F_\lambda(\mathcal{X}) \geq L_{\max} - \epsilon$. The goal is then to bound the $|\mathcal{X}|$ of the subset $\mathcal{X}$ achieving it compared to the optimal subset $\mathcal{X}^*$.

Given a set X_i which is obtained at step i of the greedy algorithm, and denote e_i to be the best gain at step i. Note that:

$$\gamma(F_\lambda(\mathcal{X}^* \cup X_i) - F_\lambda(X_i)) \leq \sum_{j\in\mathcal{X}^*} F_\lambda(j|X_i)$$
$$\leq |\mathcal{X}^*| F_\lambda(e_i|X_i) \tag{95}$$

where the last inequality holds because of the greedy algorithm. This then implies that:

$$F_\lambda(\mathcal{X}^*) - F_\lambda(X_i) \leq \frac{|\mathcal{X}^*|}{\gamma}(F_\lambda(X_{i+1}) - F_\lambda(X_i)) \tag{96}$$

We modify the second term to be $(F_\lambda(X_{i+1}) - F_\lambda(X_i)) = (F_\lambda(\mathcal{X}^*) - F_\lambda(X_i)) - (F_\lambda(\mathcal{X}^*) - F_\lambda(X_{i+1}))$ and then obtain the following recursion:

$$F_\lambda(\mathcal{X}^*) - F_\lambda(X_{i+1}) \leq (1 - \frac{\gamma}{|\mathcal{X}^*|})(F_\lambda(\mathcal{X}^*) - F_\lambda(X_i)) \tag{97}$$

We can then recursively multiply the right hand and the left hand sides, until we reach a set $\mathcal{X}$ such that $F_\lambda(\mathcal{X}) \geq L_{\max} - \epsilon$ (which is the stopping criteria). We then achieve:

$$F_\lambda(\mathcal{X}^*) - F_\lambda(\mathcal{X}) \leq (1 - \gamma/|\mathcal{X}^*|)^{|\mathcal{X}|}(F_\lambda(\mathcal{X}^*) - F_\lambda(\emptyset)) \leq (1 - \gamma/|\mathcal{X}^*|)^{|\mathcal{X}|}F_\lambda(\mathcal{X}^*) \leq (1 - \gamma/|\mathcal{X}^*|)^{|\mathcal{X}|}L_{\max} \quad (98)$$

where the second-last inequality holds since $F_\lambda(\emptyset) \geq 0$, and the last inequality holds because $F_\lambda(\mathcal{X}^*) \leq L_{\max}$.

This implies that we have the following inequality: $F_\lambda(\mathcal{X}^*) - F_\lambda(\mathcal{X}) \leq (1 - \gamma/|\mathcal{X}^*|)^{|\mathcal{X}|}L_{\max}$. Next, notice that since $F_\lambda(\mathcal{X}^*) \leq L_{\max}$, which in turn implies that $F_\lambda(\mathcal{X}^*) - F_\lambda(\mathcal{X}) \leq \epsilon$. Hence, we can pick a $\mathcal{X}$ such that $(1 - \gamma/|\mathcal{X}^*|)^{|\mathcal{X}|}L_{\max} \leq \epsilon$, which will then automatically imply that $F_\lambda(\mathcal{X}^*) - F_\lambda(\mathcal{X}) \leq \epsilon$. The above condition requires $(1 - \gamma/|\mathcal{X}^*|)^{|\mathcal{X}|} \geq \epsilon$, which in turn implies that $|\mathcal{X}| \leq |\mathcal{X}^*|/\gamma \log L_{\max}/\epsilon$. This shows the result for the standard greedy algorithm.

Finally, we prove it for the OMP case. In particular, from Lemma 4 and the proof of Theorem 5 in (Elenberg et al., 2018), we can obtain a recursion very similar to Equation 96, except that we have the ratio of the m and M corresponding to strong concavity and smoothness respectively. From the proof of Theorem 2, this is exactly the bound used for weak submodularity of F_λ, and hence the bound follows for OMP as well.

B.6. Convergence result for GRAD-MATCH using the OMP algorithm

The following result shows the convergence bound of GRAD-MATCH using OMP as the optimization algorithm.

Lemma 1 *Suppose the subsets $\mathcal{X}^t$ satisfy the condition that $E_\lambda(\mathcal{X}^t) \leq \epsilon$, for all $t = 1, \cdots, T$, then OMP based data selection achieves the following convergence result:*

- *if L_T is lipschitz continuous with parameter σ_T and $\alpha = \frac{D}{\sigma_T\sqrt{T}}$, then $\min_{t=1:T} L(\theta_t) - L(\theta^*) \leq \frac{D\sigma_T}{\sqrt{T}} + D\epsilon$,*
- *if L_T is lipschitz smooth with parameter $\mathcal{L}_T$, and L_T^i satisfies $0 \leq L_T^i(\theta) \leq \beta_T, \forall i$. Then setting $\alpha = \frac{1}{\mathcal{L}_T}$, we have $\min_{t=1:T} L(\theta_t) - L(\theta^*) \leq \frac{D^2\mathcal{L}_T + 2\beta_T}{2T} + D\epsilon$,*
- *if L_T is Lipschitz continuous (parameter σ_T) and L is strongly convex with parameter μ, then setting a learning rate $\alpha_t = \frac{2}{\mu(1+t)}$, achieves $\min_{t=1:T} L(\theta_t) - L(\theta^*) \leq \frac{2{\sigma_T}^2}{\mu(T+1)} + D\epsilon$*

PROOF We prove the first part and note that the other parts follow similarly. Notice that the stopping criteria of Algorithm 1 is $E_\lambda(\mathcal{X}^t) \leq \epsilon$. Denote w_t as the corresponding weight vector, and hence we have $E_\lambda(\mathcal{X}^t) = E_\lambda(\mathcal{X}^t, w_t) = E(\mathcal{X}^t, w_t) + \lambda||w_t||^2 \leq \epsilon$, where $E(\mathcal{X}^t, w_t) = \text{Err}(w_t, \mathcal{X}^t, L, L_T, \theta_t)$. Since $||w_t||^2 \geq 0$, this implies that $\text{Err}(w_t, \mathcal{X}^t, L, L_T, \theta_t) \leq \epsilon$, which combining with Theorem 1, immediately provides the required convergence result for OMP. Finally, note that for the third part, $\sum_{t=1}^{t=T} \frac{2Dt}{T(T+1)}\epsilon = \epsilon$ and this proves all three parts.

B.7. More details on CRAIG

B.7.1. CONNECTIONS BETWEEN GRAD-MATCH AND CRAIG

Lemma 2 *The following inequality connects $\hat{E}(\mathcal{X})$ and $E(\mathcal{X})$*

$$E(\mathcal{X}) = \min_{\mathbf{w}} Err(\mathbf{w}, \mathcal{X}, L, L_T, \theta_t) \leq \hat{E}(\mathcal{X}) \quad (99)$$

Furthermore, given the set $\mathcal{X}^t$ obtained by optimizing $\hat{E}$, the weights can be computed as: $\mathbf{w}^t = \sum_{i \in \mathcal{W}} \mathbb{I}[j = \arg\min_{s \in \mathcal{X}^t} \|\nabla_\theta L_T^i(\theta_t) - \nabla_\theta L^s(\theta_t)\|]$.

PROOF During iteration $t \in 1, \cdots, T$, we partition W by assigning every element $i \in W$ to an element $\pi_t^i \in \mathcal{X}$ as follows:

$$\pi_t^i \in \arg\min_{j \in \mathcal{X}} \|\nabla_\theta L^i(\theta) - \nabla_\theta L_T^j(\theta)\| \quad (100)$$

In other words, π_t^i denotes the representative for a specific $i \in W$ in set $\mathcal{X}$. Also recall that, $\hat{E}(\mathcal{X})$ is defined as follows:

$$\begin{aligned}\hat{E}(\mathcal{X}) &= \sum_{i \in W} \min_{j \in \mathcal{X}} \|\nabla_\theta L^i(\theta_t) - \nabla_\theta L_T^j(\theta_t)\| \\ &= \sum_{i \in W} \|\nabla_\theta L^i(\theta_t) - \nabla_\theta L_T^{\pi_t^i}(\theta_t)\| \end{aligned} \quad (101)$$

Then, for any θ_t we can write

$$\nabla_\theta L(\theta_t) = \sum_{i \in W} \left(\nabla_\theta L_T^i(\theta_t) - \nabla_\theta L^{\pi_t^i}(\theta_t) + \nabla_\theta L^{\pi_t^i}(\theta_t)\right)$$
$$= \sum_{i \in W} \left(\nabla_\theta L_T^i(\theta_t) - \nabla_\theta L^{\pi_t^i}(\theta_t)\right) + \sum_{j \in \mathcal{X}} w_j \nabla_\theta L^j(\theta_t)$$

where w_j denotes the count of number of $i \in W$ that were assigned to an element $j \in \mathcal{X}$. Subtracting the second term on the RHS, *viz.*, $\sum_{j \in \mathcal{X}} w_j \nabla_\theta L^j(\theta_t)$ from the LHS and then taking the norm of the both sides, we get the following upper bound on the error of estimating the full gradient $\text{Err}(\mathbf{w}, \mathcal{X}, L, L_T, \theta_t)$:

$$\text{Err}(\mathbf{w}, \mathcal{X}, L, L_T, \theta_t) = \left\|\nabla L(\theta_t) - \sum_{i \in \mathcal{X}} w_i \nabla L_T^i(\theta_t)\right\|$$
$$= \left\| \sum_{i \in W} \left(\nabla_\theta L_T^i(\theta_t) - \nabla_\theta L^{\pi_t^i}(\theta_t)\right)\right\|$$
$$\leq \sum_{i \in W} \|\nabla_\theta L_T^i(\theta_t) - \nabla_\theta L^{\pi_t^i}(\theta_t)\|,$$

where the inequality follows from the triangle inequality.

With $\pi_t^i \in \arg\min_{j \in \mathcal{X}} \|\nabla_\theta L^i(\theta) - \nabla_\theta L_T^j(\theta)\|$, the upper bound exactly equals the function $\hat{E}(\mathcal{X})$ defined below:

$$\hat{E}(\mathcal{X}) = \sum_{i \in W} \min_{j \in \mathcal{X}} \|\nabla_\theta L^i(\theta_t) - \nabla_\theta L_T^j(\theta_t)\|$$
$$= \sum_{i \in W} \|\nabla_\theta L^i(\theta_t) - \nabla_\theta L_T^{\pi_t^i}(\theta_t)\| \tag{102}$$

Hence it follows that $\hat{E}(\mathcal{X})$ is an upper bound of $E(\mathcal{X})$.

B.7.2. MAXIMIZATION VERSION OF CRAIG

We can similarly formulate the maximization version of this problem. Define:

$$\hat{F}(\mathcal{X}) = \sum_{i \in \mathcal{W}} L_{\max} - \min_{j \in \mathcal{X}} \|\nabla_\theta L^i(\theta_t) - \nabla_\theta L_T^j(\theta_t)\|$$
$$= \sum_{i \in \mathcal{W}} \max_{j \in \mathcal{X}} \left(L_{\max} - \|\nabla_\theta L^i(\theta_t) - \nabla_\theta L_T^j(\theta_t)\|\right)$$

Note that this function is exactly the Facility Location function considered in CRAIG (Mirzasoleiman et al., 2020a), and $\hat{F}(\mathcal{X})$ is a lower bound of $F(\mathcal{X})$. Maximizing the above expression under the constraint $|\mathcal{X}| \leq k$ is an instance of cardinality constraint submodular maximization, and a simple greedy algorithm achieves a $1 - 1/e$ approximation (Nemhauser et al., 1978).

Next, we look at the dual problem, *i.e.*, finding the minimum set size such that the error is bounded. Through the following minimization problem we obtain the smallest weighted subset $\mathcal{X}$ that approximates the full gradient by an error of at most ϵ for the current parameters θ_t:

$$\mathcal{X}^t = \min_{\mathcal{X}} |\mathcal{X}|, \text{ so that, } \hat{E}(\mathcal{X}) \leq \epsilon. \tag{103}$$

We can rewrite Equation 103 as an instance of submodular set cover:

$$\mathcal{X}^t = \min_{\mathcal{X}} |\mathcal{X}|, \text{s.t. } \hat{F}(\mathcal{X}) \geq |\mathcal{W}| L_{\max} - \epsilon. \tag{104}$$

This is an instance of submodular set cover, which can also be approximated up to a log-factor (Wolsey, 1982; Mirzasoleiman et al., 2015). In particular, denote $Q = \hat{F}(\mathcal{X}^*)$ as the optimal solution. Then the greedy algorithm is guaranteed to obtain a set $\hat{\mathcal{X}}$ such that $\hat{F}(\hat{\mathcal{X}}) \geq |\mathcal{W}| L_{\max} - \epsilon$ and $|\hat{\mathcal{X}}| \leq |\mathcal{X}^*| \log(Q/\epsilon)$. Next, note that obtaining a set $\hat{\mathcal{X}}$ such that $\hat{F}(\hat{\mathcal{X}}) \geq |\mathcal{W}| L_{\max} - \epsilon$ is equivalent to $\hat{E}(\mathcal{X}) \leq \epsilon$. Using this fact, and the convergence result of Theorem 1, we can derive convergence bounds for CRAIG. In particular, assume that using the submodular set cover, we achieve sets $\mathcal{X}^t$ such that $\hat{E}(\mathcal{X}^t) \leq \epsilon$. The following corollary provides a convergence result for the facility location based upper bound approach.

B.7.3. CONVERGENCE BOUND FOR CRAIG

Next, we state and prove a convergence bound for CRAIG.

Lemma 3 *Suppose the subsets $\mathcal{X}^t$ satisfy $\hat{E}(\mathcal{X}^t) \leq \epsilon, \forall t = 1, \cdots, T$, then using the facility location upper bound for data selection achieves the following convergence result:*

- *if L_T is lipschitz continuous with parameter σ_T and $\alpha = \frac{D}{\sigma_T \sqrt{T}}$, then $\min_{t=1:T} L(\theta_t) - L(\theta^*) \leq \frac{D\sigma_T}{\sqrt{T}} + D\epsilon$,*
- *if L_T is lipschitz smooth with parameter $\mathcal{L}_T$, and L_T^i satisfies $0 \leq L_T^i(\theta) \leq \beta_T, \forall i$. Then setting $\alpha = \frac{1}{\mathcal{L}_T}$, we have $\min_{t=1:T} L(\theta_t) - L(\theta^*) \leq \frac{D^2\mathcal{L}_T + 2\beta_T}{2T} + D\epsilon$,*
- *if L_T is Lipschitz continuous (parameter σ_T) and L is strongly convex with parameter μ, then setting a learning rate $\alpha_t = \frac{2}{\mu(1+t)}$, achieves $\min_{t=1:T} L(\theta_t) - L(\theta^*) \leq \frac{2{\sigma_T}^2}{\mu(T+1)} + D\epsilon$*

PROOF We prove the first part and note that the other parts follow similarly. Notice that the CRAIG algorithm tries to minimze the term $\hat{E}(\mathcal{X}^t) = \sum_{i \in W} \min_{j \in \mathcal{X}^t} \|\nabla_\theta L^i(\theta_t) - \nabla_\theta L_T^j(\theta_t)\|$ which is an upper bound of $E(\mathcal{X}^t) = \min_{\mathbf{w}} \text{Err}(\mathbf{w}^t, \mathcal{X}^t, L, L_T, \theta_t)$ from Lemma B.7 (i.e., $\text{Err}(\mathbf{w}^t, \mathcal{X}^t, L, L_T, \theta_t) \leq \hat{E}(\mathcal{X}^t)$). From the assumption that $\hat{E}(\mathcal{X}^t) \leq \epsilon$, we have $\text{Err}(\mathbf{w}^t, \mathcal{X}^t, L, L_T, \theta_t) \leq \epsilon$, which combining with Theorem 1, immediately proves the required convergence result for CRAIG. Finally, note that for the third part, $\sum_{t=1}^{t=T} \frac{2Dt}{T(T+1)}\epsilon = \epsilon$ and this proves all three parts.

C. More Experimental Details and Additional Results

C.1. Datasets Description

We used various standard datasets, namely, MNIST, CIFAR10, SVHN, CIFAR100, ImageNet, to demonstrate the effectiveness and stability of GRAD-MATCH.

Name	No. of classes	No. samples for training	No. samples for validation	No. samples for testing	No. of features
CIFAR10	10	50,000	-	10,000	32x32x3
MNIST	10	60,000	-	10,000	28x28
SVHN	10	73,257	-	26,032	32x32x3
CIFAR100	100	50,000	-	10,000	32x32x3
ImageNet	1000	1,281,167	50,000	100,000	224x224x3

Table 2. Description of the datasets

Table 2 gives a brief description about the datasets. Here not all datasets have an explicit validation and test set. For such datasets, 10% and 20% samples from the training set are used as validation and test set, respectively. The feature count given for the ImageNet dataset is after applying the `RandomResizedCrop` transformation function from PyTorch (Paszke et al., 2017).

C.2. Experimental Settings

We ran experiments using an SGD optimizer with an initial learning rate of 0.01, the momentum of 0.9, and a weight decay of 5e-4. We decay the learning rate using cosine annealing (Loshchilov & Hutter, 2017) for each epoch. For MNIST, we use the LeNet model (LeCun et al., 1989) and train the model for 200 epochs. For all other datasets, we use ResNet18 model (He et al., 2016) and train the model for 300 epochs (except for ImageNet, where we train the model for 350 epochs).

To demonstrate our method's effectiveness in a robust learning setting, we artificially generate class-imbalance for the above datasets by removing almost 90% of the instances from 30% of total classes available. We ran all experiments on a single V100 GPU, except for ImageNet, where we used an RTX 2080 GPU. However, for a given dataset, all experiments were run on the same GPU so that the speedup and energy comparison across techniques is fair.

C.3. Other specific settings

Here we discuss various parameters' required by Algorithm 2, their significance, and the values used in the experiments.

Data Selection Results										
			Top-1 Test accuracy of the Model(%)				Model Training time(in hrs)			
		Budget(%)	5%	10%	20%	30%	5%	10%	20%	30%
Dataset	Model	Selection Strategy								
CIFAR10	ResNet18	FULL (skyline for test accuracy)	95.09	95.09	95.09	95.09	4.34	4.34	4.34	4.34
		RANDOM (skyline for training time)	71.2	80.8	86.98	87.6	0.22	0.46	0.92	1.38
		RANDOM-WARM (skyline for training time)	83.2	87.8	90.9	92.6	0.21	0.42	0.915	1.376
		GLISTER	85.5	91.92	92.78	93.63	0.43	0.91	1.13	1.46
		GLISTER-WARM	86.57	91.56	92.98	94.09	0.42	0.88	1.08	1.40
		CRAIG	82.74	87.49	90.79	92.53	0.81	1.08	1.45	2.399
		CRAIG-WARM	84.48	89.28	92.01	92.82	0.6636	0.91	1.31	2.20
		CRAIGPB	83.56	88.77	92.24	93.58	0.4466	0.70	1.13	2.07
		CRAIGPB-WARM	86.28	90.07	93.06	93.8	0.4143	0.647	1.07	2.06
		GRADMATCH	86.7	90.9	91.67	91.89	0.40	0.84	1.42	1.52
		GRADMATCH-WARM	87.2	92.15	92.11	92.01	0.38	0.73	1.24	1.41
		GRADMATCHPB	85.4	90.01	93.34	93.75	0.36	0.69	1.09	1.38
		GRADMATCHPB-WARM	86.37	92.26	93.59	94.17	0.32	0.62	1.05	1.36
CIFAR100	ResNet18	FULL (skyline for test accuracy)	75.37	75.37	75.37	75.37	4.871	4.871	4.871	4.871
		RANDOM (skyline for training time)	19.02	31.56	49.6	58.56	0.2475	0.4699	0.92	1.453
		RANDOM-WARM (skyline for training time)	58.2	65.95	70.3	72.4	0.242	0.468	0.921	1.43
		GLISTER	29.94	44.03	61.56	70.49	0.3536	0.6456	1.11	1.5255
		GLISTER-WARM	57.17	64.95	62.14	72.43	0.3185	0.6059	1.06	1.452
		CRAIG	36.61	55.19	66.24	70.01	1.354	1.785	1.91	2.654
		CRAIG-WARM	57.44	67.3	69.76	72.77	1.09	1.48	1.81	2.4112
		CRAIGPB	38.95	54.59	67.12	70.61	0.4489	0.6564	1.15	1.540
		CRAIGPB-WARM	57.66	67.8	70.84	73.79	0.394	0.6030	1.10	1.5567
		GRADMATCH	41.01	59.88	68.25	71.5	0.5143	0.8114	1.40	2.002
		GRADMATCH-WARM	57.72	68.23	71.34	74.06	0.3788	0.7165	1.30	1.985
		GRADMATCHPB	40.53	60.39	70.88	72.57	0.3797	0.6115	1.09	1.56
		GRADMATCHPB-WARM	58.26	69.58	73.2	74.62	0.300	0.5744	1.01	1.5683
SVHN	ResNet18	FULL (skyline for test accuracy)	96.49	96.49	96.49	96.49	6.436	6.436,	6.436	6.436
		RANDOM (skyline for training time)	89.33	93.477	94.7	95.31	0.342	0.6383	1.26	1.90
		RANDOM-WARM (skyline for training time)	94.1	94.4	95.87	96.01	0.34	0.637	1.26	1.90
		GLISTER	94.78	95.37	95.5	95.82	0.5733	0.9141	1.62	2.514
		GLISTER-WARM	94.99	95.50	95.8	95.69	0.5098	0.8522	1.58	2.34
		CRAIG	94.003	94.86	95.83	96.223	1.3886	1.7566	2.39	3.177
		CRAIG-WARM	93.81	95.27	96.0	96.15	1.113	1.4599	2.15	2.6617
		CRAIGPB	94.26	95.367	95.92	96.043	0.5934	1.009	1.65	2.413
		CRAIGPB-WARM	94.339	95.724	96.06	96.385	0.5279	0.93406	1.58	2.332
		GRADMATCH	94.01	94.45	95.4	95.73	0.8153	1.1541	1.64	2.981
		GRADMATCH-WARM	94.94	95.13	96.03	95.79	0.695	0.9313	1.59	2.417
		GRADMATCHPB	94.37	95.36	96.12	96.24	0.5134	0.8438	1.6	2.52
		GRADMATCHPB-WARM	94.77	95.64	96.21	96.425	0.4618	0.7889	1.51	2.398

Table 3. Data Selection Results for CIFAR10, CIFAR100 and SVHN datasets

MNIST Data Selection Results									
			Top-1 Test accuracy of the Model(%)				Model Training time(in hrs)			
		Budget(%)	1%	3%	5%	10%	1%	3%	5%	10%
Dataset	Model	Selection Strategy								
MNIST	LeNet	FULL (skyline for test accuracy)	99.35	99.35	99.35	99.35	0.82	0.82	0.82	0.82
		RANDOM (skyline for training time)	94.55	97.14	97.7	98.38	0.0084	0.03	0.04	0.084
		RANDOM-WARM (skyline for training time)	98.8	99.1	99.1	99.13	0.0085	0.03	0.04	0.085
		GLISTER	93.11	98.062	99.02	99.134	0.045	0.0625	0.082	0.132
		GLISTER-WARM	97.63	98.9	99.1	99.15	0.04	0.058	0.078	0.127
		CRAIG	96.18	96.93	97.81	98.7	0.3758	0.4173	0.434	0.497
		CRAIG-WARM	98.48	98.96	99.12	99.14	0.2239	0.258	0.2582	0.3416
		CRAIGPB	97.72	98.47	98.79	99.05	0.08352	0.106	0.1175	0.185
		CRAIGPB-WARM	98.47	99.08	99.01	99.16	0.055	0.077	0.0902	0.1523
		GRADMATCH	98.954	99.174	99.214	99.24	0.05	0.0607	0.097	0.138
		GRADMATCH-WARM	98.86	99.22	99.28	99.29	0.046	0.057	0.089	0.132
		GRADMATCHPB	98.7	99.1	99.25	99.27	0.04	0.051	0.07	0.11
		GRADMATCHPB-WARM	99.0	99.23	99.3	99.31	0.038	0.05	0.065	0.10

Table 4. Data Selection Results for MNIST dataset

ImageNet Data Selection Results								
			Top-1 Test accuracy(%)			Model Training time(in hrs)		
		Budget(%)	5%	10%	30%	5%	10%	30%
Dataset	Model	Selection Strategy						
ImageNet	ResNet18	FULL (skyline for test accuracy)	70.36	70.36	70.36	276.28	276.28	276.28
		RANDOM (skyline for training time)	21.124	33.512	55.12	14.12	28.712	81.7
		CRAIGPB	44.28	55.36	63.52	22.24	38.9512	96.624
		GRADMATCH	47.24	56.81	66.21	18.24	35.7042	90.25
		GRADMATCH-WARM	55.86	58.21	68.241	16.48	33.024	88.248
		GRADMATCHPB	45.15	59.04	68.12	16.12	30.472	86.32
		GRADMATCHPB-WARM	56.61	61.16	69.06	15.28	29.964	86.05

Table 5. Data Selection Results for ImageNet dataset

- **k** determines the subset size with which we train the model.
- ϵ determines the extent of gradient approximation we want. We use a value of 1e-10 in our experiments.
- λ determines how much regularization we want. We set $\lambda = 0.5$.

C.4. Data Selection Results:

This section shows the results of training neural networks on subsets selected by different data selection strategies for various datasets. Table 3 shows the test accuracy and the training time of the ResNet18 model on CIFAR10, CIFAR100, and SVHN datasets for 300 epochs. Table 4 shows the test accuracy and the training time of the LeNet model on the MNIST dataset for 200 epochs. Table 5 shows the test accuracy and the training time of the ResNet18 model on the ImageNet dataset for 350 epochs. From the results, it is evident that GRAD-MATCHPB-WARM not only outperforms other baselines in terms of accuracy but is also more efficient in model training times. Furthermore, GLISTER and CRAIG could not be run on ImageNet due to large memory requirements and running time. GRAD-MATCH, GRAD-MATCHPB, and CRAIGPB were the only variants which could scale to ImageNet. Furthermore, GLISTER and CRAIG also perform poorly on CIFAR-100. Overall, we observe that GRAD-MATCH and its variants consistently outperform all baselines by achieving higher test accuracy and lower training times.

Energy Consumption Results: Table 6 shows the energy consumption (in KWH) for different subset sizes of CIFAR10, CIFAR100 datasets. The results show that GRAD-MATCHPB-WARM strategy is the most efficient in energy consumption out of all other selection strategies. Similarly, we could also observe that the PerBatch variants, i.e., CRAIGPB, GRAD-MATCHPB have better energy efficiency compared to GRAD-MATCH and CRAIG.

Energy Consumption Results						
		Budget(%)	Energy consumption for training the Model(in KWH) 5%	10%	20%	30%
Dataset	Model	Selection Strategy				
CIFAR10	ResNet18	FULL	0.5032	0.5032	0.5032	0.5032
		RANDOM (Skyline for Energy Consumption)	0.0592	0.0911	0.1281	0.18
		RANDOM-WARM (Skyline for Energy Consumption)	0.0581	0.0901	0.128	0.176
		GLISTER	0.0693	0.1012	0.1392	0.1982
		GLISTER-WARM	0.0672	0.0990	0.1360	0.1932
		CRAIG	0.0832	0.1195	0.1499	0.2063
		CRAIG-WARM	0.0770	0.1118	0.1438	0.2043
		CRAIGPB	0.0709	0.1031	0.1384	0.2005
		CRAIGPB-WARM	0.0682	0.1023	0.1355	0.2016
		GRAD-MATCH	0.0734	0.1173	0.1501	0.2026
		GRAD-MATCH-WARM	0.0703	0.1083	0.1429	0.2004
		GRAD-MATCHPB	0.0670	0.1006	0.1378	0.1927
		GRAD-MATCHPB-WARM	0.0649	0.0978	0.1354	0.1912
CIFAR100	ResNet18	FULL	0.5051	0.5051	0.5051	0.5051
		RANDOM (Skyline for Energy Consumption)	0.0582	0.0851	0.1116	0.1910
		RANDOM-WARM (Skyline for Energy Consumption)	0.0581	0.0850	0.1115	0.1910
		GLISTER	0.0674	0.0991	0.1454	0.2084
		GLISTER-WARM	0.0650	0.0940	0.1444	0.2018
		CRAIG	0.1146	0.1294	0.1795	0.2378
		CRAIG-WARM	0.0895	0.1209	0.1651	0.2306
		CRAIGPB	0.0747	0.0946	0.1443	0.2053
		CRAIGPB-WARM	0.0710	0.0916	0.1447	0.2039
		GRAD-MATCH	0.0721	0.1129	0.1577	0.2297
		GRAD-MATCH-WARM	0.0688	0.0980	0.1531	0.2125
		GRAD-MATCHPB	0.0672	0.0978	0.1477	0.2100
		GRAD-MATCHPB-WARM	0.0649	0.0928	0.1411	0.2001

Table 6. Energy consumptions results for training a ResNet18 model on CIFAR10, CIFAR100 datasets for 300 epochs

C.5. Standard deviation and statistical significance results:

Table 7 shows the standard deviation results over five training runs on CIFAR10, CIFAR100, and MNIST datasets. The results show that the GRAD-MATCHPB-WARM has the least standard deviation compared to other subset selection strategies. Note that the standard deviation of subset selection strategies is large for smaller subsets across different selection strategies. Furthermore, GLISTER has higher standard deviation values than random for smaller subsets, which partly explains the

Standard Deviation Results						
			Standard deviation of the Model(for 5 runs)			
		Budget(%)	5%	10%	20%	30%
Dataset	Model	Selection Strategy				
CIFAR10	ResNet18	FULL	0.032	0.032	0.032	0.032
		RANDOM	0.483	0.518	0.524	0.538
		RANDOM-WARM	0.461	0.348	0.24	0.1538
		GLISTER	0.453	0.107	0.046	0.345
		GLISTER-WARM	0.325	0.086	0.135	0.129
		CRAIG	0.289	0.2657	0.1894	0.1647
		CRAIG-WARM	0.123	0.1185	0.1058	0.1051
		CRAIGPB	0.152	0.1021	0.086	0.064
		CRAIGPB-WARM	0.0681	0.061	0.0623	0.0676
		GRAD-MATCH	0.192	0.123	0.112	0.1023
		GRAD-MATCH-WARM	0.1013	0.1032	0.091	0.1034
		GRAD-MATCHPB	0.0581	0.0571	0.0542	0.0584
		GRAD-MATCHPB-WARM	0.0542	0.0512	0.0671	0.0581
CIFAR100	ResNet18	FULL	0.051	0.051	0.051	0.051
		RANDOM	0.659	0.584	0.671	0.635
		RANDOM-WARM	0.359	0.242	0.187	0.175
		GLISTER	0.463	0.15	0.061	0.541
		GLISTER-WARM	0.375	0.083	0.121	0.294
		CRAIG	0.3214	0.214	0.195	0.187
		CRAIG-WARM	0.18	0.132	0.125	0.115
		CRAIGPB	0.12	0.134	0.123	0.115
		CRAIGPB-WARM	0.1176	0.1152	0.1128	0.111
		GRAD-MATCH	0.285	0.176	0.165	0.156
		GRAD-MATCH-WARM	0.140	0.134	0.142	0.156
		GRAD-MATCHPB	0.104	0.111	0.105	0.097
		GRAD-MATCHPB-WARM	0.093	0.101	0.100	0.098
			Standard deviation of the Model(for 5 runs)			
		Budget(%)	1%	3%	5%	10%
MNIST	LeNet	FULL	0.012	0.012	0.012	0.012
		RANDOM	0.215	0.265	0.224	0.213
		RANDOM-WARM	0.15	0.121	0.110	0.103
		GLISTER	0.256	0.218	0.145	0.128
		GLISTER-WARM	0.128	0.134	0.119	0.124
		CRAIG	0.186	0.178	0.162	0.125
		CRAIG-WARM	0.0213	0.0223	0.0196	0.0198
		CRAIGPB	0.021	0.0209	0.0216	0.0204
		CRAIGPB-WARM	0.023	0.0192	0.0212	0.0184
		GRAD-MATCH	0.156	0.128	0.135	0.12
		GRAD-MATCH-WARM	0.087	0.084	0.0896	0.0815
		GRAD-MATCHPB	0.0181	0.0163	0.0147	0.0129
		GRAD-MATCHPB-WARM	0.0098	0.012	0.0096	0.0092

Table 7. Standard deviation results for CIFAR10, CIFAR100 and MNIST datasets for 5 runs

fact that it does not work as well for very small subsets (e.g. 1% - 5%). We could also observe that the warm start variants of subset selection strategies have lower variance than non-warm-start ones from the standard deviation numbers, partly because of the better initialization they offer. Finally, the PerBatch variants GRAD-MATCHPB and CRAIGPB have lower standard deviation compared to GRAD-MATCH and CRAIG which proves the effectiveness of Per-Batch approximation.

In Table 8, we show the p-values of one-tailed Wilcoxon signed-rank test (Wilcoxon, 1992) performed on every single possible pair of data selection strategies to determine whether there is a significant statistical difference between the strategies in each pair, across all datasets. Our null hypothesis is that there is no difference between the data selection strategies pair. From the results, it is evident that GRAD-MATCHPB-WARM variant significantly outperforms other baselines at $p < 0.01$.

C.6. Other Results:

Gradient Errors: Table 9 shows the average gradient error obtained by various subset selection algorithms for the MNIST dataset. We observe that the gradient error of GRAD-MATCHPB is the smallest, followed closely by CRAIGPB. From the results, it is also evident that the PerBatch variants i.e., GRAD-MATCHPB and CRAIGPB achieves lower gradient error compared to GRAD-MATCH and CRAIG. Also note that GRAD-MATCH has a lower gradient error compared to CRAIG and GRAD-MATCH-PB has a lower gradient error compared to CRAIG-PB. This is expected since GRAD-MATCH directly

	RANDOM	GLISTER	GLISTER-WARM	CRAIG	CRAIG-WARM	CRAIGPB	CRAIGPB-WARM	GRAD-MATCH	GRAD-MATCH-WARM	GRAD-MATCHPB	GRAD-MATCHPB-WARM
RANDOM											
GLISTER	0.0006										
GLISTER-WARM	0.0002	0.0017									
CRAIG	0.0003	0.048	0.00866								
CRAIG-WARM	0.0002	0.0375	0.0492	0.0004							
CRAIGPB	0.0002	0.0334	0.0403	0.0010	0.0139						
CRAIGPB-WARM	0.0002	0.003	0.017	0.0002	0.0005	0.0002					
GRAD-MATCH	0.0002	0.028	0.031918	0.0030	0.0107	0.0254	0.015				
GRAD-MATCH-WARM	0.0002	0.0008	0.0018	0.0008	0.0057	0.0065	0.0117	0.0005			
GRAD-MATCHPB	0.0002	0.0048	0.0067	0.0002	0.0305	0.0002	0.0075	0.0248	0.0305		
GRAD-MATCHPB-WARM	0.0002	0.00007	0.0028	0.0002	0.0002	0.0002	0.0011	0.0005	0.0091	0.0002	

Table 8. Pairwise significance p-values using Wilcoxon signed rank test

optimizes the gradient error while CRAIG minimizes an upper bound. Also note that GLISTER has a signifcantly larger gradient error at 1% subset which partially explains the reason for bad performance of GLISTER for very small percentages.

Redundant Points: Table 10 shows the redundant points, i.e., data points that were never used for training for various subset selection algorithms on the MNIST dataset. The results give us an idea of information redundancy in the MNIST dataset while simultaneously showing that we can achieve similar performances to full training using a much smaller informative subset of the MNIST dataset.

MNIST Gradient Error Results							
			Avg.Gradient error norm				
		Budget(%)	1%	3%	5%	10%	30%
Dataset	Model	Selection Strategy					
MNIST	LeNet	RANDOM	410.1258	18.135	10.515	9.5214	6.415
		CRAIG	68.3288	19.2665	10.9991	6.5159	0.3793
		CRAIGPB	17.6352	2.9641	1.3916	0.4417	0.0825
		GLISTER	545.2769	7.9193	1.8786	2.8121	0.3249
		GRAD-MATCH	66.2003	17.6965	9.8202	2.1122	0.3797
		GRAD-MATCHPB	15.5273	2.202	1.1684	0.3793	0.0587

Table 9. Gradient approximation relative to full training gradient for various data selection strategies for different subset sizes of MNIST dataset

MNIST Redundant Points Results							
			Percentage of Redundant Points in MNIST training data				
		Budget(%)	1%	3%	5%	10%	30%
Dataset	Model	Selection Strategy					
MNIST	LeNet	CRAIG	90.381481	74.057407	60.492593	36.788889	14.425926
		CRAIGPB	90.405556	73.653704	60.327778	35.301852	2.875926
		GLISTER	90.712963	77.540741	67.544444	45.940741	7.774074
		GRAD-MATCH	91.124074	76.4	62.109259	36.114815	2.942593
		GRAD-MATCHPB	90.187037	73.468519	59.757407	36.164815	6.751852

Table 10. Redunant points(i.e., points never used for training) for various data selection strategies for different subset sizes of MNIST dataset

Comparison between variants of GRAD-MATCH: Table 11 shows the test accuracy and the training time for PerClass, PerClassPerGradient and PerBatch variants of GRAD-MATCH using ResNet18 model on different subsets of CIFAR10 and CIFAR100 datasets. First, note that even though the PerClass variant achieves higher accuracy than the PerClassPerGradient variant, it is significantly slower, having a larger training time than full data training for the 30% subset of CIFAR10 and CIFAR100. Since the PerClass variant of GRAD-MATCH is not scalable, we use the PerClassPerGradient variant, which achieves comparable accuracies while being much faster. Finally, note that the PerBatch variants performed better than the other variants in test accuracy and training efficiency.

Comparison between variants of GRAD-MATCH									
			Top-1 Test accuracy(%)				Model Training time(in hrs)			
		Budget(%)	5%	10%	20%	30%	5%	10%	20%	30%
Dataset	Model	GRAD-MATCH Variant								
CIFAR100	ResNet18	PerClassPerGradient	41.01	59.88	68.25	71.5	0.5143	0.8114	1.40	2.002
		PerClass	41.57	59.95	70.87	72.45	0.5357	1.225	1.907	3.796
		PerBatch	40.53	60.39	70.88	72.57	0.3797	0.6115	1.09	1.56
CIFAR10	ResNet18	PerClassPerGradient	86.7	90.9	91.67	91.89	0.40	0.84	1.42	1.52
		PerClass	85.12	91.04	92.12	93.69	0.4225	1.042	1.92	3.48
		PerBatch	85.4	90.01	93.34	93.75	0.36	0.69	1.09	1.38

Table 11. Top-1 test accuracy(%) and training times for variants of GRAD-MATCH for different subset sizes of CIFAR10, CIFAR100 datasets

Additional Data Selection Results			
			Top-1 Test accuracy(%)
		Budget(%)	30%
Dataset	Model	Selection strategy	
CIFAR100	ResNet164	Facility Location	91.1
		Forgetting Events	92.3
		Entropy	90.4
	ResNet18	GRAD-MATCHPB-WARM	94.17
CIFAR10	ResNet164	Facility Location	64.8
		Forgetting Events	63.4
		Entropy	60.4
	ResNet18	GRAD-MATCHPB-WARM	74.62

Table 12. Top-1 test accuracy(%) and training times for additional data selection strategies on 30% CIFAR10 and CIFAR100 subset

Comparison with additional subset selection methods: In addition to the baselines we considered so far, we compare GRAD-MATCH with additional existing subset selection strategies like Facility Location (Wolf, 2011), Entropy (Settles, 2012) and Forgetting Events (Toneva et al., 2019) on CIFAR10 and CIFAR100 datasets. The results are in Table 12. Note that we used the numbers reported in paper (Coleman et al., 2020) for comparison. The authors in (Coleman et al., 2020) used a ResNet-164 Model which is a higher complexity model compared to ResNet-18 which we use in our experiments. Even after using a lower complexity model (ResNet-18), we outperform these other baselines on both CIFAR-10 and CIFAR-100. Furthermore, we achieve this while being much faster (since we observed that the ResNet-164 model is roughly 4x slower compared to ResNet-18). Even though a much smaller model (ResNet-20) is used for data selection, the training is still done with the ResNet-164 model. Finally, note that the selection via proxy method is orthogonal to GRAD-MATCH and can also be applied to GRAD-MATCH to achieve further speedups. We expect that the accuracy of these baselines (Forgetting Events, Facility Location, and Entropy) to be even lower if they are used with a ResNet-18 model. The accuracy reported for these baselines (Table 12) are the best among the different proxy models used in (Coleman et al., 2020).