A Appendix

A.1 Additional Experiment Result

In this section, we provide additional experiment results.

Counter matrix C on D^o versus C on D^u could be a good visualization over distribution shift. As a toy example, we plot C_o and C_u on Mnist-binary with $R = 32$. We randomly sample 3 rows as it is easier to visualize. For each comparison, top row is from C_o, bottom is from C_u. We can observe the density difference.

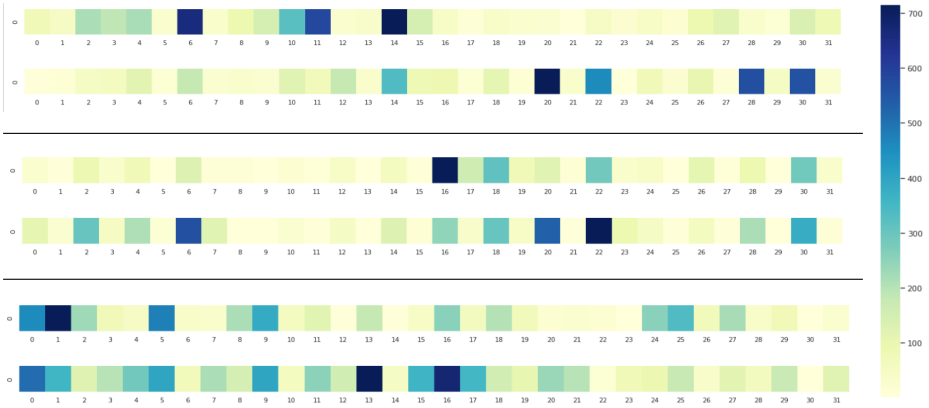

Table A.1 compare RIDDLE against two training scheme mentioned in Section 1: update and retrain. Retrain refers to retraining the model on all available data from scratch, which would yield best performance. Update refers to training on a trained model only on new data. Update shows unsatisfying accuracy, however, is more efficient than retrain. The best of both world would be only updating on new data but achieve retraining accuracy. We observe that RIDDLE's performance is significantly better than update, and close to Retrain.

Dataset	Method	Original
Mnist Binary	Update	69.1
	Retrain	98.9
	RIDDLE	98.0
Cifar10	Update	63.2
	Retrain	94.3
	RIDDLE	89.0
ImageNet	Update	55.6
	Retrain	72.0
	RIDDLE	67.1
News	RoBERTa	74.0
	Retrain	92.4
	RIDDLE	89.7

Table 5: This table place RIDDLE between the best possible performance and the worst performance. Lower bound is obtained by updating on a trained model on update dataset using SGD. Upper bound is obtained by training a model from scratch on both original dataset and update dataset. We see RIDDLE's performance is close to upper bound

We investigate the choice of number of rows L and number of cells R. L and R are the two parameters controlling the tradeoff between efficiency and expressiveness. It is clear that increase memory budget leads to better accuracy. At the same R, increase the L leads to a higher accuracy. At the same L, increase R also lead to a higher accuracy. Increase L increases the repetition, leading to more robust model. Increase R increase the hashing precision.

	Parameters	R16L500	R32L500	R64L1000	R64L2000	R32L4000
Boston Housing	Memory(M)	0.03	0.06	0.25	0.52	0.52
	Accuracy	2.30	2.23	2.20	2.16	1.96

Table 6: This table summarizes the accuracy with various choices of R, number of cells, and L, number of rows. In general, larger R and L will lead to higher accuracy with more memory consumption.

A.2 Experiment Setup

We use Adam Optimizer with a learning rate 0.0001 for all baselines. We use early stopping on all methods. We do not use any data augmentation or feature engineering, except normalizing images. We do not use regularization. Following two tables provides architecture details.

Dataset	L	LSH	α
Mnist Binary	500	SimHash	0
Cifar10	2000	SimHash	0
ImageNet	2000	SimHash	0
News	3000	SimHash	0

Table 7: This table summarize hyper-parameters for RIDDLE for Figure 2 and Table 1

Dataset	Evaluation	NN1	NN2	RIDDLE
Susy	Accuracy	300-300-300	300-300-300-300-300	R16L6000
HAR	Accuracy	300-300-300	LSTM-2Layer-32	R32L1500
Covtype	AUC	500-500	500-500-500	R128L1500
Connect4	Accuracy	500-500	500-500-500	R16L4000
Mnist	Accuracy	500-500-500	Conv2d-Conv2d-9216-128	R16L2000
Fashion Mnist	Accuracy	500-500	500-500-500	R32L1500
Boston Housing	MAE	500-500	500-500-500	R32L4000

Table 8: This table summarize information about baselines and dataset for experiments regarding to expressiveness and efficiency(Table 2 and Table 3). NN1 and NN2 summarize neural network architecture. For example, 300-300-300 refers to a three layer MLP with hidden size 300.

A.3 Theory

In this section, we provide complete proofs for our theoretical results.

A.3.1 Proof of Theorem 4.4

Theorem A.1. *Given a dataset $\mathcal{D}$ of weighted samples $\{(\alpha_{x_i}, x_i)\}$, let $h(x)$ be an LSH function drawn from an LSH family with collision probability $\mathcal{K}(\cdot,\cdot)$. Let S be a row of the sketch constructed using $h(x)$. For any query q,*

$$\mathbb{E}(S[h(q)]) = \sum_i^{|\mathcal{D}|} \alpha_{x_i} \mathcal{K}(x_i, q), \qquad \text{var}\left(S[h(q)]\right) \leq \left(\sum_i^{|\mathcal{D}|} \alpha_{x_i} \sqrt{\mathcal{K}(x_i, q)}\right)^2$$

Proof. Let $\mathbb{1}_i$ denote the indicator function $\mathbb{1}_{h(x_i)=h(q)}$. That is, $\mathbb{1}_i = 1$ when data x_i from the dataset collides with the query q. To simplify the presentation, let $Z = S[h(q)]$.

Expectation: The value in the sketch can be written as

$$Z = \sum_{i}^{|\mathcal{D}|} \alpha_{x_i} \mathbb{1}_i$$

By the linearity of the expectation

$$\mathbb{E}(Z) = \sum_{i}^{|\mathcal{D}|} \alpha_{x_i} \mathbb{E}(\mathbb{1}_i)$$

We know that $\mathbb{E}(\mathbb{1}_i)$ is the collision probability of $h(x)$, thus,

$$\mathbb{E}(Z) = \sum_{i}^{|\mathcal{D}|} \alpha_{x_i} \mathcal{K}(x_i, q)$$

Variance: The variance is bounded by the second moment. The second moment of this estimator can be written as

$$\mathbb{E}(Z^2) = \sum_{i}^{|\mathcal{D}|} \sum_{j}^{|\mathcal{D}|} \alpha_{x_i} \alpha_{x_j} \mathbb{E}(\mathbb{1}_i \mathbb{1}_j)$$

By the Cauchy-Schwarz inequality,

$$\mathbb{E}(\mathbb{1}_i \mathbb{1}_j) \leq \sqrt{\mathbb{E}(\mathbb{1}_i)} \sqrt{\mathbb{E}(\mathbb{1}_j)}$$

Then,

$$\mathbb{E}(Z^2) \leq \sum_{i}^{|\mathcal{D}|} \sum_{j}^{|\mathcal{D}|} \alpha_{x_i} \alpha_{x_j} \sqrt{\mathcal{K}(x_i, q)} \sqrt{\mathcal{K}(x_j, q)} = \left(\sum_{i}^{|\mathcal{D}|} \alpha_{x_i} \sqrt{\mathcal{K}(x_i, q)} \right)^2$$

Thus,

$$\text{var}(Z) \leq \left(\sum_{i}^{|\mathcal{D}|} \alpha_{x_i} \sqrt{\mathcal{K}(x_i, q)} \right)^2$$

□

A.3.2 Proof of Lemma A.2

We use a result from [3], where we suppose for convenience that g evenly divides N.

Lemma A.2. *Let $Z_1, ...Z_R$ be L i.i.d. random variables with mean $\mathbb{E}[Z] = \mu$ and variance $\leq \sigma^2$. Divide the L variables into g groups so that each group contains $m = L/g$ elements, and take the empirical average within each group. The median-of-means estimate $\hat{\mu}$ is the median of the g group means. If $g = 8\log(1/\delta)$ and $m = L/g$, then the following statement holds with probability $1 - \delta$.*

$$|\hat{\mu} - \mu| \leq 6 \frac{\sigma}{\sqrt{L}} \sqrt{\log 1/\delta}$$

Proof. This proof is given in [3] as the proof of Theorem 2.1 (which is a slightly more general version of the statement above). □

A.3.3 Proof of Theorem 4.5

Theorem A.3. *Let $Z(q)$ be the median-of-means estimate constructed using the L unbiased estimators of the sketch with L rows. Then with probability $1 - \delta$,*

$$|Z(q) - f_K(q)| \leq 6 \frac{\tilde{f}_K(q)}{\sqrt{L}} \sqrt{\log 1/\delta}$$

where $f_K(q)$ and $\tilde{f}_K(q)$ are the weighted KDE with kernels $\mathcal{K}(x, q)$ and $\sqrt{\mathcal{K}(x, q)}$, respectively.

Proof. From Theorem 4.4, we know that

$$\sigma \leq \sum_{i}^{|\mathcal{D}|} \alpha_{x_i} \sqrt{\mathcal{K}(x_i, q)}$$

Substituting this variance bound into Lemma A.2 proves the theorem. □

A.3.4 Proof of Lemma 4.2

Lemma A.4. *The L2 LSH kernel from [19] is shift-invariant and universal.*

Proof. To see that the kernel is shift-invariant, observe that $K(x, y) = K(\text{dist}(x, y))$ and that the Euclidean distance $\|x - y\|_2$ is a function of the difference $x - y$. Thus, $K(x, y) = K(x - y)$.

Since the kernel is shift-invariant, it is sufficient to show that the support of the Fourier transform of the kernel is the entire real line [13]. Observe from [21] that the kernel may be written as

$$k(c) = \int_0^r \left(1 - \frac{t}{r}\right) \frac{1}{c} e^{-\frac{t^2}{2c^2}} dt$$

where $c = \text{dist}(x, y)$ and r is a parameter. The Fourier transform of this quantity is

$$\int_{-\infty}^{\infty} \left(\int_0^r \frac{2}{\sqrt{2\pi}} \left(1 - \frac{t}{r}\right) \frac{1}{c} e^{-\frac{t^2}{2c^2}} dt \right) e^{i\omega c} d\omega$$

The integrand satisfies the requirements of Fubini's Theorem (absolutely integrable), so we may exchange the order of integration.

$$\int_0^r \frac{2}{\sqrt{2\pi}} \left(1 - \frac{t}{r}\right) \left(\int_{-\infty}^{\infty} \frac{1}{c} e^{-\frac{t^2}{2c^2}} e^{i\omega c} d\omega \right) dt$$

Observe that the Fourier transform of the inner integrand has full support because the exponential function is nonzero everywhere. Now observe that the outer integral is the limit of a sum of Fourier transforms, each of which has the real line as its support. Because $\frac{2}{\sqrt{2\pi}}(1 - t/r) > 0$ over the integration region, the coefficients of this sum are positive. Therefore, the Fourier transform of the kernel has the real line as its support. □

A.3.5 Proof of Theorem 4.3

Theorem A.5. *Given a continuous and bounded function $g(q)$ and $\epsilon > 0$, there exists a set of coefficients $\{\alpha_n\}$, set of points $\{x_n\}$ and an integer N such that*

$$f_N(q) = \sum_{n=1}^{N} \alpha_N \mathcal{K}(x_n, q) \qquad \|f_N(q) - g(q)\|_{\mathcal{X}} \leq \epsilon$$

where $\mathcal{K}(x_n, q)$ is the L2 LSH kernel and $\mathcal{X}$ is any compact subset of $\mathbb{R}^d$.

Proof. This follows directly from the definition of a universal kernel. □

A.3.6 Proof of Theorem 3.2

Theorem A.6. *Given a dataset D, construct a representer sketch model $f_{(S,h)}(x)$ by running Algorithm 1 for e epochs with learning rate η. Suppose the gradient norms are bounded by G. Then*

$$|\mathbb{E}_h[f_{(S,h)}(x)]| \leq e\eta G \,\text{KDE}(x)$$

where $\text{KDE}(x)$ is the kernel density of x over D using the kernel of the LSH function h.

Proof. Given a sketch S equipped with a set of hash functions h, the value of the model output for a query x is:

$$f_{(S,h)}(x) = \frac{1}{L}\sum_{l=1}^{L} S[l, h_l(x)]$$

We will examine the values of the array cells $S[l, h_l(x)]$ when trained using Algorithm 1 on a dataset D of m points (i.e. $|D| = m$). Under Algorithm 1, the sketch is updated once by each point in the dataset during every epoch. If we update the sketch at time t with the feature-target pair (z, y_z), the update takes the form:

$$S_{t+1} = S_t - \eta \nabla_S \frac{1}{m} E(f_{(S,h)}(z), y_z)$$

where $E(f_{(S,h)}(z), y_z)$ is the error function used to train the model. Observe that

$$f_{(S,h)}(z) = \frac{1}{L}\sum_{l=1}^{L} S[l, h_l(z)]$$

In other words, the only parameters in S that contribute to $f_{(S,h)}(z)$ are those at the array locations $\{(1, h_1(z)), ...(L, h_L(z))\}$. Since the other parameters do not participate in the computation of $f_{(S,h)}(z)$, each example $(z, y_z) \in D$ updates only those cells in S that correspond to the values of $\{h_1(z), ...h_L(z)\}$.

Let $U \in \mathbb{Z}^{L \times R}$ be the sparse matrix formed by placing a '1' at the locations $\{(1, h_1(z)), ...(L, h_L(z))\}$ and '0' otherwise. Since the gradient norms are bounded by G, we can bound the size of the update to the sketch.

$$S_{t+1} - S_t \leq \eta \frac{G}{m} U_t$$

If we recursively apply this inequality to all m updates in the epoch, we obtain:

$$S_m - S_0 \leq \eta G \sum_{t=1}^{m} U_t$$

Note that the inequality remains true if we swap the positions of S_t and S_{t+1}. Also note that with zero initialization, $S_0 = 0^{L \times R}$. Therefore, we have:

$$-\eta G \sum_{t=1}^{m} U_t \leq S_m \leq \eta G \sum_{t=1}^{m} U_t$$

If we obtain the sketch S after training for e epochs, we pick up an additional multiplicative factor e.

$$-e\eta G \sum_{t=1}^{m} U_t \leq S \leq e\eta G \sum_{t=1}^{m} U_t$$

This means that we can bound the absolute value of the sketch output $f_{(S,h)}(x)$:

$$|f_{(S,h)}(x)| \leq \frac{1}{L}\sum_{l=1}^{L} |S[l, h_l(x)]| \leq \frac{1}{L}\sum_{l=1}^{L} e\eta G \sum_{t=1}^{m} U_t[l, h_l(x)]$$

Taking the expectation of both sides of the inequality and applying the linearity of the expectation operator, we have:

$$\mathbb{E}[|f_{(S,h)}(x)|] \leq \frac{1}{L}\sum_{l=1}^{L} e\eta G\mathbb{E}\left[\sum_{t=1}^{m} U_t[l, h_l(x)]\right]$$

Note that we may re-write $U_t[l, h_l(x)]$ as the indicator $\mathbb{1}_{\{h_l(z_t)=h_l(x)\}}$. That is, the indicator is '1' if the example z_t from gradient descent iteration t collides with x. Since each epoch processes all elements in the dataset exactly one time, we may re-write the sum over $U_t[l, h_l(x)]$ to be over the dataset.

$$\mathbb{E}[|f_{(S,h)}(x)|] \leq \frac{1}{L}\sum_{l=1}^{L} e\eta G\mathbb{E}\left[\sum_{z\in D} \mathbb{1}_{\{h_l(z)=h_l(x)\}}\right]$$

The expected value of the indicator $\mathbb{1}_{\{h_l(z)=h_l(x)\}}$ is the collision probability $\Pr[h_l(z) = h_l(x)]$ of the LSH function h_l. Under the assumptions stated in the theorem, this collision probability is equal to the kernel $\mathcal{K}(z, x)$, leading to the following inequality:

$$\mathbb{E}[|f_{(S,h)}(x)|] \leq \frac{1}{L}\sum_{l=1}^{L} e\eta G\sum_{z\in D} \mathcal{K}(x, z)$$

To prove the theorem, recall that for any random variable Z, $|\mathbb{E}[Z]| \leq \mathbb{E}[|Z|]$. This yields the desired statement.

$$|\mathbb{E}[f_{(S,h)}(x)]| \leq e\eta G\text{KDE}(x)$$

□

A.3.7 Proof of Theorem 3.3

We want to bound the difference between the loss induced by S_o and the loss induced by S_{o+u} on D^o. In other words, we wish to bound:

$$\sum_{(x,y)\in D^o} \ell(f(x; S_{o+u}, h), y) - \sum_{(x,y)\in D^o} \ell(f(x; S_o, h), y) = \sum_{(x,y)\in D^o} \ell(f(x; S_{o+u}, h), y) - \ell(f(x; S_o, h), y)$$

Note that because S_o is the argmin, this quantity is always positive. Assume that the loss ℓ is at least locally L-Lipschitz in the first argument. That is, we suppose that the following bound holds in a ball surrounding $\hat{y}$:

$$|\ell(\hat{y}_a, y) - \ell(\hat{y}_b, y)| \leq L|\hat{y}_a - \hat{y}_b|$$

Then, applying this to the loss from earlier:

$$\sum_{(x,y)\in D^o} \ell(f(x; S_{o+u}, h), y) - \ell(f(x; S_o, h), y) \leq \sum_{(x,y)\in D^o} \leq \sum_{(x,y)\in D^o} L|f(x; S_{o+u}, h) - f(x; S_o, h)|$$

Using Theorem 3.2, we may bound the left hand side of this expression in expectation:

$$\mathbb{E}_h\left[\sum_{(x,y)\in D^o} \ell(f(x; S_{o+u}, h), y) - \ell(f(x; S_o, h), y)\right] \leq \sum_{(x,y)\in D^o} GLe\eta\text{KDE}(x, D^u)$$

Here, we used the elementary fact that $|\mathbb{E}[Z]| \leq \mathbb{E}[|\mathbb{Z}|]$ for any random variable Z.