

A Appendix Roadmap

This appendix provides supporting material organized as follows:

- **Experimental Setup** (Appendix B): Model architectures, hyperparameters, and training configurations.
- **Factorization of Representation Capacity** (Appendix C): Detailed analysis showing how representation capacity matrices can be factorized for various compression techniques including quantization, sparsity, and their combinations.
- **Ablation Studies on Law Formulation** (Appendix D): Investigation of different noise distributions (Gaussian, Logistic, Student’s t, Laplace) and functional forms (tanh, logistic) for the scaling law formulation.
- **Scaling Laws for Vector Quantization** (Appendix E): Implementation details and algorithms for vector quantization approaches, including forward and backward pass descriptions for HIGGS-based training.
- **Breaking the Scaling Law** (Appendix F): Demonstration of how training-time overparameterization with learnable block diagonal matrices can exceed FP16 performance ($\rho > 1$).
- **Theoretical Support** (Appendix G): Convergence analysis for Adam optimizer with Straight-Through Estimation (STE), including complete proofs and supporting lemmas.
- **Improved Sparse Training via RBBM** (Appendix H): Comparison of our backward mask heuristics against RigL and Gradual Magnitude Pruning, with detailed descriptions of different masking strategies.

B Experimental setup

Hyperparameters. Table 2 summarizes the architectural and training hyperparameters for each model size.

Model size	# Layers	# Heads	# Embeddings	Learning rate
30 M	6	5	640	$1.2 \cdot 10^{-3}$
50 M	7	6	768	$1.2 \cdot 10^{-3}$
100 M	8	8	1024	$6 \cdot 10^{-4}$
200 M	10	10	1280	$3 \cdot 10^{-4}$

Table 2: Key training hyperparameters for each model size.

We use 8x80GB H100 machines for efficient training, and training one model takes on average 1 hour. To produce the full set of results we ran in total approximately 250 such training runs for various compression configurations.

C Factorization of Representation Capacity

Figures 6-9 show factorization of the representation capacity matrix for various in-training compression techniques:

1. Quantized weights and activations (Fig. 6).
2. Sparsity + QuEST quantizer (Fig. 7).
3. Joint sparse & quantized weights + activations (Fig. 8), for all combinations (s_a, q_a, q_b) for sparsity $s_a \in [0.25, 0.5, 0.75]$ and bit widths $q_a, q_b \in [2, 4, 6]$.
4. Sparsity + uniform quantizer with maximum absolute value as a scale (Fig. 9).

From the factorized representation-capacity matrices we observe the following:

1. The element-wise error of the fitted coefficients ρ (from our scaling law) is of order 10^{-3} – 10^{-2} .

- 571 2. A rank-1 row-column outer product accurately approximates the matrix, confirming the
572 multiplicative property of representation capacity ρ in various scenarios.
- 573 3. Approximation error remains of the order 10^{-2} , except for the cases of *extreme* 2-bit
574 quantization, where $\rho \lesssim 0.1$. We explain this gap due to the poorer performance of the
575 optimizer in these extreme compression regimes, which is not taken into account currently
576 by our model (as it uses the same coefficients for both 16 and 2 bits).

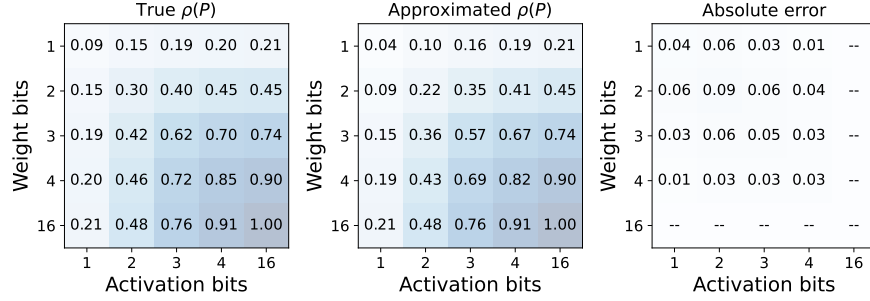


Figure 6: Representation capacity coefficients for independent quantization of weights and activations. Element-wise ρ fitting error is not greater than $5 \cdot 10^{-3}$.

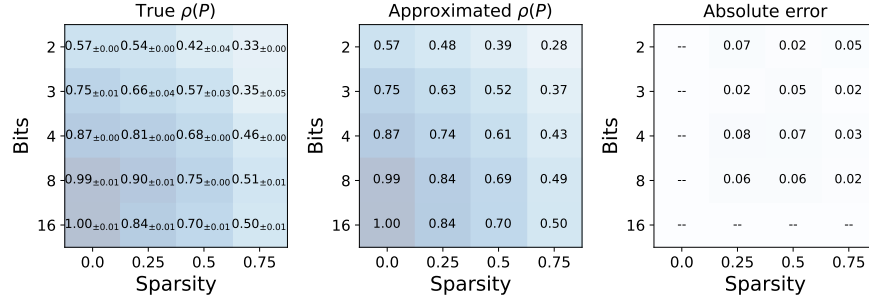


Figure 7: Representation capacity coefficients with fit errors in case of sparsity combined with the QuEST quantization.

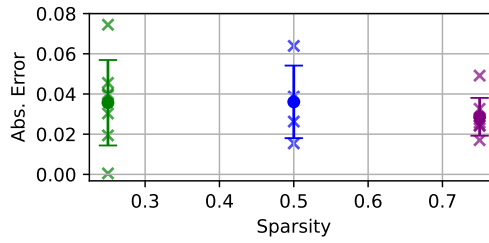


Figure 8: Representation capacity fit errors for sparse+quantized weights and quantized activations. Error bars denote ± 1 standard deviation from the mean.

577 D Ablation studies on Law Formulation

578 D.1 Evaluating RMSE across Different Distributions

579 We investigate how the choice of noise distribution used in our law formulation from Sec. 4.1 affects
580 the predicted representation capacity. In Figure 10a we plot the mapping $\rho(MSE)$ for different bit

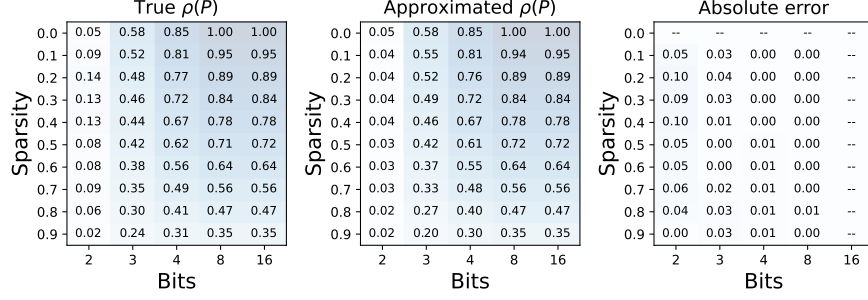
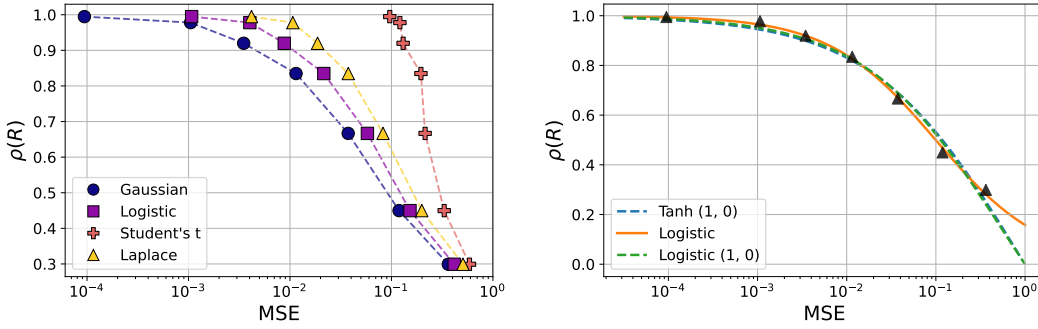


Figure 9: Representation capacity coefficients matrix for sparsity applied with uniform quantization. Element-wise ρ fitting error is not greater than $2 \cdot 10^{-3}$.



(a) Effect of input noise distribution on the mapping $\rho(MSE)$.

(b) Different functions used to fit $\rho(MSE)$

widths using Logistic, Student's t, and Laplace noise distributions. Each distribution is rescaled to have zero mean and unit variance.

We observe that, no matter which noise distribution we choose, the mapping $\rho(MSE)$ always remains strictly monotonically decreasing. In principle, one could use heavy-tailed distributions (for example, Student-t or Laplace) to give more weight to extreme outlier errors. However, this leads to a smaller range of MSE values. By contrast, assuming Gaussian noise—which we propose—produces the widest spread of MSE, which in turn allows for a better fit for the scaling law. In short, although monotonicity is preserved under various distributions, the Gaussian MSE delivers the best overall representation capacity prediction, so we adopt it as our default formulation.

Throughout this work, unless specified otherwise, MSE is computed over standard Gaussian input.

D.2 Functional form of the Law

The behavior of $\rho(GMSE)$ observed in our experiments can be captured by fitting multiple smooth, monotonically decreasing functions, with no more than 2 additional parameters. In principle, a wide range of such functions can be used to model this relationship, depending on the desired fit properties.

For lower overall fitting error, we found it beneficial to constrain the function to satisfy boundary conditions $f(0) = 1$ and $f(\infty) = 0$. For instance, the logistic form $\frac{1}{1+\exp(a \cdot \log(MSE+b))} = \frac{1}{1+B \cdot MSE^A}$ provides a good empirical fit, as shown in Figure 10b for weight quantization across 1-8 bit widths.

In cases where it is important to constrain MSE below 1, one may instead prefer the condition $f(1) = 0$. Although this typically results in a worse overall fit, it enforces the correct behavior in the high-error region $MSE \lesssim 1$, which is critical for stable predictions in the extreme compression cases. The corresponding fits, including those constrained at $f(1) = 0$, are summarized in Table 3 and visualized in Figure 10b.

The choice of functional form reflects the trade-off between global fit quality and targeted accuracy for larger MSE values. Throughout this work, we adopt the constraint $f(1) = 0$ and functional form

	Functional form	Fitting error (MSE)
Tanh	$\tanh(F \cdot \log_{1/4} \text{MSE})^C$	$1 \cdot 10^{-3}$
Logistic	$(1 + B \cdot \text{MSE}^A)^{-1}$	$1 \cdot 10^{-4}$
Logistic (1, 0)	$\frac{1 - \text{MSE}^A}{1 + B \cdot \text{MSE}^A}$	$1 \cdot 10^{-3}$

Table 3: Functional form choices and associated fitting error.

of hyperbolic tangent. As for the exact functional form, under the stated constraints, we find that the specific choice between tanh and logistic sigmoid has little effect on overall fit quality.

E Scaling Laws for Vector Quantization

In this section, we provide detailed information about the Vector Quantization approach used to produce the results in Figure 2(a). Algorithms 1 and 2 describe the forward and backward passes over a linear layer actively quantized with HIGGS for row-major weights. As was described earlier, our method combines ideas from Panferov *et al.* [24] for the gradient estimator, and Malinovskii *et al.* [22] for the lattice representation. We use the trust estimation method that zeros out gradients for any point lying outside a hypersphere of radius R : $\|x\|_2^2 > R^2$. Our experiments were conducted on 30M and 50M models using the same set of hyperparameters as in Sec. 2.

Algorithm 1 VQ Training Forward

- 1: **Input:** Input activations \mathbf{x} , row-major weight \mathbf{w}
 - 2: $\mathbf{w}_h = \text{HT}(\mathbf{w})$
 - 3: $\hat{\mathbf{w}}_h = \text{proj}_{\text{grid}} \mathbf{w}_h$
 - 4: $\mathbf{y} = \mathbf{x} \hat{\mathbf{w}}_h^T$
 - 5: **Return:** $\mathbf{y}, \mathbf{x}, \hat{\mathbf{w}}_h, M_{\text{grid}}(\mathbf{w}_h; \hat{\mathbf{w}}_h)$
-

Algorithm 2 VQ Training Backward

- 1: **Input:** $\frac{\partial L}{\partial \mathbf{y}}, \mathbf{x}, \hat{\mathbf{w}}_h, M_{\text{grid}}(\mathbf{w}_h; \hat{\mathbf{w}}_h)$
 - 2: $\frac{\partial L}{\partial \mathbf{x}} = \frac{\partial L}{\partial \mathbf{y}} \hat{\mathbf{w}}_h$
 - 3: $\frac{\partial L}{\partial \hat{\mathbf{w}}_h} = \mathbf{x}^T \frac{\partial L}{\partial \mathbf{y}}$
 - 4: $\frac{\partial L}{\partial \mathbf{w}} = \text{IHT} \left(M_{\text{grid}}(\mathbf{w}_h; \hat{\mathbf{w}}_h) \odot \frac{\partial L}{\partial \hat{\mathbf{w}}_h} \right)$
 - 5: **Return:** $\frac{\partial L}{\partial \mathbf{x}}, \frac{\partial L}{\partial \mathbf{w}}$
-

F “Breaking” the Scaling Law by Training-Time Overparametrization

One observation stemming from our law is that it is possible to overparameterize the model during training while keeping the number of inference-time parameters the same. This can be achieved by multiplying the weight matrix W by a learnable block diagonal matrix R , which is then “folded” into the model at inference time. The forward pass takes the form of $X(RW)^T$ during training and XW^T at inference. While R is omitted during evaluation, maintaining the original model size, additional parameters add flexibility during training and improve the representation quality.

For each weight matrix, we initialize our rotation matrix with a block-diagonal Hadamard, with the block sizes equal to 128, and learn it alongside W using our loss function. We train the 30M and 50M models using the same experimental setup as in Sec. 2 with rotation matrices, calculate the effective representation capacity, and compare it to baseline.

We observed that it results in a representation capacity $\rho = (1.07 \pm 0.04)$, indicating that model trained with such overparameterization outperforms the bf16 baseline ($\rho = 1$), even though their inference costs are identical.

629 G Theoretical Support

630 Here we provide the full proof of Theorem 1 giving a convergence analysis of the Adam optimizer
 631 when used with STE. For completeness, the description of the algorithm is presented in the Algorithm
 632 3.

Algorithm 3 Adam with Straight Through Estimation (STE) and AMSGrad normalization

```

1: Input: parameters  $\beta_1, \beta_2 \in (0, 1)$ ,  $\epsilon > 0$ , step-size  $\eta > 0$ ,  $\theta_1 \in \mathbb{R}^d$ ,  $m_0 = v_0 = \tilde{v}_0 = 0_N$ 
2: for  $t = \{1, 2, \dots, T\}$  do
3:    $\hat{\theta}_t = \mathcal{C}(\theta_t)$  ◇ Compress the model via quantization and/or sparsification
4:    $g_t = \tilde{\nabla}_{\theta} f(\hat{\theta}_t)$  ◇ Compute STE for compressed model
5:    $m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$  ◇ Update first-order gradient momentum
6:    $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$  ◇ Update second-order gradient momentum
7:    $\tilde{v}_t = \max(v_t, \tilde{v}_{t-1})$  ◇ Apply AMSGrad normalization
8:    $\theta_{t+1} = \theta_t - \eta \frac{m_t}{\sqrt{\tilde{v}_t + \epsilon}}$  ◇ Update the uncompressed model parameters
9: end for
  
```

Proof. Let G be the gradient bound with respect to ℓ_2 norm, that is, $\|g_t\|_2 \leq G$. Using the relationship between ℓ_2 and ℓ_∞ norms, we conclude $G \leq \sqrt{d}G_\infty$. Let $\Gamma_t = \text{Diag}^{-1/2}(\tilde{v}_t + \epsilon)$ be the preconditioning (diagonal) matrix and rewrite the main update rule as

$$\theta_{t+1} = \theta_t - \eta \Gamma_t m_t.$$

633 Letting $\theta_0 = \theta_1$, define virtual iterates x_t as follows:

$$x_t = \frac{1}{1 - \beta_1} \theta_t - \frac{\beta_1}{1 - \beta_1} \theta_{t-1}.$$

634 In particular, $x_1 = \theta_1$. Then, the update rule for the virtual iterates becomes

$$\begin{aligned}
 x_{t+1} - x_t &= \frac{1}{1 - \beta_1} (\theta_{t+1} - \theta_t) - \frac{\beta_1}{1 - \beta_1} (\theta_t - \theta_{t-1}) \\
 &= -\frac{\eta}{1 - \beta_1} \Gamma_t m_t + \frac{\eta \beta_1}{1 - \beta_1} \Gamma_{t-1} m_{t-1} \\
 &= -\frac{\eta}{1 - \beta_1} \Gamma_t m_t + \frac{\eta \beta_1}{1 - \beta_1} \Gamma_{t-1} m_{t-1} \pm \frac{\eta \beta_1}{1 - \beta_1} \Gamma_t m_{t-1} \\
 &= -\frac{\eta}{1 - \beta_1} \Gamma_t (m_t - \beta m_{t-1}) + \frac{\eta \beta_1}{1 - \beta_1} \underbrace{(\Gamma_{t-1} - \Gamma_t)}_{\stackrel{\text{def}}{=} \Delta \Gamma_t} m_{t-1} \\
 &= -\eta \Gamma_t g_t + \frac{\eta \beta_1}{1 - \beta_1} \Delta \Gamma_t m_{t-1}.
 \end{aligned}$$

635 Next we apply smoothness (Assumption 1) of the loss function f over the iterates x_t :

$$f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2.$$

636 Taking expectation and splitting the inner product into two part, we obtain

$$\begin{aligned}
& \mathbb{E}[f(x_{t+1})] - \mathbb{E}[f(x_t)] \\
& \leq -\eta \mathbb{E}[\langle \nabla f(x_t), \Gamma_t g_t \rangle] + \eta \mathbb{E} \left[\left\langle \nabla f(x_t), \frac{\beta_1}{1-\beta_1} \Delta \Gamma_t m_{t-1} \right\rangle \right] \\
& \quad + \frac{\eta^2 L}{2} \mathbb{E} \left[\left\| \Gamma_t g_t - \frac{\beta_1}{1-\beta_1} \Delta \Gamma_t m_{t-1} \right\|^2 \right] \\
& = \underbrace{-\eta \mathbb{E}[\langle \nabla f(\theta_t), \Gamma_t g_t \rangle]}_I + \underbrace{\eta \mathbb{E} \left[\left\langle \nabla f(x_t), \frac{\beta_1}{1-\beta_1} \Delta \Gamma_t m_{t-1} \right\rangle \right]}_{II} \\
& \quad + \underbrace{\frac{\eta^2 L}{2} \mathbb{E} \left[\left\| \Gamma_t g_t - \frac{\beta_1}{1-\beta_1} \Delta \Gamma_t m_{t-1} \right\|^2 \right]}_{III} + \underbrace{\eta \mathbb{E}[\langle \nabla f(\theta_t) - \nabla f(x_t), \Gamma_t g_t \rangle]}_{IV}. \tag{5}
\end{aligned}$$

637 In the following, we bound all the four terms mentioned above.

638 **Bounding term I.** Let $\|\Delta \Gamma_t\|$ be the operator norm (with respect to ℓ_2 norm) of the matrix $\Delta \Gamma_t$.
639 Since $\Delta \Gamma_t$ is diagonal, the spectral norm coincides with the largest diagonal value in magnitude.
640 Using unbiasedness of the stochastic gradients, we have

$$\begin{aligned}
I &= -\eta \mathbb{E}[\langle \nabla f(\theta_t), \Gamma_{t-1} g_t \rangle] - \eta \mathbb{E}[\langle \nabla f(\theta_t), \Delta \Gamma_t g_t \rangle] \\
&\leq -\eta \mathbb{E} \left[\left\langle \nabla f(\theta_t), \Gamma_{t-1} \nabla f(\hat{\theta}_t) \right\rangle \right] + \eta G^2 \mathbb{E}[\|\Delta \Gamma_t\|] \\
&= -\eta \mathbb{E} \left[\left\langle \nabla f(\hat{\theta}_t), \Gamma_{t-1} \nabla f(\hat{\theta}_t) \right\rangle \right] + \eta \mathbb{E} \left[\left\langle \nabla f(\hat{\theta}_t) - \nabla f(\theta_t), \Gamma_{t-1} \nabla f(\hat{\theta}_t) \right\rangle \right] + \eta G^2 \mathbb{E}[\|\Delta \Gamma_t\|] \\
&\leq -\eta \lambda_{\min}(\Gamma_{t-1}) \mathbb{E}[\|\nabla f(\hat{\theta}_t)\|^2] + \eta L G \mathbb{E}[\|\Gamma_{t-1}\| \|\hat{\theta}_t - \theta_t\|] + \eta G^2 \mathbb{E}[\|\Delta \Gamma_t\|] \\
&\leq -\frac{\eta}{C_0} \mathbb{E}[\|\nabla f(\hat{\theta}_t)\|^2] + \eta L G \mathbb{E}[\|\hat{\theta}_t - \theta_t\| \cdot \|\Gamma_{t-1}\|] + \eta G^2 \mathbb{E}[\|\Delta \Gamma_t\|], \tag{6}
\end{aligned}$$

where we used Assumption 2 and Lemma 3 to bound

$$\lambda_{\min}(\Gamma_{t-1}) \geq (\|\tilde{v}_{t-1}\|_{\max} + \epsilon)^{-1/2} \geq (G^2 + \epsilon)^{-1/2} \stackrel{\text{def}}{=} \frac{1}{C_0}.$$

641 **Bounding term II.** Splitting the inner product again and bounded each term, we get

$$\begin{aligned}
II &= \eta \mathbb{E} \left[\left\langle \nabla f(\theta_t), \frac{\beta_1}{1-\beta_1} \Delta \Gamma_t m_{t-1} \right\rangle \right] + \eta \mathbb{E} \left[\left\langle \nabla f(x_t) - \nabla f(\theta_t), \frac{\beta_1}{1-\beta_1} \Delta \Gamma_t m_{t-1} \right\rangle \right] \\
&\leq \frac{\eta \beta_1}{1-\beta_1} \mathbb{E}[\|\nabla f(\theta_t)\| \|\Delta \Gamma_t m_{t-1}\|] + \frac{\eta^2 L \beta_1^2}{(1-\beta_1)^2} \mathbb{E}[\|\Gamma_{t-1} m_{t-1}\| \cdot \|\Delta \Gamma_t m_{t-1}\|] \\
&\leq \frac{\eta \beta_1}{1-\beta_1} G^2 \mathbb{E}[\|\Delta \Gamma_t\|] + \frac{\eta^2 \beta_1^2 L G^2}{(1-\beta_1)^2 \sqrt{\epsilon}} \mathbb{E}[\|\Delta \Gamma_t\|], \tag{7}
\end{aligned}$$

642 where we used the fact that the largest eigenvalue $\lambda_{\max}(\Gamma_t) = \|\Gamma_t\| = (\|\tilde{v}_t\|_{\min} + \epsilon)^{-1/2} \leq \epsilon^{-1/2}$.
643 The second inequality is due to the smoothness of f , and the last inequality is due to Lemma 1,
644 Assumption 2 and the property of norms.

645 **Bounding term III.** This term can be bounded as follows:

$$\begin{aligned}
III &\leq \eta^2 L \mathbb{E}[\|\Gamma_t g_t\|^2] + \frac{\eta^2 L \beta_1^2}{(1-\beta_1)^2} \mathbb{E}[\|\Delta \Gamma_t m_{t-1}\|^2] \\
&\leq \frac{\eta^2 L}{\epsilon} \mathbb{E}[\|g_t - \nabla f(\hat{\theta}_t) + \nabla f(\hat{\theta}_t)\|^2] + \frac{\eta^2 L \beta_1^2}{(1-\beta_1)^2} \mathbb{E}[\|\Delta \Gamma_t m_{t-1}\|^2] \\
&\leq \frac{\eta^2 L}{\epsilon} (\mathbb{E}[\|\nabla f(\hat{\theta}_t)\|^2] + \sigma^2) + \frac{\eta^2 L \beta_1^2 G^2}{(1-\beta_1)^2} \mathbb{E}[\|\Delta \Gamma_t\|^2] \\
&\leq \frac{\eta^2 L}{\epsilon} \mathbb{E}[\|\nabla f(\hat{\theta}_t)\|^2] + \frac{\eta^2 L \sigma^2}{\epsilon} + \frac{\eta^2 L \beta_1^2 G^2}{(1-\beta_1)^2} \mathbb{E}[\|\Delta \Gamma_t\|^2], \tag{8}
\end{aligned}$$

646 where we used Assumption 3 that g_t is unbiased with bounded variance σ^2 .

647 **Bounding term IV.** Finally, for the fourth term, we have

$$\begin{aligned}
IV &= \eta \mathbb{E}[\langle \nabla f(\theta_t) - \nabla f(x_t), \Gamma_{t-1} g_t \rangle] + \eta \mathbb{E}[\langle \nabla f(\theta_t) - \nabla f(x_t), \Delta \Gamma_t g_t \rangle] \\
&\leq \eta \mathbb{E}\left[\left\langle \nabla f(\theta_t) - \nabla f(x_t), \Gamma_{t-1} \nabla f(\hat{\theta}_t) \right\rangle\right] + \frac{\eta^2 L \beta_1}{1 - \beta_1} \mathbb{E}[\|\Gamma_t m_{t-1}\| \|\Delta \Gamma_t g_t\|] \\
&\stackrel{(a)}{\leq} \frac{\eta \rho}{2\epsilon} \mathbb{E}[\|\nabla f(\hat{\theta}_t)\|^2] + \frac{\eta}{2\rho} \mathbb{E}[\|\nabla f(\theta_t) - \nabla f(x_t)\|^2] + \frac{\eta^2 \beta_1 L G^2}{(1 - \beta_1)\sqrt{\epsilon}} \mathbb{E}[\|\Delta \Gamma_t\|] \\
&\stackrel{(b)}{\leq} \frac{\eta \rho}{2\epsilon} \mathbb{E}[\|\nabla f(\hat{\theta}_t)\|^2] + \frac{\eta^3 \beta_1^2 L^2}{2(1 - \beta_1)^2 \rho} \mathbb{E}[\|\Gamma_t m_{t-1}\|^2] + \frac{\eta^2 \beta_1 L G^2}{(1 - \beta_1)\sqrt{\epsilon}} \mathbb{E}[\|\Delta \Gamma_t\|] \\
&\leq \frac{\eta \rho}{2\epsilon} \mathbb{E}[\|\nabla f(\hat{\theta}_t)\|^2] + \frac{\eta^3 \beta_1^2 L^2}{2(1 - \beta_1)^2 \rho \epsilon} \mathbb{E}[\|m_{t-1}\|^2] + \frac{\eta^2 L \beta_1 G^2}{(1 - \beta_1)\sqrt{\epsilon}} \mathbb{E}[\|\Delta \Gamma_t\|], \tag{9}
\end{aligned}$$

648 where (a) is due to Young's inequality and (b) is based on Assumption 1. Now integrating (6), (7),
649 (8), (9) into (5),

$$\begin{aligned}
I &\leq -\frac{\eta}{C_0} \mathbb{E}[\|\nabla f(\hat{\theta}_t)\|^2] + \eta L G \mathbb{E}[\|\hat{\theta}_t - \theta_t\| \cdot \|\Gamma_{t-1}\|] + \eta G^2 \mathbb{E}[\|\Delta \Gamma_t\|] \\
II &\leq \frac{\eta \beta_1}{1 - \beta_1} G^2 \mathbb{E}[\|\Delta \Gamma_t\|] + \frac{\eta^2 \beta_1^2 L G^2}{(1 - \beta_1)^2 \sqrt{\epsilon}} \mathbb{E}[\|\Delta \Gamma_t\|] \\
III &\leq \frac{\eta^2 L}{\epsilon} \mathbb{E}[\|\nabla f(\hat{\theta}_t)\|^2] + \frac{\eta^2 L \sigma^2}{\epsilon} + \frac{\eta^2 \beta_1^2 L G^2}{(1 - \beta_1)^2} \mathbb{E}[\|\Delta \Gamma_t\|^2] \\
IV &\leq \frac{\eta \rho}{2\epsilon} \mathbb{E}[\|\nabla f(\hat{\theta}_t)\|^2] + \frac{\eta^3 \beta_1^2 L^2}{2(1 - \beta_1)^2 \rho \epsilon} \mathbb{E}[\|m_{t-1}\|^2] + \frac{\eta^2 L \beta_1 G^2}{(1 - \beta_1)\sqrt{\epsilon}} \mathbb{E}[\|\Delta \Gamma_t\|],
\end{aligned}$$

650 and taking the telescoping summation over $t = 1, \dots, T$, we obtain

$$\begin{aligned}
&\mathbb{E}[f(x_{T+1})] - \mathbb{E}[f(x_1)] \\
&\leq \left(-\frac{\eta}{C_0} + \frac{\eta^2 L}{\epsilon} + \frac{\eta \rho}{2\epsilon}\right) \sum_{t=1}^T \mathbb{E}[\|\nabla f(\hat{\theta}_t)\|^2] + \frac{T \eta^2 L \sigma^2}{\epsilon} + \frac{\eta^3 \beta_1^2 L^2}{2(1 - \beta_1)^2 \rho \epsilon} \sum_{t=1}^T \mathbb{E}[\|m_{t-1}\|^2] \\
&\quad + \left(\frac{\eta G^2}{1 - \beta_1} + \frac{\eta^2 \beta_1 L G^2}{(1 - \beta_1)^2 \sqrt{\epsilon}}\right) \sum_{t=1}^T \mathbb{E}[\|\Delta \Gamma_t\|] + \frac{\eta^2 \beta_1^2 L G^2}{(1 - \beta_1)^2} \sum_{t=1}^T \mathbb{E}[\|\Delta \Gamma_t\|^2] + \frac{\eta L G}{T} \sum_{t=1}^T \mathbb{E}[\|\hat{\theta}_t - \theta_t\| \cdot \|\Gamma_{t-1}\|] \\
&\leq \left(-\frac{\eta}{C_0} + \frac{\eta^2 L}{\epsilon} + \frac{\eta \rho}{2\epsilon} + \frac{\eta^3 \beta_1^2 L^2}{2(1 - \beta_1)^2 \rho \epsilon}\right) \sum_{t=1}^T \mathbb{E}[\|\nabla f(\hat{\theta}_t)\|^2] + \frac{T \eta^2 L \sigma^2}{\epsilon} + \frac{T \eta^3 L^2 \beta_1^2 \sigma^2}{2(1 - \beta_1)^2 \rho \epsilon} \\
&\quad + \left(\frac{\eta G^2}{1 - \beta_1} + \frac{\eta^2 \beta_1 L G^2}{(1 - \beta_1)^2 \sqrt{\epsilon}}\right) \sum_{t=1}^T \mathbb{E}[\|\Delta \Gamma_t\|] + \frac{\eta^2 \beta_1^2 L G^2}{(1 - \beta_1)^2} \sum_{t=1}^T \mathbb{E}[\|\Delta \Gamma_t\|^2] + \frac{\eta L G}{T} \sum_{t=1}^T \mathbb{E}[\|\hat{\theta}_t - \theta_t\| \cdot \|\Gamma_{t-1}\|],
\end{aligned}$$

651 where we used Lemma 1. Choosing $\rho = \frac{\epsilon}{2C_0}$ and $\eta \leq \eta_0 \stackrel{\text{def}}{=} \frac{\epsilon(1-\beta_1)}{4LC_0}$ and using Lemma 2, we get

$$\begin{aligned}
\mathbb{E}[f(x_{T+1}) - f(x_1)] &\leq -\frac{\eta}{2C_0} \sum_{t=1}^T \mathbb{E}[\|\nabla f(\hat{\theta}_t)\|^2] + \frac{T \eta^2 L \sigma^2}{\epsilon} + \frac{T \eta^3 L^2 C_0 \beta_1^2 \sigma^2}{(1 - \beta_1)^2 \epsilon^2} \\
&\quad + \frac{2\eta G^2}{(1 - \beta_1)\sqrt{\epsilon}} + \frac{4\eta^2 \beta_1 L G^2}{(1 - \beta_1)^2 \epsilon} + \frac{\eta L G}{T} \sum_{t=1}^T \mathbb{E}[\|\hat{\theta}_t - \theta_t\| \cdot \|\Gamma_{t-1}\|].
\end{aligned}$$

652 Re-arranging terms, we get

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \mathbb{E}[\|\nabla f(\hat{\theta}_t)\|^2] &\leq 2C_0 \left(\frac{f(\theta_1) - f^*}{T\eta} + \frac{\eta L \sigma^2}{\epsilon} + \frac{\eta^2 L^2 C_0 \beta_1^2 \sigma^2}{(1 - \beta_1)^2 \epsilon^2} \right) \\ &\quad + 4C_0 \left(\frac{G^2}{T(1 - \beta_1)\sqrt{\epsilon}} + \frac{\eta \beta_1 L G^2}{T(1 - \beta_1)^2 \epsilon} \right) + \frac{2C_0 L G}{T} \sum_{t=1}^T \mathbb{E} \left[\frac{\|\hat{\theta}_t - \theta_t\|_2}{\sqrt{\epsilon + \|\tilde{v}_{t-1}\|_{\min}}} \right], \end{aligned}$$

653 where in the last inequality we used $x_1 = \theta_1$ and the lower bound $f^* \leq f(\theta)$ for all $\theta \in \mathbb{R}^d$. Finally,
 654 choosing $\eta = \min(\eta_0, \frac{1}{\sqrt{T}})$ and considering the two cases, we continue

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T \mathbb{E}[\|\nabla f(\hat{\theta}_t)\|^2] &\leq 2C_0 \left(\max \left(1, \frac{1}{\eta_0 \sqrt{T}} \right) \frac{f(\theta_1) - f^*}{\sqrt{T}} + \frac{L \sigma^2}{\epsilon \sqrt{T}} + \frac{L^2 C_0 \beta_1^2 \sigma^2}{(1 - \beta_1)^2 \epsilon^2 T} \right) \\ &\quad + 4C_0 \left(\frac{G^2}{T(1 - \beta_1)\sqrt{\epsilon}} + \frac{\beta_1 L G^2}{T^{3/2}(1 - \beta_1)^2 \epsilon} \right) + \frac{2C_0 L G}{T} \sum_{t=1}^T \mathbb{E} \left[\frac{\|\hat{\theta}_t - \theta_t\|_2}{\sqrt{\epsilon + \|\tilde{v}_{t-1}\|_{\min}}} \right] \\ &\leq 2C_0 \left(\frac{f(\theta_1) - f^*}{\sqrt{T}} + \frac{L \sigma^2}{\epsilon \sqrt{T}} + \frac{L^2 C_0 \beta_1^2 \sigma^2}{(1 - \beta_1)^2 \epsilon^2 T} \right) \\ &\quad + 4C_0 \left(\frac{f(\theta_1) - f^*}{2\eta_0 T} + \frac{G^2}{T(1 - \beta_1)\sqrt{\epsilon}} + \frac{\beta_1 L G^2}{T^{3/2}(1 - \beta_1)^2 \epsilon} \right) \\ &\quad + \frac{2C_0 L G}{\sqrt{\epsilon}} \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\hat{\theta}_t - \theta_t\|_2 \right], \end{aligned}$$

655 Using the bounds $G \leq \sqrt{N} G_\infty$, $C_0 \leq \frac{\sqrt{N}}{2} C$ and surpressing higher order terms, we simplify the
 656 bound to

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}[\|\nabla f(\hat{\theta}_t)\|^2] \leq \frac{C L G_\infty}{\sqrt{\epsilon}} \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\hat{\theta}_t - \theta_t\|_2 \right] \cdot N + \frac{C \sqrt{N}}{\sqrt{T}} \left(f(\theta_1) - f^* + \frac{L \sigma^2}{\epsilon} \right) + \mathcal{O} \left(\frac{N^{3/2}}{T} \right),$$

657 which completes the proof of the theorem. \square

658 **Lemma 1.** For any $t \geq 1$ the following bounds hold:

$$\|m_t\| \leq G, \quad \sum_{t=1}^T \mathbb{E}[\|m_t\|^2] \leq T \sigma^2 + \sum_{t=1}^T \mathbb{E}[\|\nabla f(\hat{\theta}_t)\|^2] \quad (10)$$

659 *Proof.* Let us start with the proof of the first bound on m_t .

$$\begin{aligned} \|m_{t+1}\|^2 &= \|\beta_1 m_t + (1 - \beta_1) g_{t+1}\|^2 \\ &\leq \beta_1 \|m_t\|^2 + (1 - \beta_1) \|g_{t+1}\|^2 \\ &\leq \beta_1^t \|m_1\| + (1 - \beta_1) \sum_{\tau=2}^{t+1} \beta_1^{t+1-\tau} \|g_\tau\|^2 = (1 - \beta_1) \sum_{\tau=1}^{t+1} \beta_1^{t+1-\tau} \|g_\tau\|^2. \end{aligned}$$

Using the bounded gradient assumption, we get

$$\|m_t\|^2 \leq (1 - \beta_1) G^2 \sum_{\tau=1}^t \beta_1^{t-\tau} \leq G^2.$$

660 To derive the bound with expectation, we apply Cauchy-Schwartz inequality and the bounded variance
 661 assumption:

$$\begin{aligned}
 \sum_{t=1}^T \mathbb{E} [\|m_t\|^2] &\leq (1 - \beta_1) \sum_{t=1}^T \sum_{\tau=1}^t \beta_1^{t-\tau} \mathbb{E} [\|g_\tau\|^2] \\
 &\leq \sum_{t=1}^T \mathbb{E} [\|g_t\|^2] = \sum_{t=1}^T \mathbb{E} [\|g_t - \nabla f(\hat{\theta}_t) + \nabla f(\hat{\theta}_t)\|^2] \\
 &\leq \sum_{t=1}^T \left(\sigma^2 + \mathbb{E} [\|\nabla f(\hat{\theta}_t)\|^2] \right) = T\sigma^2 + \sum_{t=1}^T \mathbb{E} [\|\nabla f(\hat{\theta}_t)\|^2].
 \end{aligned}$$

662

□

663 **Lemma 2.** For $\Delta\Gamma_t = \Gamma_{t-1} - \Gamma_t$ we have

$$\sum_{t=1}^T \|\Delta\Gamma_t\| \leq \frac{1}{\sqrt{\epsilon}}, \quad \sum_{t=1}^T \|\Delta\Gamma_t\|^2 \leq \frac{1}{\epsilon}.$$

664 *Proof.* From the definitions of $\Gamma_t = \text{Diag}^{-1/2}(\tilde{v}_t + \epsilon)$ and $\tilde{v}_t = \max(v_t, \tilde{v}_{t-1})$ imply that $\Delta\Gamma_t =$
 665 $\Gamma_{t-1} - \Gamma_t$ is positive semidefinite. Hence, $\|\Delta\Gamma_t\| = \lambda_{\max}(\Delta\Gamma_t) \geq 0$. Using the convexity of λ_{\max}
 666 over symmetric matrices, we get

$$\begin{aligned}
 \sum_{t=1}^T \|\Delta\Gamma_t\| &= \max_i \sum_{t=1}^T \Delta\Gamma_{t,i} \\
 &= \max_i \sum_{t=1}^T \left(\frac{1}{\sqrt{\tilde{v}_{t-1,i} + \epsilon}} - \frac{1}{\sqrt{\tilde{v}_{t,i} + \epsilon}} \right) = \max_i \left(\frac{1}{\sqrt{\tilde{v}_{0,i} + \epsilon}} - \frac{1}{\sqrt{\tilde{v}_{T,i} + \epsilon}} \right) \leq \frac{1}{\sqrt{\epsilon}}
 \end{aligned}$$

667 For the second sum of squared norms, notice that for scalars $a \geq b \geq 0$, it holds that

$$(a - b)^2 \leq (a - b)(a + b) = a^2 - b^2.$$

668 Therefore, the above derivation can be repeated without the square roots as follows:

$$\begin{aligned}
 \sum_{t=1}^T \|\Delta\Gamma_t\|^2 &= \max_i \sum_{t=1}^T \Delta\Gamma_{t,i}^2 \\
 &= \max_i \sum_{t=1}^T \left(\frac{1}{\sqrt{\tilde{v}_{t-1,i} + \epsilon}} - \frac{1}{\sqrt{\tilde{v}_{t,i} + \epsilon}} \right)^2 \\
 &= \max_i \sum_{t=1}^T \left(\frac{1}{\tilde{v}_{t-1,i} + \epsilon} - \frac{1}{\tilde{v}_{t,i} + \epsilon} \right) = \max_i \left(\frac{1}{\tilde{v}_{0,i} + \epsilon} - \frac{1}{\tilde{v}_{T,i} + \epsilon} \right) \leq \frac{1}{\epsilon},
 \end{aligned}$$

669 which completes the proof.

□

Lemma 3. For all iterates $t \geq 1$ the following bound holds

$$\|\tilde{v}_t\|_\infty \leq G^2.$$

670 *Proof.* From the update rules we get the bound for v_t using the initialization $v_0 = 0$:

$$\begin{aligned}
 \|v_t\|_\infty \leq \|v_t\|_1 &\leq \beta_2 \|v_{t-1}\|_1 + (1 - \beta_2) \|g_t\|^2 \\
 &\leq \beta_2 \|v_{t-1}\|_1 + (1 - \beta_2) G^2 \\
 &\leq \beta_2^t \|v_0\|_1 + (1 - \beta_2) G^2 \sum_{\tau=0}^{t-1} \beta_2^\tau \leq G^2.
 \end{aligned}$$

Hence, using the update rule of \tilde{v}_t and initialization $\tilde{v}_0 = 0$, we conclude

$$\|\tilde{v}_t\|_\infty \leq \max(\|v_t\|_\infty, \|\tilde{v}_{t-1}\|_\infty) \leq G^2.$$

671

□

Next, we simplify the optimization setup by considering SGD optimizer over (still generally non-convex) quadratics. In this special case, we provide improved and generally optimal asymptotic convergence rate. Moreover, we do not use the bounded gradient condition (i.e., $\|g_t\|_\infty \leq G_\infty$) of Assumption 2 in this analysis.

More formally, consider iterates $\theta_{t+1} = \theta_t - \eta \tilde{\nabla}_\theta f(\hat{\theta}_t)$, where $\hat{\theta}_t = \mathcal{C}(\theta_t)$ is the compressed model. Suppose that the loss function is quadratic with Hessian matrix $\mathbf{H} \in \mathbb{R}^{N \times N}$ and our compression scheme $\mathcal{C}: \mathbb{R}^N \rightarrow \mathbb{R}^N$ is unbiased, namely $\mathbb{E}_t[\hat{\theta}_t] = \theta_t$. Since the loss is quadratic, we have

$$\nabla f(\hat{\theta}_t) = \nabla f(\theta_t + (\hat{\theta}_t - \theta_t)) = \nabla f(\theta_t) + \mathbf{H}(\hat{\theta}_t - \theta_t).$$

Denote by $\mathbb{E}_t = \mathbb{E}[\cdot | \theta_t]$ the conditional expectation conditioned on iterate θ_t , and apply unbiasedness of the compression to get

$$\mathbb{E}_t \|\nabla f(\hat{\theta}_t)\|^2 = \|\nabla f(\theta_t)\|^2 + \mathbb{E}_t \|\mathbf{H}(\hat{\theta}_t - \theta_t)\|^2 \quad (11)$$

Therefore,

$$\begin{aligned} & \mathbb{E}_t[f(\theta_{t+1}) - f^*] \\ & \leq (f(\theta_t) - f^*) - \eta \mathbb{E}_t[\langle \nabla f(\theta_t), \tilde{\nabla} f(\hat{\theta}_t) \rangle] + \frac{L\eta^2}{2} \mathbb{E}_t[\|\tilde{\nabla} f(\hat{\theta}_t)\|^2] \\ & \leq (f(\theta_t) - f^*) - \eta \mathbb{E}_t[\langle \nabla f(\theta_t), \nabla f(\hat{\theta}_t) \rangle] + \frac{L\eta^2}{2} \mathbb{E}_t[\|\nabla f(\hat{\theta}_t)\|^2] + \frac{L\eta^2}{2} \sigma^2 \\ & = (f(\theta_t) - f^*) - \eta \mathbb{E}_t[\|\nabla f(\theta_t)\|^2] + \frac{L\eta^2}{2} \mathbb{E}_t[\|\nabla f(\theta_t)\|^2] + \frac{L\eta^2}{2} \mathbb{E}_t[\|\mathbf{H}(\hat{\theta}_t - \theta_t)\|^2] + \frac{L\eta^2}{2} \sigma^2 \\ & = (f(\theta_t) - f^*) - \eta(1 - L\eta/2) \mathbb{E}_t[\|\nabla f(\theta_t)\|^2] + \frac{L\eta^2}{2} (\mathbb{E}_t[\|\hat{\theta}_t - \theta_t\|_{\mathbf{H}^2}^2] + \sigma^2) \\ & \leq (f(\theta_t) - f^*) - \frac{\eta}{2} \mathbb{E}_t[\|\nabla f(\theta_t)\|^2] + \frac{L\eta^2}{2} (\mathbb{E}_t[\|\hat{\theta}_t - \theta_t\|_{\mathbf{H}^2}^2] + \sigma^2), \end{aligned}$$

where we used $\mathbb{E}_t[\nabla f(\hat{\theta}_t)] = \nabla f(\theta_t)$ due to the unbiasedness of compression and enforced the bound $\eta \leq \frac{1}{L}$ in the last step. Hence,

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}[\|\nabla f(\theta_t)\|^2] \leq \frac{2(f(x_1) - f^*)}{\eta T} + \eta L \left(\mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\hat{\theta}_t - \theta_t\|_{\mathbf{H}^2}^2 \right] + \sigma^2 \right).$$

Choosing the step size $\eta = \min(\frac{1}{L}, \frac{1}{\sqrt{T}})$ and applying L -smoothness, we get $\mathcal{O}(1/\sqrt{T})$ convergence rate for the uncompressed iterates θ_t :

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}[\|\nabla f(\theta_t)\|^2] \leq \frac{1}{\sqrt{T}} \left(2(f(x_1) - f^*) + L\sigma^2 + L^3 \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\hat{\theta}_t - \theta_t\|_2^2 \right] \right) \max \left(1, \frac{L}{\sqrt{T}} \right).$$

For the convergence bound with respect to the compressed iterates $\hat{\theta}_t$, we apply (11) to quantify the exact difference in average gradient norms with the following identity:

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}[\|\nabla f(\hat{\theta}_t)\|_2^2] = \frac{1}{T} \sum_{t=1}^T \mathbb{E}[\|\nabla f(\theta_t)\|_2^2] + \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\mathbf{H}(\hat{\theta}_t - \theta_t)\|_2^2 \right].$$

Thus, a randomly chosen compressed iterate $\hat{\theta}$ from $\{\hat{\theta}_1, \dots, \hat{\theta}_T\}$ satisfies

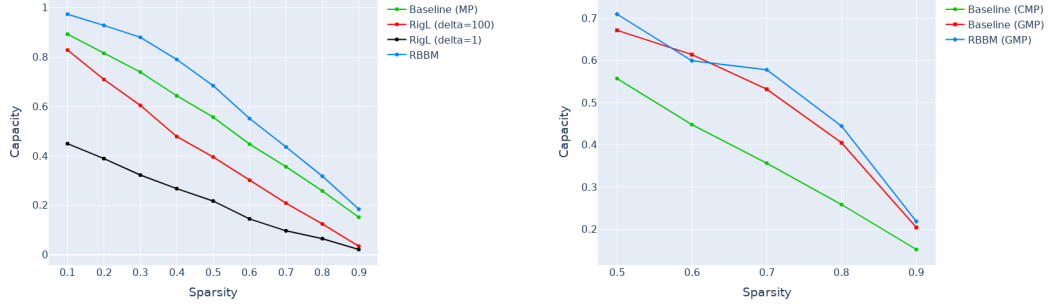
$$\mathbb{E}[\|\nabla f(\hat{\theta})\|^2] \leq L^2 \cdot \mathbb{E} \left[\frac{1}{T} \sum_{t=1}^T \|\hat{\theta}_t - \theta_t\|_2^2 \right] + \mathcal{O} \left(\frac{1}{\sqrt{T}} \right).$$

679 H Improved Sparse Training via RBBM

680 H.1 Comparison against RigL

681 In this subsection we compare our backward mask heuristic in Figure5d with the RigL method of
682 (Evci et al., 2020). We run two instances of RigL: 1) the default one that updates the mask once at

100 steps (i.e. $\Delta = 100$) and updates the mask for the last time at 75% of training and 2) a version of RigL that is closer to our RBBM setup, which changes the mask at each step (i.e. $\Delta = 1$) during the entire training. In Figure 11a we observe that both versions of RigL induce lower capacity than our naive baseline for a fixed sparsity.



(a) Pre-training Llama-30M with different sparsities using our MP baseline, RBBM heuristic and RigL variations.

(b) Pre-training Llama-30M with different sparsities using our constant MP (CMP) baseline, GMP and RBBM heuristic with GMP schedule.

Figure 11: Comparison of sparse training methods for Llama-30M.

H.2 Comparison against Gradual Magnitude Pruning (GMP)

In this section we show our results for applying the GMP sparsity schedule [32] for our setup in Figure 11b. Our first baseline is the constant Magnitude Pruning (CMP), where the backward mask is identical to the forward mask (determined by Top-K) and the sparsity is kept fixed during training. The second baseline is the original GMP where sparsity increases gradually and we compare against the gradual sparsity schedule applied to our **b-rms** heuristic.

We observe our RBBM heuristic with GMP schedule has lower capacity than both CMP and GMP baselines when sparsity is $< 40\%$. However, for sparsities $\geq 40\%$ there is no significant difference in capacity between CMP and GMP schedules.

H.3 Backward Mask Heuristics

In this section we provide more details about our backward heuristics.

Our purpose is to perform sparse training for both forward and backward passes. All models are trained with the same learning rates as in the Quest project.

Notations. Let θ be the model parameters, M_{FW} be the mask for the forward pass, and M_{BW} be the mask for the backward pass.

Forward Pass. The mask for the forward pass is computed using Top-K operator, where K is chosen based on the target sparsity. Supposing Top-K returns the indices of largest entries by magnitude, the i^{th} entry in the forward mask for a tensor x is computed using the indicator function \mathbb{I} as follows:

$$M_{FW}^i(x) = \mathbb{I}[i \in TopK(|x|)] \quad (12)$$

Backward Pass. The mask for the backward pass is computed using a few heuristics described below.

1. **fw:** the backward mask is simply set to the forward mask: $M_{BW} = M_{FW}$. This heuristic allows gradients to flow only through the largest parameters by magnitude selected by Top-K, while the low-magnitude parameters will have zero gradient
2. **rms:** we align the tensor x with the standard normal distribution by dividing x by its root mean square $RMS(x) = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2}$, which results in $\|x/RMS(x)\|_2^2 = n$. For this heuristic, the

712 user sets a median deviation parameter $p \in (0, 0.5)$, which is used to determine the threshold for
 713 the backward mask $T_{RMS}(x, p) = RMS(x) \cdot ppf(0.5 + p)$, where ppf is the inverse cumulative
 714 distribution function of the standard normal distribution (see the `scipy.stats.norm.ppf` function).
 715 The multiplication with $RMS(x)$ has the purpose of converting the threshold for the standard
 716 normal distribution to the threshold for the vector x . As a result, $M_{BW} = |x| > T_{RMS}(x, p)$.

717 3. **banded-rms (b-rms)**: the **rms** heuristic has the property that the absolute values of x that are
 718 larger than the threshold $T_{RMS}(x, p)$ will have value 1 in the mask, while the smaller ones will
 719 have value 0. This banded heuristic determines the backward mask using the threshold $T_{RMS}(x, p)$
 720 computed for the **rms** heuristic in conjunction with the Top-K threshold (which we denote by T_k).
 721 We want to allow gradients to flow for the small parameters and create a band between $T_{RMS}(x, p)$
 722 and T_k where we do not allow gradients. Concretely, the backward mask is set as follows: $M_{BW} =$
 723 $(|x| < \min(T_{RMS}(x, p), T_k)) \vee (\max(T_{RMS}(x, p), T_k) < |x|)$. Since the median deviation p is
 724 a hyper-parameter, we do not have any control over the relationship between $T_{RMS}(x, p)$ and T_k
 725 and we are using the \min and \max functions to make sure the band is valid, e.g. the parameters
 726 do not receive gradient if they lie in the interval $[\min(T_k, T_{RMS}(x, p)), \max(T_k, T_{RMS}(x, p))]$.

727 4. **area-banded-rms (a-b-rms)**: in the **b-rms** heuristic we do not have any control over the re-
 728 lationship between the Top-K threshold T_k and $T_{RMS}(x, p)$. Let us discuss the two possible
 729 cases:

730 (a) $T_k < T_{RMS}(x, p)$: $M_{BW} = (|x| < T_k) \vee (T_{RMS}(x, p) < |x|)$, which means that all
 731 values from x with a lower magnitude than T_k and larger magnitude than $T_{RMS}(x, p)$ will
 732 get gradient, while the values in the range $[T_k, T_{RMS}(x, p)]$ will not receive gradient, even
 733 though they were selected among the Top-K during the forward pass.

734 (b) $T_{RMS}(x, p) < T_k$: $M_{BW} = (|x| < T_{RMS}(x, p)) \vee (T_k < |x|)$, which is the desired case
 735 we developed the **b-rms** heuristic for: the largest entries from x according to the Top-K rule
 736 will receive gradient, as well as the entries smaller than $T_{RMS}(x, p)$. The entries lying in the
 737 interval $[T_{RMS}(x, p), T_k]$ will not receive gradient.

738 We want to make sure that case (a) above does not happen in practice and force the heuristic
 739 to behave as in the case (b). For this, we need to change the way we compute the threshold
 740 $T_{RMS}(x, p)$.

741 The **area-b-rms** heuristic uses the area hyper-parameter $a \in [0, 1]$ (instead of median deviation
 742 p) and expresses the width of the band starting from the Top-K parameter T_k towards zero
 743 to compute the threshold T_a to make sure the condition $T_a < T_k$ always holds. As a result,
 744 $M_{BW} = (|x| < T_a) \vee (T_k < |x|)$. For example, $a = 0$ yields $T_k = T_a$ and this heuristic turns
 745 into **fw**, while $a = 1$ yields $T_a = 0$ and is equivalent to $M_{BW} = \mathbf{1}_d$ (all entries set to 1, meaning
 746 all parameters get gradients). When $a \in (0, 1)$, the parameters smaller than T_a or larger than T_k
 747 get gradients, while the parameters lying in the interval $[T_a, T_k]$ do not get gradients.

748 **How to compute the threshold T_a ?** Compared to the threshold computation for the previous
 749 heuristic, the definition for T_a is slightly more complicated and it was computed graphically. Let f
 750 be the *cdf* function and f^{-1} be the *ppf* function (inverse cdf) for the standard normal distribution.

$$T_a = RMS(x) \cdot f^{-1} \left(0.5 + (1 - a) \cdot \left(f \left(\frac{T_k}{RMS(x)} \right) - 0.5 \right) \right) \quad (13)$$

751 **Explanations for the formula above.** Suppose the Top-K threshold T_k has a corresponding
 752 cdf of 0.8 and we set $a = 0.5$ (which means 50%). We need to set the threshold T_a such that
 753 $(f(T_k) - f(T_a)) / (f(T_k) - 0.5) = a$, where 0.5 is the cdf of the mean (which is identical to the
 754 median for a standard normal distribution). This ratio expresses the length of the band $[0.5, f(T_k)]$
 755 in the cdf space starting from $f(T_k)$ towards the median. As a consequence, the threshold
 756 $T_a = ppf(0.65)$ because the quantile 0.65 is the center of the interval $[0.5, cdf(T_k)] = [0.5, 0.8]$.
 757 The explanations of each term follow:

$$T_a = RMS(x) \cdot f^{-1} \left(\underbrace{0.5 + (1-a) \cdot \left(\underbrace{f \left(\underbrace{\frac{T_k}{RMS(x)}}_{=A} \right) - 0.5}_{=B} \right)}_{=C} \right)_{=D} \underbrace{\phantom{0.5 + (1-a) \cdot \left(f \left(\frac{T_k}{RMS(x)} \right) - 0.5 \right)}}_{=E} \underbrace{\phantom{0.5 + (1-a) \cdot \left(f \left(\frac{T_k}{RMS(x)} \right) - 0.5 \right)}}_{=F}$$

- 758 • **A:** $f = cdf$ computes the corresponding quantile of the Top-K threshold T_k normalized by
759 $RMS(x)$
- 760 • **B:** subtract 0.5 from term A to compute the length of the interval $[0.5, f(T_k^{RMS})]$
- 761 • **C:** multiply by $1 - a$ because we take into consideration the band length that starts at
762 $f(T_k^{RMS})$ towards 0
- 763 • **D:** compute the cdf of T_a by offsetting again by 0.5 (the quantile of the median)
- 764 • **E:** use $ppf = f^{-1}$ to obtain the value that corresponds to $cdf(T_a)$ for the standard normal
765 distribution
- 766 • **F:** multiply by $RMS(x)$ to obtain T_a in the same space as x

767 **Technical note.** One could determine the threshold T_a naively by employing the formula $T_a^{naive} =$
768 $(1 - a)T_k$. Despite simpler, this naive approach leads to a narrower band because the cdf space is
769 non-linear.

770 **Conclusion.** The mask computed using the **a-b-rms** heuristic is more straightforward to under-
771 stand because the parameter a describes the area of the red band (where parameters do not receive
772 gradients) as a percentage of the area between 0 and the Top-K threshold T_k . This heuristic can be
773 used as a replacement for **b-rms** and the parameter a should be tuned, similarly to parameter p for
774 **b-rms**, with the distinction that $a \in [0, 1]$ (for **a-b-rms**) and $p \in (0, 0.5)$ (for **b-rms**).