KL-Aware GPTQ Quantization

Haseeb-26100253, Hamza-26100130

April 2025

1. Empirical Hessian Computation

Given a calibration dataset $\{\boldsymbol{x}_t\}_{t=1}^{T} \subset \mathbb{R}^N$ of input activations to the layer, define the empirical input Hessian

$$\boldsymbol{H} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_t \, \boldsymbol{x}_t^\top \in \mathbb{R}^{N \times N}.$$

This captures the average second-moment of inputs and underlies vanilla GPTQ's reconstruction objective.

2. Hessian Damping and Inversion

For numerical stability, form the damped Hessian

$$\boldsymbol{H}' = \boldsymbol{H} + \lambda \, \frac{\operatorname{tr}(\boldsymbol{H})}{N} \, \boldsymbol{I}_N, \quad \boldsymbol{H}'^{-1} = (\boldsymbol{H}')^{-1}.$$

Factor via Cholesky: $\boldsymbol{H}' = \boldsymbol{L}\,\boldsymbol{L}^\top$, then $\boldsymbol{H}'^{-1} = \boldsymbol{L}^{-\top}\,\boldsymbol{L}^{-1}$.

3. GPTQ Reconstruction Objective

Vanilla GPTQ minimizes the weighted squared error

$$L_{\mathrm{MSE}}(\boldsymbol{Q}) = \big((\boldsymbol{W} - \boldsymbol{Q})^\top \, \boldsymbol{H} \, (\boldsymbol{W} - \boldsymbol{Q})\big).$$

We quantize column-by-column to (approximately) solve this with low complexity.

4. Column-Wise Quantization Loop

For $i = 1, \ldots, N$:

1. Diagonal scale: $h_{ii}^{-1} = (\boldsymbol{H}'^{-1})_{ii}$, set $d_i = \sqrt{h_{ii}^{-1}}$.
2. Quantization: $\boldsymbol{q}_i = d_i \, (\boldsymbol{w}_i / d_i) \in \mathcal{Q}^M$.
3. Error: $\boldsymbol{e}_i = \boldsymbol{w}_i - \boldsymbol{q}_i$.
4. Propagation: for each $j > i$,
$$\boldsymbol{w}_j \;\leftarrow\; \boldsymbol{w}_j - \frac{(\boldsymbol{H}'^{-1})_{ij}}{h_{ii}^{-1}} \, \boldsymbol{e}_i.$$

This costs $O(N^2 M)$ overall once $\boldsymbol{H}'^{-1}$ is available.

5. Activation Ordering (Optional)

Reordering columns by descending $\operatorname{diag}(\boldsymbol{H})$ can improve quantization fidelity; apply inverse permutation after loop.

6. Quantization Error Metric

Compute the average per-column loss

$$\text{Loss}_{\text{avg}} = \frac{1}{N} \sum_{i=1}^{N} \frac{\|\boldsymbol{w}_i^{\text{before}} - \boldsymbol{q}_i\|_2^2}{2\, h_{ii}^{-1}}.$$

7. KL-Augmented Objective

Let p_t be the teacher soft output and q_t the quantized soft output on each calibration input $\boldsymbol{x}_t$:

$$p_t = \operatorname{softmax}\big(\boldsymbol{W}\,\boldsymbol{x}_t/\tau\big), \quad q_t = \operatorname{softmax}\big(\boldsymbol{Q}\,\boldsymbol{x}_t/\tau\big).$$

We form the composite loss

$$L(\boldsymbol{Q}) = L_{\text{MSE}}(\boldsymbol{Q}) + \beta \sum_{t=1}^{T} (p_t \| q_t) \quad (\beta > 0,\ \tau > 0).$$

7.1 Global Second-Order KL Strategy

1. **Distillation Hessian:**
$$A = \sum_{t=1}^{T} \big[\operatorname{diag}(p_t) - p_t p_t^\top\big]\, \boldsymbol{x}_t\, \boldsymbol{x}_t^\top \in \mathbb{R}^{N \times N}.$$
2. **Combined Curvature:**
$$H_{\text{tot}} = \boldsymbol{H} + \beta\, A, \quad H_{\text{tot}} \succ 0.$$
3. **Factor and Scale:** Cholesky $H_{\text{tot}} = LL^\top$, invert to get scales $d_i = \sqrt{(H_{\text{tot}}^{-1})_{ii}}$.
4. **Column-Wise GPTQ:** Run the same loop as Sec. 4, but replace $\boldsymbol{H}$ and $\boldsymbol{H}'^{-1}$ with H_{tot} and H_{tot}^{-1}.

This adds only one extra $O(TN^2)$ pass to build A and reuses the same Cholesky ($O(N^3)$) as vanilla.

7.2 Local First-Order KL Strategy

1. Compute per-column gradient
$$g_i = \frac{\partial}{\partial q_i} \sum_{t=1}^{T} (p_t \| q_t) \in \mathbb{R}^M.$$
2. Build surrogate
$$\ell_i(q_i) \approx \|\boldsymbol{w}_i - q_i\|_{H_{ii}}^2 + \beta\, g_i^\top (q_i - \boldsymbol{w}_i) = \big\|\boldsymbol{w}_i + \tfrac{\beta}{2} H_{ii}^{-1} g_i - q_i\big\|_{H_{ii}}^2.$$
3. Shift quantize: $\tilde{w}_i = \boldsymbol{w}_i + \frac{\beta}{2} H_{ii}^{-1} g_i$, then $\boldsymbol{q}_i = d_i(\tilde{w}_i/d_i)$.
4. Propagate error $\boldsymbol{e}_i = \tilde{w}_i - \boldsymbol{q}_i$ as usual.

Cost remains $O(N^2 M)$ per layer plus one gradient pass $O(TM^2)$ for g_i.

8. Complexity Discussion

Both vanilla GPTQ and the global KL strategy share the same asymptotic costs per layer:

- Hessian assembly: $O(TN^2)$ to compute H (and A for KL).
- Cholesky factorization: $O(N^3)$.
- Column updates: $O(N^2M)$.

Thus, the global KL extension does *not* change the dominant $O(N^3)$ behavior—it only adds an extra Hessian-term build of order $O(TN^2)$, identical to vanilla GPTQ's activation covariance pass.

9. Practical Notes

- Cache all p_t once per layer before column quantization—no interleaved re-evaluation.
- Choose β to balance reconstruction vs. output fidelity; β=0 recovers vanilla GPTQ.
- Use damping λ to ensure $H_{\text{tot}} \succ 0$ when βA might be singular.
- The local strategy offers per-column flexibility at slightly lower overall cost.