BACKPACK CHEAT SHEET

- Assumptions
  - Feedforward network

$$z_n^{(0)} \xrightarrow{T^{(1)}_{\boldsymbol{\theta}^{(1)}}(\boldsymbol{z}_n^{(0)})} \boldsymbol{z}_n^{(1)} \xrightarrow{T^{(2)}_{\boldsymbol{\theta}^{(2)}}(\boldsymbol{z}_n^{(1)})} \dots \xrightarrow{T^{(L)}_{\boldsymbol{\theta}^{(L)}}(\boldsymbol{z}_n^{(L-1)})} \boldsymbol{z}^{(L)} \xrightarrow{\ell(\boldsymbol{z}_n^{(L)}, \boldsymbol{y})} \ell(\boldsymbol{\theta})$$

  - Dimension of parameter $\boldsymbol{\theta}^{(i)}$: $\dim(\boldsymbol{\theta}^{(i)}) = d^{(i)}$
  - Empirical risk

$$\mathcal{L}(\boldsymbol{\theta}) = \tfrac{1}{N}\textstyle\sum_{n=1}^{N} \ell(f(\boldsymbol{\theta}, \boldsymbol{x}_n), \boldsymbol{y}_n).$$

- Shorthands

$$\begin{aligned} \ell_n(\boldsymbol{\theta}) &= \ell(f(\boldsymbol{\theta}, \boldsymbol{x}_n), \boldsymbol{y}_n), \qquad n = 1, \dots, N, \\ f_n(\boldsymbol{\theta}) &= f(\boldsymbol{\theta}, \boldsymbol{x}_n) = \boldsymbol{z}_n^{(L)}(\boldsymbol{\theta}), \qquad n = 1, \dots, N \end{aligned}$$

- Generalized Gauss-Newton matrix

$$\boldsymbol{G}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} (\mathrm{J}_{\boldsymbol{\theta}} f_n)^\top \nabla^2_{f_n} \ell_n(\boldsymbol{\theta}) (\mathrm{J}_{\boldsymbol{\theta}} f_n)$$

- Approximative GGN via MC sampling

$$\tilde{\boldsymbol{G}}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} (\mathrm{J}_{\boldsymbol{\theta}} f_n)^\top \left[\nabla_{\boldsymbol{\theta}} \ell(f_n(\boldsymbol{\theta}), \hat{\boldsymbol{y}}) \nabla_{\boldsymbol{\theta}} \ell(f_n(\boldsymbol{\theta}), \hat{\boldsymbol{y}}_n)^\top\right]_{\hat{y}_n \sim p_{f_n(\boldsymbol{x}_n)}} (\mathrm{J}_{\boldsymbol{\theta}} f_n)$$

Table 5: Overview of the features supported in the first release of BACKPACK. The quantities are computed separately for all module parameters, i.e. $i = 1, \dots, L$.

Feature	**Details**
Individual gradients	$\frac{1}{N}\nabla_{\boldsymbol{\theta}^{(i)}} \ell_n(\boldsymbol{\theta}), \quad n = 1, \dots, N$
Batch variance	$\frac{1}{N}\sum_{n=1}^{N} \left[\nabla_{\boldsymbol{\theta}^{(i)}} \ell_n(\boldsymbol{\theta})\right]_j^2 - \left[\nabla_{\boldsymbol{\theta}^{(i)}} \mathcal{L}(\boldsymbol{\theta})\right]_j^2, \qquad j = 1, \dots, d^{(i)}$
2nd moment	$\frac{1}{N}\sum_{n=1}^{N} \left[\nabla_{\boldsymbol{\theta}^{(i)}} \ell_n(\boldsymbol{\theta})\right]_j^2, \quad j = 1, \dots, d^{(i)}$.
Indiv. gradient L_2 norm	$\left\lVert \frac{1}{N}\nabla_{\boldsymbol{\theta}^{(i)}} \ell_n(\boldsymbol{\theta}) \right\rVert_2^2, \quad n = 1, \dots, N$
DIAGGGN	$\mathrm{diag}\left(\boldsymbol{G}(\boldsymbol{\theta}^{(i)})\right)$
DIAGGGN-MC	$\mathrm{diag}\left(\tilde{\boldsymbol{G}}(\boldsymbol{\theta}^{(i)})\right)$
Hessian diagonal	$\mathrm{diag}\left(\nabla^2_{\boldsymbol{\theta}^{(i)}} \mathcal{L}(\boldsymbol{\theta})\right)$
KFAC	$\tilde{\boldsymbol{G}}(\boldsymbol{\theta}^{(i)}) \approx \boldsymbol{A}^{(i)} \otimes \boldsymbol{B}^{(i)}_{\text{KFAC}}$
KFLR	$\boldsymbol{G}(\boldsymbol{\theta}^{(i)}) \approx \boldsymbol{A}^{(i)} \otimes \boldsymbol{B}^{(i)}_{\text{KFLR}}$
KFRA	$\boldsymbol{G}(\boldsymbol{\theta}^{(i)}) \approx \boldsymbol{A}^{(i)} \otimes \boldsymbol{B}^{(i)}_{\text{KFRA}}$