Plan of the Presentation

- Introduction
- Origin of the model
  - Stochastic Gradient Descent (SGD)
  - Stochastic Weight Averaging (SWA)
  - Assumptions
- Presentation of SWAG
  - Theoretical presentation of SWAG
  - SWAG Diagonal
  - SWAG Low Rank + Diagonal Structure
- Experimental Results
- Conclusion

A Simple Baseline for Bayesian Uncertainty

Algorithm 1 Continuous learning of SWAG model

θ_{pre} : pretrained weights ; η : learning rate ; T : number of steps ; c : moment update frequency ; K : required rank ; S : number of samples

$\overline{\theta} \leftarrow \theta_0, \quad \overline{\theta^2} \leftarrow \theta_0^2$ {Initialize moments}

for $i \leftarrow 1, 2, ..., T$ **do**

  $\theta_i \leftarrow \texttt{SGD}(\theta_{i-1})$ {Perform SGD update}

  if `update_time` $=$ `True` **then**

    $n \leftarrow i/c$ {Number of models}

    $\overline{\theta} \leftarrow \dfrac{n\overline{\theta} + \theta_i}{n+1}, \quad \overline{\theta^2} \leftarrow \dfrac{n\overline{\theta^2} + \theta_i^2}{n+1}$ {Update moments}

    if `nbr_stored_param` $= K$ **then**

      `forget_first_param`

      `store_new_param`$(\theta_i - \overline{\theta})$ {Store deviation}

  if `estimate_time` $=$ `True` **then**

    for $i \leftarrow 1, 2, ..., S$ **do**

      Draw $\widetilde{\theta}_i \sim \mathcal{N}\left(\theta_{\text{SWA}}, \frac{1}{2}\Sigma_{\text{diag}} + \frac{\widehat{D}\widehat{D}^\top}{2(K-1)}\right)$

      $p(y^*|\text{Data}) \mathrel{+}= \frac{1}{S} p(y^*|\widetilde{\theta}_i)$

return $p(y^*|\text{Data})$