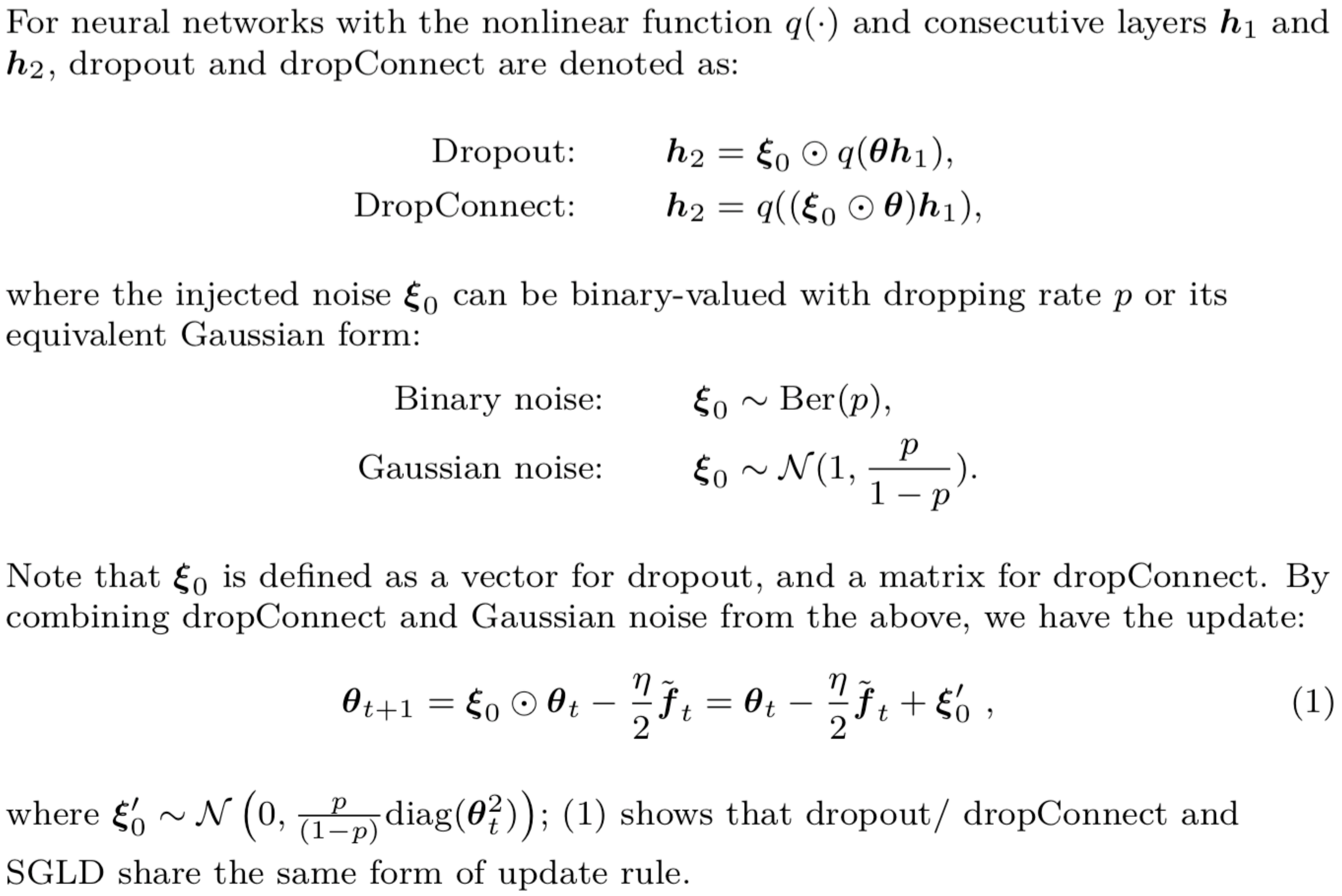
On the Equivalence between SGLD and Dropout

Chunyuan Li

Reference:
Learning Weight Uncertainty with Stochastic Gradient MCMC for Shape Classification, CVPR, 2016
Chunyuan Li, Andrew Stevens, Changyou Chen, Yunchen Pu, Zhe Gan and Lawrence Carin

Formulas on the Equivalence between SGLD and Dropout



For neural networks with the nonlinear function $q(\cdot)$ and consecutive layers $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$, dropout and dropConnect are denoted as:

$$\begin{aligned} \text{Dropout:} \qquad & \boldsymbol{h}_2 = \boldsymbol{\xi}_0 \odot q(\boldsymbol{\theta}\boldsymbol{h}_1), \\ \text{DropConnect:} \qquad & \boldsymbol{h}_2 = q((\boldsymbol{\xi}_0 \odot \boldsymbol{\theta})\boldsymbol{h}_1), \end{aligned}$$

where the injected noise $\boldsymbol{\xi}_0$ can be binary-valued with dropping rate p or its equivalent Gaussian form:

$$\begin{aligned} \text{Binary noise:} \qquad & \boldsymbol{\xi}_0 \sim \text{Ber}(p), \\ \text{Gaussian noise:} \qquad & \boldsymbol{\xi}_0 \sim \mathcal{N}(1, \frac{p}{1-p}). \end{aligned}$$

Note that $\boldsymbol{\xi}_0$ is defined as a vector for dropout, and a matrix for dropConnect. By combining dropConnect and Gaussian noise from the above, we have the update:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\xi}_0 \odot \boldsymbol{\theta}_t - \frac{\eta}{2}\tilde{\boldsymbol{f}}_t = \boldsymbol{\theta}_t - \frac{\eta}{2}\tilde{\boldsymbol{f}}_t + \boldsymbol{\xi}_0' \ , \tag{1}$$

where $\boldsymbol{\xi}_0' \sim \mathcal{N}\left(0, \frac{p}{(1-p)}\text{diag}(\boldsymbol{\theta}_t^2)\right)$; (1) shows that dropout/ dropConnect and SGLD share the same form of update rule.

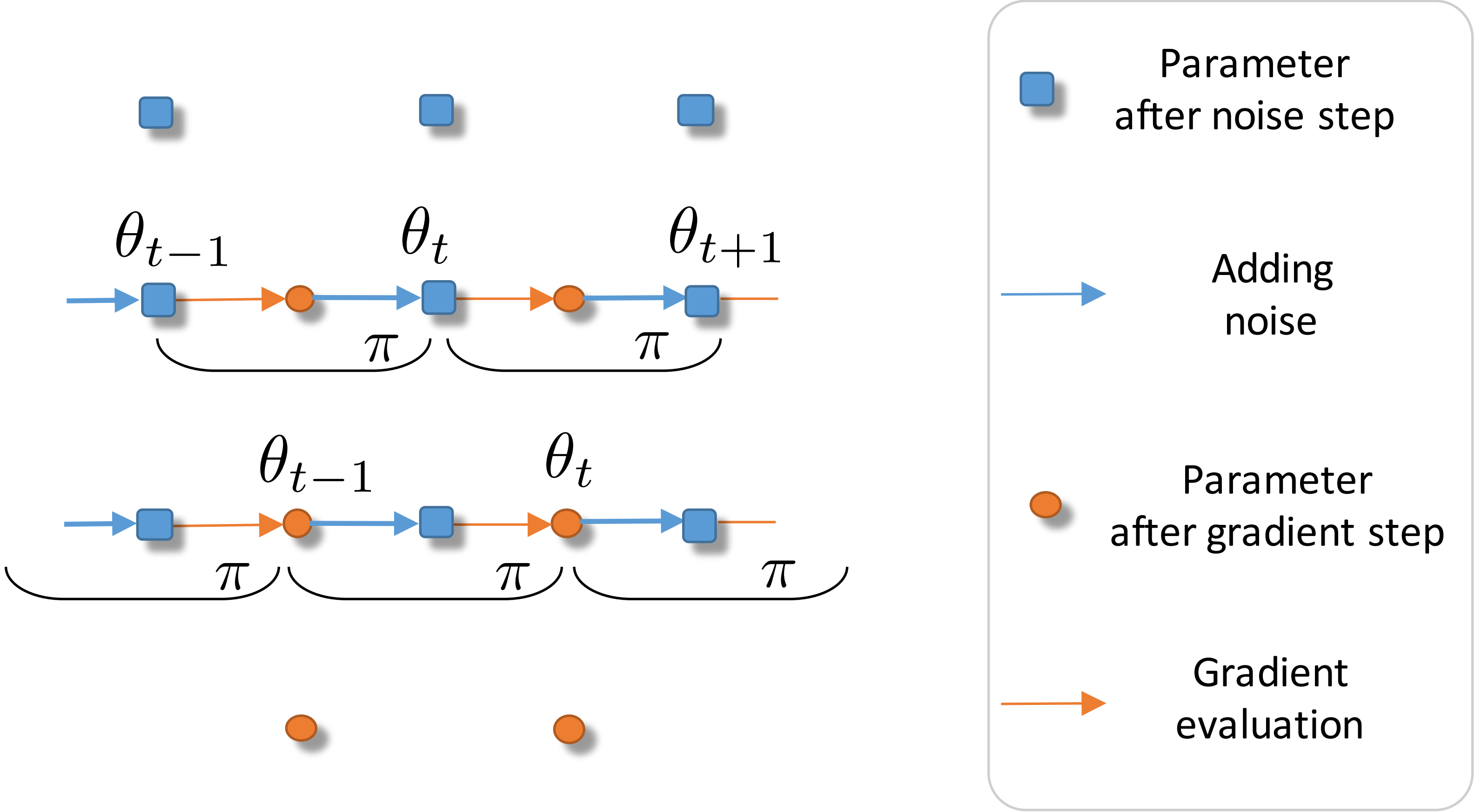
Visual Illustration of Equivalence between SGLD and DropConnect

(a) Collected samples in SGLD

(b) SGLD update iteration

(c) DropConnect update iteration

(d) Collected samples in DropConnect



■ Parameter after noise step

→ Adding noise

● Parameter after gradient step

→ Gradient evaluation

Clarification: (b) and (c) show that SGLD and DropConnect run the same Markov chains if we assume the levels of adding noise are the same. (a) and (d) show that SGLD and DropConnect may collect different samples from the same Markov chains. But one may carefully pick up the samples to make them equivalent.

Note that π indicates one iteration in the respective algorithm.