Appendix

A Implementation Details on MNIST

- **SGD**: We implement a DNN optimized by SGD with the same hyperparameters as in [1]: fixed learning rate $\eta_t = 5 \times 10^{-6}$, prior precision $\lambda = 1$.
- **SGLD**: We implement SGLD sampler and employ it as the teacher of distillation methods. We train SGLD using the same hyperparameters as in [1]: fixed learning rate of $\eta_t = 4 \times 10^{-6}$, thinning interval $\tau = 100$ and prior precision $\lambda = 1$ except the burn-in iterations is $B = 10000$. The same settings are used when SGLD plays the role of teacher model.
- **BBB**: We reimplement BBB with standard Gaussian prior for fair comparison and train BBB using Adam with the default hyperparameters according to validation set.
- **BDK**: We use fixed learning rate of $\rho = 0.005$ and a prior precision of 0.001 for the student model according to original paper [1].
- **APD**: As authors only public the code of offline APD, we reimplement the online APD for further comparison. We save the most recent 100 SGLD samples of network parameters and select randomly a batch of samples for GAN training each time. The mini-batch size and the number of iterations are the same as authors' setting for offline APD.
- **V-BDK**: According to validation set, we select SGD optimizer with learning rate $\rho = 0.01$.
- **BDPK**: According to validation set, we select Adam optimizer with the default hyperparameters.

B Implementation Details on CIFAR10

- **SGD**: $\eta_t = 1 \times 10^{-6}$ and $\lambda = 5 \times 10^{-4}$.
- **SGLD**: $\eta_t = 5 \times 10^{-7}$, $\tau = 100$, $\lambda = 5 \times 10^{-4}$, and $B = 5000$. The same settings are used when SGLD plays the role of teacher model.
- **BBB**: $\eta_t = 0.001$.

- **BDK**: $\rho = 0.03$ and $\lambda = 5 \times 10^{-6}$ for the student model.
- **V-BDK**: $\rho = 0.01$ for the student model.
- **BDPK**: $\rho = 0.01$ for the student model.

C Proof of Equation (9)

Proposition 1 *Let $\mathcal{N}_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ be two multivariate Gaussian Distributions in $\mathbb{R}^n$. Then*

$$\mathrm{KL}(\mathcal{N}_1 \| \mathcal{N}_2) = \frac{1}{2} log \frac{det(\boldsymbol{\Sigma}_2)}{det(\boldsymbol{\Sigma}_1)} - \frac{n}{2} + \frac{1}{2} tr(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) + \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

Proof of Proposition 1

$$\begin{aligned}
&\mathrm{KL}(\mathcal{N}_1 \| \mathcal{N}_2) \\
&= \mathbb{E}_{P_1}\left[log \frac{P_1}{P_2} \right] \\
&= \frac{1}{2}\mathbb{E}_{P_1}\left[logdet(\boldsymbol{\Sigma}_2) + (\boldsymbol{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2) - logdet(\boldsymbol{\Sigma}_1) - (\boldsymbol{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) \right] \\
&= \frac{1}{2} log \frac{det(\boldsymbol{\Sigma}_2)}{det(\boldsymbol{\Sigma}_1)} + \frac{1}{2}\mathbb{E}_{P_1}\left[- tr((\boldsymbol{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1)) + tr((\boldsymbol{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2)) \right] \\
&= \frac{1}{2} log \frac{det(\boldsymbol{\Sigma}_2)}{det(\boldsymbol{\Sigma}_1)} + \frac{1}{2}\mathbb{E}_{P_1}\left[- tr(\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1)(\boldsymbol{x} - \boldsymbol{\mu}_1)^T) + tr(\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2)(\boldsymbol{x} - \boldsymbol{\mu}_2)^T) \right] \\
&= \frac{1}{2} log \frac{det(\boldsymbol{\Sigma}_2)}{det(\boldsymbol{\Sigma}_1)} - \frac{1}{2} tr(\mathbb{E}_{P_1}\left[\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1)(\boldsymbol{x} - \boldsymbol{\mu}_1)^T \right]) + \frac{1}{2}\mathbb{E}_{P_1}\left[tr(\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2)(\boldsymbol{x} - \boldsymbol{\mu}_2)^T) \right] \\
&= \frac{1}{2} log \frac{det(\boldsymbol{\Sigma}_2)}{det(\boldsymbol{\Sigma}_1)} - \frac{1}{2} tr(\mathbb{E}_{P_1}\left[\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_1 \right]) + \frac{1}{2}\mathbb{E}_{P_1}\left[tr(\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}\boldsymbol{x}^T - \boldsymbol{x}\boldsymbol{\mu}_2^T - \boldsymbol{\mu}_2\boldsymbol{x}^T + \boldsymbol{\mu}_2\boldsymbol{\mu}_2^T) \right] \\
&= \frac{1}{2} log \frac{det(\boldsymbol{\Sigma}_2)}{det(\boldsymbol{\Sigma}_1)} - \frac{n}{2} + \frac{1}{2}\mathbb{E}_{P_1}\left[tr(\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}\boldsymbol{x}^T - \boldsymbol{x}\boldsymbol{\mu}_2^T - \boldsymbol{\mu}_2\boldsymbol{x}^T + \boldsymbol{\mu}_2\boldsymbol{\mu}_2^T) \right] \\
&= \frac{1}{2} log \frac{det(\boldsymbol{\Sigma}_2)}{det(\boldsymbol{\Sigma}_1)} - \frac{n}{2} + \frac{1}{2}\mathbb{E}_{P_1}\left[tr(\boldsymbol{\Sigma}_2^{-1}((\boldsymbol{x} - \boldsymbol{\mu}_1)(\boldsymbol{x} - \boldsymbol{\mu}_1)^T + \boldsymbol{x}\boldsymbol{\mu}_1^T + \boldsymbol{\mu}_1\boldsymbol{x}^T - \boldsymbol{\mu}_1\boldsymbol{\mu}_1^T - \boldsymbol{x}\boldsymbol{\mu}_2^T - \boldsymbol{\mu}_2\boldsymbol{x}^T + \boldsymbol{\mu}_2\boldsymbol{\mu}_2^T) \right] \\
&= \frac{1}{2} log \frac{det(\boldsymbol{\Sigma}_2)}{det(\boldsymbol{\Sigma}_1)} - \frac{n}{2} + \frac{1}{2}\mathbb{E}_{P_1}\left[tr(\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\Sigma}_1 + \boldsymbol{\mu}_1\boldsymbol{\mu}_1^T + \boldsymbol{\mu}_1\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_1\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_1\boldsymbol{\mu}_2^T - \boldsymbol{\mu}_2\boldsymbol{\mu}_1^T + \boldsymbol{\mu}_2\boldsymbol{\mu}_2^T) \right] \\
&= \frac{1}{2} log \frac{det(\boldsymbol{\Sigma}_2)}{det(\boldsymbol{\Sigma}_1)} - \frac{n}{2} + \frac{1}{2}\mathbb{E}_{P_1}\left[tr(\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\Sigma}_1 + \boldsymbol{\mu}_1\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_1\boldsymbol{\mu}_2^T - \boldsymbol{\mu}_2\boldsymbol{\mu}_1^T + \boldsymbol{\mu}_2\boldsymbol{\mu}_2^T) \right] \\
&= \frac{1}{2} log \frac{det(\boldsymbol{\Sigma}_2)}{det(\boldsymbol{\Sigma}_1)} - \frac{n}{2} + \frac{1}{2} tr(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) + \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)
\end{aligned}$$

Now return to equation (9). For diagonal Gaussian distribution $\mathcal{N}_1$ and $\mathcal{N}_2$, $\boldsymbol{\mu}_1 = (\mu_1, ..., \mu_n)$, $\boldsymbol{\mu}_2 = (\hat{\mu}_1, ..., \hat{\mu}_n)$, $\boldsymbol{\Sigma}_1 = diag\{\sigma_1, ..., \sigma_n\}$, $\boldsymbol{\Sigma}_1 = diag\{\hat{\sigma}_1, ..., \hat{\sigma}_n\}$.

Using Proposition 1, plug them into the last line and we have that:

$$
\begin{aligned}
\mathrm{KL}(\mathcal{N}_1\|\mathcal{N}_2) &= \frac{1}{2}log\frac{det(\mathbf{\Sigma}_2)}{det(\mathbf{\Sigma}_1)} - \frac{n}{2} + \frac{1}{2}tr(\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1) + \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
&= \frac{1}{2}log\frac{\prod_{i=1}^n \hat{\sigma}_i^2}{\prod_{i=1}^n \sigma_i^2} - \frac{n}{2} + \frac{1}{2}\sum_{i=1}^n \frac{\sigma_i^2}{\hat{\sigma}_i^2} - \frac{1}{2}\sum_{i=1}^n \frac{(\mu_i - \hat{\mu}_i)^2}{\hat{\sigma}_i^2} \\
&= \frac{1}{2}\sum_i(\frac{\sigma_i^2 + (\mu_i - \hat{\mu}_i)^2}{\hat{\sigma}_i^2} - \log\frac{\sigma_i^2}{\hat{\sigma}_i^2} - 1)
\end{aligned}
$$

D Proof of Equation (11)

Proof: For BNN's weight w_i, we have obtained SGLD samples $w_i^1, ..., w_i^{t-1}$ in the first $t-1$ iterations with mean $\hat{\mu}_i^{t-1}$ and standard error $\hat{\sigma}_i^{t-1}$. Now we run SGLD update and get the *t*-th sample w_i^t. We first compute the new mean $\hat{\mu}_i^t$ by equation (10). Then we have:

$$
\begin{aligned}
(\hat{\sigma}_i^t)^2 &= \frac{1}{t}\Big[\sum_{k=1}^{t}(w_i^k - \hat{\mu}_i^t)^2\Big] \\
&= \frac{1}{t}\Big[\sum_{k=1}^{t-1}(w_i^k - \hat{\mu}_i^t)^2 + (w_i^t - \hat{\mu}_i^t)^2\Big] \\
&= \frac{1}{t}\Big[\sum_{k=1}^{t-1}(w_i^k - \hat{\mu}_i^{t-1} + \hat{\mu}_i^{t-1} - \hat{\mu}_i^t)^2 + (w_i^t - \hat{\mu}_i^t)^2\Big] \\
&= \frac{1}{t}\Big[\sum_{k=1}^{t-1}\big[(w_i^k - \hat{\mu}_i^{t-1})^2 - 2(w_i^k - \hat{\mu}_i^{t-1})(\hat{\mu}_i^{t-1} - \hat{\mu}_i^t) + (\hat{\mu}_i^{t-1} - \hat{\mu}_i^t)^2\big] + (w_i^t - \hat{\mu}_i^t)^2\Big] \\
&= \frac{(t-1)[(\hat{\sigma}_i^{t-1})^2 + (\hat{\mu}_i^{t-1} - \hat{\mu}_i^t)^2] + (w_i^t - \hat{\mu}_i^t)^2}{t}
\end{aligned}
$$

Take the square root on the both side and we obtain equation (11).

References

[1] Anoop Korattikara Balan, Vivek Rathod, Kevin P. Murphy, and Max Welling. Bayesian dark knowledge. In *Proceedings of NIPS*, pages 3438–3446, 2015.