H. Analysis of exploration property

In this section, we analyze how L2E can collect diverse set of parameters with a single MCMC chain. In Figure 6, we analyze the behavior of L2E in downstream tasks(CIFAR-10,100) and in Figure 7, we visualize the scale of outputs of α_ϕ and β_ϕ on the regular grid of input values following Gong et al. (2018). Firstly, we plot $l2$ norm of $\Delta\theta = \theta_{t+1} - \theta_t$ at time t and training cross-entropy loss (NLL) for 200 epochs and comparing it with DE and CSGMCMC. Recall the update rule of L2E,

$$\begin{aligned} r_{t+1} &= r_t - \epsilon_t[\nabla_\theta \tilde{U}(\theta_t) + \alpha_\phi(\theta_t, r_t) + C\beta_\phi(\theta_t, r_t)] + \xi_t \\ \theta_{t+1} &= \theta_t + \epsilon_t \beta_\phi(\theta_t, r_{t+1}). \end{aligned} \tag{18}$$

where $\xi_t \sim \mathcal{N}(0, 2C\epsilon_t)$. According to the equation above, β_ϕ is responsible for updating θ, so tracking $l2$ norm of $\Delta\theta$ is same as tracking the $l2$ norm of β_ϕ since $||\beta_\phi||^2 = \frac{||\Delta\theta||^2}{\epsilon^2}$.

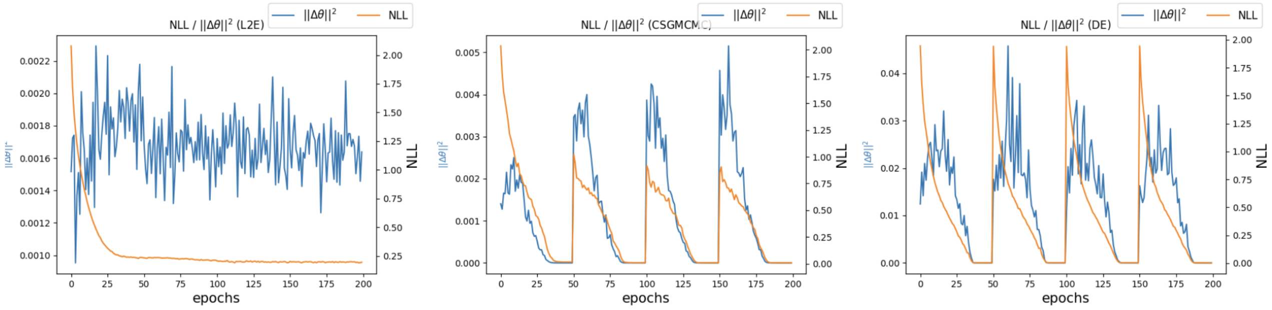

Figure 6. Plots of $||\Delta\theta||^2$ and train NLL during training of L2E,CSGMCMC, DE on CIFAR-10. Unlike other methods, L2E actively updates θ in the local minima while maintaining training NLL as nearly constant.

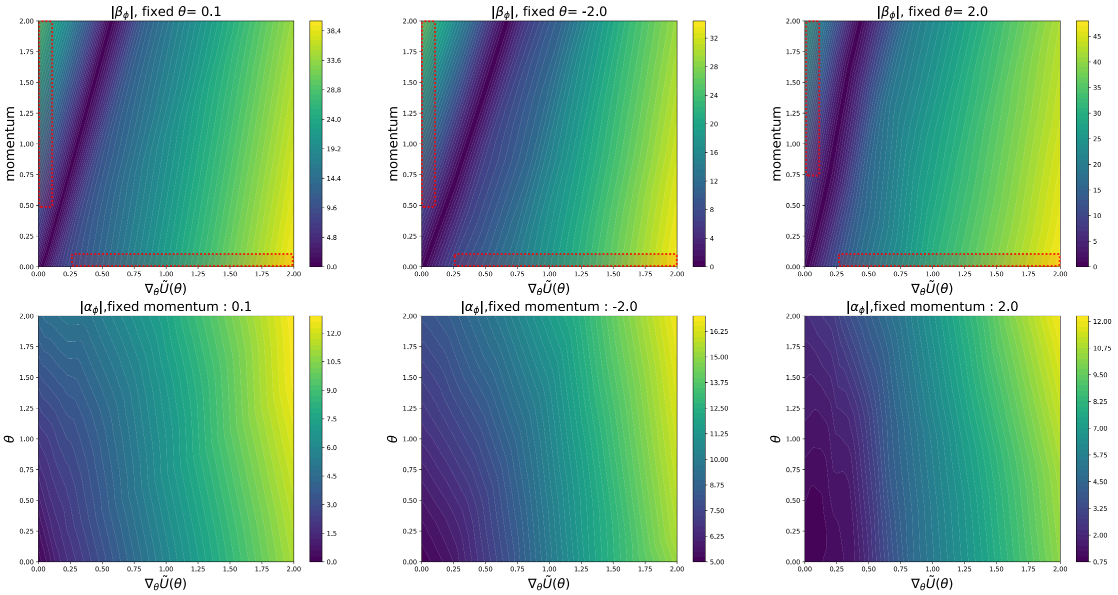

Figure 7. Countour plots of absolute value of outputs of β_ϕ(top) and α_ϕ(bottom) on the grid. β_ϕ produces large magnitude of output when $\nabla_\theta \tilde{U}(\theta)$ is high. When $\nabla_\theta \tilde{U}(\theta)$ gets smaller, the overall magnitude decreases as expected, but even when $\nabla_\theta \tilde{U}(\theta)$ is nearly zero, β_ϕ can still allow the sampler to move around posterior distribution when integrated with high momentum value. The regions marked with red dashed boxes can be beneficial for exploration in high density regions. α_ϕ is proportional to $\nabla_\theta \tilde{U}(\theta)$ in general, which helps the sampler fastly converge to the high density region.

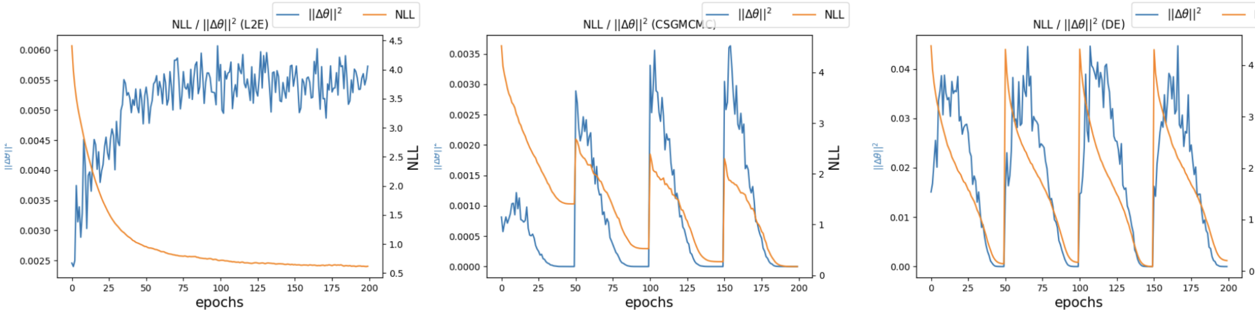

Figure 8. Plots of $||\Delta\theta||^2$ and train NLL during training of L2E,CSGMCMC, DE on CIFAR-100.

In Figures 6 and 8, we find that L2E updates θ with a larger magnitude in local minima than in the early stages of training. This tendency is different from other gradient-based optimizer or MCMC methods where the amount of update is relatively small at local minima. Additionally, we notice that L2E actively updates θ at minima while maintaining loss as nearly constant. This trend is consistently observed in both CIFAR-10 and CIFAR-100, implying that L2E learns some common knowledge of posterior information across tasks for efficient exploration in low loss regions. Various experimental results (e.g., see Figure 2b) support that L2E is good at capturing multi-modalities of BNNs posterior with a single trajectory. Since L2E produces significant amount of updates at local minima without increasing the loss, we can say that our parameterized gradients learned the general knowledge to explore high density regions among different modes.

In Figure 7, we plot absolute value of outputs of α_ϕ and β_ϕ on the regular input grid. Since α_ϕ and β_ϕ take $\theta, r, \nabla_\theta\tilde{U}(\theta)$ and running average of $\nabla_\theta\tilde{U}(\theta)$ as inputs, analyzing the function itself is a complex problem. Therefore, to simplify the analysis, we follow the approach of Gong et al. (2018), where we fix other inputs except for the statistics we are interested in. For β_ϕ, we choose three different fixed values of θ and plot the results, while for α_ϕ, we fix the momentum. We assume that there is no running average of $\nabla_\theta\tilde{U}(\theta)$ so that running average term is fixed to $\nabla_\theta\tilde{U}(\theta)$.

Since the value of θ itself does not encode the information about the posterior landscape, there is no clear distinction among contour plots in top row with different θ values. In general, the scale of outputs of α_ϕ and β_ϕ is proportional to $\nabla_\theta\tilde{U}(\theta)$ which is desirable for fast convergence to high density regions of posterior distribution. For β_ϕ, when $\nabla_\theta\tilde{U}(\theta)$ gets smaller, the overall magnitude of output decreases as expected, but even when $\nabla_\theta\tilde{U}(\theta)$ is nearly zero, β_ϕ can still make large magnitude of update of θ when integrated with high momentum value. Also, when momentum is nearly zero, it still allows sampler to explore posterior distribution when $\nabla_\theta\tilde{U}(\theta)$ is far from zero. This complementary relationship between two statistics is strength of L2E that the movement of sampler is not solely depend on a single statistic so that it can exploit more complex information about loss geometry unlike standard SGHMC. α_ϕ produces an output proportional to the scale of the gradient regardless of the momentum values. This implies that α_ϕ helps the acceleration of sampler in low-density regions as it is added to the energy gradient in Equation 18.

One limitation of our analysis is that we focus on analyzing the magnitude of the function output rather than its direction. Although the magnitude of α_ϕ and β_ϕ are closely connected to the exploration property, delving deeper into how L2E traverses complex and multi-modal BNNs posterior landscape would be an interesting direction of future research.