A Details of The Toy Example

This section discusses the details of the toy example shown in Figure (1). We construct a small 2-dimensional sample set from a mixture of 3 Gaussians, and then binarize the labels by thresholding them from the median value. The sample distribution is shown in Figure 1b. For the model we use a 5-layer MLP with sigmoid as the activation and cross entropy as the loss. There are no bias terms in the linear layers, and the weights are shared. For the shared 2-by-2 linear coefficient matrix, we treat two entries as constants and optimize the other 2 entries. In this way the whole model has only two free parameters w_1 and w_2.

The model is trained using 100 samples. Fixing the samples, we plot the loss function with respect to the model variables $\hat{L}(w_1, w_2)$, as shown in Figure 1a. Many local optima are observed even in this simple two-dimensional toy example. In particular: a sharp one, marked by the vertical green line, and a flat one, marked by the vertical red line. The colors on the loss surface display the values of the generalization metric scores (pacGen) defined in Section 7. Smaller metric value indicates better generalization power.

As displayed in the figure, the metric score around the global optimum, indicated by the vertical green bar, is high, suggesting possible poor generalization capability as compared to the local optimum indicated by the red bar. We also plot a plane on the bottom of the figure. The color projected on the bottom plane indicates an approximated generalization bound, which considers both the loss and the generalization metric.[9] The local optimum indicated by the red bar, though has a slightly higher loss, has a similar overall bound compared to the "sharp" global optimum.

On the other hand, fixing the parameter w_1 and w_2, we may also plot the labels predicted by the model given the samples. Here we plot the prediction from both the sharp minimum (Figure 1c) and the flat minimum (Figure 1d). The sharp minimum, even though it approximates the true label better, has some complex structures in its predicted labels, while the flat minimum seems to produce a simpler classification boundary.

B Details of Figure 2

In this toy example we set the perturbation level σ_i^* of u_i according to Lemma 3. For the local neighbor parameters, we let $\gamma = 5$, and $\epsilon = 0.1$. For each setting we generate the perturbation u and plot $L(w + u)$ as a function of w. The loss is averaged over 200 random draws of u. We vary the level of the perturbation for u and plot the average landscapes.

Note according to (12) when η decreases, the perturbation level σ_i^* tends to increase. Figure 2 shows as the perturbation level increases, the loss surface w.r.t. the parameters gets changed, and the sharp minimum has a higher loss value compared to the flat minimum.

C Truncated Gaussian

Because the Gaussian distribution is not bounded but the inequality (5) requires bounded perturbation, we first truncate the distribution. The procedure of truncation is similar to the proof in (Neyshabur et al., 2018) and (McAllester, 2003).

Let $u_i \sim N(0, \sigma_i^2)$. Denote the truncated Gaussian as $N_{\kappa_i}(0, \sigma_i^2)$. If $\tilde{u}_i \sim N_{\kappa_i}(0, \sigma_i^2)$ then

$$\mathbb{P}_{\kappa_i}(\tilde{u}) = \frac{1}{Z_i} \begin{cases} p(u_i) & \text{if } |u_i| < \kappa_i(w) \\ 0 & o.w. \end{cases} \tag{14}$$

Now let's look at the event

$$\mathbf{E} = \{u \mid |u_i| < \kappa_i(w) \ \ \forall \ \ i\} \tag{15}$$

If $\forall i \ \ \sigma_i < \frac{\kappa_i(w)}{\sqrt{2}\text{erf}^{-1}(\frac{1}{2m})}$, by union bound $\mathbb{P}(\mathbf{E}) \geq 1/2$. Here erf^{-1} is the inverse Gaussian error function defined as $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$, and m is the number of parameters. Following a similar procedure as in the proof of

[9] the bound was approximated with $\eta = 39$ using inequality (11)

Lemma 1 in (Neyshabur et al., 2018),

$$KL(w+\tilde{u}||\pi) \leq 2(KL(w+u||\pi)+1) \tag{16}$$

Suppose the coefficients are bounded such that $\sum_i w_i^2 \leq \tau$, where τ is a constant. Choose the prior π as $N(0, \tau I)$, and we have

$$KL(w+u||\pi) \leq \frac{1}{2}(m\log\tau - \sum_i \log\sigma_i^2 - m + \frac{1}{\tau}\sum_i \sigma_i^2 + 1) \tag{17}$$

Notice that after the truncation the variance only becomes smaller, so the bound of (7) for the truncated Gaussian becomes

$$\begin{aligned}\mathbb{E}_u[L(w+\tilde{u})] \leq \hat{L}(w) + \frac{1}{2}\sum_i \nabla_{i,i}^2 L(w)\sigma_i^2 + \frac{\rho m^{1/2}}{6}\sum_i \kappa_i(w)\sigma_i^2 \\ + \frac{m\log\tau - \sum_i \log\sigma_i^2 - m + \frac{1}{\tau}\sum_i \sigma_i^2 + 1 + 2\log\frac{1}{\delta}}{2\eta} + \frac{\eta}{2n}\end{aligned} \tag{18}$$

Again when $\hat{L}(w)$ is convex around w^* such that $\nabla^2\hat{L}(w^*) \geq 0$, solve for the best σ_i and we get the following lemma:

Lemma 5. *Suppose the loss function $l(f,x,y) \in [0,1]$, and model weights are bounded $\sum_i w_i^2 \leq \tau$. For any $\delta > 0$ and η, with probability at least $1-\delta$ over the draw of n samples, for any $w^* \in \mathbb{R}^m$ such that assumption 1 holds,*

$$\mathbb{E}_u[L(w^*+\tilde{u})] \leq \hat{L}(w^*) + \frac{m\log\tau - \sum_i \log\sigma_i^2 + 1 + 2\log\frac{1}{\delta}}{2\eta} + \frac{\eta}{2n} \tag{19}$$

where $\tilde{u_i} \sim N_{\kappa_i}(0, \sigma_i^*)$ are i.i.d. random variables distributed as truncated Gaussian,

$$\sigma_i^* = \min\left(\sqrt{\frac{1}{\eta\nabla_{i,i}^2\hat{L}(w^*) + \frac{\rho\eta m^{1/2}}{3}\kappa_i(w^*) + \frac{1}{\tau}}}, \frac{\kappa_i(w^*)}{\sqrt{2}\mathrm{erf}^{-1}(\frac{1}{2m})}\right) \tag{20}$$

and σ_i^{*2} is the i-th diagonal element in Σ^*.

Again we have an extra term η, which may be further optimized over a grid to get a tighter bound. In our algorithm we treat η as a hyper-parameter instead.

D Proof of Lemma 3

Proof. We rewrite the inequality (9) below

$$\mathbb{E}_u[L(w+u)] \leq \hat{L}(w) + \frac{1}{6}\sum_i \nabla_{i,i}^2 L(w)\sigma_i^2 + \frac{\rho m^{1/2}}{18}\sum_i \kappa_i(w)\sigma_i^2 + \frac{\sum_i \log\frac{\tau_i}{\sigma_i} + \log\frac{1}{\delta}}{\eta} + \frac{\eta}{2n} \tag{21}$$

The terms related to σ_i on the right hand side of (21) are

$$\frac{1}{6}\nabla_{i,i}^2 L(w)\sigma_i^2 + \frac{\rho m^{1/2}}{18}\kappa_i(w)\sigma_i^2 - \frac{\log\sigma_i}{\eta} \tag{22}$$

Since the assumption is $\nabla_{i,i}^2\hat{L}(w^*) \geq 0$ for all i, $\nabla_{i,i}^2\hat{L}(w) + \rho m^{1/2}\kappa_i(w)/3 > 0$. Solving for σ that minimizes the right hand side of (21), and we have

$$\sigma_i^*(w,\eta,\gamma) = \min\left(\sqrt{\frac{1}{\eta(\nabla_{i,i}^2\hat{L}(w)/3 + \rho m^{1/2}\kappa_i(w)/9)}}, \kappa_i(w)\right) \tag{23}$$

The term $\frac{1}{6}\sum_i \nabla^2_{i,i} L(w)\sigma_i^2 + \frac{\rho m^{1/2}}{18}\sum_i \kappa_i(w)\sigma_i^2$ on the right hand side of (9) is monotonically increasing w.r.t. σ^2, so

$$\begin{aligned}
&\frac{1}{6}\sum_i \nabla^2_{i,i} L(w)\sigma_i^{*2} + \frac{\rho m^{1/2}}{18}\sum_i \kappa_i(w)\sigma_i^{*2} \\
&\leq \sum_i \left(\frac{1}{6}\nabla^2_{i,i} L(w) + \frac{\rho m^{1/2}}{18}\kappa_i(w)\right) \frac{1}{\eta(\nabla^2_{i,i}\hat{L}(w)/3 + \rho m^{1/2}\kappa_i(w)/9)} \\
&= \frac{m}{2\eta}
\end{aligned} \tag{24}$$

Combine the inequality (24), and the equation (23) with (21), and we complete the proof.

□

E Proof of Theorem 2

Proof. Combining (4) and (9), we get

$$\mathbb{E}_u[L(\check{w}+u)] \leq \hat{L}(\check{w}) + \frac{1}{2}\sqrt{\frac{m}{n}} + \frac{\sum_i \log\frac{\tau_i}{\check{\sigma}_i} + \log\frac{1}{\delta}}{\eta} + \frac{\eta}{2n}$$

The following proof is similar to the proof of Theorem 6 in (Seldin et al., 2012a). Note the η in Lemma (3) cannot depend on the data. In order to optimize η we need to build a grid of the form

$$\eta_j = e^j\sqrt{2n\log\frac{1}{\delta_j}}$$

for $j \geq 0$.

For a given value of $\sum_i \log\frac{\tau_i}{\check{\sigma}_i}$, we pick η_j, such that

$$j = \left\lfloor \frac{1}{2}\log\left(\frac{\sum_i \log\frac{\tau_i}{\check{\sigma}_i}}{\log\frac{1}{\delta_j}} + 1\right)\right\rfloor$$

where $\lfloor x \rfloor$ is the largest integer value smaller than x. Set $\delta_j = \delta 2^{-(j+1)}$, and take a weighted union bound over η_j-s with weights $2^{-(j+1)}$, and we have with probability at least $1-\delta$,

$$\mathbb{E}_u[L(\check{w}+u)] \leq \hat{L}(\check{w}) + \frac{1}{2}\sqrt{\frac{m}{n}} + (1+1/e)\sqrt{\frac{\sum_i \log\frac{\tau_i}{\check{\sigma}_i} + \log\frac{1}{\delta} + \frac{\log 2}{2}\left(2 + \log\left(\frac{\sum_i \log\frac{\tau_i}{\check{\sigma}_i}}{\log\frac{1}{\delta}} + 1\right)\right)}{2n}}$$

Simplify the right hand side and we complete the proof.

□

F Proof of Lemma 5

Proof. We first rewrite the inequality (18) below:

$$\begin{aligned}
\mathbb{E}_u[L(w^*+\tilde{u})] \leq \hat{L}(w^*) &+ \frac{1}{2}\sum_i \nabla^2_{i,i} L(w^*)\sigma_i^2 + \frac{\rho m^{1/2}}{6}\sum_i \kappa_i(w^*)\sigma_i^2 \\
&+ \frac{m\log\tau - \sum_i \log\sigma_i^2 - m + \frac{1}{\tau}\sum_i \sigma_i^2 + 1 + 2\log\frac{1}{\delta}}{2\eta} + \frac{\eta}{2n}
\end{aligned}$$

The terms related to σ_i on the right hand side of (18) is

$$\left(\frac{1}{2}\nabla^2_{i,i}L(w^*) + \frac{\rho m^{1/2}}{6}\kappa_i(w^*) + \frac{1}{2\tau\eta}\right)\sigma_i^2 - \frac{\log\sigma_i^2}{2\eta} \tag{25}$$

Take gradients w.r.t. σ_i, when $\nabla_i^2\hat{L} \geq 0$, we get the optimal σ_i^*,

$$\sigma_i^* = \min\left(\sqrt{\frac{1}{\eta\nabla^2_{i,i}\hat{L}(w^*) + \frac{\rho\eta m^{1/2}}{3}\kappa_i(w^*) + \frac{1}{\tau}}}, \frac{\kappa_i(w^*)}{\sqrt{2}\text{erf}^{-1}(\frac{1}{2m})}\right)$$

Note the first term in (25) is monotonously increasing w.r.t. σ_i, so

$$\begin{aligned}
&\left(\frac{1}{2}\nabla^2_{i,i}L(w^*) + \frac{\rho m^{1/2}}{6}\kappa_i(w^*) + \frac{1}{2\tau\eta}\right)\sigma_i^{*2} \\
&\leq \left(\frac{1}{2}\nabla_i^2 L(w^*) + \frac{\rho m^{1/2}}{6}\kappa_i(w^*) + \frac{1}{2\tau\eta}\right)\frac{1}{\eta\nabla_i^2\hat{L}(w^*) + \frac{\rho\eta m^{1/2}}{3}\kappa_i(w^*) + \frac{1}{\tau}} \\
&= \frac{1}{2\eta}
\end{aligned} \tag{26}$$

Summing over m parameters and combine (18), we complete the proof.

□

G About the Re-parameterization Invariance

If the loss is cross entropy and RELU-MLP is used, ideally we expect a bound invariant to re-parameterization. Unfortunately without further assumptions our proposed bound in this case is not invariant. On the other hand, the bound can be made invariant to re-parameterization with additional assumptions. Basically we need a layer-wise Hessian Lipschitz constant ρ^l as well as a layer-wise weight bound τ_i^l for different layers. In that case if the weight gets scaled w'=cw, then by the chain rule

$$\nabla^2 L_{ii}(w') = \nabla L_{ii}(w')\nabla L_{ii}^T(w') = \nabla L_{ii}(w)\frac{\partial w}{w'}\nabla L_{ii}^T(w)\frac{\partial w}{w'} = \nabla^2 L_{ii}(w)/c^2. \tag{27}$$

Similarly the layer-wise Lipschitz constant ρ^l is scaled by $1/c^3$, since

$$\|\nabla^2 f(w_1') - \nabla^2 f(w_2')\| = \|\nabla^2 f(w_1) - \nabla^2 f(w_2)\|/c^2 \leq \rho\|w_1 - w_2\|/c^2 = \frac{\rho}{c^3}\|w_1' - w_2'\|, \tag{28}$$

κ is scaled by c (ignoring the ϵ term). As a consequence σ in (4) is scaled by c.

Note after the scaling it is natural to assume the corresponding weight bound τ be scaled by c. In this way the ratio τ/σ keeps invariant during re-parameterization if we ignore the epsilon term in κ. Similarly we can check the bound does not change for the layer of weights scaled by $1/c$.

H A Lemma about Eigenvalues of Hessian and Generalization

By extrema of the Rayleigh quotient, the quadratic term on the right hand side of inequality (5) is further bounded by

$$u^T\nabla^2\hat{L}(w)u \leq \lambda_{max}(\nabla^2\hat{L}(w))\|u\|^2. \tag{29}$$

This is consistent with the empirical observations of Keskar et al. (2017) that the generalization ability of the model is related to the eigenvalues of $\nabla^2\hat{L}(w)$. The inequality (29) still holds even if the perturbations u_i and u_j are correlated. We add another lemma about correlated perturbations below.

Lemma 6. *Suppose the loss function $l(f, x, y) \in [0, 1]$. Let π be any distribution on the parameters that is independent from the data. Given $\delta > 0$ $\eta > 0$, with probability at least $1 - \delta$ over the draw of n samples, for any local optimal w^* such that $\nabla \hat{L}(w^*) = 0$, $\hat{L}(w)$ satisfies the local ρ-Hessian Lipschitz condition in $Neigh_\kappa(w^*)$, and any random perturbation u, s.t., $|u_i| \leq \kappa_i(w^*)$ $\forall i$, we have*

$$\mathbb{E}_u[L(w^* + u)] \leq \hat{L}(w^*) + \frac{1}{2}\lambda_{max}\left(\nabla^2 \hat{L}(w^*)\right)\sum_i \mathbb{E}[u_i^2] + \frac{\rho}{6}\mathbb{E}[\|u\|^3] + \frac{KL(w^* + u||\pi) + \log\frac{1}{\delta}}{\eta} + \frac{\eta}{2n}. \tag{30}$$

Proof. The proof of the Lemma 6 is straightforward. Since $\nabla \hat{L}(w^*) = 0$, the first order term is zero at the local optimal point even if $\mathbb{E}[u] \neq 0$. By extrema of the Rayleigh quotient, the quadratic term on the right hand side of inequality (5) is further bounded by

$$u^T \nabla^2 \hat{L}(w) u \leq \lambda_{max}\left(\nabla^2 \hat{L}(w)\right)\|u\|^2. \tag{31}$$

Due to the linearity of the expected value,

$$\mathbb{E}[u^T \nabla^2 \hat{L}(w) u] \leq \lambda_{max}\left(\nabla^2 \hat{L}(w)\right)\sum_i \mathbb{E}[u_i^2], \tag{32}$$

which does not assume independence among the perturbations u_i and u_j for $i \neq j$.

□