Loss-Landscape Geometry and Optimization

1 Key Concepts and Measurable Geometric Quantities

The empirical risk is defined as:

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell\big(f_\theta(x_i), y_i\big) \tag{1}$$

Gradient and Hessian:

$$g(\theta) = \nabla_\theta L(\theta) \tag{2}$$
$$H(\theta) = \nabla_\theta^2 L(\theta) \tag{3}$$

Common geometric summaries:

- Maximum eigenvalues of H: $\lambda_{\max}, \lambda_2, \ldots$ (sharpness indicator)
- $\mathrm{Tr}(H)$ approximations (aggregate curvature)
- Hessian-vector product (Hv) probes using Lanczos
- Directional scans $L(\theta + \alpha v)$ and plane scans
- Mode connectivity barriers along minimum-energy paths
- Sharpness approximations:
$$s_\epsilon(\theta) = \max_{\|\delta\| \le \epsilon} L(\theta + \delta) - L(\theta) \tag{4}$$
- Log-determinant flatness: $\log\det(H + \epsilon I)$
- Mini-batch gradient covariance:
$$\Sigma(\theta) = \mathbb{E}\big[(\nabla \ell_i)(\nabla \ell_i)^\top\big] - gg^\top \tag{5}$$
- Intrinsic dimension via random subspace training

2 Theoretical Research Directions

2.1 SGD as Stochastic Differential Equation

SGD with batch noise, step size η, and batch size B can be modeled as:

$$d\theta_t = -\nabla L(\theta_t)\,dt + \sqrt{\frac{\eta}{B}}\,\Sigma(\theta_t)^{1/2} dW_t \tag{6}$$

Objective: characterize the stationary measure and show concentration around flat basins. Target theorem sketch:

$$\Pr(\text{basin } i) \propto \frac{\exp\left(-L(\theta_i)/T_{\text{eff}}\right)}{\sqrt{\det(H_i)}} \tag{7}$$

2.2 Architecture's Effect on Hessian Spectrum

Hypothesis: skip connections and normalization layers compress the Hessian spectrum and reduce extreme eigenvalues. Possible analysis via NTK linearization and random matrix theory.

2.3 Geometry-Aware Generalization Bounds

Use PAC-Bayes bounds where posterior covariance is informed by inverse Fisher/Hessian:

$$q(\theta) = \mathcal{N}\big(\theta^\star,\ c(H + \epsilon I)^{-1}\big) \tag{8}$$

Goal: prove bounds based on effective rank/trace instead of raw parameter count.

2.4 Landscape Difficulty Index

Define difficulty index:

$$D(\theta) = \frac{\lambda_{\max}(H(\theta))^2}{\mathrm{Tr}(H(\theta))} \cdot \kappa(\Sigma, H) \tag{9}$$

where κ measures gradient-noise alignment with sharp directions.

3 Efficient Landscape Probing Methods

3.1 Hessian-Vector Product (PyTorch)

```
grads = torch.autograd.grad(loss, model.parameters(), create_graph=True)
grads_flat = torch.cat([g.contiguous().view(-1) for g in grads])

def Hv(v):
    grad_v = torch.dot(grads_flat, v)
    Hv_grads = torch.autograd.grad(grad_v, model.parameters(),
        retain_graph=True)
    Hv_flat = torch.cat([h.contiguous().view(-1) for h in Hv_grads]).
        detach()
    return Hv_flat
```

Plug this Hv function into a Lanczos or LOBPCG eigen-estimation routine.

3.2 Spectral Density Estimation

Apply stochastic Lanczos quadrature (SLQ) using random Gaussian probes to estimate spectral histograms and density.

3.3 Loss Scans

Line scan interpolation between θ_0 and θ_1:

```
alphas = np.linspace(-0.5, 1.5, 60)
losses = []
for a in alphas:
    set_model_flat(theta0 + a*(theta1-theta0))
    losses.append(eval_loss(dataloader_eval))
plot(alphas, losses)
```

3.4 Mode Connectivity

Optimize nonlinear paths (e.g., Bezier control points $p_1 ... p_{k-1}$):

$$\arg \min_{p_1 \dots p_{k-1}} \max_t L(\text{Bezier}(t; p_0 = \theta_a, p_k = \theta_b)) \tag{10}$$

Solve with Adam on path control points, track barrier height.

4 Engineering for Scalability

- Mixed precision + gradient checkpointing
- Hv estimation on data subsets (e.g., 1k samples)
- Parallelized probes over GPUs
- Logging via TensorBoard or Weights & Biases

5 Experimental Plan

5.1 Datasets

- Controlled: synthetic 2D loss surface, MNIST
- Mid-scale: CIFAR-10/100, Tiny-ImageNet subset
- Domain split: medical or shifted subsets if possible

5.2 Architectures

- MLP (depth/width sweep)
- Small CNN / VGG-like
- ResNet18/34 variants
- ViT-Small transformer models
- With/without BatchNorm, LayerNorm, skip connections

5.3 Sweeps

Batch size B, learning rate η, momentum, weight decay, augmentations, initialization scale.

5.4 Checkpoint Measurements

Top Hessian eigenvalues, trace, gradient norm, covariance eigenvectors, intrinsic dimension, mode connectivity barriers, validation generalization metrics (5 seeds per setting).

6 Concrete Hypotheses

- **H1**: Smaller B (more noise) $\rightarrow$ lower $\lambda_{\max}$ at minima $\rightarrow$ better test error for same train loss.
- **H2**: Skip connections flatten the landscape $\rightarrow$ smaller spectral outliers and lower barriers.
- **H3**: Alignment between top eigendirections of Hessian and gradient covariance predicts faster escape from sharp minima.
- **H4**: PAC-Bayes bounds with Hessian-informed covariance beat naive parameter-count bounds.

Evaluation via paired statistical tests, regression from curvature summaries $\rightarrow$ test error (R^2, p-values).

7 Risks and Limitations

- Hessian is local; combine with path probes
- Spectral estimates are noisy for large models
- Correlation causation $\rightarrow$ controlled ablations needed

8 Reproducibility

- Fix seeds but also measure variability
- Release runnable notebooks + full probe utilities
- Log full hyperparameters and training/eigen curves

9 Suggested Theoretical Keywords

- SGD-as-SDE, Freidlin–Wentzell large deviations
- PAC-Bayes geometry aware bounds
- Random matrix theory and Hessian spectra
- Mode connectivity and minimum energy paths
- Intrinsic dimension via random subspace SGD