Rethinking Parameter Counting in Deep Models: Effective Dimensionality Revisited

Nikita Okhotnikov

MIPT

2024

Introduction

- Number of parameters – poor generalisation metric
- Effective dimensionality might be a better one
- There's a connection to variance of posterior distribution

Theory behind

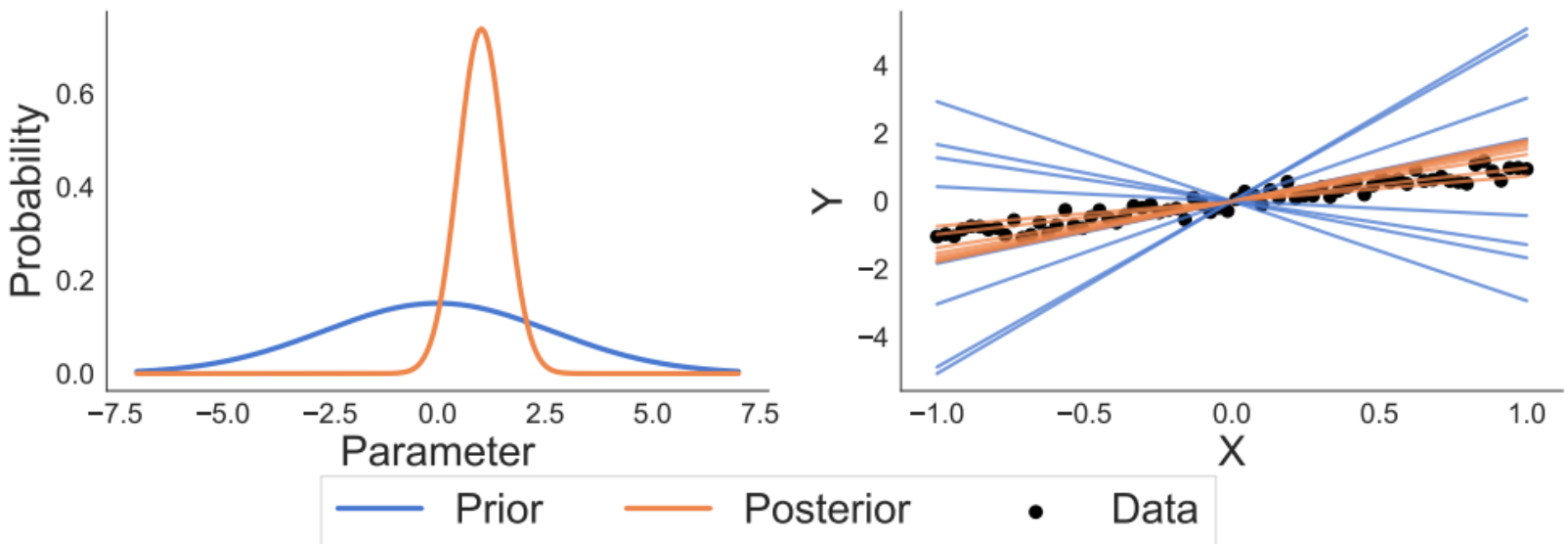
Posterior Contraction:

$$\Delta_{post}(\theta) = tr(Cov_{p(\theta)}(\theta)) - tr(Cov_{p(\theta|\mathcal{D})}(\theta))$$

For Bayesian linear regression:

$$\Delta_{post}(\theta) = \alpha^2 \sum_{i=1}^{N} \frac{\lambda_i}{\lambda_i + \alpha^{-2}}$$

Params vs functions distribution



Theory behind

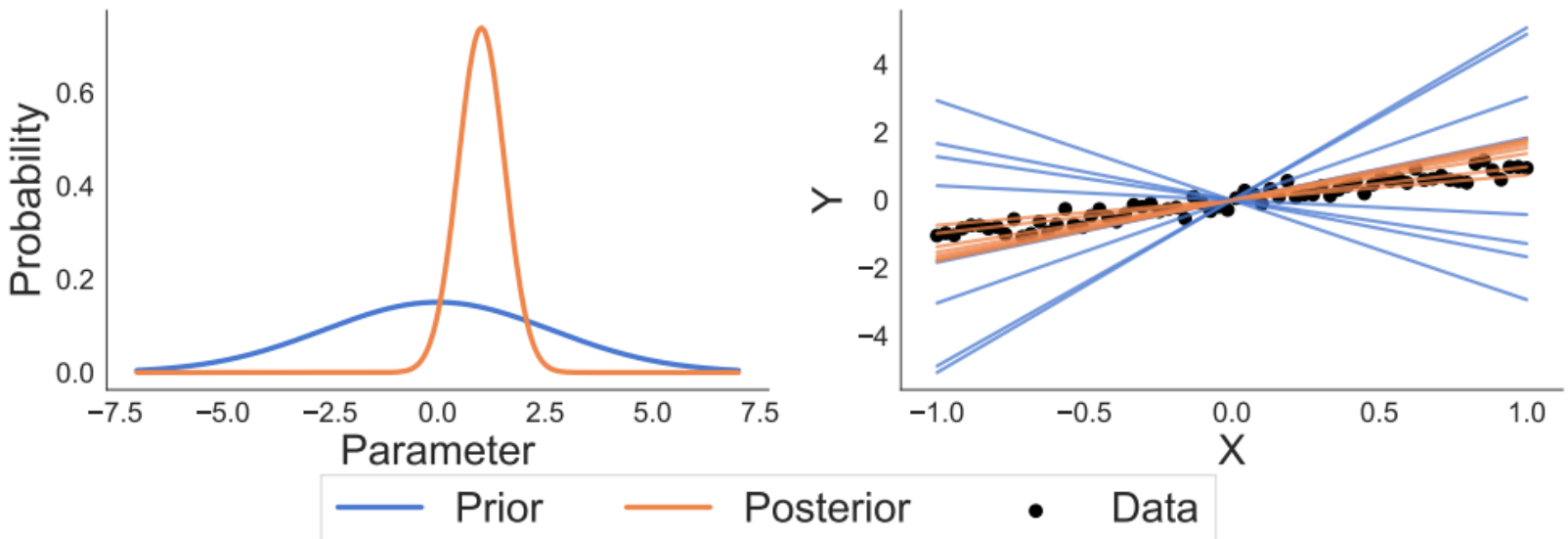
Effective Dimensionality of symmentric matrix $A \in \mathbb{R}^{k \times k}$:

$$N_{eff}(A, z) = \sum_{i=1}^{k} \frac{\lambda_i}{\lambda_i + z}, \text{ where } \lambda_i - \text{eigenvalues of A and } z > 0$$

Laplace approximation

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}}\, p(\theta|\mathcal{D}), \quad \theta \sim \mathcal{N}(\theta_{MAP}, (\mathcal{H}_\theta + A)^{-1}),$$

where $A = -\nabla\nabla_\theta \log p(\theta), \quad \mathcal{H}_\theta = -\nabla\nabla_\theta \mathcal{L}(\theta, \mathcal{D}) = \nabla\nabla_\theta \log p(\theta|\mathcal{D})$



Posterior contraction and effective dimensionality

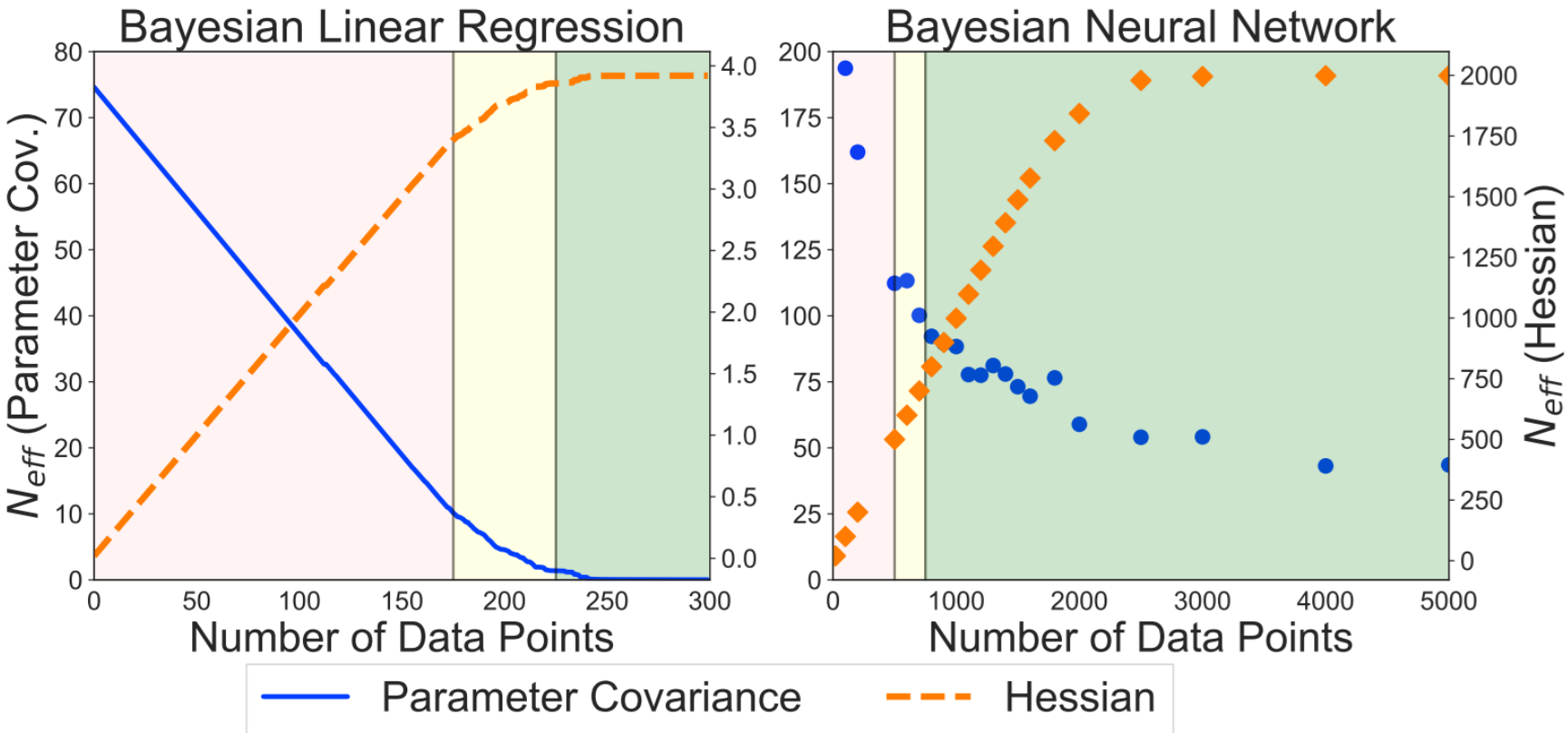
Theorem

$\Phi = \Phi(x) \in \mathbb{R}^{n \times k}$ – feature map of n observations, $n < k$.
$\beta \sim \mathcal{N}(0_k, \alpha^2, I_k)$ – parameters prior.
Model – $y \sim \mathcal{N}(\Phi\beta, \sigma^2 I_n)$.
Then the posterior distribution of β has a $k - n$ directional subspace in which the variance is identical to the prior variance.

Posterior contraction and effective dimensionality

$$\Phi = \Phi(x) \in \mathbb{R}^{n \times k} = [\cos(\pi x), \sin(\pi x), \cos(2\pi x), \sin(2\pi x) \dots], \quad \beta \sim \mathcal{N}(0, I)$$

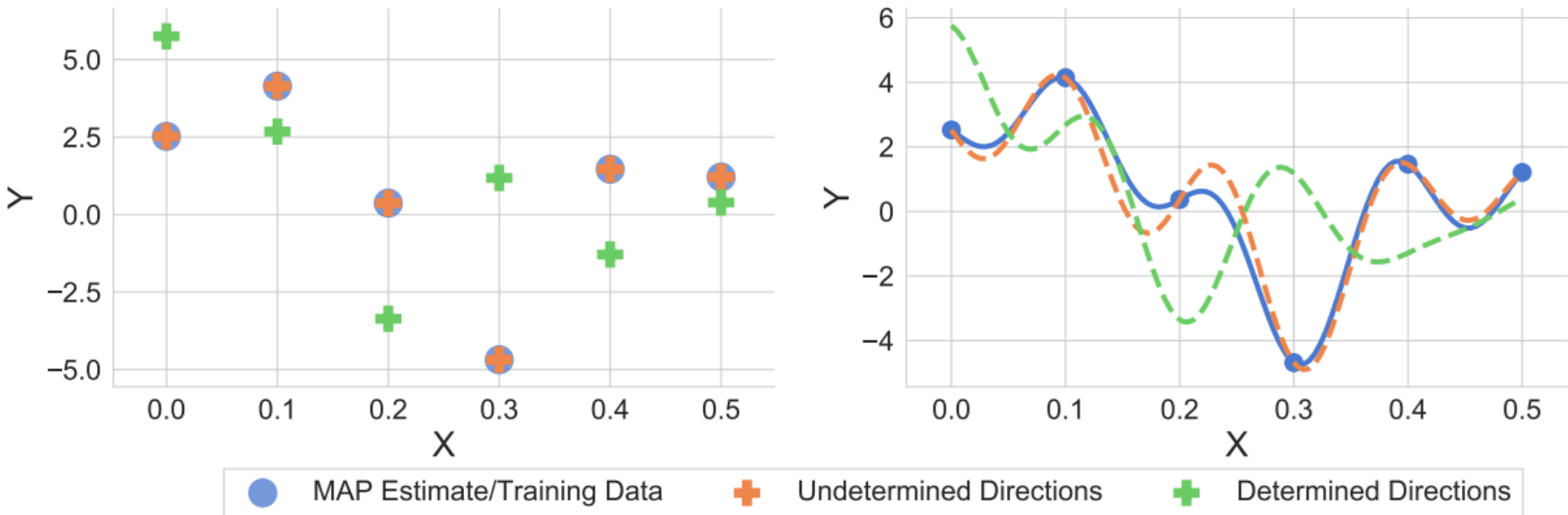
ground truth parameters β^* from the same distribution. $y \sim \mathcal{N}(\Phi\beta, \sigma^2 I)$



Function-space homogenenity in linear models

Theorem

$\Phi = \Phi(x) \in \mathbb{R}^{n \times k}$ – feature map of n observations, $n < k$.
$\beta \sim \mathcal{N}(0_k, \alpha^2, I_k)$ – parameters prior.
The minimal eigenvectors of the Hessian define a $k - n$ dimensional subspace in which parameters can be perturbed without changing the training predictions in function space. (Proved for generalized linear models)



Loss Surfaces in Orthogonalized basis

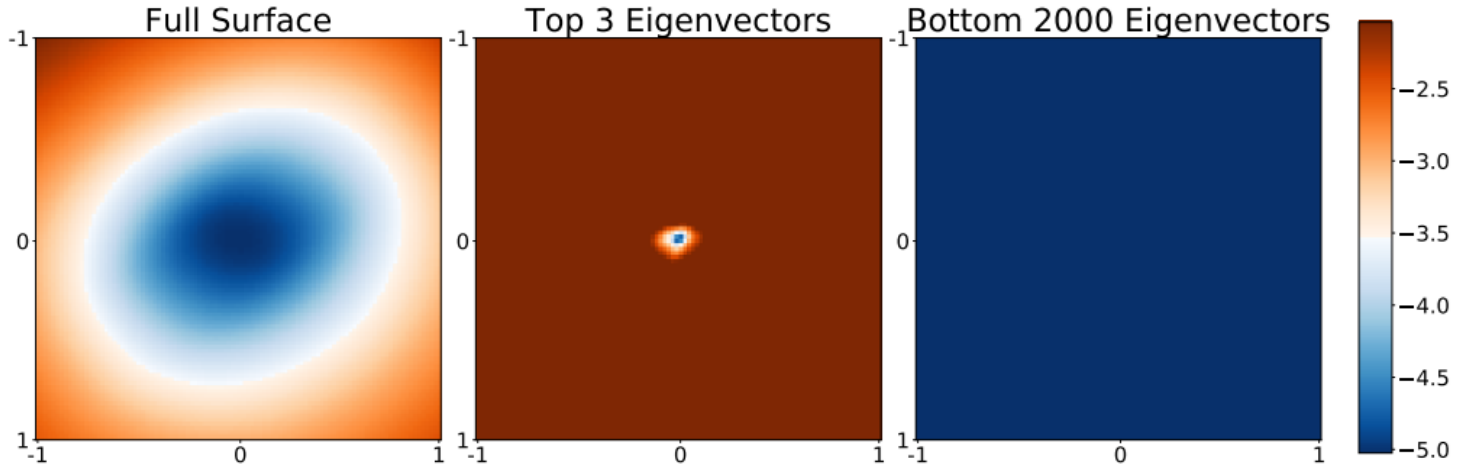


Figure 6. **Left:** A random projection of the loss surface. **Center:** A projection of the loss surface in the top 3 directions in which parameters have been determined. **Right:** A projection of the loss surface in the 2000 (out of 2181) directions in which parameters have been determined the least. The rightmost plot shows that in degenerate parameter directions the loss is constant.

Degenerate Parameters Lead to Homogeneous Models

Perturbing some parameters in over-parameterized model leads to model equivalent in function space

$$\theta \longleftarrow \theta^* + s\frac{Bv}{\|Bv\|_2}, \quad v \sim \mathcal{N}(0, I_d), \; B - \text{d-dimensional basis}, \quad s \leqslant \|\theta^*\|_2/2$$

Thus, we might probably ignore subspaces of degenerate parameters for model compression.

Effective dimensionality and compression

$$p(\mathcal{D}|\mathcal{M}_i) \approx \underbrace{p(\mathcal{D}|\theta_{MP}, \mathcal{M})}_{\text{Likelihood}} \times \underbrace{p(\theta_{MP}|\mathcal{M}) \det{}^{-\frac{1}{2}}(\mathcal{H}_\theta/2\pi)}_{\text{Occam Factor}},$$

$$\text{Evidence} \approx \quad \text{Likelihood} \quad \times \quad \text{Occam Factor}$$

Lower eigenvalues $\rightarrow$ lower effective dimensionality and higher Occam Factor $\rightarrow$ lower description length and better compression

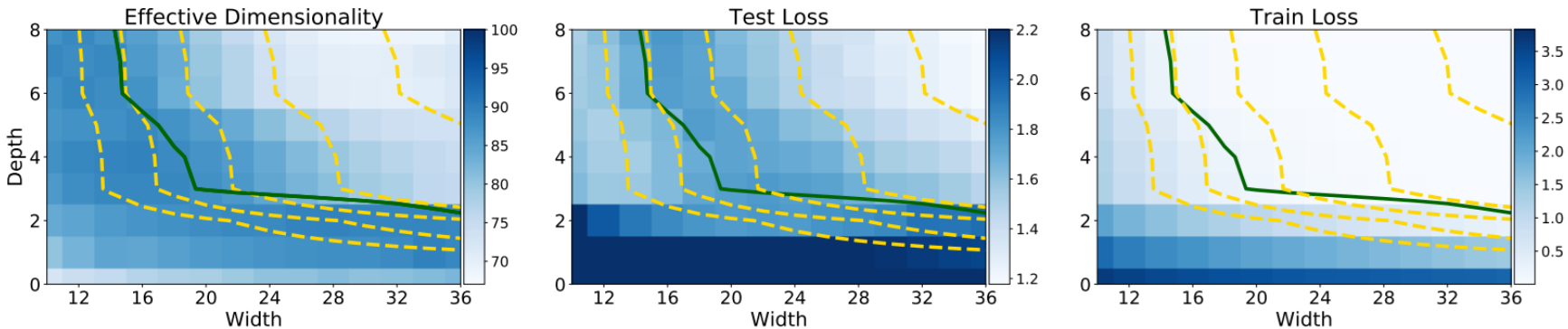


Figure 2. **Left:** Effective dimensionality as a function of model width and depth for a CNN on CIFAR-100. **Center:** Test loss as a function of model width and depth. **Right:** Train loss as a function of model width and depth. Yellow level curves represent equal parameter counts ($1e5$, $2e5$, $4e5$, $1.6e6$). The green curve separates models with near-zero training loss. Effective dimensionality serves as a good proxy for generalization for models with low train loss. We see wide but shallow models overfit, providing low train loss, but high test loss and high effective dimensionality. For models with the same train loss, lower effective dimensionality can be viewed as a better compression of the data at the same fidelity. Thus depth provides a mechanism for compression, which leads to better generalization.

Effective dimensionality and generalization

	Test loss	Test Error	Gen. Gap
N_{eff}(Hessian)	**0.9434**	0.9188	**0.9429**
PAC-Bayes	−0.8443	−0.7372	−0.8597
Mag. PAC-Bayes	0.7066	0.8270	0.6805
Path-Norm	0.5598	0.7216	0.5259
Log Path-Norm	0.9397	**0.9846**	0.9257

Table 1. Sample Pearson correlation with generalization on double descent for ResNet18s of varying width on CIFAR-100 with a training loss below 0.1.

	Test loss	Test Error	Gen. Gap
N_{eff}(Hessian)	0.9305	**0.9461**	0.9060
PAC-Bayes	−0.8619	−0.7916	−0.8873
Mag. PAC-Bayes	0.8724	0.9225	0.8330
Path-Norm	0.7996	0.7721	0.7511
Log Path-Norm	**0.9781**	0.9402	**0.9602**

Table 2. Sample Pearson correlation with generalization for CNNs of varying width and depth on CIFAR-100 with a training loss below 0.1.