Simplicity Suffices for Parameter Noise Injection in Stochastic Gradient Descent

Benjamin Leblanc, Louis-Jacob Lebel
Computer Science & Electronic Department, Université Laval, Québec, Canada

August 16th, 2026

The machine learning optimization problem

Modern machine learning $\Rightarrow$ Non-convex optimization in millions or billions of dimensions

Model parameters

$$\boldsymbol{\theta} \in \mathbb{R}^{|\boldsymbol{\theta}|}$$

Training set

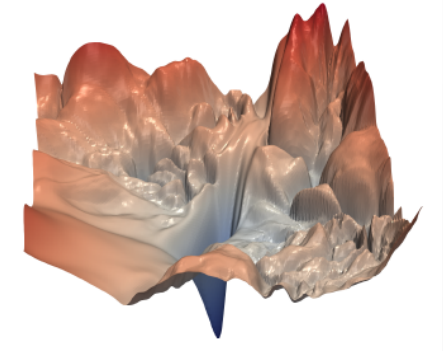
$$S = \{x_i, y_i\}_{i=1}^{m}$$

Training loss

$$\mathcal{L}_S(\boldsymbol{\theta})$$

Learning objective

$$\mathcal{L}_S(\boldsymbol{\theta})$$



The machine learning optimization problem

Modern machine learning $\Rightarrow$ Non-convex optimization in millions or billions of dimensions

Model parameters

$$\boldsymbol{\theta} \in \mathbb{R}^{|\boldsymbol{\theta}|}$$

Learning objective

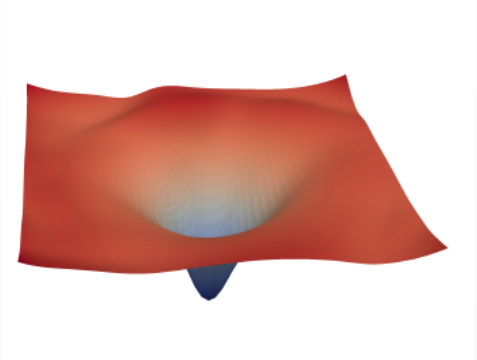
?

Training set

$$S = \{x_i, y_i\}_{i=1}^{m}$$

Training loss

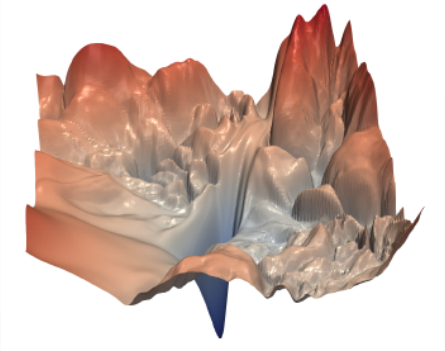
$$\mathcal{L}_S(\boldsymbol{\theta})$$



In search of an adequate learning objective

How can we find a better learning objective?

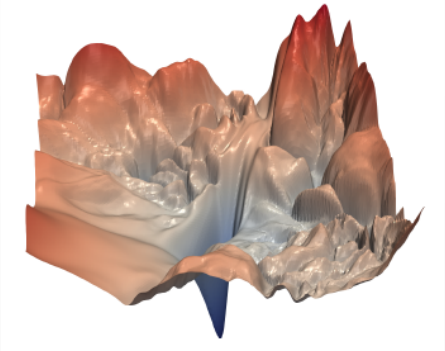
- Keep the idea of optimizing the training loss;
- We are interested in the **neighbouring** training loss;
- Idea: weighted average of the training loss around a point of interest!



In search of an adequate learning objective

Learning objective

$$\mathcal{L}_S(\theta)$$

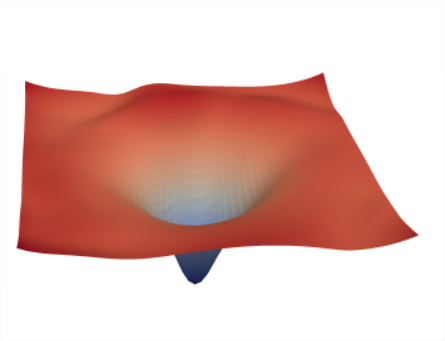


In search of an adequate learning objective

Learning objective

$\mathbb{E}_{z \sim \mathcal{Z}} \mathcal{L}_S(\boldsymbol{\theta} - z)$, for some distribution $\mathcal{Z}$.

$\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_S(\boldsymbol{\theta} - z_i)$, where $z_i \sim \mathcal{Z}$.



Notice that:

$$\mathbb{E}_{z \sim \mathcal{Z}} \mathcal{L}_S(\boldsymbol{\theta} - z) = (\mathcal{L}_S \star p_{\mathcal{Z}})(\boldsymbol{\theta}),$$

where $p_{\mathcal{Z}}$ is the probability density function of $\mathcal{Z}$, and $\star$ is the convolution operator.

Noisy stochastic gradient descent

The "noisy" version of the Stochastic Gradient Descent (SGD) algorithm is well-known and commonly used:

- PAC-Bayes [McAllester, 1998];
- Stochastic Gradient Langevin Dynamics [Welling and Teh, 2011];
- Gradient smoothing [Starnes et al., 2024];
- Perturbed Gradient Descent [Cui et al., 2025].

Learning objective

$\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_S(\boldsymbol{\theta} - \mathrm{z}_i)$, where $\mathrm{z}_i \sim \mathcal{Z}$.

Our question marks

The general learning objective being found, an intuitive choice for the perturbations distribution $\mathcal{Z}$ is a **centered Gaussian distribution** $\mathcal{N}(0, \mathbf{\Omega})$.

Still, some questions remain:

- What is the best covariance matrix $\mathbf{\Omega}$?
- What is the best number of noise samples per gradient step n?

Learning objective

$\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_S(\boldsymbol{\theta} - \mathrm{z}_i)$, where $\mathrm{z}_i \sim \mathcal{Z}$.

What is the best covariance matrix $\mathbf{\Omega}$?

Inspired by the noisy gradient descent literature, we tested various changing covariance matrices $\mathbf{\Omega}_t$, where t corresponds to the current optimization step:

- The *moving* method;
- The *squared gradient* method;
- The *inverse squared gradient* method.

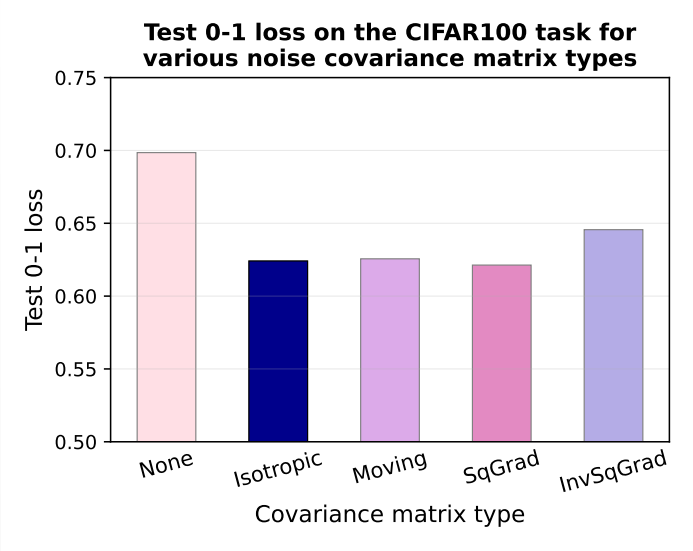
We compared these sophisticated methods to some vanilla approach:

- An isotropic covariance matrix $\mathbf{\Omega} = \sigma \cdot \mathrm{I}$.

Noise distribution

$\mathrm{z}_i \sim \mathcal{Z}$, where $\mathcal{Z} = \mathcal{N}(0, \mathbf{\Omega})$.

What is the best covariance matrix $\mathbf{\Omega}$?



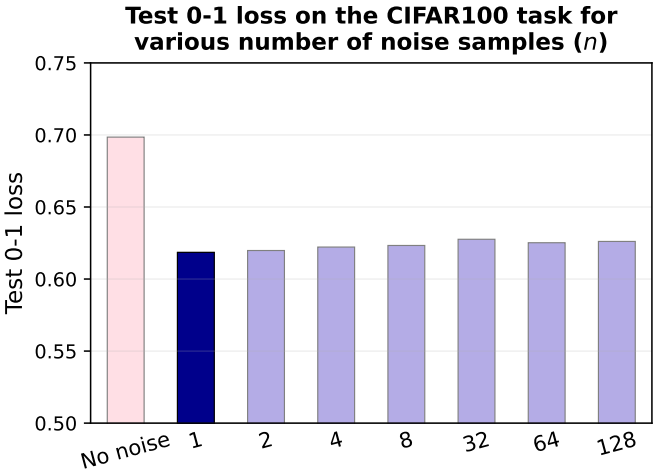
What is the best number of noise samples?

Recall that for every gradient step, we need n forward passes, made with n different noise samples.

Learning objective

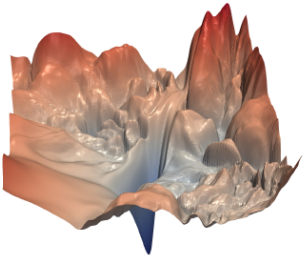
$\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_S(\boldsymbol{\theta} - \mathrm{z}_i)$, where $\mathrm{z}_i \sim \mathcal{Z}$.

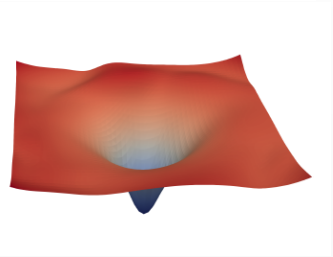
What is the best number of noise samples?



Key takeaway

- Adding noise to parameters during modern machine learning optimization is clearly useful...
- ... But over-engineering the process leads to no significant advantage.





Q&A

Thank you for attending!