The effect of gradient confusion on the convergence of SGD in neural nets and other over-parameterized problems

Karthik A. Sankararaman*, Soham De*, Zheng Xu, W. Ronny Huang, Tom Goldstein

University of Maryland, College Park

Main points

- Why does constant step-size SGD work so well on neural nets?
- "Gradient Confusion": measures how aligned the gradients are.
- Low gradient confusion $\rightarrow$ fast convergence of SGD.
- When is gradient confusion low? Overparameterized models.
- Increasing width of neural net decreases gradient confusion; Increasing depth increases gradient confusion.

Problem formulation

Consider a function f and n data points in d-dimension, $\{\mathbf{x}_i\}_{i=1}^n$. Let $f_i(\mathbf{w}) := f(\mathbf{w}, \mathbf{x}_i)$. For a given parameter $\mathbf{w}$, the *empirical risk* F is defined as,

$$F(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{w}).$$

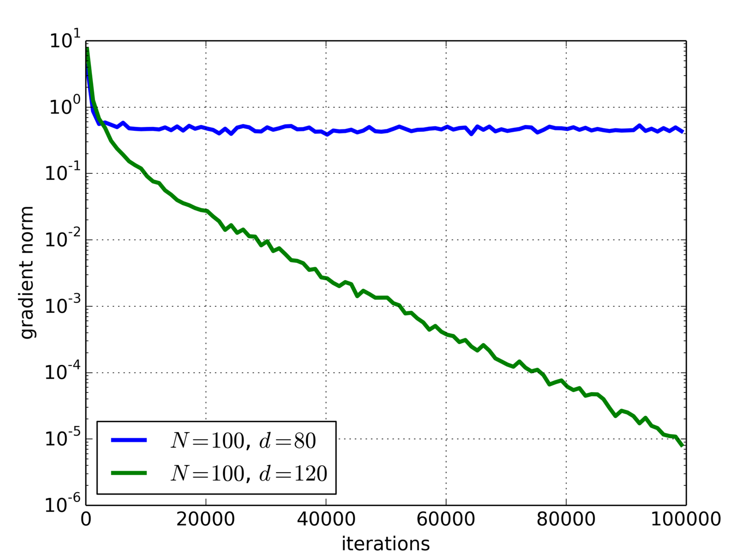
We want to minimize the empirical risk: $\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w})$.

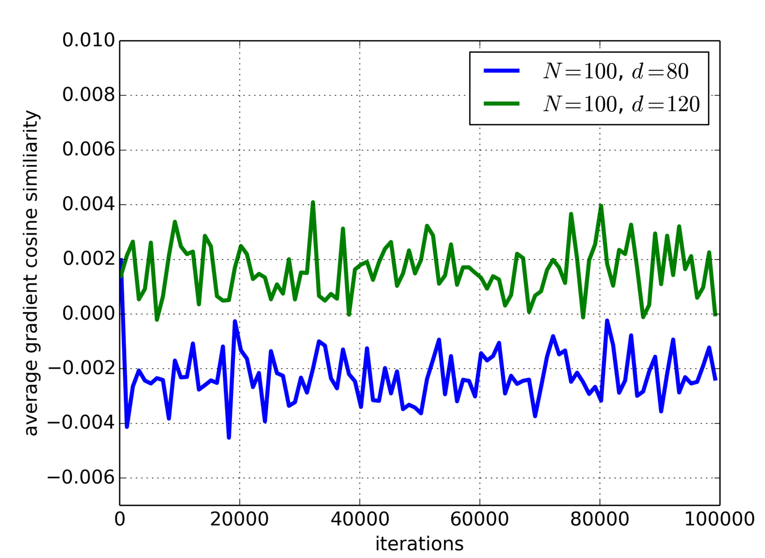
Stochastic Gradient Descent (SGD)

Input: $f_1, f_2, \ldots, f_n$, number of iterations T, learning rate α.
Output: $\mathbf{w}_T$.
$\mathbf{w}_0 \leftarrow \mathcal{N}(0, \frac{1}{d})$;
for $k = 1$ *to* T **do**
  $\tilde{f}_k \leftarrow$ Uniform random sample from $\{f_i\}_{i=1}^n$;
  $\mathbf{w}_{k+1} = \mathbf{w}_k - \alpha \nabla \tilde{f}_k(\mathbf{w}_k)$;
end

Simulations on Toy Problem

Linear regression with random Gaussian data (avg. over 3 runs):





Overparameterized model *converges linearly*, and has on average *positive cosine similarity* between the individual gradients.

Gradient confusion

The set $\{f_i\}_{i=1}^n$ has gradient confusion $\eta \geq 0$ at $\mathbf{w}$, if:

$$\forall i, j \in [n] \qquad \langle \nabla f_i(\mathbf{w}), \nabla f_j(\mathbf{w})\rangle \geq -\eta.$$

Informal: Low gradient confusion $\rightarrow$ pair-wise vectors ∇f_i and ∇f_j do not point in opposite directions.

Low gradient confusion $\rightarrow$ fast convergence

Consider logistic regression with orthogonal data.

$$\langle \nabla f_i(\mathbf{w}), \nabla f_j(\mathbf{w})\rangle = 0.$$

SGD decouples into GD on each term separately.

Assumptions: $\{f_i\}$ are L-Lipshitz smooth and μ-strongly convex.

Theorem. SGD converges *linearly* to a neighborhood of the minimizer with *constant step size* α as:

$$F(\mathbf{w}_k) - F^\star \leq \rho^k (F(\mathbf{w}_0) - F^\star) + \frac{\alpha\eta}{1-\rho},$$

where step size $\alpha \leq 2/nL$ and $\rho = 1 - \frac{2\mu}{n}\left(\alpha - \frac{nL\alpha^2}{2}\right)$.

If the objective function satisfies the *strengthened* bound:

$$\langle \nabla f_i(\mathbf{w}), \nabla f_j(\mathbf{w}')\rangle \geq -\eta, \;\; \forall i, j, \mathbf{w}, \mathbf{w}',$$

SGD converges to the noise floor at a *faster rate*:

$$F(\mathbf{w}_k) - F^\star \leq \rho^k (F(\mathbf{w}_0) - F^\star) + \frac{\alpha\eta}{1-\rho},$$

where the step-size $\alpha \leq 2/L$ and $\rho = 1 - 2\mu\alpha/n + \mu L \alpha^2/n$.

When is gradient confusion low?

Informal: Random vectors in high dimensions are nearly orthogonal. So, over-parameterized linear models are expected to have low gradient confusion.

Formal results proved for a *random data model*:

- Random data $\mathcal{D} = \{(\mathbf{x}_i, \mathcal{C}(\mathbf{x}_i))\}_{i=1}^n$, for some labeling function $\mathcal{C}$.
- $\{\mathbf{x}_i\}$ are drawn iid from surface of a d-dimensional unit sphere.
- $f_i(\mathbf{w}) = \frac{1}{2}(g_{\mathbf{w}}(\mathbf{x}_i) - \mathcal{C}(\mathbf{x}_i))^2$, where $g_{\mathbf{w}}(\mathbf{x}_i)$ is over-parameterized.
- $\mathcal{C}$ needs to satisfy *mild* conditions, such as boundedness and bounded first derivative.

Linear regression bounds

- $g_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x}\rangle$.
- Consider a given dimension $d \geq \Omega(\log n)$. Let $\{\mathbf{W}^{(k)}\}_{k=1}^T \in [-1,1]^d$ be the set of realized weights vectors in a run of SGD. With probability at least $1-\delta$, gradient confusion of η holds for a constant $\eta < 1$.

Informal: Number of parameters being large enough $\rightarrow$ low gradient confusion with high-probability.

Neural net bounds

- Depth β and width ℓ neural nets. $\mathbf{W}_0 \in [-1/\ell, 1/\ell]^{d\times\ell}$, $\mathbf{W}_1, \ldots, \mathbf{W}_\beta \in [-1/\ell, 1/\ell]^{\ell\times\ell}$ and $\mathbf{W}_{\beta+1} \in [-1/\ell, 1/\ell]^{\ell\times 1}$.
- $g(\mathbf{x}) := \sigma(\mathbf{W}_\beta \cdot \sigma(\mathbf{W}_{\beta-1} \ldots \sigma(\mathbf{W}_1 \cdot \sigma(\mathbf{W}_0 \mathbf{x}))))$.
- σ point-wise non-linearity. Bounded, bounded 1st and 2nd derivatives. Examples: *sigmoid*, *tanh* and *softmax* (not *relu*).

Consider a given dimension d. Let $\{(\mathbf{W}_i^{(k)})_{i=0}^{\beta+1}\}_{k=1}^T$ be the set of realized weights vectors in a run of SGD. With probability at least $1-\delta$ over the randomness in the data, gradient confusion of η holds for all weights and for a constant $\eta < 5$ as long as $\frac{\ell}{\beta^2} \geq \tilde{\Omega}\left(\frac{1}{\sqrt{d}}\right)$.

Informal: Increasing width lowers gradient confusion; Increasing depth increases gradient confusion.

Experiments

SGD on Wide ResNets for image classification on CIFAR-10 with diff. width and depth. Tuned step-sizes. [width here denotes width factor]