Continuous Time Stochastic Gradient Descent and Flat Minimum Selection

Stephan Wojtowytsch

Texas A&M University

October 28, 2021, Simons Institute

Dynamics and Discretization: PDEs, Sampling, and Optimization

Outline

Introduction

Classical interpolation.

- Data set $\{(x_i, y_i) : 0 \leq i \leq n\}$ with $x_i, y_i \in \mathbb{R}$.
- Optimal approximation/Runge phenomenon.
    1. Polynomial of degree $n+1$
    2. Piecewise polynomial splines
    3. Least squares approximation
    4. ...

Modern interpolation.

- Data set $\{(x_i, y_i) : 0 \leq i \leq n\}$ with $x_i \in \mathbb{R}^d$ for $d \gg 1$.
- Neural networks
    1. Overparametrized
    2. Non-linear in both parameters and data
    3. Statistical learning guarantees (?)

$h(\theta, x)$ parametrized function: Minimize

$$L_y(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \left| h(\theta, x_i) - y_i \right|^2 .$$

Theorem (Cooper '18)

1. $\{(x_i, y_i) : 1 \leq i \leq n\}$ *a data set*
2. $h : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}$ *a parametrized function,* $h(\theta, x)$ *is* C^{m-n}*-smooth in* θ*.*
3. *For any* $y_1', \ldots, y_n'$ *there exists* $\theta \in \mathbb{R}^m$ *such that* $h(\theta, x_i) = y_i'$ *for all* i*.*

Then for almost all $y' \in \mathbb{R}^n$*, the set of minimizers*

$$N_{y'} = L_{y'}^{-1}(0) = \{\theta \in \mathbb{R}^m : h(\theta, x_i) = y_i'\}$$

is an $m-n$*-dimensional submanifold of* $\mathbb{R}^m$*. If* h *is Lipschitz-continuous in* θ *and can fit random data at* $n+1$ *data puts, then* $N_{y'}$ *is non-compact.*

Proof.

Consider $\Phi : \mathbb{R}^m \to \mathbb{R}^n$, $\Phi(\theta) = \big(h(\theta, x_1), \ldots, h(\theta, x_n)\big)$ and apply regular value theorem + Sard's theorem.

For non-compactness: $|\theta - \tilde{\theta}| \geq \frac{1}{L}|h(\theta, x_{n+1}) - h(\tilde{\theta}, x_{n+1})| = \frac{1}{L}|y_{n+1} - \tilde{y}_{n+1}|$. $\square$

Implicit regularization

There are many minimizers of $L(\theta) = \frac{1}{2n}\sum_{i=1}^n \big|h(\theta, x_i) - y_i\big|^2$.

- Some memorize the data set.
- Some extract the underlying structure of the data set.

Question: Which one do we find when 'training' the function model?

Conjecture 1: We find minimizers where the energy landscape is 'flat' in some sense.

Conjecture 2: Flat minimizers generalize better.

(Hochreiter & Schmidhuber '97, ...)

Loss landscape and stochastic gradient descent

$$f(x) = \sum_{i=1}^{m} a_i\,\sigma(w_i \cdot x + b_i).$$

Theorem

If $m \geq n$, then f can fit any values $y_1, \dots, y_n$ at $x_1, \dots, x_m$.

Proof.

- Clear in one dimension.
- Choose $w_1 = \cdots = w_n$ such that $z_j = w \cdot x_j$ are all different.

$\square$

Remark: Any two-layer network can be approximated by certain deep neural networks with at most four times more parameters.

Corollary (W '21)

- *Under the same assumptions as above, L is convex if and only if $\theta \mapsto h(\theta, x)$ is linear.*
- *If h is non-linear enough, then: For every $\theta \in N_y$ and every $\varepsilon > 0$, there exists $\theta' \in B_\varepsilon(\theta)$ such that $D^2 L(\theta')$ has a negative eigenvalue.*

$$\sigma(z) = \max\{z, 0\}, \qquad f(x) = \sum_{i=1}^{m} a_i\, \sigma(w_i \cdot x + b_i).$$

1. Permutation of i.
2. $a_1\sigma(w_1 \cdot x + b_1) + a_2\sigma(w_2 \cdot x + b_2) = 0$ if $a_2 = -a_1$ and $(w_2, b_2) = (w_1, b_1)$.
3. $\sigma(z) = \frac{1}{\mu}\,\sigma(\mu z)$
4. $z = \sigma(z) - \sigma(-z)$

$$\begin{aligned} \Rightarrow z + 1 &= \sigma(z) - \sigma(-z) + \sigma(1) \\ &= \sigma(z+1) - \sigma(-(z+1)) \end{aligned}$$

5. If $f \in W^{2,1}(\mathbb{R})$ and $x > 0$, then

$$f(x) = f(0) + f'(0)\,\sigma(x) + \int_0^\infty f''(t)\,\sigma(t - x)\,\mathrm{d}t.$$

Represent $\|x\|_2^2$ along coordinate axes/rotated coordinate system/rotationally symmetrically…

Gradient descent

$$\theta_{t+1} = \theta_t - \eta_t \nabla L(\theta_t) = \theta_t - \frac{\eta_t}{n} \sum_{i=1}^{n} \big(h(\theta, x_i) - y_i\big) \nabla_\theta h(\theta, x_i)$$

Stochastic gradient descent

$$\theta_{t+1} = \theta_t - \frac{\eta_t}{b} \sum_{j=1}^{b} \big(h(\theta, x_{i_j}) - y_{i_j}\big) \nabla_\theta h(\theta, x_{i_j})$$

Stochastic gradient descent (general)

$$\theta_{t+1} = \theta_t - \eta g(\theta_t), \qquad \mathbb{E} g(\theta_t) = \nabla f(\theta_t)$$

and g satisfies some moment bounds.

$$L(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \left| h(\theta, x_i) - y_i \right|^2$$

$$D^2 L(\theta) = \sum_{i=1}^{n} \Big[\nabla_\theta h(\theta, x_i) \otimes \nabla_\theta h(\theta, x_i) + \big(h(\theta, x_i) - y_i \big) D_\theta^2 h(\theta, x_i) \Big]$$

$$\Sigma(\theta) = \sum_{i=1}^{n} \big(h(\theta, x_i) - y_i \big)^2 \big(\nabla_\theta h(\theta, x_i) - \nabla L(\theta) \big) \otimes \big(\nabla_\theta h(\theta, x_i) - \nabla L(\theta) \big)$$

- Gradient estimator noise intensity scales with loss
- Gradient estimator noise has low rank $n \ll m$

$\Sigma \approx L \cdot D^2 L$?

$$\begin{aligned}
\mathbb{E}_{(x_i,y_i)}\left[\left|\nabla\left(h(\theta,x_i)-y_i\right)^2-\nabla L(\theta)\right|^2\right] &\le \mathbb{E}_{(x_i,y_i)}\left[\left|\nabla\left(h(\theta,x_i)-y_i\right)^2\right|^2\right] \\
&\le \mathbb{E}\left[|h(\theta,x_i)-y_i|^2\,|\nabla_\theta h|^2(\theta,x_i)\right] \\
&\le \|\nabla_\theta h\|_{L^\infty}^2\,\mathbb{E}\left[|h(\theta,x_i)-y_i|^2\right] \\
&= \|\nabla_\theta h\|_{L^\infty}^2\,L(\theta),
\end{aligned}$$

so at least locally

$$\mathbb{E}\left[|g(\theta,\omega)-\nabla f(\theta)|^2\right] \le \sigma\, f(\theta).$$

Discrete time convergence of SGD

Lemma (W '21)

Let $f : \mathbb{R}^m \to [0, \infty)$ be an objective function such that

- *∇f is C_L-Lipschitz, and*
- *the energy/energy-dissipation inequality $\Lambda f(\theta) \leq |\nabla f|^2(\theta)$ holds.*

Let g be a family of gradient estimators such that

$$\mathbb{E} g(\theta) = \nabla f(\theta), \qquad \mathbb{E}\big[|(g - \nabla f)(\theta)|^2\big] \leq \sigma f(\theta).$$

Then if

$$\eta < \frac{\Lambda}{\Lambda + \sigma} \frac{2}{C_L} \quad \text{and} \quad \rho_\eta = 1 - \Lambda \eta + \frac{C_L(1+\sigma)}{2\Lambda} \eta^2,$$

the estimate

$$\mathbb{E}\big[f(\theta_t)\big] \leq \rho_\eta^t \, \mathbb{E}\big[f(\theta_0)\big]$$

holds for

$$\theta_t = \theta_{t-1} - \eta g(\theta_{t-1})$$

and there exists a random variable θ_∞ such that

$$\mathbb{E}\big[|\theta_t - \theta_\infty|^2\big] \leq C\, \rho_\eta^t.$$

Theorem (W '21)

Let $f : \mathbb{R}^m \to \mathbb{R}$ *be a function such that*

1. ∇f *is Lipschitz-continuous*
2. f *satisfies an energy/energy dissipation inequality on the set* $\{f < \varepsilon\}$.
3. f *satisfies a energy/energy dissipation inequality on the set* $\{f > S\}$.
4. *The set* $\{f \leq S\}$ *is contained in a bounded tube around the set* $\{f = 0\}$.

Consider gradient estimators g *such that*

- $\mathbb{E}\big[|g - \nabla f|^2(\theta)\big] \leq \sigma f(\theta)$ *and*
- $g(\theta) = \nabla f(\theta) + \sqrt{f(\theta)}\, Y(\theta, \omega)$ *where* Y *is 'uniformly spread out' (e.g. standard Gaussian).*

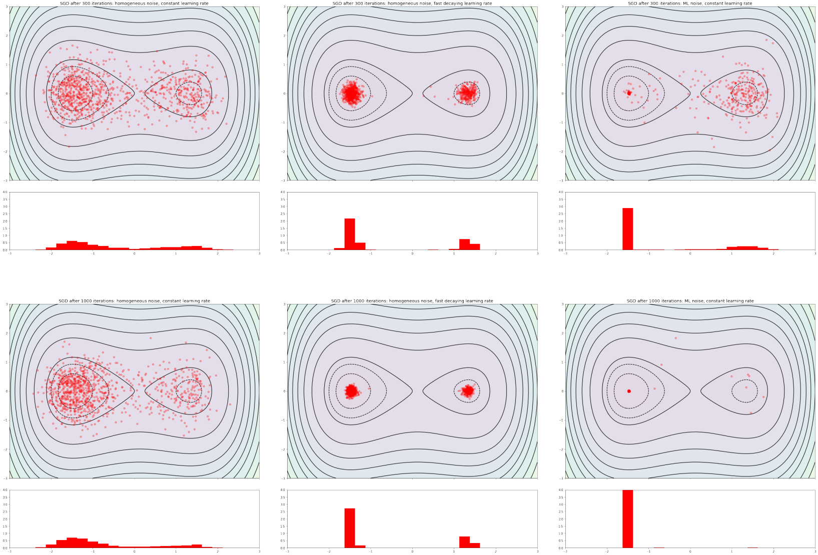
If η *is small enough and* $\rho_\eta < \beta \leq 1$, *then*

$$\limsup_{t\to\infty} \frac{\mathbb{E}\big[f(\theta_t)\big]}{\beta^t} = 0,$$

almost surely, and the iterates θ_t *converge exponentially fast to a limit* θ_∞ *in the set of minimizers.*

- We expect convergence for small, strictly positive step size.
  Different from classical SGD where the noise is bounded and $\eta_t \to 0$!
- We expect convergence to a global minimizer.
- The limiting point depends on the initial condition due to exponential convergence.

Numerical comparison



Implicit Bias of SGD: Continuous Time Analysis

$$\theta_{t+1} - \theta_t = -\eta_t \, g(\theta_t, \omega_t) \quad \to \quad \mathrm{d}\theta_t = -\nabla f(\theta_t)\,\mathrm{d}t + \sqrt{\eta_t \, \Sigma(\theta_t)}\,\mathrm{d}B_t \qquad (1)$$

Lemma

Assume that θ_t solves follows continuous time SGD (1)*. Then the law ρ_t of θ_t solves the PDE*

$$\partial_t \rho = \operatorname{div}\big(\rho \, \nabla f\big) + \partial_i \partial_j \big(\rho \Sigma_{ij}\big).$$

Consider $\Sigma = \sigma f \, I$, so

$$\begin{aligned} \partial_t \rho &= \operatorname{div}\big(\rho \, \nabla f\big) + \eta\sigma \, \partial_i \partial_j \big(\rho f \, \delta_{ij}\big) \\ &= \operatorname{div}\big(\eta\sigma \, f \nabla \rho + (1 + \eta\sigma)\rho \, \nabla f\big) \\ &= \eta\sigma \operatorname{div}\Big(f^{-\frac{1}{\eta\sigma}} \nabla \big(f^{1 + \frac{1}{\eta\sigma}} \rho\big)\Big) \end{aligned}$$

(isotropic, not homogeneous).

Trivially, we have

$$\eta\sigma\,\mathrm{div}\big(f^{-\frac{1}{\eta\sigma}}\nabla\big(f^{1+\frac{1}{\eta\sigma}}\rho\big)\big) = 0$$

if $\rho = c\,f^{-1-\frac{1}{\eta\sigma}}$ (also Liu-Ziyin -Ueda '20).

Lemma (W '21)

If $\inf f > 0$ *and* $\frac{|\nabla f|}{f}(\theta) \leq \frac{C}{1+|\theta|}$, $\rho = f^{-1-\frac{1}{\eta\sigma}}$ *is the only non-negative solution (up to multiplication by constant).*

Corollary

There exists no invariant distribution unless $f^{1+\frac{1}{\eta\sigma}}$ *grows fast enough at* ∞.

Proof of Lemma.

Liouville theorem of [Edmunds-Peletier '73]. $\square$

For comparison, if $\Sigma = \sigma I$, then $\tilde{\rho}'_{\sigma\eta} = \frac{1}{Z}\exp\left(-\frac{f(\theta)}{\eta\sigma}\right)$.

Lemma

Assume that

1. *the set* $\{\theta : f(\theta) = 0\}$ *is a compact n-manifold* N,
2. $D^2 f(\theta)$ *has full rank on* N, *and*
3. *there exist* $\gamma > \frac{2m}{m-n}$, $R > 0$ *such that*

$$f(\theta) \geq |\theta|^\gamma \qquad \forall\ |\theta| \geq R.$$

If $\frac{m}{\gamma} < 1 + \frac{1}{\eta\sigma} < \frac{m-n}{2}$, *then* $\tilde{\rho}_{\eta\sigma} = f^{-1-\frac{1}{\eta\sigma}}$ *is integrable.*

Theorem (W '21)

For $\frac{m}{\gamma} < 1 + \frac{1}{\eta\sigma} < \frac{m-n}{2}$, let $\pi_{\eta\sigma}$ be the probability distribution with density proportional to $f^{-1-\frac{1}{\eta\sigma}}$. As $\eta\sigma \searrow \frac{2}{m-n-2}$, the distributions $\pi_{\eta\sigma}$ converge to a distribution π^* on N and π^* has density proportional to

$$\tilde{\rho}^*(\theta) = \int_{S^{m-n-1}} \left(\nu^T \widehat{D^2 f(\theta)} \nu\right)^{-\frac{m-n}{2}} \mathrm{d}\theta.$$

Theorem (W '21)

For $\sigma\eta > 0$, let $\pi'_{\eta\sigma}$ be the probability distribution with density proportional to $\exp(-f/\eta\sigma)$. As $\eta\sigma \searrow 0$, the distributions $\pi_{\eta\sigma}$ converge to a distribution π' on N and π' has density proportional to

$$\tilde{\rho}'(\theta) = \det(\widehat{D^2 f(\theta)})^{-\frac{1}{2}}.$$

- Both functions of $D^2 f$ have the same homogeneity, but

$$\widehat{D^2 f}(\theta) = \mathrm{diag}(1,\lambda) \quad \Rightarrow \quad \tilde{\rho}'(\theta) = \lambda^{-1/2}, \qquad \tilde{\rho}^*(\theta) = \mathrm{agm}^{-1}(1,\lambda).$$

  The *algebraic-geometric mean* satisfies $\lim_{\lambda\to 0} |\log|(\lambda)\,\mathrm{agm}(1,\lambda) = \frac{\pi}{2}$.

- If $\inf f > 0$, the limit of invariant distributions is the same in both cases.

Theorem (W '21)

Assume that $c(1+|\theta|^2) \leq f(\theta) \leq C(1+|\theta|^2)$. *If* ρ_0 *is smooth and compactly supported, there exists a unique solution of the evolution equation*

$$\partial_t \rho = \eta\sigma \,\mathrm{div}\big(f^{-\frac{1}{\eta\sigma}} \nabla \big(f^{1+\frac{1}{\eta\sigma}} \rho\big)\big)$$

and

$$\int_{\mathbb{R}^m} \left|\rho f^{1+\frac{1}{\eta\sigma}} - \left\langle \rho f^{1+\frac{1}{\eta\sigma}} \right\rangle\right|^2 f^{-1-\frac{1}{\eta\sigma}} \,\mathrm{d}\theta$$

decays exponentially fast. In particular

$$\lim_{t\to\infty} \rho = c f^{-1-\frac{1}{\eta\sigma}}.$$

Proof.
Consider an equation for $u = f^{1+\frac{1}{\eta\sigma}}\rho$.

$$\|u\|^2_{L^2_{\eta\sigma}} = \int_{\mathbb{R}^m} u^2 f^{-1-\frac{1}{\eta\sigma}}\,\mathrm{d}x$$

$$\|u\|_{H^1_{\eta\sigma}} = \int_{\mathbb{R}^m} |\nabla u|^2 f^{-\frac{1}{\eta\sigma}}\,\mathrm{d}x$$

and

$$Au = f^{1+\frac{1}{\eta\sigma}}\,\mathrm{div}\left(f^{-\frac{1}{\eta\sigma}}\nabla u\right).$$

Then $\langle Au, v\rangle_{L^2_{\eta\sigma}} = -\langle u, v\rangle_{H^1_{\eta\sigma}}$ and the Poincaré-Hardy inequality

$$\left\|u - \langle u\rangle_{\eta\sigma}\right\|_{L^2_{\eta\sigma}} \leq C\,\|u\|_{H^1_{\eta\sigma}}$$

holds (Bonforte-Dolbeault-Grillo-Vazquez 2010). □

Theorem (W '21)

Assume that there exists a finite set of points $\Theta = \{\theta_1, \ldots, \theta_n\}$ *where* f *vanishes. Assume furthermore that* $m \geq 3$ *and* $1 + \frac{1}{\eta\sigma} = \frac{m}{2}$,

$$f(\theta) \sim |\theta - \theta_i|^2 \log^2(|\theta - \theta_i|)$$

close to θ_i *and* $f(\theta) \sim |\theta|^2 \log^2(|\theta|)$ *at infinity. If* ρ_0 *is smooth and compactly supported, there exists a unique solution of the evolution equation*

$$\partial_t \rho = \eta\sigma \, \mathrm{div}\Big(f^{-\frac{1}{\eta\sigma}} \nabla \Big(f^{1+\frac{1}{\eta\sigma}} \rho\Big)\Big)$$

and

$$\int_{\mathbb{R}^m} \left| \rho f^{1+\frac{1}{\eta\sigma}} - \langle \rho f^{1+\frac{1}{\eta\sigma}} \rangle \right|^2 f^{-1-\frac{1}{\eta\sigma}} \, \mathrm{d}\theta$$

decays exponentially fast. In particular

$$\lim_{t \to \infty} \rho = c f^{-1-\frac{1}{\eta\sigma}}.$$

Heuristic summary:

1. Noise in machine learning has low rank and the intensity depends on the loss.
2. A small, positive step size is admissible in SGD and leads to linear convergence (under assumptions).
3. Toy-SGD prefers minima where $D^2 f$ is small in a precise sense. The geometry of $\{f = 0\}$ does not matter.

Open problems:

1. Validity of continuum model
2. Convergence of continuous time SGD in overparametrized loss landscape
3. Analysis of continuous time SGD with low rank diffusion
   - Existence of the invariant distribution
   - Asymptotics
   - Convergence of SGD
4. The analysis of cross-entropy classification problems must be entirely different since minimizers do not exist.
5. Realistic growth, Lipschitz and convexity assumptions
6. Random pass SGD vs random choice SGD

Thank you for your attention!