Introduction to (Stochastic) Gradient Flow

Yixuan Yang
yixuanyang@zju.edu.cn

2025 Spring

1 Modified Equation(Backward Error Analysis)(Griffiths and Higham, 2010)

The modified equation is a technique in numerical analysis. By constructing new differential equations that closely approximate numerical methods, it can provide valuable insights for our analysis.
Example(ODE):
Consider the autonomous IVP :

$$x'(t) = f(x(t)) \quad \text{with } x(0) = \eta. \tag{1}$$

Forward Euler's method of 1:

$$x_{n+1} = x_n + hf(x_n) \tag{2}$$

where h is the step size.
We consider the local truncation error(LTE) under the localizing assumption that $x_n = x(t_n)$:

$$\begin{aligned} T_{n+1} = x(t_{n+1}) - x_{n+1} &= x(t_{n+1}) - x(t_n) - hf(x(t_n)) \\ &\overset{Taylor}{=} \frac{h^2}{2}x''(t_n) + \mathcal{O}(h^3) \\ &= \mathcal{O}(h^2) \end{aligned}$$

We now construct a modified equation—or more correctly a modified IVP of the form:

$$\begin{aligned} \hat{x}'(t) &= f(\hat{x}(t)) - \frac{h}{2}f'(\hat{x}(t))f(\hat{x}(t)) \\ \hat{x}(0) &= \eta \end{aligned} \tag{3}$$

It follows from the localizing assumption $x_n = \hat{x}(t_n)$ that $x_{n+1} = \hat{x}(t_n) + hf(\hat{x}(t_n))$ and so,

$$\begin{aligned}
\hat{T}_{n+1} &= \hat{x}(t_{n+1}) - x_{n+1} \\
&= \hat{x}(t_{n+1}) - \hat{x}(t_n) - hf(\hat{x}(t_n)) \\
&\overset{Taylor}{=} \hat{x}(t_n) + h\left[f(\hat{x}(t_n)) - \frac{h}{2}f'(\hat{x}(t_n))f(\hat{x}(t_n))\right] \\
&+ \frac{h^2}{2}\frac{d\hat{x}}{dt}\left[f'(\hat{x}(t_n)) - \frac{h}{2}\left(f''(\hat{x}(t_n))f(\hat{x}(t_n)) + f'^2(\hat{x}(t_n))\right)\right] + \mathcal{O}(h^3) \\
&- \hat{x}(t_n) - hf(\hat{x}(t_n))
\end{aligned}$$

Substituting $\frac{d\hat{x}(t)}{dt} = f(\hat{x}(t)) - \frac{h}{2}f'(\hat{x}(t))f(\hat{x}(t))$, we have

$$\begin{aligned}
\hat{T}_{n+1} &= -\frac{h^2}{2}f'(\hat{x}(t_n))f(\hat{x}(t_n)) \\
&+ \frac{h^2}{2}f(\hat{x}(t))f'(\hat{x}(t)) - \frac{h^3}{4}f'^2(\hat{x}(t_n))f(\hat{x}(t_n)) + \mathcal{O}(h^3) \\
&= \mathcal{O}(h^3)
\end{aligned}$$

Therefore, while the method $x_{n+1} = x_n + hf_n$ is a first-order approximation to 1, it is a second-order approximation of the modified equation 3.

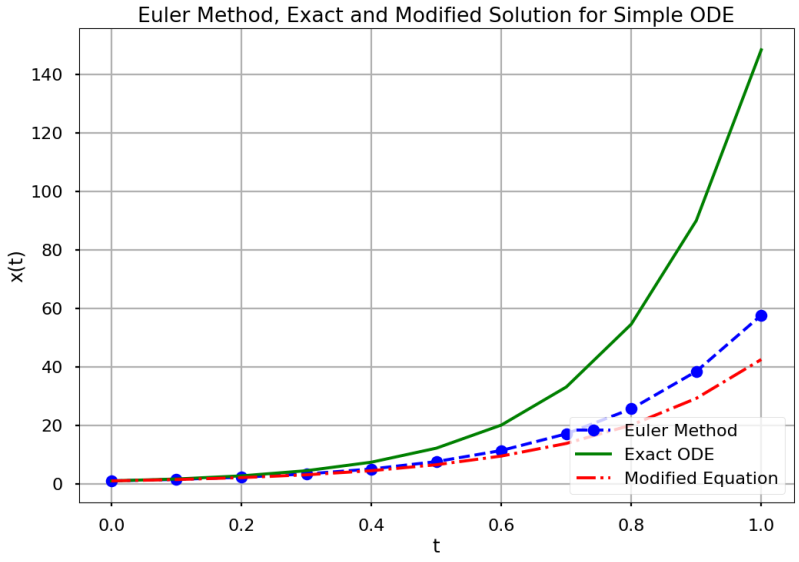

Figure 1: Let step size $h = 0.1$. The modified equation $y'(t) = (5 - 25h/2)x$ is closer to the numerical Euler solution for the original equation $x'(t) = 5x$.

Remark 1.1. *Modified equations are not unique, each numerical method has an unlimited number of modified equations of any given order of accuracy. Similarly, we can find the modified equation for the numerical method of PDE, SDE.*

2 Classical Statistical Learning Theory

$\mathcal{X}$- input space
$\mathcal{Y}$- output space
$\mathcal{D}$- distribution over $\mathcal{X} \times \mathcal{Y}$
$\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$(all functions)- hypothesis space
$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$- loss function
$S \equiv \{(x_i, y_i)\}_{i=1}^{n}$- training dataset
Goal: Find $h : \mathcal{X} \to \mathcal{Y}$ that minimizes the population loss:

$$L_{\mathcal{D}}(h) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(y, h(x))]$$

Approach: Empirical Risk Minimization (ERM) i.e. Given training set S drawn i.i.d. from $\mathcal{D}$, find $h \in \mathcal{H}$(given) that minimizes the empirical loss:

$$L_S(h) := \frac{1}{m}\sum_{i=1}^{m} \ell(y_i, h(x_i))$$

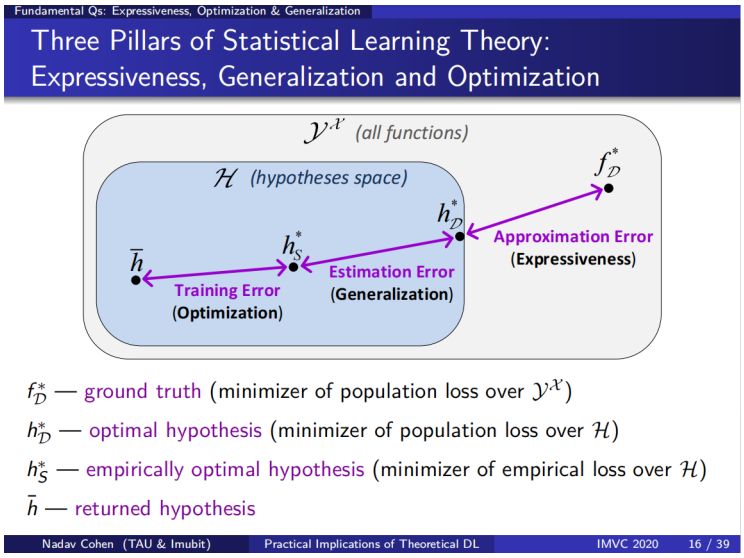

Figure 2: Expressiveness, Generalization, and Optimization(Cohen, 2020)

Approximation: (Li et al., 2022a), (Li et al., 2022b)
Optimization & Generalization:(Liao et al., 2024), (Hu et al., 2021), (Ali et al., 2019), (Ali et al., 2020), (Woodworth et al., 2020), (Pesme et al., 2021)

3 Gradient Flow(Bach, 2020; Jinyuma, 2024)

Minimize the empirical risk:

$$\min_{x\in\mathbb{R}^d} f(x) \tag{4}$$

The gradient descent iteration is written as:

$$x_{k+1} = x_k - \eta\nabla f(x_k) \tag{5}$$

The gradient flow can be viewed as a continuous-time version of the gradient descent algorithm:

$$\dot{X}(t) = -\nabla f(X(t)) \tag{6}$$

It's well-defined for a wide variety of conditions on the function f. The approximation result of GF can be based on the classical Euler discretization of ODEs.

Proposition 3.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function and $X(t)_{t\geq 0}$ follow the gradient flow 6 with any initialization. Then $f(X(t))$ is non-increasing.*

$$\frac{d}{dt}f(X(t)) = \nabla f(X(t))^\top \frac{dX(t)}{dt} = -\|\nabla f(X(t))\|_2^2 \leq 0. \tag{7}$$

Remark 3.1. *Note that:*
1. If f is bounded from below, then $f(X(t))$ will always converge (Easily prove bounded below + monotonically deceasing =convergence.).
2. In general, $X(t)$ may not always converge without any further assumptions, e.g., it may oscillate forever.

Theorem 3.1 (Convergence of gradient flow for convex function)**.** *f is a convex and differentiable function, i.e. $f \in \mathcal{F}^1(\mathbb{R}^n)$. x^* is Optimal value point, then the gradient flow has a sublinear convergence rate:*

$$f(x_t) - f(x^*) \leq \frac{1}{2t}\|x_0 - x^*\|_2^2 \tag{8}$$

Proof. 1. For differentiable f is convex but not strongly convex,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \quad \Longrightarrow \quad f(y) - f(x_t) \geq \langle -\nabla f(x_t), x - y \rangle \tag{9}$$

2. Consider

$$\begin{aligned} \frac{d}{dt}\|x_t - x^*\|_2^2 &\overset{\text{chain rule}}{=} \left\langle \frac{d}{d(x_t - x^*)}\|x_t - x^*\|_2^2, \frac{d}{dt}\big(x_t - x^*\big) \right\rangle \\ &= \left\langle 2\big(x_t - x^*\big), \frac{d}{dt}x_t \right\rangle \\ &\overset{\dot{x}_t = -\nabla f}{=} 2\Big\langle x_t - x^*, -\nabla f\big(x_t\big) \Big\rangle \\ &\overset{(6)}{\leq} 2\Big(f^* - f\big(x_t\big)\Big) = -2\Big(f\big(x_t\big) - f^*\Big). \end{aligned} \tag{10}$$

We now have

$$\frac{d}{dt}\|x_t - x^*\|_2^2 \leq -2\Big(f(x_t) - f^*\Big).$$

Rearrange the inequality gives

$$\begin{aligned} f(x_t) - f^* &\leq \frac{-1}{2}\frac{d}{dt}\|x_t - x^*\|_2^2 \\ \Longleftrightarrow \int_0^t f(x_t) - f^* dt &\leq \int_0^t \frac{-1}{2}\frac{d}{dt}\|x_t - x^*\|_2^2 dt \\ \Longleftrightarrow \int_0^t f(x_t)dt - tf^* &\leq \frac{-1}{2}\|x_t - x^*\|_2^2 - \frac{-1}{2}\|x_0 - x^*\|_2^2 \\ \Longrightarrow \int_0^t f(x_t)dt - tf^* &\leq \frac{1}{2}\|x_0 - x^*\|_2^2 \\ \Longleftrightarrow \frac{1}{t}\int_0^t f(x_t)dt - f^* &\leq \frac{1}{2t}\|x_0 - x^*\|_2^2 \\ \overset{(7)}{\Longrightarrow} f(x_t) - f^* &\leq \frac{1}{2t}\|x_0 - x^*\|_2^2 \end{aligned} \tag{11}$$

Then, we have

$$f(x_t) - f^* \leq O(\frac{1}{t})$$

i.e. For convergence, we need iterate $T \geq \Omega(\frac{1}{\epsilon})$. □

Theorem 3.2 (Convergence of gradient flow for strongly convex function)**.** *f is a m-strongly convex and differentiable function, x^* is Optimal value point, then the gradient flow has a linear convergence rate:*

$$f\big(x_t\big) - f(x^*) \leq e^{-2mt}\bigg(f(x_0) - f(x^*)\bigg). \tag{12}$$

Proof. By Lojasiewicz inequality (4), we have

$$-\|\nabla f(x_t)\|_2^2 \leq -2m\big(f(x_t) - f^*\big).$$

Plug in the Lojasiewicz inequality,

$$\begin{aligned} \frac{d}{dt}f(x_t) &= -\|\nabla f(x_t)\|_2^2 \\ &\leq -2m\big(f(x_t) - f^*\big) \\ \iff \frac{d}{dt}\big(f(x_t) - f^*\big) &\leq -2m\big(f(x_t) - f^*\big) \\ \overset{Rearrange}{\iff} \frac{1}{\big(f(x_t) - f^*\big)}\frac{d}{dt}\big(f(x_t) - f^*\big) &\leq -2m \\ \iff f\big(x_t\big) - f(x^*) \leq e^{-2mt}\Big(f(x_0) - f(x^*)\Big) & \end{aligned} \tag{13}$$

Then, we have

$$f(x_t) - f^* \leq O(e^{-2mt})$$

i.e. For convergence, we need iterate $T \geq \Omega(\log \frac{1}{\epsilon})$. □

4 Stochastic Gradient Flow(Li et al., 2017, 2019)

Minimize the empirical risk:

$$\min_{x\in\mathbb{R}^d} f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) \tag{14}$$

where $f_i(x)$ is the loss function of the i-th sample.
The plain SGD iteration is written as:

$$x_{k+1} = x_k - \eta\nabla f_{\gamma_k}(x_k) \tag{15}$$

where $k \geq 0$ and $\{\gamma_k\}$ are i.i.d uniform variates taking values in $\{1, 2, \ldots, n\}$.
We now introduce the SGF approximation. First, rewrite the SGD iteration rule (15) as:

$$x_{k+1} - x_k = -\eta\nabla f(x_k) + \sqrt{\eta}V_k, \tag{16}$$

where $V_k = \sqrt{\eta}(\nabla f(x_k) - \nabla f_{\gamma_k}(x_k))$ is a d-dimensional random vector.
As seen in equation 16, evaluated at x_k, the stochastic noise V_k has two main characteristics which we want to preserve:

- The expectation of V_k is zero, i.e. $\mathbb{E}[V_k] = 0$.
- The covariance matrix of V_k is $\eta\Sigma(x_k)$, where $\Sigma(x) = \frac{1}{n}\sum_{i=1}^{n}(\nabla f(x) - \nabla f_i(x))(\nabla f(x) - \nabla f_i(x))^T$.

Guided by the previous considerations, we study the following stochastic gradient flow:

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t, \quad X_0 = x_0, \tag{17}$$

whose Euler discretization $X_{k+1} = X_k + \Delta t b(X_k) + \sqrt{\Delta t}\sigma(X_k)Z_k$, $Z_k \sim \mathcal{N}(0, I)$ resembles 16 if we set $\Delta t = \eta$, $b \sim -\nabla f$ and $\sigma \sim (\eta\Sigma)^{1/2}$.
It is now important to discuss the precise meaning of "an approximation".

Definition 4.1. *(Weak Approximation) Let $0 < \eta < 1$, $T > 0$ and set $N = \lfloor T/\eta \rfloor$. Let G denote the set of functions of polynomial growth, i.e. $g \in G$ if there exists constants $K, \kappa > 0$ such that $|g(x)| < K(1 + |x|^\kappa)$. we say that the SDE (17) is an order α weak approximation to the SGD (15) if for every $g \in G$, there exists $C > 0$, independent of η, such that for all $k = 0, 1, \ldots, N$,*

$$|\mathbb{E}g(X_{k\eta}) - \mathbb{E}g(x_k)| < C\eta^\alpha.$$

Intuitively, weak approximations are close to the original process not in terms of individual sample paths, but their distributions.
We now state informally the approximation theorem.

Theorem 4.1. *(Informal Statement) Let $T > 0$. Assume f, f_i are Lipschitz continuous, have at most linear asymptotic growth and have sufficiently high derivatives belonging to G. Then the stochastic process X_t, $t \in [0, T]$ satisfying*

$$dX_t = -\nabla f(X_t)dt + (\eta\Sigma(X_t))^{\frac{1}{2}}dW_t, \tag{18}$$

is an order 1 weak approximation of the SGD.

5 Example

5.1 Linear Regression

Let the predictor matrix X be arbitrary and fixed, and assume the response y follows a standard regression model,

$$y = X\beta_0 + \eta, \tag{19}$$

for some fixed underlying coefficients $\beta_0 \in \mathbb{R}^p$, and noise $\eta \sim (0, \sigma^2 I)$.
Consider the standard (linear) least squares problem

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \frac{1}{2n} \|y - X\beta\|_2^2. \tag{20}$$

where $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$ are a response vector and a matrix of predictors or features, respectively.
Now consider the ℓ_2 regularized version of (20), called ridge regression:

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \frac{1}{n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2, \tag{21}$$

where $\lambda > 0$ is a tuning parameter. The explicit ridge solution is

$$\hat{\beta}^{\text{ridge}}(\lambda) = (X^T X + n\lambda I)^{-1} X^T y. \tag{22}$$

Next, We will consider the statistical (estimation) risk of an estimator $\hat{\beta} \in \mathbb{R}^p$,

$$\text{Risk}(\hat{\beta}; \beta_0) = \mathbb{E}_{\eta, Z} \|\hat{\beta} - \beta_0\|_2^2.$$

Here Z denotes any potential randomness inherent to $\hat{\beta} \in \mathbb{R}^p$ (e.g., due to mini-batching).
Recall the bias-variance decomposition for risk,

$$\begin{aligned} &\text{Risk}(\hat{\beta}(t); \beta_0) \\ &= \|\mathbb{E}_{\eta, Z}(\hat{\beta}(t)) - \beta_0\|_2^2 + \text{tr}\, \text{Cov}_{\eta, Z}(\hat{\beta}(t)) \\ &\equiv \text{Bias}^2(\hat{\beta}(t); \beta_0) + \text{Var}_{\eta, Z}(\hat{\beta}(t)) \end{aligned}$$

5.1.1 GF(Ali et al., 2019)

The gradient flow differential equation for the least squares problem 20

$$\dot{\beta}(t) = \frac{X^T}{n}(y - X\beta(t)), \tag{23}$$

over time $t \geq 0$, subject to an initial condition $\beta(0) = 0$.

Lemma 5.1. *Fix a response y and predictor matrix X. Then the gradient flow problem (23), subject to $\beta(0) = 0$, admits the exact solution*

$$\hat{\beta}^{\text{gf}}(t) = (X^T X)^+ (I - \exp(-tX^T X/n)) X^T y, \tag{24}$$

for all $t \geq 0$. Here A^+ is the Moore-Penrose generalized inverse of a matrix A, and $\exp(A) = I + A + A^2/2! + A^3/3! + \ldots$ is the matrix exponential.

Theorem 5.1. *Let $t \geq 0$. Consider the data model (19). Then,*
(a) $\text{Bias}^2(\hat{\beta}^{\text{gf}}(t); \beta_0) \leq \text{Bias}^2(\hat{\beta}^{\text{ridge}}(1/t); \beta_0)$.
(b) $\text{Var}(\hat{\beta}^{\text{gf}}(t); \beta_0) \leq 1.6862\ \text{Var}(\hat{\beta}^{\text{ridge}}(1/t); \beta_0)$.
(c) So that $\text{Risk}(\hat{\beta}^{\text{gf}}(t); \beta_0) \leq 1.6862\ \text{Risk}(\hat{\beta}^{\text{ridge}}(1/t); \beta_0)$.

5.1.2 SGF(Ali et al., 2020)

Mini-batch SGD applied to (20) is the iteration

$$\begin{aligned} \beta^{(k)} = & \quad \beta^{(k-1)} + \frac{\epsilon}{m} \sum_{i \in \mathcal{I}_k} (y_i - x_i^T \beta^{(k-1)}) x_i \\ = & \quad \beta^{(k-1)} + \frac{\epsilon}{m} X_{\mathcal{I}_k}^T (y_{\mathcal{I}_k} - X_{\mathcal{I}_k} \beta^{(k-1)}), \end{aligned} \tag{25}$$

for $k = 1, 2, 3, \ldots$, where $\epsilon > 0$ is a fixed step size, m is the mini-batch size, and $\mathcal{I}_k \subseteq \{1, \ldots, n\}$ denotes the minibatch on iteration k with $|\mathcal{I}_k| = m$, for all k. For simplicity, we assume the mini-batches are sampled with replacement.
Now, adding and subtracting the negative gradient of the loss in (25) yields

$$\begin{aligned} \beta^{(k)} = \beta^{(k-1)} + \frac{\epsilon}{n} \cdot X^T (y - X\beta^{(k-1)}) \\ + \epsilon \cdot \left(\frac{1}{m} X_{\mathcal{I}_k}^T (y_{\mathcal{I}_k} - X_{\mathcal{I}_k} \beta^{(k-1)}) - \frac{1}{n} X^T (y - X\beta^{(k-1)}) \right). \end{aligned} \tag{26}$$

This may be recognized as gradient descent, plus the deviation between the sample average of m i.i.d. random variables and their mean, which motivates the continuous-time dynamics (stochastic differential equation)

$$d\beta(t) = \frac{1}{n} X^T (y - X\beta(t)) dt + Q_\epsilon(\beta(t))^{1/2} dW(t), \tag{27}$$

with $\beta(0) = 0$. We denote the diffusion coefficient

$$Q_\epsilon(\beta) = \epsilon \cdot \mathrm{Cov}_{\mathcal{I}} \left(\frac{1}{m} X_{\mathcal{I}}^T (y_{\mathcal{I}} - X_{\mathcal{I}} \beta) \right),$$

where the randomness is due to $\mathcal{I} \subseteq \{1, \ldots, n\}$.

Lemma 5.2. *Fix y, X, and $\epsilon > 0$. Let $t \geq 0$. Then*

$$\hat{\beta}^{\mathrm{sgf}}(t) = \hat{\beta}^{\mathrm{gf}}(t) + \exp(-t\hat{\Sigma}) \cdot \int_0^t \exp(\tau\hat{\Sigma}) Q_\epsilon(\hat{\beta}^{\mathrm{sgf}}(\tau))^{1/2} dW(\tau) \tag{28}$$

is the unique solution to the differential equation 27. Here $\hat{\Sigma} = X^T X/n$.

Theorem 5.2. *Fix X. Let $t > 0$ and ϵ be small enough.*

- *Then, relative to gradient flow,*

$$\mathrm{Risk}(\hat{\beta}^{\mathrm{sgf}}(t); \beta_0) \leq \mathrm{Bias}^2(\hat{\beta}^{\mathrm{gf}}(t); \beta_0) + \mathrm{Var}_\eta(\hat{\beta}^{\mathrm{gf}}(t)) + \epsilon \cdot \frac{n}{m} \sum_{i=1}^p \mathbb{E}_\eta \nu_i(t). \tag{29}$$

- *Relative to ridge regression,*

$$\mathrm{Risk}(\hat{\beta}^{\mathrm{sgf}}(t);\beta_0) \le \mathrm{Bias}^2(\hat{\beta}^{\mathrm{ridge}}(1/t);\beta_0) + 1.6862\,\mathrm{Var}_\eta(\hat{\beta}^{\mathrm{gf}}(1/t)) + \epsilon\cdot\frac{n}{m}\sum_{i=1}^{p}\mathbb{E}_\eta\nu_i(t). \tag{30}$$

Proof Sketch: Note that
1. $\mathrm{Bias}^2(\hat{\beta}^{\mathrm{sgf}}(t);\beta_0) = \mathrm{Bias}^2(\hat{\beta}^{\mathrm{gf}}(t);\beta_0)$.
2. $\mathrm{tr}\,\mathrm{Cov}_{\eta,Z}(\hat{\beta}^{\mathrm{sgf}}(t)) = \mathrm{tr}\,\mathbb{E}_\eta\left[\mathrm{Cov}_Z(\hat{\beta}^{\mathrm{sgf}}(t)\mid\eta)\right] + \mathrm{Var}_\eta(\hat{\beta}^{\mathrm{gf}}(t))$.

Then we just need to bound the first term of the right-hand side of 2.

5.2 Diagonal Linear Neural Network(toy model of neural network)

Consider 2-layer diagonal linear network:

$$f(\mathbf{w},\mathbf{x}) = \sum_{i=1}^{d}(\mathbf{w}_{+,i}^{\odot 2} - \mathbf{w}_{-,i}^{\odot 2})\mathbf{x}_i = \langle\beta_{\mathbf{w}},\mathbf{x}\rangle, \text{where } \mathbf{w} = \begin{bmatrix}\mathbf{w}_+\\ \mathbf{w}_-\end{bmatrix} \in \mathbb{R}^{2d}, \text{ and } \beta_{\mathbf{w}} = \mathbf{w}_+^{\odot 2} - \mathbf{w}_-^{\odot 2}.$$

Although this model appears simple, it is inherently non-convex.

Remark 5.1. *The standard 2-layer neural network architecture is as follows:*

$$f(W,x) = W_2\sigma(W_1x) \tag{31}$$

where $W_1 \in \mathbb{R}^{d\times d}$, $W_2 \in \mathbb{R}^{1\times d}$ are weight matrices, σ is the activation function.
Let σ be the identity mapping and W_1 be a diagonal matrix, then the model becomes a diagonal linear network:

$$f(W,x) = \langle w_1 \odot w_2, x\rangle \tag{32}$$

where $w_1, w_2 \in \mathbb{R}^d$ are the diagonal elements of W_1, W_2 respectively.
If $w_1(0) = w_2(0)$ at initialization, then their magnitudes will remain equal throughout GF training.
Therefore, we can equivalently parametrize the model in terms of a single shared input and output weight w_i for each hidden unit, yielding the model $f(w,x) = \langle w^{\odot 2}, x\rangle$.
The reason for using two weights w_+ and w_- is two-fold:

- *It ensures that the elements of the weight vector w can take negative values, rather than being restricted to positive values.*

- *It allows for initialization at $f(w_0) = 0$ (by choosing $w_+(0) = w_-(0)$) without this being a saddle point from which gradient flow will never escape.*

Another perspective is that we can interpret DLNN as a reparameterization of the linear model.

5.2.1 GF(Woodworth et al., 2020)

Consider the squared loss of the DLNN over a training set $\{x_i, y_i\}_{i=1}^n$:

$$L(\mathbf{w}) = \sum_{i=1}^{n} \left(f(\mathbf{w}, x_i) - y_i\right)^2 \tag{33}$$

We study the underdetermined $N << d$ case where there are many possible solutions $X\beta = y$. Then the gradient flow dynamics:

$$\dot{\mathbf{w}}(t) = -\nabla L(\mathbf{w}(t)). \tag{34}$$

with the initial condition $\mathbf{w}_+(0) = \mathbf{w}_-(0) = \alpha \in \mathbb{R}^d$.

Theorem 5.3. *Let α with no zero entries, if the gradient flow solution β_α^∞ satisfies $X\boldsymbol{\beta}_\alpha^\infty = \mathbf{y}$, then*

$$\beta_\alpha^\infty = \underset{\beta \in \mathbb{R}^d \text{ s.t. } X\beta = y}{\arg\min} \phi_\alpha(\beta) := \frac{1}{4}\left[\sum_{i=1}^{d} \beta_i \text{arcsinh}(\frac{\beta_i}{2\alpha_i^2}) - \sqrt{\beta_i^2 + 4\alpha_i^4}\right]. \tag{35}$$

where ϕ is hyperbolic entropy function.

Proof Sketch: First, we show that if $X\boldsymbol{\beta}_{\alpha,\mathbf{1}}^\infty = \mathbf{y}$, then $\boldsymbol{\beta}_{\alpha,\mathbf{1}}^\infty = b_\alpha(X^\top \nu)$ for a certain function b_α and vector v. Next, we suppose that there is some function ϕ_α such that (35) holds. The KKT optimality conditions for (35) are $X\boldsymbol{\beta}^* = y$ and $\exists \nu$ s.t. $\nabla\phi_\alpha(\boldsymbol{\beta}^*) = X^\top\nu$. Therefore, if indeed $\boldsymbol{\beta}_{\alpha,1}^\infty = \boldsymbol{\beta}^*$ and $X\boldsymbol{\beta}_{\alpha,1}^\infty = \mathbf{y}$ then $\nabla\phi_\alpha\left(\beta_{\alpha,1}^\infty\right) = \nabla\phi_\alpha\left(b_\alpha(X^\top\nu)\right) = X^\top\nu$. We solve the differential equation $\nabla\phi_\alpha = b_\alpha^{-1}$ to yield ϕ_α.

Though the hyperbolic entropy function has a non-trivial expression, its principal characteristic is that it interpolates between the ℓ_1 and the ℓ_2 norms according to the scale of α.

Theorem 5.4. *(Special case) For any $0 < \epsilon < d$, under the setting of Theorem 5.3 with $\mathbf{w}_+(0) = \mathbf{w}_-(0) = \alpha\mathbf{1} \in (\mathbb{R}_+)^d$,*

$$\alpha \leq \min\left\{\left(2(1+\epsilon)\|\beta_{\ell_1}^*\|_1\right)^{-\frac{2+\epsilon}{2\epsilon}}, \exp\left(-d/(\epsilon\|\beta_{\ell_1}^*\|_1)\right)\right\} \Longrightarrow \|\beta_{\alpha,1}^\infty\|_1 \leq (1+\epsilon)\|\beta_{\ell_1}^*\|_1$$

$$\alpha \geq \sqrt{2(1+\epsilon)(1+2/\epsilon)\|\beta_{\ell_2}^*\|_2} \Longrightarrow \|\beta_{\alpha,1}^\infty\|_2^2 \leq (1+\epsilon)\|\beta_{\ell_2}^*\|_2^2$$

.

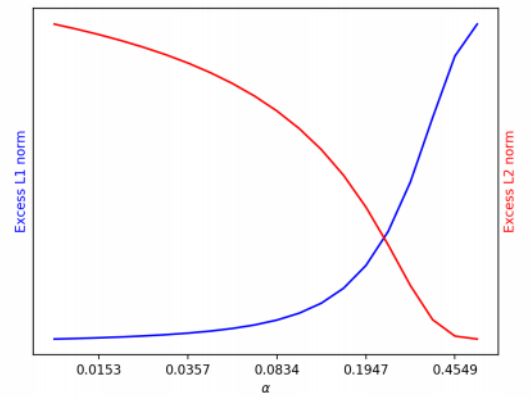

Figure 3: We plot $\|\beta^{\infty}_{\alpha,\mathbf{1}}\|_1 - \|\beta^*_{\ell_1}\|_1$ in blue and $\|\beta^{\infty}_{\alpha,\mathbf{1}}\|_2 - \|\beta^*_{\ell_2}\|_2$ in red vs. α.

5.2.2 SGF(Pesme et al., 2021)

We study the quadratic loss and the overall loss is written as:

$$L(w) = L(\beta_w) \equiv \frac{1}{4n}\sum_{i=1}^{n}\left(f(\mathbf{w}, x_i) - y_i\right)^2 \tag{36}$$

where $\{(x_i, y_i)\}_{i=1}^n$ are linear measurements, i.e. $y_i = \langle \beta^*, x_i\rangle$, $i = 1, 2, \ldots, n$. By abuse of notation we use $L(w) = L(\beta_w)$.
Stochastic gradient flow dynamics:

$$\mathrm{d}w_{t,+} = -\nabla_{w_+} L(w_t)\mathrm{d}t + 2\sqrt{\gamma n^{-1} L(w_t)}\, w_{t,+} \odot [X^\top \mathrm{d}B_t] \tag{37}$$

$$\mathrm{d}w_{t,-} = -\nabla_{w_-} L(w_t)\mathrm{d}t - 2\sqrt{\gamma n^{-1} L(w_t)}\, w_{t,-} \odot [X^\top \mathrm{d}B_t], \tag{38}$$

where $\gamma > 0$ is the step size. The initial condition is $w_+(0) = w_-(0) = \alpha \in (\mathbb{R}_+)^d$.

Remark 5.2. *Compared with the SGF 18 introduced earlier, the SGF here simplifies the diffusion term through an approximation to facilitate the analysis. See (Pesme et al., 2021) for more details. Is the simplification reasonable, and do the approximation results exist? These are key questions that require thorough analysis and justification.*

Theorem 5.5. *For $p \le \frac{1}{2}$, let $(w_t)_{t\ge 0}$ follow the stochastic gradient flow (37),(38) with step size $\gamma \le O\left(\left[\ln(\frac{4}{p})\lambda_{\max}\max\{\|\beta^*_{\ell_1}\|_1 \ln\left(\frac{\|\beta^*_{\ell_1}\|_1}{\min_i \alpha_i^2}\right), \|\alpha\|_2^2\}\right]^{-1}\right)$ where $\beta^*_{\ell_1} = \underset{\beta\in\mathbb{R}^d s.t. X\beta = y}{\arg\min} \|\beta\|_1$ and $\lambda_{\max}$ is the largest eigenvalue of $X^\top X/n$. Then with probability at least $1-p$:*

- $(\beta_t)_{t\geq 0}$ *converges towards a zero-training error solution* β_∞^α
- *the solution* β_∞^α *satisfies*

$$\beta_\infty^\alpha = \underset{\beta\in\mathbb{R}^d \text{ s.t. } X\beta=y}{\arg\min} \phi_{\alpha_\infty}(\beta) \quad \textit{where} \quad \alpha_\infty = \alpha \odot \exp\left(-2\gamma \operatorname{diag}\left(\frac{X^\top X}{n}\right) \int_0^{+\infty} L(\beta_s)\mathrm{d}s\right). \tag{39}$$

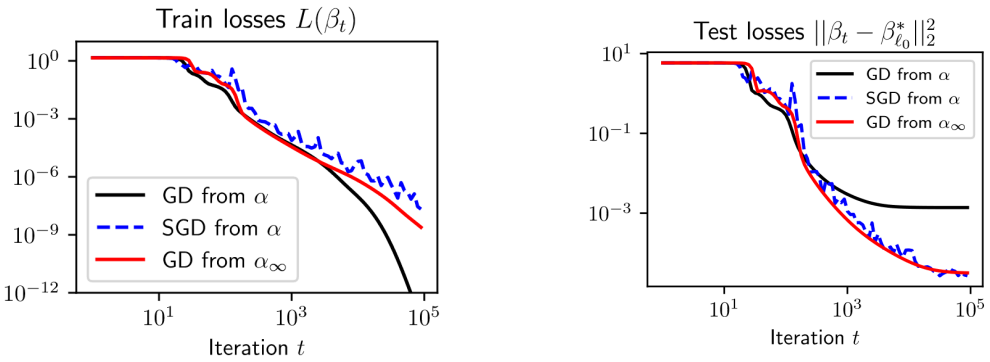

Figure 4: Left and right: SGD initialised at $\alpha 1$ converges towards the same point as GD initialised at $\alpha_\infty = \alpha \odot \exp\left(-2\gamma \operatorname{diag}\left(X^\top X/n\right) \int_0^{+\infty} L(\beta_s^{\mathrm{SGD}})\mathrm{d}s\right)$.

References

Alnur Ali, J Zico Kolter, and Ryan J Tibshirani. A continuous-time view of early stopping for least squares regression. In *The 22nd international conference on artificial intelligence and statistics*, pages 1370–1378. PMLR, 2019.

Alnur Ali, Edgar Dobriban, and Ryan Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In *International conference on machine learning*, pages 233–244. PMLR, 2020.

Francis Bach. Effortless optimization through gradient flows, 2020. URL https://francisbach.com/gradient-flows/.

Nadav Cohen. Practical implications of theoretical deep learning, 2020. Presented at the Israel Machine Vision Conference (IMVC), Expo Tel Aviv, October 2020.

David Francis Griffiths and Desmond J Higham. *Numerical methods for ordinary differential equations: initial value problems*, volume 5. Springer, 2010.

Kaitong Hu, Zhenjie Ren, David Šiška, and Łukasz Szpruch. Mean-field langevin dynamics and energy landscape of neural networks. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, volume 57, pages 2043–2065. Institut Henri Poincaré, 2021.

Jinyuma. Gradient flow and differential equations in optimization algorithms, 2024. URL https://zhuanlan.zhihu.com/p/664979479.

Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pages 2101–2110. PMLR, 2017.

Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019.

Qianxiao Li, Ting Lin, and Zuowei Shen. Deep learning via dynamical systems: An approximation perspective. *Journal of the European Mathematical Society*, 25(5):1671–1709, 2022a.

Zhong Li, Jiequn Han, E Weinan, and Qianxiao Li. Approximation and optimization theory for linear continuous-time recurrent neural networks. *Journal of Machine Learning Research*, 23(42):1–85, 2022b.

Huafu Liao, Alpár R Mészáros, Chenchen Mou, and Chao Zhou. Convergence analysis of controlled particle systems arising in deep learning: from finite to infinite sample size. *arXiv preprint arXiv:2404.05185*, 2024.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230, 2021.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR, 2020.