Stochastic Gradient Descent

Dr. Dingzhu Wen

School of Information Science and Technology (SIST)
ShanghaiTech University

wendzh@shanghaitech.edu.cn

March 13, 2024

Overview

SGD from A Learning Perspctive

Ideal Learning Objective:

$$E(f) = \mathcal{L}(\mathbf{w}) = \int \ell(\mathbf{w}; \mathbf{z}) \mathrm{d}P(\mathbf{z}),$$

- $\mathbf{z} = [\mathbf{x}^T, y]^T$, the training data sample,
- $\mathbf{w}$, the machine learning model parameter vector,
- $P(\mathbf{z})$, the distribution of $\mathbf{z}$,
- $L(\mathbf{w})$, the loss function,
- $\ell\left(f_{\mathbf{w}}(\mathbf{x}), y\right)$, the learning loss regarding one sample, e.g.,

$$\ell\left(f_{\mathbf{w}}(\mathbf{x}), y\right) = \left|y - f_{\mathbf{w}}(\mathbf{x})\right|^2,$$

- $f_{\mathbf{w}}(\mathbf{x})$, the machine learning model, e.g., linear regression $f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$.

SGD from A Learning Perspctive

Empirical Risk Function:

$$E(f) = \mathcal{L}(\mathbf{w}) = \frac{1}{M}\sum_{i=1}^{M} \ell(\mathbf{w}; \mathbf{z}_i),$$

- $\mathbf{z}_i = [\mathbf{x}_i^T, y_i]^T$, the training data sample,
- $\mathbf{w}$, the machine learning model parameter vector,
- $L(\mathbf{w})$, the loss function,
- $\ell\left(f_{\mathbf{w}}(\mathbf{x}_i), y_i\right)$, the learning metric regarding one sample, e.g.,

$$\ell\left(f_{\mathbf{w}}(\mathbf{x}_i), y_i\right) = \left|y_i - f_{\mathbf{w}}(\mathbf{x}_i)\right|^2,$$

- $f_{\mathbf{w}}(\mathbf{x}_i)$, the machine learning model.

Gradient Descent (GD) for Training

Gradient Descent Method:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \frac{1}{M} \sum_{i=1}^{M} \nabla_{\mathbf{w}_t} \ell(\mathbf{w}_t; \mathbf{z}_i),$$

- t, the number of training iteration,
- η, learning rate(step size).
- Convergece under (strong) convexity and L-smoothness (refer to Lecture 2).

Stochastic Gradient Descent (SGD) for Training

Stochastic Gradient Descent Method:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla_{\mathbf{w}_t} \ell(\mathbf{w}_t; \mathbf{z}_{i^t}),$$

- t, the number of training iteration,
- η, learning rate(step size),
- $\mathbf{z}_{i^t}$, the sample selected in this iteration.

A Single Sample Picked in Each Iteration

- A simplification of GD with noise
- Hope that SGD behave similarly as GD in expectation,
- i^t should be randomly selected.

SGD Applications

Loss	Stochastic gradient algorithm
Adaline $Q_{\text{adaline}} = \frac{1}{2}\big(y - w^\top \Phi(x)\big)^2$ Features $\Phi(x) \in \mathbb{R}^d$, Classes $y = \pm 1$	$w \leftarrow w + \gamma_t \big(y_t - w^\top \Phi(x_t)\big)\, \Phi(x_t)$
Perceptron $Q_{\text{perceptron}} = \max\{0, -y\, w^\top \Phi(x)\}$ Features $\Phi(x) \in \mathbb{R}^d$, Classes $y = \pm 1$	$w \leftarrow w + \gamma_t \begin{cases} y_t\, \Phi(x_t) & \text{if } y_t\, w^\top \Phi(x_t) \le 0 \\ 0 & \text{otherwise} \end{cases}$
K-Means $Q_{\text{kmeans}} = \min_k \frac{1}{2}(z - w_k)^2$ Data $z \in \mathbb{R}^d$ Centroids $w_1 \dots w_k \in \mathbb{R}^d$ Counts $n_1 \dots n_k \in \mathbb{N}$, initially 0	$k^* = \arg\min_k (z_t - w_k)^2$ $n_{k^*} \leftarrow n_{k^*} + 1$ $w_{k^*} \leftarrow w_{k^*} + \frac{1}{n_{k^*}}(z_t - w_{k^*})$ (counts provide optimal learning rates!)
SVM $Q_{\text{svm}} = \lambda w^2 + \max\{0, 1 - y\, w^\top \Phi(x)\}$ Features $\Phi(x) \in \mathbb{R}^d$, Classes $y = \pm 1$ Hyperparameter $\lambda > 0$	$w \leftarrow w - \gamma_t \begin{cases} \lambda w & \text{if } y_t\, w^\top \Phi(x_t) > 1, \\ \lambda w - y_t\, \Phi(x_t) & \text{otherwise.} \end{cases}$
Lasso $Q_{\text{lasso}} = \lambda \lvert w \rvert_1 + \frac{1}{2}\big(y - w^\top \Phi(x)\big)^2$ $w = (u_1 - v_1, \dots, u_d - v_d)$ Features $\Phi(x) \in \mathbb{R}^d$, Classes $y = \pm 1$ Hyperparameter $\lambda > 0$	$u_i \leftarrow \big[u_i - \gamma_t\big(\lambda - (y_t - w^\top \Phi(x_t))\Phi_i(x_t)\big)\big]_+$ $v_i \leftarrow \big[v_i - \gamma_t\big(\lambda + (y_t - w^\top \Phi(x_t))\Phi_i(x_t)\big)\big]_+$ with notation $[x]_+ = \max\{0, x\}$.

Convergence Under Convexity

Assumption (Unbiased Estimation)

The expectation of the stochastic gradient in each iteration is an unbiased estimation of the ground-true gradient:

$$\mathbb{E}\left[\nabla_{\mathbf{w}_t}\ell(\mathbf{w}_t;\mathbf{z}_{i^t})\middle|\mathbf{w}_t\right]=\frac{1}{M}\sum_{i=1}^{M}\nabla_{\mathbf{w}_t}\ell(\mathbf{w}_t;\mathbf{z}_i)=\nabla_{\mathbf{w}_t}\mathcal{L}(\mathbf{w}_t).$$

Assumption (Convexity)

$\{\ell(\mathbf{w};\mathbf{z}_i),\ 1\le i\le M\}$ *are convex.*

Assumption (Bounded Stochastic Gradient Norm)

$\|\nabla_{\mathbf{w}}\ell(\mathbf{w};\mathbf{z}_i)\|\le B,\ 1\le i\le M,\ \forall\mathbf{w}$.

Question: How to use this assumptions to show the convergence?

Convergence Under Convexity

SGD Iteration:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla_{\mathbf{w}_t} \ell(\mathbf{w}_t; \mathbf{z}_{i^t}).$$

Let's mimic the proof in the deterministic case:

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2 &= \|\mathbf{w}_t - \eta \nabla_{\mathbf{w}_t} \ell(\mathbf{w}; \mathbf{z}_{i^t}) - \mathbf{w}_*\|^2, \\ &= \|\mathbf{w}_t - \mathbf{w}_*\|^2 - 2\eta \nabla_{\mathbf{w}_t} \ell(\mathbf{w}; \mathbf{z}_{i^t})^T (\mathbf{w}_t - \mathbf{w}_*) \\ &\quad + \eta^2 \|\nabla_{\mathbf{w}_t} \ell(\mathbf{w}; \mathbf{z}_{i^t})\|^2.\end{aligned}$$

With the fact that i^t is independently seclected of $\mathbf{w}_t$ and based on the assumptions of convexity and unbiased estimation,

$$\mathbb{E}\left[\nabla_{\mathbf{w}_t} \ell(\mathbf{w}; \mathbf{z}_{i^t})^T (\mathbf{w}_t - \mathbf{w}_*) | \mathbf{w}_t\right] \geq \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_*).$$

$\downarrow \mathbb{E}$

$\nabla \mathcal{L}(\vec{w}_t)$

First-order equivalent condition

Convergence Under Convexity

SGD Iteration:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla_{\mathbf{w}_t} \ell(\mathbf{w}_t; \mathbf{z}_{i^t}).$$

From the perspective of expectation,

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2 \middle| \mathbf{w}_t\right] =& \mathbb{E}\left[\|\mathbf{w}_t - \eta \nabla_{\mathbf{w}_t} \ell(\mathbf{w}; \mathbf{z}_{i^t}) - \mathbf{w}_*\|^2 \middle| \mathbf{w}_t\right], \\
\leq& \mathbb{E}\left[\|\mathbf{w}_t - \mathbf{w}_*\|^2 \middle| \mathbf{w}_{t-1}\right] \\
& - 2\eta \mathbb{E}\left[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_*) \middle| \mathbf{w}_{t-1}\right] \\
& + \eta^2 B^2.
\end{aligned}$$

Summation over t from 0 to $T-1$.

Convergence Under Convexity

SGD Iteration:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla_{\mathbf{w}_t} \ell(\mathbf{w}_t; \mathbf{z}_{i^t}).$$

Theorem (Convergence of SGD under Convexity)

Let $\mathcal{L}(\mathbf{w}) = \frac{1}{M} \sum_{i=1}^{M} \ell(\mathbf{w}; \mathbf{z_i})$ *and* $\{\ell(\mathbf{w}; \mathbf{z}_i),\ 1 \leq i \leq M\}$ *be convex, then the sequence* $\{\mathbf{w}_t\}_{t \in \mathbb{N}}$ *generated by SGD with step size* $\eta = \frac{\|\mathbf{w}_0 - \mathbf{w}_*\|}{\sqrt{T+1}B}$ *satisfies*

$$\mathbb{E}\left[\bar{\mathcal{L}}(\mathbf{w}_T) - \mathcal{L}(\mathbf{w}_*)\right] \leq \frac{\|\mathbf{w}_0 - \mathbf{w}_*\| B}{\sqrt{T+1}}.$$

$\bar{L}(W_T) = \frac{1}{T+1} \sum_{t=0}^{T} L(W_t)$

Convergence Under Strong Convexity & Smoothness

Assumption (Unbiased Estimation)

The expectation of the stochastic gradient in each iteration is an unbiased estimation of the ground-true gradient:

$$\mathbb{E}\left[\nabla_{\mathbf{w}_t}\ell(\mathbf{w}_t;\mathbf{z}_{i^t})\middle|\mathbf{w}_t\right] = \frac{1}{M}\sum_{i=1}^{M}\nabla_{\mathbf{w}_t}\ell(\mathbf{w}_t;\mathbf{z}_i) = \nabla_{\mathbf{w}_t}\mathcal{L}(\mathbf{w}_t).$$

Assumption (Strong Convexity and Smoothness)

$\{\ell(\mathbf{w};\mathbf{z}_i),\ 1 \le i \le M\}$ *are* μ*-strongly convex and* L*-smooth.*

Question: How to use this assumptions to show the convergence?

Convergence Under Strong Convexity & Smoothness

Descent Lemma of L-smooth:

$$\mathcal{L}(\mathbf{w}_{t+1}) \leq \mathcal{L}(\mathbf{w}_t) + \nabla_{\mathbf{w}_t}\mathcal{L}(\mathbf{w}_t)^T(\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{L}{2}\left\|\mathbf{w}_{t+1} - \mathbf{w}_t\right\|^2.$$

For the t-th SGD iteration,

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\nabla_{\mathbf{w}_t}\ell(\mathbf{w}_t; \mathbf{z}_{i^t}).$$

Then,

$$\mathcal{L}(\mathbf{w}_{t+1}) \leq \mathcal{L}(\mathbf{w}_t) - \eta\nabla_{\mathbf{w}_t}\mathcal{L}(\mathbf{w}_t)^T\nabla_{\mathbf{w}_t}\ell(\mathbf{w}_t; \mathbf{z}_{i^t}) + \frac{L\eta^2}{2}\left\|\nabla_{\mathbf{w}_t}\ell(\mathbf{w}_t; \mathbf{z}_{i^t})\right\|^2.$$

Convergence Under Strong Convexity & Smoothness

$$EX^2 = (EX)^2 + Var(X)$$

Conditioned on all past iterations,

$$\mathbb{E}\left[\mathcal{L}(\mathbf{w}_{t+1})\middle|\mathbf{w}_t\right] \leq \mathcal{L}(\mathbf{w}_t) - \eta \left\|\nabla_{\mathbf{w}_t}\mathcal{L}(\mathbf{w}_t)\right\|^2 + \frac{L\eta^2}{2}\mathbb{E}\left[\left\|\nabla_{\mathbf{w}_t}\ell(\mathbf{w}_t;\mathbf{z}_{i^t})\right\|^2\middle|\mathbf{w}_t\right].$$

Quite unfortunately,

$$\leq \left\|\nabla L_{\vec{w}_t}(\vec{w}_t)\right\|^2 + \sigma^2$$

$$\mathbb{E}\left[\left\|\nabla_{\mathbf{w}_t}\ell(\mathbf{w}_t;\mathbf{z}_{i^t})\right\|^2\middle|\mathbf{w}_t\right] = \frac{1}{M}\sum_{i=1}^{M}\left\|\nabla_{\mathbf{w}_t}\ell(\mathbf{w}_t;\mathbf{z}_{i^t})\right\|^2 \geq \left\|\nabla_{\mathbf{w}_t}\mathcal{L}(\mathbf{w}_t;\mathbf{z}_{i^t})\right\|^2.$$

$$\nabla_{\vec{w}_t}\ell(\vec{w}_t;\vec{z}_{i^t}) = \nabla_{\vec{w}_t}L(\vec{w}_t) + \vec{n}.$$

$$E(\vec{n}) = \vec{0}, \quad Var(\vec{n}) \neq \vec{0}.$$

Convergence Under Strong Convexity & Smoothness

Assumption (Bounded Variance)

$$\mathbb{E}\left[\left\|\nabla_{\mathbf{w}_t}\ell(\mathbf{w}_t;\mathbf{z}_{i^t})-\nabla_{\mathbf{w}_t}\mathcal{L}(\mathbf{w}_t;\mathbf{z}_{i^t})\right\|^2\middle|\mathbf{w}_t\right]\leq\sigma^2.$$

As a result,

$$\sum_{t=0}^{T}\mathbb{E}\left[\mathcal{L}(\mathbf{w}_{t+1})\middle|\mathbf{w}_t\right]\leq\sum_{t=0}^{T}\mathcal{L}(\mathbf{w}_t)-\eta\left(1-\frac{L\eta}{2}\right)\left\|\nabla_{\mathbf{w}_t}\mathcal{L}(\mathbf{w}_t)\right\|^2+\frac{L\eta^2\sigma^2}{2}.$$

$\mathcal{L}(\mathbf{w}_{t+1}) \geq \mathcal{L}_*$

Convergence Under Strong Convexity & Smoothness

Theorem (Convergence under L-Smoothness and Bounded Variance)

Let $\mathcal{L}(\cdot)$ be L-smooth with bounded stochastic gradient variance, then the sequence $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ generated by SGD with step size $\eta = 1/L$ satisifies

$$\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla_{\mathbf{w}_t}\mathcal{L}(\mathbf{w}_t)\|^2 \leq \frac{2L}{T+1}\left\{\mathbb{E}\left[\mathcal{L}(\mathbf{w}_{t+1})\middle|\mathbf{w}_t\right] - \mathcal{L}_*\right\} + \sigma^2.$$

With fixed step size, the gradient won't converge to 0.

Convergence Under Strong Convexity & Smoothness

Property of Strong Convexity:

$$\mathcal{L}_* \geq \mathcal{L}(\mathbf{w}) - \frac{1}{2\mu} \|\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w})\|^2, \quad \forall \mathbf{w}.$$

Try to prove it using the definition of μ-strongly convex.

$$\mathcal{L}(\mathbf{w}_1) \geq \mathcal{L}(\mathbf{w}) + \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w})^T (\mathbf{w}_1 - \mathbf{w}) + \frac{\mu}{2} \|\mathbf{w}_1 - \mathbf{w}\|^2.$$

$G(\vec{w}_i)$

$L(a) = L_* \geq G_i(a) \geq G_*$, $G_i(\vec{w}_i)$ is quadratic and convex.

$\nabla G(\vec{w}_i^*) = 0 = \nabla_{\vec{w}} L(\vec{w}) + \mu(\vec{w}_i^* - \vec{w})$

$\Rightarrow \vec{w}_i^* = \vec{w} - \frac{1}{\mu} \nabla_{\vec{w}} L(\vec{w})$

Convergence Under Strong Convexity & Smoothness

Under Smoothness and Strong Convexity:

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}(\mathbf{w}_{t+1})\middle|\mathbf{w}_t\right] &\leq \mathcal{L}(\mathbf{w}_t) - \eta\left(1-\frac{L\eta}{2}\right)\|\nabla_{\mathbf{w}_t}\mathcal{L}(\mathbf{w}_t)\|^2 + \frac{L\eta^2\sigma^2}{2},\\ &\leq \mathcal{L}(\mathbf{w}_t) - \eta\left(1-\frac{L\eta}{2}\right)2\mu[\mathcal{L}(\mathbf{w}_t)-\mathcal{L}_*] + \frac{L\eta^2\sigma^2}{2}.\end{aligned}$$

property of strong convexity

In conclusion,

$$\mathbb{E}\left[\mathcal{L}(\mathbf{w}_{t+1})\middle|\mathbf{w}_t\right] - \mathcal{L}_* \leq \left[1-2\mu\eta\left(1-\frac{L\eta}{2}\right)\right][\mathcal{L}(\mathbf{w}_t)-\mathcal{L}_*] + \frac{L\eta^2\sigma^2}{2}.$$

Convergence Under Strong Convexity & Smoothness

Theorem

Let $\mathcal{L}(\cdot)$ be μ-strongly convex and L-smooth with bounded stochastic gradient variance, then the sequence $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ generated by SGD with step size η satisifies

$$\mathbb{E}\left[\mathcal{L}(\mathbf{w}_{t+1})\middle|\mathbf{w}_t\right] - \mathcal{L}_* \leq \left[1 - 2\mu\eta\left(1 - \frac{L\eta}{2}\right)\right][\mathcal{L}(\mathbf{w}_t) - \mathcal{L}_*] + \frac{L\eta^2\sigma^2}{2}.$$

- Optimal value can not be achieved under fixed step size!
- Diminishing η can reduce the bad term but sacrifies the rate.

Convergence Under Strong Convexity & Smoothness

Theorem

Let $\mathcal{L}(\cdot)$ be μ-strongly convex and L-smooth with bounded stochastic gradient variance, and let η_t be chosen such that

$$\eta_t = \frac{\beta}{c+t}, \text{ for some } \beta > \frac{1}{\mu}, \ c > 0, \text{ such that } \eta_0 \leq \frac{1}{L},$$

then the sequence $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ generated by SGD satisifies

$$\mathbb{E}\left[\mathcal{L}(\mathbf{w}_{T+1})\right] - \mathcal{L}_* \leq \frac{1}{c+T} \max\left\{\frac{\beta^2\sigma^2 L}{2(\beta\mu-1)}, (c+1)(\mathcal{L}(\mathbf{w}_0) - \mathcal{L}_*)\right\}.$$

Shortages of SGD

Key Inequality:

$$\mathbb{E}\left[\mathcal{L}(\mathbf{w}_{t+1})\middle|\mathbf{w}_t\right] \leq \mathcal{L}(\mathbf{w}_t) - \eta\left(1-\frac{L\eta}{2}\right)\|\nabla_{\mathbf{w}_t}\mathcal{L}(\mathbf{w}_t)\|^2 + \frac{L\eta^2\sigma^2}{2}.$$

- Diminishing η leads to lower (sub-linear) convergence rate.
- Constant η
  - SGD iteration: $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\nabla_{\mathbf{w}_t}\ell(\mathbf{w}_t;\mathbf{z}_{i^t})$,
  - Sanity check: assume $\mathbf{w}_t \to \mathbf{w}_*$ (not granted), then $\eta\nabla_{\mathbf{w}_t}\ell(\mathbf{w}_t;\mathbf{z}_{i_t}) \to 0$,
  - $\mathbf{w}_*$ cannot be staionary: $\nabla_{\mathbf{w}_*}\ell(\mathbf{w}_*;\mathbf{z}_{i_t}) \neq 0$ for any i.
- Solution: correct the stochatic gradient to kill σ^2.
- Basic idea: replace $\nabla_{\mathbf{w}_t}\ell(\mathbf{w}_t;\mathbf{z}_{i_t})$ by $\mathbf{g}_t$ such that $\mathbf{g}_t \to \nabla_{\mathbf{w}_t}\mathcal{L}(\mathbf{w}_t)$.

Stochastic Average Gradient (SAG)

In the t-the iteration, the true gradient can be written as

$$\mathcal{L}(\mathbf{w}_t) = \frac{1}{M}\left(\nabla_{\mathbf{w}_t}\ell(\mathbf{w}_t;\mathbf{z}_{i_t}) + \sum_{j\neq i_t}\nabla_{\mathbf{w}_t}\ell(\mathbf{w}_t;\mathbf{z}_j)\right),$$

Replace $\nabla_{\mathbf{w}_t}\ell(\mathbf{w}_t;\mathbf{z}_j)$ with its latest evaluation $\nabla_{\mathbf{w}_{t-d_j}}\ell(\mathbf{w}_{t-d_j};\mathbf{z}_j)$!

- $(t-d_j)$ is the latest iteration which the j-th sample is selected.

Implementation

1. Matain a gradient table $\mathbf{v}_{j,t}$ storing the latest evaluation of $\nabla_{\mathbf{w}_{t-d_j}}\ell(\mathbf{w}_{t-d_j};\mathbf{z}_j)$.
2. At the t-th iteration, update the table

$$\mathbf{v}_{i,t} = \begin{cases}\nabla_{\mathbf{w}_t}\ell(\mathbf{w}_t;\mathbf{z}_{i_t}), & \text{if } i = i_t,\\ \mathbf{v}_{i,t-1}, & \text{otherwise.}\end{cases}$$

3. Average over $\{\mathbf{v}_{i,t}\}$ (Recycle pervious computations).

SAG-Convergence Rate

SAG Iteration:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{\eta}{M}\sum_{i=1}^{M}\mathbf{v}_{i,t}.$$

$E(\|\vec{n}\|^2) = \|E(\vec{n})\|^2 + \|Var(\vec{n})\|^2$

Theorem

Let $\mathcal{L}(\cdot)$ be μ-strongly convex and each $\ell_i(\mathbf{w};\mathbf{z})$ be $L_{\max}$-smooth, then the sequence $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ generated by SAG with step size $\eta = \frac{1}{L_{\max}}$ satisifies

$$\mathbb{E}\left[\mathcal{L}(\mathbf{w}_t)\right] - \mathcal{L}_* \leq \left(1 - \min\left\{\frac{\mu}{L_{\max}}, \frac{1}{8M}\right\}\right)^t \times \left(\frac{3}{2}(\mathcal{L}(\mathbf{w}_0) - \mathcal{L}_*) + \frac{4L_{\max}}{M}\|\mathbf{w}_0 - \mathbf{w}_*\|^2\right).$$

SAG-Convergence Rate

SAG Iteration:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{\eta}{M}\sum_{i=1}^{M}\mathbf{v}_{i,t}.$$

- Linear convergence rate: $\mathcal{O}\left(M + \frac{L_{\max}}{\mu}\log\frac{1}{\epsilon}\right)$,
- Comparable to GD with convergence rate of $\mathcal{O}\left(M + \frac{L_{\max}}{\mu}\log\frac{1}{\epsilon}\right)$,
- Which one is better? (Show $ML \leq L_{\max}$),
- The proof of SAG's convergence is complex, as $\frac{1}{M}\sum_{i=1}^{M}\mathbf{v}_{i,t}$ is biased.

  Please refer to the paper "A Stochastic Gradient Method with an Exponential Convergence Rate for Finite Training Sets".

SAGA–Control Variants

Basic Idea

- Target: Estimate $\mu = \mathbb{E}(X)$,
- Get some estimation $Y \approx X$,
- The mean of Y is ζ.
- Given (X_i, Y_i), an variance reduced estimation is

$$\tilde{X}_i = X_i - Y_i + \zeta,$$

then

$$\text{(Unbiased)} \qquad \mathbb{E}(\tilde{X}_i) = \mathbb{E}(X_i) = \mu,$$

$$\text{(Reduced Variance)} \quad \mathbb{V}(\tilde{X}_i) \leq \mathbb{E}(|X_i - Y_i|^2) \approx 0.$$

SAGA–Control Variants

Apply the idea of variance reduction to the gradient estimator,

$$\nabla_{\mathbf{w}_t}\mathcal{L}(\mathbf{w}_t) = \frac{1}{M}\sum_{i=1}^{M}\left(\nabla_{\mathbf{w}_t}\ell_i(\mathbf{w}_t) - \mathbf{v}_{i,t-1} + \mathbf{v}_{i,t-1}\right),$$

$$= \frac{1}{M}\sum_{i=1}^{M}\left(\nabla_{\mathbf{w}_t}\ell_i(\mathbf{w}_t) - \mathbf{v}_{i,t-1} + \bar{\mathbf{v}}_{t-1}\right).$$

$\vec{V}_{t-1} = \frac{1}{M}\sum_{i=1}^{M}\vec{V}_{i,t-1}$

Let the stochastic estimation be

$$\mathbf{g}_t = \nabla_{\mathbf{w}_t}\ell_{i_t}(\mathbf{w}_t) - (\mathbf{v}_{i_t,t-1} - \bar{\mathbf{v}}_{t-1}).$$

- i_t is the selected sample,
- $\mathbf{g}_t$ is unbiased,
- Select i_t such that $\mathbf{v}_{i_t,t-1}$ approaches $\nabla_{\mathbf{w}_t}\ell_{i_t}(\mathbf{w}_t)$.

SAGA–Implementation

1. Matain a gradient table $\mathbf{v}_{j,t}$ storing the latest evaluation of $\nabla_{\mathbf{w}_{t-d_j}} \ell(\mathbf{w}_{t-d_j}; \mathbf{z}_j)$.
2. At the t-th iteration, update the table

$$\mathbf{v}_{i,t} = \begin{cases} \nabla_{\mathbf{w}_t} \ell(\mathbf{w}_t; \mathbf{z}_{i_t}), & \text{if } i = i_t, \\ \mathbf{v}_{i,t-1}, & \text{otherwise.} \end{cases}$$

3. SAGA gradient estimator:

$$\mathbf{g}_t = \nabla_{\mathbf{w}_t} \ell_{i_t}(\mathbf{w}_t) - \mathbf{v}_{i_t,t-1} + \frac{1}{M} \sum_{i=1}^{M} \mathbf{v}_{i,t-1}.$$

4. Update: $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{g}_t$.

SAGA is very similar to SAG.

- $\eta = \mathcal{O}(1/L_{\max})$, linear convergence rate: $\mathcal{O}\left(M + \frac{L_{\max}}{\mu} \log \frac{1}{\epsilon}\right)$,
- $\mathbf{g}_t$ is unbiased simplifies the proof.

SVRG

Drawback of SAG and SAGA: Table maintainance cost $\mathcal{O}(MN)$.
How to reduce the memory requirement without sacrificing the rate?
The idea of SVRG: Align the reference points of the $\mathbf{v}_i$'s.
For every k iterations, do

- Store $\bar{\mathbf{w}} = \mathbf{w}_t$,
- Compute true gradient $\bar{\mathbf{v}} = \nabla_{\mathbf{w}_t} \mathcal{L}(\mathbf{w}_t)$.

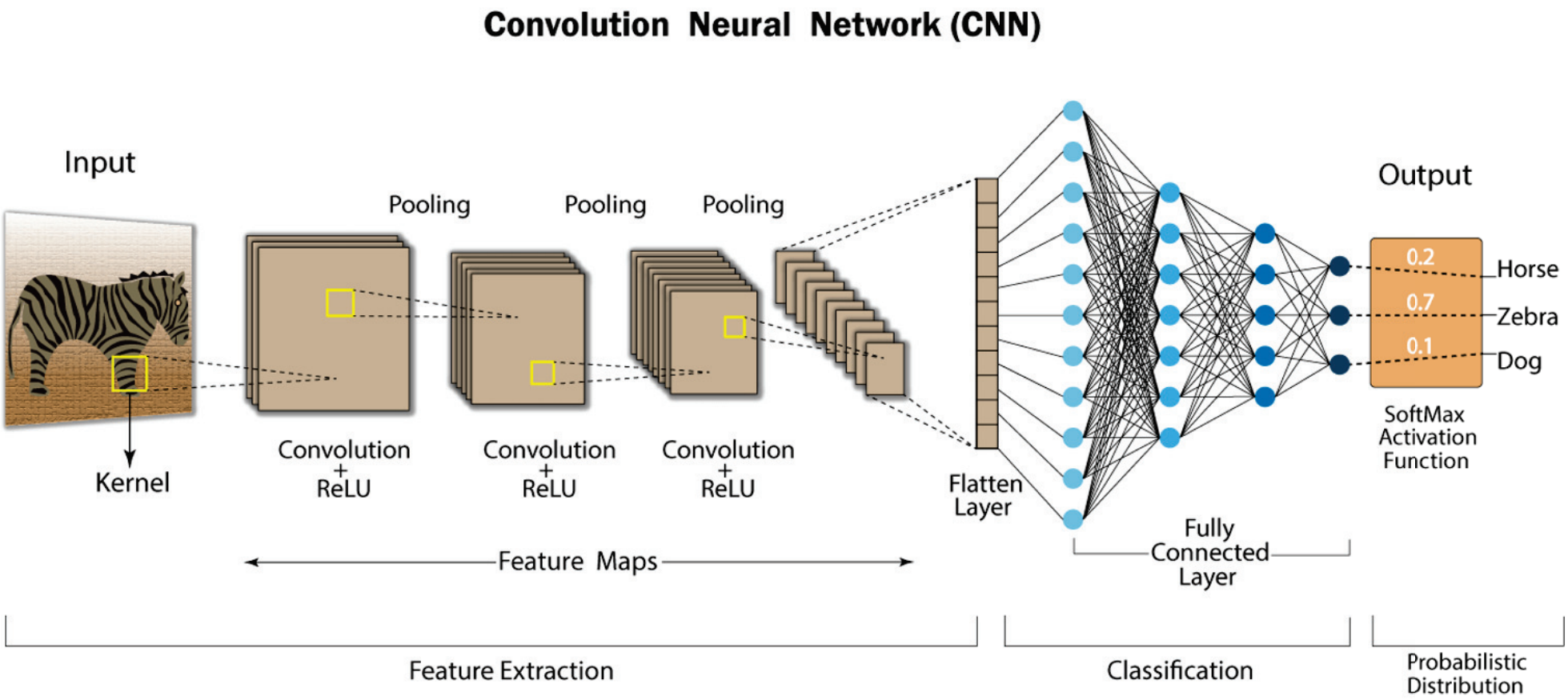
SVRG gradient estimator:

$$\mathbf{g}_t = \nabla_{\mathbf{w}_t} \ell_{i_t}(\mathbf{w}_t) - \nabla_{\mathbf{w}_t} \ell_{i_t}(\bar{\mathbf{w}}_t) + \bar{\mathbf{v}}.$$

- Linear convergence rate: $\mathcal{O}\left(M + \frac{L_{\max}}{\mu} \log \frac{1}{\epsilon}\right)$,
- Memory requirement: $\mathcal{O}(d)$,
- GD computation once in a while.

Complex Learning Models

An Example: A CNN Model

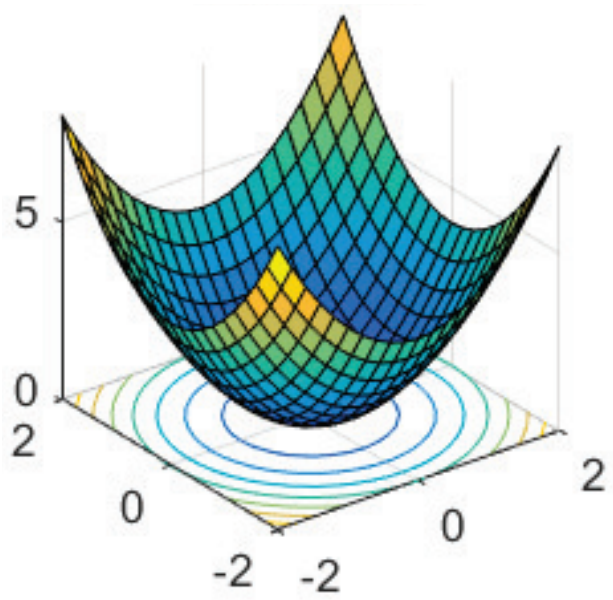


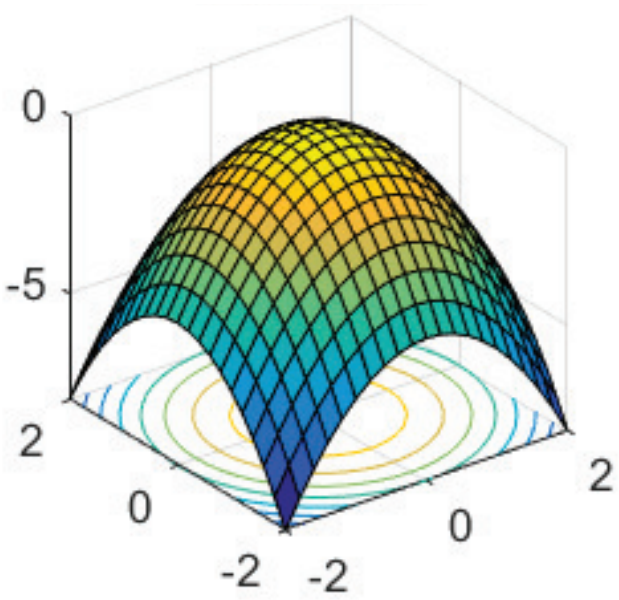
- High Dimension Model Vector $\mathbf{w}$,
- Non-Convex: Large Number of Stationary Points $\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}) = 0$.

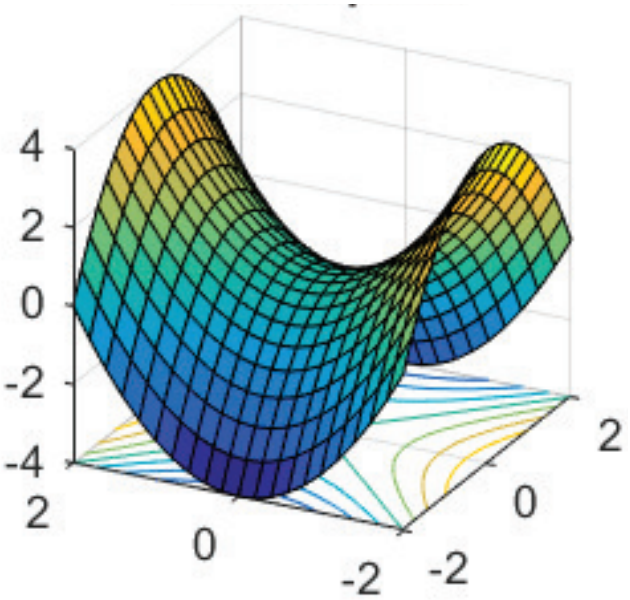
Stationary Points

$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = 0$:

- Local Minimum,
- Local Maximum,
- Saddle Points.



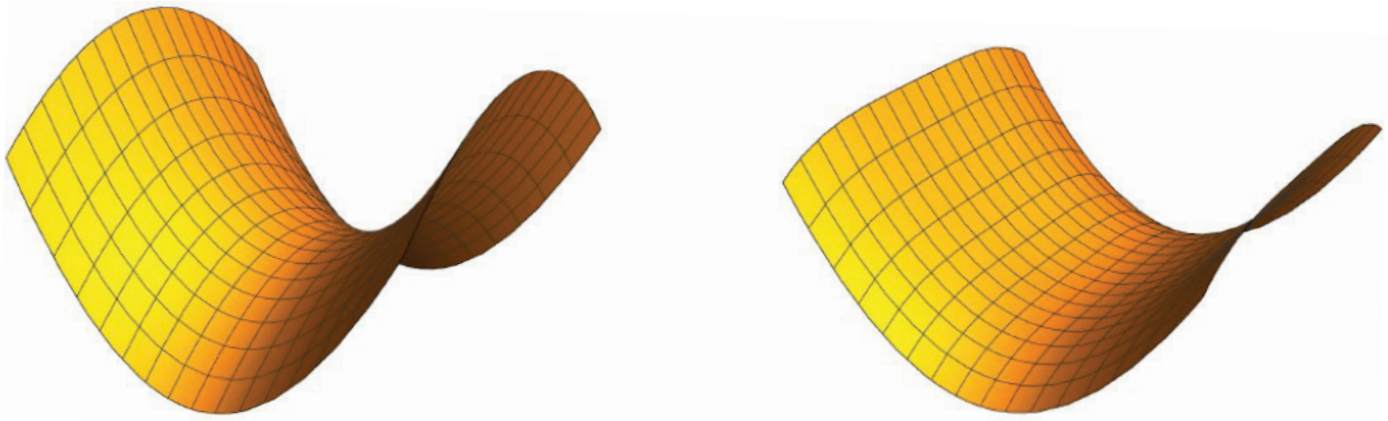




Stationary Points

$\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}) = 0$:

- Local Minimum,
  - Hessain matrix is positive-definite.
- Local Maximum,
  - Hessain matrix is negative-definite.
- Saddle Points.
  - **Strict Saddle Point**: Eigenvales of Hessian matrix should either be positive and negative.
  - **Non-strict Saddle Point**: Eigenvales of Hessian matrix can be positive, 0, and negative.



Convergence Under Strong Convexity & Smoothness

Theorem (Convergence under L-Smoothness and Bounded Variance)

Let $\mathcal{L}(\cdot)$ be L-smooth with bounded stochastic gradient variance, then the sequence $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ generated by SGD with step size $\eta = 1/L$ satisifies

$$\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla_{\mathbf{w}_t}\mathcal{L}(\mathbf{w}_t)\|^2 \leq \frac{2L}{T+1}\left\{\mathbb{E}\left[\mathcal{L}(\mathbf{w}_{t+1})\middle|\mathbf{w}_t\right] - \mathcal{L}_*\right\} + \sigma^2.$$

It's unknown where the loss function converge.

Question: How to escape from the (strict) saddle points?

SGD: Escape from Strict Saddle Points

Gradient Descent: Trapped in any point with $\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}) = 0$.
Properties of Functions with Strict Saddle Points:

- Large gradient norm (far from stationary points),
- Small gradient norm with non-negative Hessian matrix (near a local optimum),
- Small gradient norm with negative eigenvalues of Hessian matrix (near a strict saddle point).

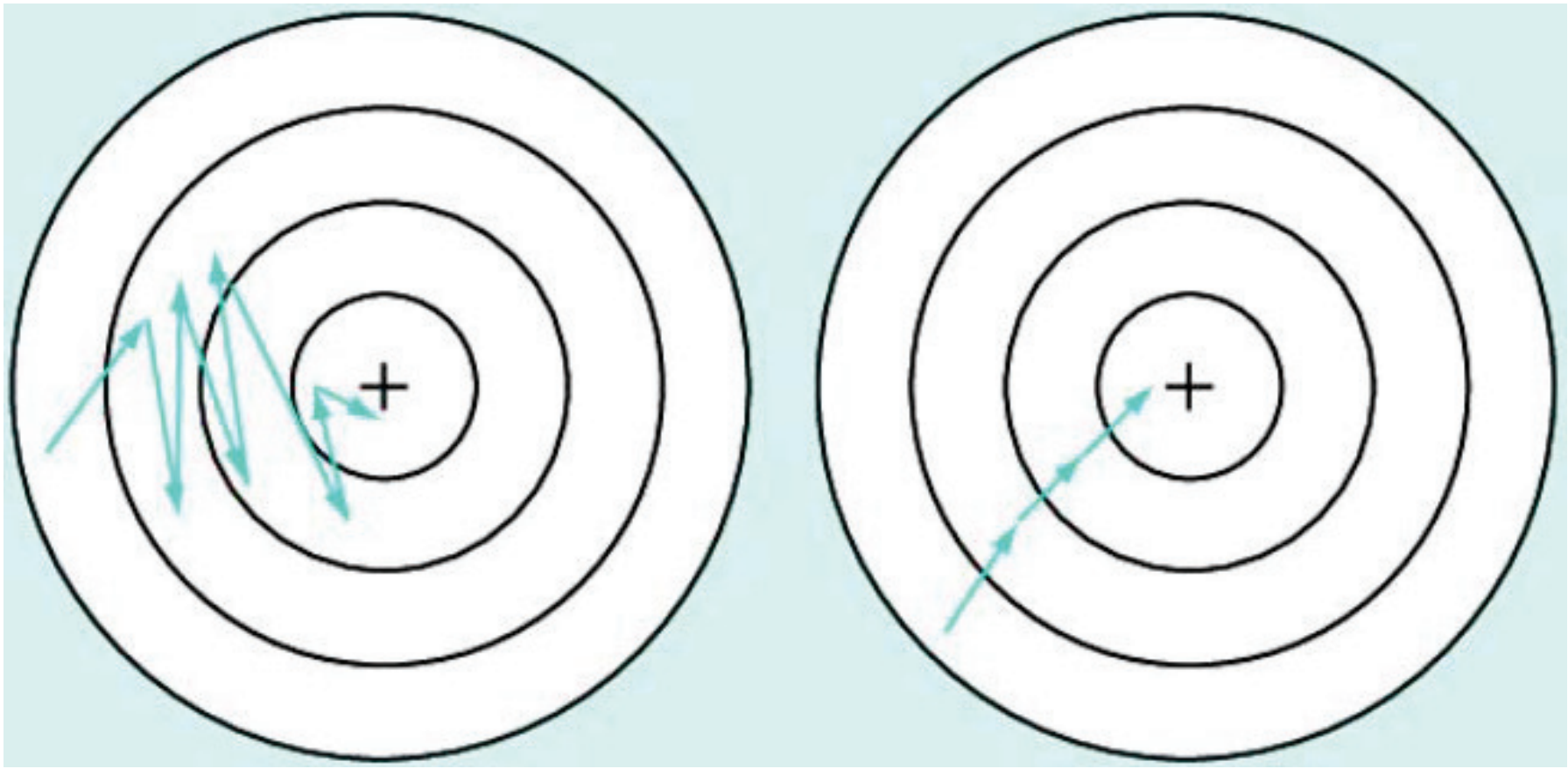
Why SGD can escape from strict saddle points?

$$\hat{\mathbf{g}} = \mathbf{g} + \boldsymbol{\delta},$$

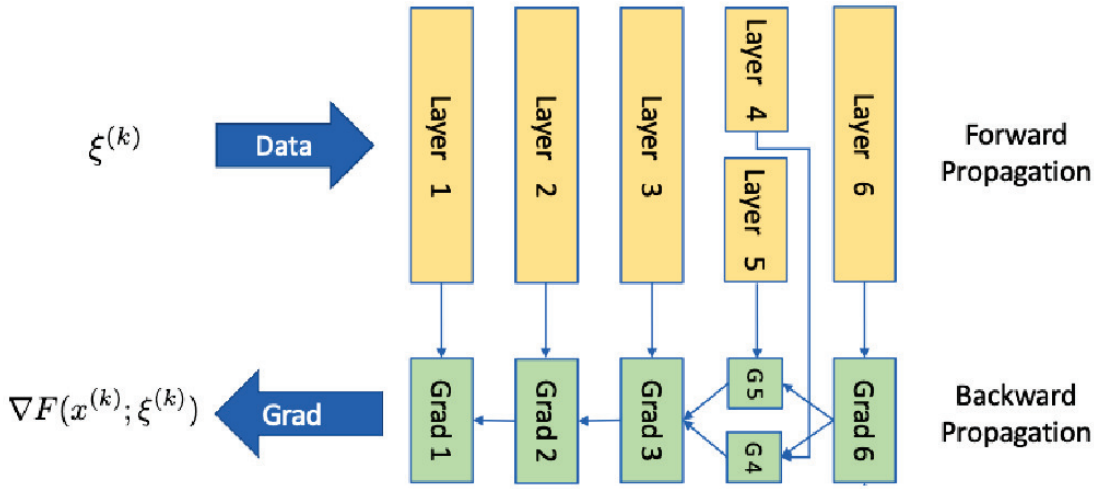
where $\boldsymbol{\delta}$ is a Gassuain distortion.

SGD: Escape from Strict Saddle Points

SGD v.s. GD



Implementing SGD in Deep Neural Networks



- Stochastic gradient can be calculated via forward-backward propagation.
- Stochastic gradient can be achieved automatically via Pytorch/Tensorflow.
- DNN training typically utilizes GPUs.

GD v.s. SGD for Training

	GD	**SGD**
Single-Interation Training	All samples	One sample
Comp. Efficiency	Slow	Faster
Memory Cost	High	Low
Convergece Latnecy	Low	High
Scalabilty	Low	High
Convex Functions	Optimal	Deviation from Optimum
Saddle Points	Trapped	Esacape
Local Optimum	Trapped	Can escape from shallow one

Thank you!

wendzh@shanghaitech.edu.cn