北京大学
PEKING UNIVERSITY

Towards Understanding Deep Learning:

Two Theories of Stochastic Gradient Langevin Dynamics

王立威

北京大学 信息科学技术学院

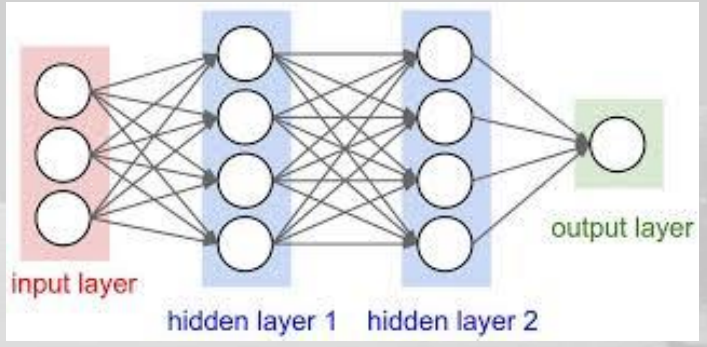
Joint work with: 牟文龙 翟曦雨 郑凯

Content

1. Mystery Behind the Success of Deep Learning
2. Stochastic Gradient Descent Method
3. Stochastic Gradient Langevin Dynamics
4. Our Results
5. Conclusions

Deep Learning Method

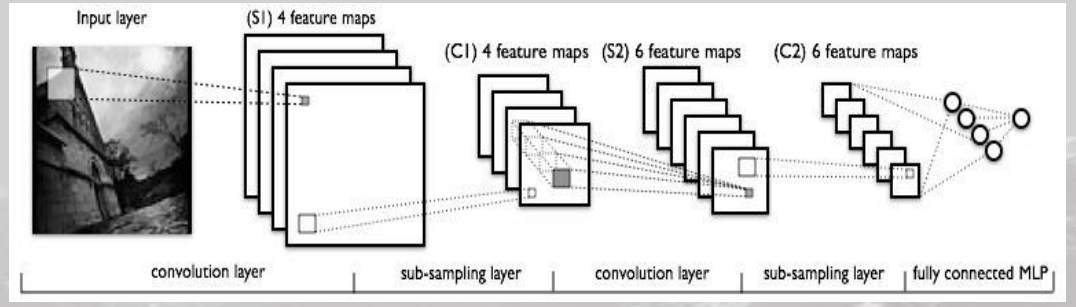
Main idea: using layer-wise structures to represent hypothesis set



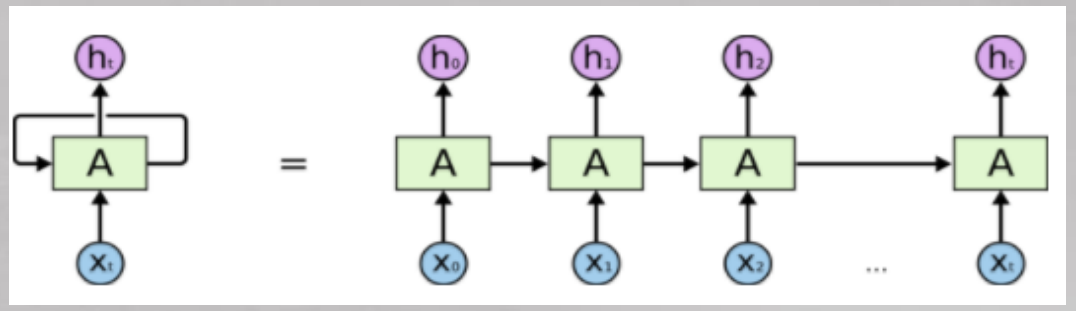
Spatial

Time

CNN



RNN



composite function

$$f(x) = \sigma_L(w_L\sigma_{L-1}(\cdots\sigma_1(w_1x)))$$

Mystery of Deep Learning

Y. LeCun, CVPR2015, invited talk: **What's wrong with deep learning**

One important piece: **missing some theory**!

Such as:

- Faster training algorithms?
- Better net structures?
- ……

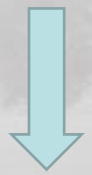
Fundamental problem:

Generalization ability of DNN

Supervised Learning

Collect data: $\{(x_i, y_i) | \ i = 1, 2, \dots, n\}$

Learn a model: $f: \mathcal{X} \rightarrow \mathcal{Y}, \ f \in \mathcal{H}$



Predict new data: $x \rightarrow f(x)$

All $(x, y) \sim \mathcal{D}$, where $\mathcal{D}$ is unknown

A common approach to learn:

ERM (Empirical Risk Minimization)

$$\min R_n(w) := \frac{1}{n} \sum_i l(w; x_i, y_i)$$

w: model parameters
$l(w; x, y)$: loss function w.r.t. data

Population Risk: $R(w) := \mathrm{E}[l(w; x, y)]$

Three Components of Deep Learning

- Model (Architecture)
  - CNN for images, RNN for speech…
  - Shallow (but wide) networks are universal approximator (Cybenko, 1989)
  - Deep (and thin) ReLU networks are universal approximator (LPWH**W**, 2017)
- Optimization on Training Data
  - Learning by optimizing the empirical loss, nonconvex optimization
- Generalization to Test Data
  - Generalization theory

Generalization: the Core of Learning Theory

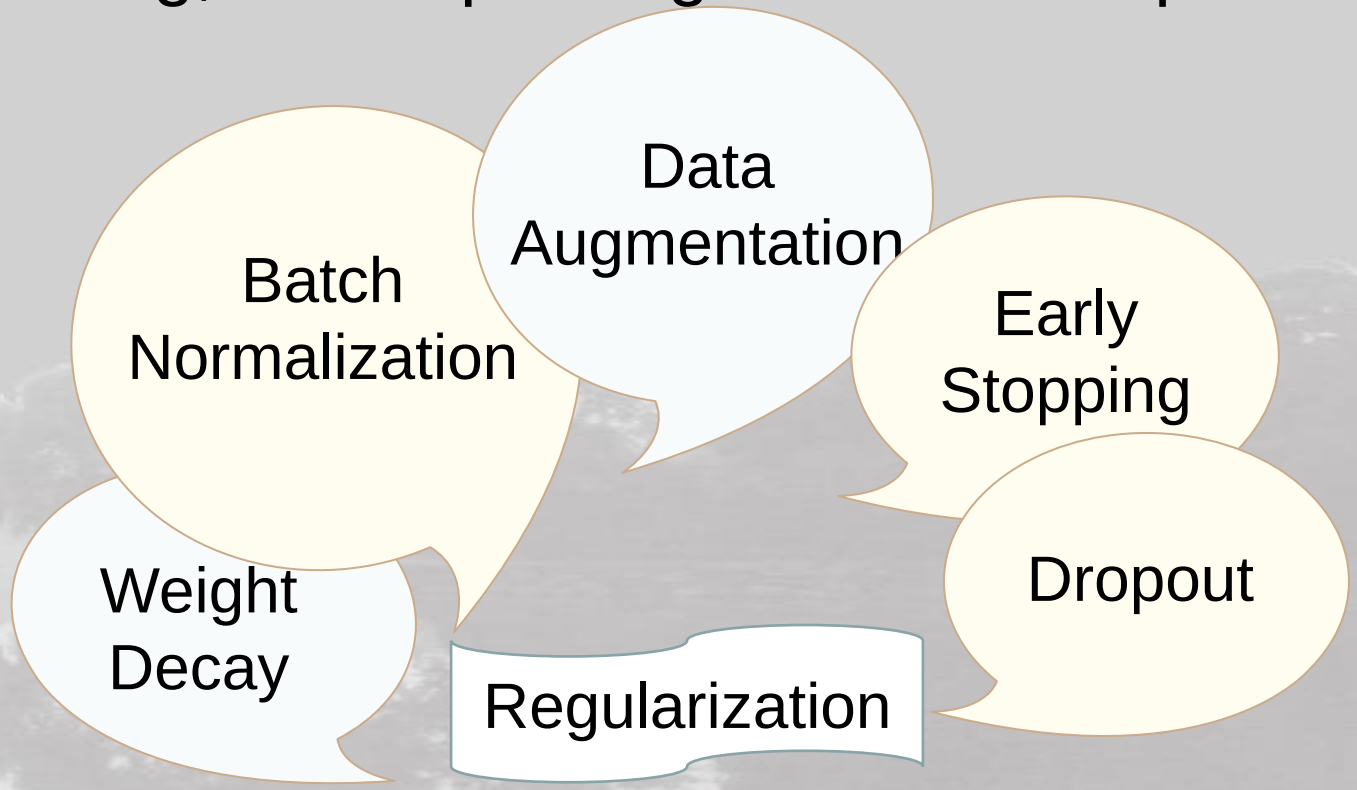
To train a model, we only have training data.

Model should fit unknown data well, not just training data

To Learn, Not To Memory

Regularization Techniques

To avoid overfitting, and improve generalization performance



Some Observations of Deep Nets

- # of parameters >> # of data, hence easy to fit data
- Without regularization, deep nets also have benign generalization
- For random label or random feature, deep nets converge with 0 training error but without any generalization

How to explain these phenomena?

ICLR 2017 Best Paper:

"Understanding deep learning requires rethinking generalization"

Traditional Learning Theory Fails

Common form of generalization bound (in expectation or high probability)

$$R(w) \leq R_n(w) + \sqrt{\frac{Capacity\ Measure}{n}}$$

Capacity Measurement	Complexity
VC-dimension	$VC \leq O(\lvert E\rvert \log \lvert E\rvert)$
$\mathcal{E}$-Covering number	$\log_2 N_{l_1}(\mathcal{F}, \epsilon, m) \leq O\left(\frac{(AL_\phi)^{L(L+1)}}{\epsilon^{2L}}\right)$
Rademacher Average	$R_m(\mathcal{F}) \leq O(\mu^L)$

$|E|$: # of edges

L: # of layers

All these measurements are far beyond the number of data points!

Recent Advances Using the Idea of Margin

Bartlett et al. (NIPS17):

Main idea	Normalize Lipschitz constant (product of spectral norms of weighted matrices) by margin
Final bound	$\Pr\left[\arg\max_j F_{\mathcal{A}}(x)_j \neq y\right] \leq \widehat{\mathcal{R}}_\gamma(F_{\mathcal{A}}) + \widetilde{\mathcal{O}}\left(\frac{\|X\|_2 R_{\mathcal{A}}}{\gamma n}\ln(n)\ln(W) + \sqrt{\frac{\ln(1/\delta)}{n}}\right)$

where $R_{\mathcal{A}}$ is the spectral complexity

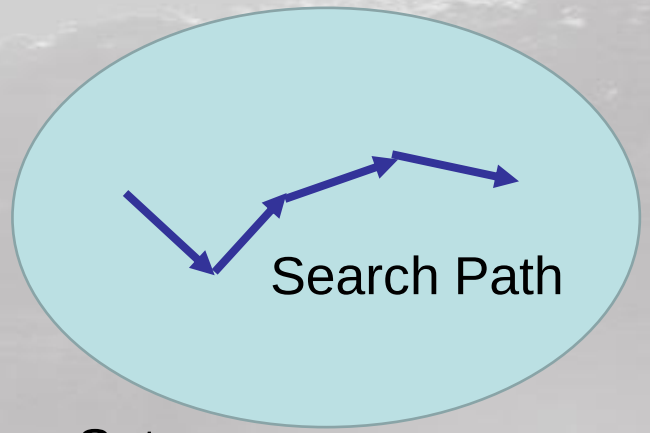
Remark:

(1) nearly has no dependence on # of parameters

(2) a multiclass bound, with no explicit dependence on # of classes

Content

- Traditional generalization bounds mainly focus on the complexity of the whole function set represented by deep nets.
- Another pivotal reason to the success of DL is the training algorithm we use.
- A training algorithm may only search a small subspace of the function set.



Training Algorithms for Deep Learning

Commonly Used Algorithms

Non-adaptive

- SGD
- SGD with momentum
- ……

Easy to implement

Adaptive

- Adam
- AdaGrad
- AdaDelta
- ……

Automatically tune parameters

Learn the curvature adaptively

Stochastic Gradient Descent

Objective loss function:

$$\min R_n(w) := \frac{1}{n}\sum_i l(w; x_i, y_i)$$

where $(x_i,\ y_i)$ is the data, w is the parameter vector.

Gradient Descent: $w_{t+1} = w_t - \frac{\eta}{n}\sum \nabla l(w_t; x_i, y_i)$

O(n) time complexity

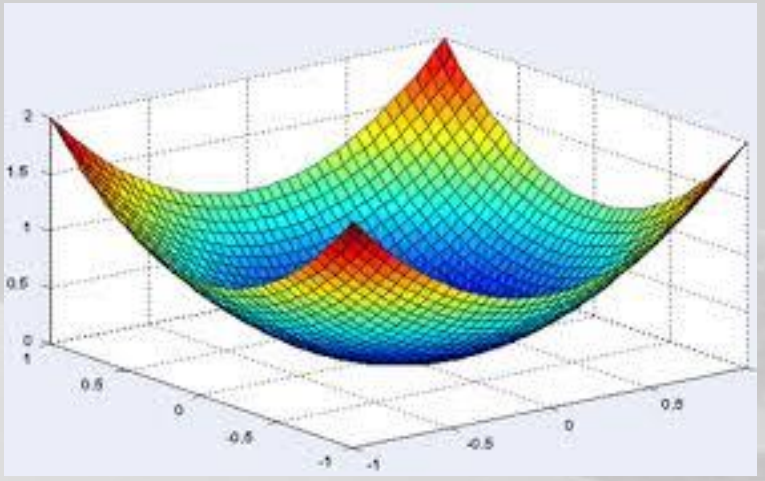
SGD: $w_{t+1} = w_t - \eta \nabla l(w_t; x_{i_t}, y_{i_t})$, where i_t is uniform in {1, …, n}

O(1) time complexity

Extensions: mini-batch SGD

The Generalization Ability Induced by SGD

If we are in convex case, things are easier



From classical stochastic optimization:

SGM converges in $O(\frac{1}{\sqrt{T}})$

but requires one-pass over training data, unrealistic

From uniform stability, which implies generalization

SGM converges in $O(\frac{\sum_t \alpha_t}{n})$

α_t is the step size of SGM.

This result applies to multi-pass SGM.

The Generalization Ability Induced by SGD

In nonconvex case, there are some results, but very weak

Hardt et al. (ICML15) , for SGM:

Assuming Lipschitz and β-smooth, then

$$\varepsilon_{stab} \leq O\left(\frac{T^{1-\frac{1}{\beta c+1}}}{n}\right)$$

which maybe linear dependent on training iterations

Content

Stochastic Gradient Langevin Dynamics (SGLD)

SGLD is a variant of SGD.

$$w_{t+1} = w_t - \eta \nabla l\left(w_t; x_{i_t}, y_{i_t}\right) + \sqrt{\frac{2\eta}{\beta}} z_t, \text{ where } z_t \sim \mathcal{N}(0, I_d)$$

Injection of Gaussian noise makes SGLD completely different with SGD

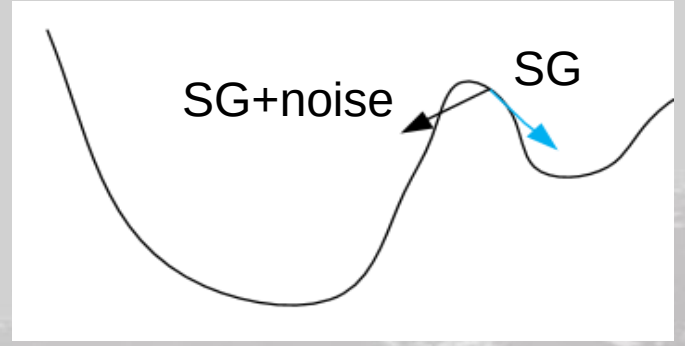
For small enough step size η_t, Gaussian noise will dominate the stochastic gradient.

Distinctions of SGLD

Intuitively, injected Gaussian noise helps escape saddle points or local minimum

SGLD is the discretization of following SDE

$$dW(t) = -\nabla F\big(W(t)\big)dt + \sqrt{\frac{2}{\beta}}\,dB(t)$$



where $F(\cdot)$ is the empirical loss function, $B(t)$ is the standard Brownian motion

Its distribution converges to Gibbs distribution $\propto \exp(-\beta F(w))$

Large β will concentrate on the global minimizer of $F(w)$

Asymptotic VS. Non-asymptotic for SGLD

Asymptotic result requires to run exponential time.

In practice, we can only run the algorithm for poly or even log time.

Thus, we should focus on ***non-asymptotic*** properties:

1. Finite time approximation
2. Finite discretization

Non-asymptotic results can also guide us to improve existing algorithms.

Raginsky et al. (COLT17) decomposed the generalization error as:

$$E[F(w_k)] - F^*$$

$$= \boxed{(E[F(w_k)] - E[F(\bar{w})])} + \boxed{(E[F(\bar{w})] - F^*)}$$

$\bar{w}$: output of Gibbs

F^*: global optimal.

discretization error of SGLD:

$O(\epsilon * poly(\beta, d, \frac{1}{\gamma}))$

for any $\epsilon > 0$, γ is spectral gap

generalization error of Gibbs algorithm

$O(\frac{(\beta+d)^2}{\gamma n} + \frac{d \log \beta}{\beta})$

However, $\frac{1}{\gamma}$ may be ***exponential in d***, unacceptable!

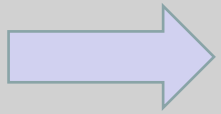
Content

Two Classical Learning Theories for Generalization Error

- Algorithmic Stability theory

Stable Algorithm

$$sup_z |E[l(w_S, z)] - E[l(w_{S'}, z)]| \leq \epsilon$$



Generalization Performance

$$E[l(w_S, z)] - E_S[l(w_S, z)] \leq \epsilon$$

- PAC-Bayesian theory

Generalization Performance

$$E[E[l(w, z)]] - E_S[E[l(w, z)]] \leq O\left(\sqrt{\frac{KL(Q||P)}{n}}\right)$$

Q is the distribution of w
P is the prior of w
Inner expectation is over Q

Our Results

From the view of stability theory:

Under mild conditions of (surrogate) loss function, the generalization error of SGLD at N-th round satisfies

$$E[l(w_S, z)] - E_S[l(w_S, z)] \leq O\left(\frac{1}{n}\left(k_0 + L\sqrt{\beta \sum_{k=k_0+1}^{N} \eta_k}\right)\right)$$

where L is the Lipschitz constant, and $k_0 := \min\{k: \eta_k \beta L^2 < 1\}$

If consider high probability form, there is an additional $\tilde{O}\left(\sqrt{1/n}\right)$ term

Our Results

From the view of PAC-Bayesian theory:

For regularized ERM with $R(w) = \lambda||w||^2/2$. Under mild conditions, with high probability, the generalization error of SGLD at *N-th* round satisfies

$$E[E[l(w_S, z)]] - E_S[E[l(w_S, z)]] \leq O\left(\sqrt{\frac{\beta}{n}\sum_{k=1}^{N}\eta_k e^{-\lambda(T_N - T_k)/2} E[||g_k||^2]}\right)$$

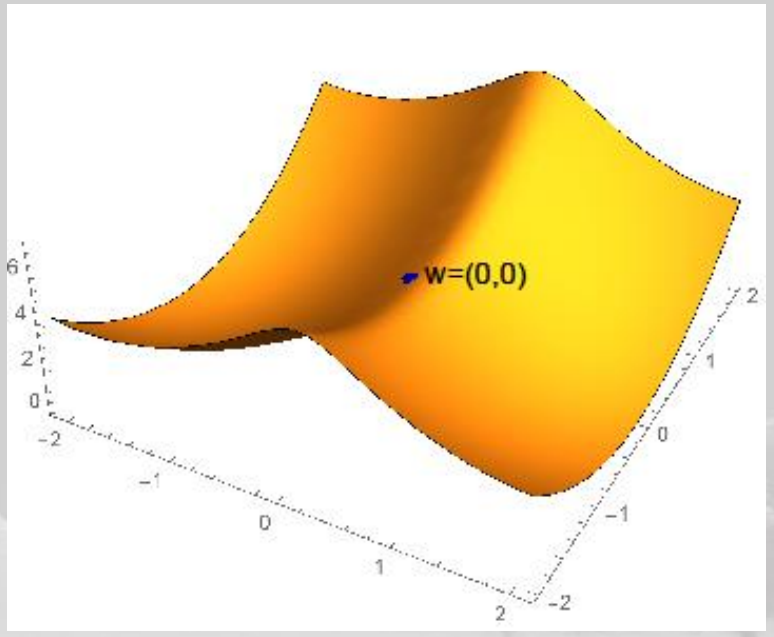
where $T_k = \sum_{j=1}^{k}\eta_k$, g_k is the stochastic gradient in each round.

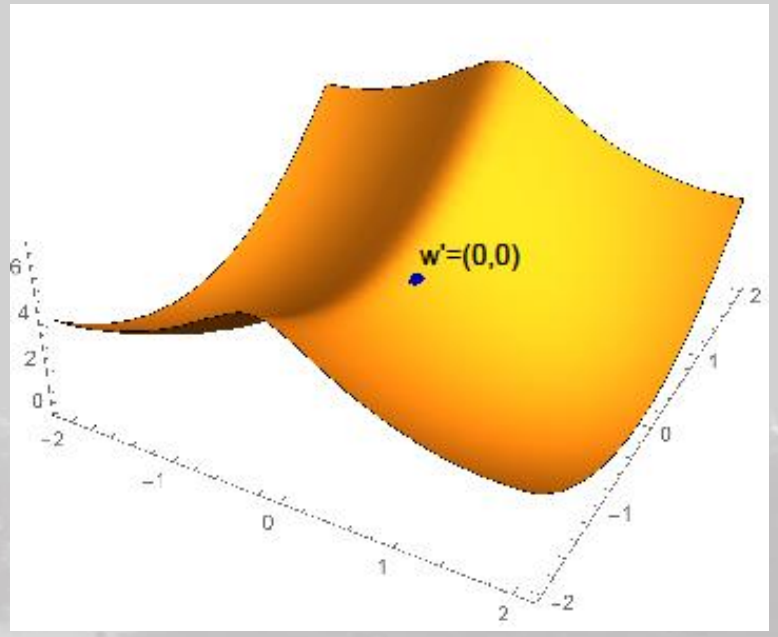
Comparison Two Results

Both bounds suggest “train faster, generalize better”, which explain the random label experiments in ICLR17

- In expectation, stability bound has a faster $O\left(\frac{1}{n}\right)$ rate.
- PAC-Bayes bound is data dependent, and doesn’t rely on Lipschitz condition.
- Effect of step sizes in PAC-Bayes exponentially decay with time.

Motivation





Iteration point w (w') w.r.t. data S (S')

For nonconvex loss

$E[||w - w'||]$ may become large

But their f-divergence doesn't vary a lot, as Gaussian noise helps smooth the distribution

Main Ideas

Suppose $F_z(W)$ is the loss function, $\mathrm{E}[g_k] = \nabla F_z(W_k)$, $\xi_k \sim \mathcal{N}(0, I_d)$

SGLD: $W_{k+1} = W_k - \eta g_k + \sqrt{\frac{2\eta}{\beta}}\xi_k$

Interpolation

Langevin: $dW(t) = -\nabla F_z\big(W(t)\big)dt + \sqrt{\frac{2}{\beta}}dB(t)$

Consider an equivalent process with such evolution w.r.t. density function π

Fokker-Planck: $\frac{\partial \pi}{\partial t} = \nabla \cdot \left(\frac{1}{\beta}\nabla \pi + \pi \nabla F_z\right)$

Techniques for Stability Bound

For neighboring datasets S, S', consider SGLD at each iteration:

With probability $1 - 1/n$

choose the same data, as using same gradient mapping, Hellinger distance does not increase

With probability $1/n$

choose different data, as probability of this case is small, hence the increase of Hellinger distance is limited

Construct appropriate PDEs → Control the change of Hellinger between consecutive steps → Recursively obtain final result

Techniques for PAC-Bayesian Bound

Different with stability, there is a $||w||_2$ term in the prior, which is dimension-dependent

Add regularization term $\rightarrow$ Cancel $||w||_2$ dependent term in generalization

To cancel them perfectly, allow prior to vary with iterations in a data-independent way

Construct appropriate PDEs

Recursively control the change

Content

Conclusions

Deep learning has received huge success ***empirically***.

But many fundamental problems remain unsolved:

- Generalization
- Loss surface
- Faster and automatic training algorithms

Only when we have some understanding, then it is possible to design better architectures.

Thanks!

W. Mou, L. Wang, X. Zhai, K. Zheng, Generalization Bounds of SGLD for Non-convex Learning: Two Theoretical Viewpoints, arXiv:1707.05947