HPC Class
A Bayesian Perspective on Generalization and Stochastic Gradient Descent

Smith, Samuel L., and Quoc V. Le. "A bayesian perspective on generalization and stochastic gradient descent." (2018)

In this paper…

To explain “SGD” and “Generalization” in terms of Baysian,

this paper consists of 2 part

1, What decide ‘Good’ model?

2, To make ’Good’ model, what we should do?

ABSTRACT

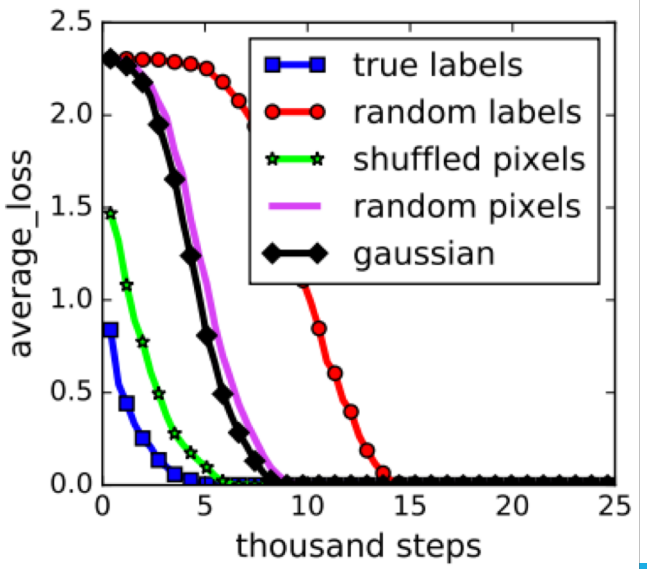
We consider two questions at the heart of machine learning; how can we predict if a minimum will generalize to the test set, and why does stochastic gradient descent find minima that generalize well? Our work responds to Zhang et al. (2016), who showed deep neural networks can easily memorize randomly labeled

Understanding deep learning requires rethinking generalization

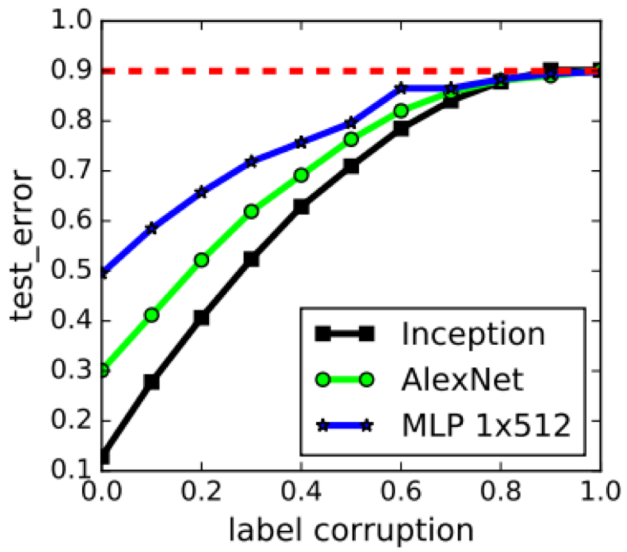
This paper compared two situations.

1) Pure-train data and Pure-test data
2) Randomized-train(randomized labels or pixels) and Pure-test data

1 and 2 can memorize whole train-data(train loss=0) but (2) cannot generalize well



(a) learning curves



(c) generalization error growth

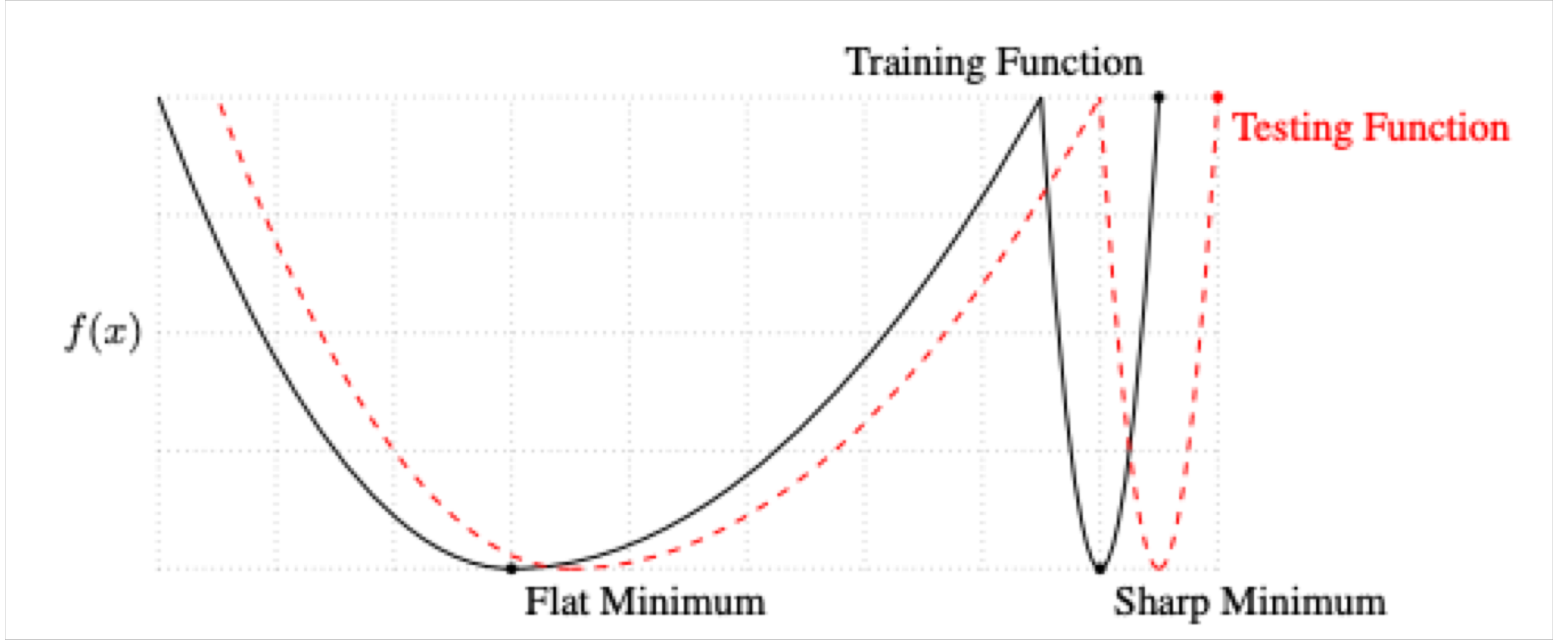
-> Deep convolutional networks can memorize whole (train)data???

Source: [Zhang et al.(2016)]

On large-batch training for deep learning: Generalization gap and sharp minima.

They found that if we hold fixed learning rate(lr) and increase the batch-size(bs).
=> The “test” accuracy falls down.

Large Batch tends go to sharp minimum and Small Batch go to Flat ?



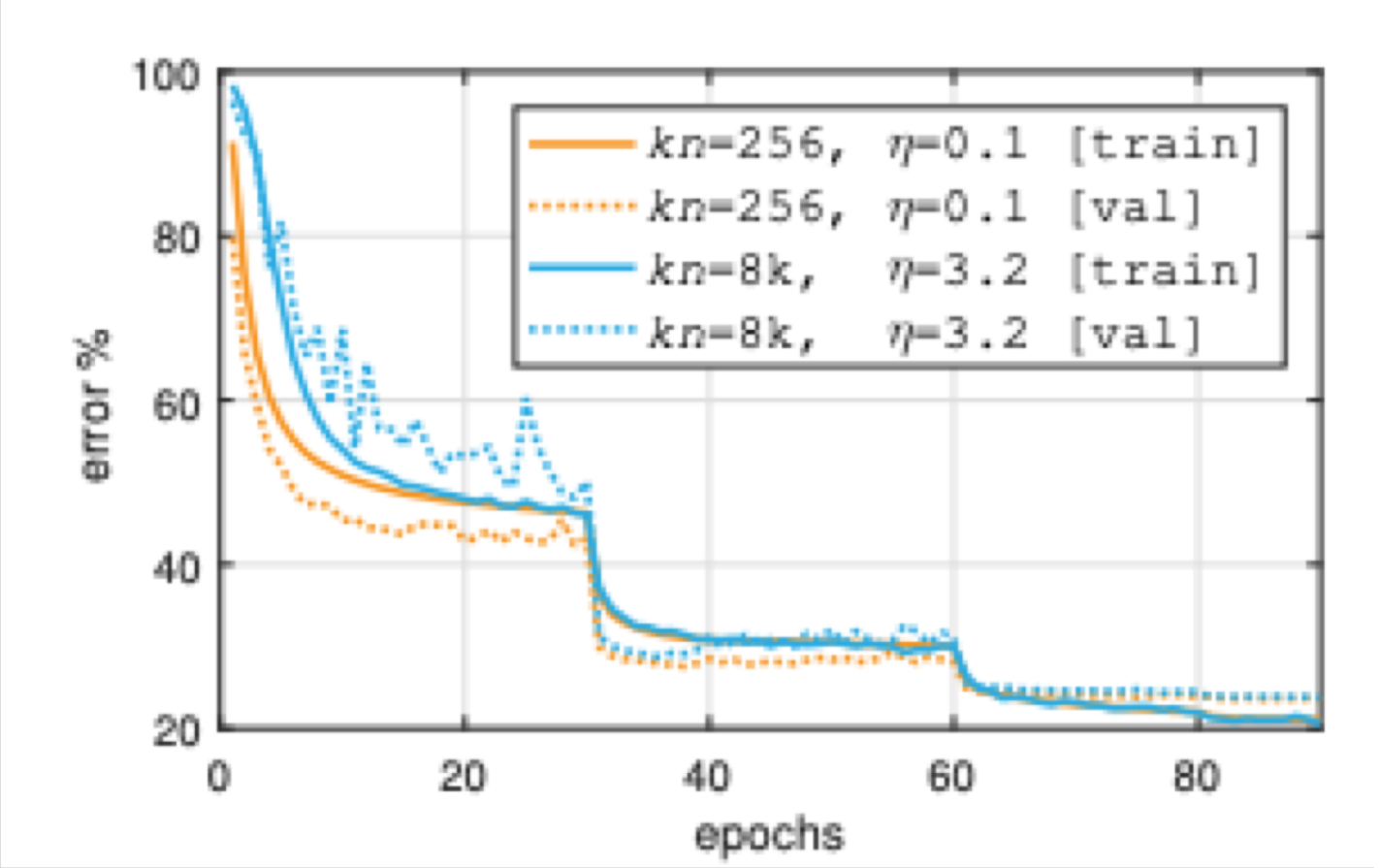
Flat Minimum vs Sharp minumu (Source: [Keskar et al.])

Accurate, Large Minibatch SGD: Training ImageNet in 1Hour

They sets a linear scaling rule between learning rate and batch size.

$$\frac{lr}{BS} = Constant$$

$$\left(\frac{0.1}{256} = \frac{3.2}{8\text{K}}\right)$$



What is Sharp minima? Flat minima?

There are many papers what is the relationship between small and large batch, flat minima and sharp minima.

- Entropy-SGD: Biasing gradient descent into wide valleys.
- Sharp minima can generalize for deep nets.
- …

In this paper…

This paper show two:

1) The result of [Zhang et al(2016)] is not unique to deep learning,

this paper demonstrate that this phenomenon is straightforwardly understood by evaluating the Bayesian evidence in favor of each model.

2) SGD integrates a stochastic differential equation whose “noise scale” $g \approx \varepsilon N/B$, where ε is the learning rate, N training set size and B batch size.

BAYESIAN MODEL COMPARISON

Classification model M with a single parameter ω, training inputs x and training labels y.

$$P(\omega|y,x;M) = \frac{\mathrm{P(y|\omega,x;M)P(\omega;M)}}{\mathrm{P(y|x;M)}}$$

$$P(y|\omega,x;M) = \prod_i P(y_i|\omega_i,x_i;M) = e^{-H(\omega;M)}$$

- $H(\omega;\mathrm{M}) = -\sum_i In(P(y_i|\omega,x_i;M)$

$$\mathrm{P}(\omega;\mathrm{M}) = \sqrt{\frac{\lambda}{2\pi}}\mathrm{e}^{-\lambda\omega^2/2}$$ (=>set model with gaussian prior)

$$P(\omega|y,x;M) \propto \sqrt{\frac{\lambda}{2\pi}}e^{-C(\omega;M)}(where\ C(\omega;M) = H(\omega;\mathrm{M}) + \lambda\omega^2/2)$$

BAYESIAN MODEL COMPARISON

To predict an unknown label y_t of a new input x_t, we use Bayesian predictive distribution($P(y|x) = \int P(y|w)P(w|x)dw$)

$$\mathrm{P}(\mathrm{y_t}|\mathrm{x_t}, \mathrm{y}, \mathrm{x}; \mathrm{M}) = \int d\omega \; P(y_t|\omega, x_t; M)P(\omega|y, x; M)$$

$$= \{\int d\omega \; P(y_t|\omega, x_t; M)e^{-C(\omega;M)}\} / \{\int d\omega \; e^{-C(\omega;M)}\}$$

Since $\mathrm{P}(\mathrm{y_t}|\omega\,, x_t; \mathrm{M})$ is smooth near ω_0,

- (ω_0 is the point which minimize cost functions)

we can approximate $\mathrm{P}(\mathrm{y_t}|\mathrm{x_t}, \mathrm{x}, \mathrm{y}; \mathrm{M}) \approx \mathrm{P}(\mathrm{y_t}|\omega_0, \mathrm{x_t}; \mathrm{M})$

BAYESIAN MODEL COMPARISON

To compare two model, we set probability ratio.

$\frac{\mathrm{P}(M_1|y,x)}{\mathrm{P}(M_2|y,x)} = \frac{\mathrm{P}(y|x;M_1)}{\mathrm{P}(y|x;M_2)} \frac{\mathrm{P}(M_1)}{\mathrm{P}(M_2)}$ = (Evidence ratio)*(Prior ratio)

From this point, this paper set P(M)=1

Evidence ratio controls how much the training data changes our prior beliefs.

$$\mathrm{P}(\mathrm{y}|\mathrm{x};\mathrm{M}) = \int d\omega\, P(y|\omega,x;M)P(\omega;M) = \sqrt{\frac{\lambda}{2\pi}} \int d\omega\ e^{-C(\omega;M)}$$

$$= \exp\{-(C(\omega_0) + \frac{1}{2}\sum_{i=1}^{p} In(\lambda i\ /\lambda))\}$$

(p is the number of parameters and λ_{i} is the eigenvalues of Hessian)

MODEL COMPARISON BAYESIAN

They will compare the evidence against a null model which assumes the labels are entirely random. This unusual model has no parameters and so the evidence is $\mathrm{P(y|x;NULL)} = \left(\frac{1}{\mathrm{n}}\right)^{\mathrm{N}} = \mathrm{e}^{-\mathrm{N\ In(n)}}$

(n is the number of model classes, N is the number of train labels)

$$\frac{\mathrm{P}(y|x;M)}{\mathrm{P}(y|x;NULL)} = e^{-E(\omega_0)},$$

$$where\ E(\omega_0) = \mathrm{C}(\omega_0) + \frac{1}{2}\sum_{i=1}^{p} In(\lambda i\ /\lambda)) - N * In(n)$$

(If $E(\omega_0)$<0(= $e^{-E(\omega_0)}$>1), Evidence ratio is larger than 1 and this imply that the model with $\mathrm{P}(y|x;M)$ is reliable)

BAYES THEOREM AND GENERALIZATION

[Zhang et al.(2016)] showed that DNN generalize well on training input and can overfit on same input with randomized labels.

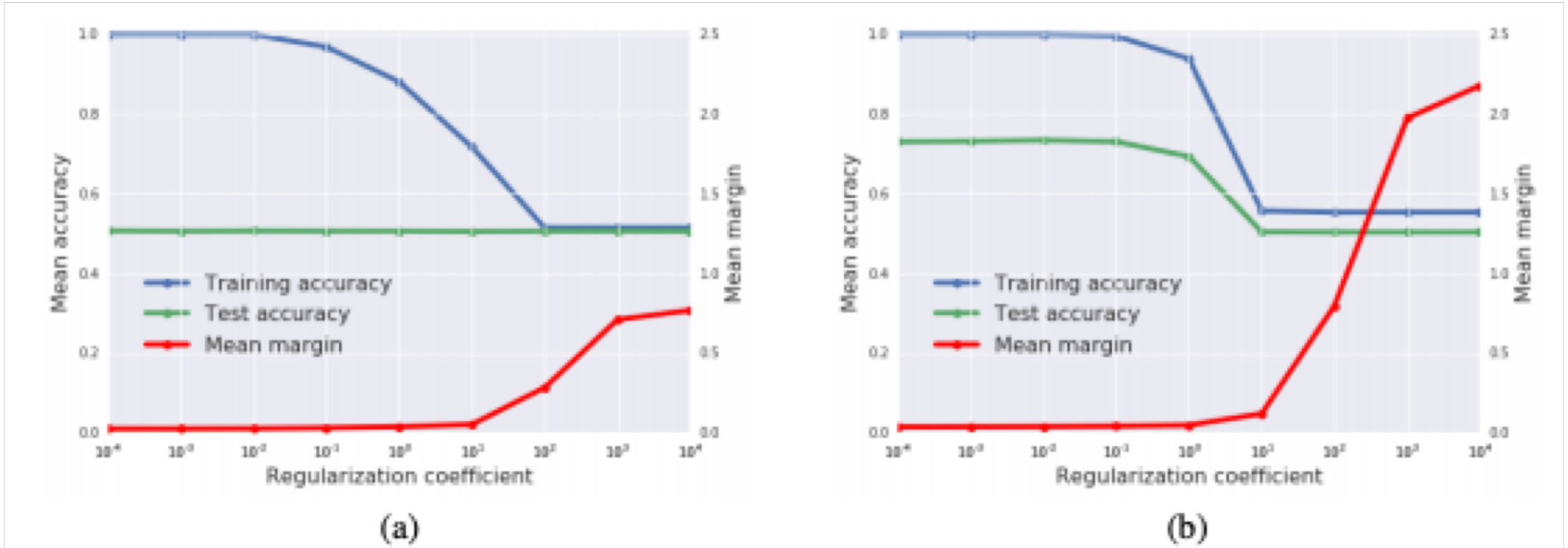
To show that this overfit is not unique to DNN, this paper consider easy model(= logistic regression)

- Using MNIST as input(labels are only 0 and 1)

they test two tasks

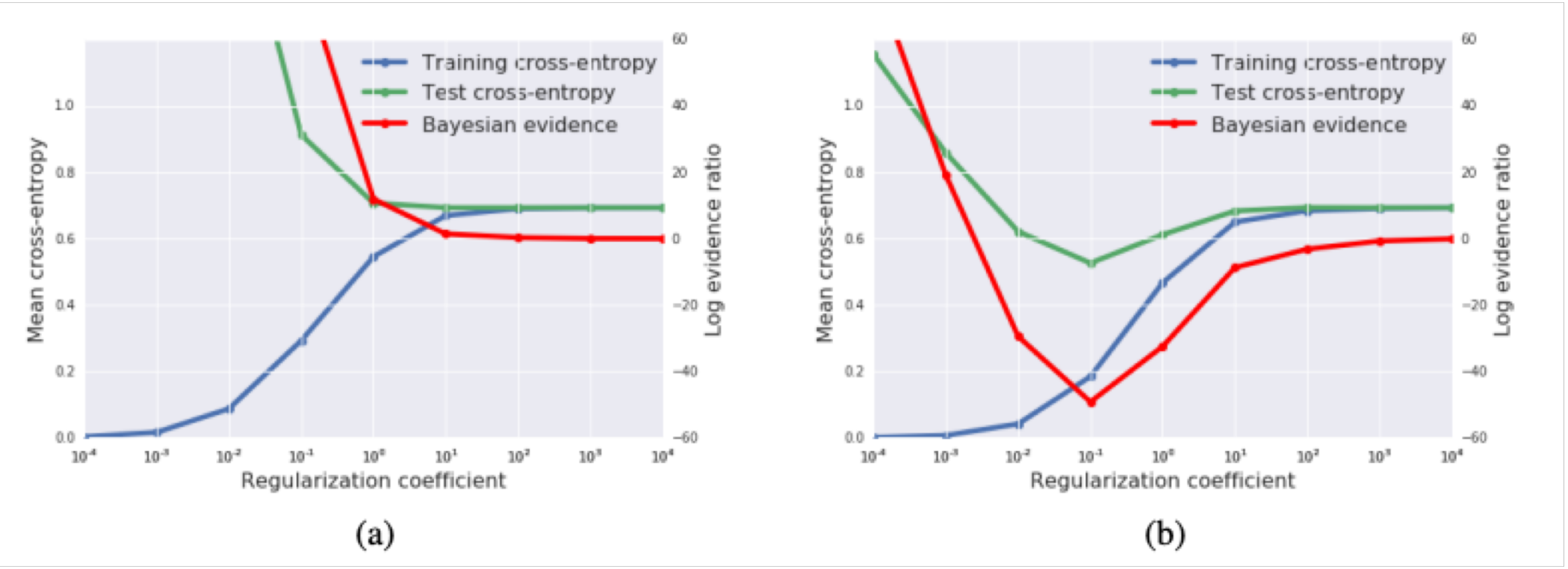
- the labels of both train and test sets are randomized
- the labels are not randomized

Train with randomized or not …



Left one is trained with RANDOMIZED labels, Right one is NOT RANDOMIZED
=> The phenomenon that model can memorize all data occurs not only DNN but also easy-model(like logistic regression)

Compare cross-entropy



Left one is trained with RANDOMIZED labels, Right one is NOT RANDOMIZED

=>Log Evidence ratio(Red calculated from Train data) is same as test cross-entropy

=>Beyesian Evidence and test cross entropy are strongly correlated

Supplement) Noise of SGD

SGD repeats the equation below and finds minima.(ℓ is loss function)

$$w' = w + \eta \times \frac{1}{BS} \sum \nabla \ell(x_i)$$

focus on right term…

$$\frac{\eta}{BS} \sum \nabla \ell(x_i) = \eta \times \nabla \ell + \frac{\eta}{BS} \sum (\nabla \ell(x_i) - \nabla \ell)$$

where $\nabla \ell = \frac{1}{N} \sum \nabla \ell(x_i)$

(this convert like “x” = “y + (x-y)”)

SGD’s gradient can divide into two terms, whole data’s gradient(left) and the noise which derived from the choice of data(right)

BAYES THEOREM AND SGD

This paper showed strong correlation between Bayesian evidence and generalization.

=> We should train DNN‘s minima with large evidence

Bayesian add isotropic Gaussian noise to gradient[Welling & Teh +]

They said that

- In small batch training, the noise of SGD is not isotropic(=large noise?)
- The noise of SGD can go away from sharp minima([Keskar +2016] found empirically)
- Gradient drives the SGD towards deep minima, noise drives the SGD towards broad minima

BAYES THEOREM AND SGD

Considering a shallow network with 800 hidden units and relu activations.

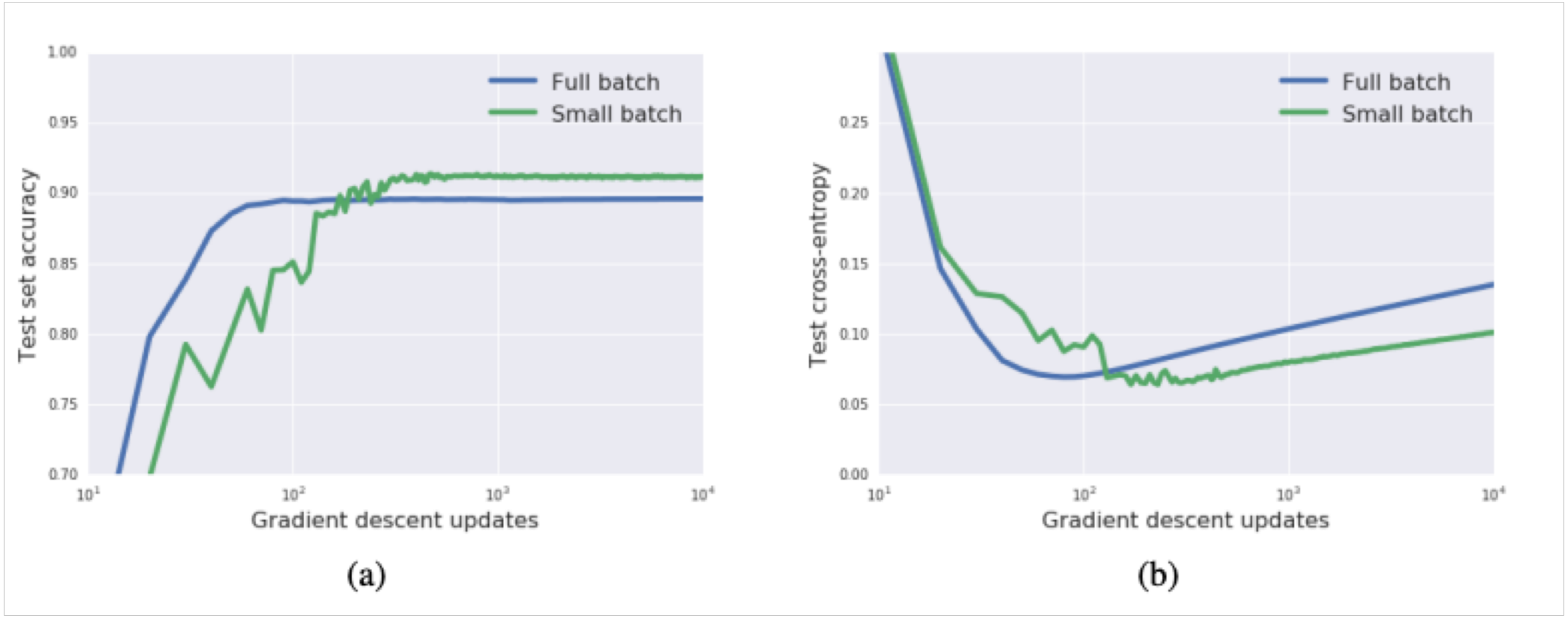
To see generalization gap, this paper set two conditions

- Small batch (BS=30)
- Full batch

Train data is 1000 images randomly selected from MNIST.

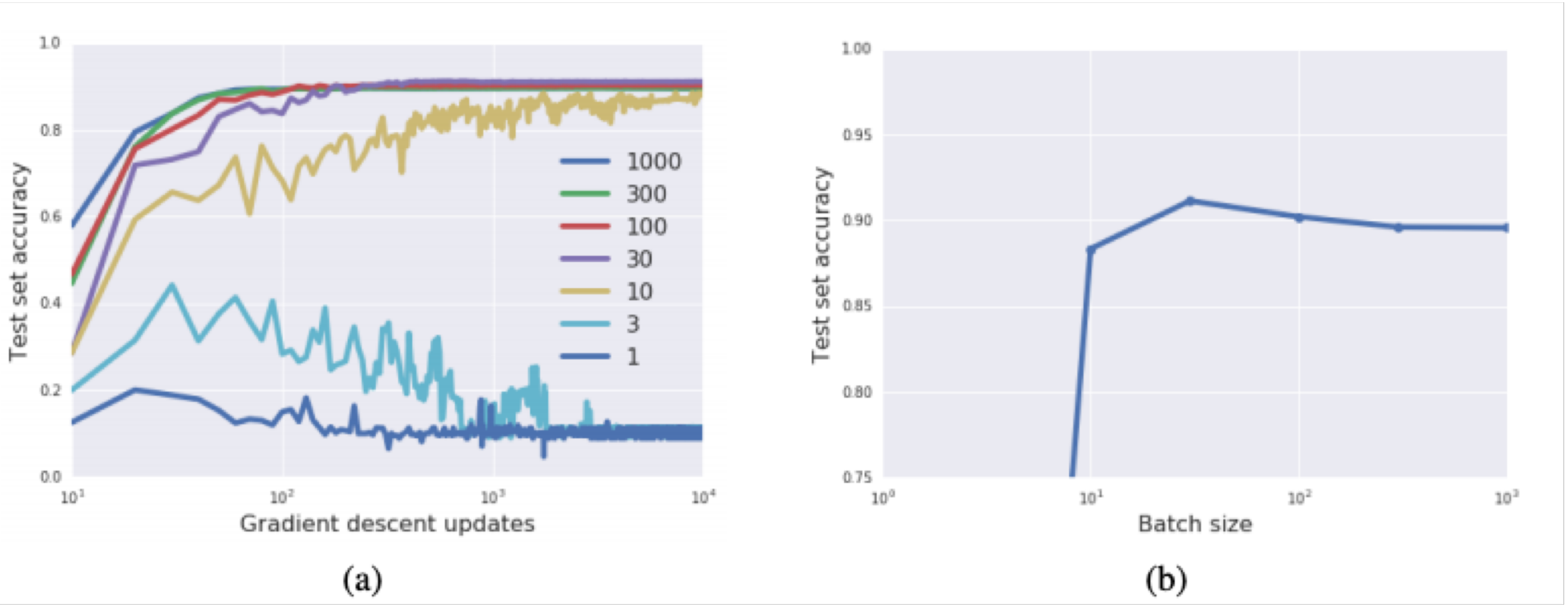
Train network with momentum-SGD (momentum=0.9, lr=1.0)

Train with large batch and small batch…



We can see “generalization gap” between large and small from left.
And we can see from right that the cross-entropy is increasing, indicative of overfitting.

Train with some batch size…



Right picture is correlation between batch size and TEST accuracy after 10000 training steps.

We can see that there is a “optimum” batch size which maximize test acc.

STOCHASTIC DIFFERENCE EQUATIONA AND THE SCALING RULES

This paper said that there is a optimal batch size.
Next, they show how optimal batch size depends on lr, training set size(=N), momentum.

$$w' = w + \Delta w = w + \varepsilon \times \frac{1}{BS} \sum \nabla \ell(x_i)$$

$$\Delta w = \frac{\varepsilon}{N} \left(\frac{dC}{d\omega} + \left(\frac{dC'}{d\omega} - \frac{dC}{d\omega} \right) \right)$$

There interpret this equation as the discrete update of a stochastic differential equation

$$\frac{\mathrm{d}\omega}{dt} = -\frac{dC}{d\omega} + \eta(t)$$

STOCHASTIC DIFFERENCE EQUATIONA AND THE SCALING RULES

And skip steps…

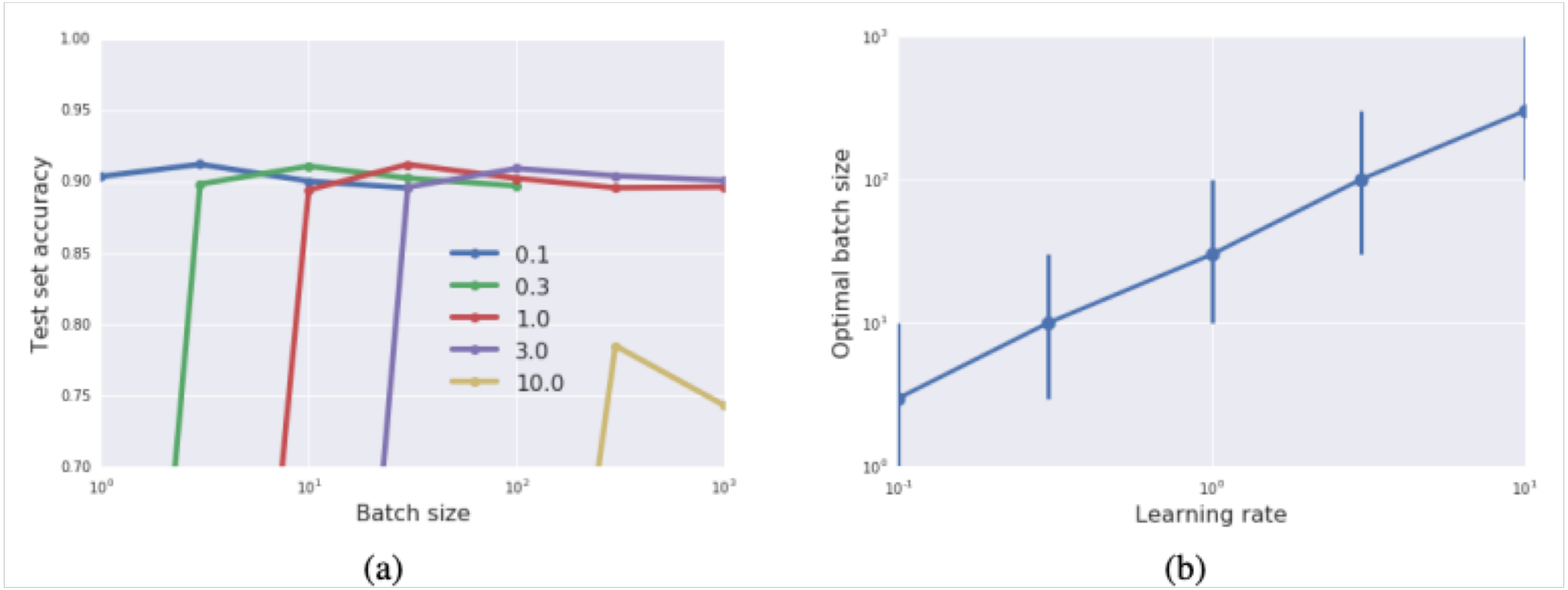
We can approximate the noise scale(=g)

$g = \varepsilon\left(\frac{N}{B} - 1\right) \approx \frac{\varepsilon \mathrm{N}}{\mathrm{B}}$, because N>>>B

This approximate says that the noise of sgd falls when the batch size increases.

=>$\mathrm{B}_{\mathrm{optimal}} \approx \frac{\varepsilon \mathrm{N}}{\mathrm{g}}$???

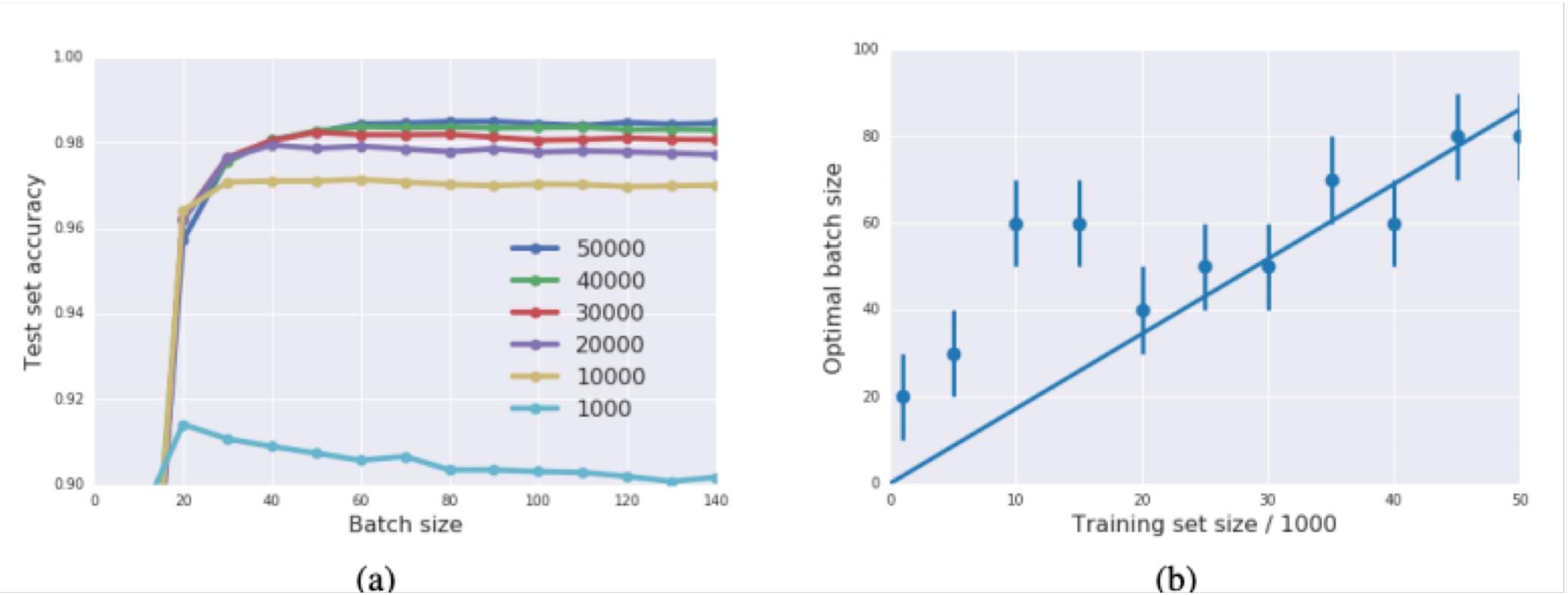
STOCHASTIC DIFFERENCE EQUATIONA AND THE SCALING RULES



Comparing lr, BS and test acc(left), we can see that optimal batch size is in proportion to learning rate

=> $B_{optimal} \propto \varepsilon$

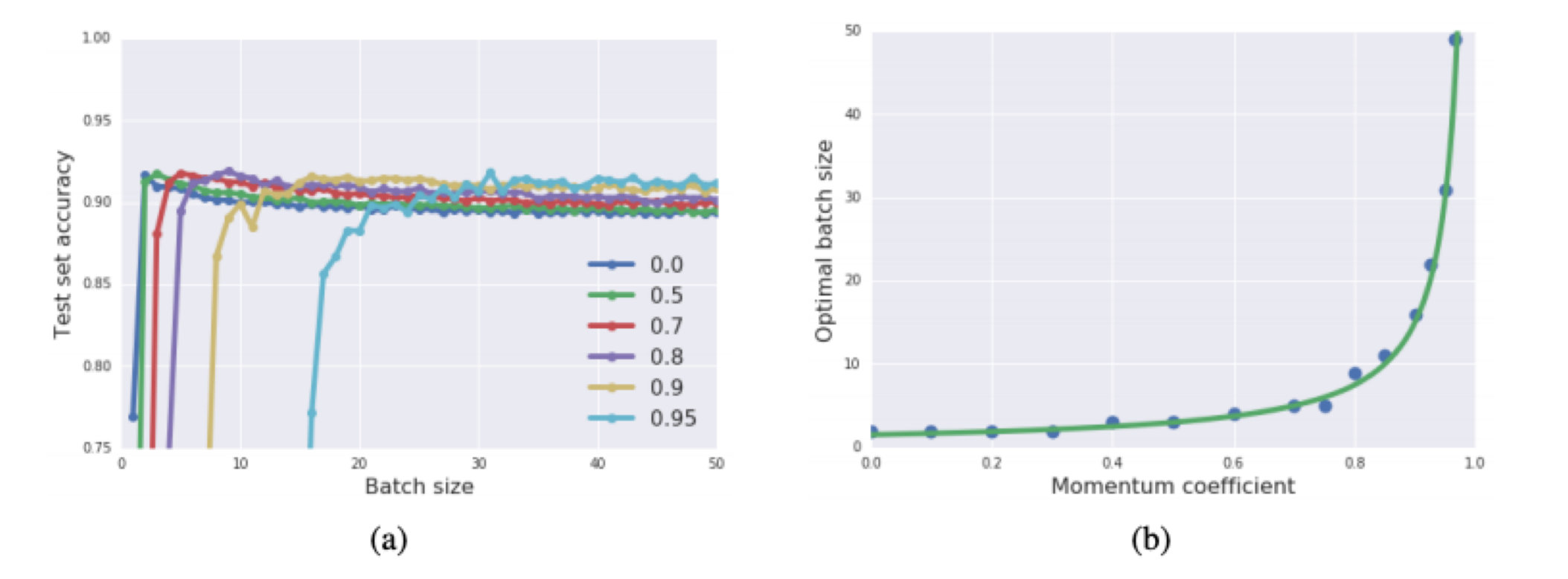
STOCHASTIC DIFFERENCE EQUATIONA AND THE SCALING RULES



Comparing N, BS and test acc(left), we can see that optimal batch size is in proportion to N (=data set size)

=> $B_{optimal} \propto N$

STOCHASTIC DIFFERENCE EQUATIONA AND THE SCALING RULES



Comparing BS, momentum terms(=m) and test acc(left), we can see that optimal batch size is in inverse proportion to N (=data set size)

=> $B_{optimal} \propto \frac{1}{1-m}$

STOCHASTIC DIFFERENCE EQUATIONA AND THE SCALING RULES

We can explain from these there correlations between $\mathrm{B} \propto \frac{\varepsilon \mathrm{N}}{g(1-m)}$.

We can increase batch size with increasing lr.

=> Speed up without accuracy drop.

This match the fact that we should keep $\frac{\mathrm{lr}}{BS}$ constant.([Goyal +2017])

Conclusion

1)DNN or linear regression which generalize well on informative labels can memorize whole dataset with randomized labels of same input.

These observations are explained by Beyesian evidence

2)Mini-batch noise drives SGD away from sharp minima, therefore there is an optimum batch size which maximizes test acc.

3)Interpreting SGD as the discretization of stochastic differential equation, we predict optimum batch size and $\mathrm{B}_{\mathrm{opt}} \propto \frac{\varepsilon \mathrm{N}}{g(1-m)}$.

出典

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. arXiv preprint arXiv:1609.04836

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., ... & He, K. (2017). Accurate, large minibatch SGD: training imagenet in 1 hour. arXiv preprint arXiv:1706.02677.

Welling, M., & Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (pp. 681-688).