ML/DS 지식 따라가기

Optimization & Overfitting Prevention

WHAT IS GRADIENT?

스칼라장의 최대의 증가율을 나타내는 벡터장

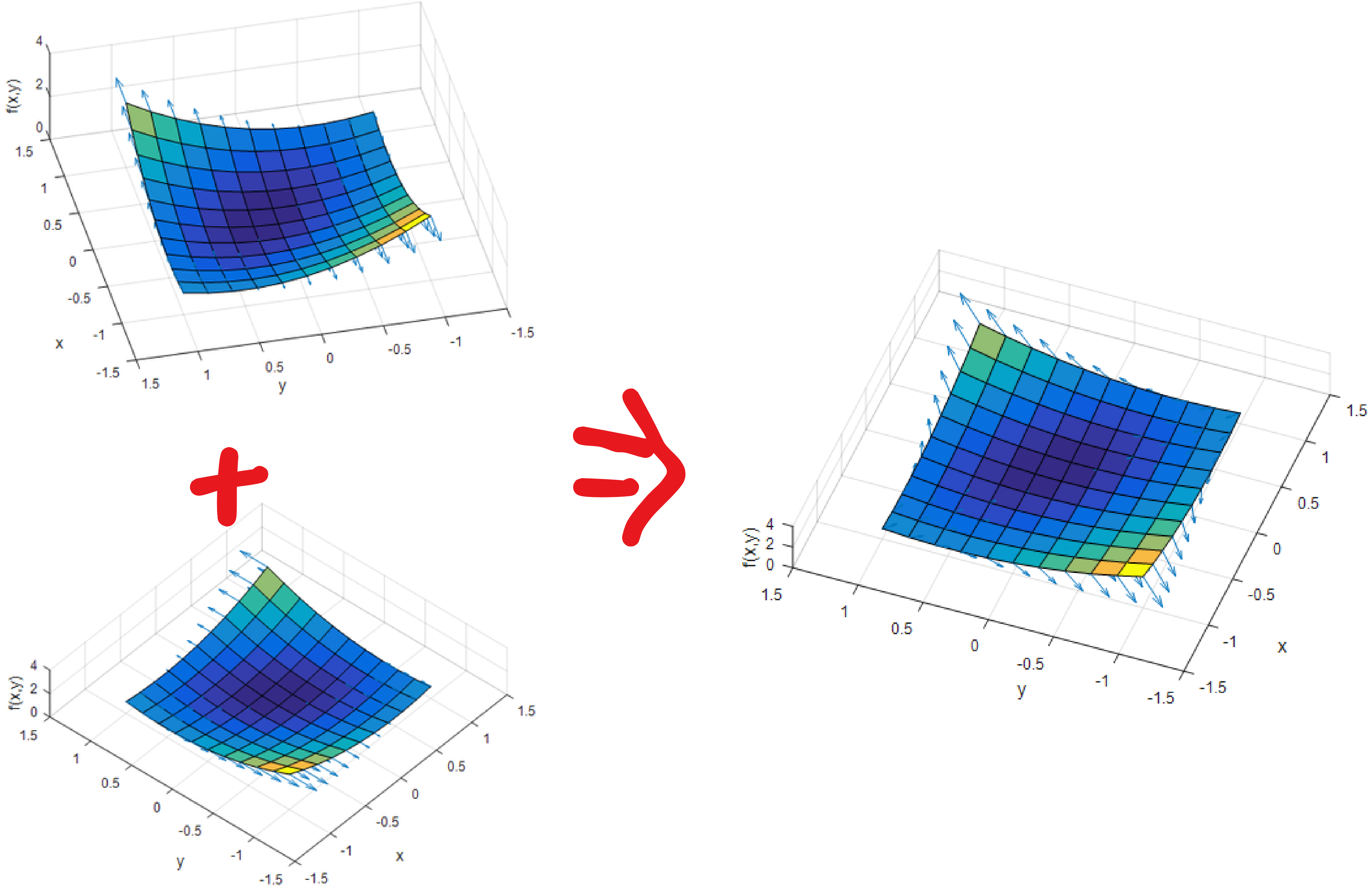
For a differentiable scalar field $f(\mathbf{x})$,

$$\nabla f = \frac{\partial f}{\partial x_1}\mathbf{e}_1 + \frac{\partial f}{\partial x_2}\mathbf{e}_2 + \frac{\partial f}{\partial x_3}\mathbf{e}_3 = \sum_{i=1}^{3} \frac{\partial f}{\partial x_i} e_i = \frac{\partial f}{\partial x_i}\mathbf{e}_i$$

∇f is called the gradient of a scalar field $f(\mathbf{x})$
and $\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + x_3\mathbf{e}_3$ (a position vector)

GRADIENT의 특성

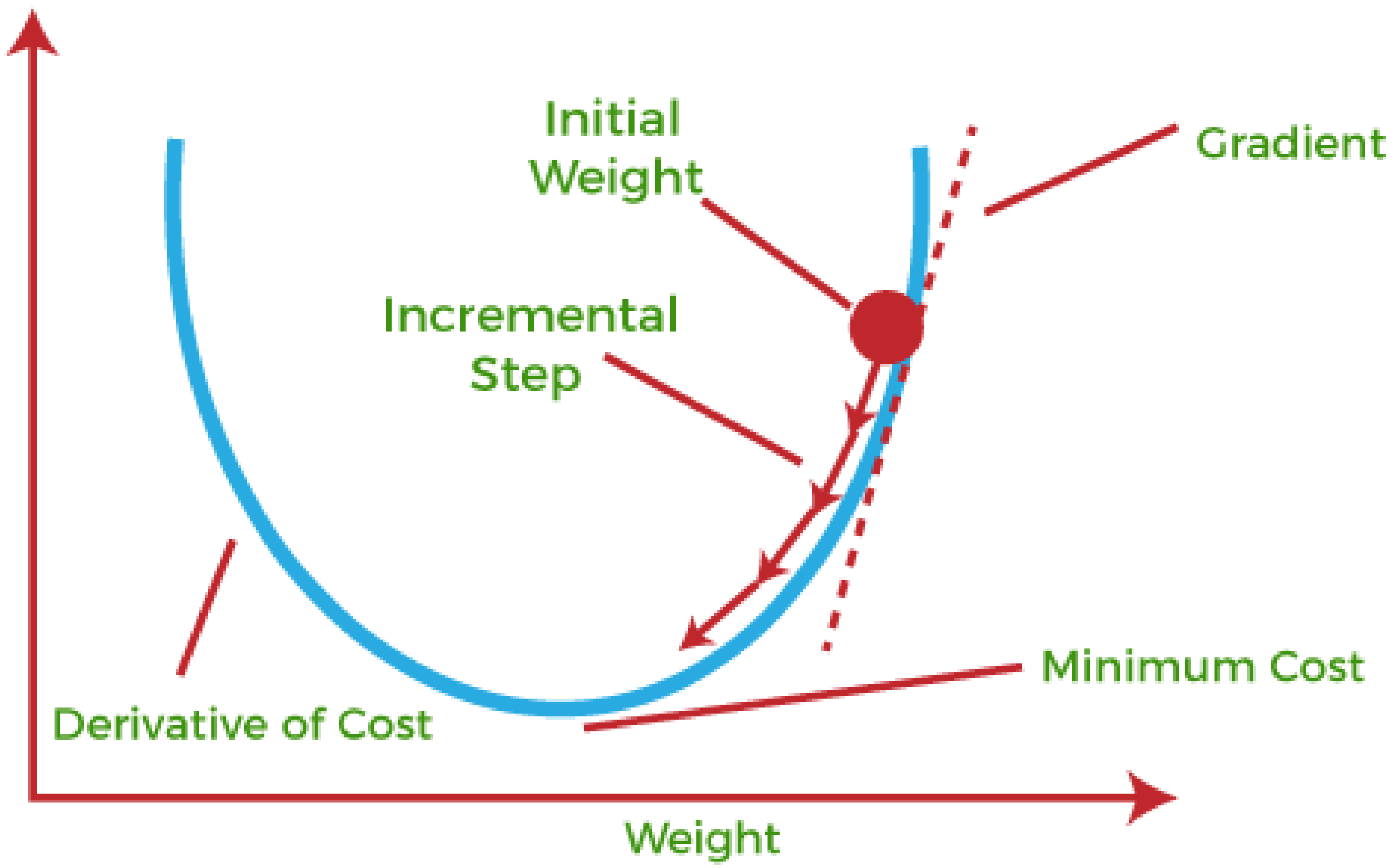
방향성을 가짐 / 등위선과 직교함



$$\frac{\partial f}{\partial x}x' + \frac{\partial f}{\partial y}y' + \frac{\partial f}{\partial z}z' = (grad f)\cdot r' = 0$$

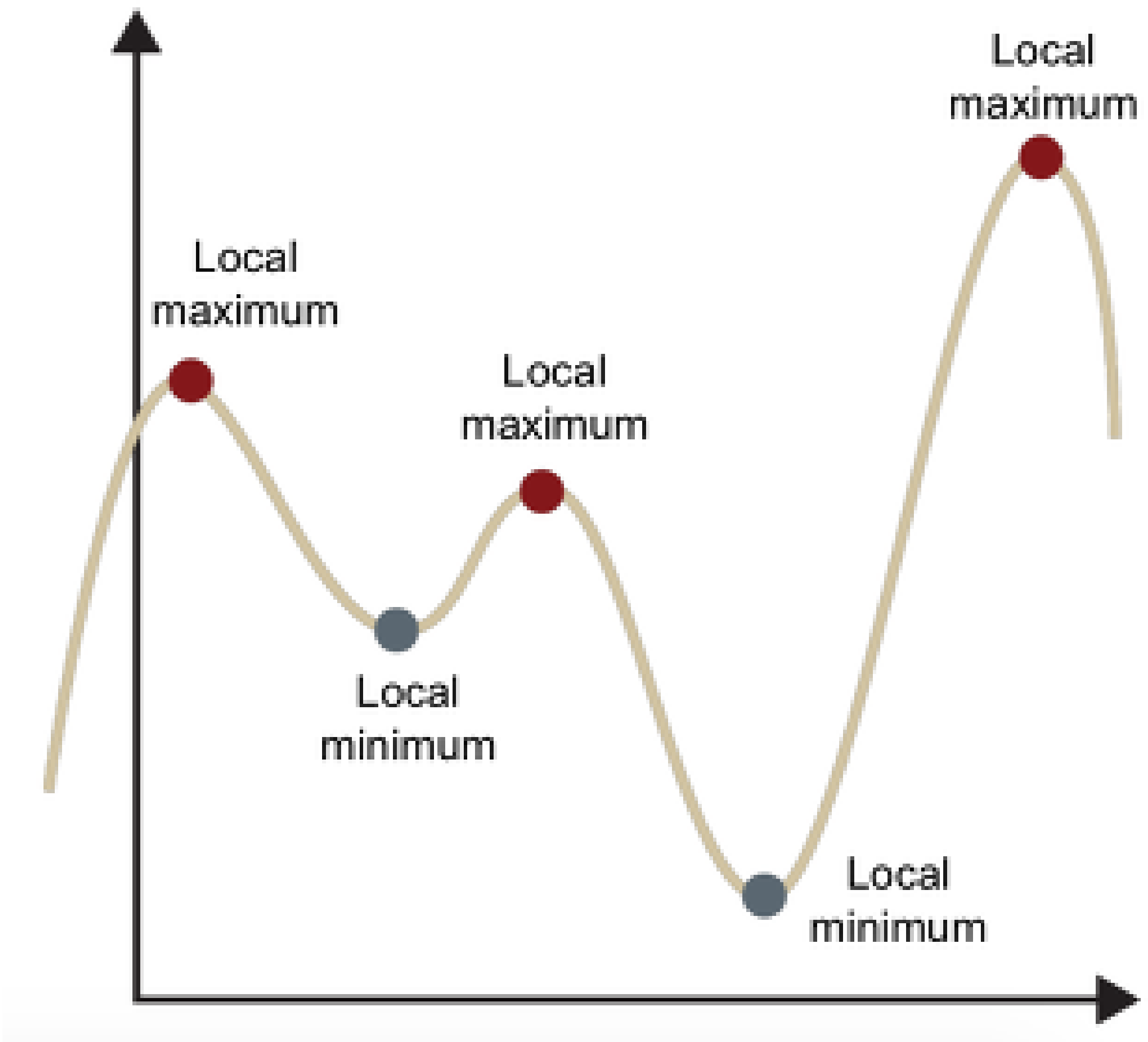
GRADIENT DESCENT란

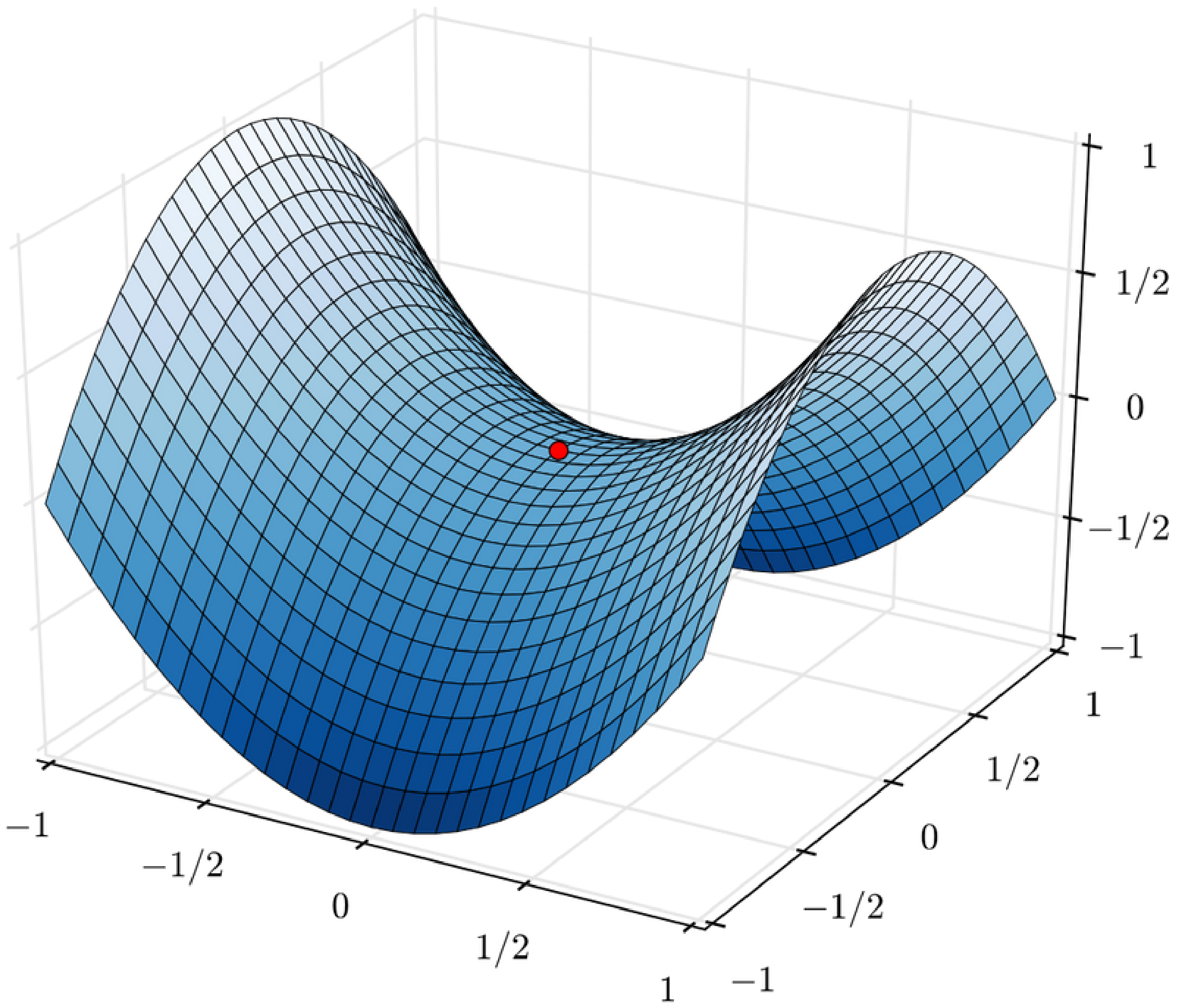
GRADIENT를 계산하고 그 반대 방향으로 계속 가는 최적화 기법



GRADIENT DESCENT의 한계

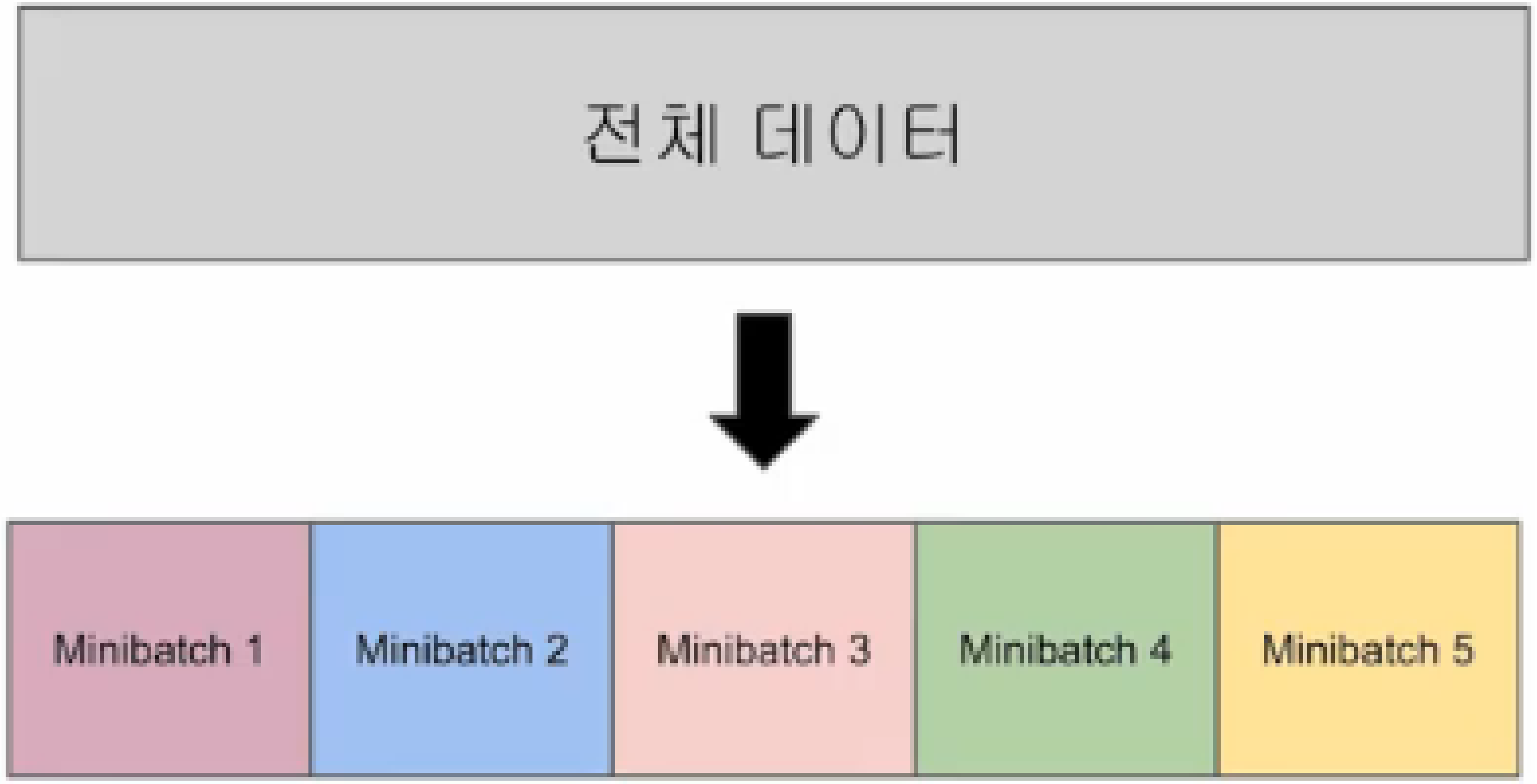
Local Minima, Saddle Point에 빠질 수 있음





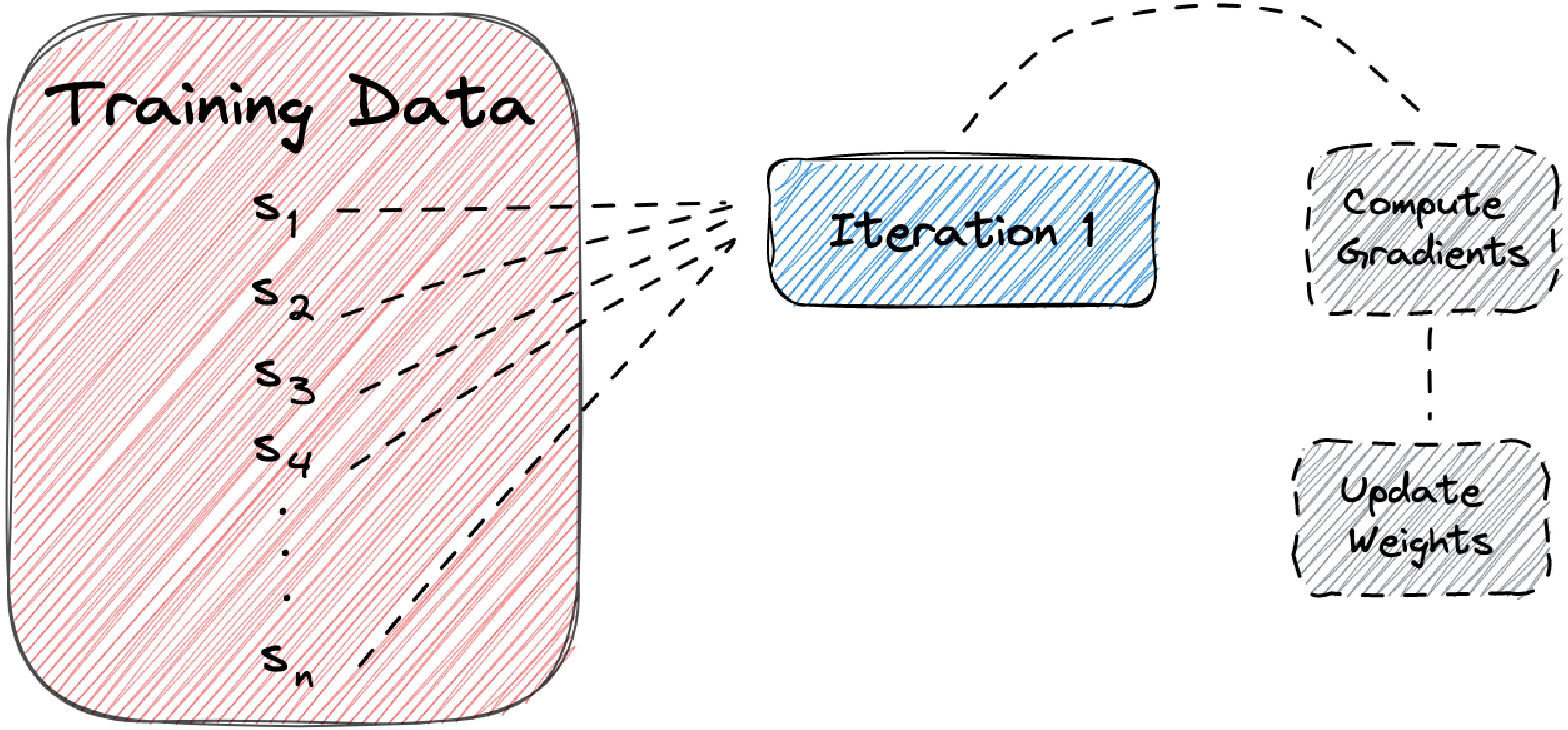
MINI-BATCH GRADIENT DESCENT

Batch와 Mini-batch



MINI-BATCH GRADIENT DESCENT

Mini-batch의 장점과 한계



STOCHASTIC GRADIENT DESCENT

Stochastic 이란

영어사전 다른 어학정보 2

stochastic

미국·영국[stəkǽstik]

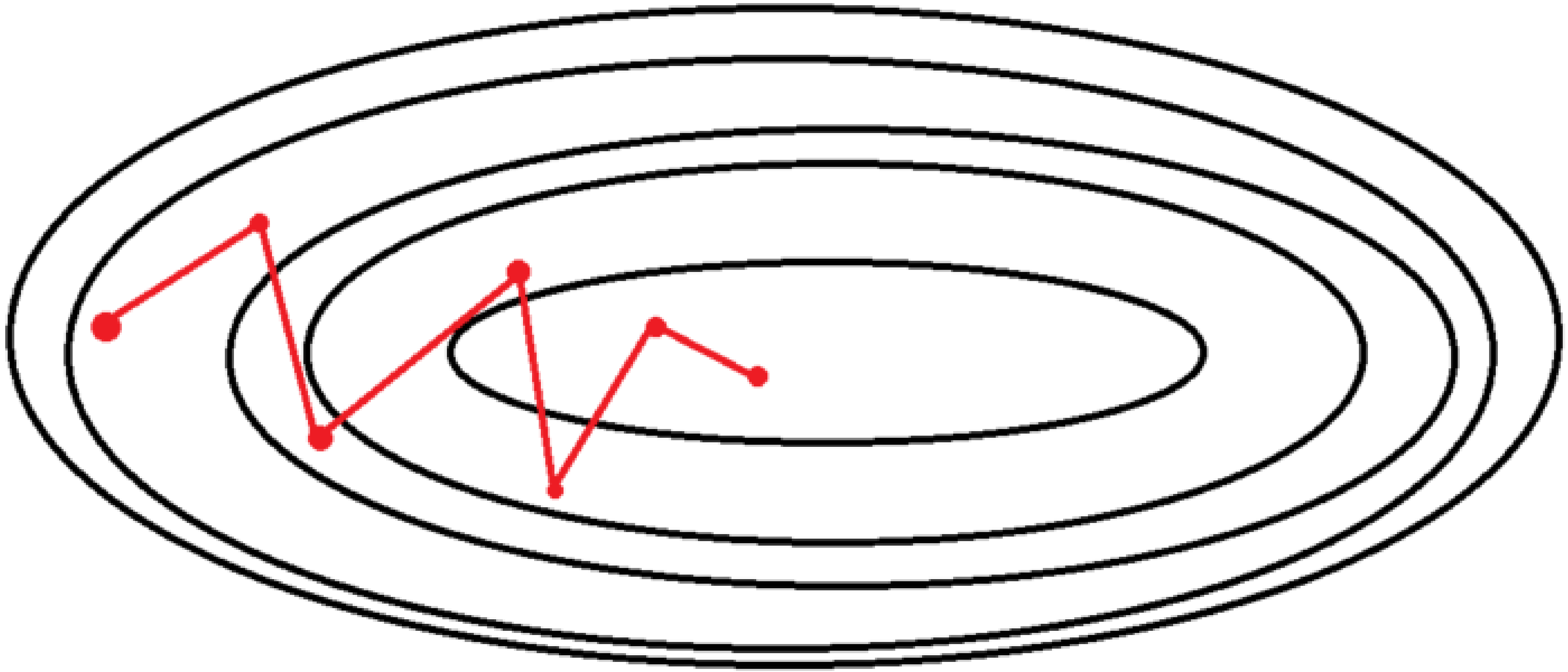
추계학(推計學)의, 확률(론)적인

영어사전 결과 더보기

어학사전 더보기 →

STOCHASTIC GRADIENT DESCENT

SGD란



STOCHASTIC GRADIENT DESCENT

Robbins-Monro Algorithm

$$\begin{aligned}
S_N &= \frac{1}{N}\sum_{i=1}^{N} X_i \\
&= \frac{1}{N}\sum_{i=1}^{N-1} X_i + \frac{1}{N}X_N \\
&= \frac{N-1}{N}\frac{1}{N-1}\sum_{i=1}^{N-1} X_i + \frac{1}{N}X_N \\
&= (1-\frac{1}{N})S_{N-1} + \frac{1}{N}X_N \\
&= S_{N-1} + \frac{1}{N}(X_N - S_{N-1})
\end{aligned}$$

STOCHASTIC GRADIENT DESCENT

GD계열 비교하기

	GD	SGD
Update frequency	After training all data	After training mini-batch
	slow	fast
Accuracy for one step	Optimized	Not optimized

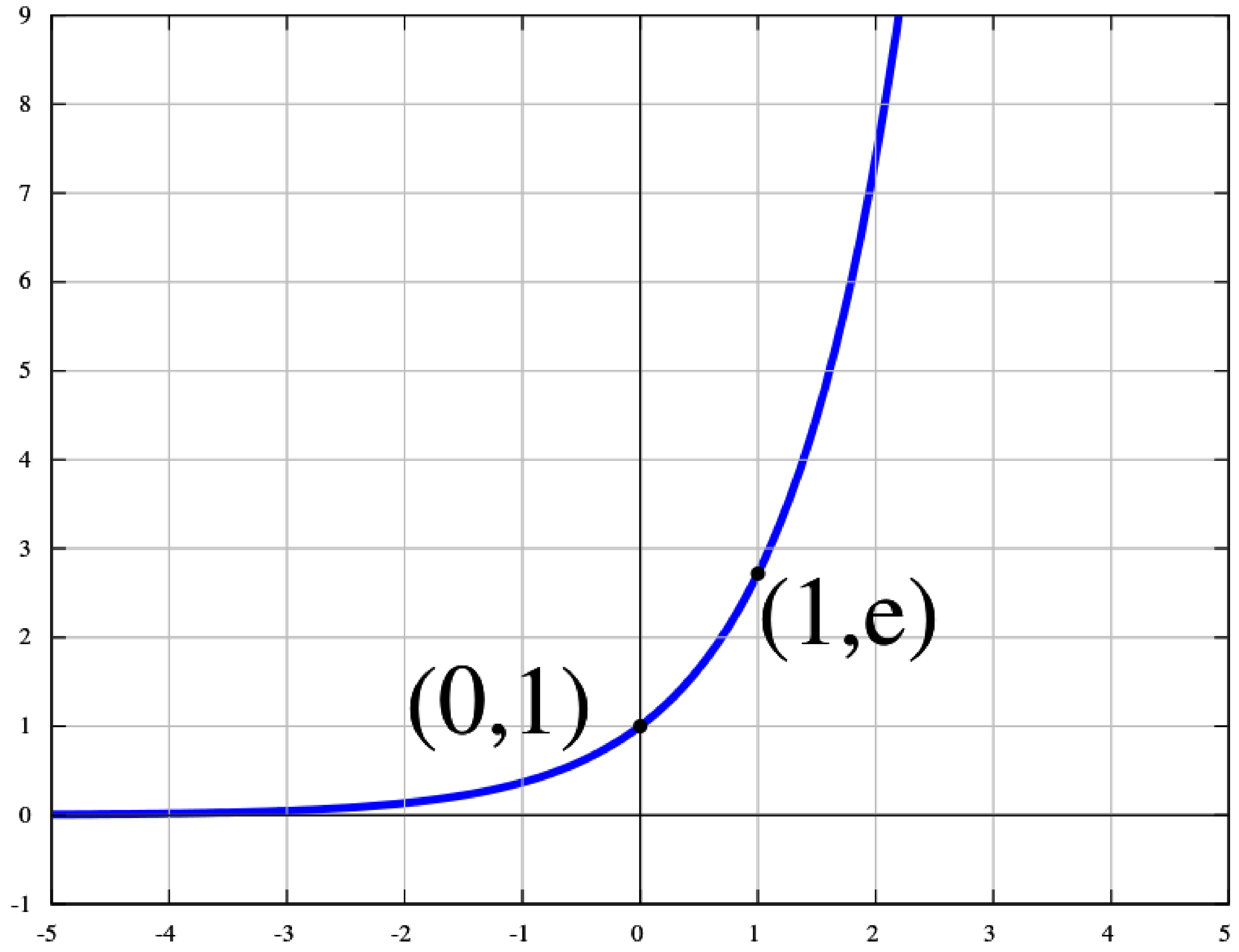
STOCHASTIC GRADIENT DESCENT

GD계열의 한계점

- LOCAL MINIMA 문제를 완벽하게 해결할 수 없음
- LEARNING RATE를 설정하는데 어려움이 있음
- 초기값에 민감함
- 비선형 문제를 처리하기 어려울 수 있음

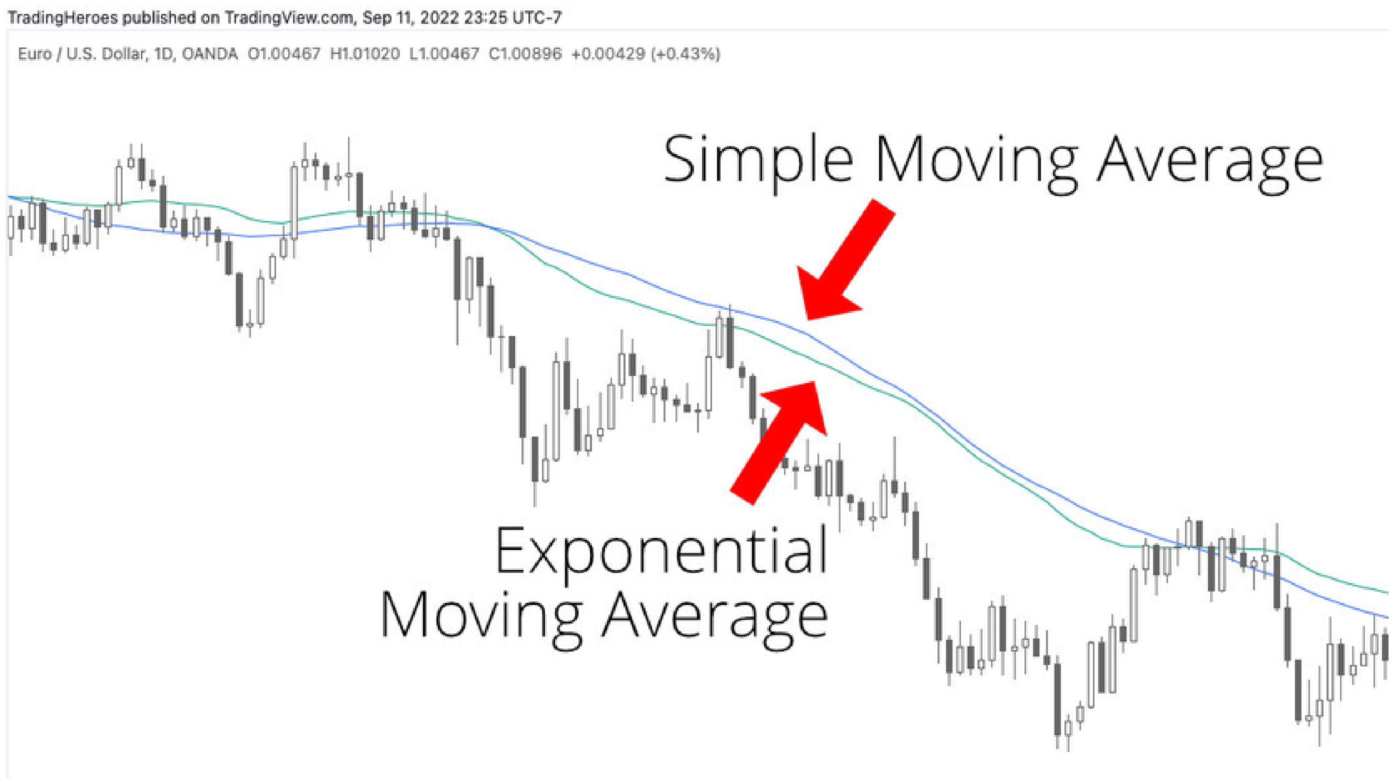
EMA (EXPONENTIAL MOVING AVERAGE)

Exponential이란



EMA (EXPONENTIAL MOVING AVERAGE)

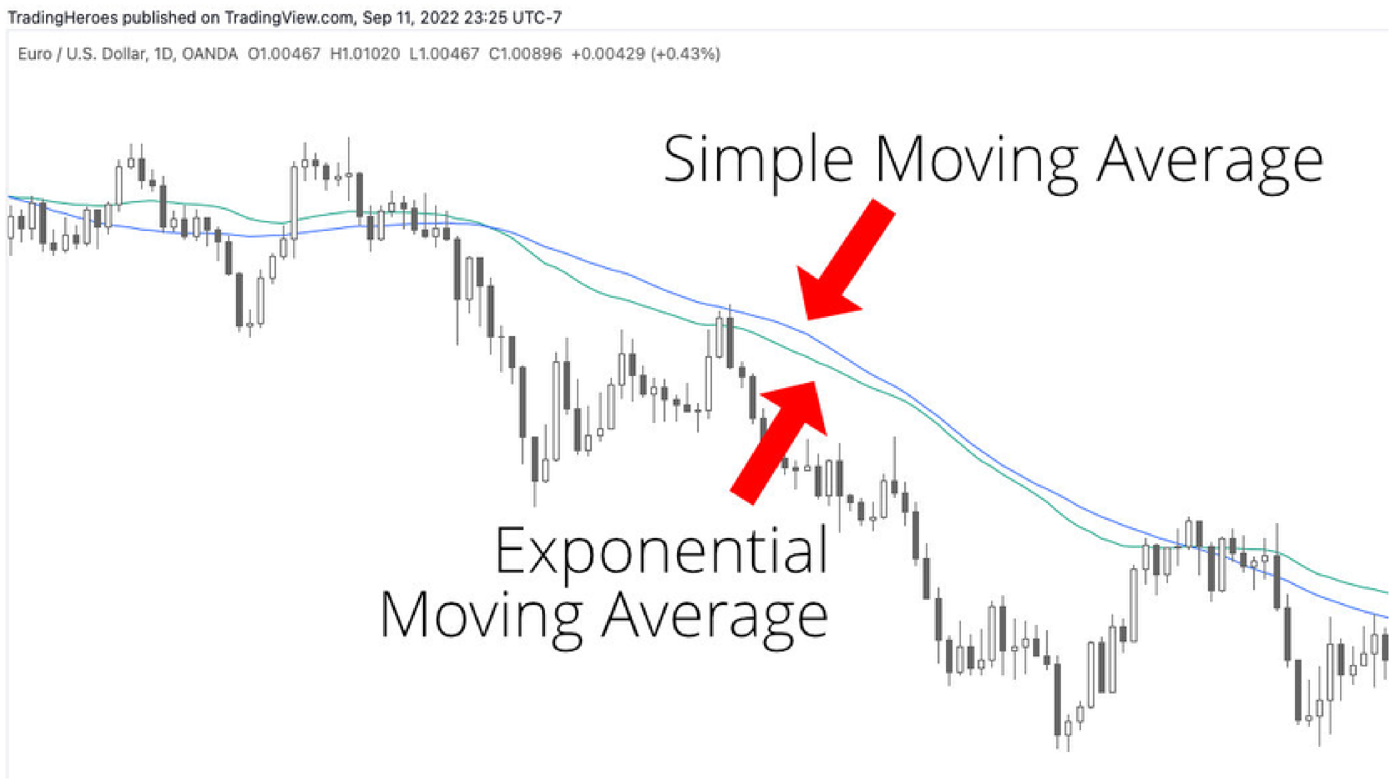
EMA란



$$V_t = \beta \times V_{t-1} + (1-\beta) \times \Theta_t$$

EMA (EXPONENTIAL MOVING AVERAGE)

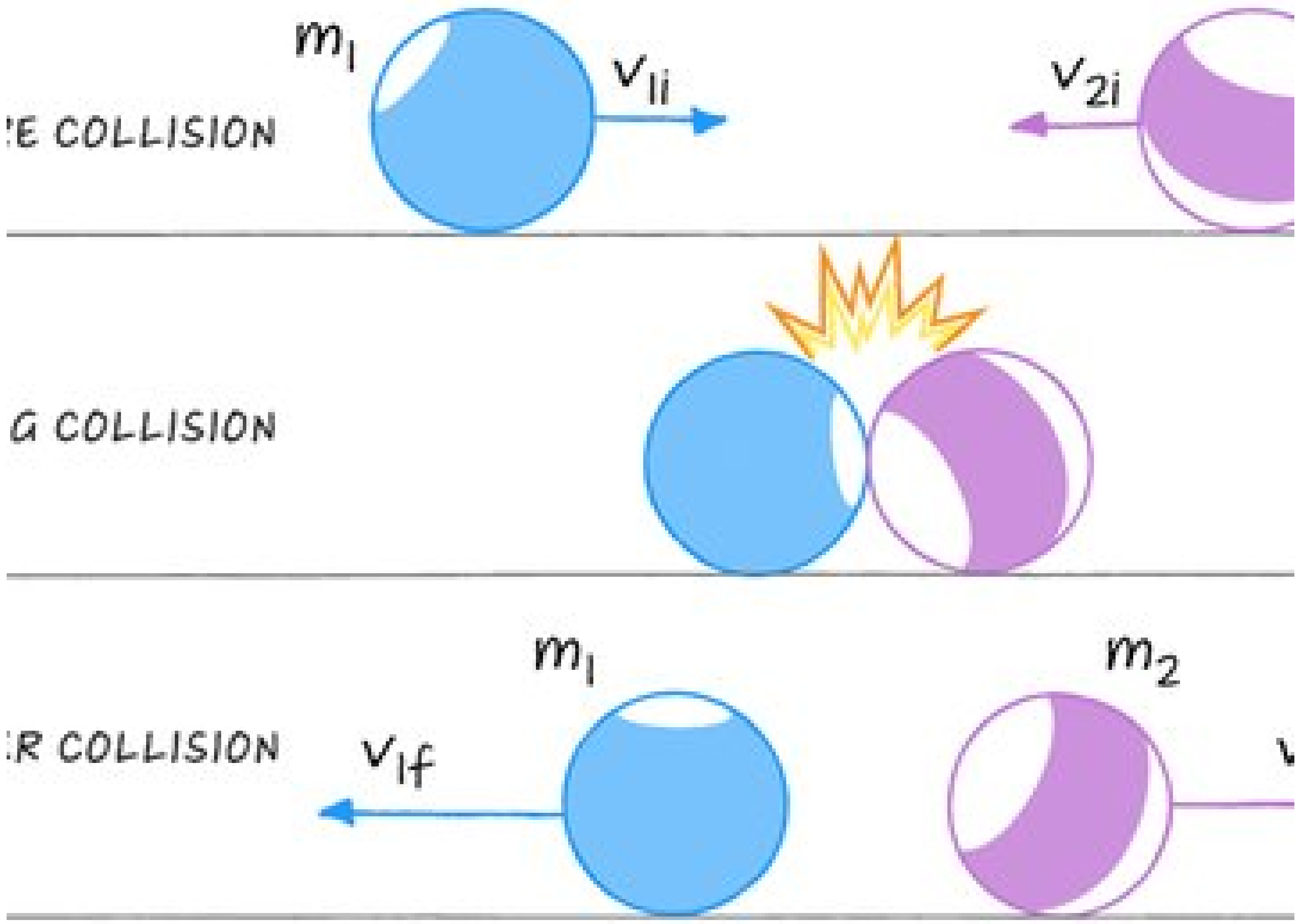
EMA란



$$V_t = \beta \times V_{t-1} + (1-\beta) \times \Theta_t$$

MOMENTUM

Momemtum in Physics



EMA (EXPONENTIAL MOVING AVERAGE)

Exponentially Decaying Moving Average

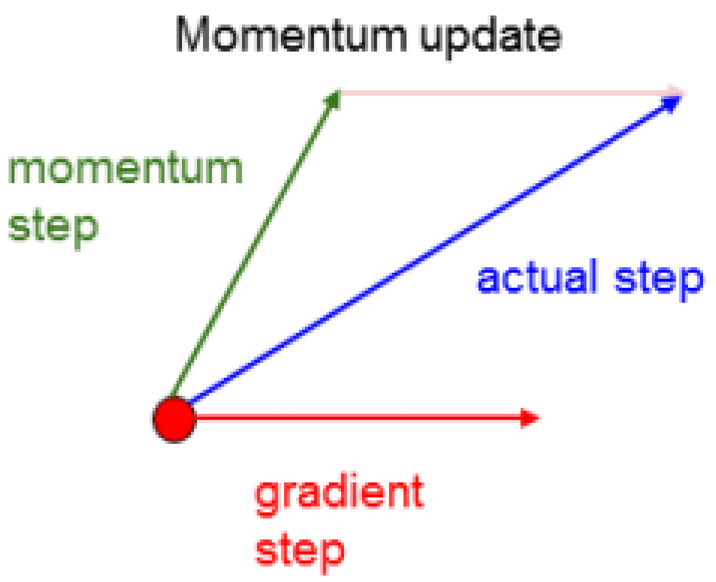
$$y_0 = x_0$$

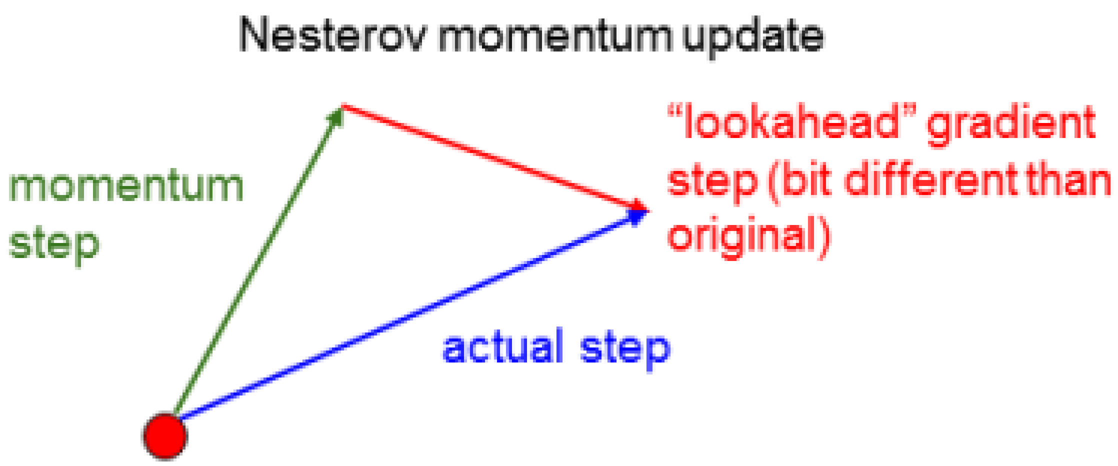
$$y_t = \frac{x_t+(1-\alpha)x_{t-1}+(1-\alpha)^2 x_{t-2}+\ldots+(1-\alpha)^t x_0}{1+(1-\alpha)+(1-\alpha)^2+\ldots+(1-\alpha)^t}$$

NAG

Nesterov Accekerated Gradient

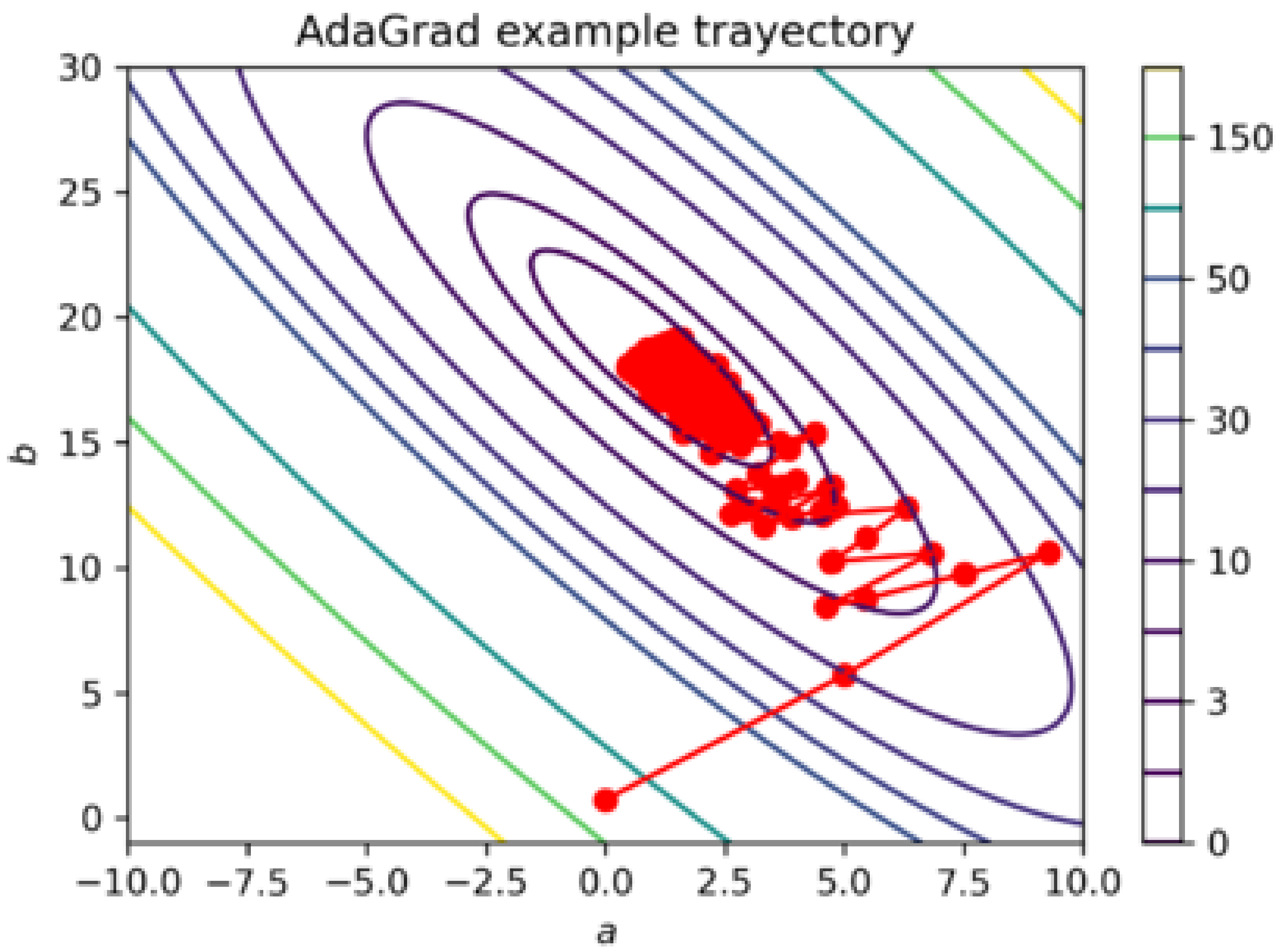
$$v_t = \gamma v_{t-1} + \eta \nabla_\theta J(\theta - \gamma v_{t-1})$$
$$\theta = \theta - v_t$$





ADAGRAD

Adpative Gradient



$$v_t^w = v_{t-1}^w + (\nabla w_t)^2$$

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{v_t^w + \epsilon}} * \nabla w_t$$

$$v_t^b = v_{t-1}^b + (\nabla b_t)^2$$

$$b_{t+1} = b_t - \frac{\eta}{\sqrt{v_t^b + \epsilon}} * \nabla b_t$$

ADADELTA & RMSPROP

$$g_{t+1} = \gamma g_t + (1-\gamma)\bigtriangledown \mathcal{L}(\theta)^2$$

$$x_{t+1} = \gamma x_t + (1-\gamma) v_{t+1}^2$$

$$v_{t+1} = -\frac{\sqrt{x_t + \epsilon}\delta L(\theta_t)}{\sqrt{g_{t+1} + \epsilon}}$$

$$\theta_{t+1} = \theta_t + v_{t+1}$$

$$g_{t+1} = g_t + \delta L(\theta_t)^2$$

$$\theta_{t+1} = \theta_t - \frac{\alpha \delta L(\theta)^2}{\sqrt{g_{t+1}} + \epsilon}$$

$$v_{dw} = \beta \cdot v_{dw} + (1-\beta) \cdot dw^2$$

$$v_{db} = \beta \cdot v_{dw} + (1-\beta) \cdot db^2$$

$$W = W - \alpha \cdot \frac{dw}{\sqrt{v_{dw}} + \epsilon}$$

$$b = b - \alpha \cdot \frac{db}{\sqrt{v_{db}} + \epsilon}$$

ADAM

Adaptive Moment Estimation

$$\nu_t = \beta_1 * \nu_{t-1} - (1 - \beta_1) * g_t$$
$$s_t = \beta_2 * s_{t-1} - (1 - \beta_2) * g_t^2$$

$$\Delta\omega_t = -\eta \frac{\nu_t}{\sqrt{s_t + \epsilon}} * g_t$$

$$\omega_{t+1} = \omega_t + \Delta\omega_t$$

$\eta : Initial\ Learning\ rate$

$g_t : Gradient\ at\ time\ t\ along\ \omega^j$

$\nu_t : Exponential\ Average\ of\ gradients\ along\ \omega_j$

$s_t : Exponential\ Average\ of\ squares\ of\ gradients\ along\ \omega_j$

$\beta_1, \beta_2 : Hyperparameters$

ADAM

Moment

Moment	Uncentered	Centered
1st	$E(X) = \mu$	
2nd	$E(X^2)$	$E((X-\mu)^2)$
3rd	$E(X^3)$	$E((X-\mu)^3)$
4th	$E(X^4)$	$E((X-\mu)^4)$

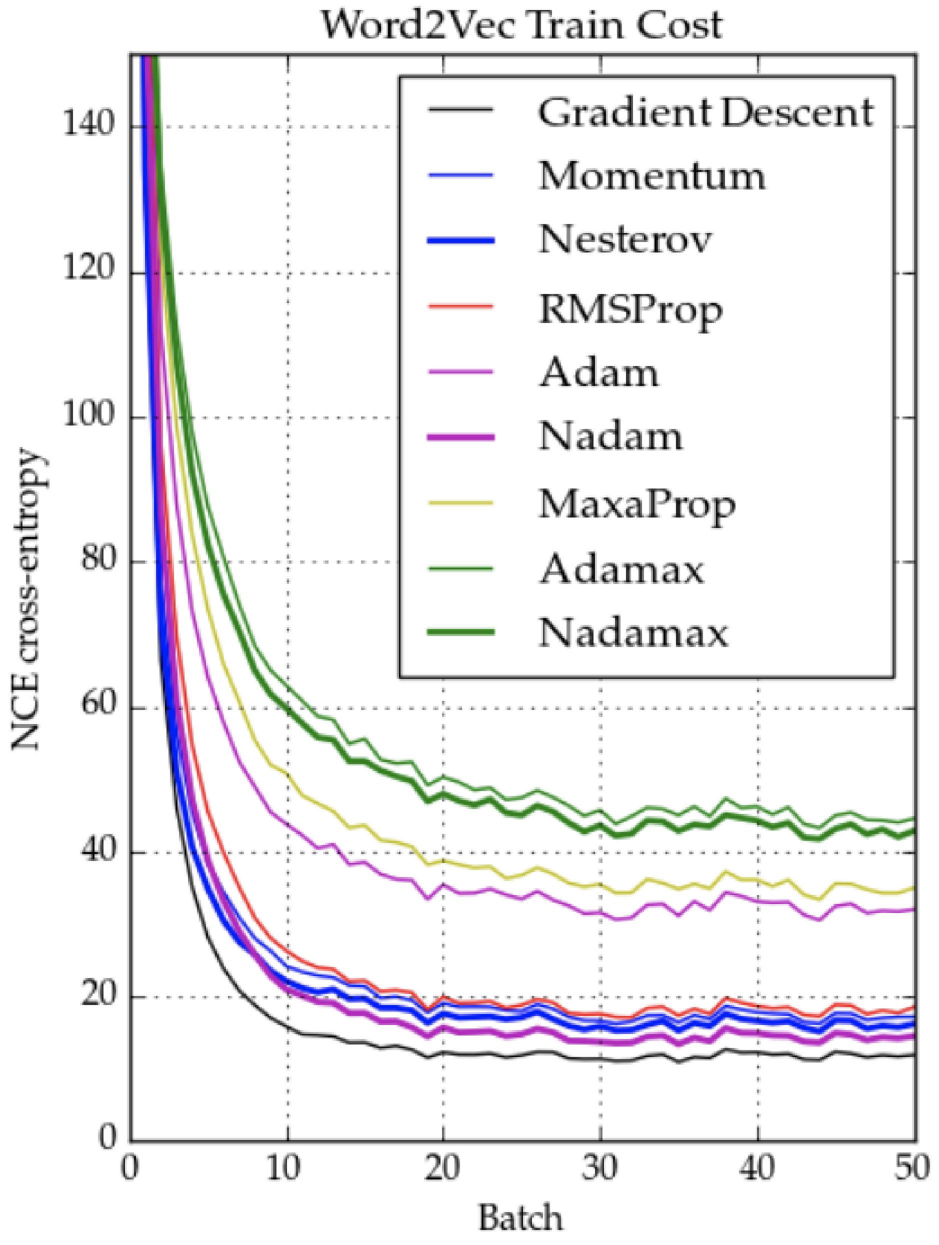
$$\mathrm{Mean}(X) = E(X)$$

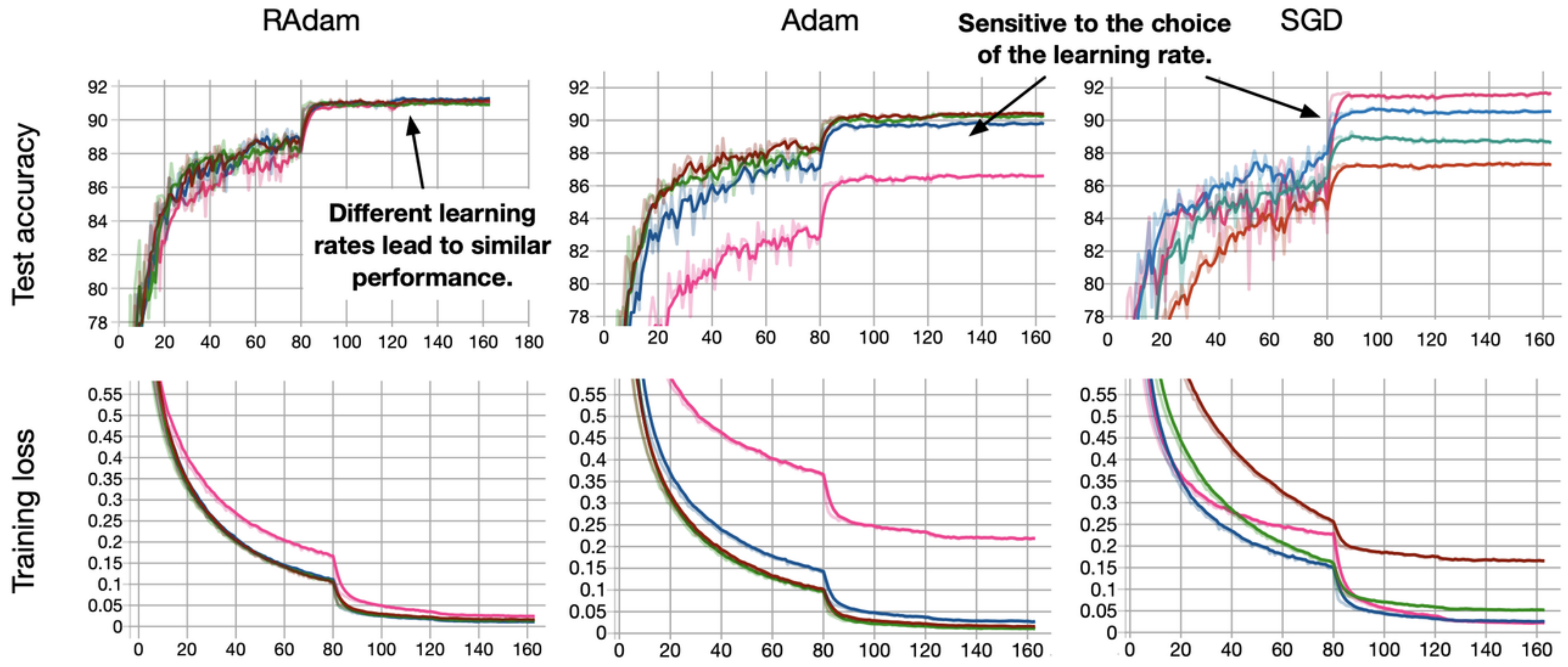
$$\mathrm{Var}(X) = E((X-\mu)^2) = \sigma^2$$

$$\mathrm{Skewness}(X) = E((X-\mu)^3) / \sigma^3$$

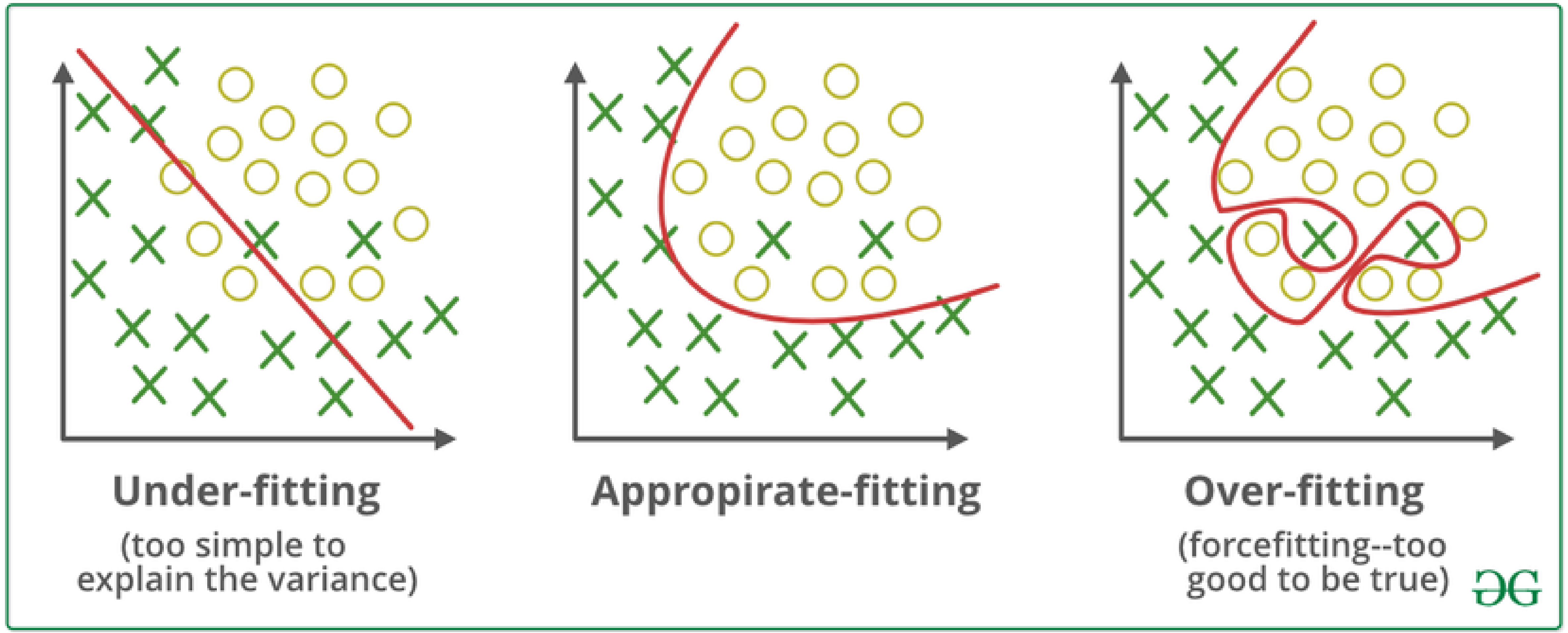
$$\mathrm{Kurtosis}(X) = E((X-\mu)^4) / \sigma^4$$

ADAM SERIES

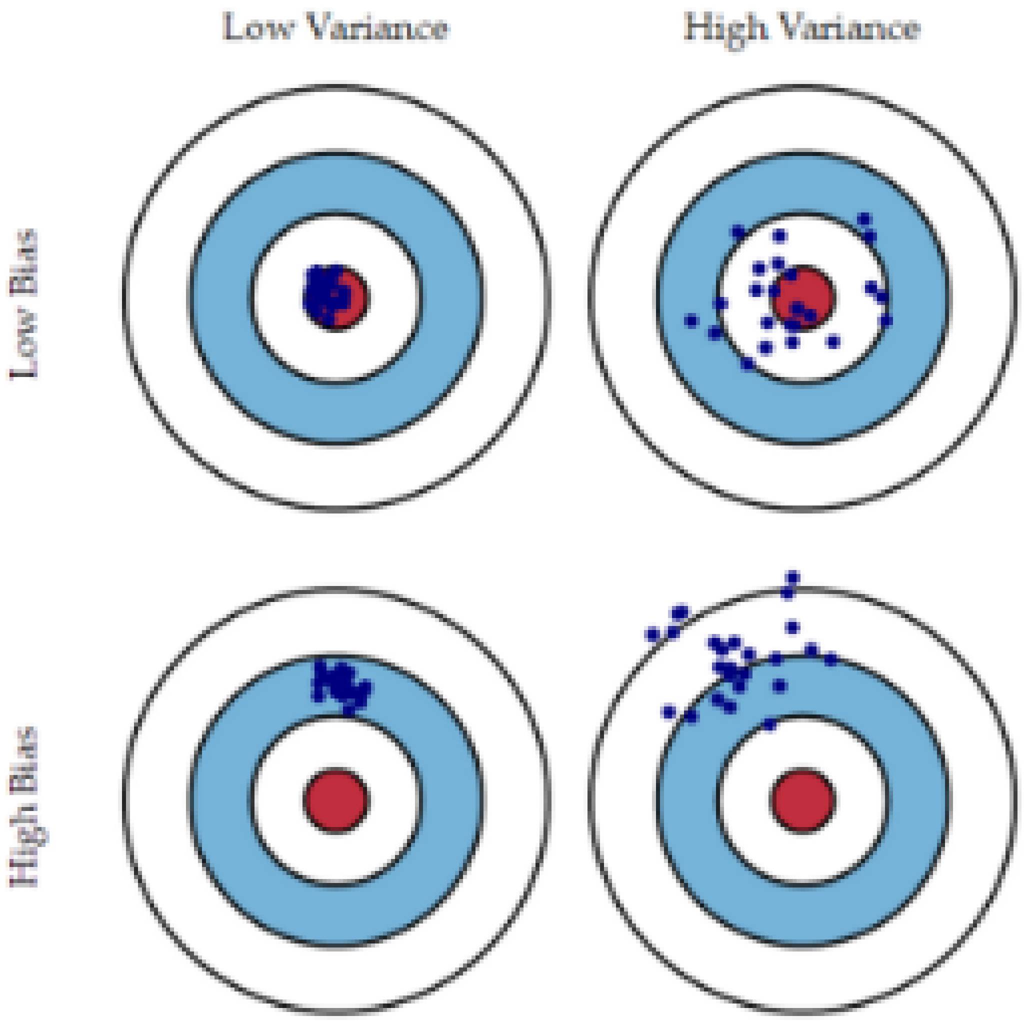

Word2Vec Train Cost
NCE cross-entropy
Batch
Gradient Descent
Momentum
Nesterov
RMSProp
Adam
Nadam
MaxaProp
Adamax
Nadamax


RAdam
Adam
SGD
Sensitive to the choice of the learning rate.
Different learning rates lead to similar performance.
Test accuracy
Training loss

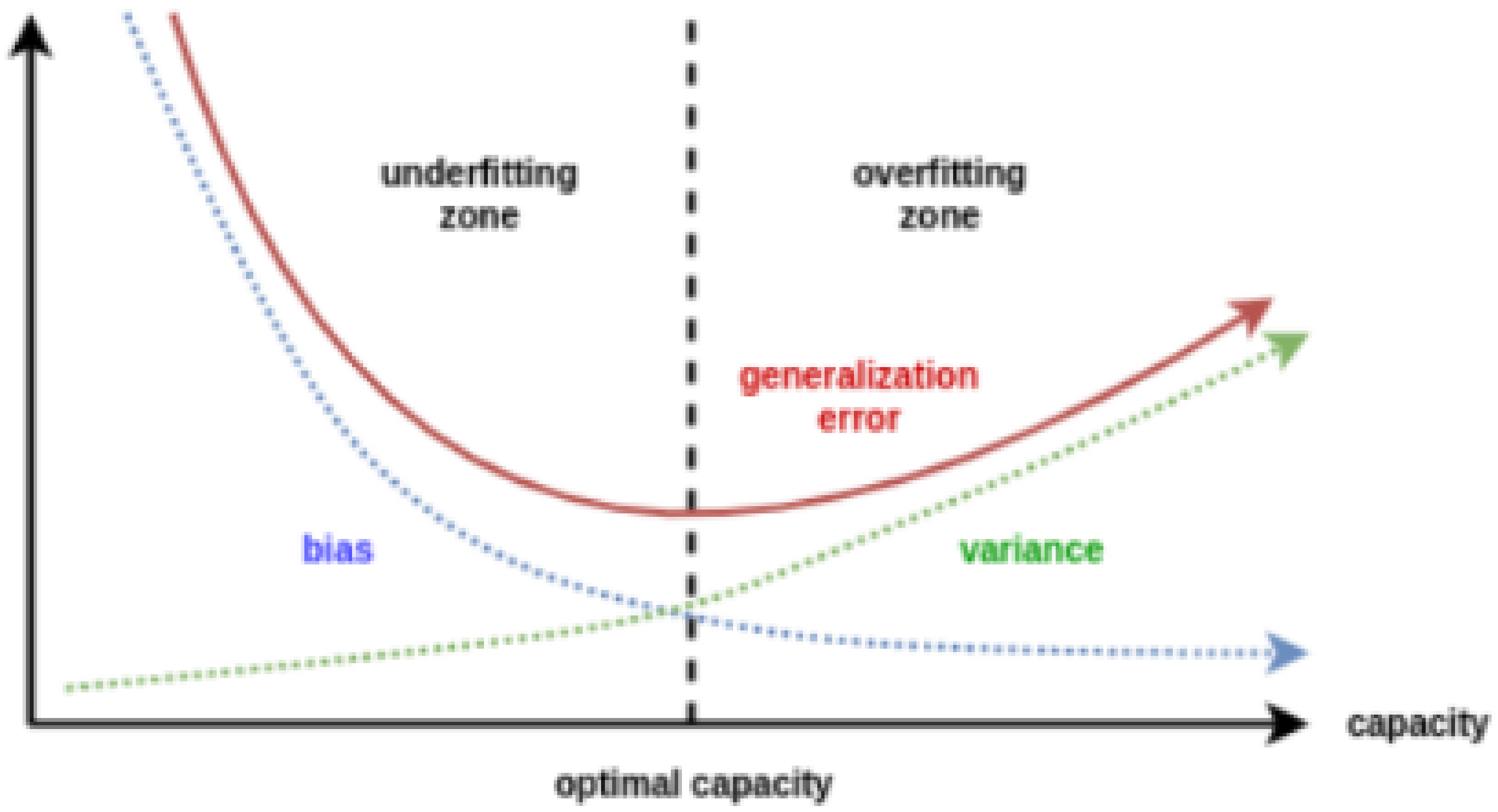
OVERFITTING과 UNDERFITTING



BIAS와 VARIANCE



OPTIMAL CAPACITY



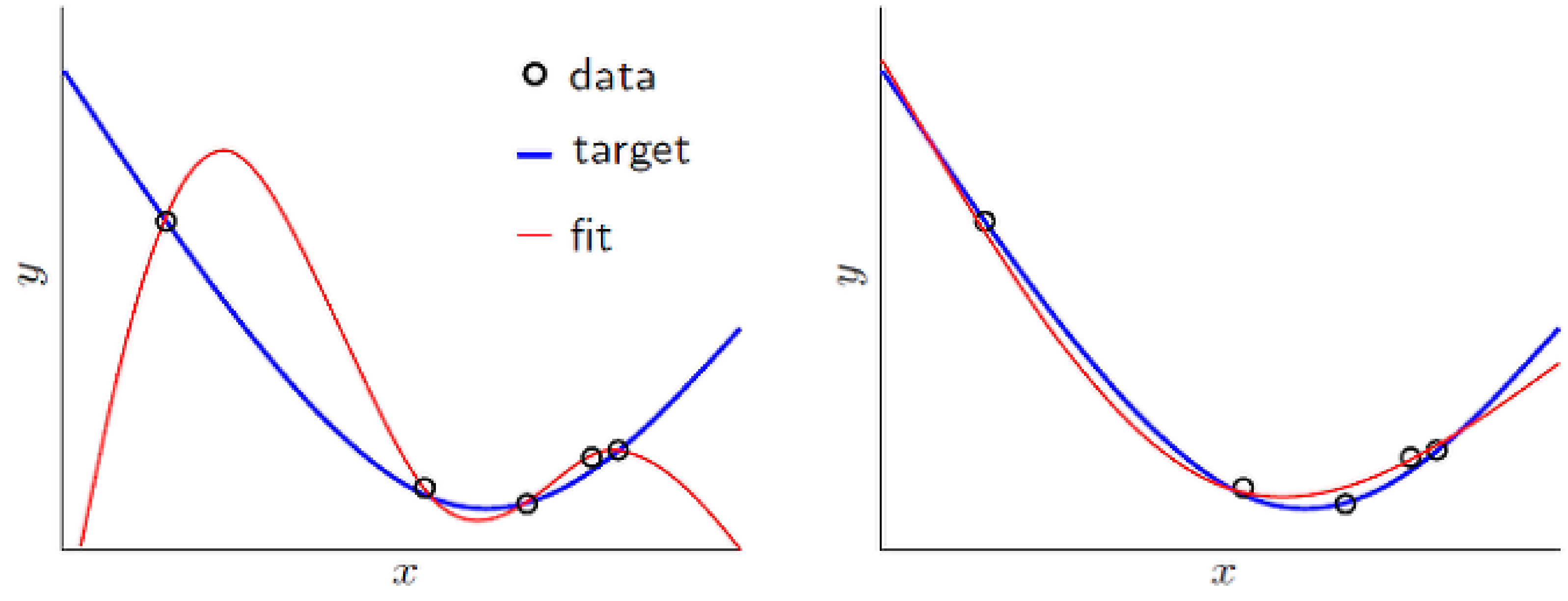
OVERFITTING PREVENTION

Naive Approach

- Increase Sample
- Reduce Complexity

Regularization
Dropout
Batch Normalization

REGULARIZATION



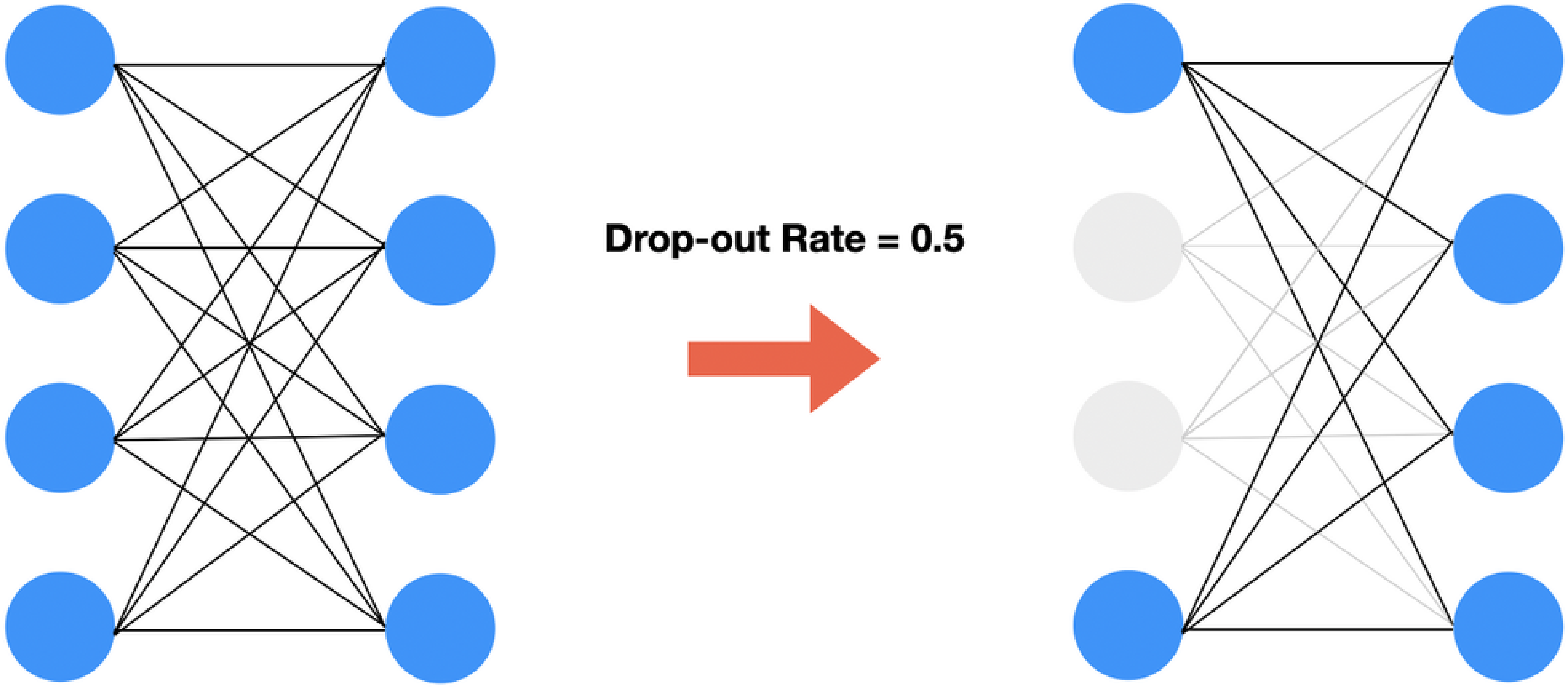
(a) without regularization (b) with regularization

$$C = C_0 + \frac{\lambda}{n} \sum_w |w|$$

$$C = C_0 + \frac{\lambda}{2n} \sum_w w^2$$

$$\frac{\sum_{i=1}^n (y_i - x_i^J \hat{\beta})^2}{2n} + \lambda \left(\frac{1-\alpha}{2} \sum_{j=1}^m \hat{\beta}_j^2 + \alpha \sum_{j=1}^m |\hat{\beta}_j| \right)$$

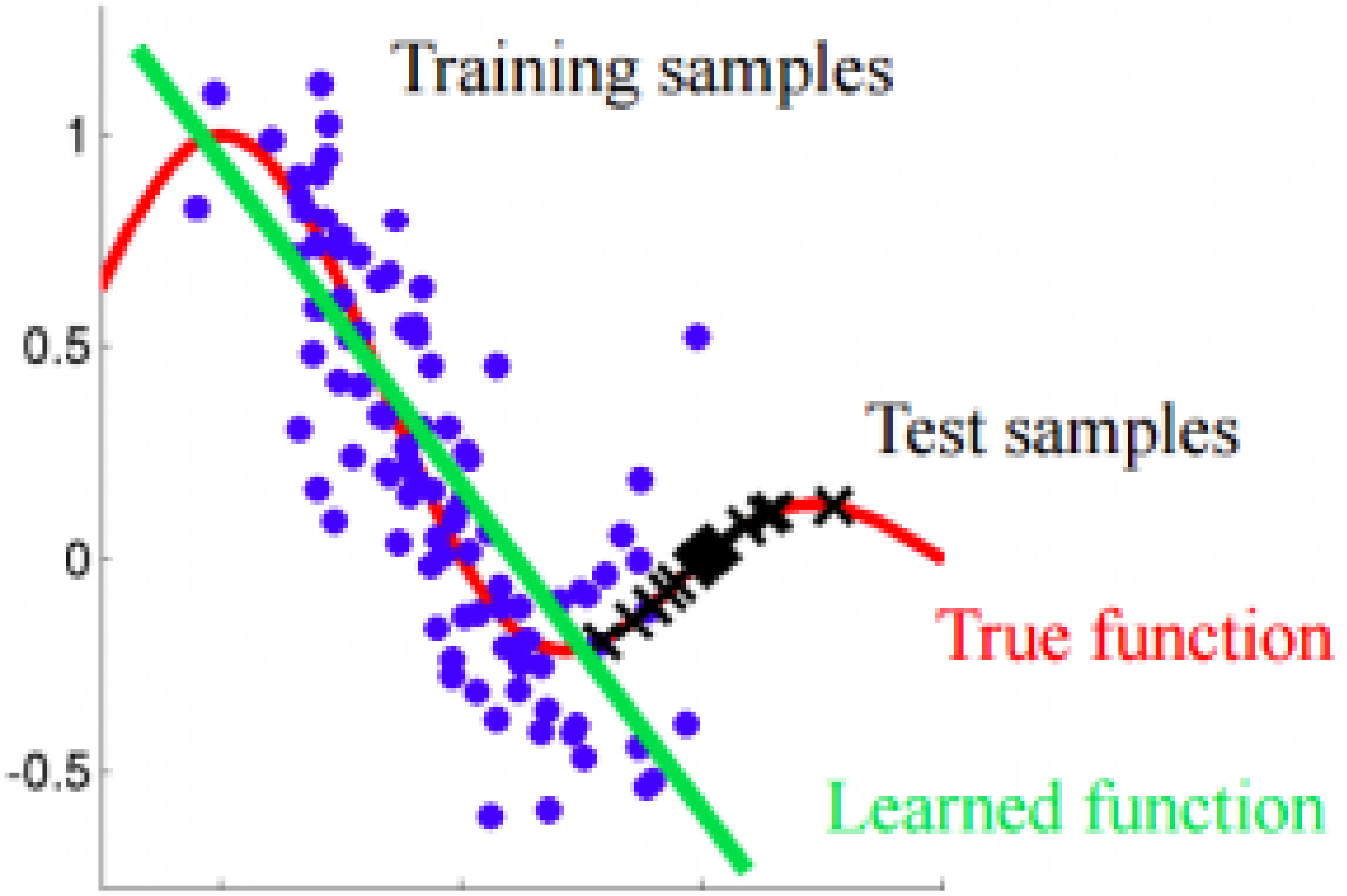
DROPOUT



https://heytech.tistory.com/

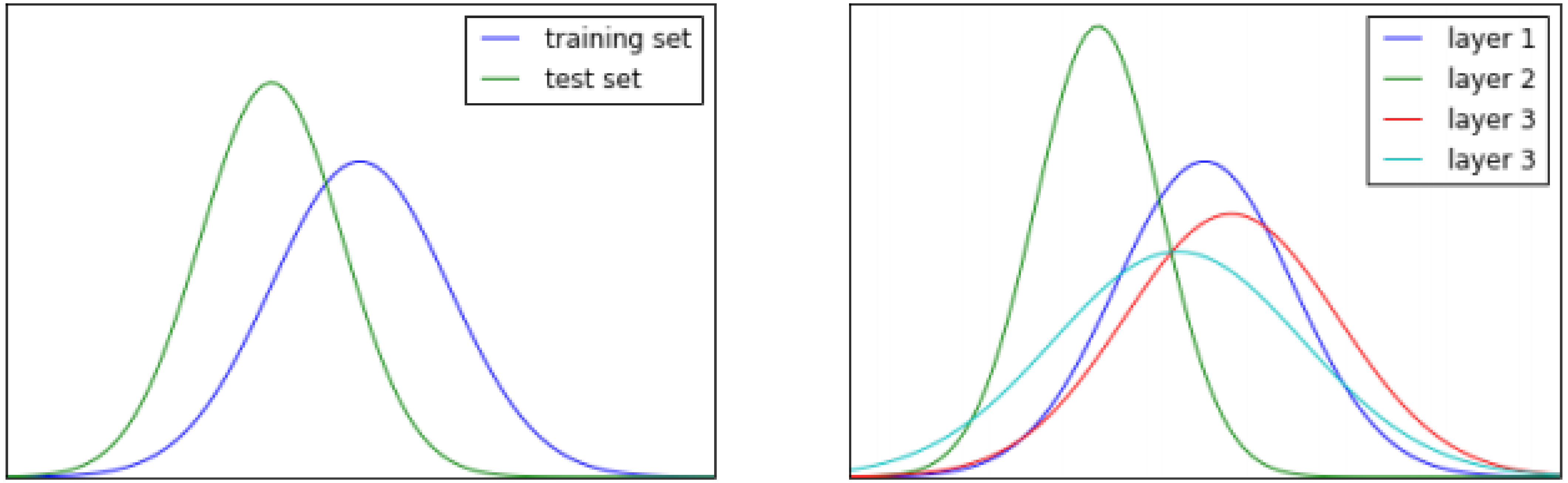
BATCH NORMALIZATION

Covariance Shift



BATCH NORMALIZATION

Internal Covariate Shift



(a) Covariate shift (b) Internal covariate shift

Figure 3.1: Covariate shift vs. internal covariate shift

BATCH NORMALIZATION

Batch Normalization

$$\hat{x}^k = \frac{x^k - E[x^k]}{\sqrt{Var[x^k]}}$$

$$y^k = \gamma^k \hat{x} + \beta^k$$

Thank you