Implicit regularization in Heavy-ball momentum accelerated stochastic gradient descent

Avrajit Ghosh[1], He Lyu[1], Xitong Zhang, Rongrong Wang

Dept. of Computational Math., Science and Engineering, Michigan State University, USA.

Dept of Mathematics, Michigan State University, USA.

April 21, 2023

ICLR
International Conference On Learning Representations

[1] denotes equal contribution

Background

- Momentum methods (SGD+M) empirically outperforms traditional stochastic gradient descent (SGD). [2]

[2] Sutskever, I., Martens, J., Dahl, G., and Hinton, G. On the importance of initialization and momentum in deep learning. In International conference on machine learning, pp.1139–1147. PMLR, 2013

[3] Defazio, A. Understanding the role of momentum in non-convex optimization: Practical insights from a lyapunov analysis

Background

- Momentum methods (SGD+M) empirically outperforms traditional stochastic gradient descent (SGD). [2]

- From optimization perspective, convergence of momentum [3] is well studied.

[2] Sutskever, I., Martens, J., Dahl, G., and Hinton, G. On the importance of initialization and momentum in deep learning. In International conference on machine learning, pp.1139–1147. PMLR, 2013

[3] Defazio, A. Understanding the role of momentum in non-convex optimization: Practical insights from a lyapunov analysis

Background

- Momentum methods (SGD+M) empirically outperforms traditional stochastic gradient descent (SGD). [2]
- From optimization perspective, convergence of momentum [3] is well studied.
- No theoretical answer on how momentum helps generalization for deep neural networks.

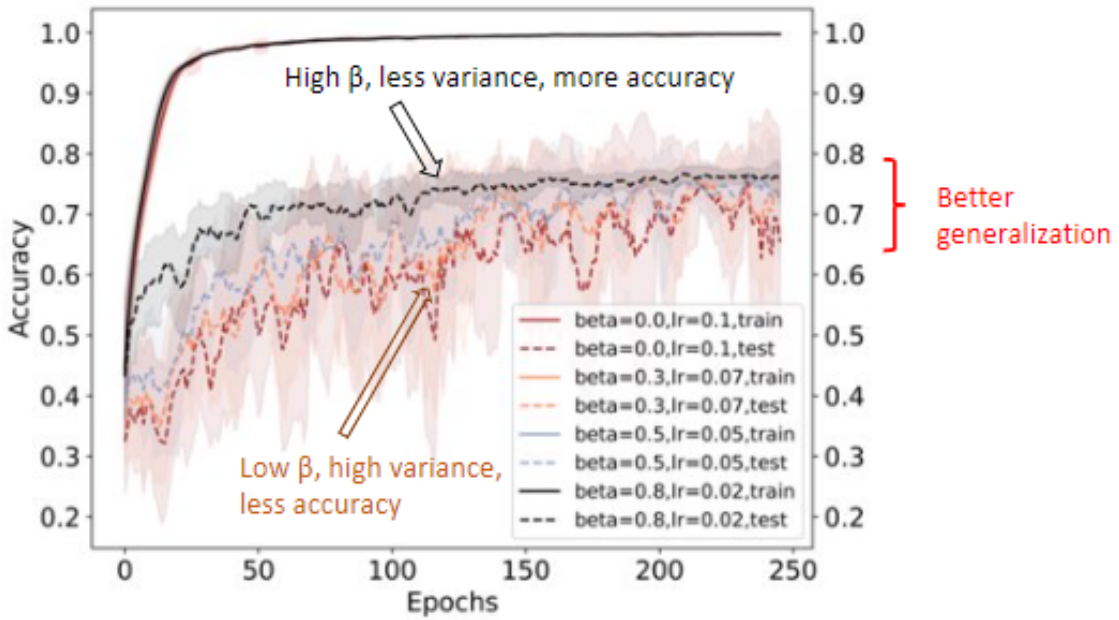
[2] Sutskever, I., Martens, J., Dahl, G., and Hinton, G. On the importance of initialization and momentum in deep learning. In International conference on machine learning, pp.1139–1147. PMLR, 2013

[3] Defazio, A. Understanding the role of momentum in non-convex optimization: Practical insights from a lyapunov analysis

CIFAR-10 classification with momentum

Heavy ball momentum update

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - h\nabla E(\boldsymbol{x}^k) + \beta(\boldsymbol{x}^k - \boldsymbol{x}^{k-1}) \quad \forall k = 1, 2, ..., n$$



Some prevalent observations in deep learning

1. SGD/GD: Test accuracy improves with higher learning rate (h) ?

   $\Longrightarrow$ Implicit Gradient Regularization[4] [5]

[4] Barrett and Dherin, "Implicit gradient regularization", ICLR-21

[5] Smith et al, "On the Origin of Implicit Regularization in Stochastic Gradient Descent", ICLR-21

Some prevalent observations in deep learning

1. SGD/GD: Test accuracy improves with higher learning rate (h) ?

   $\Longrightarrow$ Implicit Gradient Regularization[4] [5]

2. (SGD+M): $\beta \uparrow$, variation in test-accuracy $\downarrow$ across epochs .

[4] Barrett and Dherin, "Implicit gradient regularization", ICLR-21

[5] Smith et al, "On the Origin of Implicit Regularization in Stochastic Gradient Descent", ICLR-21

Some prevalent observations in deep learning

1. SGD/GD: Test accuracy improves with higher learning rate (h) ?

   $\Longrightarrow$ Implicit Gradient Regularization[4] [5]

2. (SGD+M): $\beta \uparrow$, variation in test-accuracy $\downarrow$ across epochs .

3. (SGD+M): $\beta \uparrow$, test accuracy $\uparrow$.

[4] Barrett and Dherin, "Implicit gradient regularization", ICLR-21

[5] Smith et al, "On the Origin of Implicit Regularization in Stochastic Gradient Descent", ICLR-21

Some prevalent observations in deep learning

1. SGD/GD: Test accuracy improves with higher learning rate (h) ?

   $\Longrightarrow$ Implicit Gradient Regularization[4] [5]

2. (SGD+M): $\beta \uparrow$, variation in test-accuracy $\downarrow$ across epochs .

3. (SGD+M): $\beta \uparrow$, test accuracy $\uparrow$.

[4] Barrett and Dherin, "Implicit gradient regularization", ICLR-21

[5] Smith et al, "On the Origin of Implicit Regularization in Stochastic Gradient Descent", ICLR-21

Some prevalent observations in deep learning

1. SGD/GD: Test accuracy improves with higher learning rate (h) ?

   $\Longrightarrow$ Implicit Gradient Regularization[4] [5]

2. (SGD+M): $\beta \uparrow$, variation in test-accuracy $\downarrow$ across epochs .

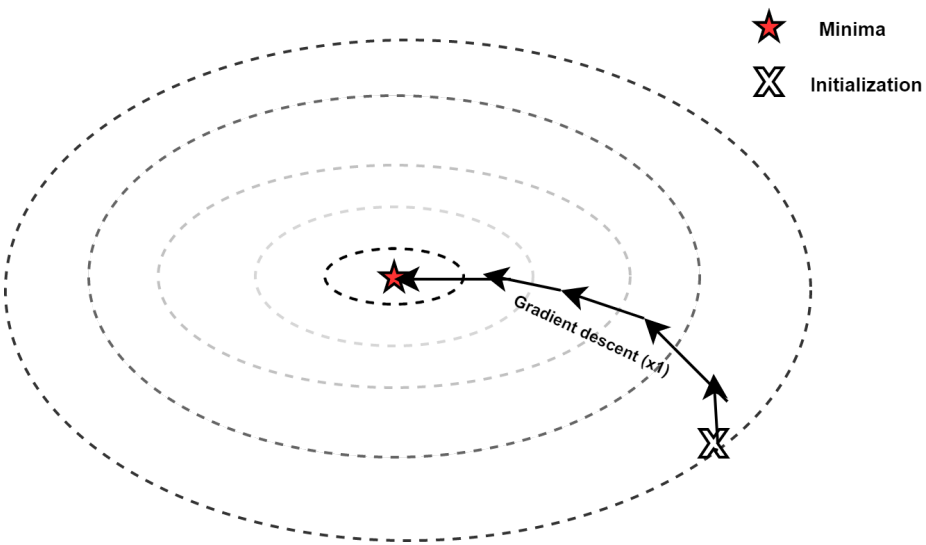
3. (SGD+M): $\beta \uparrow$, test accuracy $\uparrow$.

   Contribution: We study why generalization improves and variance reduces for (SGD+M) with increasing β.

[4] Barrett and Dherin, "Implicit gradient regularization", ICLR-21

[5] Smith et al, "On the Origin of Implicit Regularization in Stochastic Gradient Descent", ICLR-21

Implicit Gradient Regularization[6]

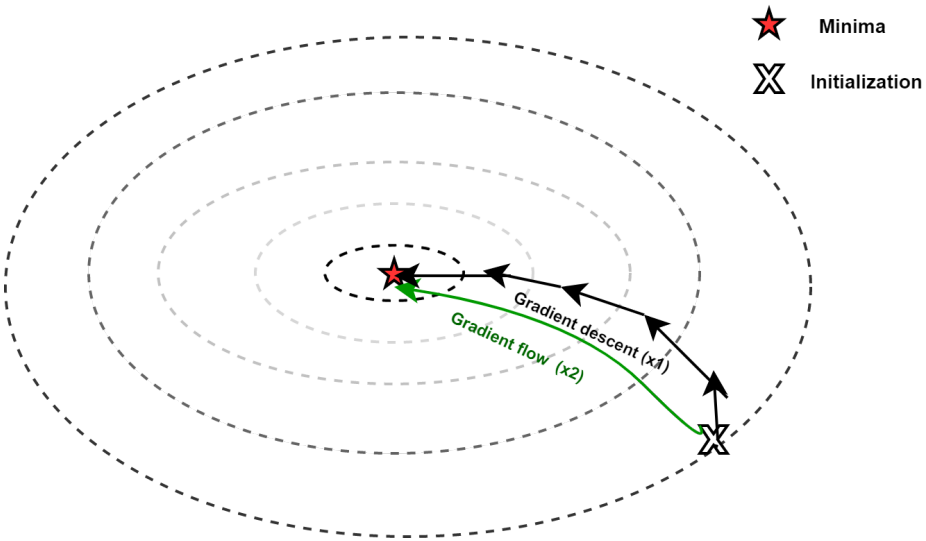


- Gradient Descent:
  $\boldsymbol{x}_1^{k+1} = \boldsymbol{x}_1^k - h\nabla E(\boldsymbol{x}_1^k)$

[6] Barrett and Dherin, "Implicit gradient regularization", ICLR-21

Implicit Gradient Regularization[6]

$\boldsymbol{x}_1$ is $O(h)$ close to $\boldsymbol{x}_2$.

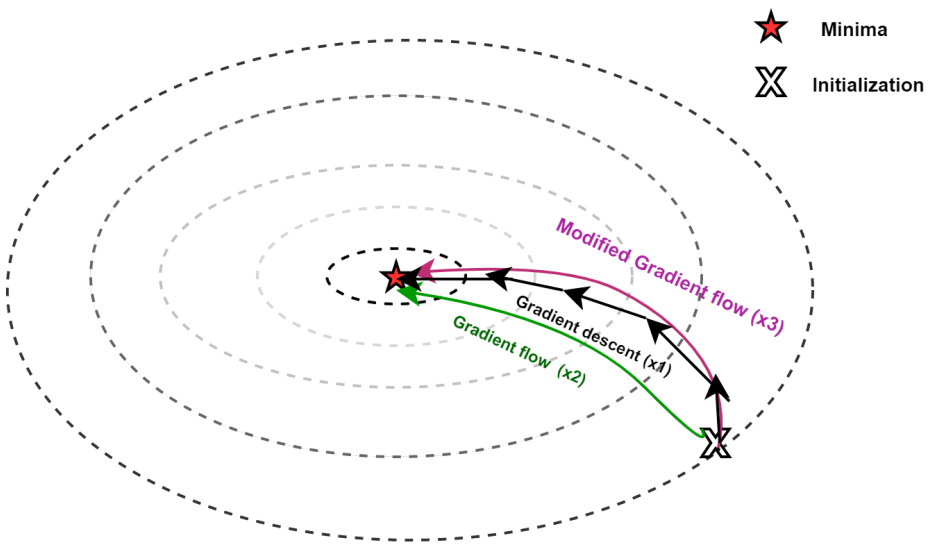


- Gradient Descent:
  $\boldsymbol{x}_1^{k+1} = \boldsymbol{x}_1^k - h\nabla E(\boldsymbol{x}_1^k)$
- Gradient Flow ($h \longrightarrow 0$):
  $\boldsymbol{x}_2'(t) = -\nabla E(\boldsymbol{x}_2(t))$

[6] Barrett and Dherin, "Implicit gradient regularization", ICLR-21

Implicit Gradient Regularization[6]

x_1 is $O(h)$ close to x_2.



x_1 is $O(h^2)$ close to x_3.

- Gradient Descent:
  $x_1^{k+1} = x_1^k - h\nabla E(x_1^k)$
- Gradient Flow ($h \longrightarrow 0$):
  $x_2'(t) = -\nabla E(x_2(t))$
- Modified flow:
  $x_3'(t) = -\nabla(E(x_3) + \underbrace{\frac{h}{4}\|\nabla E(x_3)\|^2}_{\text{IGR}})$

[6] Barrett and Dherin, "Implicit gradient regularization", ICLR-21

Can we study Heavy-ball momentum gradient descent trajectory using modified continuous flow ?

IGR for (GD+M)

Heavy ball momentum update

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - h\nabla E(\boldsymbol{x}^k) + \beta(\boldsymbol{x}^k - \boldsymbol{x}^{k-1}) \quad \forall k = 1, 2, ..., n$$

[7] Kovachki,Stuart, "Continuous time analysis of momentum methods". JMLR 2021

IGR for (GD+M)

Heavy ball momentum update

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - h\nabla E(\boldsymbol{x}^k) + \beta(\boldsymbol{x}^k - \boldsymbol{x}^{k-1}) \quad \forall k = 1, 2, ..., n$$

is globally $\underline{O(h)}$ close to the ODE[7]

$$\widetilde{\boldsymbol{x}}'(t) = -\frac{1}{1-\beta}\nabla E(\widetilde{\boldsymbol{x}}(t)), \quad t \in [0, T] \tag{1}$$

[7] Kovachki,Stuart, "Continuous time analysis of momentum methods". JMLR 2021

IGR for (GD+M)

Heavy ball momentum update

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - h\nabla E(\boldsymbol{x}^k) + \beta(\boldsymbol{x}^k - \boldsymbol{x}^{k-1}) \quad \forall k = 1, 2, ..., n$$

is globally $\underline{O(h)}$ close to the ODE[7]

$$\widetilde{\boldsymbol{x}}'(t) = -\frac{1}{1-\beta}\nabla E(\widetilde{\boldsymbol{x}}(t)), \quad t \in [0, T] \tag{1}$$

but globally $\underline{O(h^2)}$ closer to the ODE

$$\widetilde{\boldsymbol{x}}'(t) = -\frac{1}{1-\beta}\nabla(E(\widetilde{\boldsymbol{x}}(t)) + \underbrace{\frac{(1+\beta)h}{4(1-\beta)^2}\|\nabla E(\widetilde{\boldsymbol{x}}(t))\|_2^2}_{IGR-M}), \quad t \in [0, T] \tag{2}$$

[7] Kovachki,Stuart, "Continuous time analysis of momentum methods". JMLR 2021

Gradient Descent[4]

$$x'(t) = -\nabla \left(E(x(t)) + \frac{h}{4} \|\nabla E(x(t))\|^2 \right)$$

$$x'(t) = -\nabla \left(\frac{1}{1-\beta} E(x(t)) + \frac{(1+\beta)h}{4(1-\beta)^3} \|\nabla E(x(t))\|_2^2 \right)$$

Gradient Descent + Momentum

4) Barrett and Dherin, "Implicit gradient regularization", ICLR-21
5) Smith et al, "On the Origin of Implicit Regularization in Stochastic Gradient Descent", ICLR-21

Gradient Descent[4]

$$x'(t) = -\nabla \left(E(x(t)) + \frac{h}{4} \|\nabla E(x(t))\|^2 \right)$$

$\frac{1}{1-\beta} \times$ force

$\frac{(1+\beta)}{(1-\beta)^3} \times$ IGR

$$x'(t) = -\nabla \left(\frac{1}{1-\beta} E(x(t)) + \frac{(1+\beta)h}{4(1-\beta)^3} \|\nabla E(x(t))\|_2^2 \right)$$

Gradient Descent + Momentum

4) Barrett and Dherin, "Implicit gradient regularization", ICLR-21
5) Smith et al, "On the Origin of Implicit Regularization in Stochastic Gradient Descent", ICLR-21

Gradient Descent[4]

$$x'(t) = -\nabla \left(E(x(t)) + \frac{h}{4} \|\nabla E(x(t))\|^2 \right)$$

$\frac{1}{1-\beta} \times$ force

$\frac{(1+\beta)}{(1-\beta)^3} \times$ IGR

$$x'(t) = -\nabla \left(\frac{1}{1-\beta} E(x(t)) + \frac{(1+\beta)h}{4(1-\beta)^3} \|\nabla E(x(t))\|_2^2 \right)$$

Faster convergence

Better generalization

Gradient Descent + Momentum

4) Barrett and Dherin, "Implicit gradient regularization", ICLR-21
5) Smith et al, "On the Origin of Implicit Regularization in Stochastic Gradient Descent", ICLR-21

IGR $\left(\frac{h}{4}\|\nabla E(\boldsymbol{x})\|^2\right)$ promotes flatness

- Penalizes norm of the gradient.

[8] Foret et al, "Sharpness Aware Minimization", ICLR-21

IGR $(\frac{h}{4}\|\nabla E(x)\|^2)$ promotes flatness

- Penalizes norm of the gradient.
- IGR penalizes "sharpness" $\|\nabla^2 E\|_2$.

[8] Foret et al, "Sharpness Aware Minimization", ICLR-21

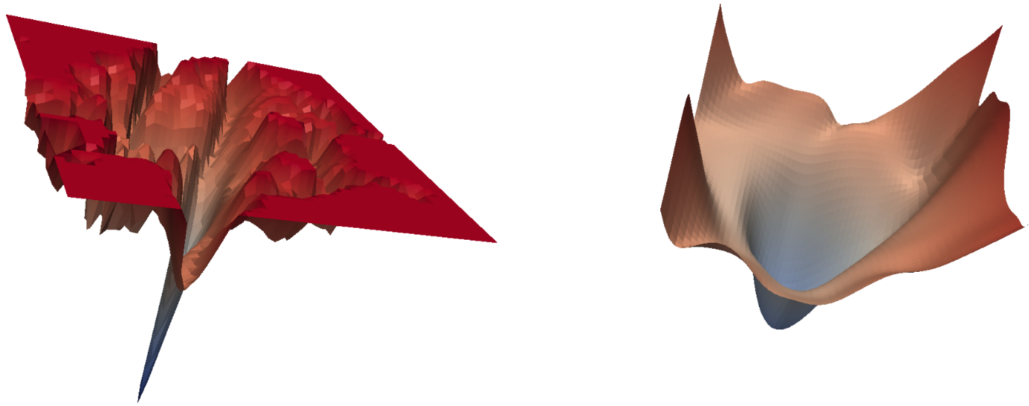
IGR $(\frac{h}{4}\|\nabla E(x)\|^2)$ promotes flatness

- Penalizes norm of the gradient.
- IGR penalizes "sharpness" $\|\nabla^2 E\|_2$.
- Smaller sharpness $\rightarrow$ flatter minima.

[8] Foret et al, "Sharpness Aware Minimization", ICLR-21

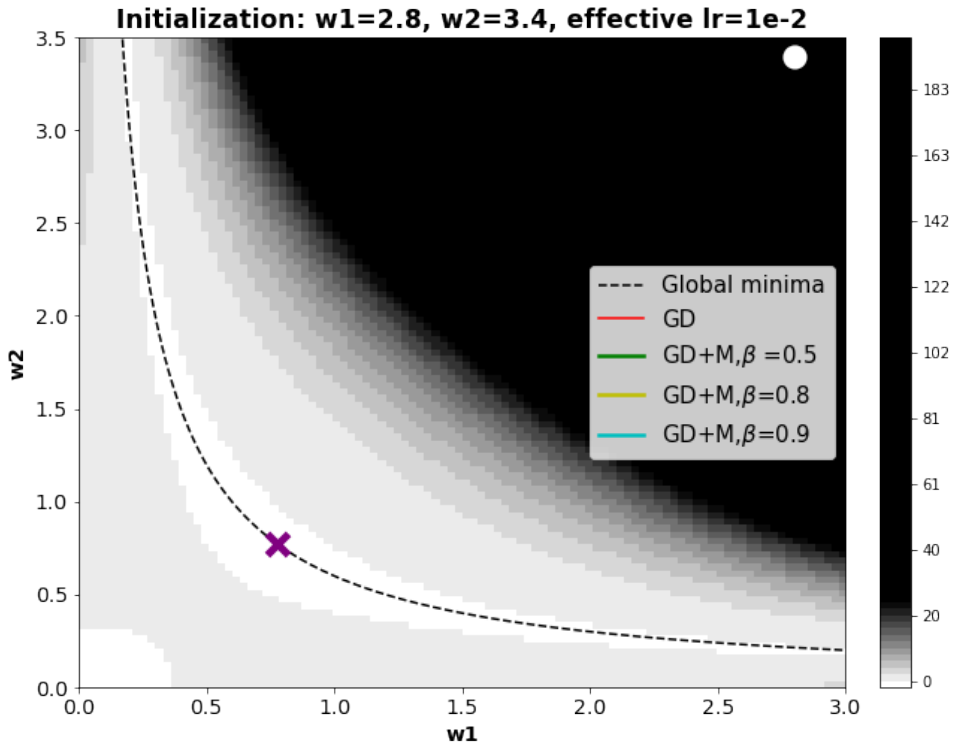
IGR $(\frac{h}{4}\|\nabla E(x)\|^2)$ promotes flatness

- Penalizes norm of the gradient.
- IGR penalizes "sharpness" $\|\nabla^2 E\|_2$.
- Smaller sharpness $\rightarrow$ flatter minima.
- Flatter minima $\rightarrow$ better generalization[8].



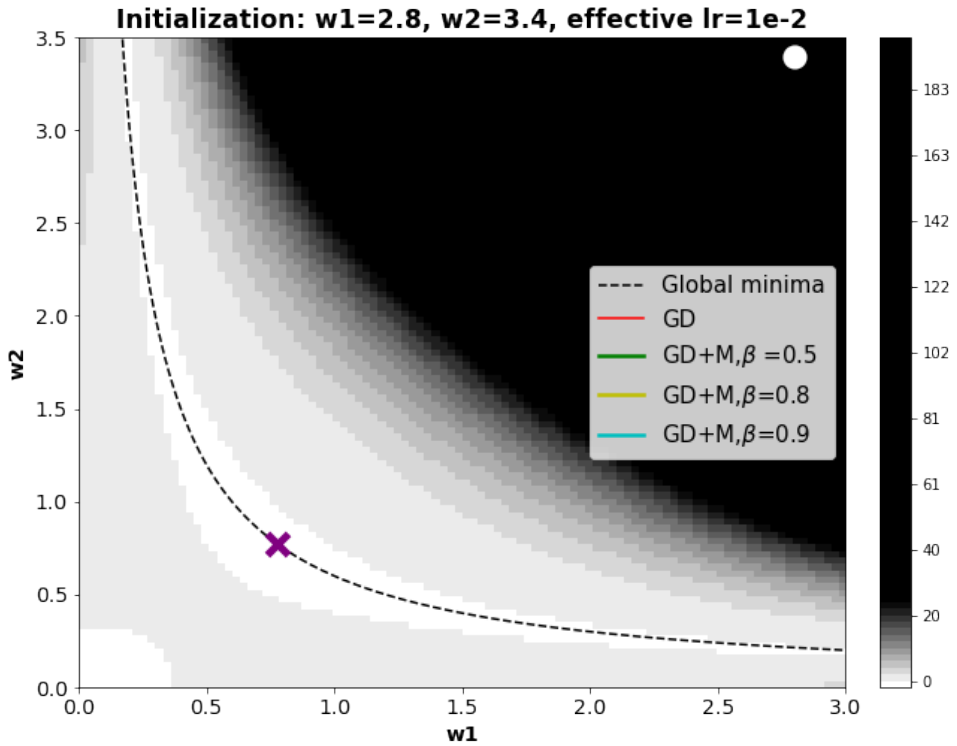
[8] Foret et al, "Sharpness Aware Minimization", ICLR-21

$\min_{w_1,w_2} E(w_1, w_2) = \frac{1}{2}(y - w_1 w_2 x)^2$



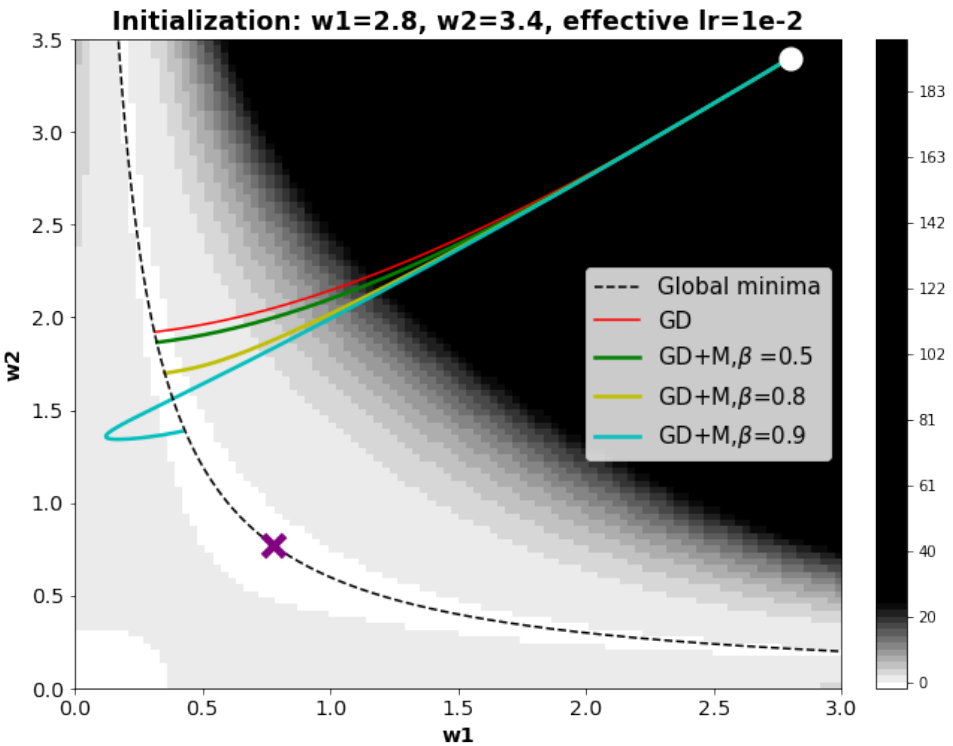
- Global minima: $w_1 w_2 = \frac{y}{x}$

$\min_{w_1,w_2} E(w_1, w_2) = \frac{1}{2}(y - w_1 w_2 x)^2$



- Global minima: $w_1 w_2 = \frac{y}{x}$
- Which minima just (GD+M) prefer?

$min_{w_1,w_2} E(w_1, w_2) = \frac{1}{2}(y - w_1 w_2 x)^2$



- Global minima: $w_1 w_2 = \frac{y}{x}$
- Which minima just (GD+M) prefer?

Larger β finds flatter minima. Recall IGR magnified by $\frac{1+\beta}{1-\beta}$.

How does this implicit regularization change in Stochastic Gradient Descent with momentum ?

Modified flow for SGD+M

Heavy-ball momentum SGD updates

$$\begin{cases} \boldsymbol{x}^{k+1} = \boldsymbol{x}^k - h\nabla E_k(\boldsymbol{x}^k) + \beta(\boldsymbol{x}^k - \boldsymbol{x}^{k-1}) \quad k = 1, 2, ..., n \\ \boldsymbol{x}^1 = \boldsymbol{x}^0 - h\nabla E_0(\boldsymbol{x}^0) \\ \boldsymbol{x}^0 = \boldsymbol{x}^{-1} = \boldsymbol{0} \end{cases} \tag{3}$$

Modified flow for SGD+M

Heavy-ball momentum SGD updates

$$\begin{cases} \boldsymbol{x}^{k+1} = \boldsymbol{x}^k - h\nabla E_k(\boldsymbol{x}^k) + \beta(\boldsymbol{x}^k - \boldsymbol{x}^{k-1}) \quad k = 1, 2, ..., n \\ \boldsymbol{x}^1 = \boldsymbol{x}^0 - h\nabla E_0(\boldsymbol{x}^0) \\ \boldsymbol{x}^0 = \boldsymbol{x}^{-1} = \boldsymbol{0} \end{cases} \tag{3}$$

is $O(h^2)$ close to the modified loss trajectory

Modified flow for SGD+M

Heavy-ball momentum SGD updates

$$\begin{cases} \boldsymbol{x}^{k+1} = \boldsymbol{x}^k - h\nabla E_k(\boldsymbol{x}^k) + \beta(\boldsymbol{x}^k - \boldsymbol{x}^{k-1}) \quad k = 1, 2, ..., n \\ \boldsymbol{x}^1 = \boldsymbol{x}^0 - h\nabla E_0(\boldsymbol{x}^0) \\ \boldsymbol{x}^0 = \boldsymbol{x}^{-1} = \boldsymbol{0} \end{cases} \tag{3}$$

is $O(h^2)$ close to the modified loss trajectory

$$\begin{cases} \boldsymbol{x}'(t) = -\nabla \left(\underbrace{G_k(\boldsymbol{x}(t))}_{force} + \underbrace{\frac{h}{4} \left(\|\nabla G_k(\boldsymbol{x}(t))\|_2^2 + 2\sum_{r=0}^{k-1} \beta^{k-r} \|\nabla G_r(\boldsymbol{x}(t))\|_2^2 \right)}_{IGR_s} \right) \\ \text{for } t_k \le t < t_{k+1}, \qquad \text{where} \quad \underbrace{G_k(\boldsymbol{x}(t)) = \sum_{r=0}^{k} \beta^{k-r} E_r(\boldsymbol{x}(t))}_{\text{Historical gradients}} \end{cases}$$

Stochastic Gradient Descent[5]

$$x'(t) = -\nabla \left(E_k(x(t)) + \frac{h}{4} \|\nabla E_k(x(t))\|_2^2 \right)$$

for $t_k \leq t < t_{k+1}$

$E_k = E$

Gradient Descent[4]

$$x'(t) = -\nabla \left(E(x(t)) + \frac{h}{4} \|\nabla E(x(t))\|^2 \right)$$

$\frac{1}{1-\beta} \times$ force

$\frac{(1+\beta)}{(1-\beta)^3} \times$ IGR

β=0

Stochastic Gradient Descent + Momentum

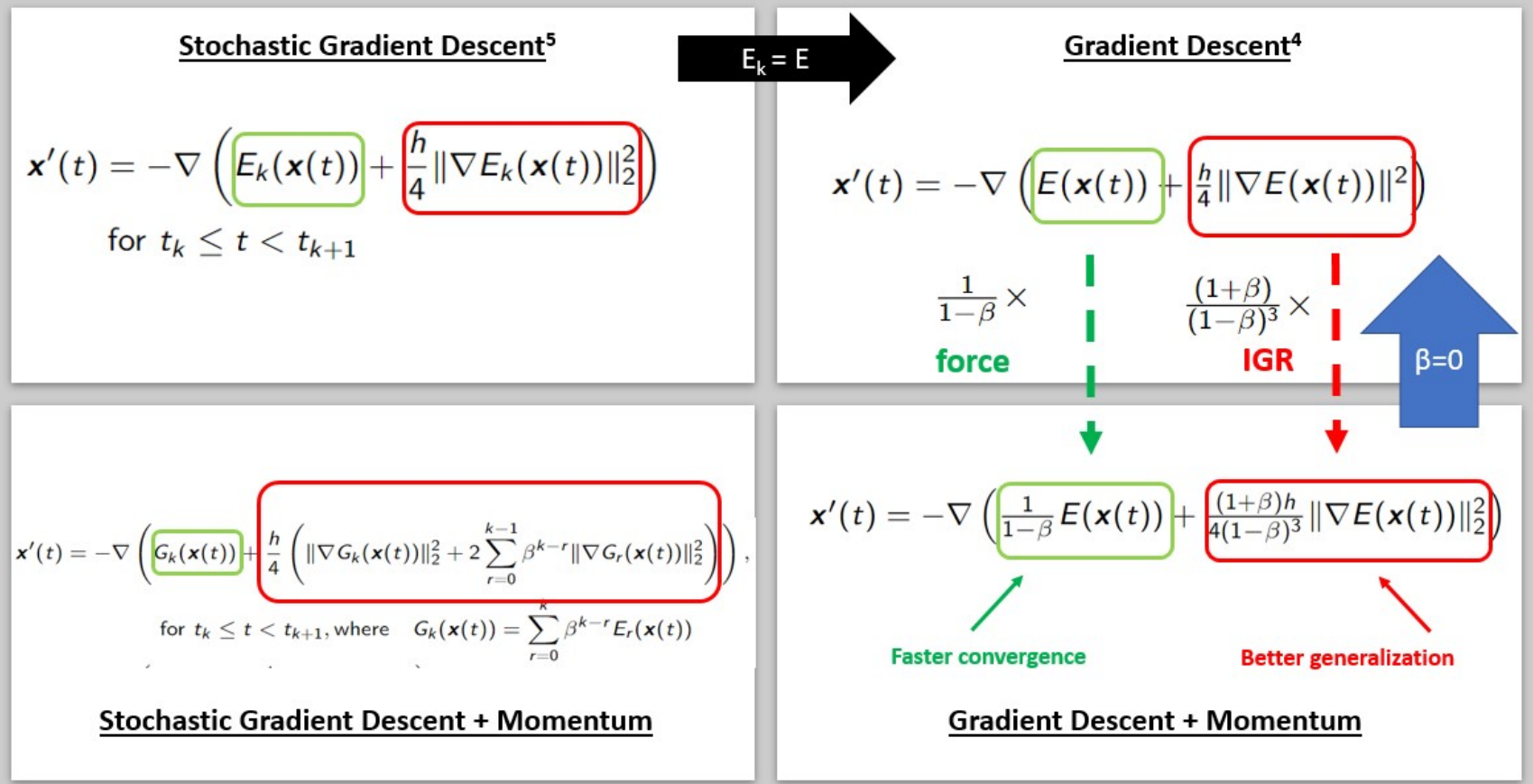
$$x'(t) = -\nabla \left(\frac{1}{1-\beta} E(x(t)) + \frac{(1+\beta)h}{4(1-\beta)^3} \|\nabla E(x(t))\|_2^2 \right)$$

Faster convergence

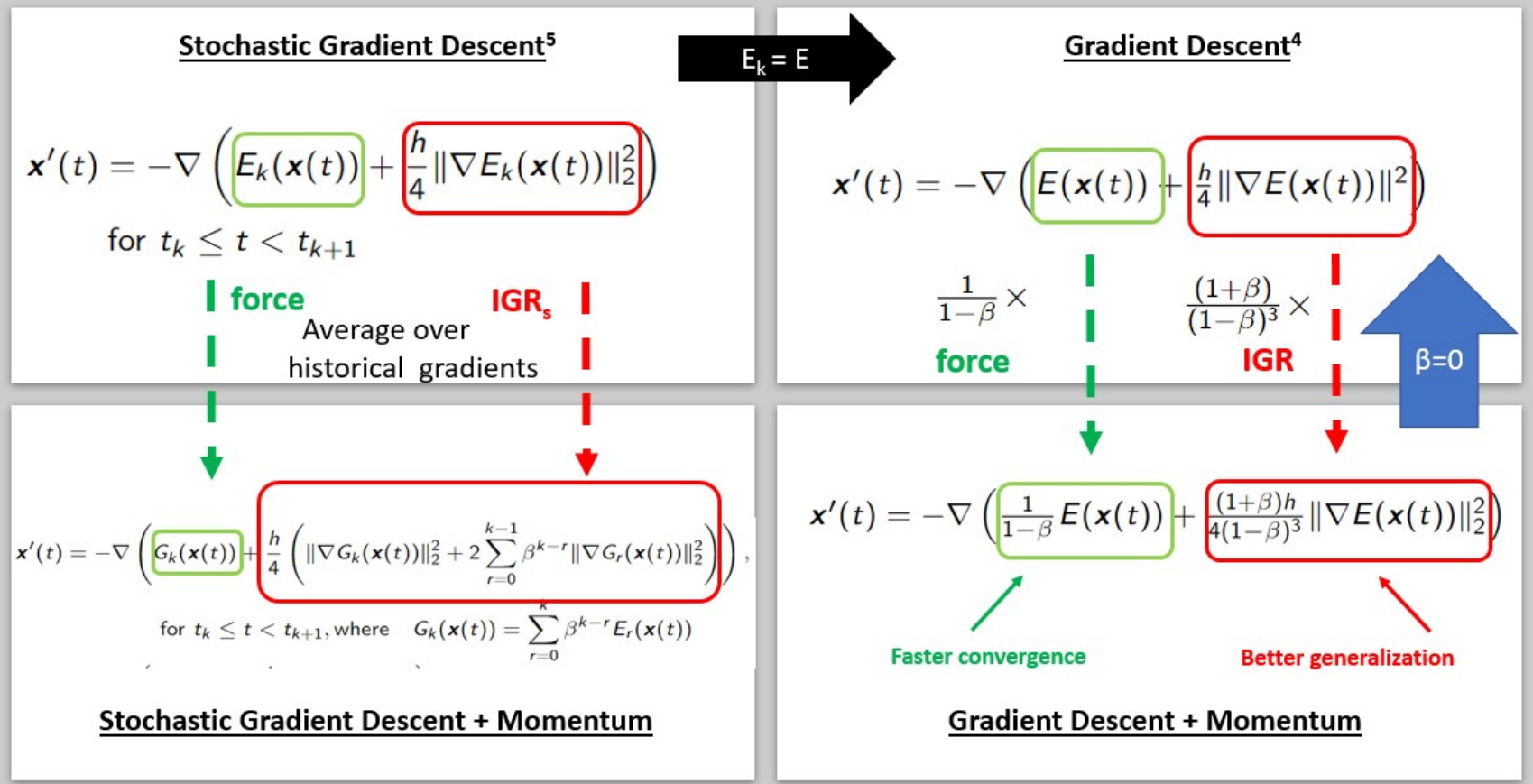
Better generalization

Gradient Descent + Momentum

4) Barrett and Dherin, "Implicit gradient regularization", ICLR-21
5) Smith et al, "On the Origin of Implicit Regularization in Stochastic Gradient Descent", ICLR-21

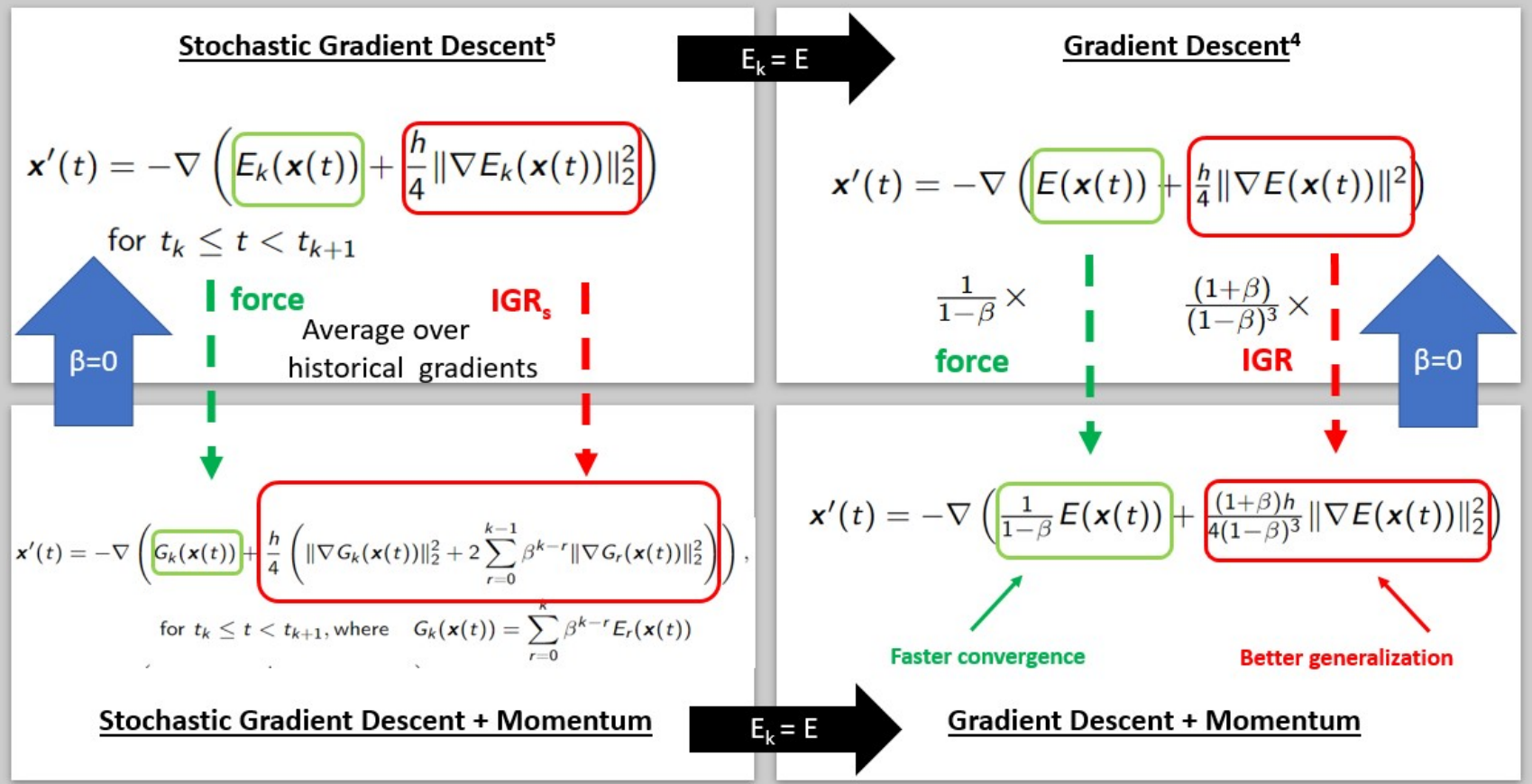


4) Barrett and Dherin, "Implicit gradient regularization", ICLR-21
5) Smith et al, "On the Origin of Implicit Regularization in Stochastic Gradient Descent", ICLR-21



4) Barrett and Dherin, "Implicit gradient regularization", ICLR-21
5) Smith et al, "On the Origin of Implicit Regularization in Stochastic Gradient Descent", ICLR-21

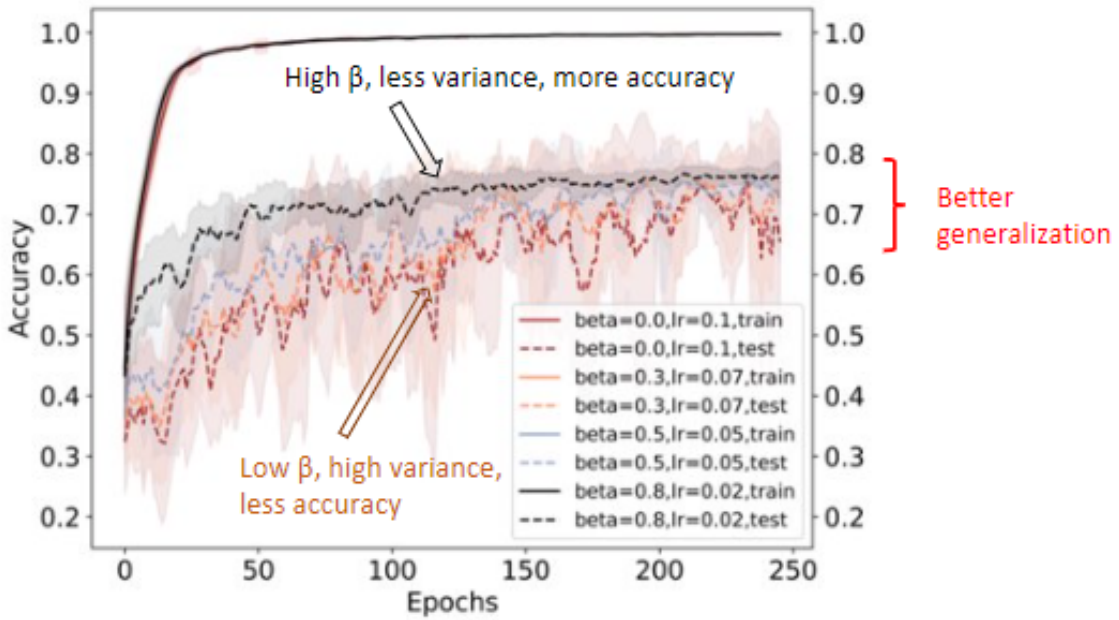
Family of $O(h^2)$ modified flows



4) Barrett and Dherin, "Implicit gradient regularization", ICLR-21
5) Smith et al, "On the Origin of Implicit Regularization in Stochastic Gradient Descent", ICLR-21

Remarks

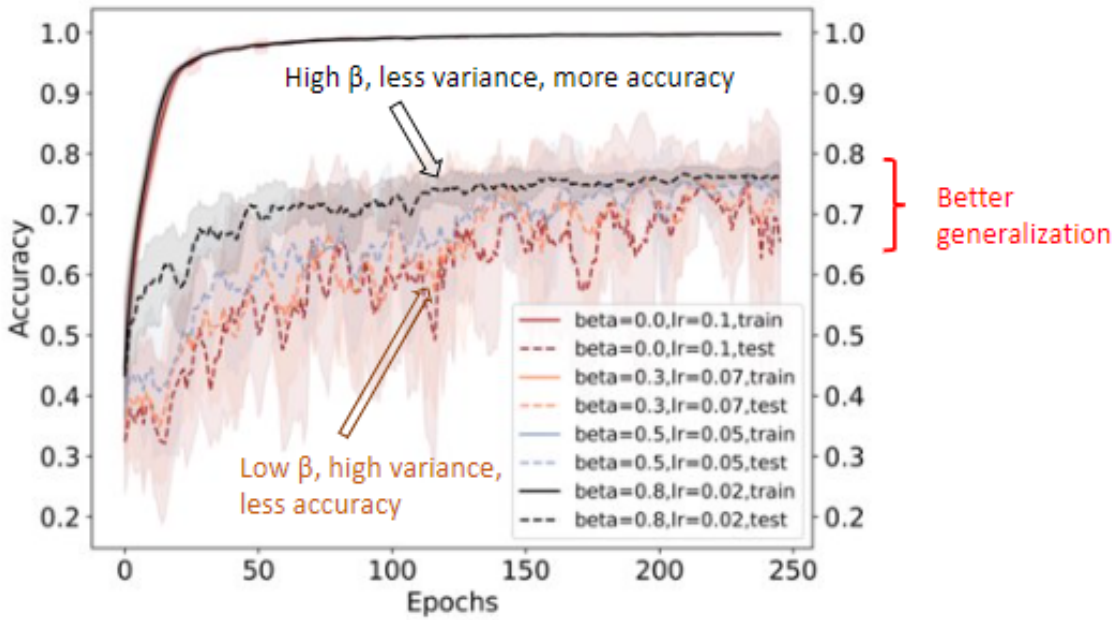
Obs-1: $\beta \uparrow$, variance in test-accuracy $\downarrow$ across epochs .



Remarks

Obs-1: $\beta \uparrow$, variance in test-accuracy $\downarrow$ across epochs .

$\Longrightarrow$ **Variance reduction**: $Cov(G_k(\boldsymbol{x})) = \frac{1-\beta}{1+\beta} Cov(E_k(\boldsymbol{x}))$

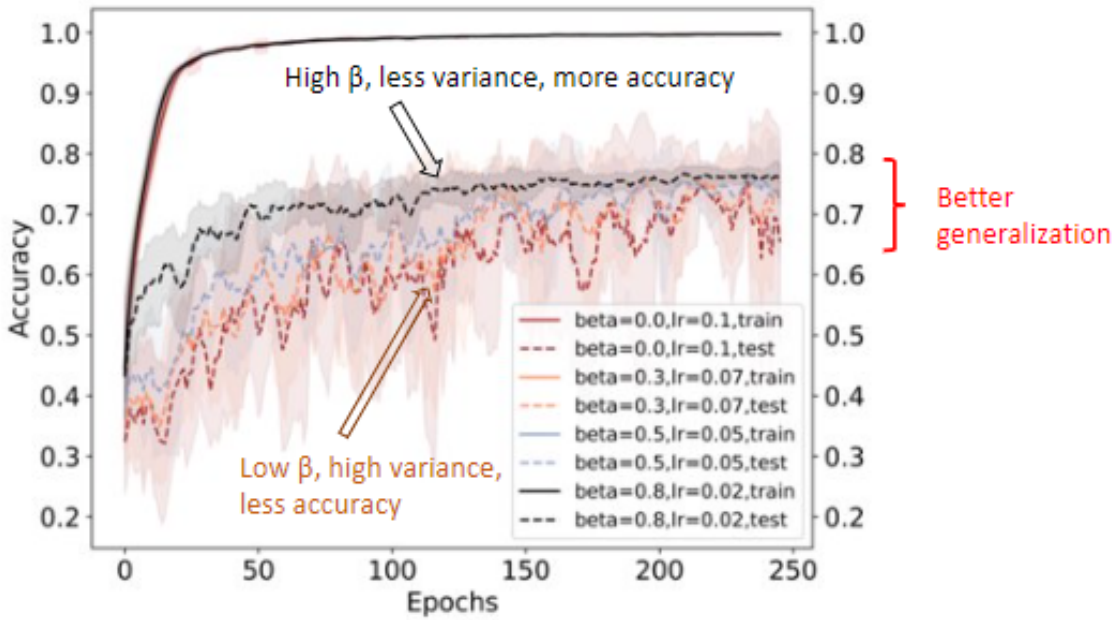


Remarks

Obs-1: $\beta \uparrow$, variance in test-accuracy $\downarrow$ across epochs .

$\Longrightarrow$ **Variance reduction**: $Cov(G_k(\boldsymbol{x})) = \frac{1-\beta}{1+\beta} Cov(E_k(\boldsymbol{x}))$

Obs-2: $\beta \uparrow$, test accuracy $\uparrow$



Remarks

Obs-1: $\beta \uparrow$, variance in test-accuracy $\downarrow$ across epochs .

$\Longrightarrow$ **Variance reduction**: $Cov(G_k(\boldsymbol{x})) = \frac{1-\beta}{1+\beta} Cov(E_k(\boldsymbol{x}))$

Obs-2: $\beta \uparrow$, test accuracy $\uparrow$

$\Longrightarrow$ **Stronger regularization:** $\mathbb{E}(IGRM_s)(\boldsymbol{x}) = \frac{1}{(1-\beta)} \mathbb{E}(IGR_s)(\boldsymbol{x})$