Understanding Adaptive Optimizers with Central Flows

ELLIT Symposium 2026

Alex Damian

Kempner Institute → MIT

Central Flows: A Brief Recap

Central Flows: A Brief Recap

Challenge: optimizers operate at the **edge of stability** and their trajectories are oscillatory

Central Flows: A Brief Recap

Challenge: optimizers operate at the **edge of stability** and their trajectories are oscillatory

Approach: model the **time-averaged** optimization trajectory

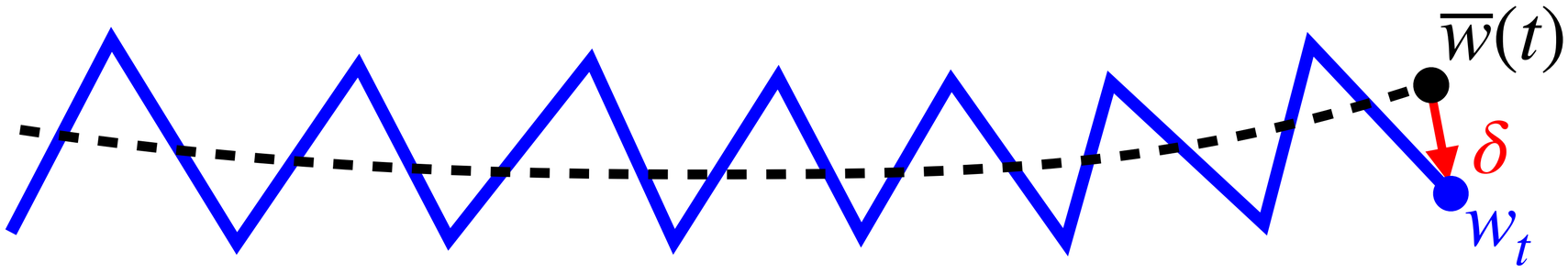
Central Flows: A Brief Recap

Challenge: optimizers operate at the **edge of stability** and their trajectories are oscillatory

Approach: model the **time-averaged** optimization trajectory

- Pretend w_t oscillates around a ***central flow*** $\overline{w}(t)$ with covariance $\Sigma(t)$

$$w_t = \overline{w}(t) + \delta_t \qquad \mathbb{E}[\delta_t] = 0 \qquad \mathbb{E}[\delta_t \delta_t^T] = \Sigma(t)$$



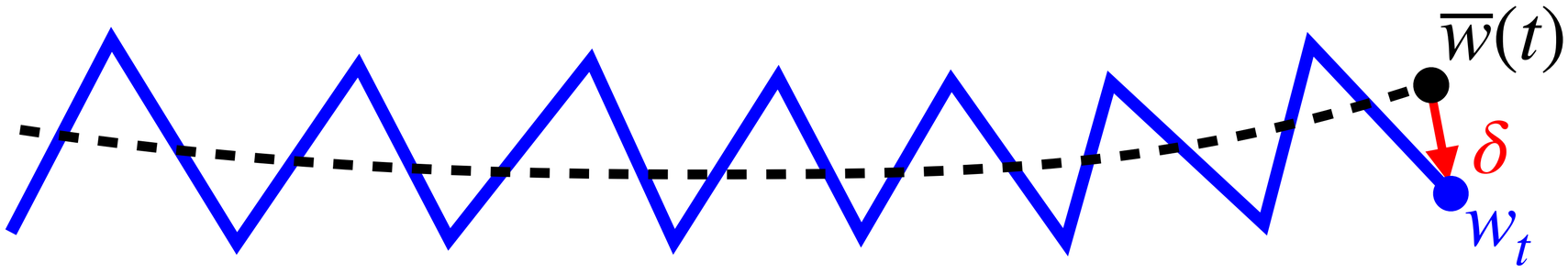
Central Flows: A Brief Recap

Challenge: optimizers operate at the **edge of stability** and their trajectories are oscillatory

Approach: model the **time-averaged** optimization trajectory

- Pretend w_t oscillates around a ***central flow*** $\overline{w}(t)$ with covariance $\Sigma(t)$

$$w_t = \overline{w}(t) + \delta_t \qquad \mathbb{E}[\delta_t] = 0 \qquad \mathbb{E}[\delta_t \delta_t^T] = \Sigma(t)$$



- Derive necessary conditions for $(\overline{w}(t), \Sigma(t))$ which allow us to solve for $\frac{d}{dt}\overline{w}(t)$

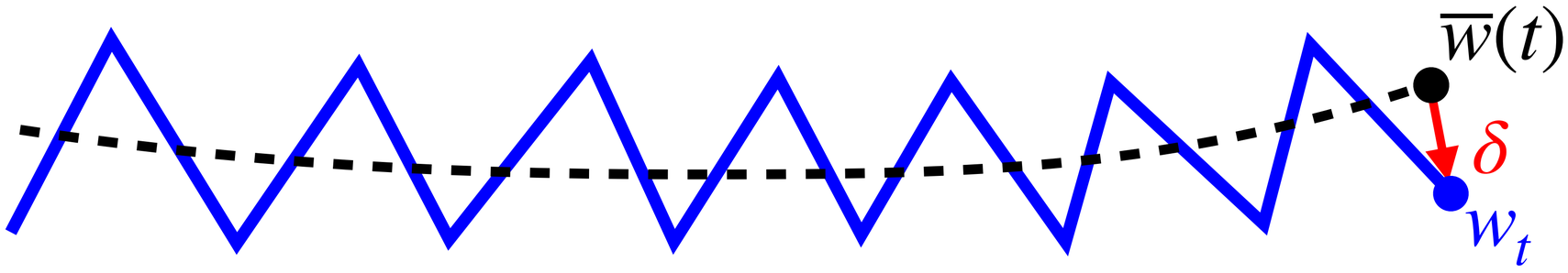
Central Flows: A Brief Recap

Challenge: optimizers operate at the **edge of stability** and their trajectories are oscillatory

Approach: model the **time-averaged** optimization trajectory

- Pretend w_t oscillates around a ***central flow*** $\overline{w}(t)$ with covariance $\Sigma(t)$

$$w_t = \overline{w}(t) + \delta_t \qquad \mathbb{E}[\delta_t] = 0 \qquad \mathbb{E}[\delta_t \delta_t^T] = \Sigma(t)$$



- Derive necessary conditions for $(\overline{w}(t), \Sigma(t))$ which allow us to solve for $\frac{d}{dt}\overline{w}(t)$
- Empirically verify these central flows in realistic deep learning settings

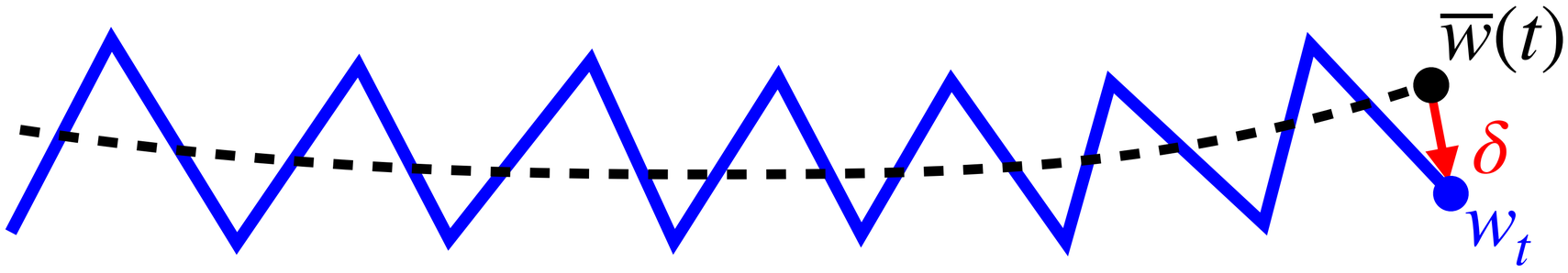
Central Flows: A Brief Recap

Challenge: optimizers operate at the **edge of stability** and their trajectories are oscillatory

Approach: model the **time-averaged** optimization trajectory

- Pretend w_t oscillates around a ***central flow*** $\overline{w}(t)$ with covariance $\Sigma(t)$

$$w_t = \overline{w}(t) + \delta_t \qquad \mathbb{E}[\delta_t] = 0 \qquad \mathbb{E}[\delta_t \delta_t^T] = \Sigma(t)$$



- Derive necessary conditions for $(\overline{w}(t), \Sigma(t))$ which allow us to solve for $\frac{d}{dt}\overline{w}(t)$
- Empirically verify these central flows in realistic deep learning settings
- Interpret the central flow to understand the implicit behavior of the optimizer

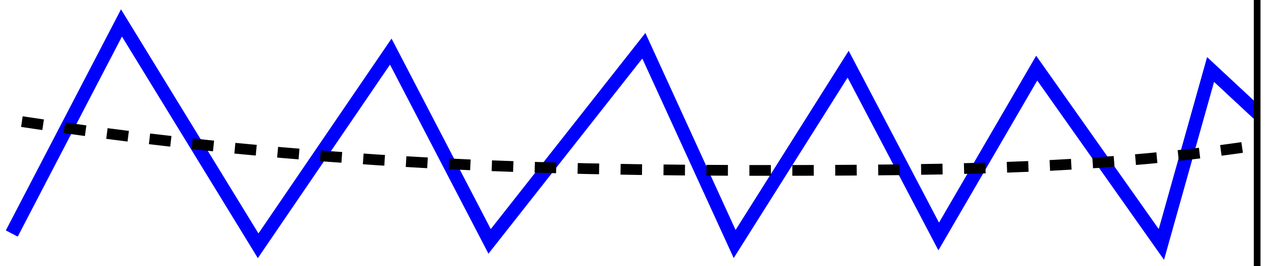
Central Flows: A Brief Recap

Challenge: optimizers operate at the **edge of stability** and their trajectories are oscillatory

Approach: model the **time-averaged** optimization trajectory

- Pretend w_t oscillates around a ***central flow*** $\overline{w}(t)$ with covarian

$$w_t = \overline{w}(t) + \delta_t \qquad \mathbb{E}[\delta_t] = 0 \qquad \mathbb{E}$$



- Derive necessary conditions for $(\overline{w}(t), \Sigma(t))$ which allow us to
- Empirically verify these central flows in realistic deep learning
- Interpret the central flow to understand the implicit behavior of t

but it is **very** predictive!

RMSProp (Adam without momentum)

$$\nu_t = \beta_2 \nu_{t-1} + (1 - \beta_2) \nabla L(w_t)^{\odot 2}$$

maintain an EMA of the squared gradient

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\nu_t}} \nabla L(w_t)$$

effective step sizes

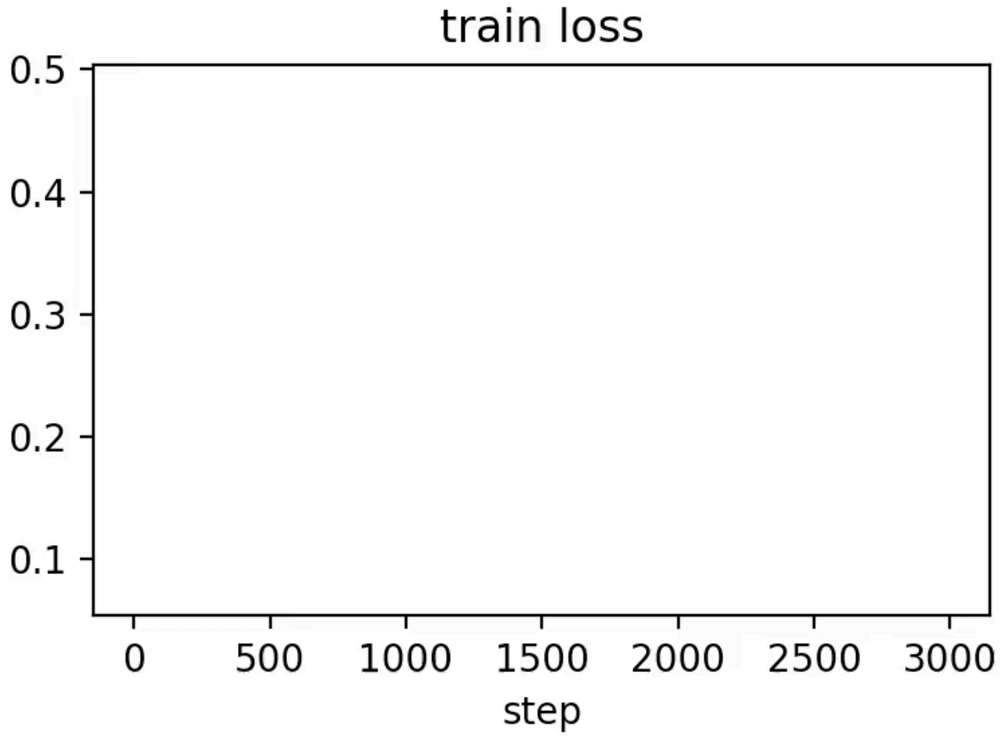
RMSProp (Adam without momentum)

$$\nu_t = \beta_2 \nu_{t-1} + (1 - \beta_2) \nabla L(w_t)^{\odot 2}$$

maintain an EMA of the squared gradient

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\nu_t}} \nabla L(w_t)$$

effective step sizes



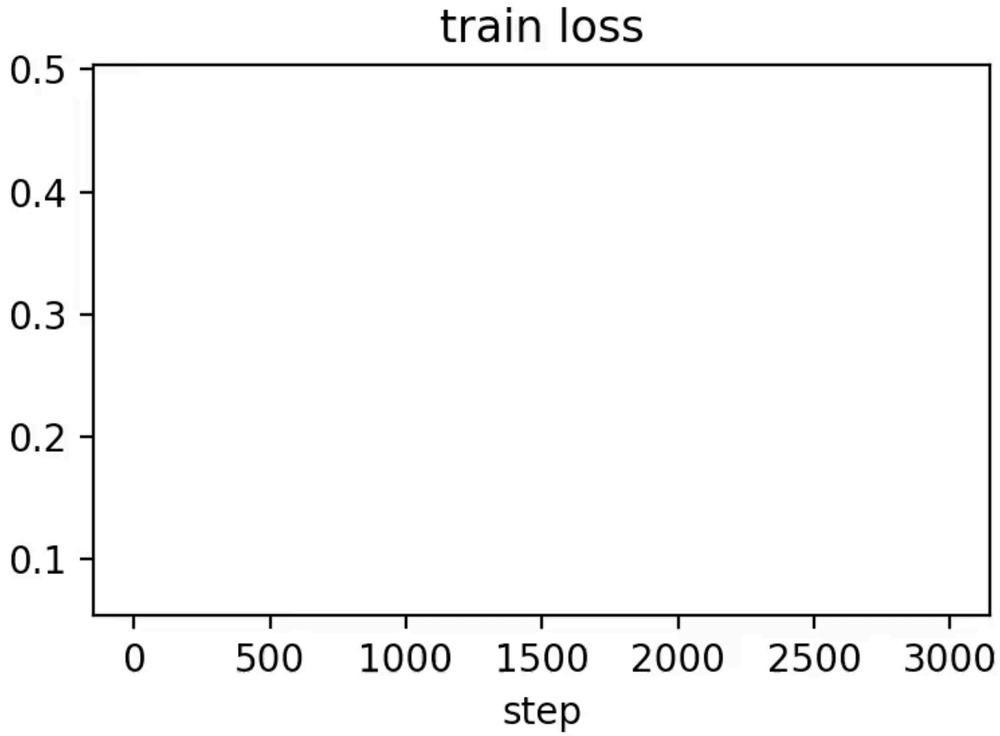
RMSProp (Adam without momentum)

$$\nu_t = \beta_2 \nu_{t-1} + (1 - \beta_2) \nabla L(w_t)^{\odot 2}$$

maintain an EMA of the squared gradient

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\nu_t}} \nabla L(w_t)$$

effective step sizes



If Adam is adaptive,

- what does it adapt **to**?
- what does the learning rate do?

RMSProp (Adam without momentum)

$$\nu_t = \beta_2 \nu_{t-1} + (1 - \beta_2) \nabla L(w_t)^{\odot 2}$$

maintain an EMA of the squared gradient

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\nu_t}} \nabla L(w_t)$$

effective step sizes

- what does it adapt **to**?

RMSProp (Adam without momentum)

$$\nu_t = \beta_2 \nu_{t-1} + (1 - \beta_2) \nabla L(w_t)^{\odot 2}$$

maintain an EMA of the squared gradient

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\nu_t}} \nabla L(w_t)$$

effective step sizes

- what does it adapt **to**?

A: it adapts to the gradient scale… duh?

RMSProp (Adam without momentum)

$$\nu_t = \beta_2 \nu_{t-1} + (1 - \beta_2) \nabla L(w_t)^{\odot 2}$$

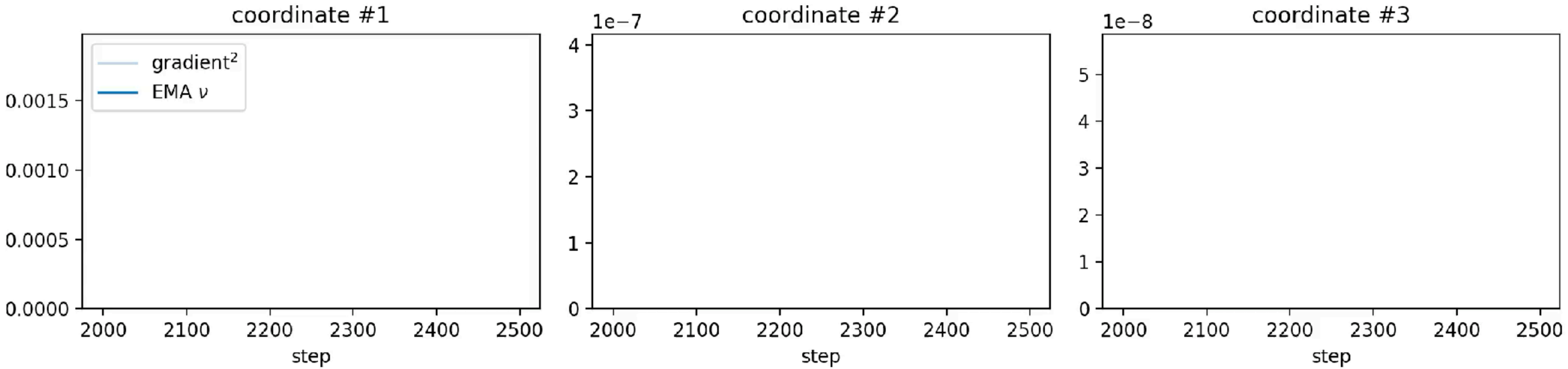
maintain an EMA of the squared gradient

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\nu_t}} \nabla L(w_t)$$

effective step sizes

- what does it adapt **to**?

A: it adapts to the gradient scale… duh?



RMSProp (Adam without momentum)

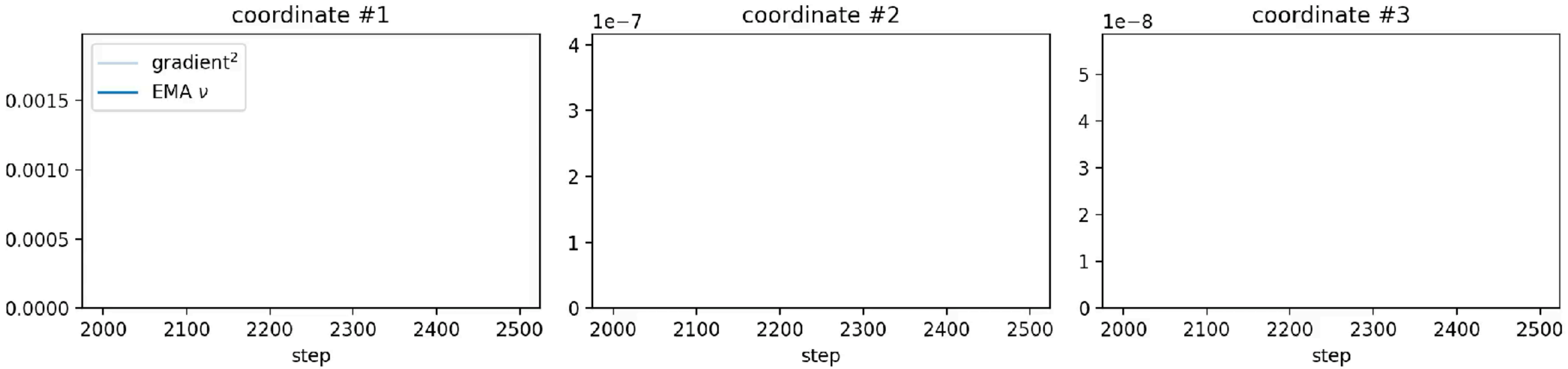
$$\nu_t = \beta_2 \nu_{t-1} + (1-\beta_2) \nabla L(w_t)^{\odot 2}$$

maintain an EMA of the squared gradient

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\nu_t}} \nabla L(w_t)$$

effective step sizes

Like Gradient Descent, RMSProp operates at the **edge of stability**



RMSProp (Adam without momentum)

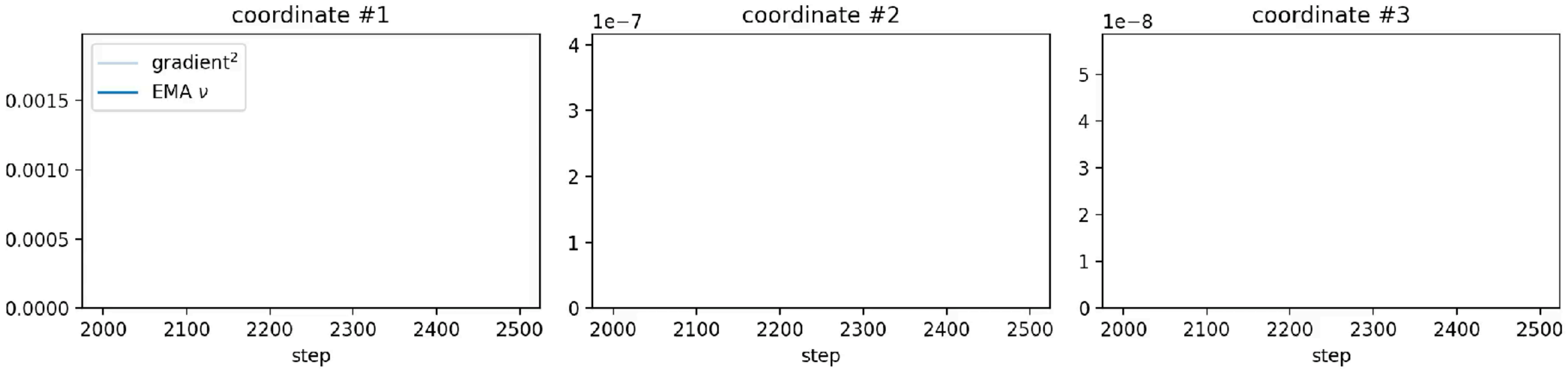
$$\nu_t = \beta_2 \nu_{t-1} + (1-\beta_2) \nabla L(w_t)^{\odot 2}$$

maintain an EMA of the squared gradient

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\nu_t}} \nabla L(w_t)$$

effective step sizes

Like Gradient Descent, RMSProp operates at the **edge of stability**



RMSProp (Adam without momentum)

$$\nu_t = \beta_2 \nu_{t-1} + (1-\beta_2)\nabla L(w_t)^{\odot 2}$$

maintain an EMA of the squared gradient

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\nu_t}} \nabla L(w_t)$$

effective step sizes

- If $P_t^{-1} := \operatorname{diag}\left(\frac{\eta}{\sqrt{\nu_t}}\right)$, RMSProp is just GD with an evolving preconditioner P_t:

$$w_{t+1} = w_t - P_t^{-1} \nabla L(w_t)$$

RMSProp (Adam without momentum)

$$\nu_t = \beta_2 \nu_{t-1} + (1-\beta_2)\nabla L(w_t)^{\odot 2} \qquad w_{t+1} = w_t - \frac{\eta}{\sqrt{\nu_t}}\nabla L(w_t)$$

maintain an EMA of the squared gradient

effective step sizes

- If $P_t^{-1} := \operatorname{diag}\left(\frac{\eta}{\sqrt{\nu_t}}\right)$, RMSProp is just GD with an evolving preconditioner P_t:

$$w_{t+1} = w_t - P_t^{-1}\nabla L(w_t)$$

- On a quadratic $L(w) = \frac{1}{2}w^T H w$, the update is $w \leftarrow (I - P_t^{-1}H)w$

RMSProp (Adam without momentum)

$$\nu_t = \beta_2 \nu_{t-1} + (1-\beta_2)\nabla L(w_t)^{\odot 2} \qquad w_{t+1} = w_t - \frac{\eta}{\sqrt{\nu_t}}\nabla L(w_t)$$

maintain an EMA of the squared gradient

effective step sizes

- If $P_t^{-1} := \operatorname{diag}\left(\frac{\eta}{\sqrt{\nu_t}}\right)$, RMSProp is just GD with an evolving preconditioner P_t:

$$w_{t+1} = w_t - P_t^{-1}\nabla L(w_t)$$

- On a quadratic $L(w) = \frac{1}{2}w^T H w$, the update is $w \leftarrow (I - P_t^{-1}H)w$

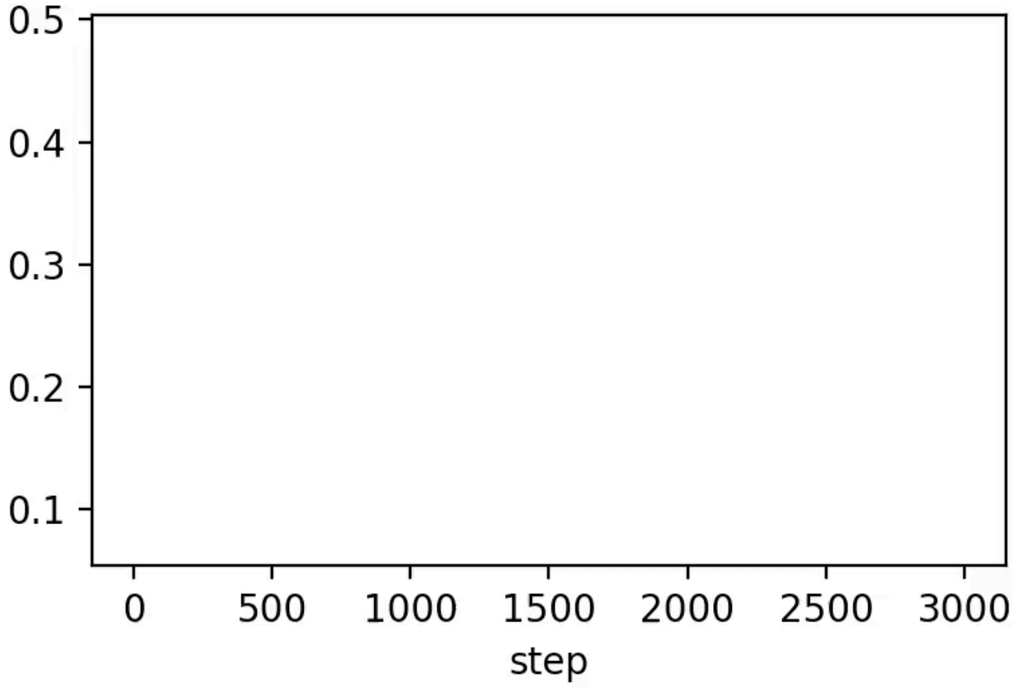
- In general, local stability is given by the eigenvalues of the effective Hessian:

$$\lambda_1(H_{\text{eff}}) \le 2 \qquad \text{where} \qquad H_{\text{eff}} = P_t^{-1}H(w_t) \qquad \text{and} \qquad H = \nabla^2 L$$

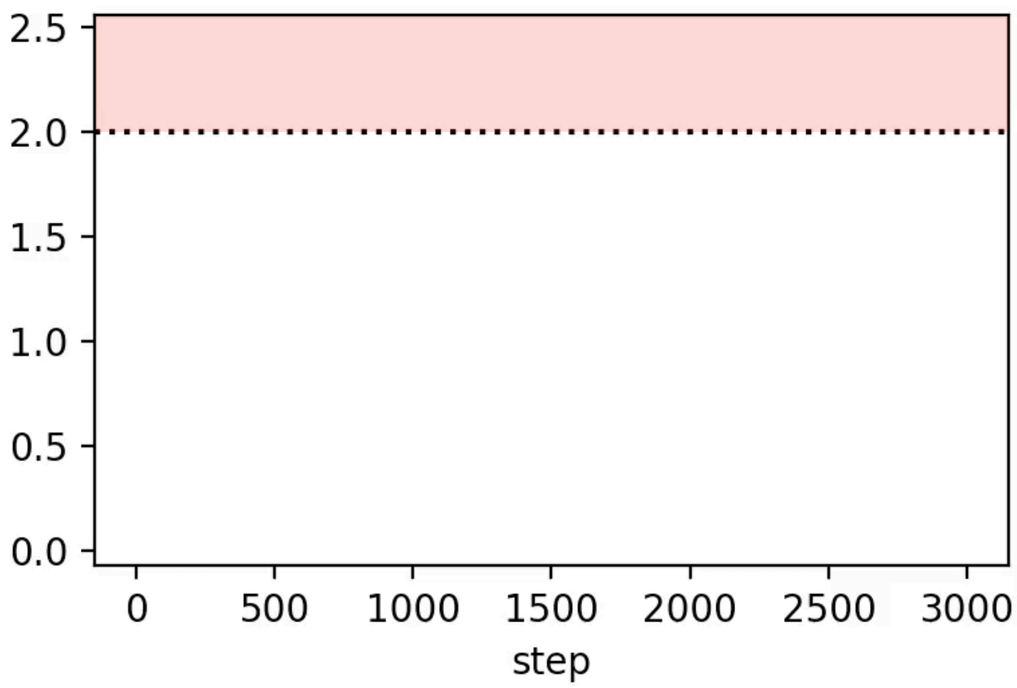
RMSProp in Deep Learning

RMSProp on a ResNet, trained on CIFAR10, $\eta = 2 \times 10^{-5}, \beta = 0.99$

Loss $L(w)$



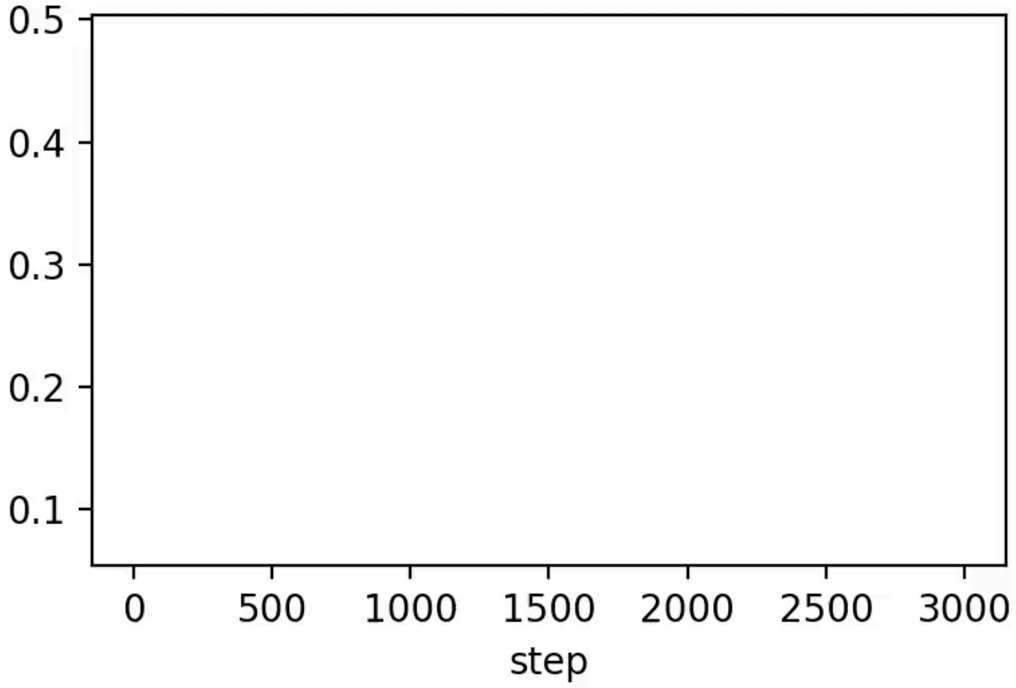
Top 4 evals of $H_{\text{eff}} := P_t^{-1} H(w_t)$



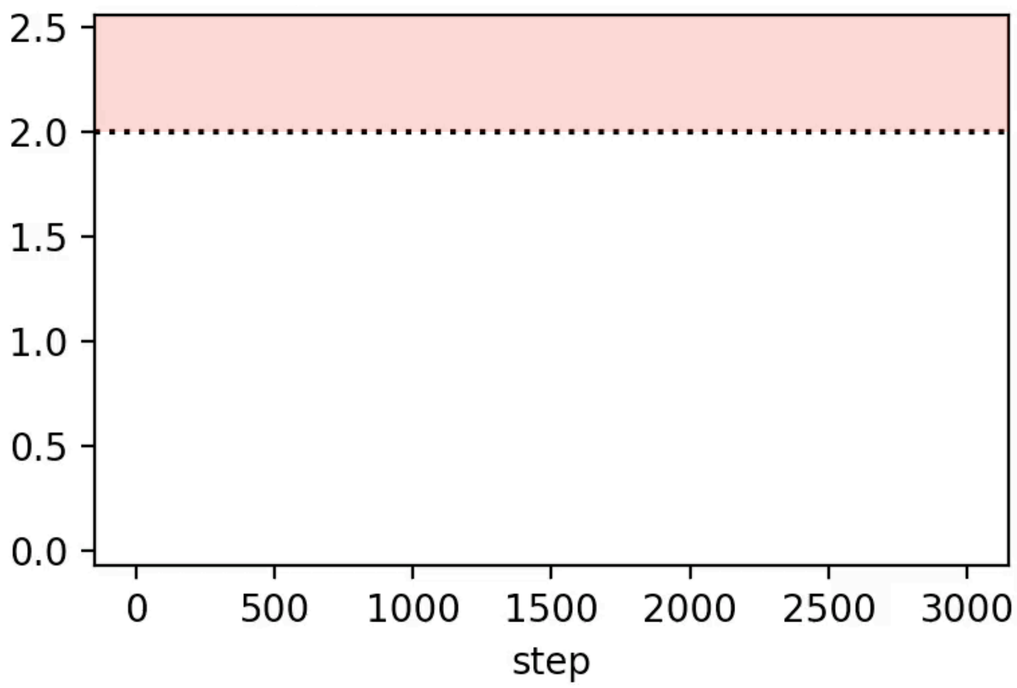
RMSProp in Deep Learning

RMSProp on a ResNet, trained on CIFAR10, $\eta = 2 \times 10^{-5}, \beta = 0.99$

Loss $L(w)$

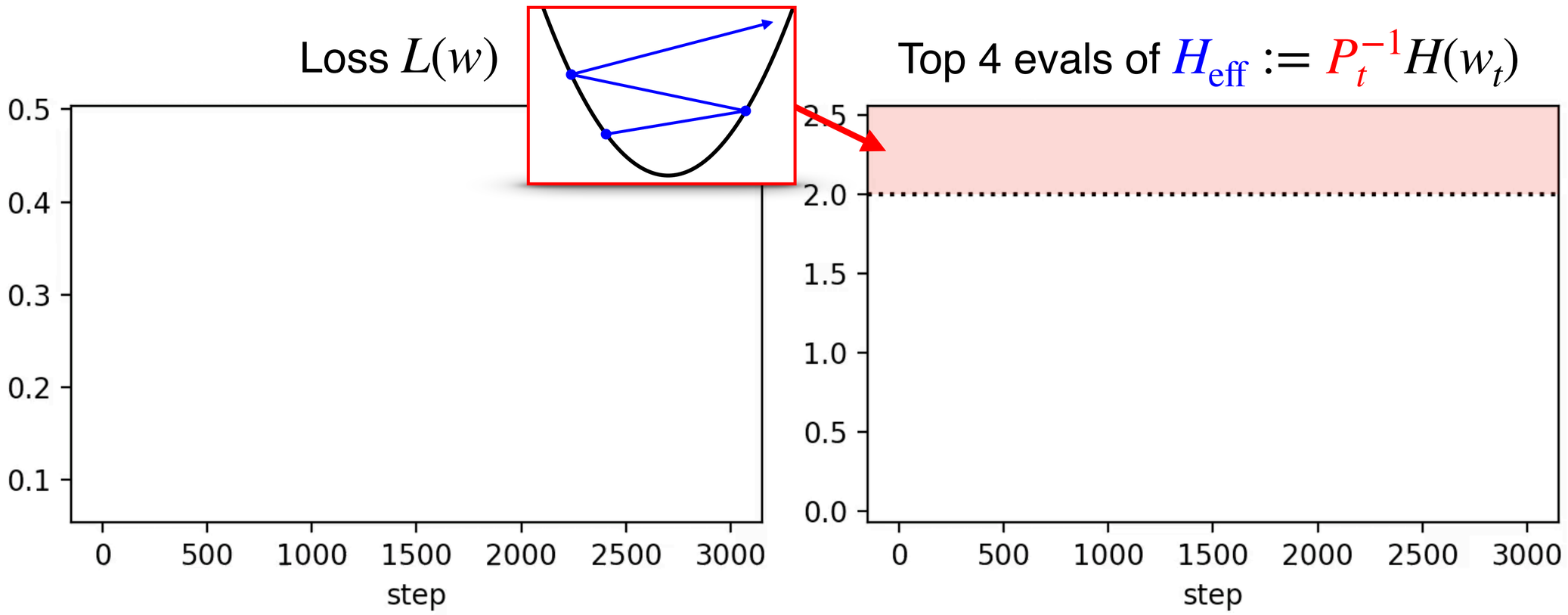


Top 4 evals of $H_{\text{eff}} := P_t^{-1} H(w_t)$



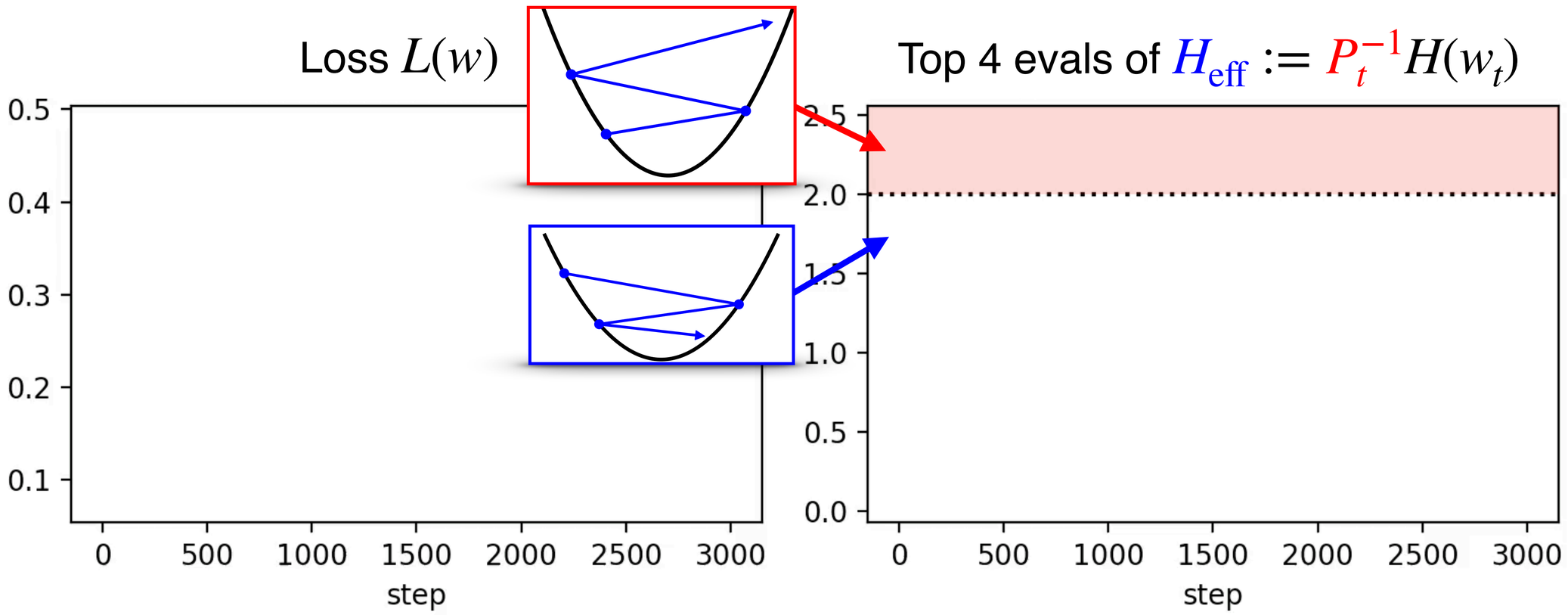
RMSProp in Deep Learning

RMSProp on a ResNet, trained on CIFAR10, $\eta = 2 \times 10^{-5}, \beta = 0.99$



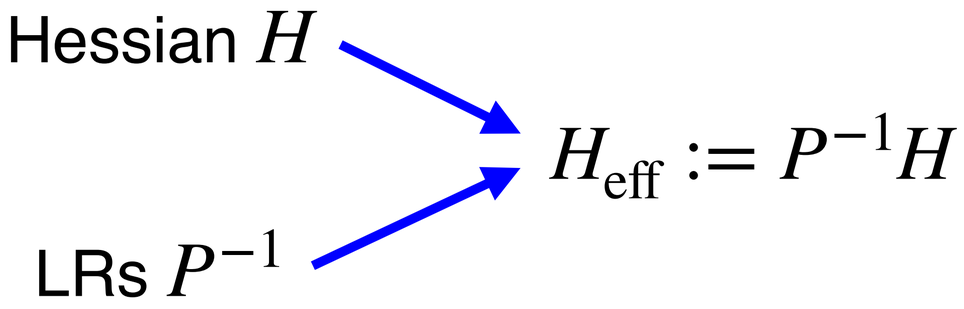
RMSProp in Deep Learning

RMSProp on a ResNet, trained on CIFAR10, $\eta = 2 \times 10^{-5}, \beta = 0.99$



RMSProp at the Edge of Stability

$$\nu = \mathrm{EMA}\left(\nabla L(w)^{\odot 2}\right) \qquad P_t^{-1} = \mathrm{diag}\left(\frac{\eta}{\sqrt{\nu}}\right) \qquad w_{t+1} = w_t - P_t^{-1}\nabla L(w_t)$$

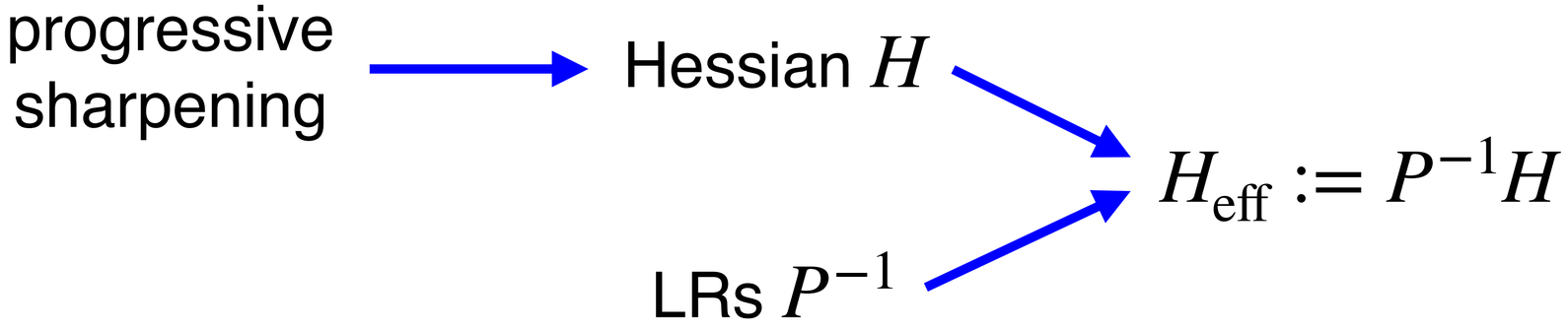


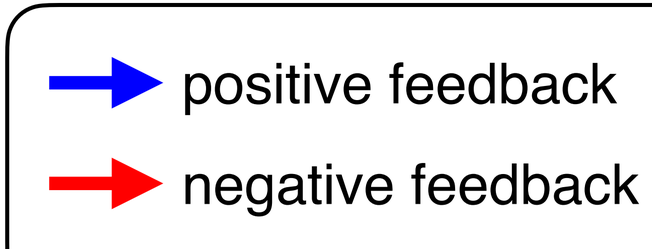
positive feedback

negative feedback

RMSProp at the Edge of Stability

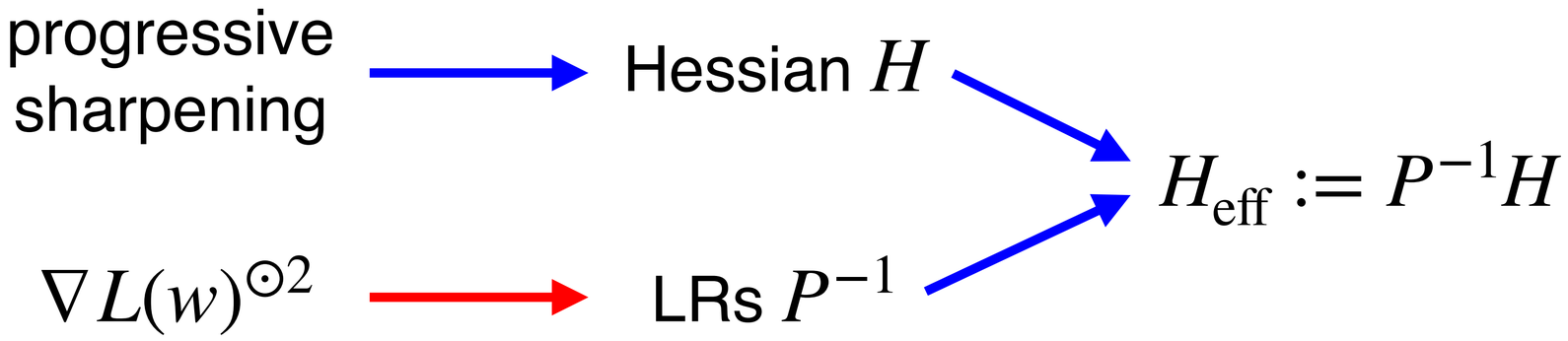
$$\nu = \mathrm{EMA}\left(\nabla L(w)^{\odot 2}\right) \qquad P_t^{-1} = \mathrm{diag}\left(\frac{\eta}{\sqrt{\nu}}\right) \qquad w_{t+1} = w_t - P_t^{-1}\nabla L(w_t)$$





RMSProp at the Edge of Stability

$$\nu = \mathrm{EMA}\left(\nabla L(w)^{\odot 2}\right) \qquad P_t^{-1} = \mathrm{diag}\left(\frac{\eta}{\sqrt{\nu}}\right) \qquad w_{t+1} = w_t - P_t^{-1}\nabla L(w_t)$$

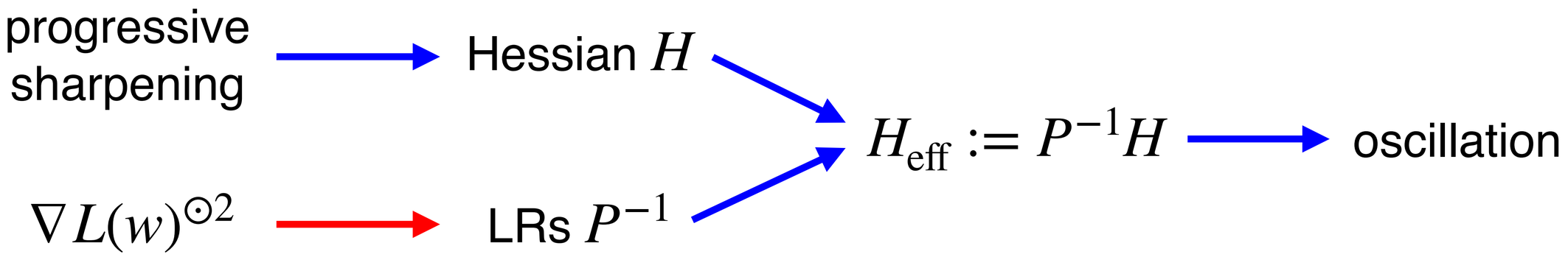


positive feedback

negative feedback

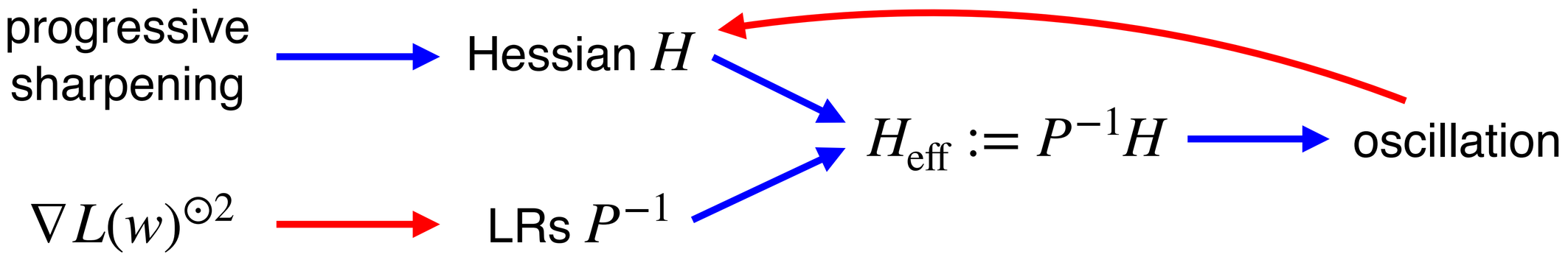
RMSProp at the Edge of Stability

$$\nu = \mathrm{EMA}\left(\nabla L(w)^{\odot 2}\right) \qquad P_t^{-1} = \mathrm{diag}\left(\frac{\eta}{\sqrt{\nu}}\right) \qquad w_{t+1} = w_t - P_t^{-1}\nabla L(w_t)$$



RMSProp at the Edge of Stability

$$\nu = \mathrm{EMA}\left(\nabla L(w)^{\odot 2}\right) \qquad P_t^{-1} = \mathrm{diag}\left(\frac{\eta}{\sqrt{\nu}}\right) \qquad w_{t+1} = w_t - P_t^{-1}\nabla L(w_t)$$

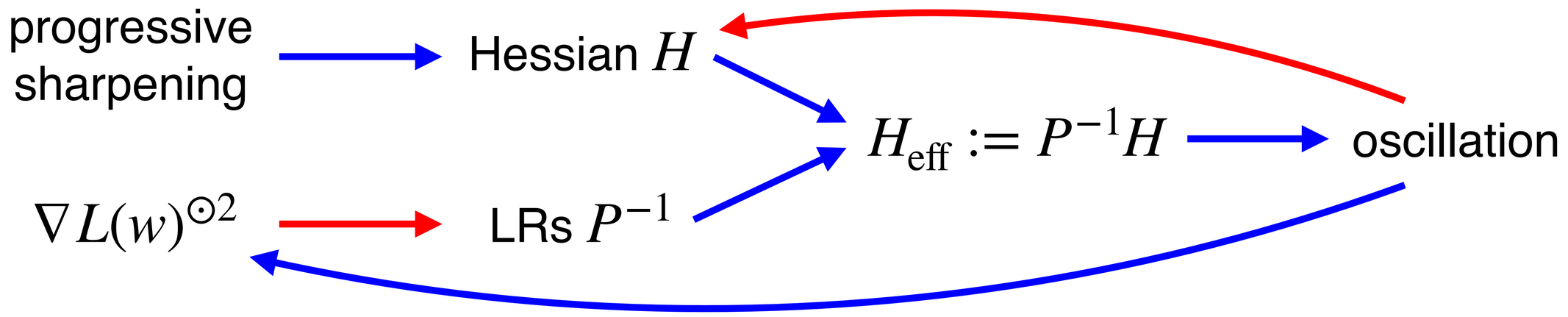


positive feedback

negative feedback

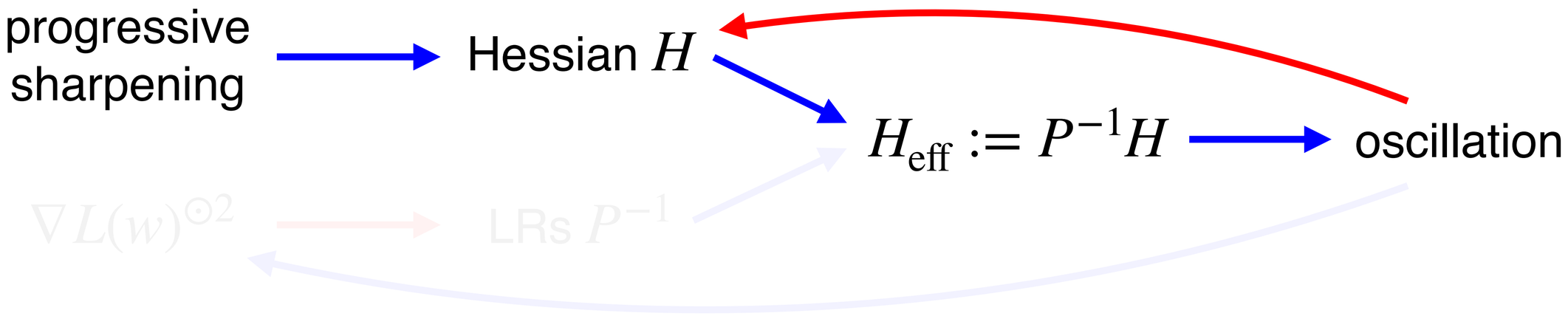
RMSProp at the Edge of Stability

$$\nu = \mathrm{EMA}\left(\nabla L(w)^{\odot 2}\right) \qquad P_t^{-1} = \mathrm{diag}\left(\frac{\eta}{\sqrt{\nu}}\right) \qquad w_{t+1} = w_t - P_t^{-1}\nabla L(w_t)$$



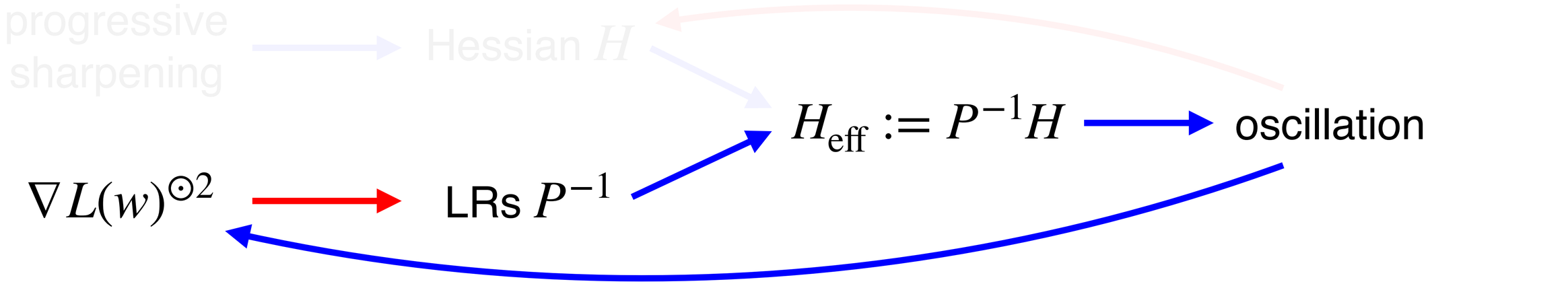
RMSProp at the Edge of Stability

$$\nu = \mathrm{EMA}\left(\nabla L(w)^{\odot 2}\right) \qquad P_t^{-1} = \mathrm{diag}\left(\frac{\eta}{\sqrt{\nu}}\right) \qquad w_{t+1} = w_t - P_t^{-1}\nabla L(w_t)$$



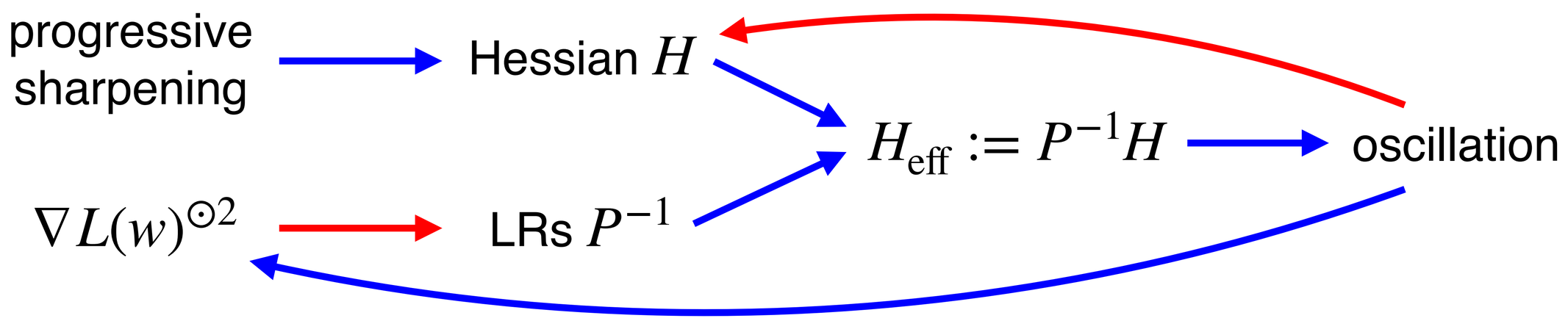
RMSProp at the Edge of Stability

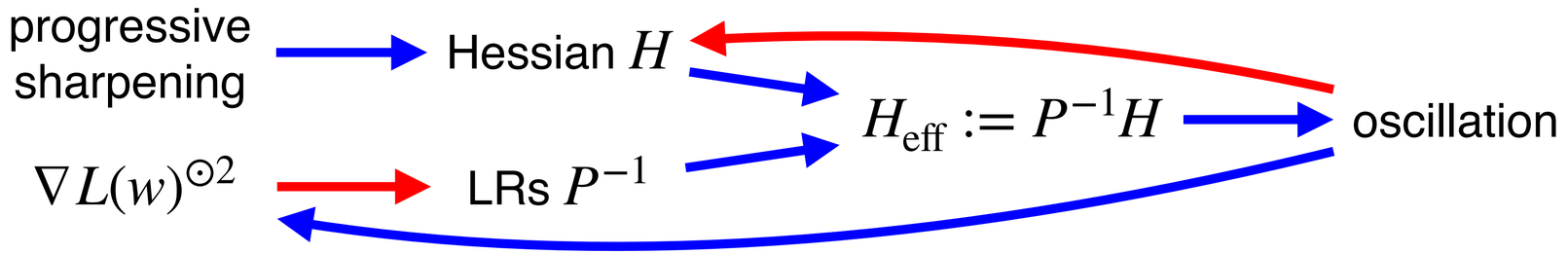
$$\nu = \mathrm{EMA}\left(\nabla L(w)^{\odot 2}\right) \qquad P_t^{-1} = \mathrm{diag}\left(\frac{\eta}{\sqrt{\nu}}\right) \qquad w_{t+1} = w_t - P_t^{-1}\nabla L(w_t)$$

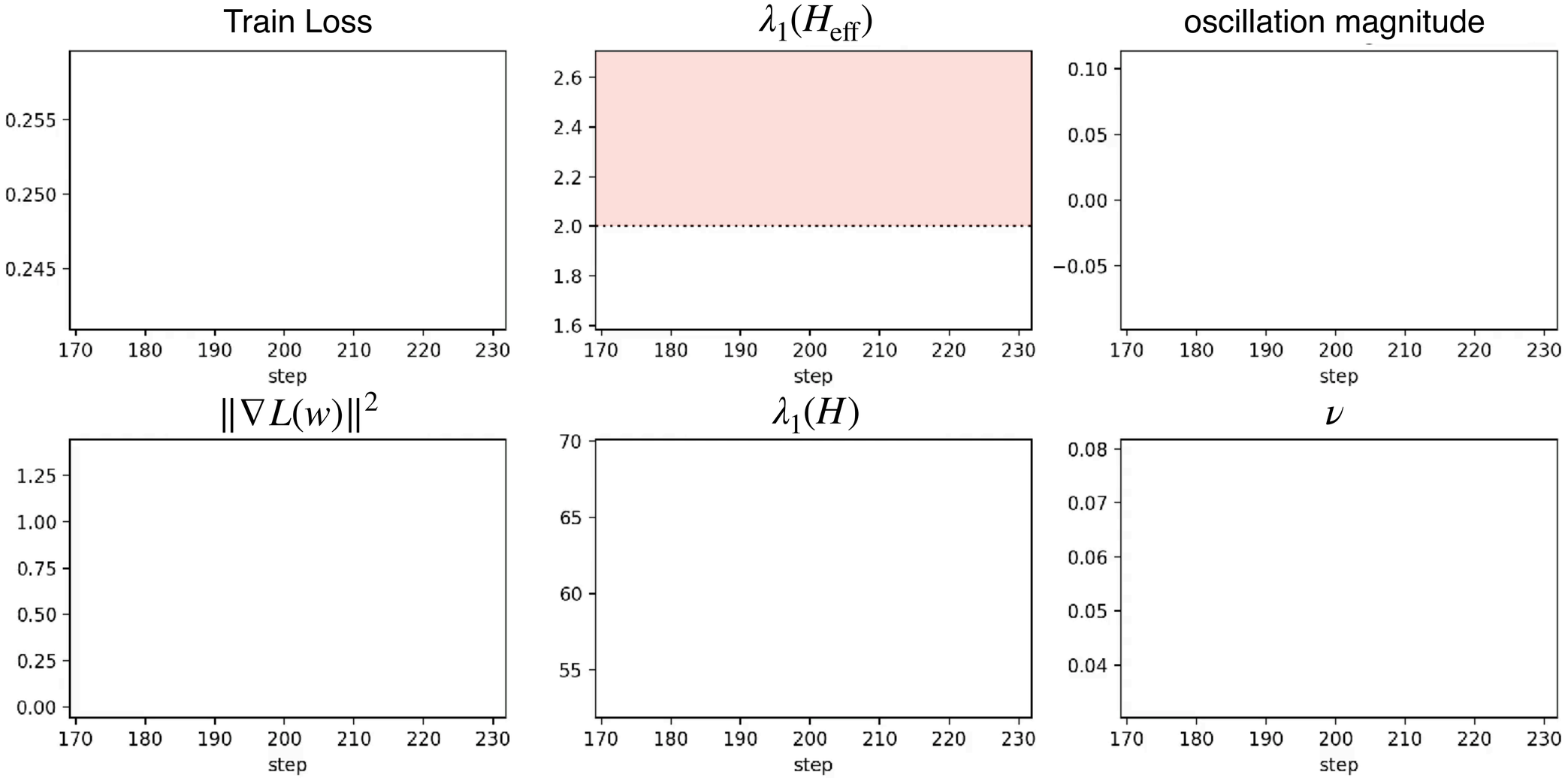


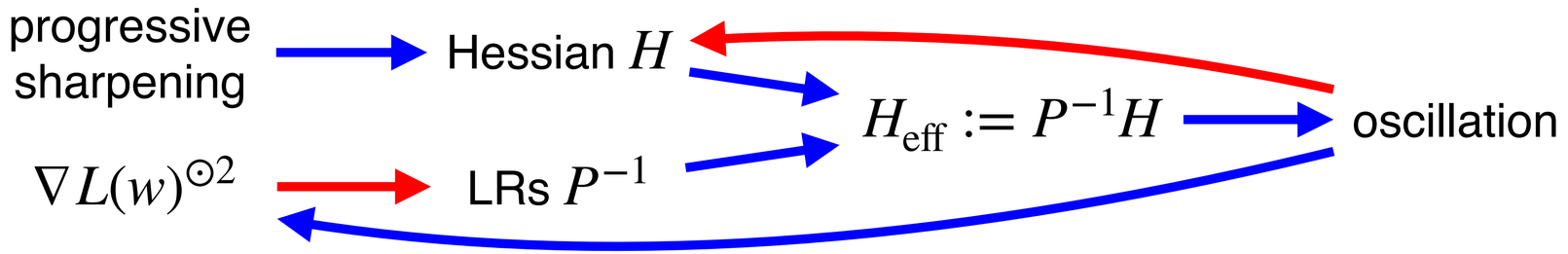
(Scalar) RMSProp at the Edge of Stability

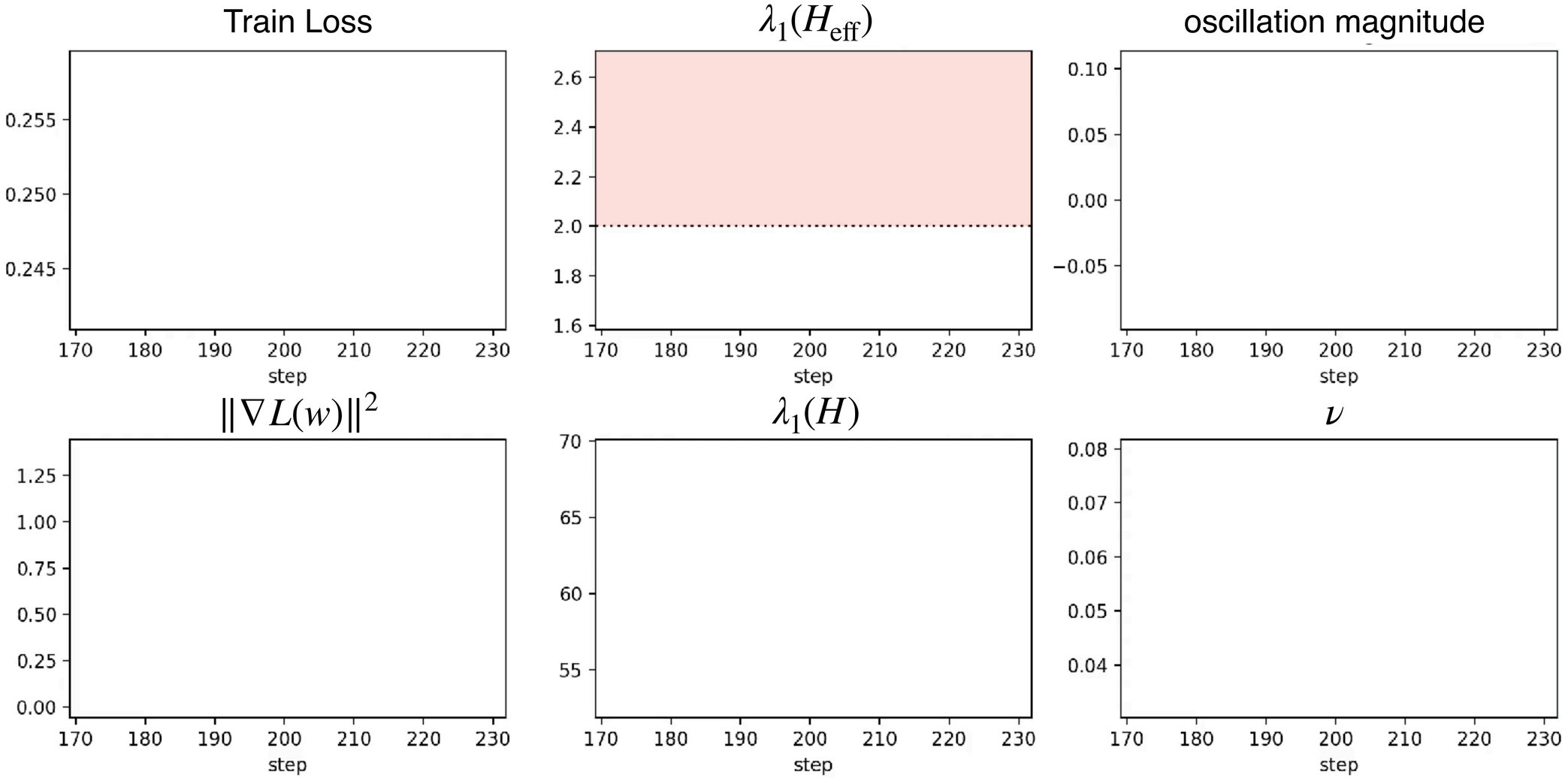
$$\nu = \mathrm{EMA}\left(\|\nabla L(w)\|^2\right) \qquad P_t^{-1} = \frac{\eta}{\sqrt{\nu}} \qquad w_{t+1} = w_t - P_t^{-1}\nabla L(w_t)$$




progressive sharpening
Hessian H
$H_{\text{eff}} := P^{-1}H$
oscillation
$\nabla L(w)^{\odot 2}$
LRs P^{-1}


Train Loss
$\lambda_1(H_{\text{eff}})$
oscillation magnitude
$\|\nabla L(w)\|^2$
$\lambda_1(H)$
ν
step

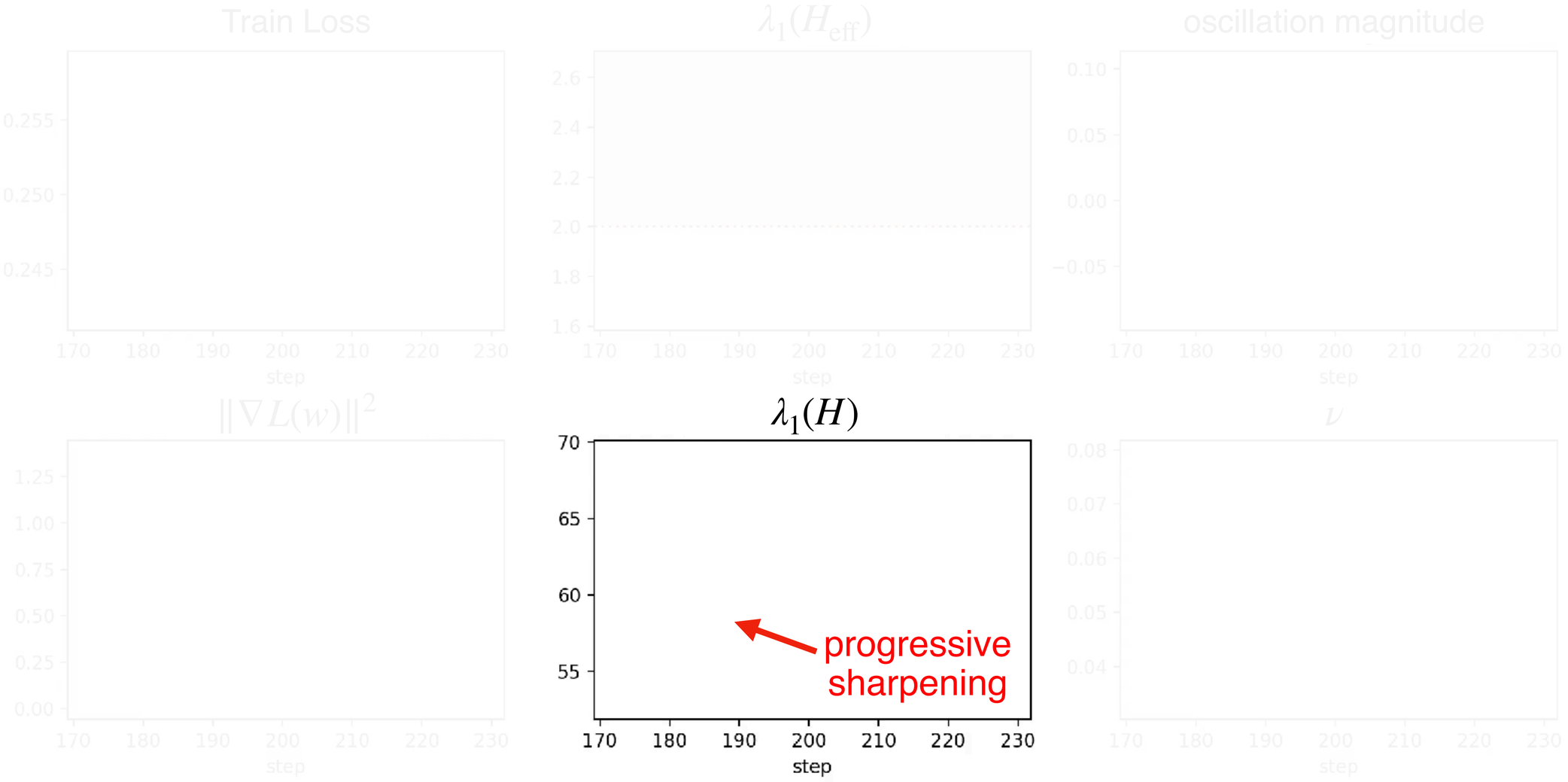

progressive sharpening
Hessian H
$H_{\text{eff}} := P^{-1}H$
oscillation
$\nabla L(w)^{\odot 2}$
LRs P^{-1}


Train Loss
$\lambda_1(H_{\text{eff}})$
oscillation magnitude
$\|\nabla L(w)\|^2$
$\lambda_1(H)$
ν
step

progressive sharpening → Hessian H

$H_{\text{eff}} := P^{-1}H$ → oscillation

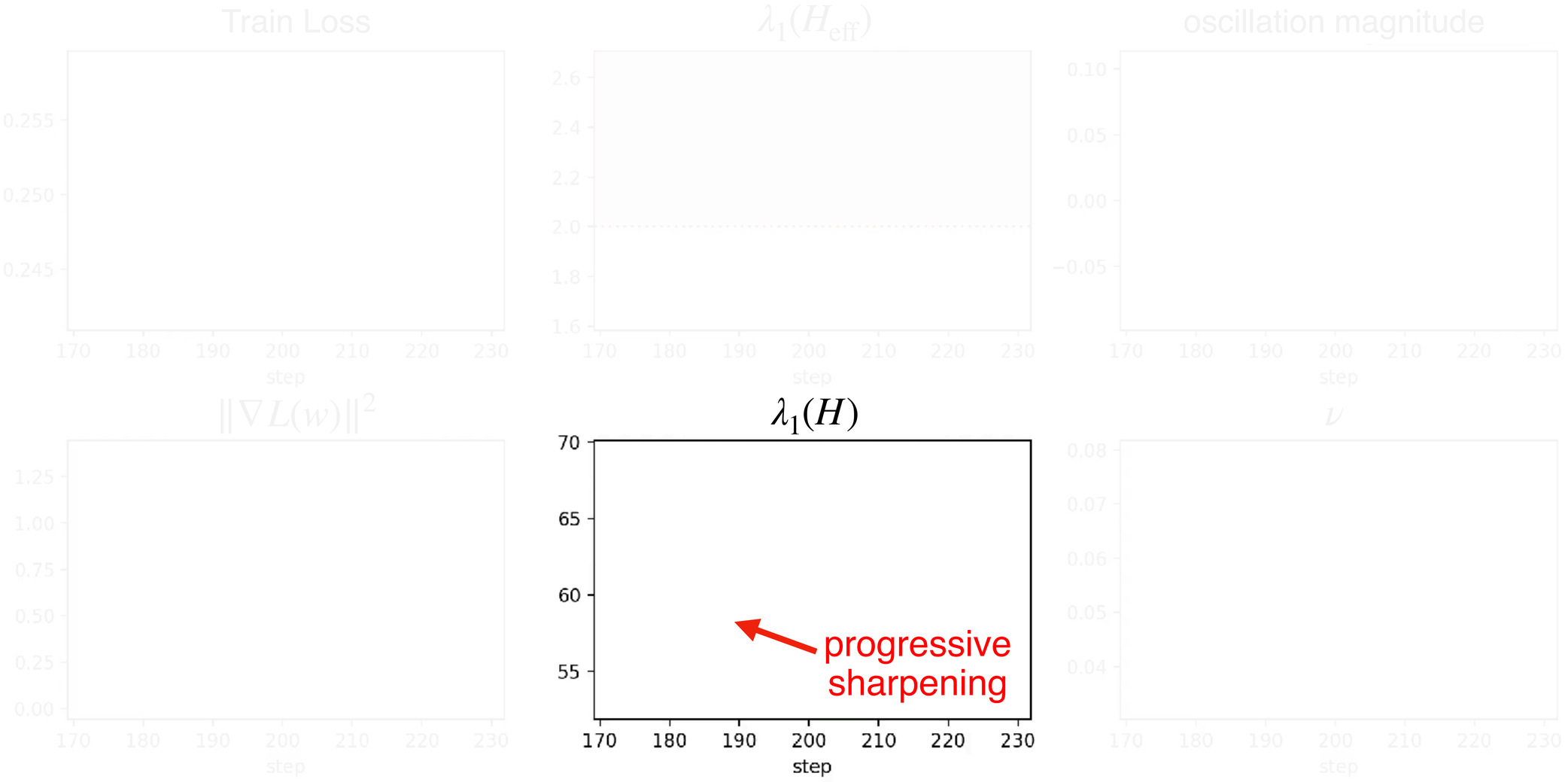
$\nabla L(w)^{\odot 2}$ → LRs P^{-1}



progressive sharpening → Hessian H

$H_{\text{eff}} := P^{-1}H$ → oscillation

$\nabla L(w)^{\odot 2}$ → LRs P^{-1}



progressive sharpening → Hessian H

$H_{\text{eff}} := P^{-1}H$ → oscillation

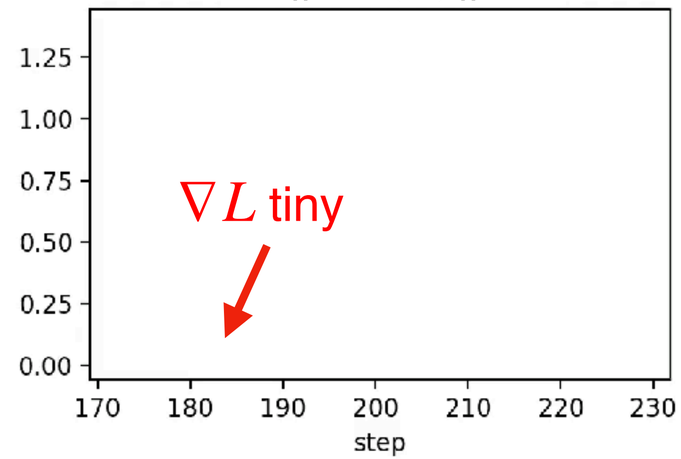
$\nabla L(w)^{\odot 2}$ → LRs P^{-1}

Train Loss

$\lambda_1(H_{\text{eff}})$

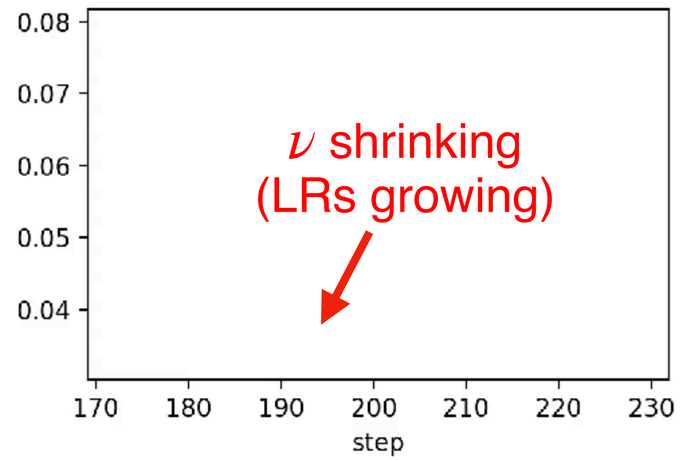
oscillation magnitude

$\|\nabla L(w)\|^2$



$\lambda_1(H)$

ν



progressive sharpening → Hessian H

$H_{\text{eff}} := P^{-1}H$ → oscillation

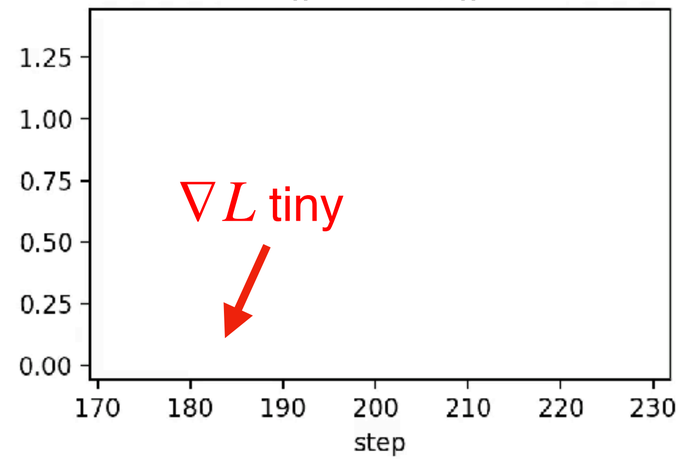
$\nabla L(w)^{\odot 2}$ → LRs P^{-1}

Train Loss

$\lambda_1(H_{\text{eff}})$

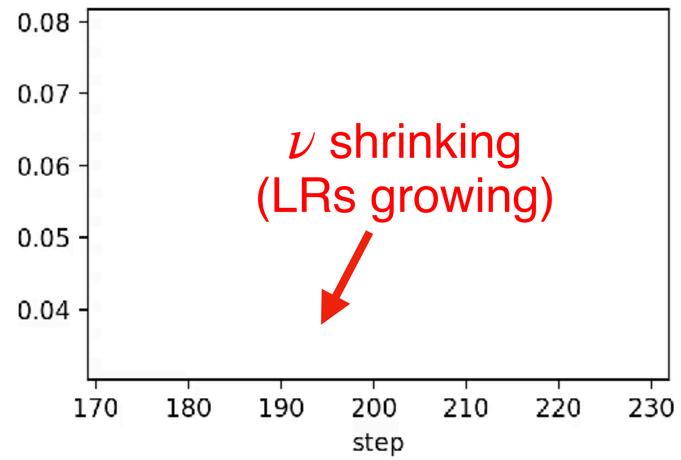
oscillation magnitude

$\|\nabla L(w)\|^2$



$\lambda_1(H)$

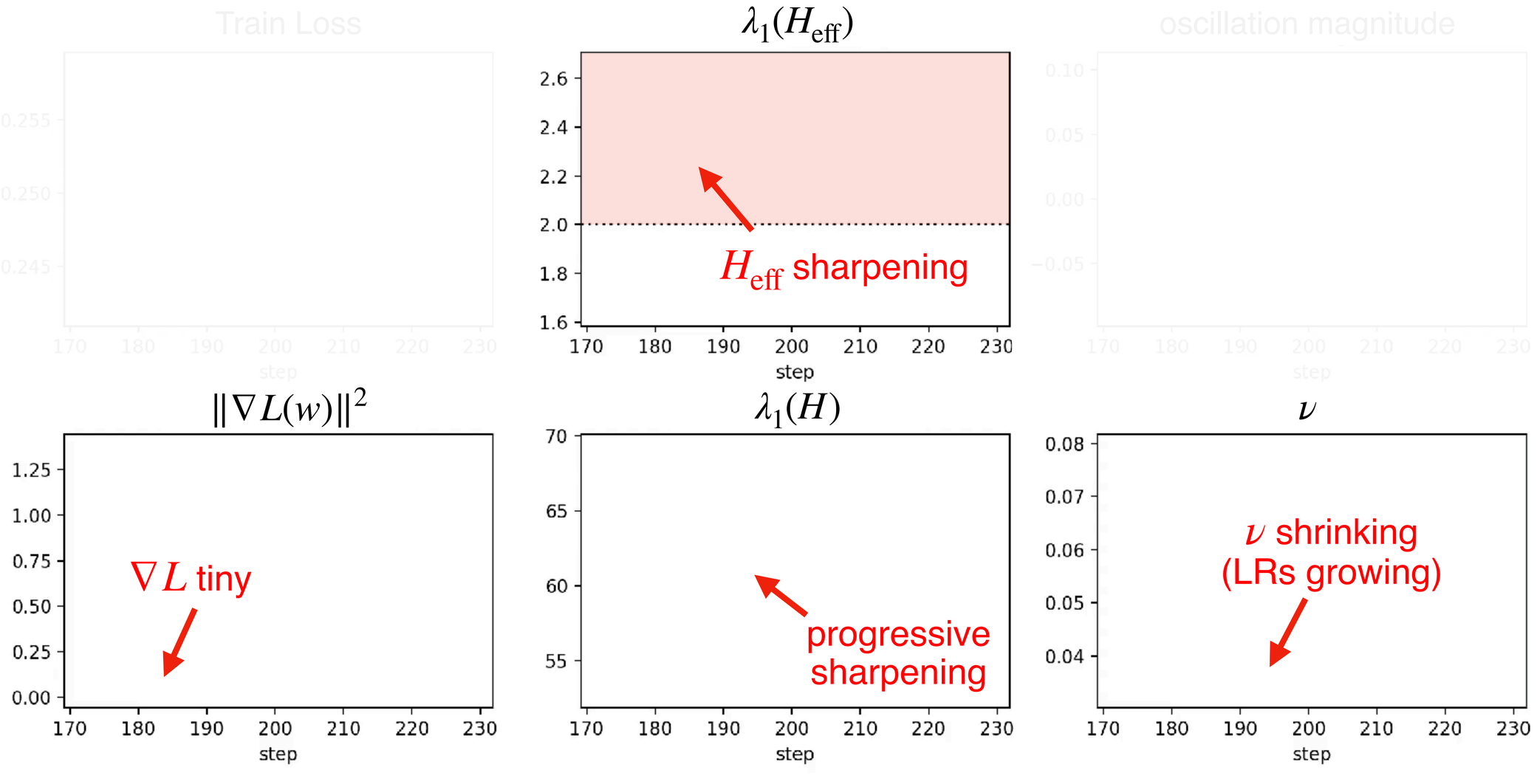
ν



progressive sharpening → Hessian H

$\nabla L(w)^{\odot 2}$ → LRs P^{-1}

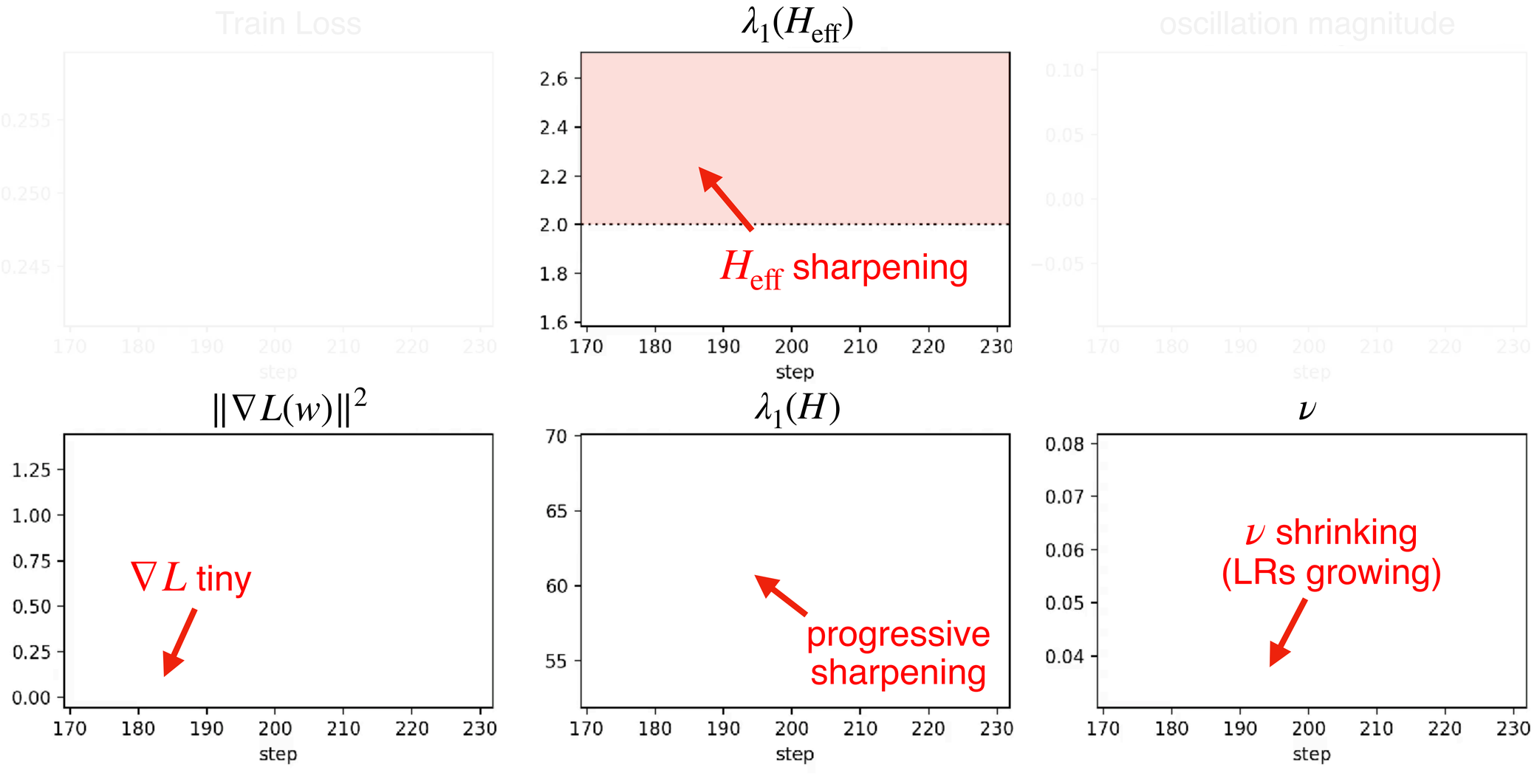
$H_{\text{eff}} := P^{-1}H$ → oscillation



progressive sharpening → Hessian H

$\nabla L(w)^{\odot 2}$ → LRs P^{-1}

$H_{\text{eff}} := P^{-1}H$ → oscillation



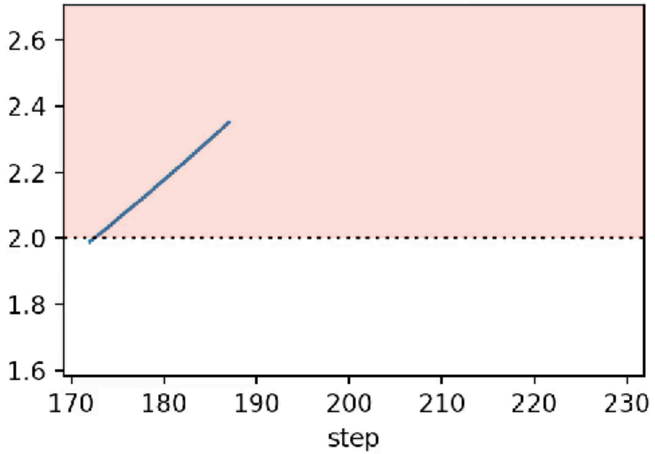
progressive sharpening → Hessian H

$\nabla L(w)^{\odot 2}$ → LRs P^{-1}

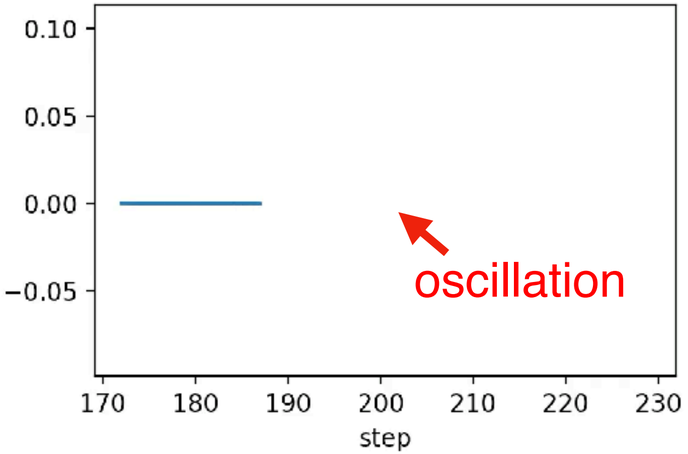
$H_{\text{eff}} := P^{-1}H$ → oscillation

Train Loss

$\lambda_1(H_{\text{eff}})$



oscillation magnitude



$\|\nabla L(w)\|^2$

$\lambda_1(H)$

ν

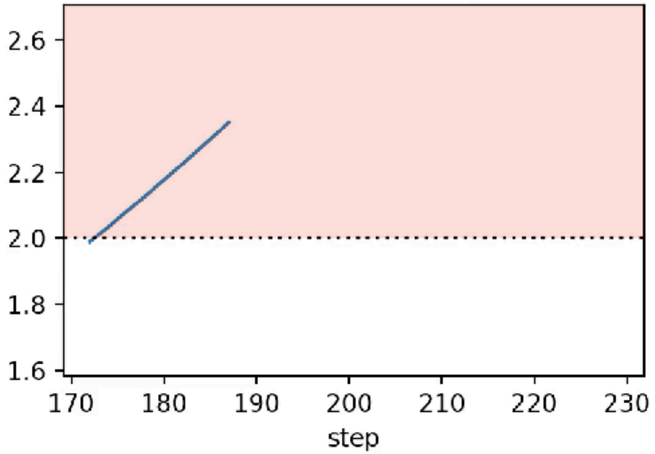
progressive sharpening → Hessian H

$\nabla L(w)^{\odot 2}$ → LRs P^{-1}

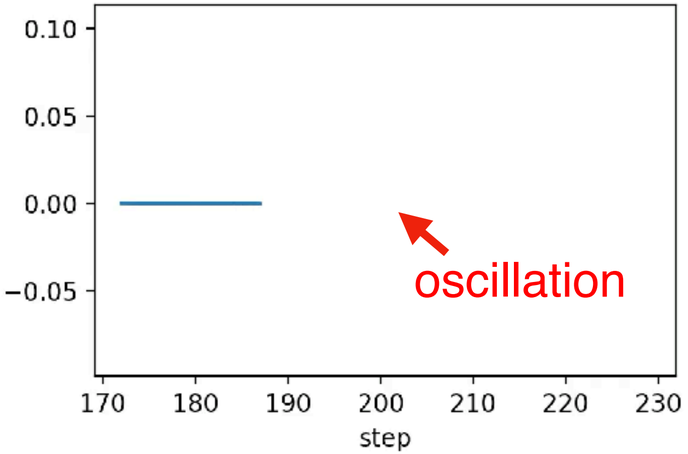
$H_{\text{eff}} := P^{-1}H$ → oscillation

Train Loss

$\lambda_1(H_{\text{eff}})$



oscillation magnitude



$\|\nabla L(w)\|^2$

$\lambda_1(H)$

ν

progressive sharpening

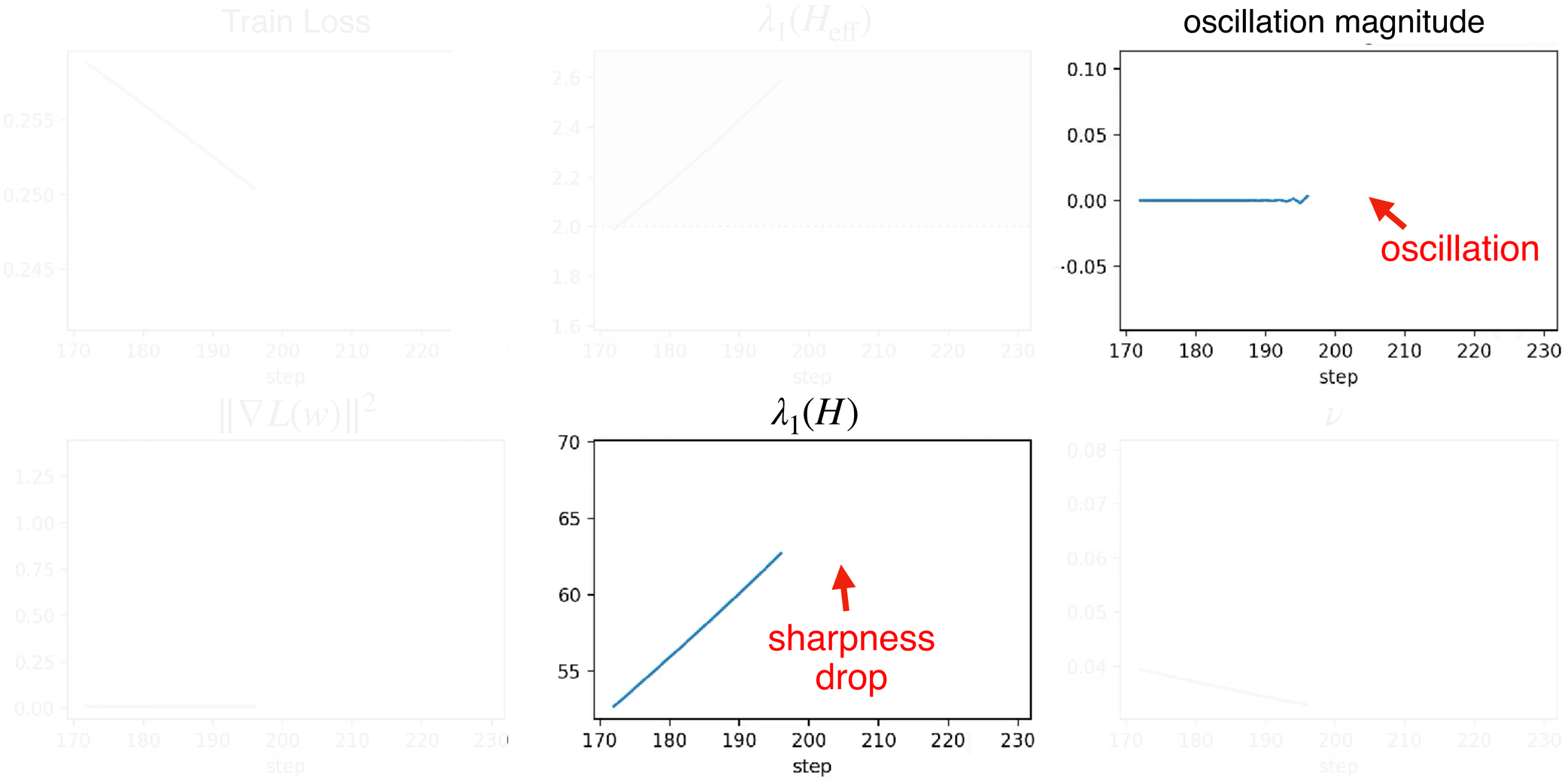
Hessian H

$H_{\text{eff}} := P^{-1}H$

oscillation

$\nabla L(w)^{\odot 2}$

LRs P^{-1}



progressive sharpening

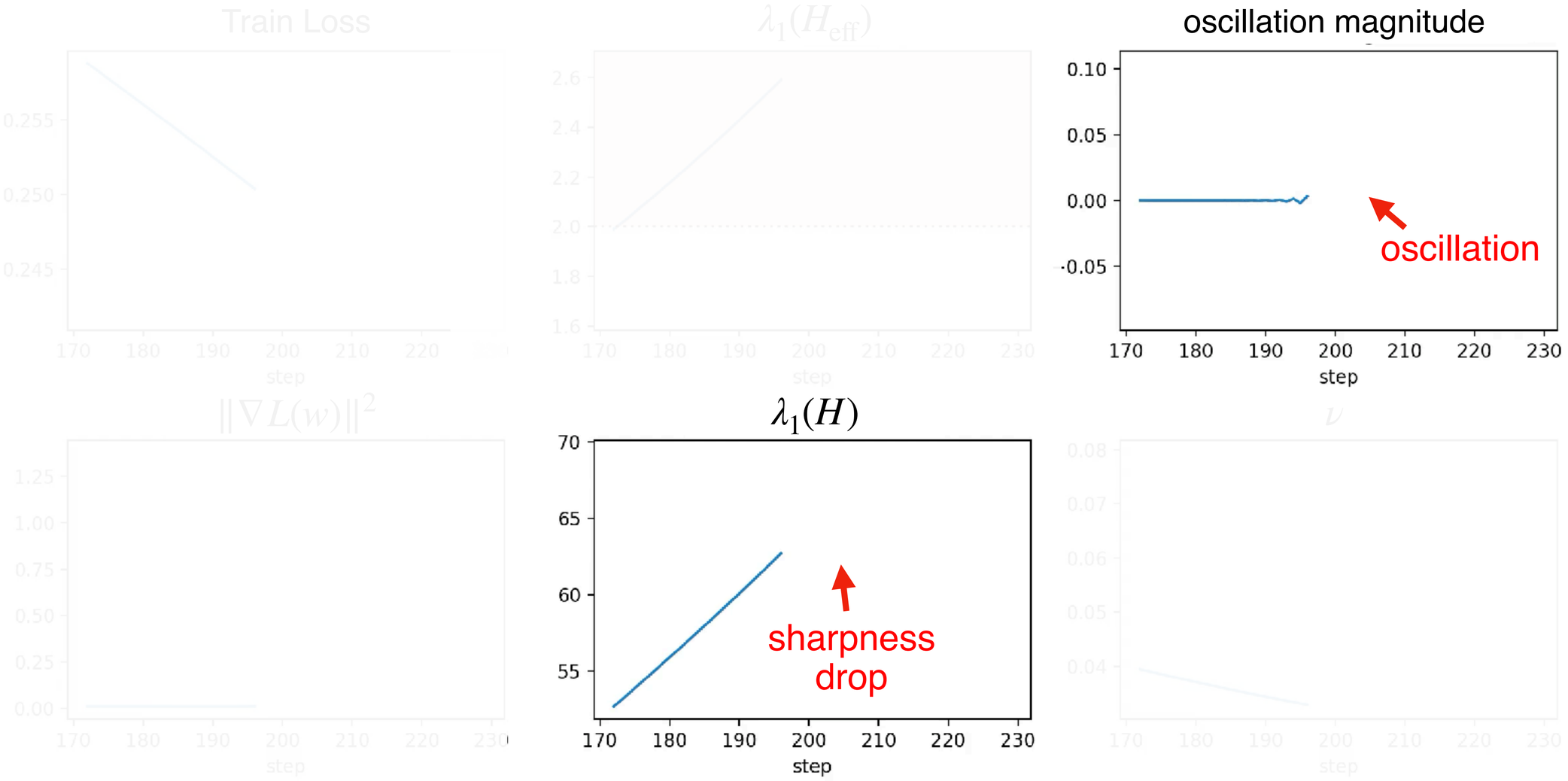
Hessian H

$H_{\text{eff}} := P^{-1}H$

oscillation

$\nabla L(w)^{\odot 2}$

LRs P^{-1}

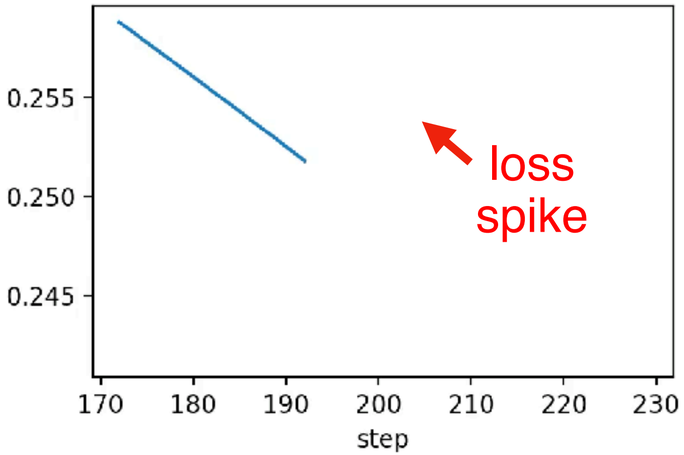


progressive sharpening → Hessian H

$H_{\text{eff}} := P^{-1}H$ → oscillation

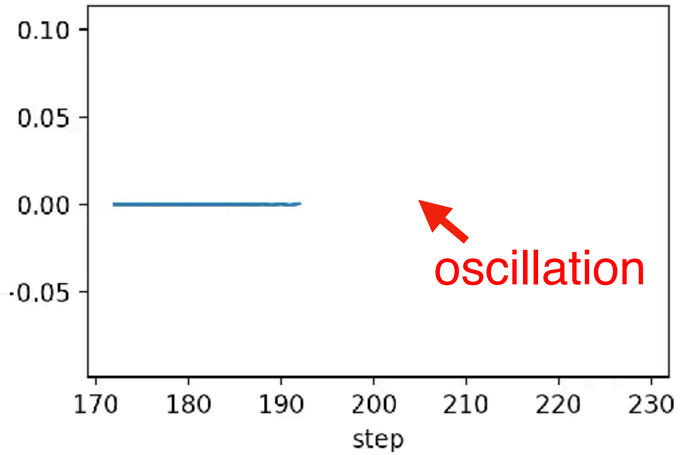
$\nabla L(w)^{\odot 2}$ → LRs P^{-1}

Train Loss

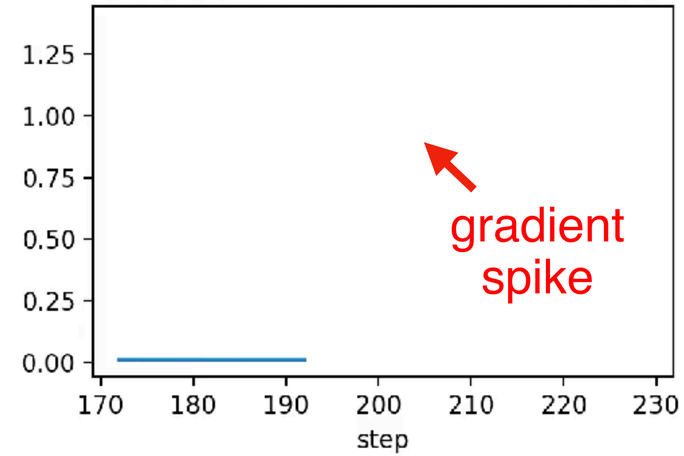


$\lambda_1(H_{\text{eff}})$

oscillation magnitude



$\|\nabla L(w)\|^2$



$\lambda_1(H)$

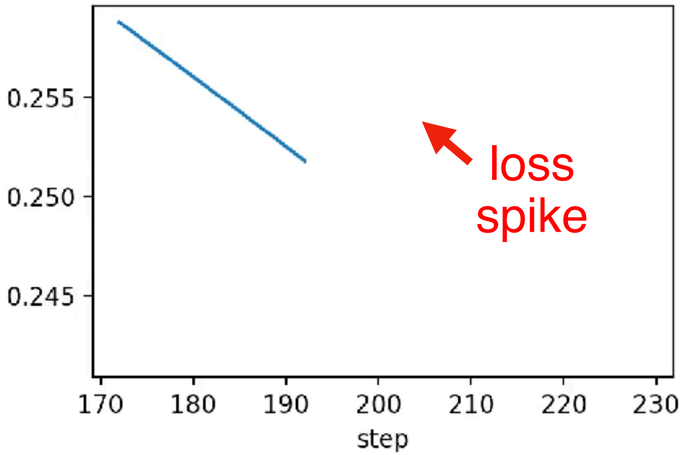
ν

progressive sharpening → Hessian H

$H_{\text{eff}} := P^{-1}H$ → oscillation

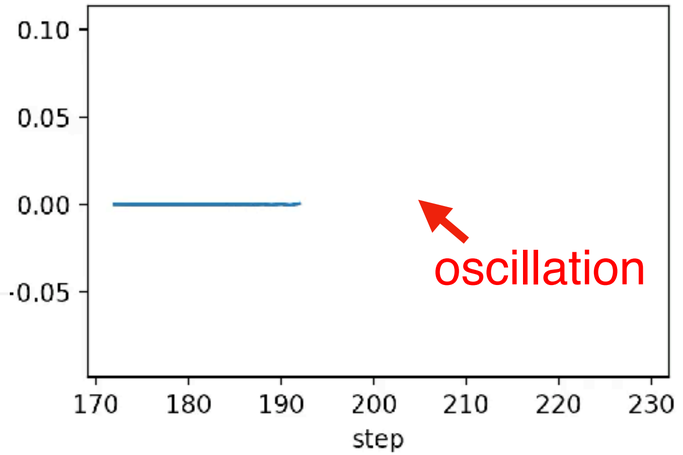
$\nabla L(w)^{\odot 2}$ → LRs P^{-1}

Train Loss

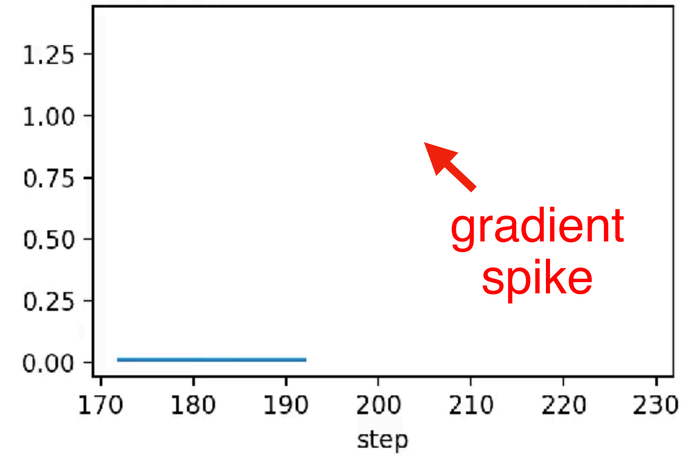


$\lambda_1(H_{\text{eff}})$

oscillation magnitude



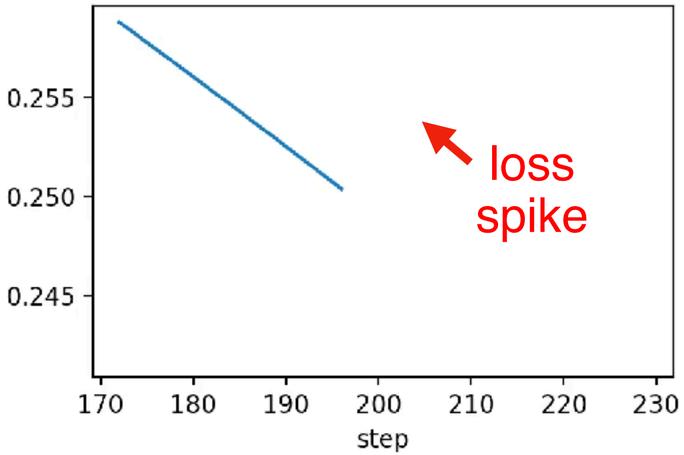
$\|\nabla L(w)\|^2$



$\lambda_1(H)$

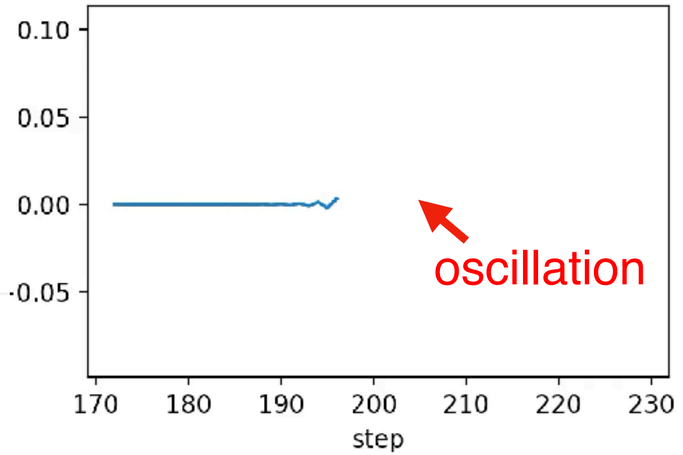
ν

Train Loss

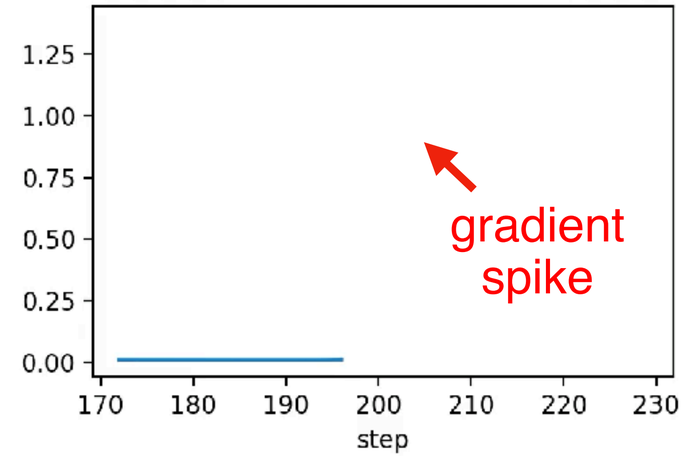


$\lambda_1(H_{\text{eff}})$

oscillation magnitude

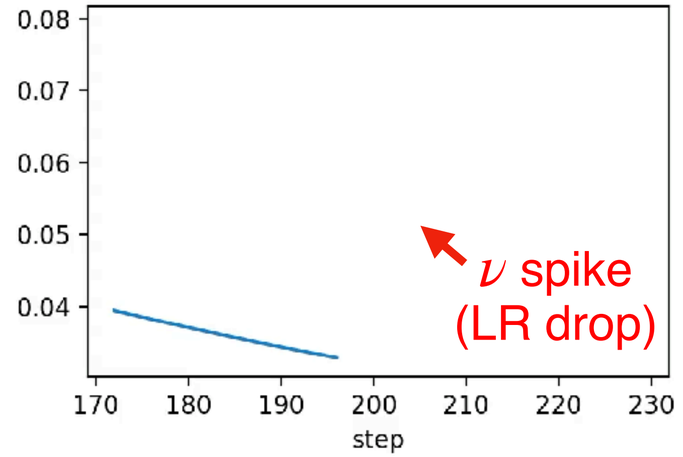


$\|\nabla L(w)\|^2$

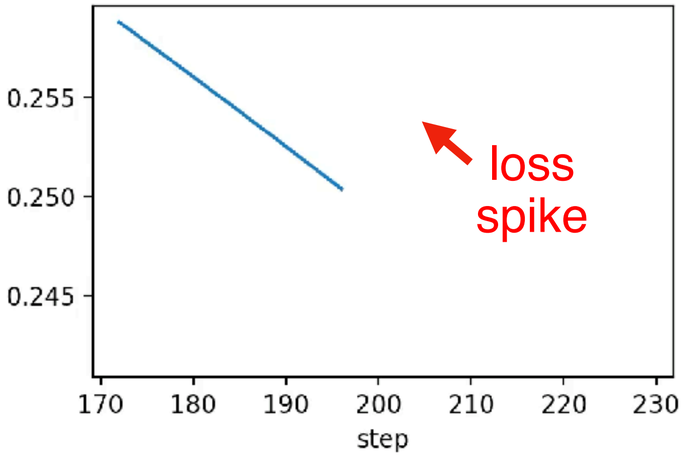


$\lambda_1(H)$

ν

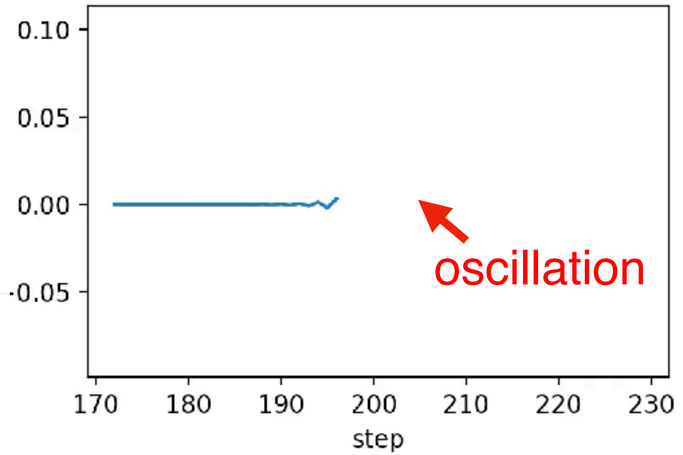


Train Loss

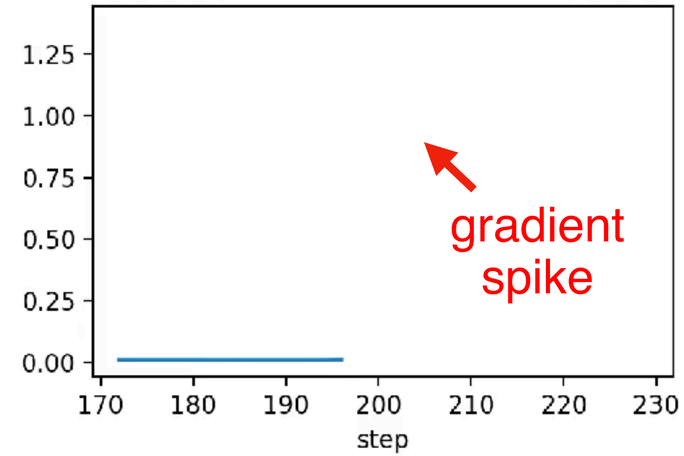


$\lambda_1(H_{\text{eff}})$

oscillation magnitude

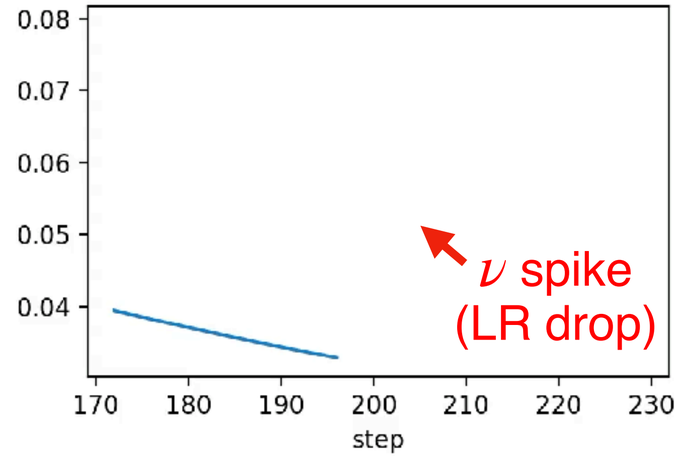


$\|\nabla L(w)\|^2$



$\lambda_1(H)$

ν



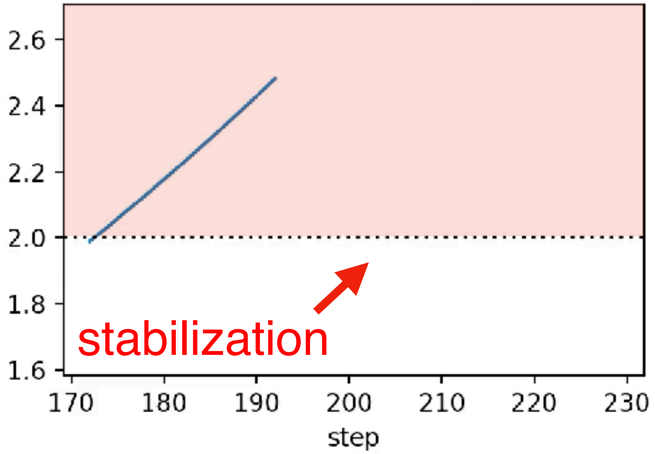
progressive sharpening → Hessian H

$\nabla L(w)^{\odot 2}$ → LRs P^{-1}

$H_{\text{eff}} := P^{-1}H$ → oscillation

Train Loss

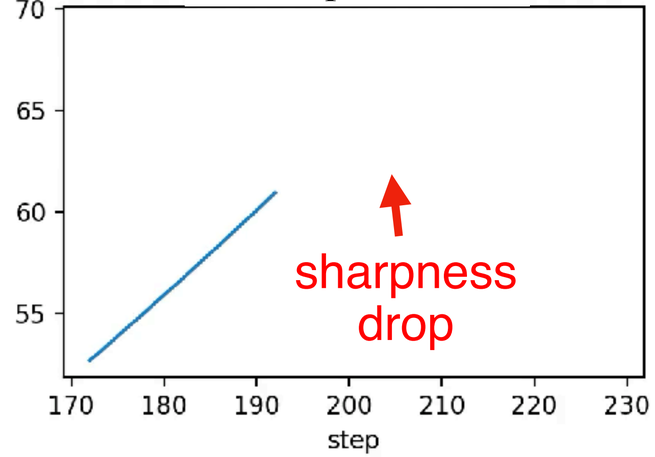
$\lambda_1(H_{\text{eff}})$



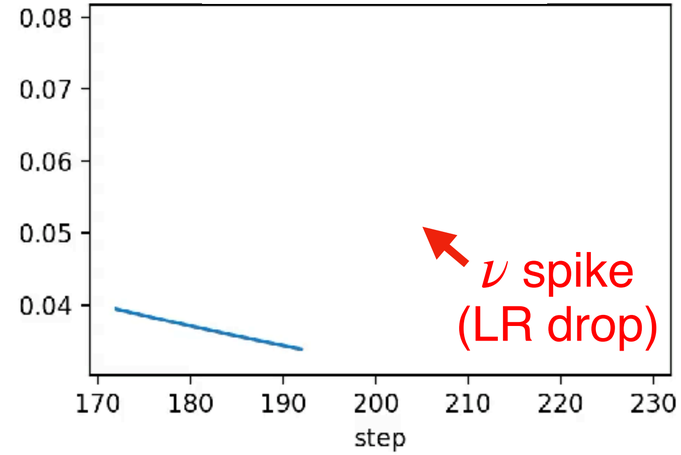
oscillation magnitude

$\|\nabla L(w)\|^2$

$\lambda_1(H)$



ν



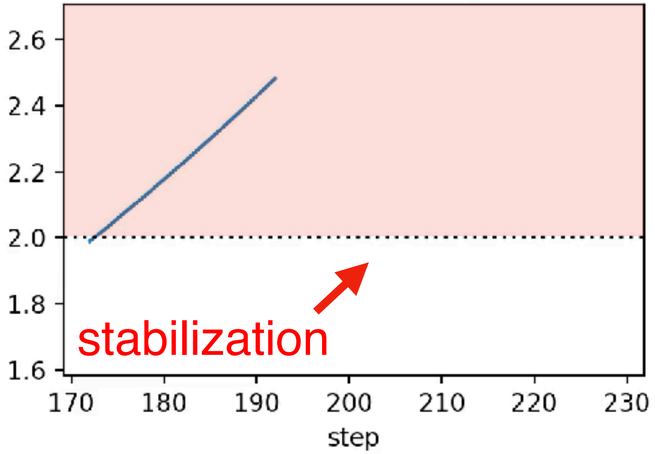
progressive sharpening → Hessian H

$\nabla L(w)^{\odot 2}$ → LRs P^{-1}

$H_{\text{eff}} := P^{-1}H$ → oscillation

Train Loss

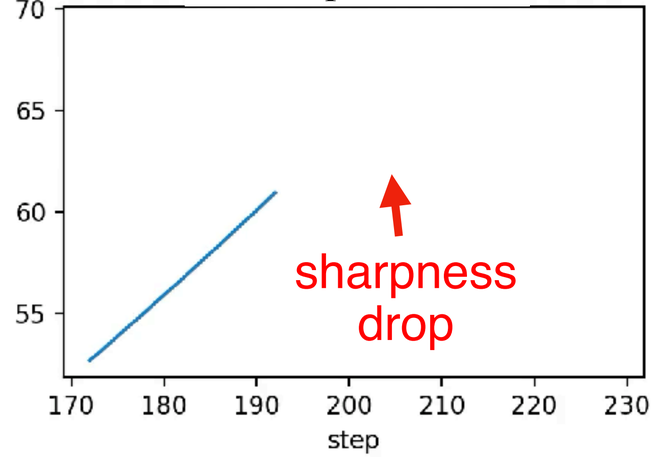
$\lambda_1(H_{\text{eff}})$



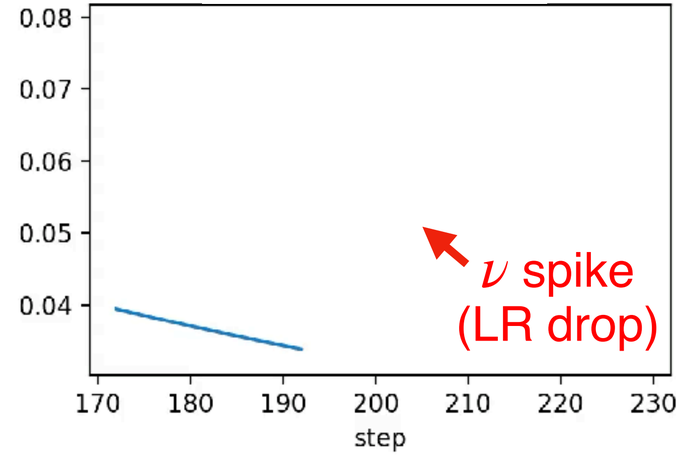
oscillation magnitude

$\|\nabla L(w)\|^2$

$\lambda_1(H)$



ν



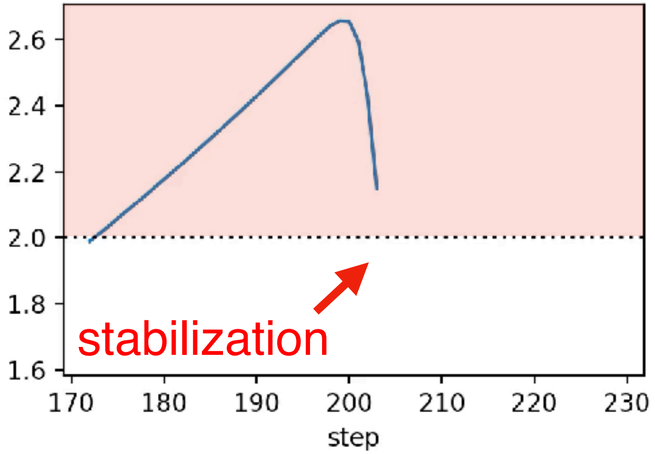
progressive sharpening → Hessian H

$\nabla L(w)^{\odot 2}$ → LRs P^{-1}

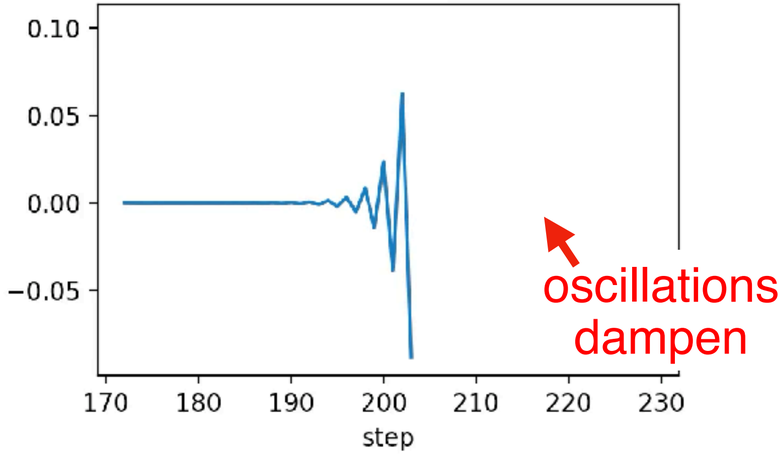
$H_{\text{eff}} := P^{-1}H$ → oscillation

Train Loss

$\lambda_1(H_{\text{eff}})$



oscillation magnitude



$\|\nabla L(w)\|^2$

$\lambda_1(H)$

ν

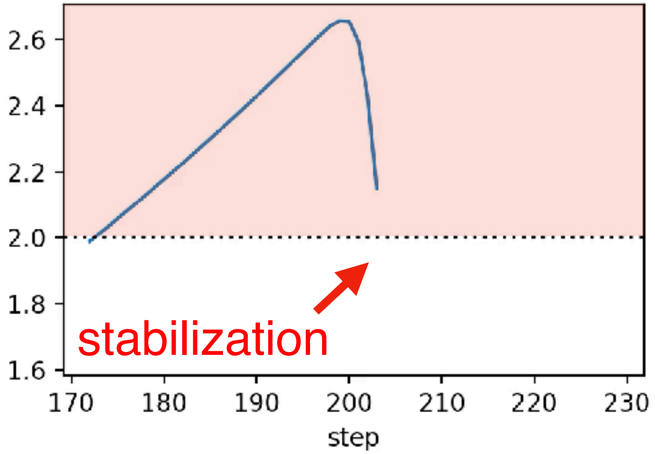
progressive sharpening → Hessian H

$\nabla L(w)^{\odot 2}$ → LRs P^{-1}

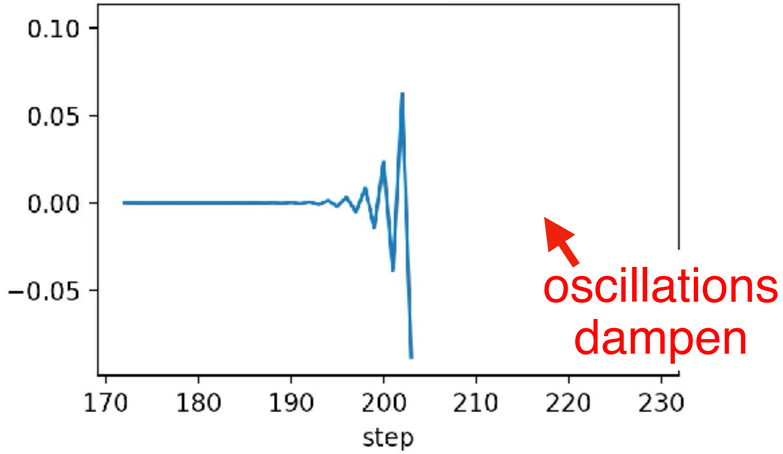
$H_{\text{eff}} := P^{-1}H$ → oscillation

Train Loss

$\lambda_1(H_{\text{eff}})$



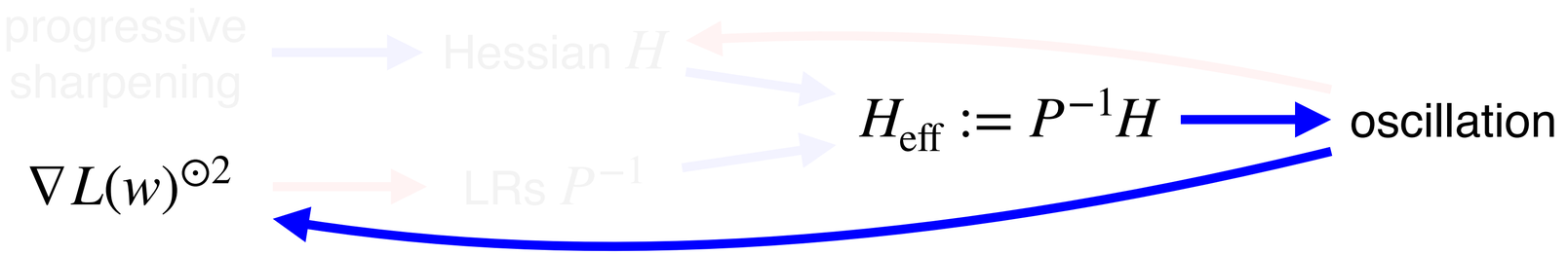
oscillation magnitude



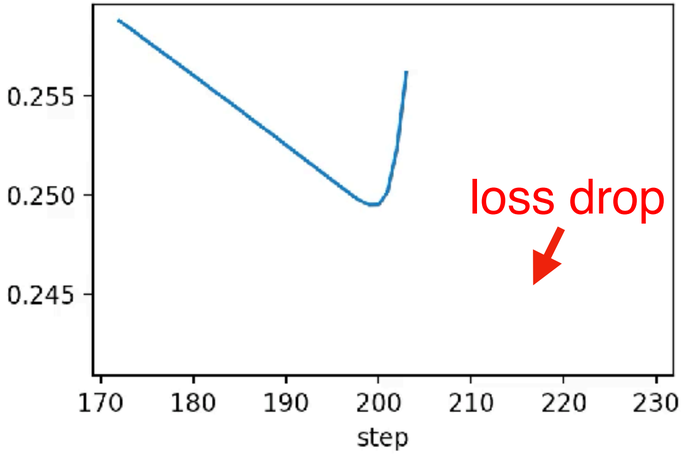
$\|\nabla L(w)\|^2$

$\lambda_1(H)$

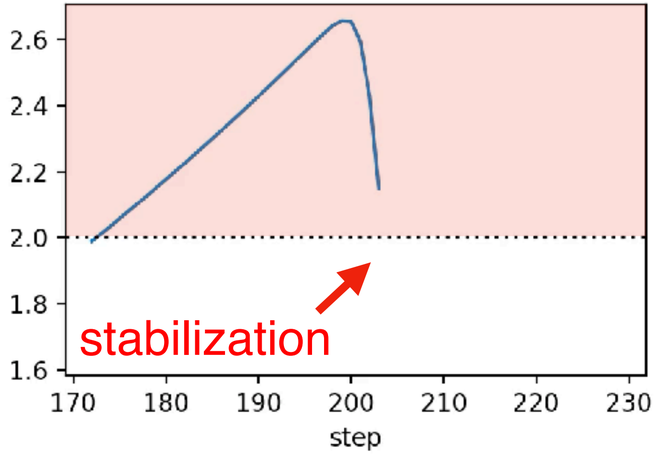
ν



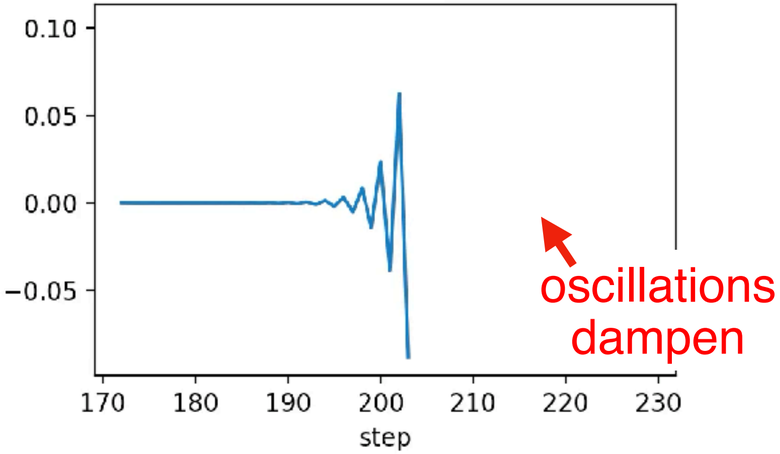
Train Loss



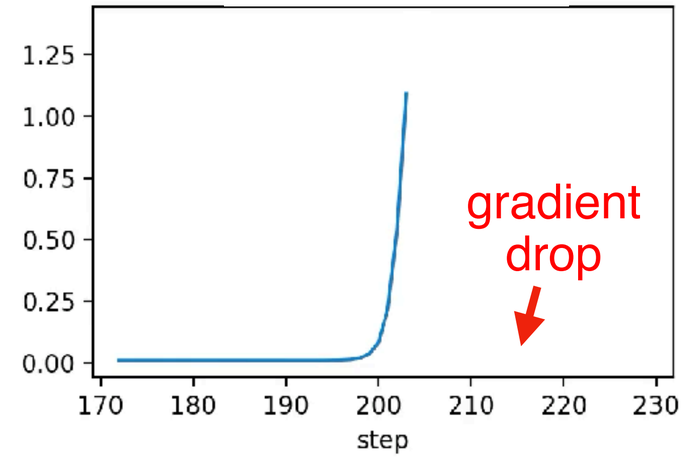
$\lambda_1(H_{\text{eff}})$

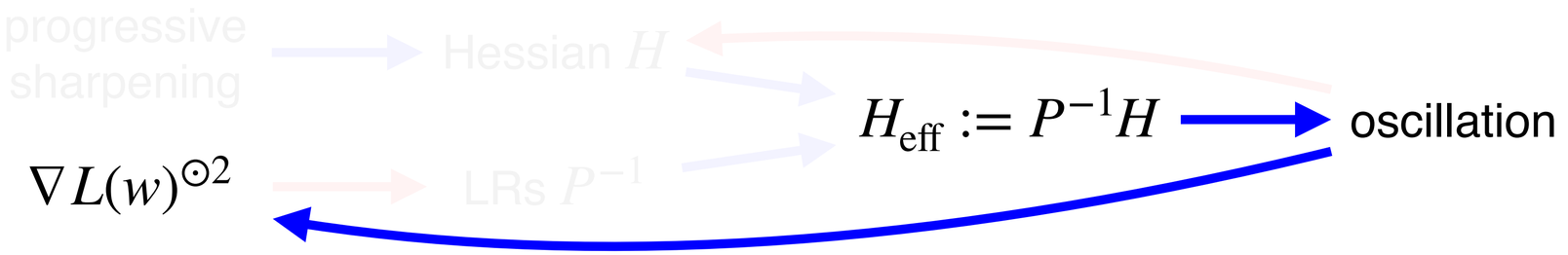


oscillation magnitude

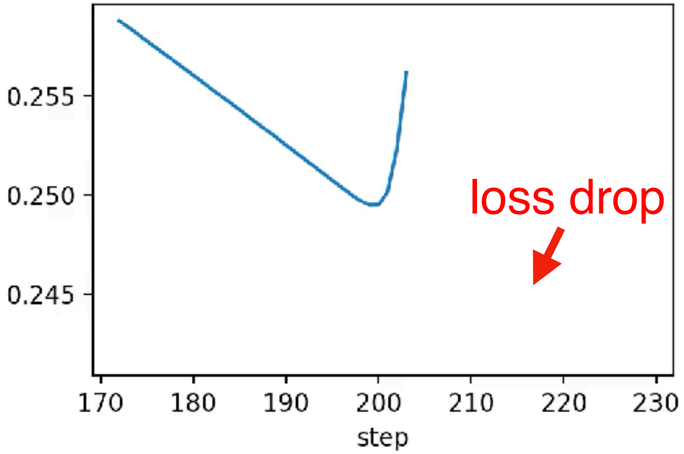


$\|\nabla L(w)\|^2$

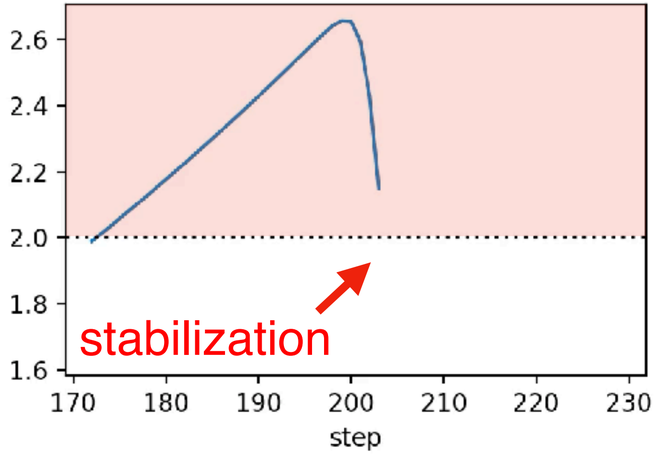




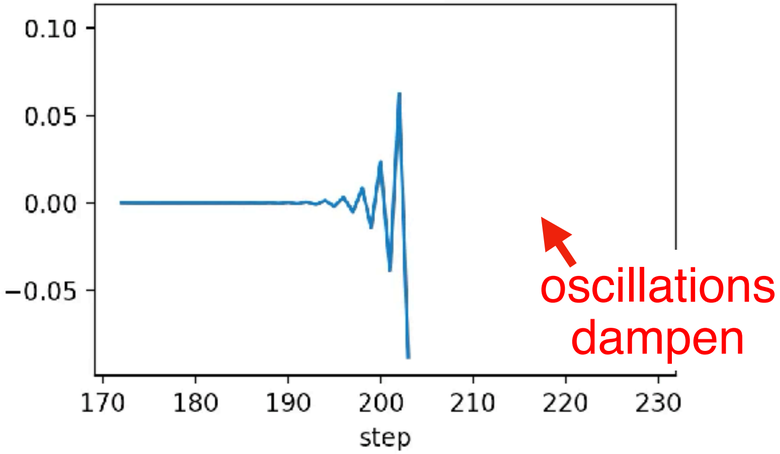
Train Loss



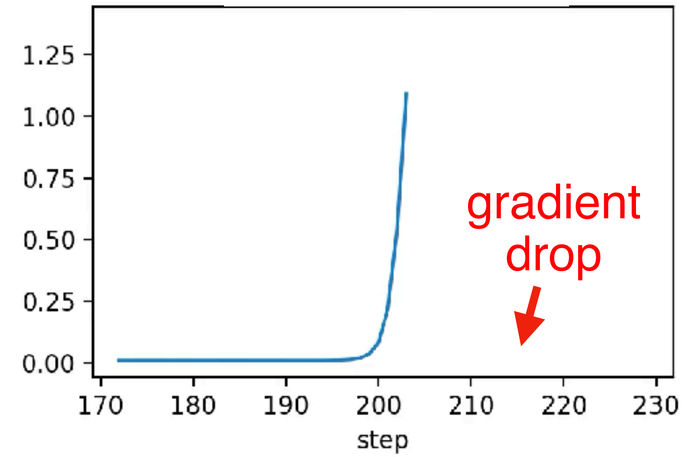
$\lambda_1(H_{\text{eff}})$

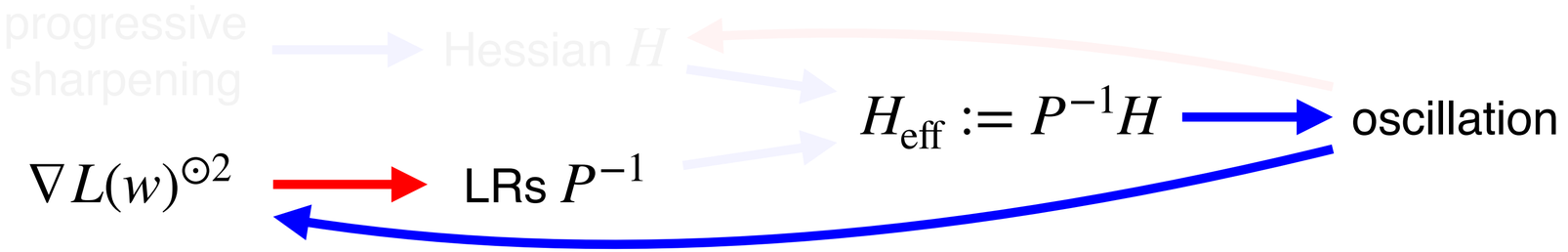


oscillation magnitude

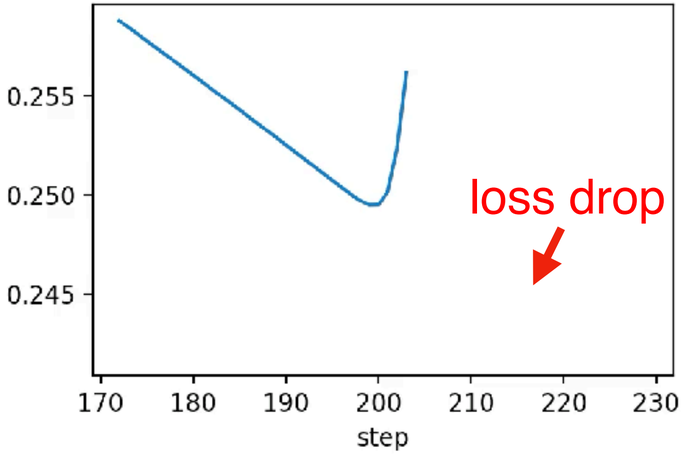


$\|\nabla L(w)\|^2$

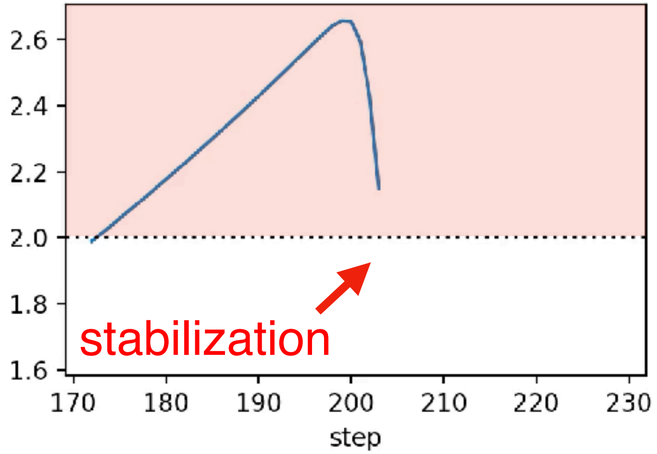




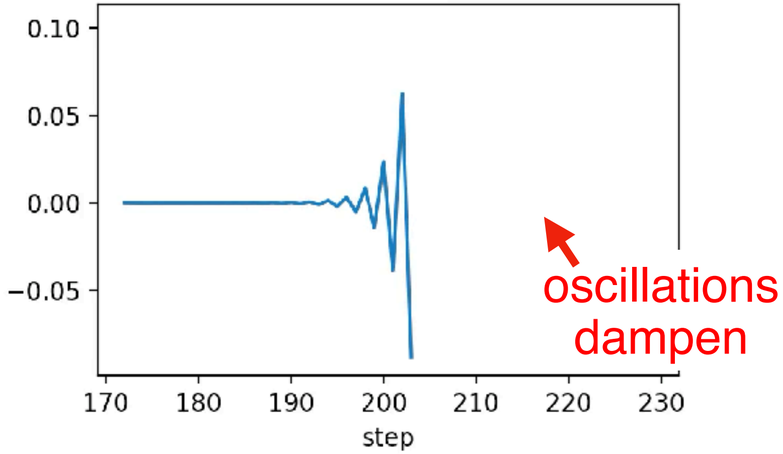
Train Loss



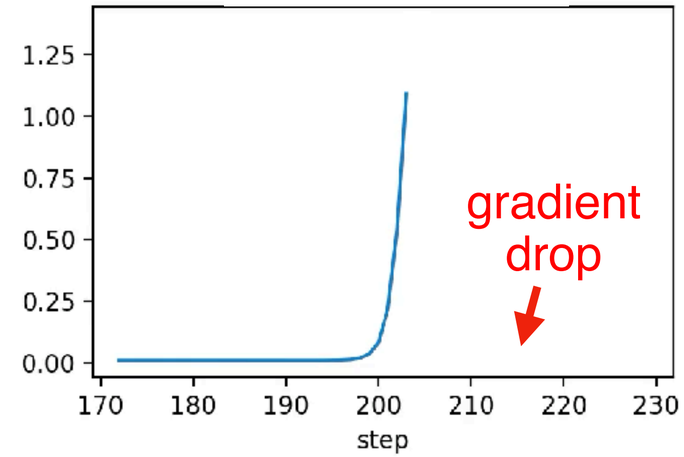
$\lambda_1(H_{\text{eff}})$



oscillation magnitude

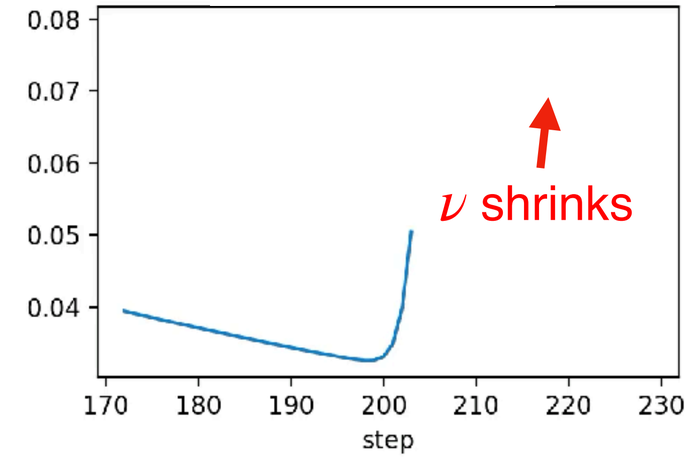


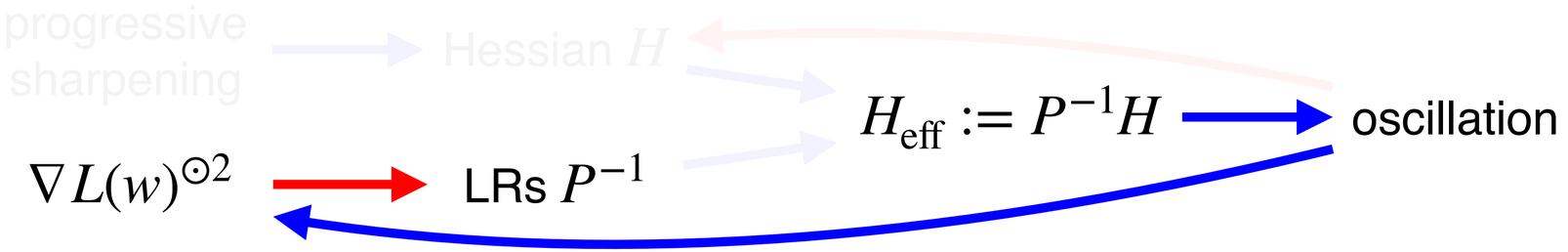
$\|\nabla L(w)\|^2$



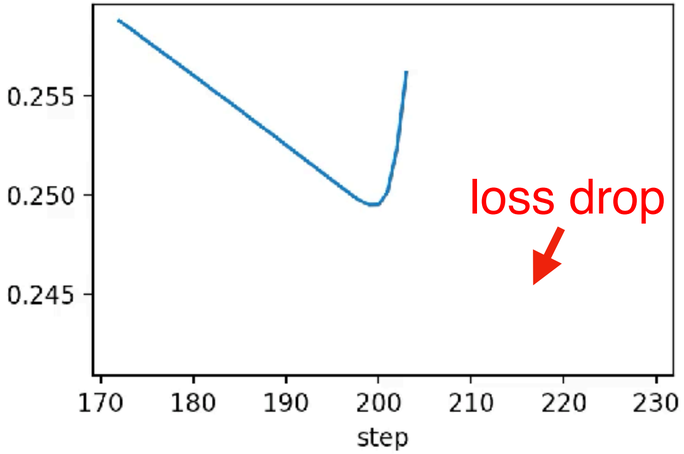
$\lambda_1(H)$

ν

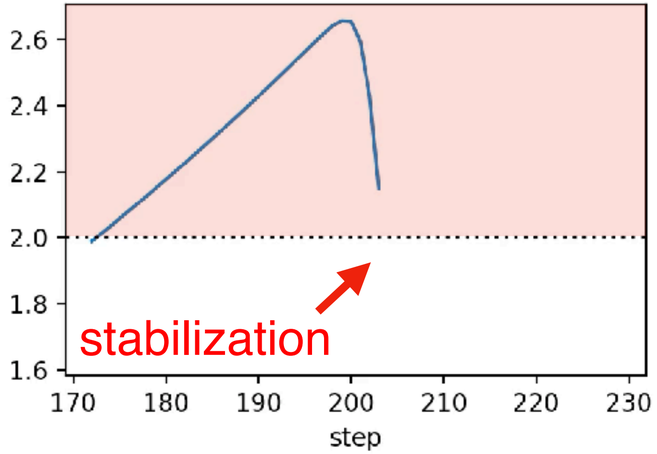




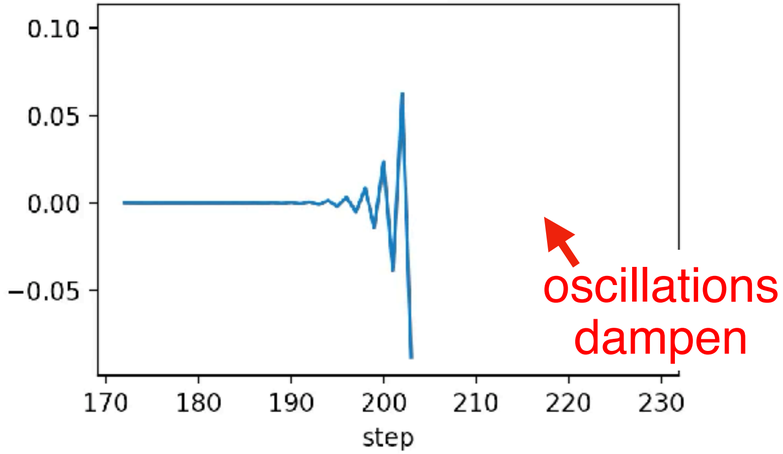
Train Loss



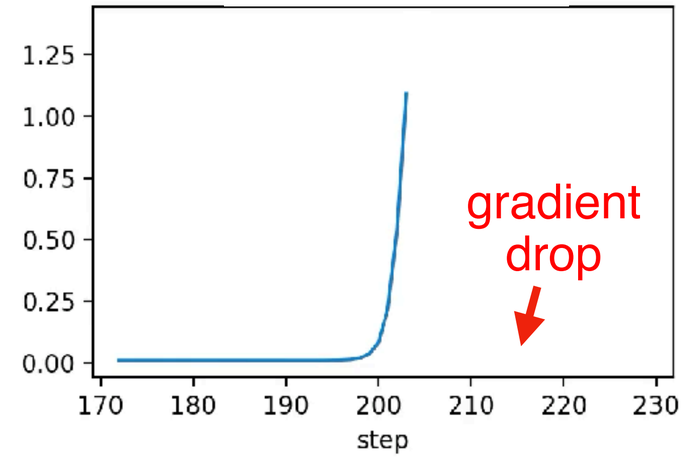
$\lambda_1(H_{\text{eff}})$



oscillation magnitude

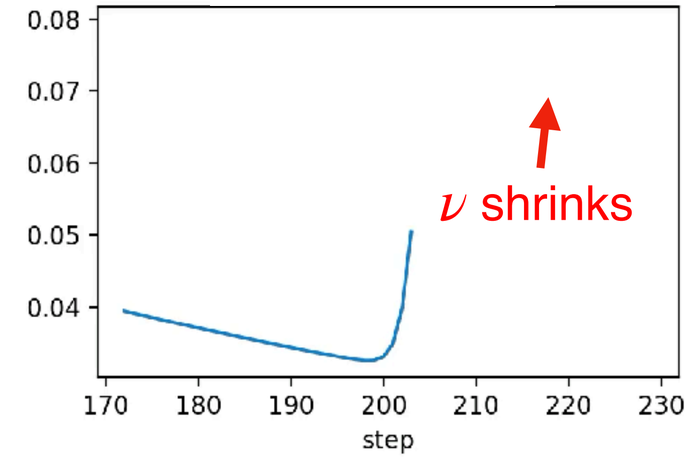


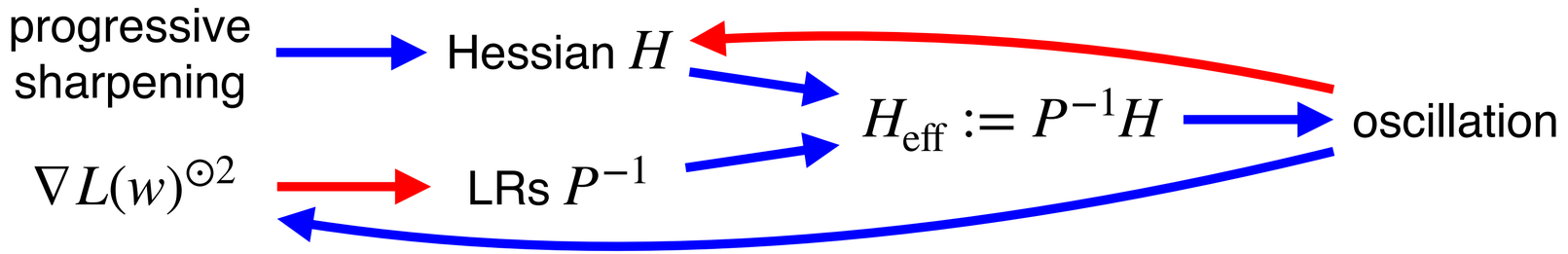
$\|\nabla L(w)\|^2$

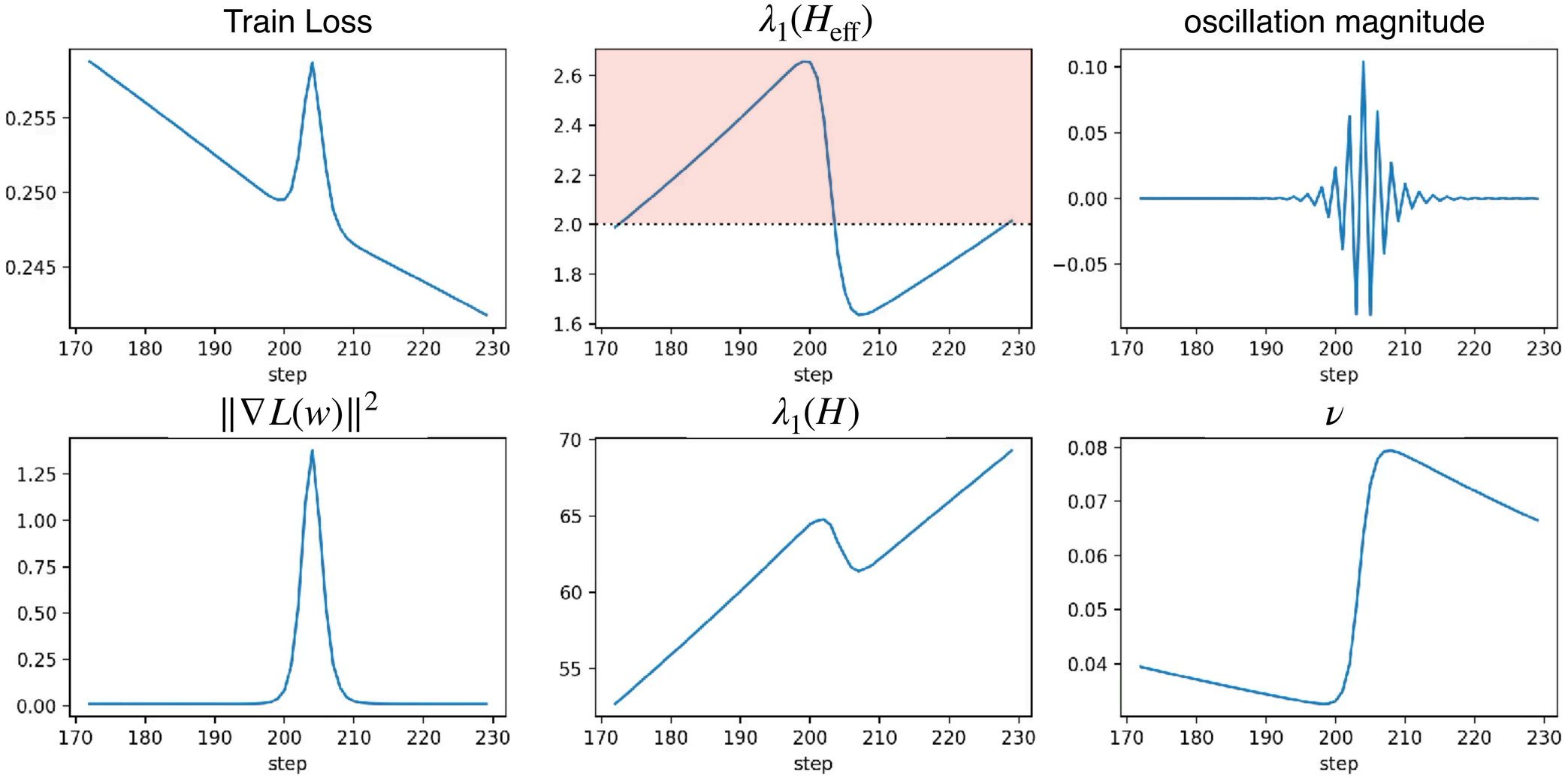


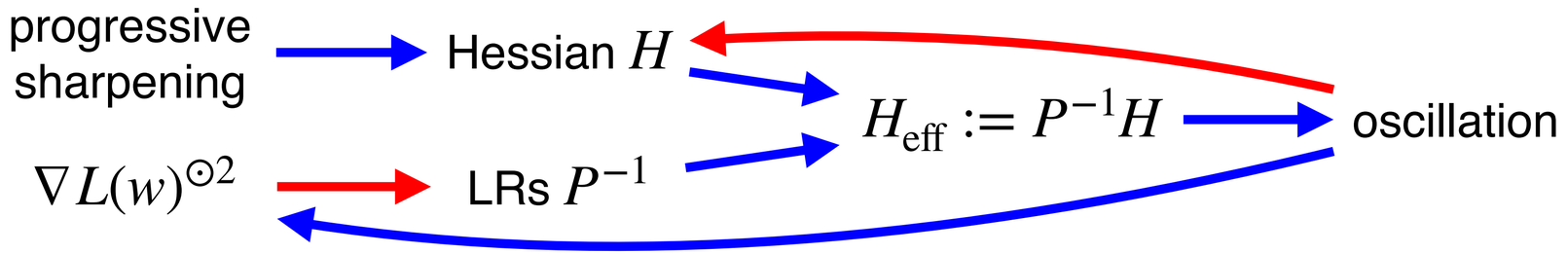
$\lambda_1(H)$

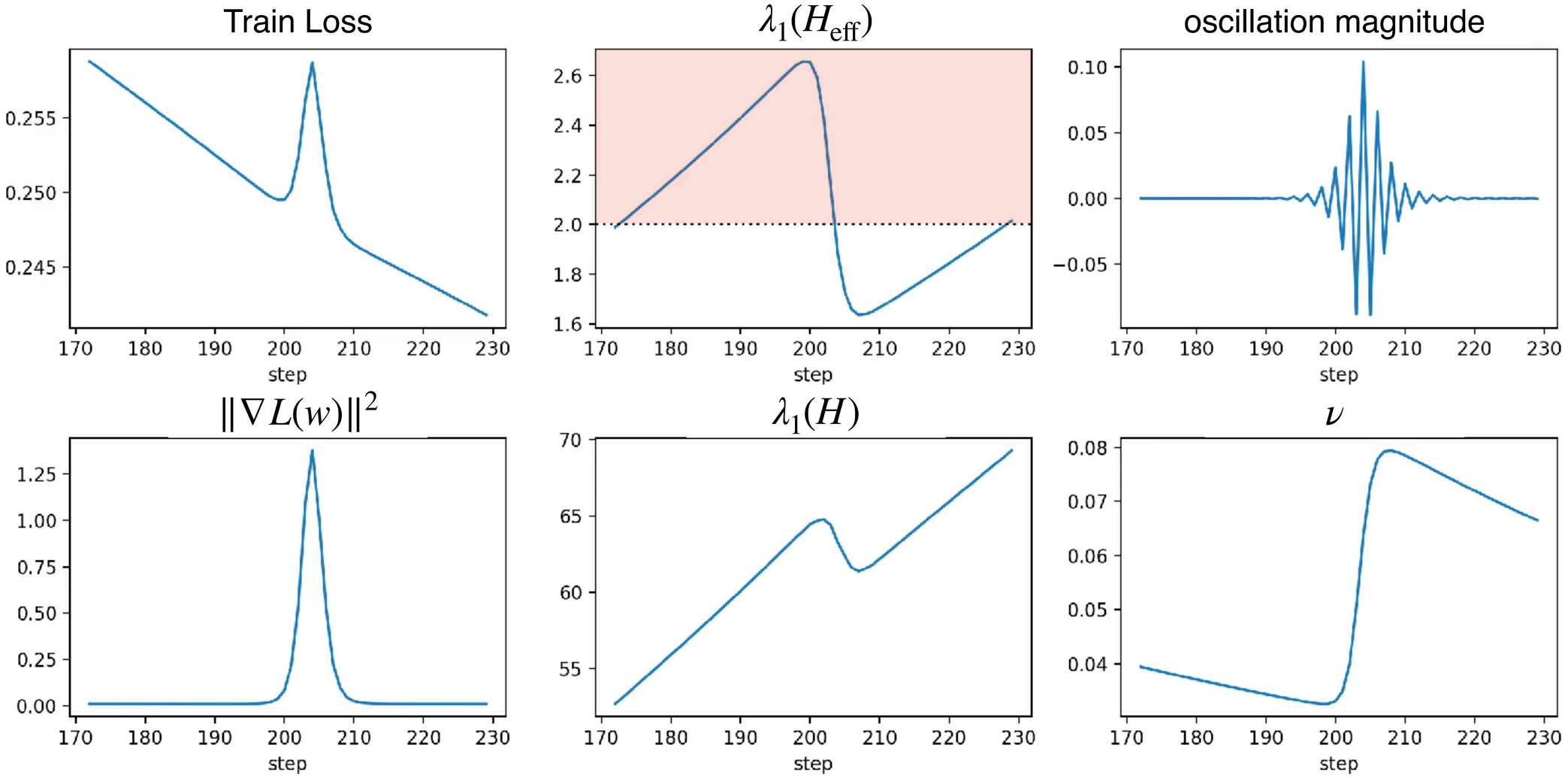
ν

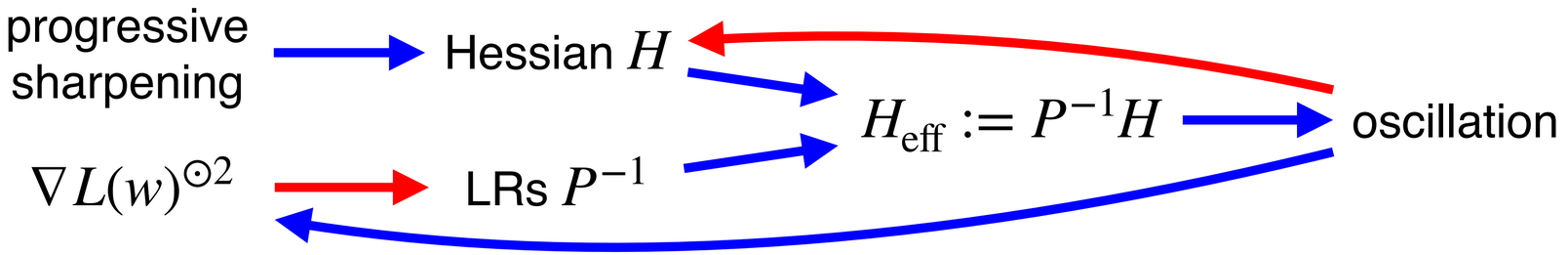



progressive sharpening
Hessian H
$H_{\text{eff}} := P^{-1}H$
oscillation
$\nabla L(w)^{\odot 2}$
LRs P^{-1}

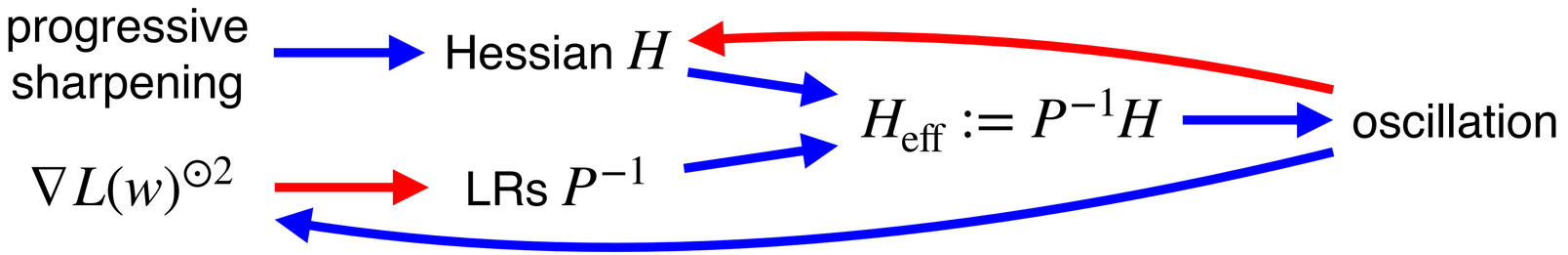

Train Loss
$\lambda_1(H_{\text{eff}})$
oscillation magnitude
$\|\nabla L(w)\|^2$
$\lambda_1(H)$
ν
step


progressive sharpening
Hessian H
$H_{\text{eff}} := P^{-1}H$
oscillation
$\nabla L(w)^{\odot 2}$
LRs P^{-1}


Train Loss
$\lambda_1(H_{\text{eff}})$
oscillation magnitude
$\|\nabla L(w)\|^2$
$\lambda_1(H)$
ν
step

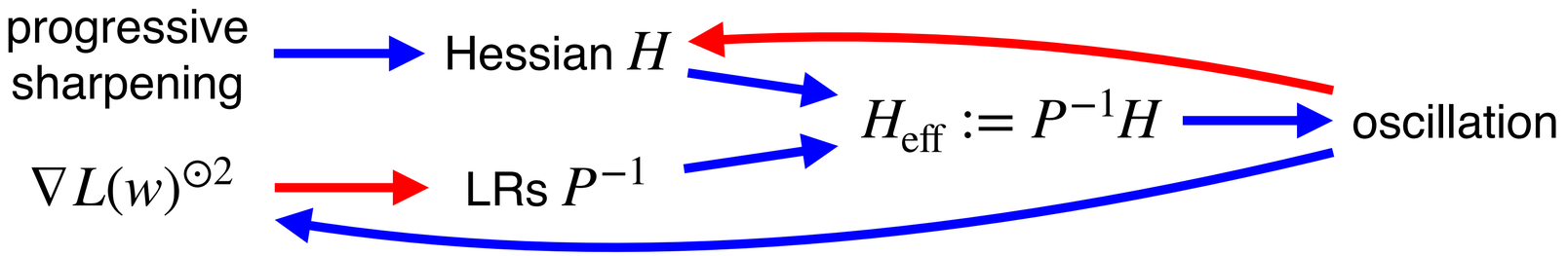


How can we analyze this system?



How can we analyze this system?

- The fine-grained dynamics are super complicated… 🤕



How can we analyze this system?

- The fine-grained dynamics are super complicated… 🤕
- Luckily, the central flows time-averaging argument generalizes easily!

Sketch: Deriving the RMSProp Central Flow

- Pretend w_t oscillates around a ***central flow*** $\overline{w}(t)$ with covariance $\Sigma(t)$:

$$w_t = \overline{w}(t) + \delta_t \qquad \text{where} \qquad \mathbb{E}[\delta_t \delta_t^T] = \Sigma(t)$$

Sketch: Deriving the RMSProp Central Flow

- Pretend w_t oscillates around a ***central flow*** $\overline{w}(t)$ with covariance $\Sigma(t)$:

$$w_t = \overline{w}(t) + \delta_t \qquad \text{where} \qquad \mathbb{E}[\delta_t \delta_t^T] = \Sigma(t)$$

- We again assume $\overline{w}(t)$ follows the average gradient:

$$\frac{d\overline{w}}{dt} = -\frac{\eta}{\sqrt{\nu}} \mathbb{E}\Big[\nabla L(w)\Big]$$

Sketch: Deriving the RMSProp Central Flow

- Pretend w_t oscillates around a ***central flow*** $\overline{w}(t)$ with covariance $\Sigma(t)$:

$$w_t = \overline{w}(t) + \delta_t \qquad \text{where} \qquad \mathbb{E}[\delta_t \delta_t^T] = \Sigma(t)$$

- We again assume $\overline{w}(t)$ follows the average gradient:

$$\frac{d\overline{w}}{dt} = -\frac{\eta}{\sqrt{\nu}} \mathbb{E}\Big[\nabla L(\overline{w} + \delta)\Big]$$

Sketch: Deriving the RMSProp Central Flow

- Pretend w_t oscillates around a ***central flow*** $\overline{w}(t)$ with covariance $\Sigma(t)$:

$$w_t = \overline{w}(t) + \delta_t \qquad \text{where} \qquad \mathbb{E}[\delta_t \delta_t^T] = \Sigma(t)$$

- We again assume $\overline{w}(t)$ follows the average gradient:

$$\frac{d\overline{w}}{dt} = -\frac{\eta}{\sqrt{\nu}} \mathbb{E}\left[\nabla L(\overline{w}) + \nabla^2 L(\overline{w})\delta + \frac{1}{2}\nabla \left\langle \nabla^2 L(\overline{w}), \delta\delta^T \right\rangle \right] + \mathcal{O}(\delta^3)$$

Sketch: Deriving the RMSProp Central Flow

- Pretend w_t oscillates around a ***central flow*** $\overline{w}(t)$ with covariance $\Sigma(t)$:

$$w_t = \overline{w}(t) + \delta_t \qquad \text{where} \qquad \mathbb{E}[\delta_t \delta_t^T] = \Sigma(t)$$

- We again assume $\overline{w}(t)$ follows the average gradient:

$$\frac{d\overline{w}}{dt} = -\frac{\eta}{\sqrt{\nu}} \mathbb{E}\Big[\nabla L(\overline{w}) + \nabla^2 L(\overline{w}) \overbrace{\delta}^{0} + \frac{1}{2} \nabla \langle \nabla^2 L(\overline{w}), \delta\delta^T \rangle \Big] + \mathcal{O}(\delta^3)$$

Sketch: Deriving the RMSProp Central Flow

- Pretend w_t oscillates around a ***central flow*** $\overline{w}(t)$ with covariance $\Sigma(t)$:

$$w_t = \overline{w}(t) + \delta_t \qquad \text{where} \qquad \mathbb{E}[\delta_t \delta_t^T] = \Sigma(t)$$

- We again assume $\overline{w}(t)$ follows the average gradient:

$$\frac{d\overline{w}}{dt} = -\frac{\eta}{\sqrt{\nu}} \mathbb{E}\Big[\underbrace{\nabla L(\overline{w})}_{\text{gradient}} + \underbrace{\frac{1}{2} \nabla \big\langle \nabla^2 L(\overline{w}), \Sigma \big\rangle}_{\text{curvature penalty}} \Big]$$

Sketch: Deriving the RMSProp Central Flow

- Pretend w_t oscillates around a ***central flow*** $\overline{w}(t)$ with covariance $\Sigma(t)$:

$$w_t = \overline{w}(t) + \delta_t \qquad \text{where} \qquad \mathbb{E}[\delta_t \delta_t^T] = \Sigma(t)$$

- We again assume $\overline{w}(t)$ follows the average gradient:

$$\frac{d\overline{w}}{dt} = -\frac{\eta}{\sqrt{\nu}} \mathbb{E}\Big[\underbrace{\nabla L(\overline{w})}_{\text{gradient}} + \underbrace{\frac{1}{2}\nabla \langle \nabla^2 L(\overline{w}), \Sigma \rangle}_{\text{curvature penalty}} \Big]$$

- We assume $\nu(t)$ follows the average **squared** gradient:

$$\frac{d\nu}{dt} = \frac{1-\beta_2}{\beta_2}\Big[\mathbb{E}\big[\nabla L(w)^{\odot 2}\big] - \nu\Big]$$

Sketch: Deriving the RMSProp Central Flow

- Pretend w_t oscillates around a ***central flow*** $\overline{w}(t)$ with covariance $\Sigma(t)$:

$$w_t = \overline{w}(t) + \delta_t \qquad \text{where} \qquad \mathbb{E}[\delta_t \delta_t^T] = \Sigma(t)$$

- We again assume $\overline{w}(t)$ follows the average gradient:

$$\frac{d\overline{w}}{dt} = -\frac{\eta}{\sqrt{\nu}} \mathbb{E}\Big[\underbrace{\nabla L(\overline{w})}_{\text{gradient}} + \underbrace{\frac{1}{2} \nabla \big\langle \nabla^2 L(\overline{w}), \Sigma \big\rangle}_{\text{curvature penalty}} \Big]$$

- We assume $\nu(t)$ follows the average **squared** gradient:

$$\frac{d\nu}{dt} = \frac{1-\beta_2}{\beta_2} \Big[\mathbb{E}\big[\nabla L(\overline{w} + \delta)^{\odot 2} \big] - \nu \Big]$$

Sketch: Deriving the RMSProp Central Flow

- Pretend w_t oscillates around a ***central flow*** $\overline{w}(t)$ with covariance $\Sigma(t)$:

$$w_t = \overline{w}(t) + \delta_t \qquad \text{where} \qquad \mathbb{E}[\delta_t \delta_t^T] = \Sigma(t)$$

- We again assume $\overline{w}(t)$ follows the average gradient:

$$\frac{d\overline{w}}{dt} = -\frac{\eta}{\sqrt{\nu}} \mathbb{E}\Big[\underbrace{\nabla L(\overline{w})}_{\text{gradient}} + \underbrace{\frac{1}{2}\nabla \langle \nabla^2 L(\overline{w}), \Sigma \rangle}_{\text{curvature penalty}} \Big]$$

- We assume $\nu(t)$ follows the average **squared** gradient:

$$\frac{d\nu}{dt} = \frac{1-\beta_2}{\beta_2}\Big[\nabla L(\overline{w})^{\odot 2} + \operatorname{diag}[H(\overline{w})\Sigma H(\overline{w})] - \nu \Big]$$

Sketch: Deriving the RMSProp Central Flow

- If we knew $\Sigma(t)$, then $\overline{w}, \nu$ would evolve by:

$$\frac{d\overline{w}}{dt} = -\frac{\eta}{\sqrt{\nu}} \mathbb{E}\Big[\nabla L(\overline{w}) + \frac{1}{2} \nabla \big\langle \nabla^2 L(\overline{w}), \Sigma \big\rangle \Big]$$

$$\frac{d\nu}{dt} = \frac{1-\beta_2}{\beta_2} \Big[\nabla L(\overline{w})^{\odot 2} + \mathrm{diag}[H(\overline{w}) \Sigma H(\overline{w})] - \nu \Big]$$

Sketch: Deriving the RMSProp Central Flow

- If we knew $\Sigma(t)$, then $\overline{w}, \nu$ would evolve by:

$$\frac{d\overline{w}}{dt} = -\frac{\eta}{\sqrt{\nu}}\mathbb{E}\left[\nabla L(\overline{w}) + \frac{1}{2}\nabla\left\langle \nabla^2 L(\overline{w}), \Sigma\right\rangle\right]$$

$$\frac{d\nu}{dt} = \frac{1-\beta_2}{\beta_2}\left[\nabla L(\overline{w})^{\odot 2} + \operatorname{diag}[H(\overline{w})\Sigma H(\overline{w})] - \nu\right]$$

- We show there is only one choice of $\Sigma(t)$ which satisfies the constraints:
  1. **Positivity:** As a covariance matrix, $\Sigma(t) \succeq 0$

Sketch: Deriving the RMSProp Central Flow

- If we knew $\Sigma(t)$, then $\overline{w}, \nu$ would evolve by:

$$\frac{d\overline{w}}{dt} = -\frac{\eta}{\sqrt{\nu}}\mathbb{E}\left[\ \nabla L(\overline{w}) \ + \ \frac{1}{2}\nabla\left\langle \nabla^2 L(\overline{w}), \Sigma\right\rangle\ \right]$$

$$\frac{d\nu}{dt} = \frac{1-\beta_2}{\beta_2}\left[\nabla L(\overline{w})^{\odot 2} + \operatorname{diag}[H(\overline{w})\Sigma H(\overline{w})] - \nu\right]$$

- We show there is only one choice of $\Sigma(t)$ which satisfies the constraints:
  1. **Positivity:** As a covariance matrix, $\Sigma(t) \succeq 0$
  2. **Stability:** $H(\overline{w}) \preceq 2P(\nu)$ where $P(\nu) = \operatorname{diag}(\sqrt{\nu}/\eta)$

Sketch: Deriving the RMSProp Central Flow

- If we knew $\Sigma(t)$, then $\overline{w}, \nu$ would evolve by:

$$\frac{d\overline{w}}{dt} = -\frac{\eta}{\sqrt{\nu}}\mathbb{E}\left[\nabla L(\overline{w}) + \frac{1}{2}\nabla\langle\nabla^2 L(\overline{w}), \Sigma\rangle\right]$$

$$\frac{d\nu}{dt} = \frac{1-\beta_2}{\beta_2}\left[\nabla L(\overline{w})^{\odot 2} + \operatorname{diag}[H(\overline{w})\Sigma H(\overline{w})] - \nu\right]$$

- We show there is only one choice of $\Sigma(t)$ which satisfies the constraints:
  1. **Positivity:** As a covariance matrix, $\Sigma(t) \succeq 0$
  2. **Stability:** $H(\overline{w}) \preceq 2P(\nu)$ where $P(\nu) = \operatorname{diag}(\sqrt{\nu}/\eta)$
  3. **Complementarity:** $\operatorname{span}\Sigma(t) \subseteq \operatorname{span}[\text{unstable evecs}]$

Sketch: Deriving the RMSProp Central Flow

- If we knew $\Sigma(t)$, then $\overline{w}, \nu$ would evolve by:

$$\frac{d\overline{w}}{dt} = -\frac{\eta}{\sqrt{\nu}}\mathbb{E}\left[\ \nabla L(\overline{w}) \ + \ \frac{1}{2}\nabla\left\langle \nabla^2 L(\overline{w}), \Sigma\right\rangle\ \right]$$

$$\frac{d\nu}{dt} = \frac{1-\beta_2}{\beta_2}\left[\nabla L(\overline{w})^{\odot 2} + \operatorname{diag}[H(\overline{w})\Sigma H(\overline{w})] - \nu\right]$$

- We show there is only one choice of $\Sigma(t)$ which satisfies the constraints:

$$\underbrace{0 \ \preceq \ \overbrace{2P(\nu) - H(\overline{w})}^{\text{Stability}} \ \perp \ \overbrace{\Sigma \ \succeq \ 0}^{\text{Positivity}}}_{\text{Complementarity}} \qquad \text{for all } t$$

Sketch: Deriving the RMSProp Central Flow

- If we knew $\Sigma(t)$, then $\overline{w}, \nu$ would evolve by:

$$\frac{d\overline{w}}{dt} = -\frac{\eta}{\sqrt{\nu}}\mathbb{E}\left[\ \nabla L(\overline{w}) \ + \ \frac{1}{2}\nabla\left\langle \nabla^2 L(\overline{w}), \Sigma\right\rangle\ \right]$$

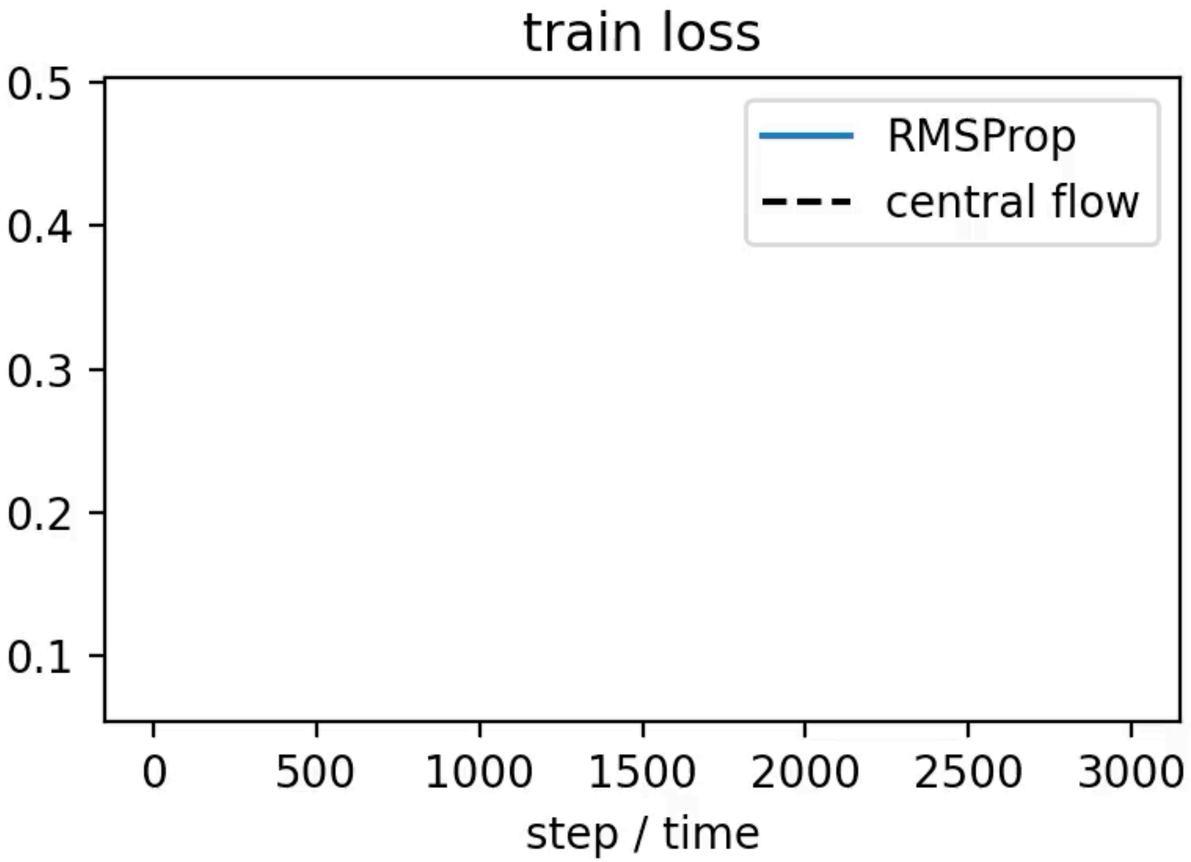
$$\frac{d\nu}{dt} = \frac{1-\beta_2}{\beta_2}\left[\nabla L(\overline{w})^{\odot 2} + \operatorname{diag}[H(\overline{w})\Sigma H(\overline{w})] - \nu\right]$$

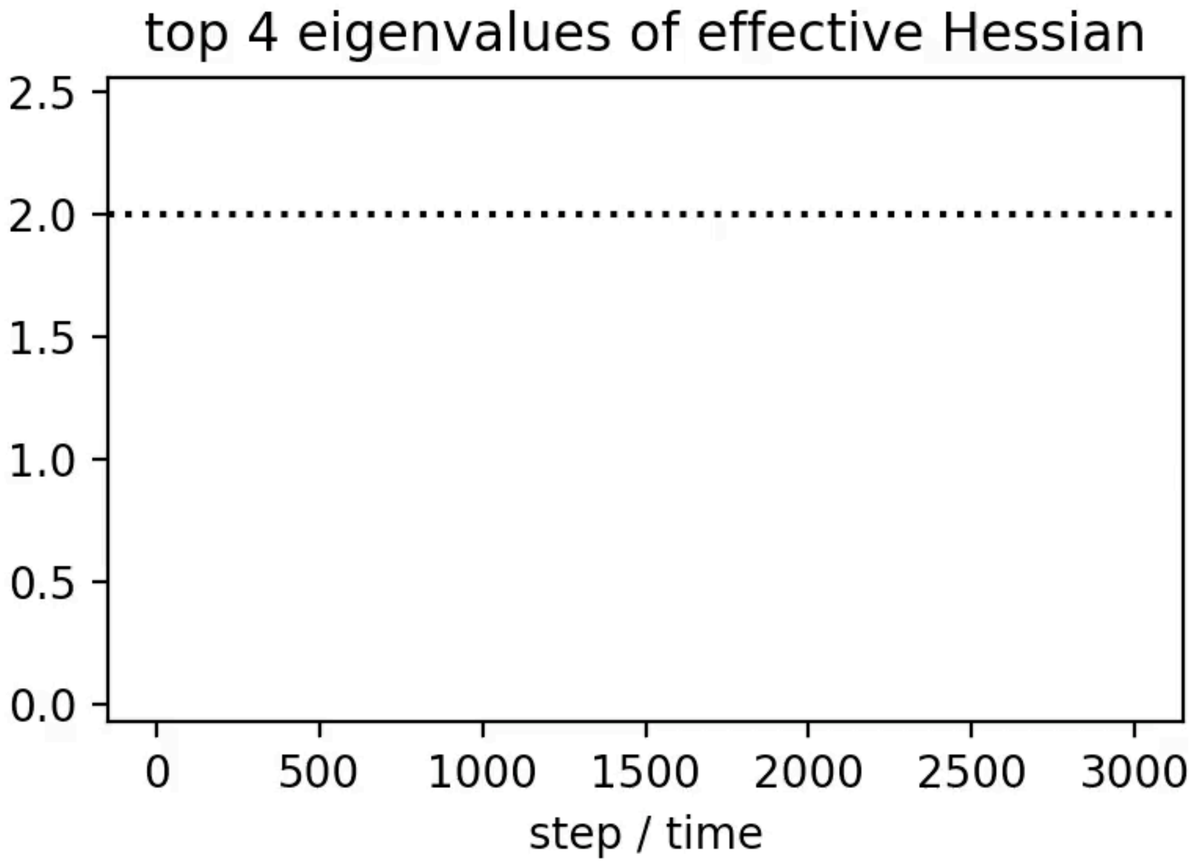
- We show there is only one choice of $\Sigma(t)$ which satisfies the constraints:

$$\underbrace{0 \ \preceq \ \overbrace{2P(\nu) - H(\overline{w})}^{\text{Stability}} \ \perp \ \overbrace{\Sigma \ \succeq \ 0}^{\text{Positivity}}}_{\text{Complementarity}} \qquad \text{for all } t$$

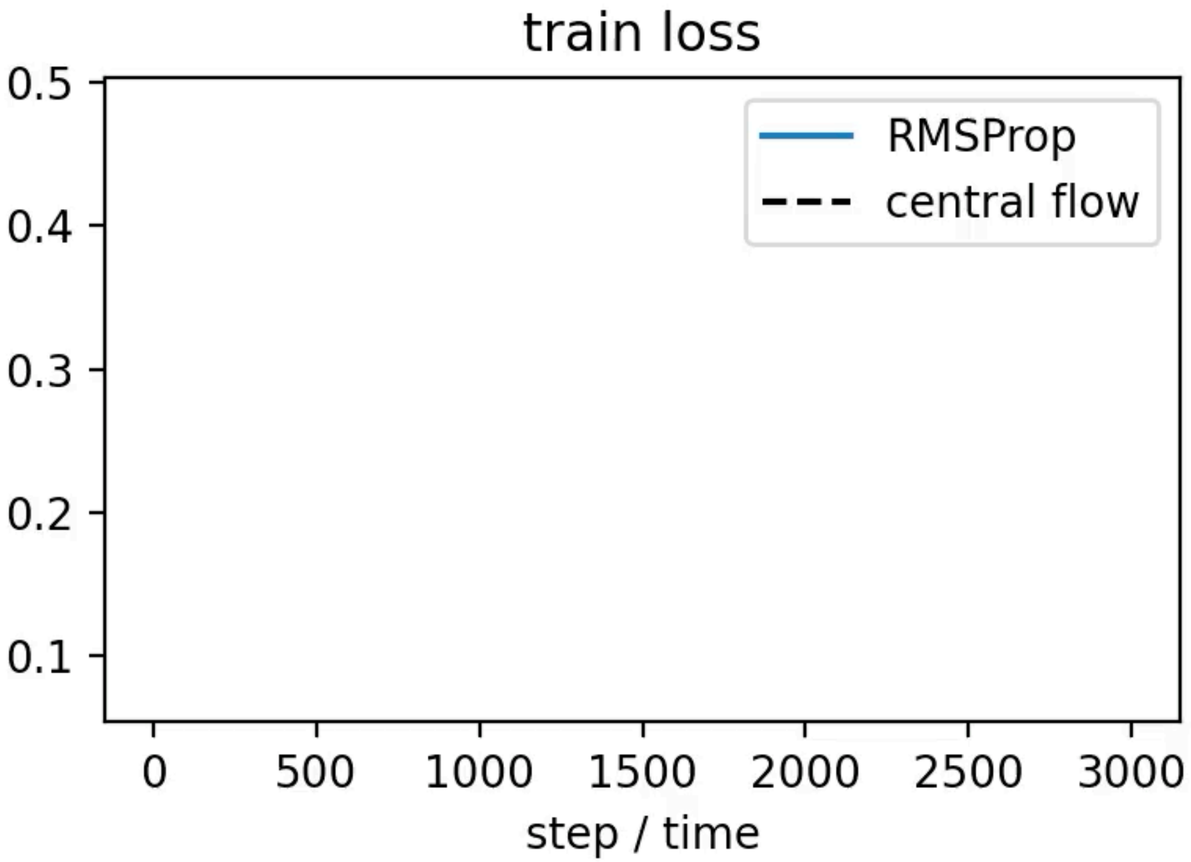
- This defines a **differential complementarity problem** (DCP) for $(\overline{w}, \nu, \Sigma)$ which uniquely determines the central flow $\{(\overline{w}(t), \nu(t), \Sigma(t))\}_{t\geq 0}$ and can be efficiently simulated

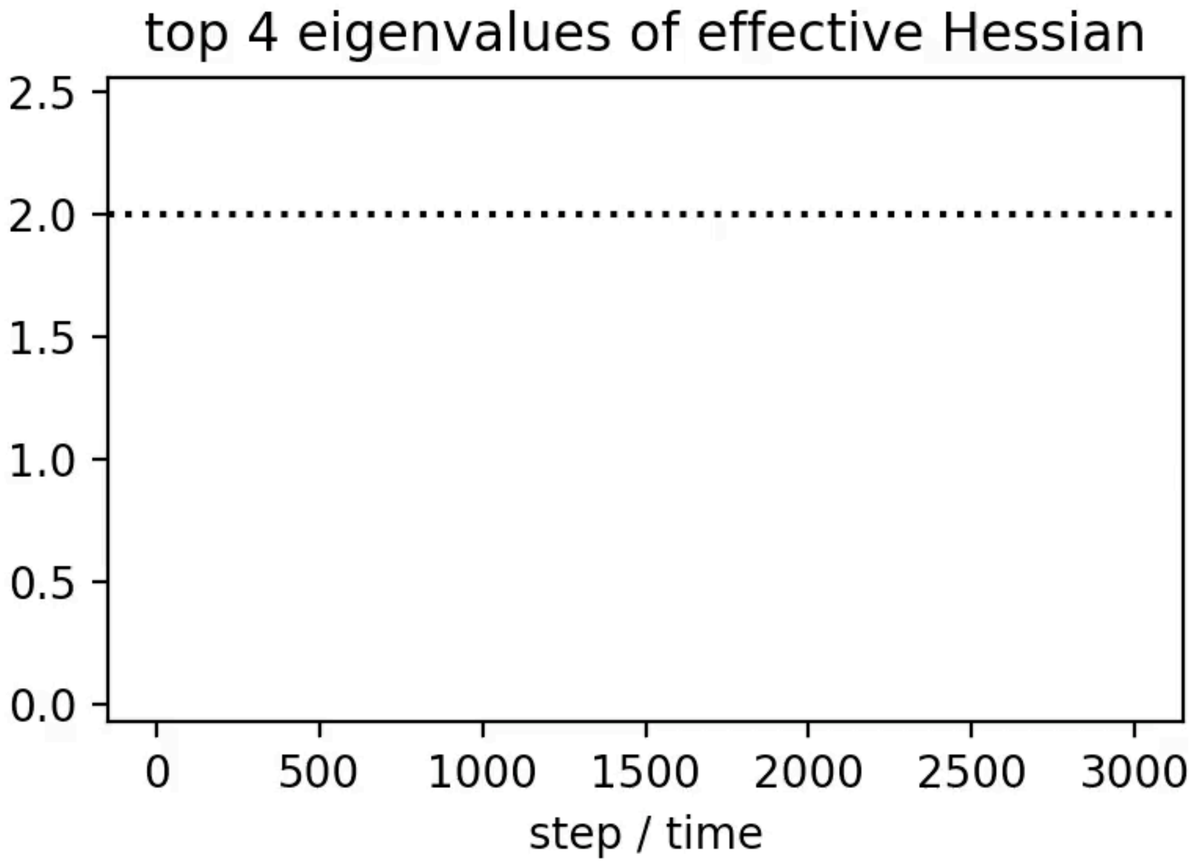
Validating The RMSProp Central Flow



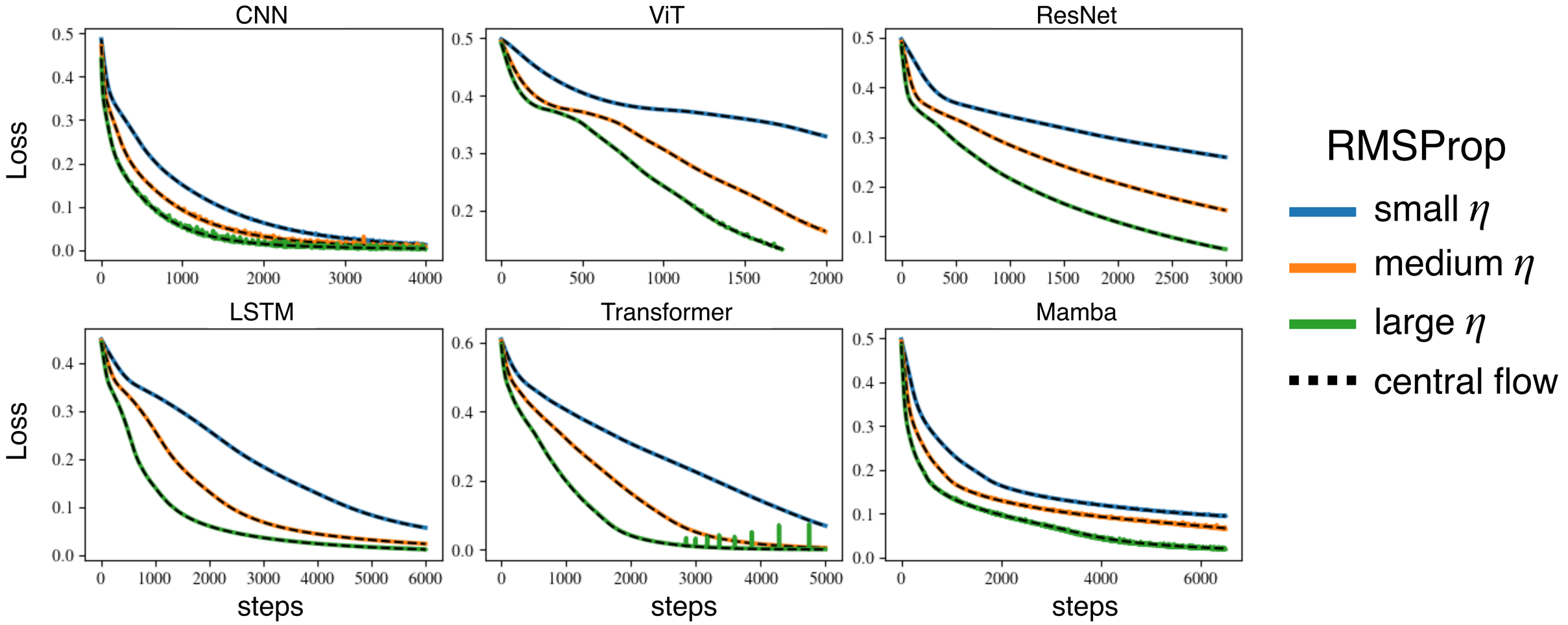


Validating The RMSProp Central Flow



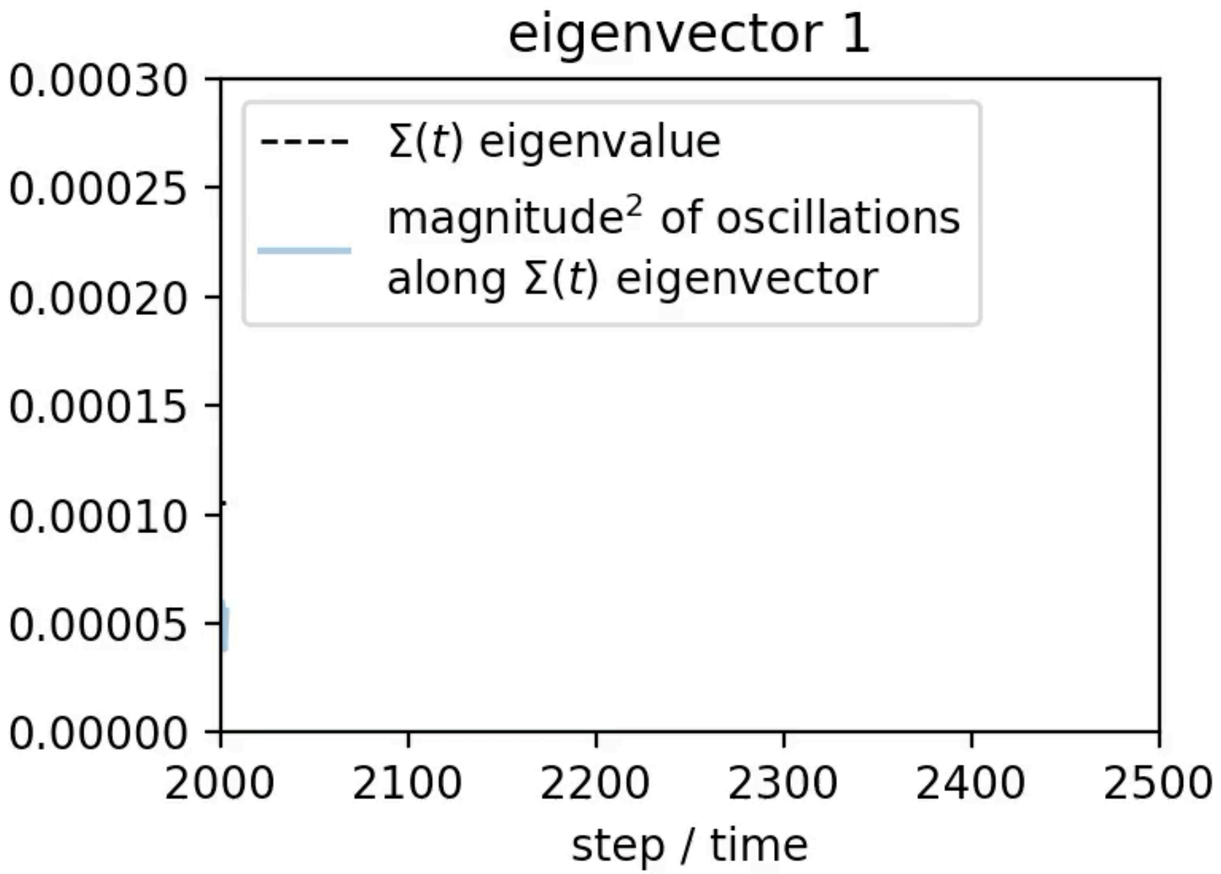


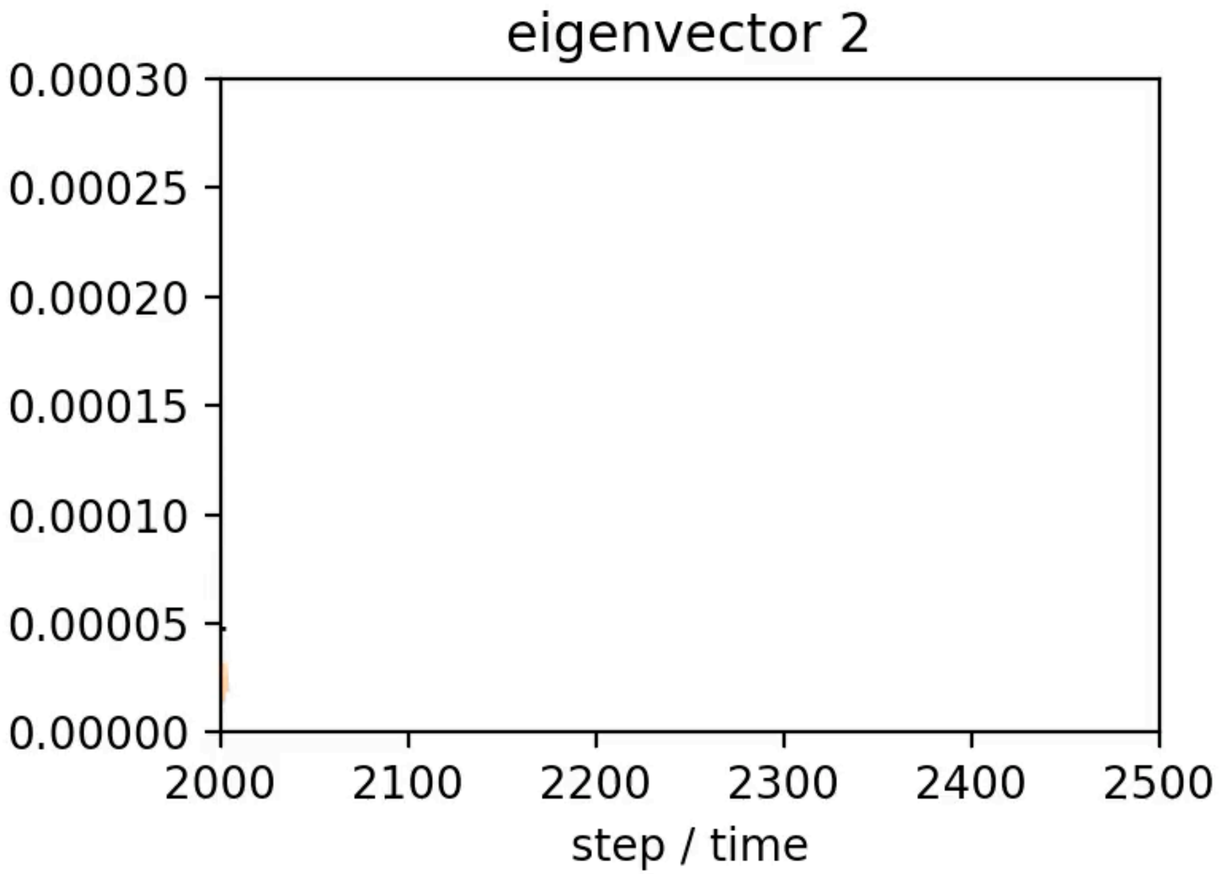
The RMSProp Central Flow Holds Across Architectures



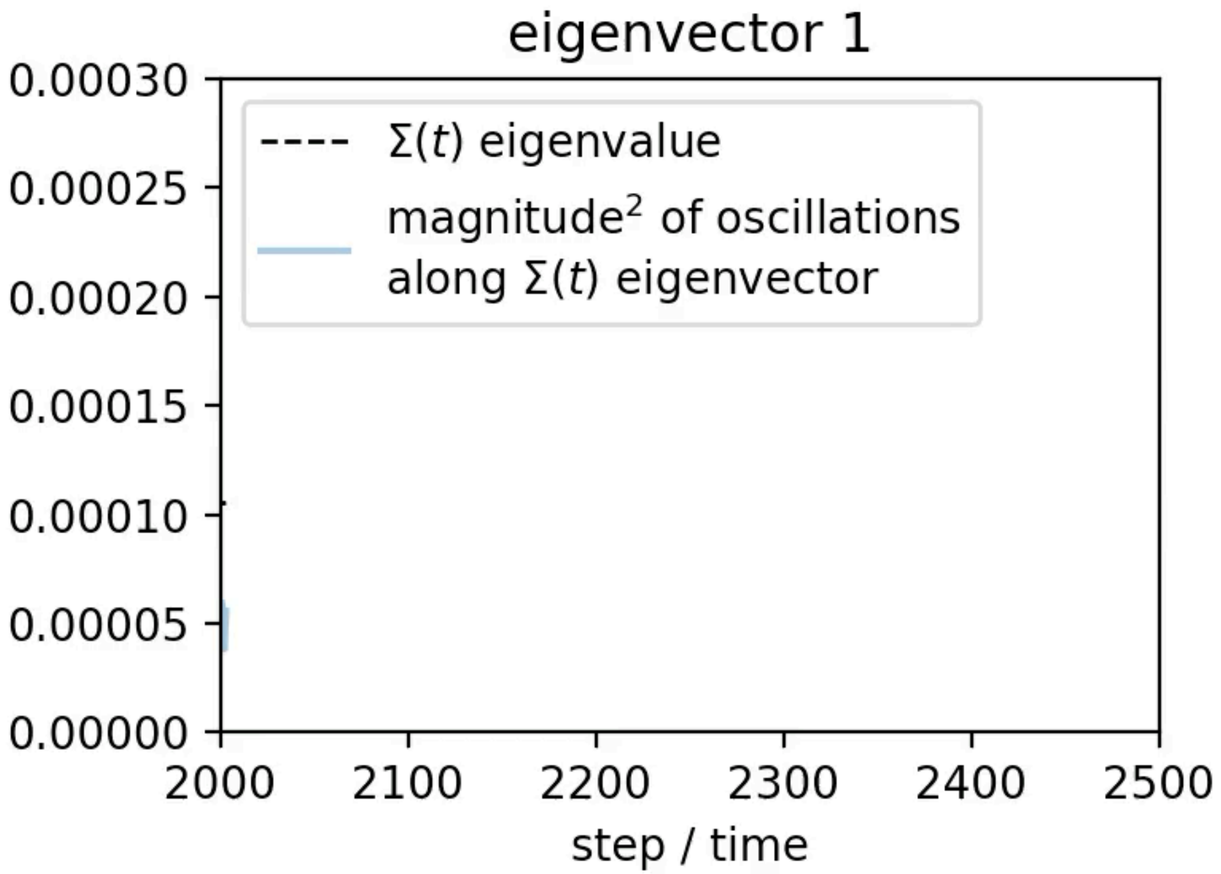
Strong agreement across architectures!

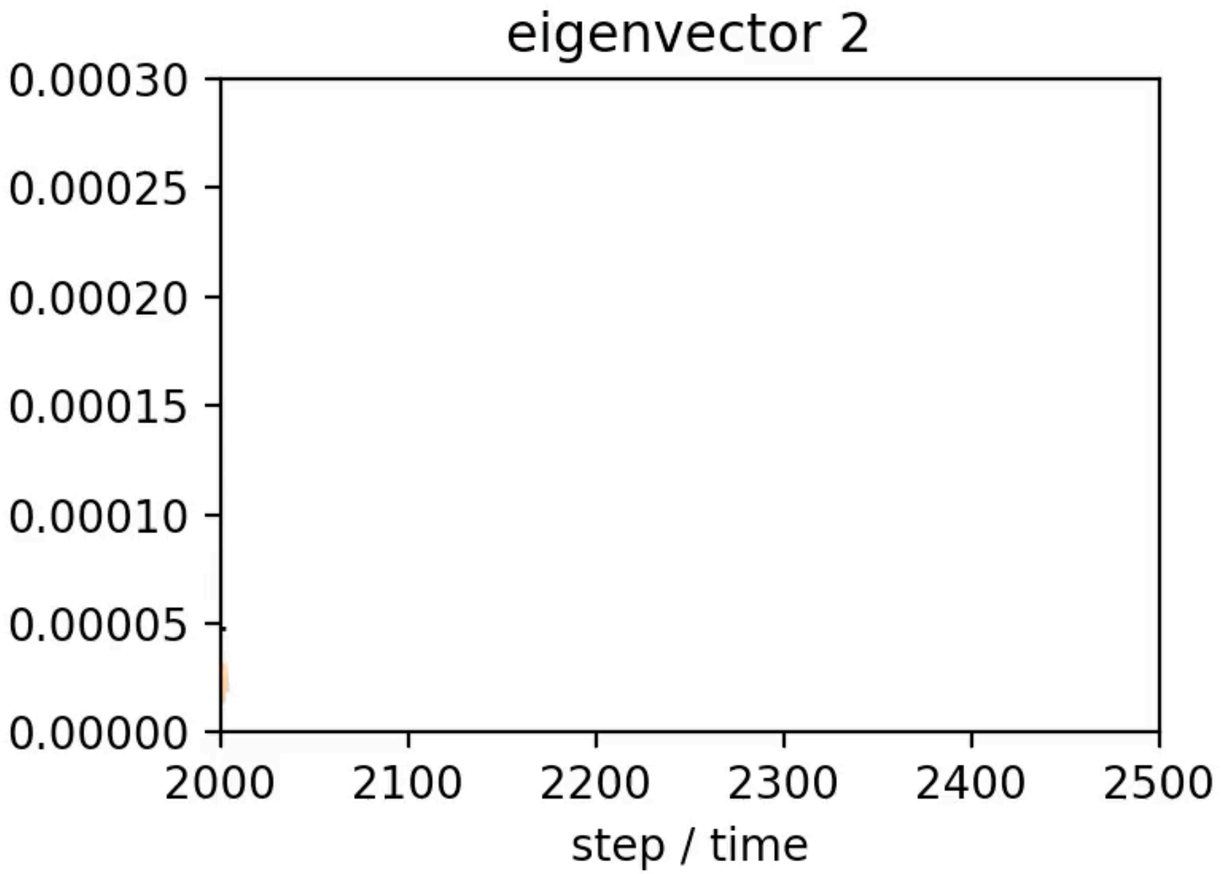
The RMSProp Central Flow Predicts Σ



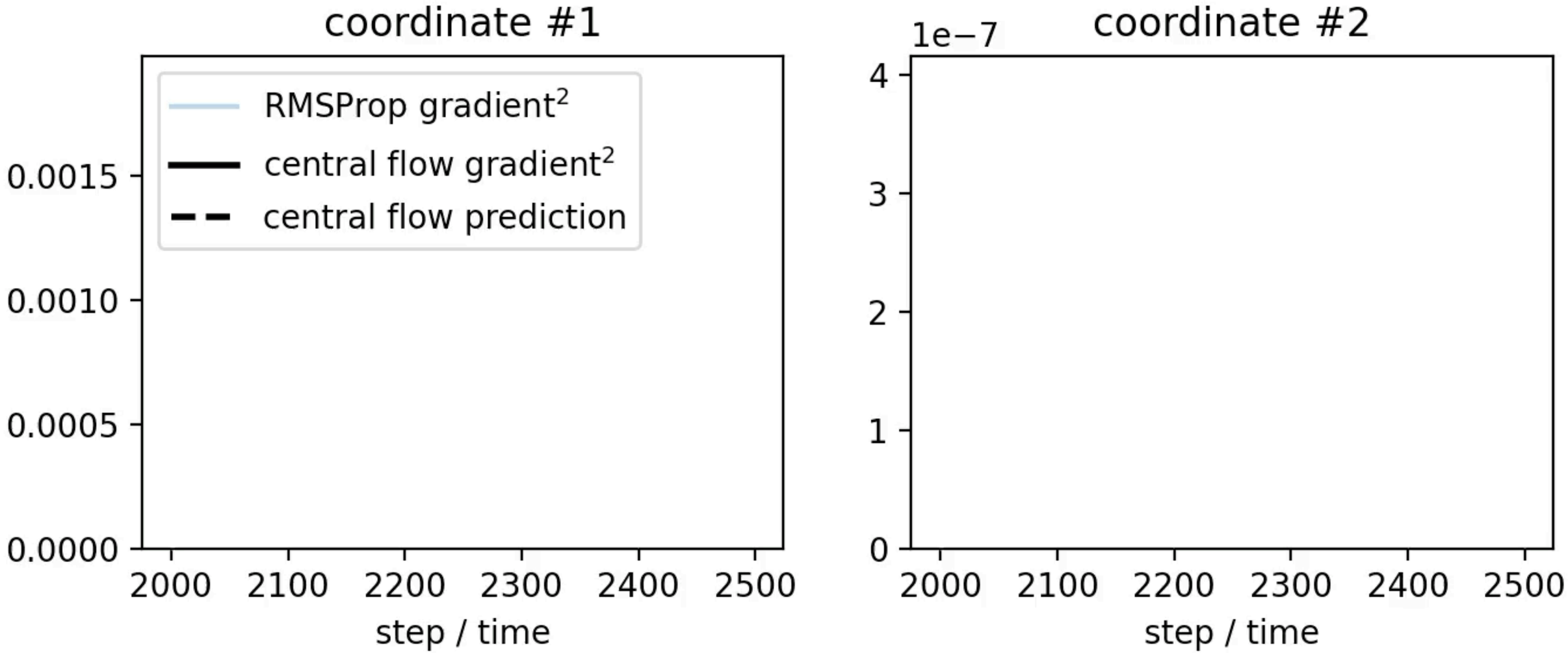


The RMSProp Central Flow Predicts Σ

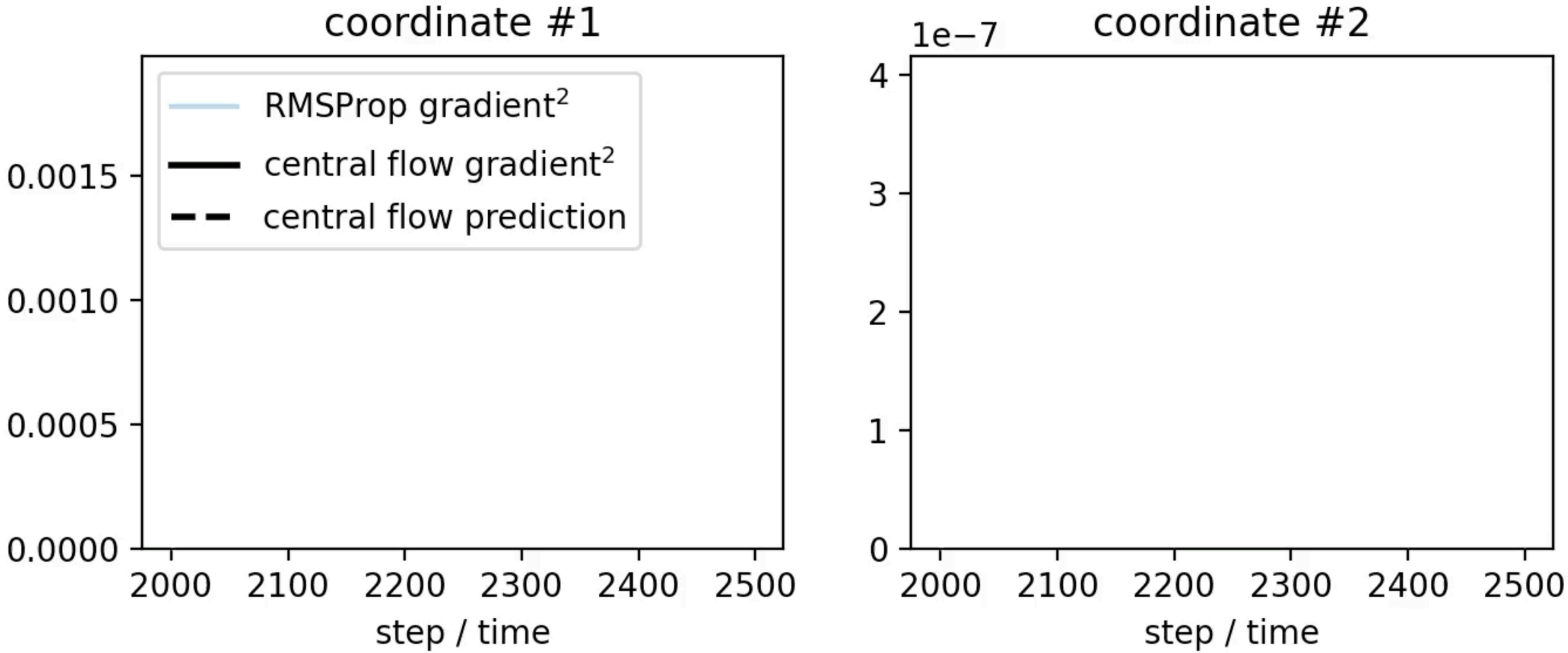




The RMSProp Central Flow Predicts Gradient Norms



The RMSProp Central Flow Predicts Gradient Norms



Interpreting the Flow: The Stationary Preconditioner

$$\frac{d\overline{w}}{dt} = -\frac{\eta}{\sqrt{\nu}}\,\mathbb{E}\Big[\ \nabla L(\overline{w}) \ + \ \frac{1}{2}\nabla\big\langle\, \nabla^2 L(\overline{w}),\, \Sigma\big\rangle\ \Big]$$

$$\frac{d\nu}{dt} = \frac{1-\beta_2}{\beta_2}\Big[\, \nabla L(\overline{w})^{\odot 2} + \mathrm{diag}[H(\overline{w})\Sigma H(\overline{w})] - \nu \Big]$$

Interpreting the Flow: The Stationary Preconditioner

$$\frac{d\overline{w}}{dt} = -\frac{\eta}{\sqrt{\nu}}\,\mathbb{E}\Big[\ \nabla L(\overline{w}) \ + \ \frac{1}{2}\nabla\big\langle\, \nabla^2 L(\overline{w}),\, \Sigma\big\rangle\ \Big]$$

$$\frac{d\nu}{dt} = \frac{1-\beta_2}{\beta_2}\Big[\, \nabla L(\overline{w})^{\odot 2} + \mathrm{diag}[H(\overline{w})\Sigma H(\overline{w})] - \nu \Big]$$

- Freeze $\overline{w}$ and evolve (ν, Σ) until they reach stationary values $\nu^\star(\overline{w})$, $\Sigma^\star(\overline{w})$

Interpreting the Flow: The Stationary Preconditioner

$$\frac{d\overline{w}}{dt} = -\frac{\eta}{\sqrt{\nu}}\,\mathbb{E}\Big[\ \nabla L(\overline{w}) \ + \ \frac{1}{2}\nabla\langle\,\nabla^2 L(\overline{w}),\,\Sigma\rangle\ \Big]$$

$$\frac{d\nu}{dt} = \frac{1-\beta_2}{\beta_2}\Big[\,\nabla L(\overline{w})^{\odot 2} + \mathrm{diag}[H(\overline{w})\Sigma H(\overline{w})] - \nu\Big]$$

- Freeze $\overline{w}$ and evolve (ν, Σ) until they reach stationary values $\nu^\star(\overline{w})$, $\Sigma^\star(\overline{w})$
- Gives a stationary preconditioner $P^\star(\overline{w}) := \mathrm{diag}[\nu^\star(\overline{w})^{1/2}/\eta]$

Interpreting the Flow: The Stationary Preconditioner

$$\frac{d\overline{w}}{dt} = -P^\star(\overline{w})^{-1}\,\mathbb{E}\Big[\ \nabla L(\overline{w}) + \frac{1}{2}\nabla\big\langle \nabla^2 L(\overline{w}), \Sigma^\star(\overline{w})\big\rangle\ \Big]$$

- Freeze $\overline{w}$ and evolve (ν, Σ) until they reach stationary values $\nu^\star(\overline{w})$, $\Sigma^\star(\overline{w})$
- Gives a stationary preconditioner $P^\star(\overline{w}) := \mathrm{diag}[\nu^\star(\overline{w})^{1/2}/\eta]$

Interpreting the Flow: The Stationary Preconditioner

$$\frac{d\overline{w}}{dt} = -P^\star(\overline{w})^{-1}\,\mathbb{E}\Big[\ \nabla L(\overline{w}) + \frac{1}{2}\nabla\big\langle \nabla^2 L(\overline{w}), \Sigma^\star(\overline{w})\big\rangle\ \Big]$$

- Freeze $\overline{w}$ and evolve (ν, Σ) until they reach stationary values $\nu^\star(\overline{w})$, $\Sigma^\star(\overline{w})$
- Gives a stationary preconditioner $P^\star(\overline{w}) := \mathrm{diag}[\nu^\star(\overline{w})^{1/2}/\eta]$

Lemma: The stationary preconditioner $P^\star(\overline{w})$ solves the following SDP:

$$P^\star(w) = \underset{P\ \text{diagonal}}{\operatorname{argmin}}\ \ \mathrm{tr}(P) + \frac{1}{\eta^2}\|\nabla L(w)\|_{P^{-1}} \quad \text{such that} \quad H \preceq 2P$$

Interpreting the Flow: The Stationary Preconditioner

$$\frac{d\overline{w}}{dt} = -P^\star(\overline{w})^{-1}\,\mathbb{E}\Big[\ \nabla L(\overline{w}) + \frac{1}{2}\nabla\big\langle \nabla^2 L(\overline{w}), \Sigma^\star(\overline{w})\big\rangle\ \Big]$$

- Freeze $\overline{w}$ and evolve (ν, Σ) until they reach stationary values $\nu^\star(\overline{w})$, $\Sigma^\star(\overline{w})$
- Gives a stationary preconditioner $P^\star(\overline{w}) := \mathrm{diag}[\nu^\star(\overline{w})^{1/2}/\eta]$

Lemma: The stationary preconditioner $P^\star(\overline{w})$ solves the following SDP:

$$P^\star(w) = \underset{P\ \text{diagonal}}{\operatorname{argmin}}\ \ \mathrm{tr}(P) + \frac{1}{\eta^2}\|\nabla L(w)\|_{P^{-1}} \quad \text{such that} \quad H \preceq 2P$$

Interpreting the Flow: The Stationary Preconditioner

$$\frac{d\overline{w}}{dt} = -P^\star(\overline{w})^{-1}\,\mathbb{E}\Big[\ \nabla L(\overline{w}) + \frac{1}{2}\nabla\big\langle \nabla^2 L(\overline{w}), \Sigma^\star(\overline{w})\big\rangle\ \Big]$$

- Freeze $\overline{w}$ and evolve (ν, Σ) until they reach stationary values $\nu^\star(\overline{w})$, $\Sigma^\star(\overline{w})$
- Gives a stationary preconditioner $P^\star(\overline{w}) := \mathrm{diag}[\nu^\star(\overline{w})^{1/2}/\eta]$

Lemma: The stationary preconditioner $P^\star(\overline{w})$ solves the following SDP:

$$P^\star(w) = \underset{P\ \text{diagonal}}{\operatorname{argmin}}\ \operatorname{tr}(P) + \frac{1}{\eta^2}\|\nabla L(w)\|_{P^{-1}} \quad \text{such that} \quad H \preceq 2P$$

local stability

Interpreting the Flow: The Stationary Preconditioner

$$\frac{d\overline{w}}{dt} = -P^\star(\overline{w})^{-1}\,\mathbb{E}\Big[\ \nabla L(\overline{w}) + \frac{1}{2}\nabla\big\langle \nabla^2 L(\overline{w}), \Sigma^\star(\overline{w})\big\rangle\ \Big]$$

- Freeze $\overline{w}$ and evolve (ν, Σ) until they reach stationary values $\nu^\star(\overline{w})$, $\Sigma^\star(\overline{w})$
- Gives a stationary preconditioner $P^\star(\overline{w}) := \mathrm{diag}[\nu^\star(\overline{w})^{1/2}/\eta]$

Lemma: The stationary preconditioner $P^\star(\overline{w})$ solves the following SDP:

$$P^\star(w) = \underset{P\ \text{diagonal}}{\operatorname{argmin}}\ \operatorname{tr}(P) + \frac{1}{\eta^2}\|\nabla L(w)\|_{P^{-1}} \quad \text{such that} \quad H \preceq 2P$$

max ***harmonic mean*** of the learning rates

local stability

Interpreting the Flow: The Stationary Preconditioner

$$\frac{d\overline{w}}{dt} = -P^\star(\overline{w})^{-1}\,\mathbb{E}\Big[\ \nabla L(\overline{w}) + \frac{1}{2}\nabla\big\langle \nabla^2 L(\overline{w}), \Sigma^\star(\overline{w})\big\rangle\ \Big]$$

- Freeze $\overline{w}$ and evolve (ν, Σ) until they reach stationary values $\nu^\star(\overline{w})$, $\Sigma^\star(\overline{w})$
- Gives a stationary preconditioner $P^\star(\overline{w}) := \mathrm{diag}[\nu^\star(\overline{w})^{1/2}/\eta]$

Lemma: The stationary preconditioner $P^\star(\overline{w})$ solves the following SDP:

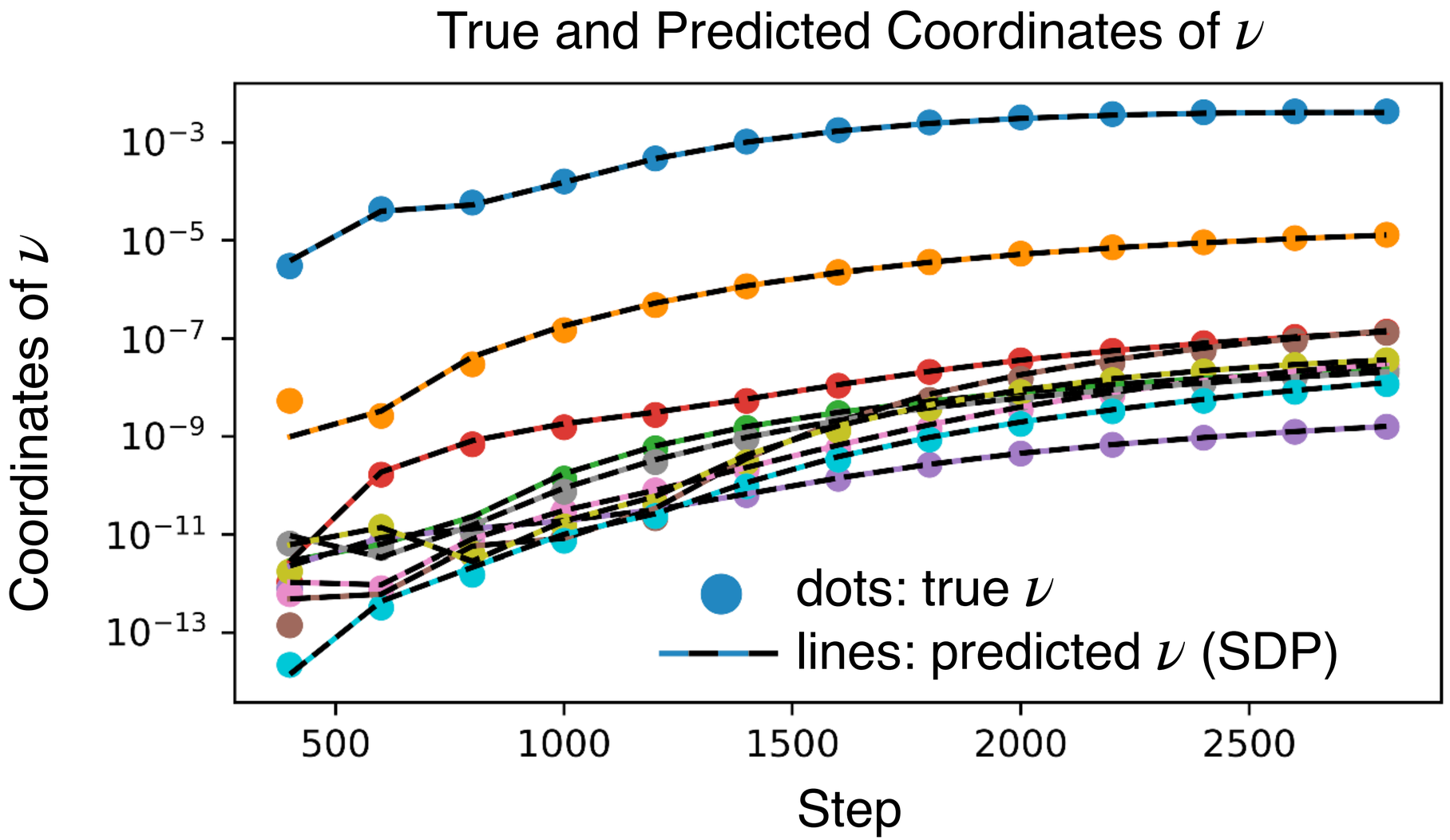
$$P^\star(w) = \underset{P \text{ diagonal}}{\operatorname{argmin}}\ \operatorname{tr}(P) + \frac{1}{\eta^2}\|\nabla L(w)\|_{P^{-1}} \quad \text{such that} \quad H \preceq 2P$$

max ***harmonic mean*** of the learning rates

optimization speed

local stability

Interpreting the Flow: The Stationary Preconditioner



The Stationary RMSProp Central Flow*

$$\frac{d\overline{w}}{dt} = -P^{\star}(\overline{w})^{-1}\left[\nabla L(\overline{w}) + \frac{\eta^2}{4}\nabla \operatorname{tr} P^{\star}(\overline{w})\right]$$

stationary preconditioner

* this is a slightly simplified form which gives clearer intuition

The Stationary RMSProp Central Flow*

$$\frac{d\overline{w}}{dt} = -\,P^{\star}(\overline{w})^{-1}\left[\nabla L(\overline{w}) + \frac{\eta^2}{4}\nabla\,\mathrm{tr}\,P^{\star}(\overline{w})\right]$$

stationary preconditioner

curvature penalty

* this is a slightly simplified form which gives clearer intuition

The Stationary RMSProp Central Flow*

$$\frac{d\overline{w}}{dt} = -\,P^{\star}(\overline{w})^{-1}\left[\nabla L(\overline{w}) + \frac{\eta^2}{4}\nabla\,\mathrm{tr}\,P^{\star}(\overline{w})\right]$$

stationary preconditioner

curvature penalty

Acceleration via Regularization:

Use the largest step sizes you can **(adapt)**, but avoid regions of the loss landscape where you are forced to take smaller steps **(regularize)**.

* this is a slightly simplified form which gives clearer intuition

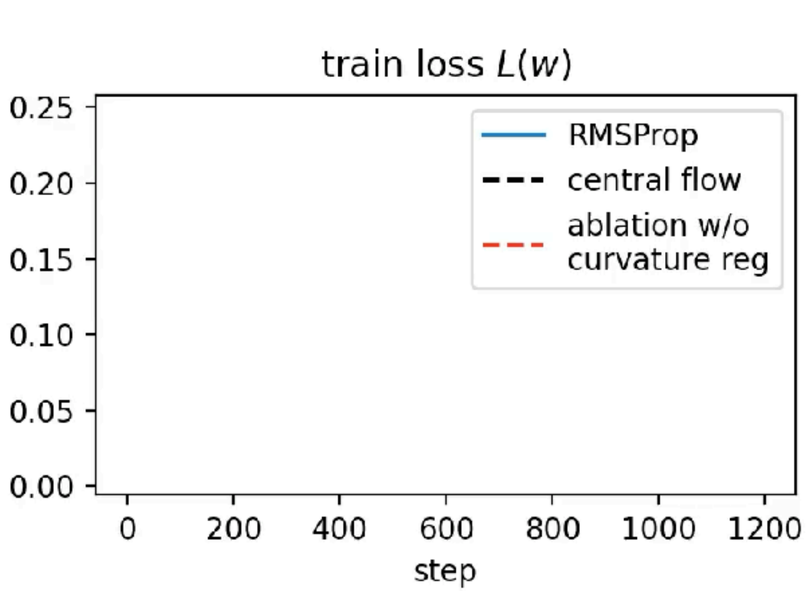
Acceleration via Regularization:

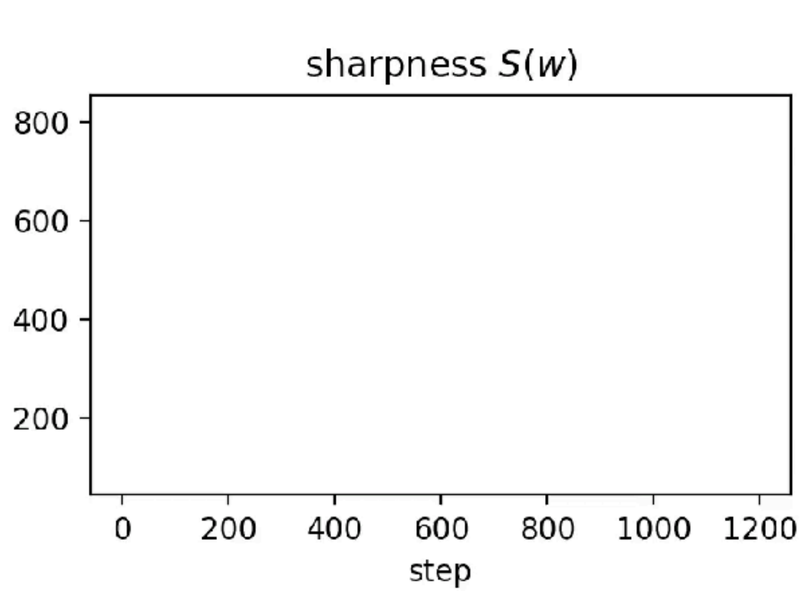
Use the largest step sizes you can **(adapt)**, but avoid regions of the loss landscape where you are forced to take smaller steps **(regularize)**.

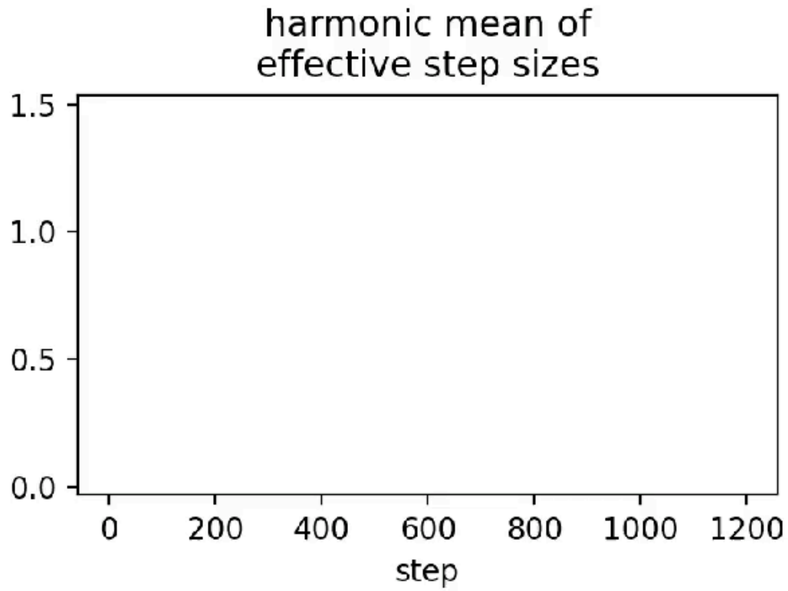
Acceleration via Regularization:

Use the largest step sizes you can **(adapt)**, but avoid regions of the loss landscape where you are forced to take smaller steps **(regularize)**.

$$\frac{d\overline{w}}{dt} = -P^{\star}(\overline{w})^{-1}\left[\nabla L(\overline{w}) + \frac{\eta^2}{4}\nabla \operatorname{tr} P^{\star}(\overline{w})\right]$$



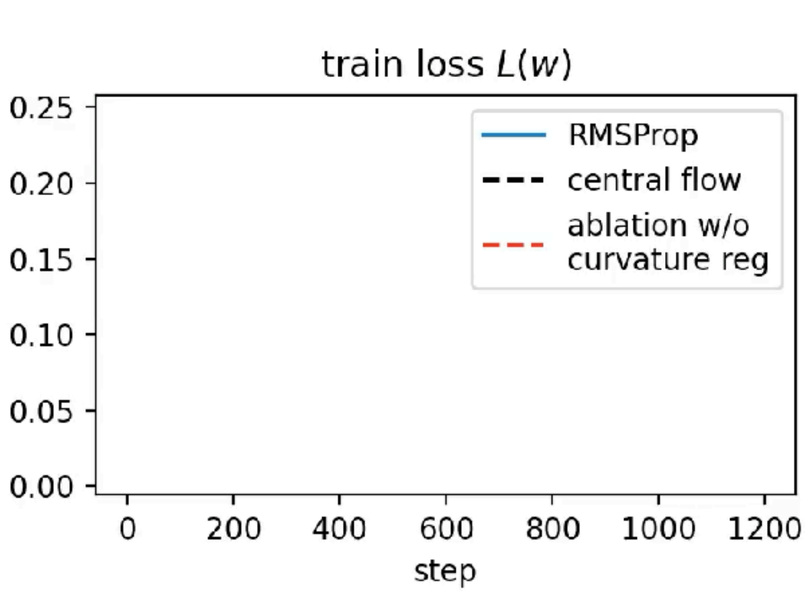


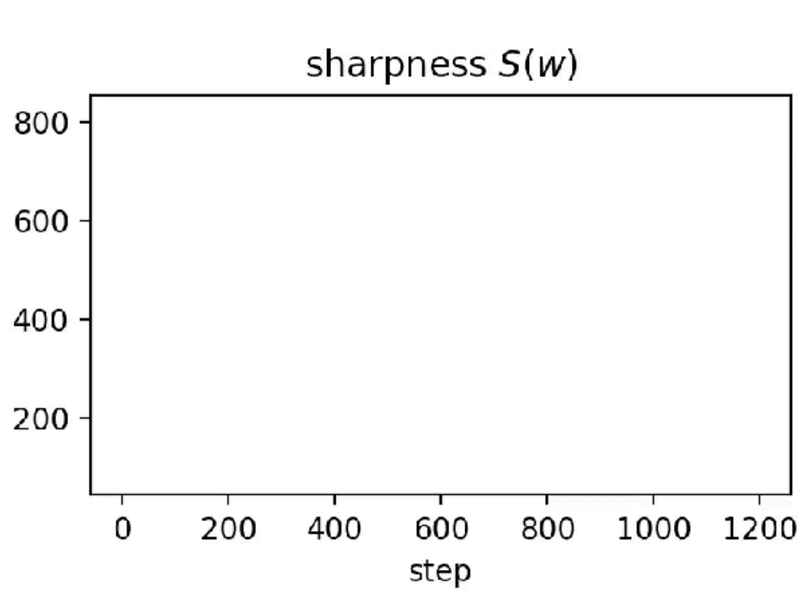


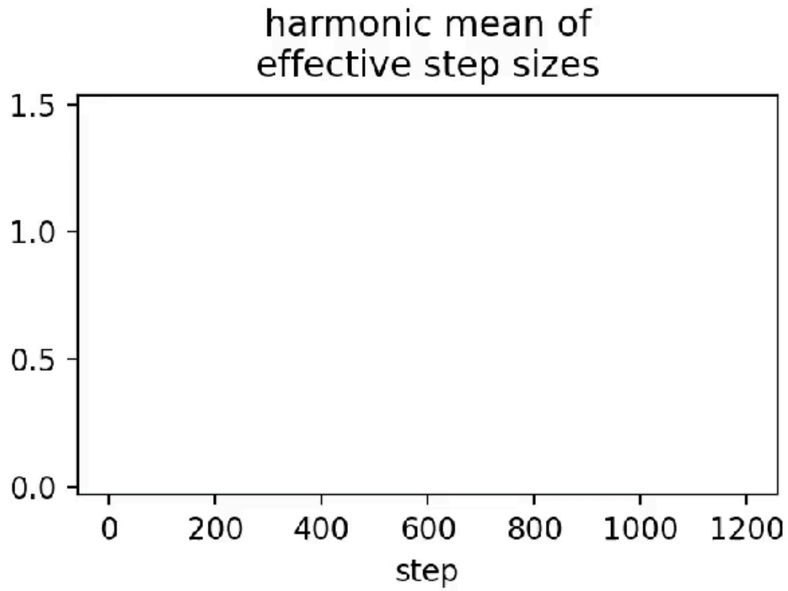
Acceleration via Regularization:

Use the largest step sizes you can **(adapt)**, but avoid regions of the loss landscape where you are forced to take smaller steps **(regularize)**.

$$\frac{d\overline{w}}{dt} = -P^{\star}(\overline{w})^{-1}\left[\nabla L(\overline{w}) + \frac{\eta^2}{4}\nabla \operatorname{tr} P^{\star}(\overline{w})\right]$$





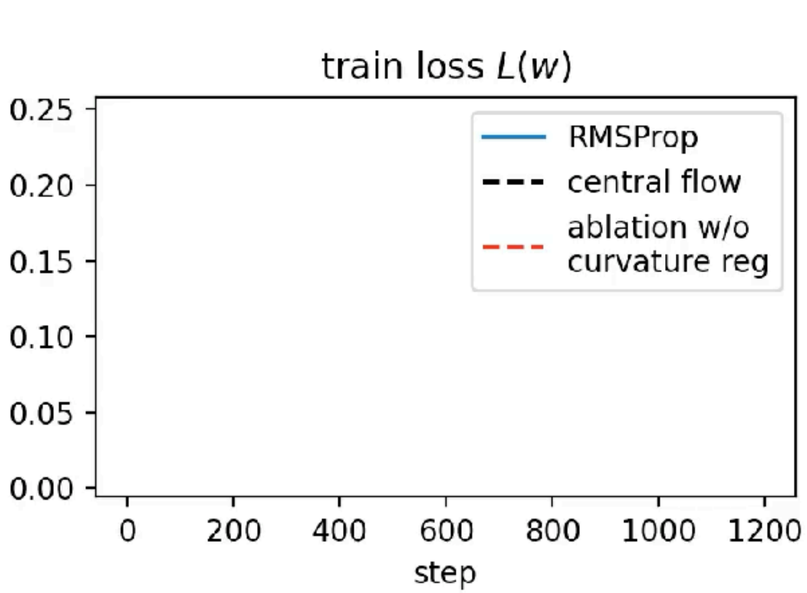


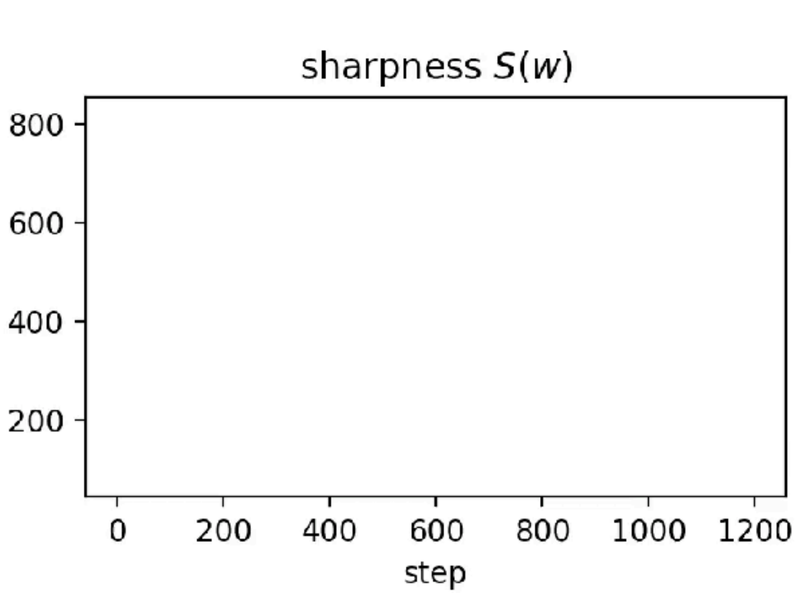
Acceleration via Regularization:

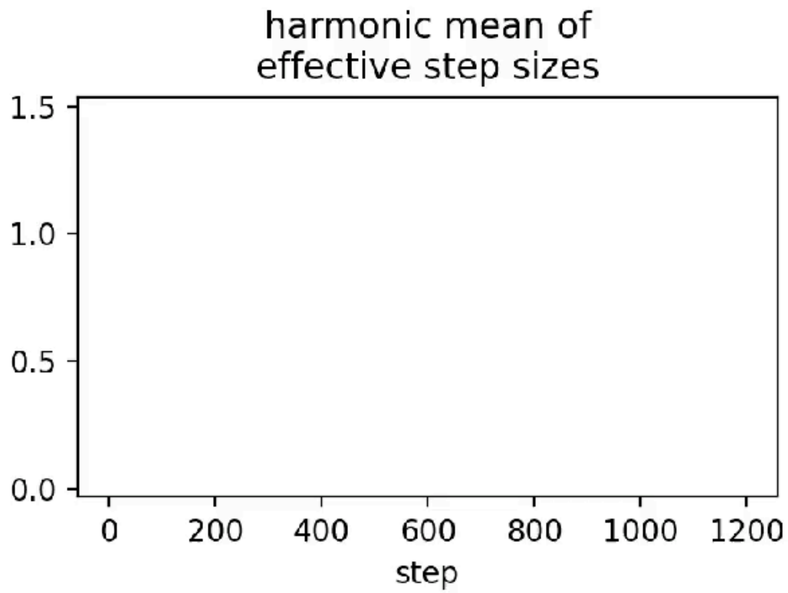
Use the largest step sizes you can **(adapt)**, but avoid regions of the loss landscape where you are forced to take smaller steps **(regularize)**.

$$\frac{d\overline{w}}{dt} = -P^{\star}(\overline{w})^{-1}\left[\nabla L(\overline{w}) + \frac{\eta^2}{4}\nabla \operatorname{tr} P^{\star}(\overline{w})\right]$$

controlling this reg. is the main role of η!







Takeaways

Takeaways

- We derive ***central flows,*** which model time-averaged trajectories of oscillatory optimizers

Takeaways

- We derive ***central flows,*** which model time-averaged trajectories of oscillatory optimizers

- We empirically verify these flows predict long-term optimization trajectories

Takeaways

- We derive ***central flows,*** which model time-averaged trajectories of oscillatory optimizers
- We empirically verify these flows predict long-term optimization trajectories
- By interpreting these flows, we can easily read off an optimizer's behavior

Takeaways

- We derive ***central flows,*** which model time-averaged trajectories of oscillatory optimizers
- We empirically verify these flows predict long-term optimization trajectories
- By interpreting these flows, we can easily read off an optimizer's behavior

---

- **First order** optimizers implicitly implement sophisticated **second-order** strategies

Takeaways

- We derive ***central flows,*** which model time-averaged trajectories of oscillatory optimizers
- We empirically verify these flows predict long-term optimization trajectories
- By interpreting these flows, we can easily read off an optimizer's behavior

---

- **First order** optimizers implicitly implement sophisticated **second-order** strategies
- **Acceleration via regularization:** good optimizers not only adapt to the loss landscape, but also avoid high curvature regions where they are forced to take smaller steps

Takeaways

- We derive ***central flows,*** which model time-averaged trajectories of oscillatory optimizers
- We empirically verify these flows predict long-term optimization trajectories
- By interpreting these flows, we can easily read off an optimizer's behavior

---

- **First order** optimizers implicitly implement sophisticated **second-order** strategies
- **Acceleration via regularization:** good optimizers not only adapt to the loss landscape, but also avoid high curvature regions where they are forced to take smaller steps
- RMSProp implicitly solves an SDP related to max-cut to determine its preconditioner

Promising Future Work

Promising Future Work

Rigor:

- Under what assumptions is the central flow approximation provably correct?
- What do “reasonable” assumptions for deep learning optimization look like?

Promising Future Work

Rigor:

- Under what assumptions is the central flow approximation provably correct?
- What do “reasonable” assumptions for deep learning optimization look like?

Noise:

- Can similar arguments be used to derive a central flow for SGD?
- When does an SGD central flow exist?

Promising Future Work

Rigor:

- Under what assumptions is the central flow approximation provably correct?
- What do “reasonable” assumptions for deep learning optimization look like?

Noise:

- Can similar arguments be used to derive a central flow for SGD?
- When does an SGD central flow exist?

Applications:

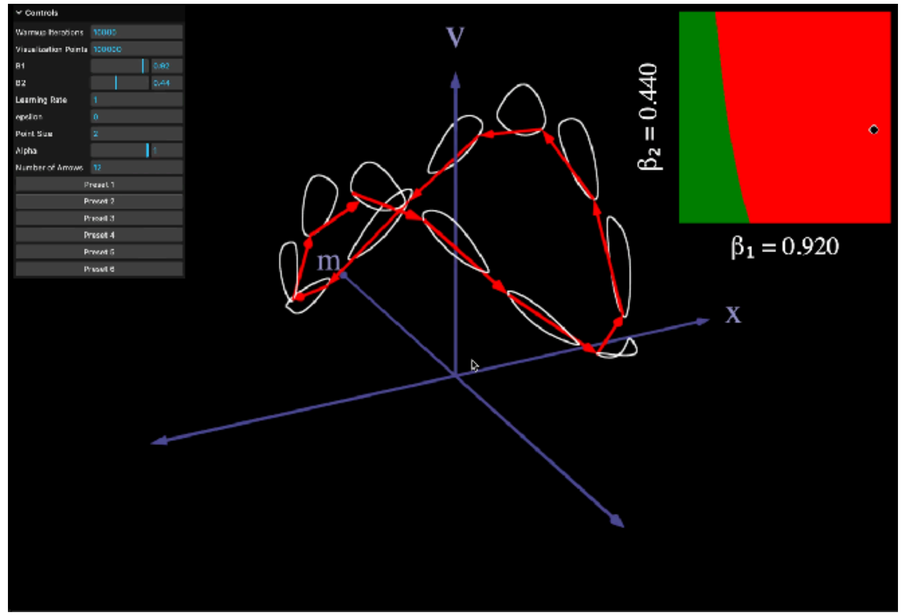
- Can these ideas be used to design new optimizers?
- How can we design architectures that are easy to optimize?

Thanks!

Jeremy Cohen
Flatiron Institute

https://centralflows.github.io

Animated Blog Post!



https://alex-damian.github.io/adam/

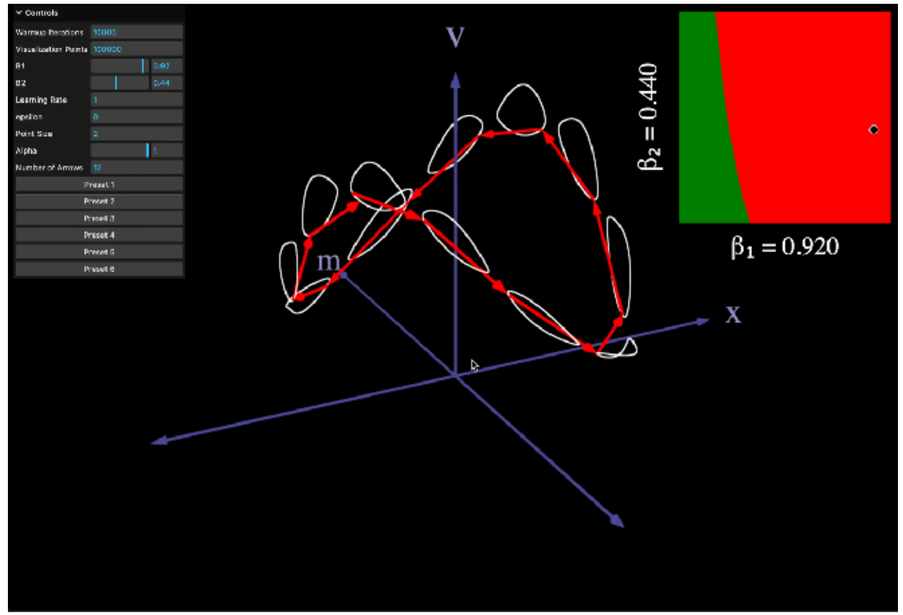
Bonus: crazy limit cycles in Adam

Thanks!

Jeremy Cohen
Flatiron Institute

https://centralflows.github.io

Animated Blog Post!



https://alex-damian.github.io/adam/

Bonus: crazy limit cycles in Adam