Stability of Gradient-Based Bilevel Algorithms for Hyperparameter Optimization

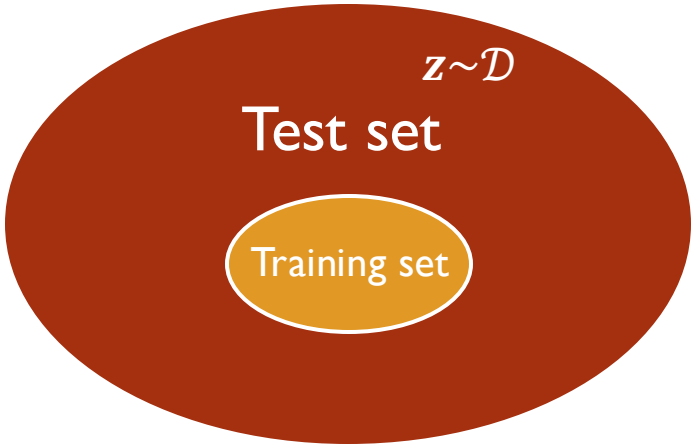
Rongzhen Wang

2024.3.28

Generalization and stability

- We care about:
  - Expected risk: $R = \mathbb{E}_{z\sim D}[\ell(A(S); z)]$
- This can be estimated by:
  - Empirical risk: $R_S = \frac{1}{n}\sum_{i=1}^{n} \ell(A(S); z_i)$

- $R - R_S$ ➔ generalization error



Stability and generalization

- Stability:
  - The change in the performance of the algorithm's output if a single data point is replaced
  - $S = \{\boldsymbol{z}_1, \cdots, \boldsymbol{z}_i, \cdots, \boldsymbol{z}_n\}, \tilde{S}^i = \{\boldsymbol{z}_1, \cdots, \tilde{\boldsymbol{z}}_i, \cdots, \boldsymbol{z}_n\}$ ➔$\ell\left(A\left(\tilde{S}^i\right)\right)$? $\ell(A(S))$

- In expectation, generalization equals stability.
- $\mathbb{E}_S[R - R_S] = \mathbb{E}_S\left[\mathbb{E}_{z\sim D}[\ell(A(S); \boldsymbol{z})] - \frac{1}{n}\sum_{i=1}^{n} \ell(A(S); \boldsymbol{z}_i)\right]$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{S,\tilde{S}}\left[\ell\left(A\left(\tilde{S}^i\right); \boldsymbol{z}_i\right) - \ell(A(S); \boldsymbol{z}_i)\right]$$

Recap: Lipchitz conditions

- L-Lipschitz continuous:

$$\forall \boldsymbol{u}, \boldsymbol{v} \in \Omega\,, \|f(\boldsymbol{u}) - f(\boldsymbol{v})\| \leq L\|\boldsymbol{u} - \boldsymbol{v}\| \text{ or } \forall \boldsymbol{x} \in \Omega, \|\nabla f(\boldsymbol{x})\| \leq L.$$

- γ-smooth (i.e., gradient Lipschitz continuous):

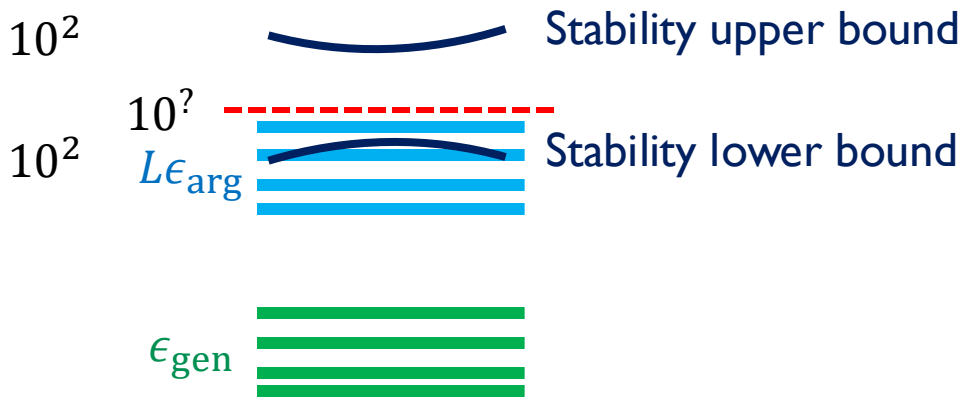
$$\forall \boldsymbol{u}, \boldsymbol{v} \in \Omega\,, \|\nabla f(\boldsymbol{u}) - \nabla f(\boldsymbol{v})\| \leq \gamma.$$

- $\mathbb{E}_S[R - R_S] \leq \sup_{S,\tilde{S},\boldsymbol{z}} \left|\ell\left(A\left(\tilde{S}\right); z\right) - \ell(A(S); z)\right|$ ➔ uniform stability ϵ_{arg}

$$\leq \sup_{S,\tilde{S}} L\left\|A(S) - A\left(\tilde{S}\right)\right\|$$ ➔ uniform argument stability ϵ_{stab}

Stability based generalization bound for Lipschitz loss functions

1. Generalization upper bound via stability: $\epsilon_{\text{gen}}(A) \leq \epsilon_{\text{stab}}(A) \leq L\epsilon_{\text{arg}}(A)$
2. Stability upper bound
   - About key factors: sample size, learning rate, number of steps
3. Stability lower bound
   - Construct an example



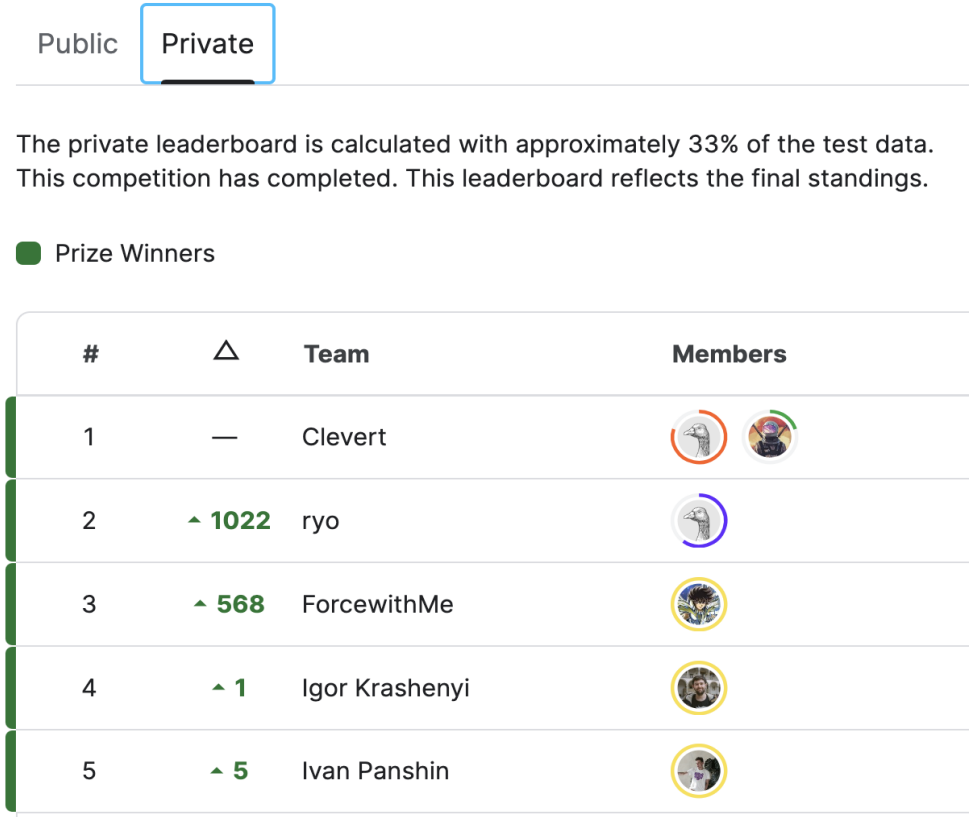
Outline

- Background on hyperparameter optimization (HO)
- Generalization and stability of HO algorithms
- Stability bounds of gradient-based HO algorithms
- Conclusion and discussion

Hyperparameter optimization (HO)

- Hyperparameters:
  - `num_layers, lr …`
  - Influential to training efficiency and results
  - Fixed during the training phase

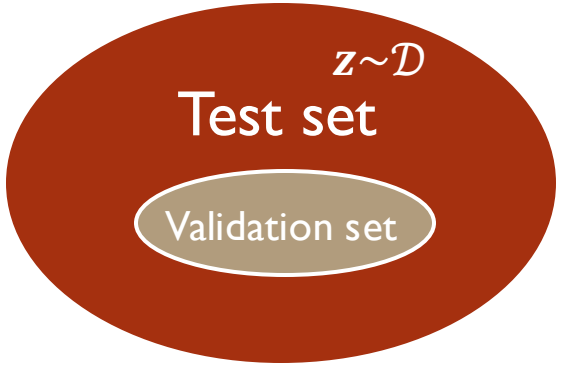
- Three phases in machine learning
  - training
  - validation
  - test



Generalization and stability of HO

- Recap: Single-level optimization
  - Generalization via uniform argument stability

$$\epsilon_{\text{gen}}(A) \triangleq \mathbb{E}_S[R - R_S] \leq \sup_{S,\tilde{S}} L\mathbb{E}_A[\|A(S) - A(\tilde{S})\|] \triangleq L\epsilon_{\text{arg}}(A)$$

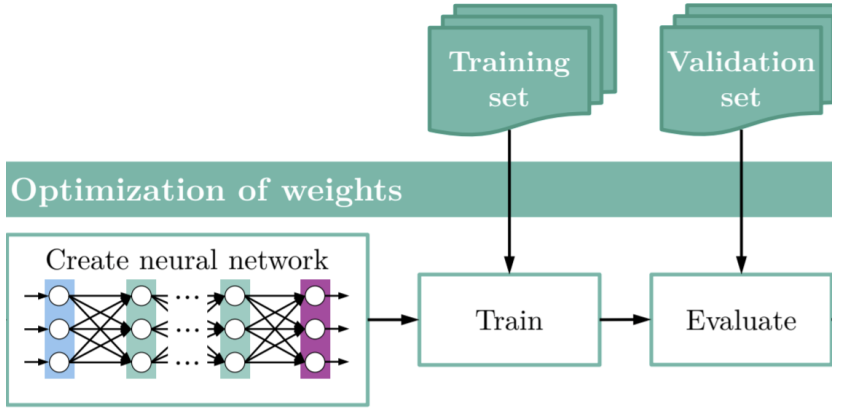


- Hyperparameter optimization:
  - Generalization on validation via uniform argument stability

$$\epsilon_{\text{gen}}(A) \triangleq \mathbb{E}_{S^{\text{val}},S^{\text{tr}}}[R - R_{S^{\text{val}}}] \leq \sup_{S^{\text{val}},\tilde{S}^{\text{val}},S^{\text{tr}}} L\mathbb{E}_A[\|A(S^{\text{val}},S^{\text{tr}}) - A(\tilde{S}^{\text{val}},S^{\text{tr}})\|] \triangleq L\epsilon_{\text{arg}}(A)$$

Formulation of HO

- Model parameters $\boldsymbol{\theta}$, hyperparameters $\boldsymbol{\lambda}$
- Training loss ℓ^{tr}, validation loss ℓ^{val}
- Training set $S^{\mathrm{tr}} = \{\boldsymbol{z}_i^{\mathrm{tr}}\}_{i=1}^{n}$, validation set $S^{\mathrm{val}} = \{\boldsymbol{z}_i^{\mathrm{val}}\}_{i=1}^{m}$
- Example $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$



Outer level: $\boldsymbol{\lambda}^* = \underset{\boldsymbol{\lambda}\in\Lambda}{\operatorname{argmin}} R_{S^{\mathrm{val}}}(\boldsymbol{\lambda}) \triangleq \frac{1}{m}\sum_{i=1}^{m} \ell^{\mathrm{val}}\left(\boldsymbol{\lambda}, \boldsymbol{\theta}^*(\boldsymbol{\lambda}); \boldsymbol{z}_i^{\mathrm{val}}\right) \mathcal{L}^{\mathrm{val}}\left(\boldsymbol{\lambda}; \boldsymbol{z}_i^{\mathrm{val}}\right)$

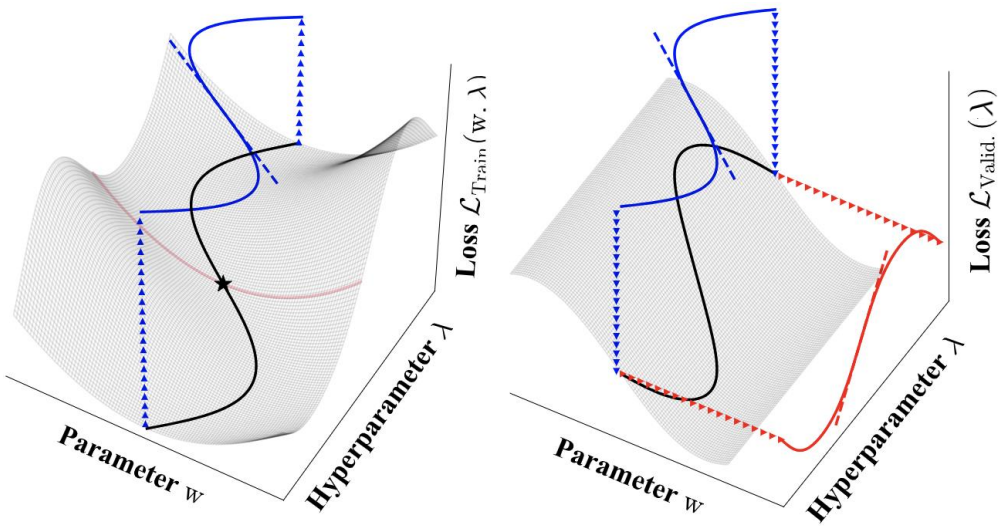
Inner level: where $\boldsymbol{\theta}^*(\boldsymbol{\lambda}) = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} R_{S^{\mathrm{tr}}}(\boldsymbol{\lambda}, \boldsymbol{\theta}) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell^{\mathrm{tr}}\left(\boldsymbol{\lambda}, \boldsymbol{\theta}; \boldsymbol{z}_i^{\mathrm{tr}}\right)$

$$\boldsymbol{\lambda}^* = \underset{\boldsymbol{\lambda}\in\Lambda}{\operatorname{argmin}} \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}^{\mathrm{val}}\left(\boldsymbol{\lambda}; \boldsymbol{z}_i^{\mathrm{val}}\right)$$

Approaches for solving HO

Gradient-based methods

- How to get the hypergradient $\nabla_{\boldsymbol{\lambda}} \mathcal{L}^{\text{val}}$
  - Modeling $\mathcal{L}^{\text{val}}$ to be differentiable w.r.t. $\boldsymbol{\lambda}$
    - e.g., architecture search, data reweighting
  - Calculate $\nabla_{\boldsymbol{\lambda}} \mathcal{L}^{\text{val}}$
    - difficult for $\nabla_{\boldsymbol{\lambda}} \mathcal{L}^{\text{val}}(\boldsymbol{\lambda}; \boldsymbol{z}) \neq \nabla_{\boldsymbol{\lambda}} \ell^{\text{val}}(\boldsymbol{\lambda}, \widehat{\boldsymbol{\theta}}; \boldsymbol{z})$



Algorithm 1 Gradient-based bilevel HO

Input: Initialization $\boldsymbol{\lambda}_0$ and $\boldsymbol{\theta}_0$; training set S^{tr} and validation set S^{val}; learning rate scheme α and η.

for $t = 1$ **to** T **do**

  for $k = 1$ **to** K **do**

    uniformly sampling j_k from $[n]$

    $\boldsymbol{\theta}_k \leftarrow \boldsymbol{\theta}_{k-1} - \eta_k \nabla_{\boldsymbol{\theta}} \ell^{\text{tr}}(\boldsymbol{\lambda}_{t-1}, \boldsymbol{\theta}_{k-1}; \boldsymbol{z}^{\text{tr}}_{j_k})$

  end for

  Inner optimization

  uniformly sampling i_t from $[m]$

  $\boldsymbol{\lambda}_t \leftarrow \boldsymbol{\lambda}_{t-1} - \alpha_t \nabla_{\boldsymbol{\lambda}} \mathcal{L}^{\text{val}}(\boldsymbol{\lambda}_{t-1}; \boldsymbol{z}^{\text{val}}_{i_t})$

end for

Outer optimization

return $\boldsymbol{\lambda}_T$ and $\boldsymbol{\theta}_K$

Figure credit: Lorraine J, Vicol P, Duvenaud D. Optimizing millions of hyperparameters by implicit differentiation[C]//International conference on artificial intelligence and statistics. PMLR, 2020: 1540-1552.

Approaches for solving HO

Gradient-based methods

- $\nabla_x f\big(\boldsymbol{u}(\boldsymbol{x}), \boldsymbol{v}(\boldsymbol{x})\big) = \nabla_{\boldsymbol{u}} f(\boldsymbol{u}, \boldsymbol{v}) \nabla_x \boldsymbol{u}(\boldsymbol{x}) + \nabla_{\boldsymbol{v}} f(\boldsymbol{u}, \boldsymbol{v}) \nabla_x \boldsymbol{v}(\boldsymbol{x})$
- $\nabla_{\boldsymbol{\lambda}} \mathcal{L}^{\mathrm{val}}(\boldsymbol{\lambda}; \boldsymbol{z}) = \nabla_{\boldsymbol{\lambda}} \ell^{\mathrm{val}}\big(\boldsymbol{\lambda}, \widehat{\boldsymbol{\theta}}(\boldsymbol{\lambda}); \boldsymbol{z}\big) = \nabla_{\boldsymbol{\lambda}} \ell^{\mathrm{val}}\big(\boldsymbol{\lambda}, \widehat{\boldsymbol{\theta}}; \boldsymbol{z}\big) + \nabla_{\boldsymbol{\theta}} \ell^{\mathrm{val}}\big(\boldsymbol{\lambda}, \widehat{\boldsymbol{\theta}}; \boldsymbol{z}\big) \nabla_{\boldsymbol{\lambda}} \widehat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$

  Explicit, easy to be auto-calculated

  Implicit, involving the inner optimization

- An example of optimizing the regularization coefficient in ridge regression
  - $\mathcal{L}^{\mathrm{val}}(\lambda; \boldsymbol{z}) = \frac{1}{2}\left(y - \boldsymbol{x}^T \widehat{\boldsymbol{\theta}}(\lambda)\right)^2 + \frac{1}{2}\lambda \left\|\widehat{\boldsymbol{\theta}}(\lambda)\right\|_2^2$
  - $\nabla_\lambda \mathcal{L}^{\mathrm{val}}(\lambda; \boldsymbol{z}) = \frac{1}{2}\left\|\widehat{\boldsymbol{\theta}}(\lambda)\right\|_2^2 + \left[\left(y - \boldsymbol{x}^T \widehat{\boldsymbol{\theta}}(\lambda)\right)(-\boldsymbol{x}) + \lambda \widehat{\boldsymbol{\theta}}(\lambda)\right] \nabla_\lambda \widehat{\boldsymbol{\theta}}(\lambda)$
- A direct approach: unrolling differentiation
  - $\widehat{\boldsymbol{\theta}}_k(\boldsymbol{\lambda}) = \widehat{\boldsymbol{\theta}}_{k-1}(\boldsymbol{\lambda}) - \eta \nabla_{\boldsymbol{\theta}} \ell^{\mathrm{tr}}\left(\boldsymbol{\lambda}, \widehat{\boldsymbol{\theta}}_{k-1}(\boldsymbol{\lambda})\right),$
  - $\nabla_{\boldsymbol{\lambda}} \widehat{\boldsymbol{\theta}}_k(\boldsymbol{\lambda}) = \nabla_{\boldsymbol{\lambda}} \widehat{\boldsymbol{\theta}}_{k-1}(\boldsymbol{\lambda}) - \eta \nabla^2{}_{\boldsymbol{\theta}\boldsymbol{\lambda}} \ell^{\mathrm{tr}}\big(\boldsymbol{\lambda}, \widehat{\boldsymbol{\theta}}_{k-1}\big) - \eta \nabla^2{}_{\boldsymbol{\theta}\boldsymbol{\theta}} \ell^{\mathrm{tr}}\big(\boldsymbol{\lambda}, \widehat{\boldsymbol{\theta}}_{k-1}\big) \nabla_{\boldsymbol{\lambda}} \widehat{\boldsymbol{\theta}}_{k-1}(\boldsymbol{\lambda})$

Tight stability bounds

In the case of the constructed example

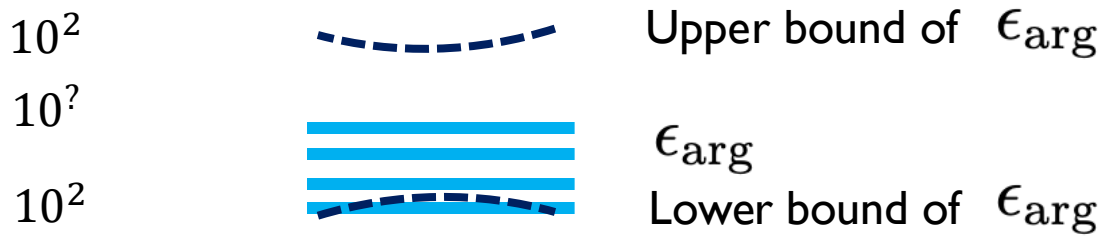
- When $\alpha_t = c/t$,

Thm 4 (Deformed argument stability bounds, [2]).

$\mathcal{L}^{\mathrm{val}}(\boldsymbol{\lambda})$ is L-Lipschitz continuous and γ-smooth, the outer step size is $\alpha_t \leq \frac{c}{t}$, $\ell^{\mathrm{tr}}(\boldsymbol{\lambda}, \boldsymbol{\theta})$ is γ^{tr}-smooth w.r.t. $\boldsymbol{\theta}$ and the inner step size $\eta_k = \eta$, then for all S^{tr}, S^{val} and $\tilde{S}^{\mathrm{val}}$, we have

$$\Omega\left(\frac{T^{\ln\left(1+\left(1-\frac{1}{m}\right)c\gamma\right)}}{m}\right) \leq \epsilon_{\mathrm{arg}} \leq \mathcal{O}\left(\frac{T^{\left(1-\frac{1}{m}\right)c\gamma}}{m}\right),$$

where $\gamma = \Theta\left((1+\eta\gamma^{tr})^{2K}\right)$.



Conclusion

- The fundamental relation between generalization and stability:
  - In expectation, generalization equals stability
  - $\epsilon_{\text{gen}}(A) \leq \epsilon_{\text{stab}}(A) \leq L\epsilon_{\text{arg}}(A)$

- Hyperparameter optimization (HO) problems and gradient-based algorithms
  - The core is to calculate the hypergradient

- Matching argument stability upper and lower bounds
  - $\Omega\left(\frac{T^{\ln\left(1+\left(1-\frac{1}{m}\right)c\gamma\right)}}{m}\right) \leq \epsilon_{\text{arg}} \leq \mathcal{O}\left(\frac{T^{\left(1-\frac{1}{m}\right)c\gamma}}{m}\right)$, where $\gamma = \Theta\left((1+\eta\gamma^{tr})^{2K}\right)$
  - The current upper bound cannot be improved

Stability upper bound

Recursion formula

- $\epsilon_{\text{arg}}(A) = \sup_{S^{\text{val}}, \tilde{S}^{\text{val}}, S^{\text{tr}}} L\mathbb{E}_A\left[\left\|A(S^{\text{val}}, S^{\text{tr}}) - A(\tilde{S}^{\text{val}}, S^{\text{tr}})\right\|\right] = \sup_{S^{\text{val}}, \tilde{S}^{\text{val}}, S^{\text{tr}}} \mathbb{E}_A\left[\left\|\boldsymbol{\lambda}_T - \tilde{\boldsymbol{\lambda}}_T\right\|\right]$
- $\mathbb{E}_A\left[\left\|\boldsymbol{\lambda}_T - \tilde{\boldsymbol{\lambda}}_T\right\|\right] = \mathbb{E}_{\mathcal{A}}\left[\left\|\boldsymbol{\lambda}_{t-1} - \alpha_t \nabla_{\boldsymbol{\lambda}} \mathcal{L}^{\text{val}}\left(\boldsymbol{\lambda}_{t-1}; \boldsymbol{z}_{i_t}^{\text{val}}\right) - \left(\tilde{\boldsymbol{\lambda}}_{t-1} - \alpha_t \nabla_{\boldsymbol{\lambda}} \mathcal{L}^{\text{val}}\left(\tilde{\boldsymbol{\lambda}}_{t-1}; \tilde{\boldsymbol{z}}_{i_t}^{\text{val}}\right)\right)\right\|\right]$

$$\leq \mathbb{E}_A\left[\left\|\boldsymbol{\lambda}_{t-1} - \tilde{\boldsymbol{\lambda}}_{t-1}\right\|\right] + \alpha_t \mathbb{E}_A\left[\left\|\nabla_{\boldsymbol{\lambda}} \mathcal{L}^{\text{val}}\left(\boldsymbol{\lambda}_{t-1}; \boldsymbol{z}_{i_t}^{\text{val}}\right) - \nabla_{\boldsymbol{\lambda}} \mathcal{L}^{\text{val}}\left(\tilde{\boldsymbol{\lambda}}_{t-1}; \tilde{\boldsymbol{z}}_{i_t}^{\text{val}}\right)\right\|\right]$$

If $\boldsymbol{z}_{i_t}^{\text{val}} = \tilde{\boldsymbol{z}}_{i_t}^{\text{val}}$, leverage smoothness $\leq \gamma\left\|\boldsymbol{\lambda}_{t-1} - \tilde{\boldsymbol{\lambda}}_{t-1}\right\|$

If $\boldsymbol{z}_{i_t}^{\text{val}} \neq \tilde{\boldsymbol{z}}_{i_t}^{\text{val}}$, leverage continuity $\leq 2L$

$$\leq \mathbb{E}_A\left[\left\|\boldsymbol{\lambda}_{t-1} - \tilde{\boldsymbol{\lambda}}_{t-1}\right\|\right] + \alpha_t \left[\left(1 - \frac{1}{m}\right)\mathbb{E}_A\left[\left\|\boldsymbol{\lambda}_{t-1} - \tilde{\boldsymbol{\lambda}}_{t-1}\right\|\right] + \frac{1}{m} 2L\right]$$

$$= \left[1 + \alpha_t\left(1 - \frac{1}{m}\right)\gamma\right]\mathbb{E}_A\left[\left\|\boldsymbol{\lambda}_{t-1} - \tilde{\boldsymbol{\lambda}}_{t-1}\right\|\right] + \alpha_t \frac{1}{m} 2L$$

Stability upper bound

- Unroll the recursion:

- **Thm 1** (Argument stability upper bound, [1]).

Suppose $\mathcal{L}^{\mathrm{val}}(\boldsymbol{\lambda})$ *is L-Lipschitz continuous and* γ*-smooth, the outer step size is* $\alpha_t \leq \frac{c}{t}$*, the gradient of* $\ell^{\mathrm{tr}}(\boldsymbol{\lambda}, \boldsymbol{\theta})$ *is* γ^{tr}*-smooth w.r.t.* $\boldsymbol{\theta}$ *and the inner step size* $\eta_k = \eta$*, then for all* S^{tr}, S^{val} *and* $\tilde{S}^{\mathrm{val}}$*, we have*

$$\epsilon_{\mathrm{arg}} \leq \sum_{t=1}^{T} \prod_{s=t+1}^{T+1} \left(1 + \left(1 - \frac{1}{m}\right) \alpha_s \gamma\right) \frac{2\alpha_t L}{m} \ ,$$

where $L = \mathcal{O}\left((1 + \eta\gamma^{tr})^K\right), \gamma = \mathcal{O}\left((1 + \eta\gamma^{tr})^{2K}\right)$.

- Considering T and m:When $\alpha_t = \alpha, \epsilon_{\mathrm{arg}} = \mathcal{O}\left(\left(1 + \left(1 - \frac{1}{m}\right) \alpha\gamma\right)^T / m\right)$

Stability upper bound

- When $\alpha_t = c/t$,

- **Thm 2** (Deformed argument stability upper bound, [2]).

Suppose $\mathcal{L}^{\mathrm{val}}(\boldsymbol{\lambda})$ *is L-Lipschitz continuous and* γ*-smooth, the outer step size is* $\alpha_t \leq \frac{c}{t}$*, the gradient of* $\ell^{\mathrm{tr}}(\boldsymbol{\lambda}, \boldsymbol{\theta})$ *is* γ^{tr}*-smooth w.r.t.* $\boldsymbol{\theta}$ *and the inner step size* $\eta_k = \eta$*, then for all* S^{tr}, S^{val} *and* $\tilde{S}^{\mathrm{val}}$*, we have*

$$\epsilon_{\mathrm{arg}} \leq \left(1 + \left(1 - \frac{1}{m}\right) c\gamma\right) T^{\left(1-\frac{1}{m}\right)c\gamma} \frac{2L}{(m-1)\gamma} = \mathcal{O}\left(\frac{T^{\left(1-\frac{1}{m}\right)c\gamma}}{m}\right),$$

where $\gamma = \mathcal{O}\left((1 + \eta\gamma^{tr})^{2K}\right)$.

Deformed stability upper bound

$$
\begin{aligned}
\epsilon_{\arg} &\leq \sum_{t=1}^{T} \prod_{s=t+1}^{T} \big(1+\alpha_s(1-1/m)\gamma\big)\frac{2\alpha_t L}{m} \\
&\leq \sum_{t=1}^{T-1} \prod_{s=t+1}^{T} \exp\left[\left(1-\frac{1}{m}\right)\frac{\gamma c}{s}\right]\frac{2cL}{tm} + \frac{2cL}{Tm} \\
&= \sum_{t=1}^{T-1} \exp\left[\left(1-\frac{1}{m}\right)\gamma c \sum_{s=t+1}^{T}\frac{1}{s}\right]\frac{2cL}{tm} + \frac{2cL}{Tm} \\
&\leq \sum_{t=1}^{T-1} \exp\left[\left(1-\frac{1}{m}\right)\gamma c \ln\frac{T}{t}\right]\frac{2cL}{tm} + \frac{2cL}{Tm} && (\forall t_2 > t_1 > 1, \textstyle\sum_{t=t_1}^{t_2}\frac{1}{t} \leq \ln\frac{t_2}{t_1-1}) \\
&= \sum_{t=1}^{T}\left(\frac{T}{t}\right)^{(1-1/m)\gamma c}\frac{2cL}{tm} \\
&= \frac{2cL}{m}T^{(1-1/m)\gamma c}\sum_{t=1}^{T} t^{-(1-1/m)\gamma c-1} \\
&\leq \frac{2cL}{m}T^{(1-1/m)\gamma c}\left(1+\int_1^T t^{-(1-1/m)\gamma c-1}dt\right) && (\forall a>0, \textstyle\sum_{t=1}^{T} t^{-a-1} \leq 1+\int_1^T t^{-a-1}dt) \\
&= \frac{2cL}{m(1-1/m)c\gamma}\left[\Big(1+(1-1/m)\gamma c\Big)T^{(1-1/m)\gamma c}-1\right] \\
&= \frac{2L}{(m-1)\gamma}\left[\Big(1+(1-1/m)\gamma c\Big)T^{(1-1/m)\gamma c}-1\right] \\
&= \mathcal{O}(\frac{T^{(1-\frac{1}{m})c\gamma}}{m}),
\end{aligned}
$$

Integral and exponential inqualities

Stability upper bound

- $\epsilon_{\text{arg}}(A) = \sup_{S^{\text{val}}, \tilde{S}^{\text{val}}, S^{\text{tr}}} L\mathbb{E}_A\left[\left\|A(S^{\text{val}}, S^{\text{tr}}) - A(\tilde{S}^{\text{val}}, S^{\text{tr}})\right\|\right] = \sup_{S^{\text{val}}, \tilde{S}^{\text{val}}, S^{\text{tr}}} \mathbb{E}_A\left[\left\|\boldsymbol{\lambda}_T - \tilde{\boldsymbol{\lambda}}_T\right\|\right]$
- $\mathbb{E}_A\left[\left\|\boldsymbol{\lambda}_T - \tilde{\boldsymbol{\lambda}}_T\right\|\right] = \mathbb{E}_{\mathcal{A}}\left[\left\|\boldsymbol{\lambda}_{t-1} - \alpha_t \nabla_{\boldsymbol{\lambda}} \mathcal{L}^{\text{val}}(\boldsymbol{\lambda}_{t-1}; \boldsymbol{z}_{i_t}^{\text{val}}) - \left(\tilde{\boldsymbol{\lambda}}_{t-1} - \alpha_t \nabla_{\boldsymbol{\lambda}} \mathcal{L}^{\text{val}}(\tilde{\boldsymbol{\lambda}}_{t-1}; \tilde{\boldsymbol{z}}_{i_t}^{\text{val}})\right)\right\|\right]$

$$\text{①} \quad \leq \mathbb{E}_A\left[\left\|\boldsymbol{\lambda}_{t-1} - \tilde{\boldsymbol{\lambda}}_{t-1}\right\|\right] + \alpha_t \mathbb{E}_A\left[\left\|\nabla_{\boldsymbol{\lambda}} \mathcal{L}^{\text{val}}(\boldsymbol{\lambda}_{t-1}; \boldsymbol{z}_{i_t}^{\text{val}}) - \nabla_{\boldsymbol{\lambda}} \mathcal{L}^{\text{val}}(\tilde{\boldsymbol{\lambda}}_{t-1}; \tilde{\boldsymbol{z}}_{i_t}^{\text{val}})\right\|\right]$$

If $\boldsymbol{z}_{i_t}^{\text{val}} = \tilde{\boldsymbol{z}}_{i_t}^{\text{val}}$, leverage smoothness

$$\text{②} \quad \leq \gamma\left\|\boldsymbol{\lambda}_{t-1} - \tilde{\boldsymbol{\lambda}}_{t-1}\right\|$$

If $\boldsymbol{z}_{i_t}^{\text{val}} \neq \tilde{\boldsymbol{z}}_{i_t}^{\text{val}}$, leverage continuity

$$\text{③} \quad \leq 2L$$

$$\leq \mathbb{E}_A\left[\left\|\boldsymbol{\lambda}_{t-1} - \tilde{\boldsymbol{\lambda}}_{t-1}\right\|\right] + \alpha_t \left[\left(1 - \frac{1}{m}\right)\mathbb{E}_A\left[\left\|\boldsymbol{\lambda}_{t-1} - \tilde{\boldsymbol{\lambda}}_{t-1}\right\|\right] + \frac{1}{m} 2L\right]$$

$$= \left[1 + \alpha_t\left(1 - \frac{1}{m}\right)\gamma\right]\mathbb{E}_A\left[\left\|\boldsymbol{\lambda}_{t-1} - \tilde{\boldsymbol{\lambda}}_{t-1}\right\|\right] + \alpha_t \frac{1}{m} 2L$$

Equality conditions

$$\mathbb{E}_A\left[\left\|\boldsymbol{\lambda}_{t-1}-\tilde{\boldsymbol{\lambda}}_{t-1}-\alpha_t\left(\nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\mathrm{val}}\left(\boldsymbol{\lambda}_{t-1};\boldsymbol{z}_{i_t}^{\mathrm{val}}\right)-\nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\mathrm{val}}\left(\tilde{\boldsymbol{\lambda}}_{t-1};\tilde{\boldsymbol{z}}_{i_t}^{\mathrm{val}}\right)\right)\right\|\right]$$
$$\leq \mathbb{E}_A\left[\left\|\boldsymbol{\lambda}_{t-1}-\tilde{\boldsymbol{\lambda}}_{t-1}\right\|\right]+\alpha_t\mathbb{E}_A\left[\left\|\nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\mathrm{val}}\left(\boldsymbol{\lambda}_{t-1};\boldsymbol{z}_{i_t}^{\mathrm{val}}\right)-\nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\mathrm{val}}\left(\tilde{\boldsymbol{\lambda}}_{t-1};\tilde{\boldsymbol{z}}_{i_t}^{\mathrm{val}}\right)\right\|\right]$$

- ① Triangle inequality
  - $\boldsymbol{\lambda}_{t-1}-\tilde{\boldsymbol{\lambda}}_{t-1}$ and $-\left(\nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\mathrm{val}}\left(\boldsymbol{\lambda}_{t-1};\boldsymbol{z}_{i_t}^{\mathrm{val}}\right)-\nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\mathrm{val}}\left(\tilde{\boldsymbol{\lambda}}_{t-1};\tilde{\boldsymbol{z}}_{i_t}^{\mathrm{val}}\right)\right)$ are parallel and in the same direction
    - $\boldsymbol{\lambda}_0=\tilde{\boldsymbol{\lambda}}_0, \boldsymbol{\lambda}_1-\tilde{\boldsymbol{\lambda}}_1=-\alpha_1\left(\nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\mathrm{val}}\left(\boldsymbol{\lambda}_0;\boldsymbol{z}_1^{\mathrm{val}}\right)-\nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\mathrm{val}}\left(\tilde{\boldsymbol{\lambda}}_0;\tilde{\boldsymbol{z}}_1^{\mathrm{val}}\right)\right)\triangleq\boldsymbol{v}$
    - $\boldsymbol{v}_1\triangleq\boldsymbol{v}_2$, iff $\exists a\geq 0$, s. t. $\boldsymbol{v}_1=a\boldsymbol{v}_2$
  - ← For all $1\leq t\leq T$, $-\left(\nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\mathrm{val}}\left(\boldsymbol{\lambda}_{t-1};\boldsymbol{z}_{i_t}^{\mathrm{val}}\right)-\nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\mathrm{val}}\left(\tilde{\boldsymbol{\lambda}}_{t-1};\tilde{\boldsymbol{z}}_{i_t}^{\mathrm{val}}\right)\right)\triangleq\boldsymbol{v}$, so that we also have $\boldsymbol{\lambda}_t-\tilde{\boldsymbol{\lambda}}_t\triangleq\boldsymbol{v}$

Equality conditions

$$\text{If } \boldsymbol{z}_{i_t}^{\text{val}} = \tilde{\boldsymbol{z}}_{i_t}^{\text{val}}, \left\|\nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\text{val}}\left(\boldsymbol{\lambda}_{t-1}; \boldsymbol{z}_{i_t}^{\text{val}}\right) - \nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\text{val}}\left(\tilde{\boldsymbol{\lambda}}_{t-1}; \boldsymbol{z}_{i_t}^{\text{val}}\right)\right\| \leq \gamma\left\|\boldsymbol{\lambda}_{t-1} - \tilde{\boldsymbol{\lambda}}_{t-1}\right\|$$

- ② Smoothness:
  - Based on ①, $\nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\text{val}}\left(\boldsymbol{\lambda}_{t-1}; \boldsymbol{z}_{i_t}^{\text{val}}\right) - \nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\text{val}}\left(\tilde{\boldsymbol{\lambda}}_{t-1}; \boldsymbol{z}_{i_t}^{\text{val}}\right) = -\gamma\left(\boldsymbol{\lambda}_{t-1} - \tilde{\boldsymbol{\lambda}}_{t-1}\right)$ for $\boldsymbol{\lambda}_t - \tilde{\boldsymbol{\lambda}}_t \triangleq \boldsymbol{v}$
    - The gradient of $\mathcal{L}^{\text{val}}$ is linear w.r.t. $\boldsymbol{\lambda}$
  - $\leftarrow \mathcal{L}^{\text{val}}$ is quadratic w.r.t. $\boldsymbol{\lambda}$ with term: $\frac{1}{2}\boldsymbol{\lambda}^{\boldsymbol{T}}\boldsymbol{H}\boldsymbol{\lambda}$, where $\boldsymbol{H}\boldsymbol{v} = -\gamma\boldsymbol{v}$

Equality conditions

$$\text{If } \boldsymbol{z}_{i_t}^{\text{val}} \neq \tilde{\boldsymbol{z}}_{i_t}^{\text{val}}, \left\|\nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\text{val}}\left(\boldsymbol{\lambda}_{t-1}; \boldsymbol{z}_{i_t}^{\text{val}}\right) - \nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\text{val}}\left(\tilde{\boldsymbol{\lambda}}_{t-1}; \tilde{\boldsymbol{z}}_{i_t}^{\text{val}}\right)\right\| \leq 2L$$

- ③ Continuity:
  - $\left\|\nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\text{val}}\left(\boldsymbol{\lambda}_{t-1}; \boldsymbol{z}_{i_t}^{\text{val}}\right)\right\| = \left\|\nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\text{val}}\left(\tilde{\boldsymbol{\lambda}}_{t-1}; \tilde{\boldsymbol{z}}_{i_t}^{\text{val}}\right)\right\| = L = \sup_{\boldsymbol{\lambda},\boldsymbol{z}}\left\|\nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\text{val}}(\boldsymbol{\lambda}; \boldsymbol{z})\right\|$
    - Contradictory to ②: require $\mathcal{L}^{\text{val}}$ to be linear
    - $\left\|\nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\text{val}}\left(\boldsymbol{\lambda}_{t-1}; \boldsymbol{z}_{i_t}^{\text{val}}\right) - \nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\text{val}}\left(\tilde{\boldsymbol{\lambda}}_{t-1}; \tilde{\boldsymbol{z}}_{i_t}^{\text{val}}\right)\right\| \geq 2L'$, where $L' \leq L$
  - Based on ① and ②,

$$\begin{aligned}
&\nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\text{val}}\left(\boldsymbol{\lambda}_{t-1}; \boldsymbol{z}_{i_t}^{\text{val}}\right) - \nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\text{val}}\left(\tilde{\boldsymbol{\lambda}}_{t-1}; \tilde{\boldsymbol{z}}_{i_t}^{\text{val}}\right) \\
&= \nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\text{val}}\left(\boldsymbol{\lambda}_{t-1}; \boldsymbol{z}_{i_t}^{\text{val}}\right) - \nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\text{val}}\left(\tilde{\boldsymbol{\lambda}}_{t-1}; \boldsymbol{z}_{i_t}^{\text{val}}\right) + \nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\text{val}}\left(\tilde{\boldsymbol{\lambda}}_{t-1}; \boldsymbol{z}_{i_t}^{\text{val}}\right) - \nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\text{val}}\left(\tilde{\boldsymbol{\lambda}}_{t-1}; \tilde{\boldsymbol{z}}_{i_t}^{\text{val}}\right) \\
&= -\gamma\left(\boldsymbol{\lambda}_{t-1} - \tilde{\boldsymbol{\lambda}}_{t-1}\right) + \nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\text{val}}\left(\tilde{\boldsymbol{\lambda}}_{t-1}; \boldsymbol{z}_{i_t}^{\text{val}}\right) - \nabla_{\boldsymbol{\lambda}}\mathcal{L}^{\text{val}}\left(\tilde{\boldsymbol{\lambda}}_{t-1}; \tilde{\boldsymbol{z}}_{i_t}^{\text{val}}\right)
\end{aligned}$$

    - $\mathcal{L}^{\text{val}}$ has a linear cross term of $\boldsymbol{\lambda}$ and $\boldsymbol{z}$
  - ← $\mathcal{L}^{\text{val}}$ has a linear term: $-y\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\lambda}$ and $y_m\boldsymbol{x}_m - \tilde{y}_m\tilde{\boldsymbol{x}}_m \triangleq \boldsymbol{v}$

Constructed example

- The composite validation loss: $\mathcal{L}^{\mathrm{val}}(\boldsymbol{\lambda}; \boldsymbol{z}) = \frac{1}{2}\boldsymbol{\lambda}^{\boldsymbol{T}}\boldsymbol{H}\boldsymbol{\lambda} - y\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\lambda}$
- $\boldsymbol{H}$ is a real symmetric matric
  - With the maximum absolute eigenvalue γ
  - With eigenvalue $-\gamma$ and a corresponding unit eigenvector $\boldsymbol{v}$
- S^{val} and $\tilde{S}^{\mathrm{val}}$ only differ in the last example, where $\boldsymbol{z}_m^{\mathrm{val}} = (\boldsymbol{x}, y) = (\boldsymbol{v}, 1)$ and $\tilde{\boldsymbol{z}}_m^{\mathrm{val}} = (\tilde{\boldsymbol{x}}, \tilde{y}) = (-\boldsymbol{v}, 1)$

Constructed example

- An HO problem with: $\ell^{\mathrm{val}}(\boldsymbol{\lambda}, \boldsymbol{\theta}; \boldsymbol{z}) = \ell^{\mathrm{tr}}(\boldsymbol{\lambda}, \boldsymbol{\theta}; \boldsymbol{z}) = \frac{1}{2}\boldsymbol{\theta}^{\boldsymbol{T}}\boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\lambda}^{\boldsymbol{T}}\boldsymbol{\theta} - y\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\theta}$
- $\boldsymbol{A}$ is a real symmetric matric
  - With the maximum absolute eigenvalue $\gamma_{\boldsymbol{A}}$
  - With eigenvalue $-\gamma_{\boldsymbol{A}}$ and a corresponding unit eigenvector $\boldsymbol{v}$
- S^{val} and $\tilde{S}^{\mathrm{val}}$ only differ in the last example, where $\boldsymbol{z}_m^{\mathrm{val}} = (\boldsymbol{x}, y) = (\boldsymbol{v}, 1)$ and $\tilde{\boldsymbol{z}}_m^{\mathrm{val}} = (\tilde{\boldsymbol{x}}, \tilde{y}) = (-\boldsymbol{v}, 1)$

Stability lower bound

- $\mathbb{E}_{\mathcal{A}}[\|\boldsymbol{\lambda}_t - \tilde{\boldsymbol{\lambda}}_t\|] = \mathbb{E}_{\mathcal{A}}\left[\left\|\boldsymbol{\lambda}_{t-1} - \alpha_t(\boldsymbol{H}\boldsymbol{\lambda}_{t-1} - y\boldsymbol{x}_{\boldsymbol{i_t}}) - \left(\tilde{\boldsymbol{\lambda}}_{t-1} - \alpha_t(\boldsymbol{H}\boldsymbol{\lambda}_{t-1} - y\tilde{\boldsymbol{x}}_{\boldsymbol{i_t}})\right)\right\|\right]$

$= \mathbb{E}_{\mathcal{A}}[\|\boldsymbol{\lambda}_{t-1} - \tilde{\boldsymbol{\lambda}}_{t-1} - \alpha_t[\boldsymbol{H}(\boldsymbol{\lambda}_{t-1} - \tilde{\boldsymbol{\lambda}}_{t-1}) + (y\boldsymbol{x}_{\boldsymbol{i_t}} - y\tilde{\boldsymbol{x}}_{\boldsymbol{i_t}})]\|]$

① $= \mathbb{E}_{\mathcal{A}}[\|\boldsymbol{\lambda}_{t-1} - \tilde{\boldsymbol{\lambda}}_{t-1}\|] + \alpha_t \mathbb{E}_{\mathcal{A}}[\|-\boldsymbol{H}(\boldsymbol{\lambda}_{t-1} - \tilde{\boldsymbol{\lambda}}_{t-1})\| + \|y\boldsymbol{x}_{\boldsymbol{i_t}} - y\tilde{\boldsymbol{x}}_{\boldsymbol{i_t}}\|]$

If $\boldsymbol{z}_{i_t}^{\text{val}} = \tilde{\boldsymbol{z}}_{i_t}^{\text{val}}$, ② $= \gamma\|(\boldsymbol{\lambda}_{t-1} - \tilde{\boldsymbol{\lambda}}_{t-1})\|$

If $\boldsymbol{z}_{i_t}^{\text{val}} \neq \tilde{\boldsymbol{z}}_{i_t}^{\text{val}}$ ③ $\geq 2L'$

$\geq \mathbb{E}_{\mathcal{A}}[\|\boldsymbol{\lambda}_{t-1} - \tilde{\boldsymbol{\lambda}}_{t-1}\|] + \alpha_t\left[\left(1 - \frac{1}{m}\right)\gamma\mathbb{E}_{\mathcal{A}}[\|\boldsymbol{\lambda}_{t-1} - \tilde{\boldsymbol{\lambda}}_{t-1}\|] + \frac{1}{m}2L'\right]$

$= \left[1 + \alpha_t\left(1 - \frac{1}{m}\right)\gamma\right]\mathbb{E}_{\mathcal{A}}[\|\boldsymbol{\lambda}_{t-1} - \tilde{\boldsymbol{\lambda}}_{t-1}\|] + \alpha_t\frac{1}{m}2L'$

Tight stability bounds

In the case of the constructed example

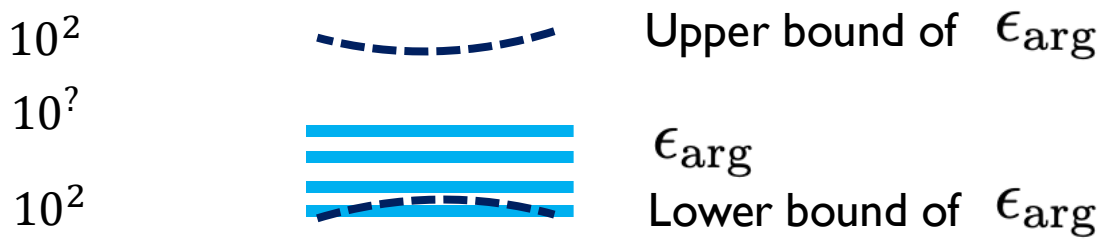
- Unroll the recursion:

- **Thm 3** (Argument stability bounds, [2]).

$\mathcal{L}^{\mathrm{val}}(\boldsymbol{\lambda})$ is L-Lipschitz continuous and γ-smooth, the outer step size is α_t, $\ell^{\mathrm{tr}}(\boldsymbol{\lambda}, \boldsymbol{\theta})$ is γ^{tr}-smooth w.r.t. $\boldsymbol{\theta}$ and the inner step size $\eta_k = \eta$, then for all S^{tr}, S^{val} and $\tilde{S}^{\mathrm{val}}$, we have

$$\sum_{t=1}^{T} \prod_{s=t+1}^{T+1} \left(1 + \left(1 - \frac{1}{m}\right) \alpha_s \gamma\right) \frac{2\alpha_t L'}{m} \leq \epsilon_{\mathrm{arg}} \leq \sum_{t=1}^{T} \prod_{s=t+1}^{T+1} \left(1 + \left(1 - \frac{1}{m}\right) \alpha_s \gamma\right) \frac{2\alpha_t L},$$

where $L' \backsimeq L = \Theta\left((1 + \eta\gamma^{tr})^{K}\right), \gamma = \Theta\left((1 + \eta\gamma^{tr})^{2K}\right).$



Tight stability bounds

In the case of the constructed example

- When $\alpha_t = c/t$,

- **Thm 4** (Deformed argument stability bounds, [2]).

$\mathcal{L}^{\mathrm{val}}(\boldsymbol{\lambda})$ is L-Lipschitz continuous and γ-smooth, the outer step size is $\alpha_t \leq \frac{c}{t}$, $\ell^{\mathrm{tr}}(\boldsymbol{\lambda}, \boldsymbol{\theta})$ is γ^{tr}-smooth w.r.t. $\boldsymbol{\theta}$ and the inner step size $\eta_k = \eta$, then for all S^{tr}, S^{val} and $\tilde{S}^{\mathrm{val}}$, we have

$$\Omega\left(\frac{T^{\ln\left(1+\left(1-\frac{1}{m}\right)c\gamma\right)}}{m}\right) \leq \epsilon_{\mathrm{arg}} \leq \mathcal{O}\left(\frac{T^{\left(1-\frac{1}{m}\right)c\gamma}}{m}\right),$$

where $\gamma = \Theta\left((1+\eta\gamma^{tr})^{2K}\right)$.

Deformed stability upper bound

$$
\begin{aligned}
\epsilon_{\arg} &\geq \sum_{t=1}^{T} \prod_{s=t+1}^{T} \left(1+\alpha_s(1-1/m)\gamma'\right)\frac{2\alpha_t L'}{m} \\
&\geq \sum_{t=1}^{T} \prod_{s=t+1}^{T} \exp\left[r\left(1-\frac{1}{m}\right)\alpha_s\gamma'\right]\frac{2\alpha_t L'}{m} \\
&= \sum_{t=1}^{T-1} \prod_{s=t+1}^{T} \exp\left[r\left(1-\frac{1}{m}\right)\frac{c\gamma'}{s}\right]\frac{2cL'}{tm} + \frac{2cL'}{Tm} && (\alpha_t = c/t) \\
&= \sum_{t=1}^{T-1} \exp\left[r\left(1-\frac{1}{m}\right)c\gamma' \sum_{s=t+1}^{T}\frac{1}{s}\right]\frac{2cL'}{tm} + \frac{2cL'}{Tm} \\
&\geq \sum_{t=1}^{T-1} \exp\left[r\left(1-\frac{1}{m}\right)c\gamma' \ln\frac{T+1}{t+1}\right]\frac{2cL'}{tm} + \frac{2cL'}{Tm} && (\forall t_2 > t_1 > 0, \textstyle\sum_{t=t_1}^{t_2}\frac{1}{t} \geq \ln\frac{t_2+1}{t_1}) \\
&= \sum_{t=1}^{T}\left(\frac{T+1}{t+1}\right)^{r\left(1-1/m\right)c\gamma'}\frac{2cL'}{tm} \\
&\geq \frac{2cL'}{m}(T+1)^{r\left(1-1/m\right)c\gamma'} \sum_{t=2}^{T+1} t^{-r\left(1-1/m\right)c\gamma'-1} \\
&\geq \frac{2cL'}{m}(T+1)^{r\left(1-1/m\right)c\gamma'} \int_{2}^{T+2} t^{-r\left(1-1/m\right)c\gamma'-1}dt && (\forall a > 0, \textstyle\sum_{t=2}^{T+1} t^{-a-1} \geq \int_{2}^{T+2} t^{-a-1}dt) \\
&= \frac{2cL'}{m}(T+1)^{r\left(1-1/m\right)c\gamma'}\left[\frac{2^{-r\left(1-1/m\right)c\gamma'} - (T+2)^{-r\left(1-1/m\right)c\gamma'}}{r\left(1-1/m\right)c\gamma'}\right] \\
&= \frac{2L'}{r(m-1)\gamma'}(T+1)^{r\left(1-1/m\right)c\gamma'}\left[2^{-r\left(1-1/m\right)c\gamma'} - (T+2)^{-r\left(1-1/m\right)c\gamma'}\right] \\
&\geq \frac{2cL'}{m\ln\left(1+(1-1/m)c\gamma'\right)}\left[\left(\frac{T+1}{2}\right)^{\ln\left(1+(1-1/m)c\gamma'\right)} - 1\right], && (r = \tfrac{\ln\left(1+(1-1/m)c\gamma'\right)}{(1-1/m)c\gamma'}) \\
&= \Theta(\frac{T^{\ln\left(1+(1-\frac{1}{m})c\gamma'\right)}}{m}).
\end{aligned}
$$

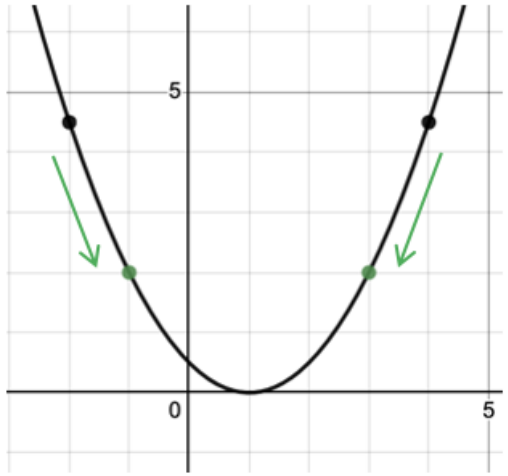
Conclusion

- The fundamental relation between generalization and stability:
  - In expectation, generalization equals stability
  - $\epsilon_{\text{gen}}(A) \leq \epsilon_{\text{stab}}(A) \leq L\epsilon_{\text{arg}}(A)$

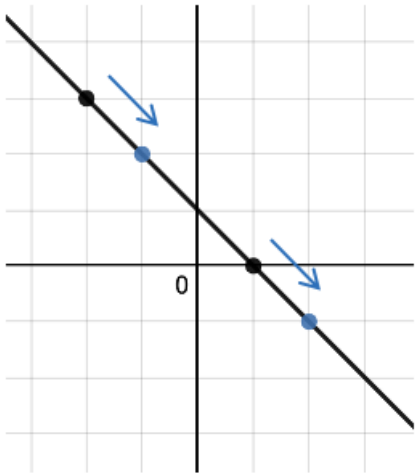
- Hyperparameter optimization (HO) problems and gradient-based algorithms
  - The core is to calculate the hypergradient

- Matching argument stability upper and lower bounds
  - $\Omega\left(\frac{T^{\ln\left(1+\left(1-\frac{1}{m}\right)c\gamma\right)}}{m}\right) \leq \epsilon_{\text{arg}} \leq \mathcal{O}\left(\frac{T^{\left(1-\frac{1}{m}\right)c\gamma}}{m}\right)$, where $\gamma = \Theta\left((1+\eta\gamma^{tr})^{2K}\right)$
  - The current upper bound cannot be improved

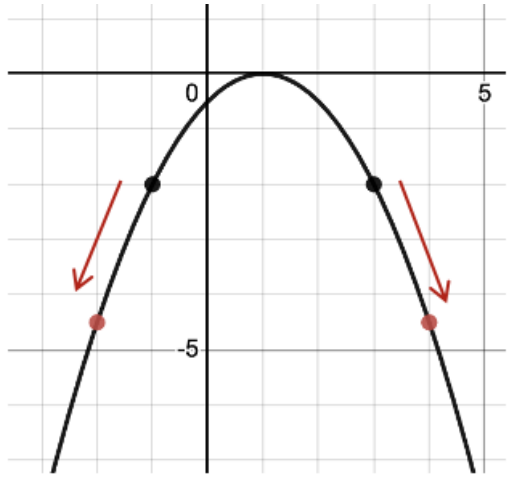
Recap: Uniform stability of SGD



strongly convex



convex



non-convex