Advanced deep learning optimizers and convergence

Acknowledgement

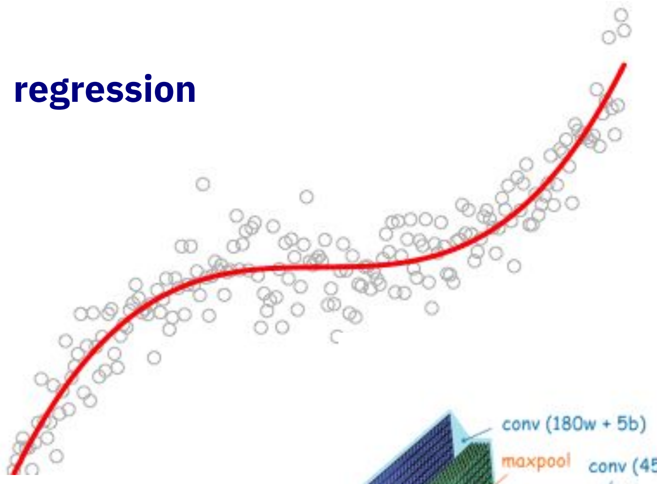
Special thanks to the work of Dr. Haitham Bou Ammar

Part of this lecture’s content is based on the presentation of Machine Learning and AI Academy

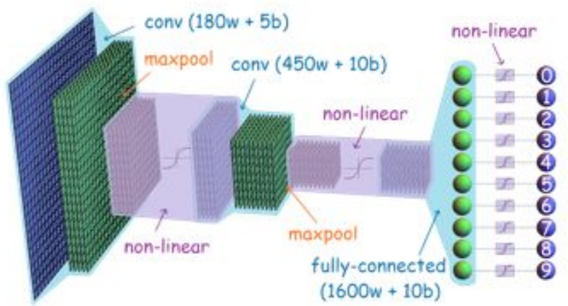
- A group of experts in ML and AI with PhDs from top-tier schools and universities

Why Optimization?

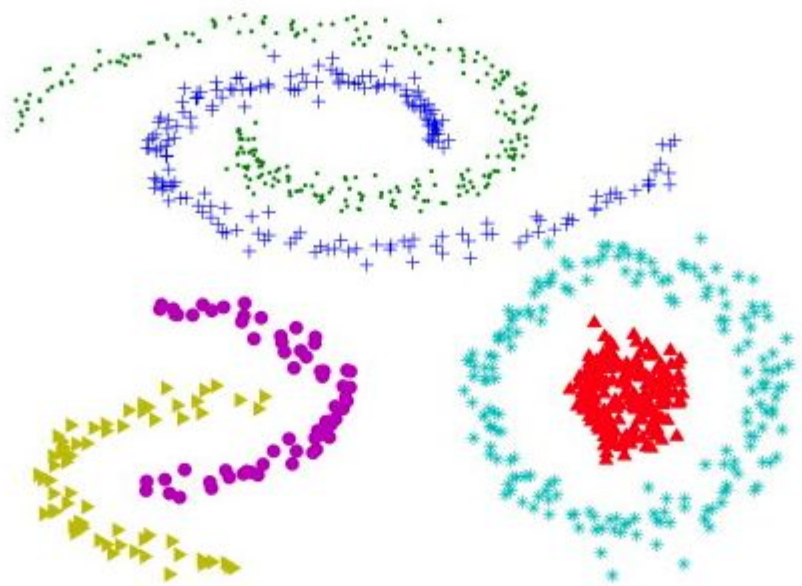
regression



classification



clustering/density estimation



computer games

robotics

Supervised Learning

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{j=1}^{n} \mathcal{L}_{\boldsymbol{\theta}} \left(\mathbf{x}^{(i)}, y^{(i)} \right)$$

Unsupervised Learning

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{j=1}^{n} \mathcal{L}_{\boldsymbol{\theta}} \left(\mathbf{x}^{(i)} \right)$$

Reinforcement Learning

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{\tau} \sim p_{\boldsymbol{\theta}}(\boldsymbol{\tau})} \left(\mathcal{R}_{\text{total}}(\boldsymbol{\tau}) \right)$$

... all these involve a minimization of some function ...

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} f(\boldsymbol{\theta})$$

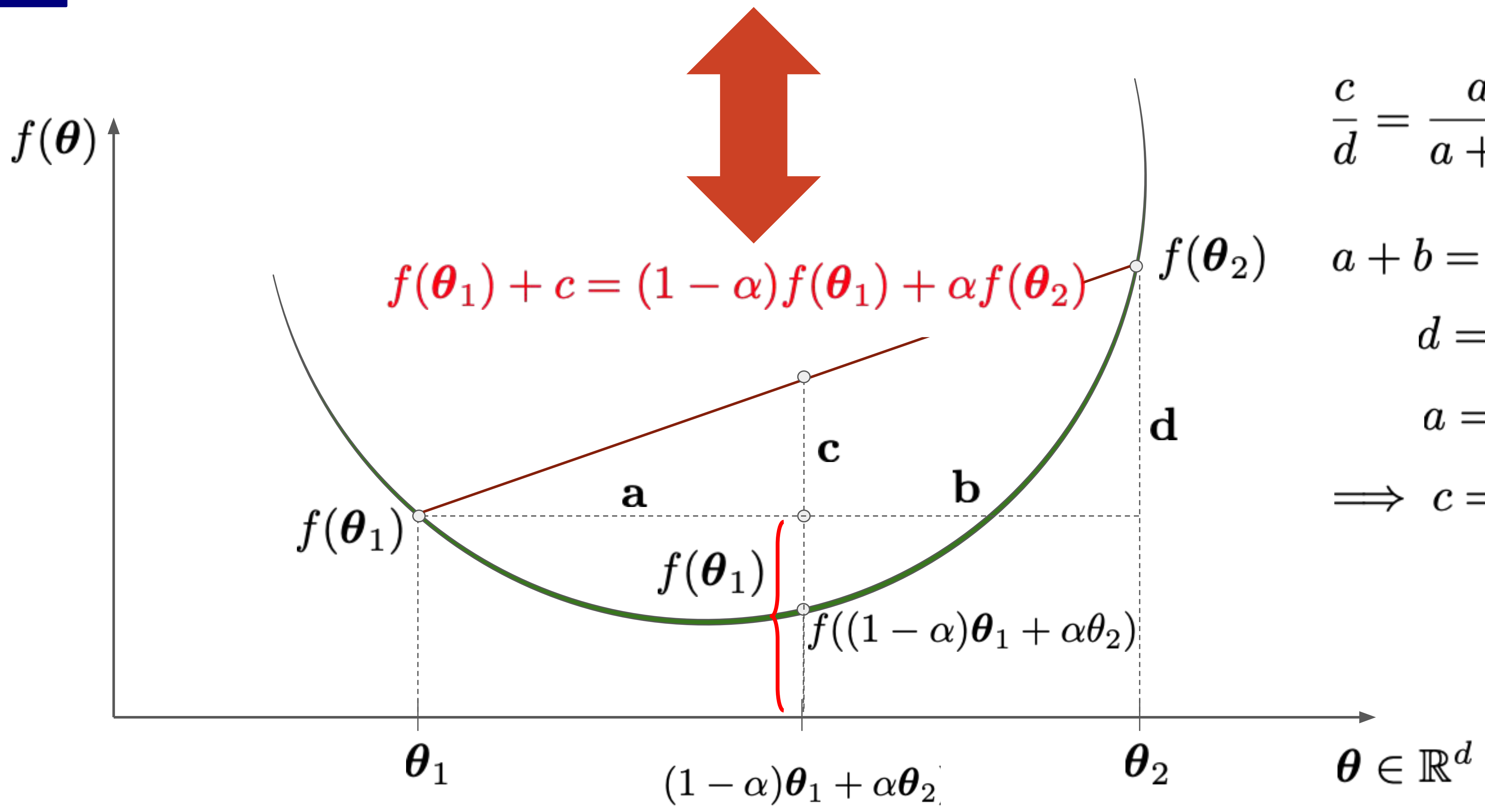
Optimization in Deep Learning

- Training Neural Network Models
  - Will your method hurt the convergence (optimization process)?
- Distributed Machine Learning
  - Will your method hurt the convergence (optimization process)?
- Gradients Compression
  - Will your method hurt the convergence (optimization process)?
- Collaboration with application people
  - We are not allowed to change the model or dataset.
  - We can only improve the optimization part

Function types, and what one can hope for ...

... optimizing for unknown parameters depends on the type of function under study ...

Convex $f((1-\alpha)\boldsymbol{\theta}_1 + \alpha\boldsymbol{\theta}_2) \leq (1-\alpha)f(\boldsymbol{\theta}_1) + \alpha f(\boldsymbol{\theta}_2) \quad \forall\alpha \in (0,1) \quad \forall\boldsymbol{\theta}_1,\ \boldsymbol{\theta}_2 \in \mathbb{R}^d$



any point in between

$$\frac{c}{d} = \frac{a}{a+b}$$

$$a + b = ||\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1||_2$$

$$d = f(\boldsymbol{\theta}_2) - f(\boldsymbol{\theta}_1)$$

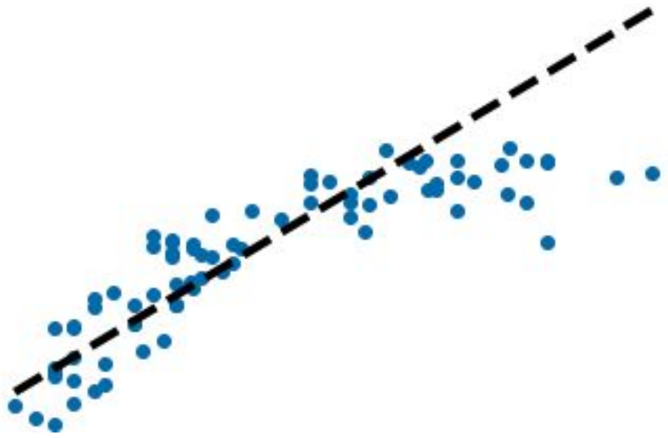
$$a = \alpha||\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1||_2$$

$$\implies c = \alpha f(\boldsymbol{\theta}_2) - \alpha f(\boldsymbol{\theta}_1)$$

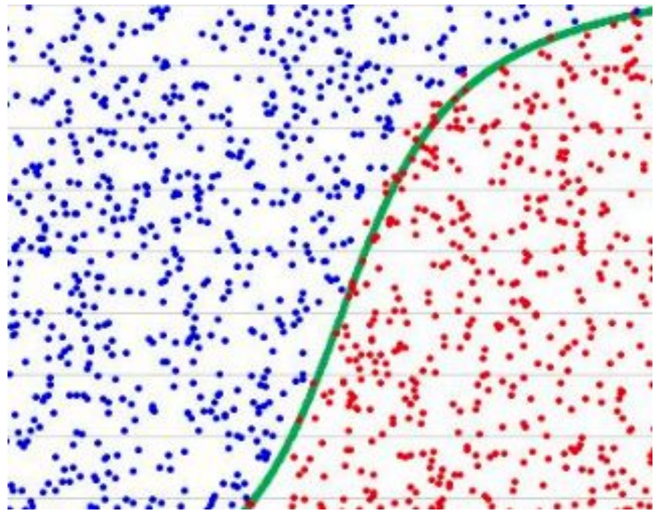
Function types, and what one can hope for …

… optimizing for unknown parameters depends on the type of function under study …

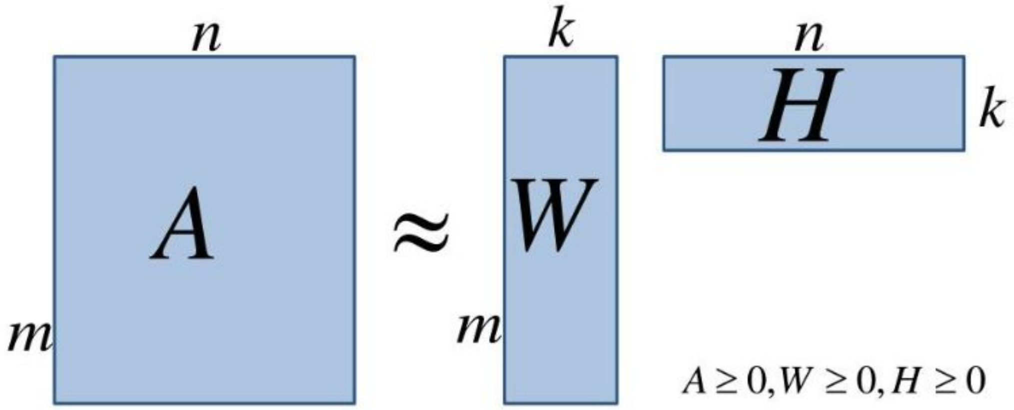
Convex $f((1-\alpha)\boldsymbol{\theta}_1 + \alpha\boldsymbol{\theta}_2) \leq (1-\alpha)f(\boldsymbol{\theta}_1) + \alpha f(\boldsymbol{\theta}_2) \quad \forall\alpha \in (0,1) \quad \forall\boldsymbol{\theta}_1,\ \boldsymbol{\theta}_2 \in \mathbb{R}^d$



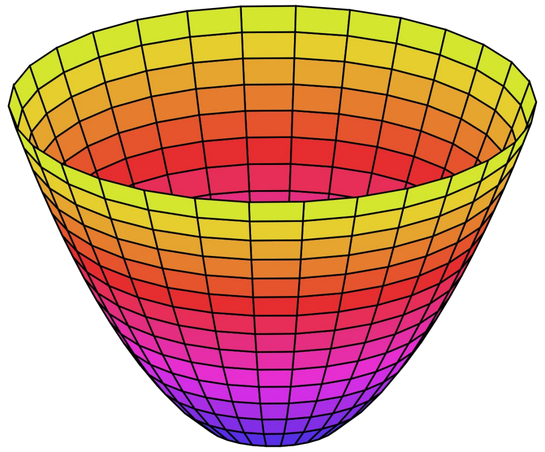
Linear Regression



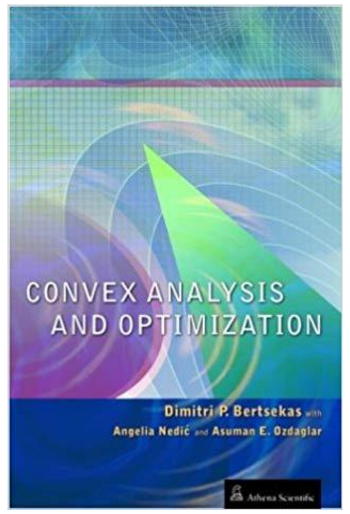
Classification with Hinge Loss



Non-Negative Matrix Factorisation



Unique global minimum



Function types, and what one can hope for ...

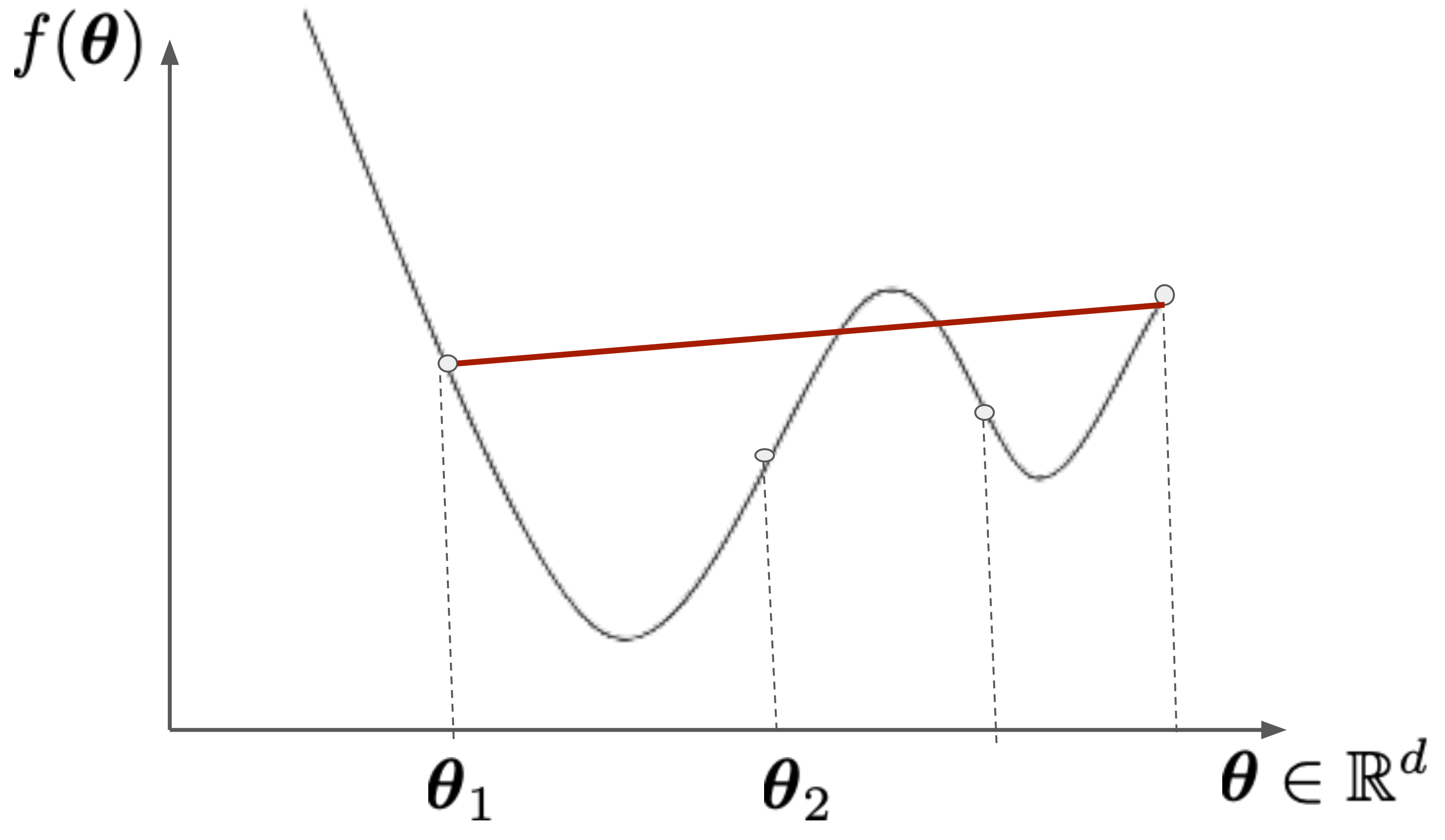
... optimising for unknown parameters depends on the type of function under study ...

Non-Convex **... we want to negate the convex definition (and avoid concave definition) ...**

$$\exists \;\; \boldsymbol{\theta}_1, \;\; \boldsymbol{\theta}_2, \;\; \text{and} \;\; \alpha \in (0,1) \;\; \text{such that} \;\; f\left((1-\alpha_1)\boldsymbol{\theta}_1 + \alpha\boldsymbol{\theta}_2\right) > (1-\alpha)f(\boldsymbol{\theta}_1) + \alpha f(\boldsymbol{\theta}_2)$$

&

$$\exists \;\; \tilde{\boldsymbol{\theta}}_1, \;\; \tilde{\boldsymbol{\theta}}_2, \;\; \text{and} \;\; \tilde{\alpha} \in (0,1) \;\; \text{such that} \;\; f\left((1-\tilde{\alpha}_1)\tilde{\boldsymbol{\theta}}_1 + \tilde{\alpha}\tilde{\boldsymbol{\theta}}_2\right) < (1-\tilde{\alpha})f(\tilde{\boldsymbol{\theta}}_1) + \tilde{\alpha} f(\tilde{\boldsymbol{\theta}}_2)$$

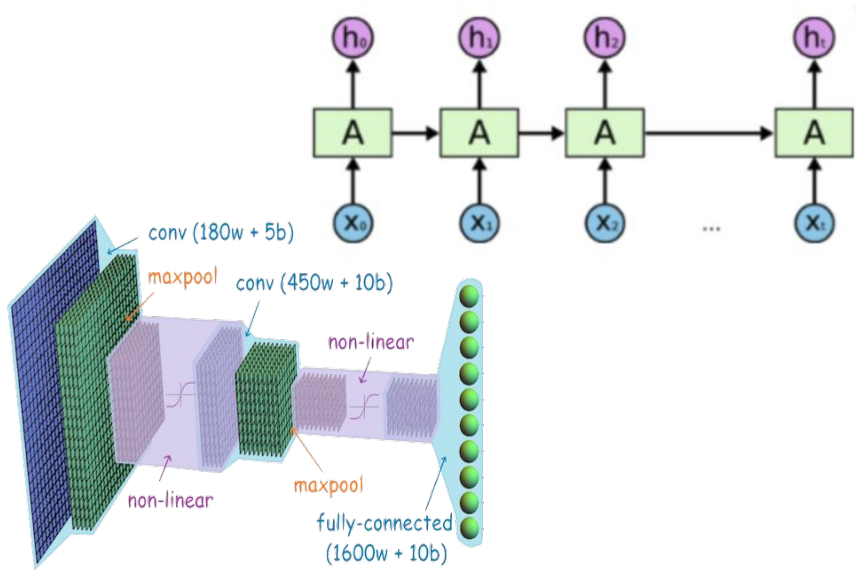


Function types, and what one can hope for …

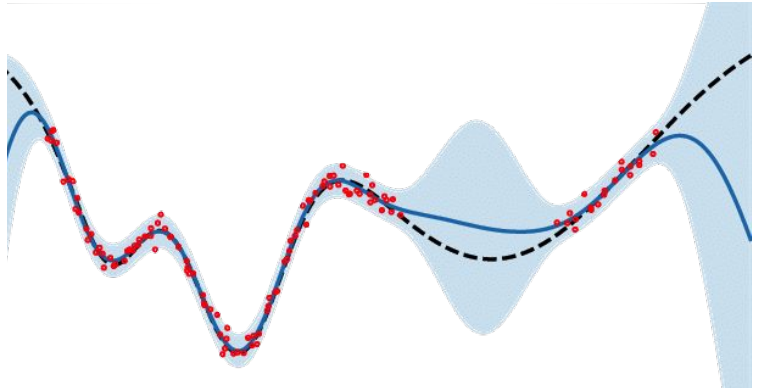
… optimising for unknown parameters depends on the type of function under study …

Non-Convex **… we want to negate the convex definition (and avoid concave definition) …**

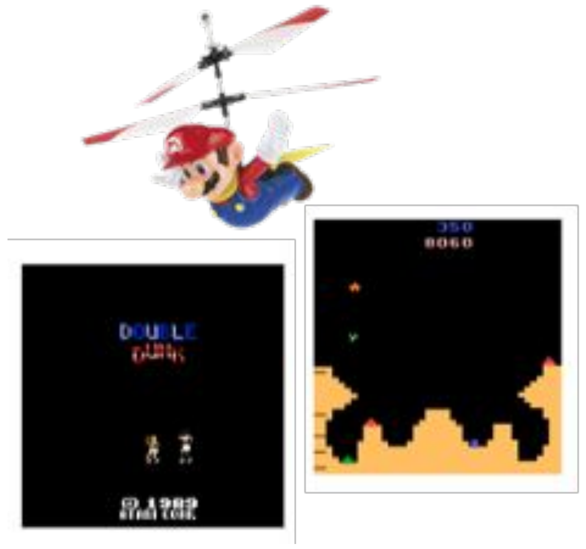
$$\begin{rcases}\exists\ \boldsymbol{\theta}_1,\ \boldsymbol{\theta}_2,\ \text{and}\ \alpha \in (0,1)\ \text{such that}\ f\left((1-\alpha_1)\boldsymbol{\theta}_1 + \alpha\boldsymbol{\theta}_2\right) > (1-\alpha)f(\boldsymbol{\theta}_1) + \alpha f(\boldsymbol{\theta}_2) \\ \exists\ \tilde{\boldsymbol{\theta}}_1,\ \tilde{\boldsymbol{\theta}}_2,\ \text{and}\ \tilde{\alpha} \in (0,1)\ \text{such that}\ f\left((1-\tilde{\alpha}_1)\tilde{\boldsymbol{\theta}}_1 + \tilde{\alpha}\tilde{\boldsymbol{\theta}}_2\right) < (1-\tilde{\alpha})f(\tilde{\boldsymbol{\theta}}_1) + \tilde{\alpha} f(\tilde{\boldsymbol{\theta}}_2)\end{rcases} \&$$



Deep Learning



Gaussian Processes & Bayesian Models

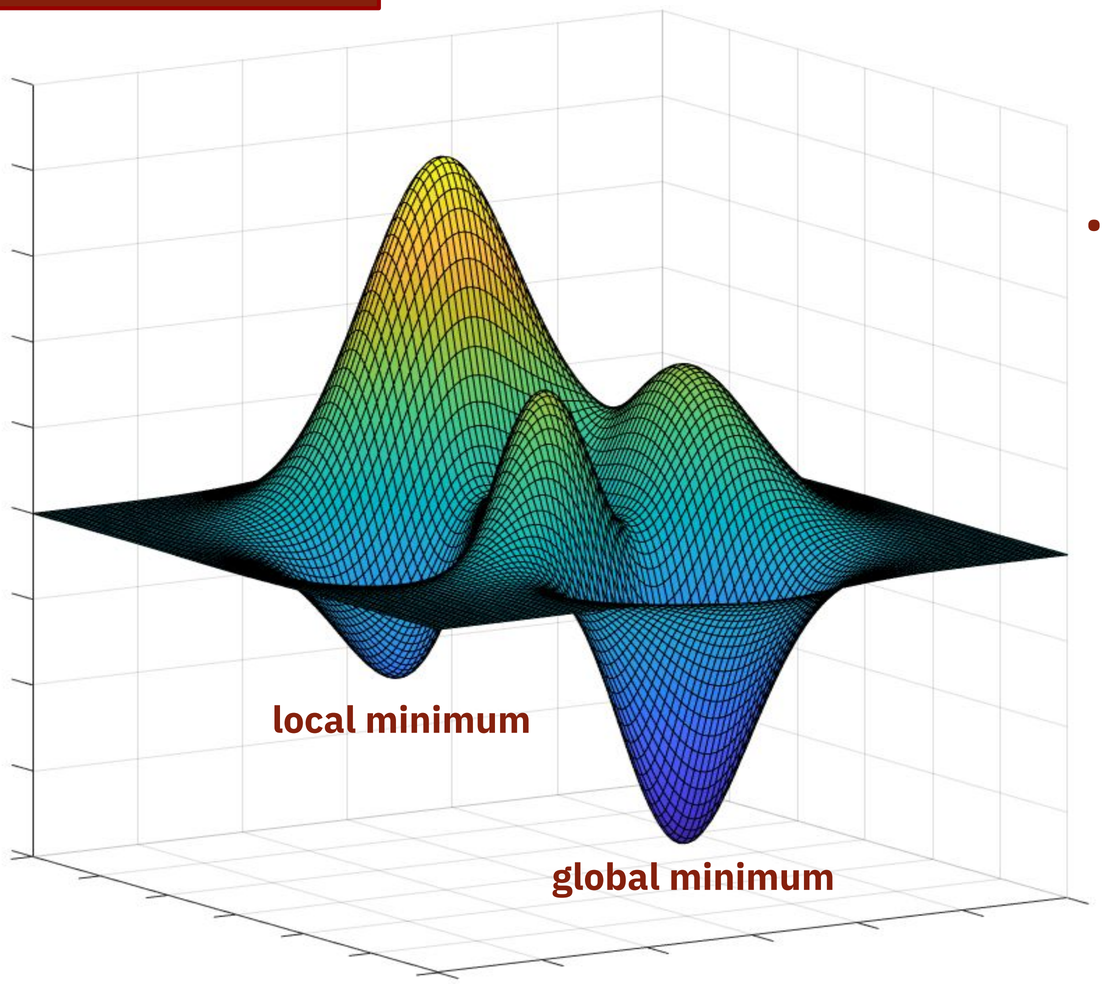


Reinforcement Learning

Function types, and what one can hope for ...

... optimising for unknown parameters depends on the type of function under study ...

Non-Convex ... global and local minima (checking) are NP-Hard, we look for other types of points ...



$$\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_{\text{stationary}}) = \mathbf{0}$$

... so instead, we are searching for stationary points ...

1. ϵ -First-Order-Stationary Point (FOSP): $||\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}_{\text{FOSP}})||_2 \leq \epsilon$

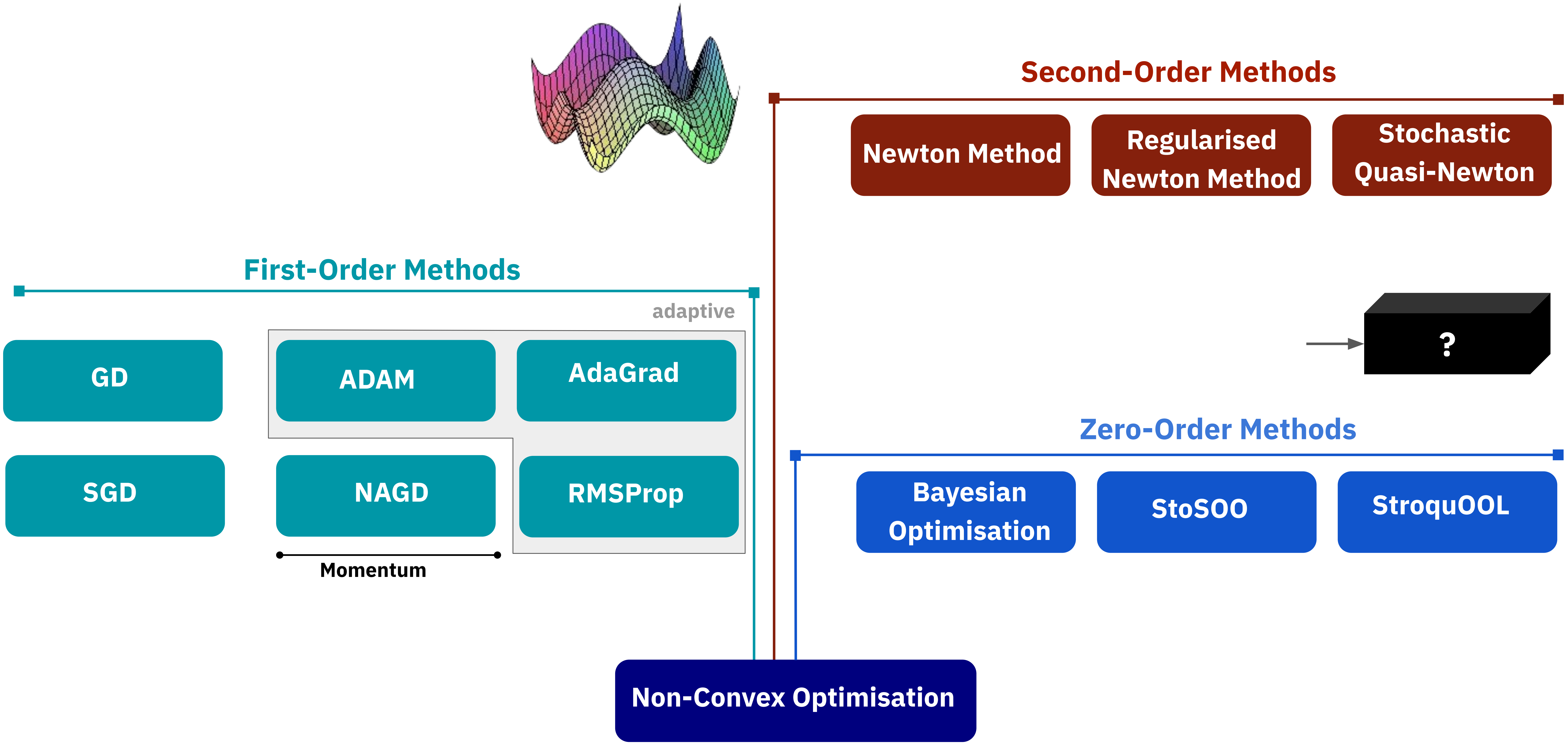
[e.g., all global and local minima, saddle points, plateau points]

2. ϵ- Second-Order-Stationary Point (SOSP):

$$||\nabla_{\boldsymbol{\theta}} f\left(\boldsymbol{\theta}_{\text{SOSP}}\right)||_2 \leq \epsilon \quad \text{and} \quad \lambda_{\min}\left(\nabla^2_{\boldsymbol{\theta},\boldsymbol{\theta}} f(\boldsymbol{\theta}_{\text{SOSP}})\right) \geq -\sqrt{\epsilon}$$

[e.g., all global and local minima, plateau points]

Algorithms vary in type of information used ...



A review on first-order optimizers

An example of loss function (ignore bias correction)

Algorithm's Inputs:

$$\underbrace{\boldsymbol{\theta}_1 \in \mathbb{R}^d}_{\text{Initial param value}} \quad \overbrace{\{\eta_t\}_{t=1}^{T}}^{\text{Learning Rates}} \quad \underbrace{0 \le \beta_1, \beta_2 \le 1}_{\text{Decay parameters}} \quad \overbrace{\delta > 0}^{\text{Stability param}}$$

Update Procedure:

Set $\mathbf{m}_0 = \mathbf{0}$, and $\mathbf{v}_0 = \mathbf{0}$

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

Draw a sample ξ_t from $\mathbb{P}$

Compute $\mathbf{g}_t = \nabla \mathcal{L}(\boldsymbol{\theta}_t, \xi_t)$

Update $\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1)\mathbf{g}_t$

Update $\mathbf{v}_t = \mathbf{v}_{t-1} - (1 - \beta_2)(\mathbf{v}_{t-1} - \mathbf{g}_t^2)$

Update $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \dfrac{\mathbf{g}_t}{(\sqrt{\mathbf{v}_t} + \delta)}$

end for

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_{\boldsymbol{\theta}}(\mathbf{x}_i))^2$$

Sample $\xi_t = i_t \in \{1, \ldots, n\}$

$$\implies \mathcal{L}(\boldsymbol{\theta}, i_t) = (y_{i_t} - f_{\boldsymbol{\theta}}(\mathbf{x}_{i_t}))^2$$

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, i_t) = \nabla_{\boldsymbol{\theta}}(y_{i_t} - f_{\boldsymbol{\theta}}(\mathbf{x}_{i_t}))^2$$

$$= -2(y_{i_t} - f_{\boldsymbol{\theta}}(\mathbf{x}_{i_t}))\nabla f_{\boldsymbol{\theta}}(\mathbf{x}_{i_t})$$

Explainable AI is important

How can we prove Adam can converge?

Knowing this will help you to prove your own algorithm in the future.

From ML to ERM (Empirical risk minimization)

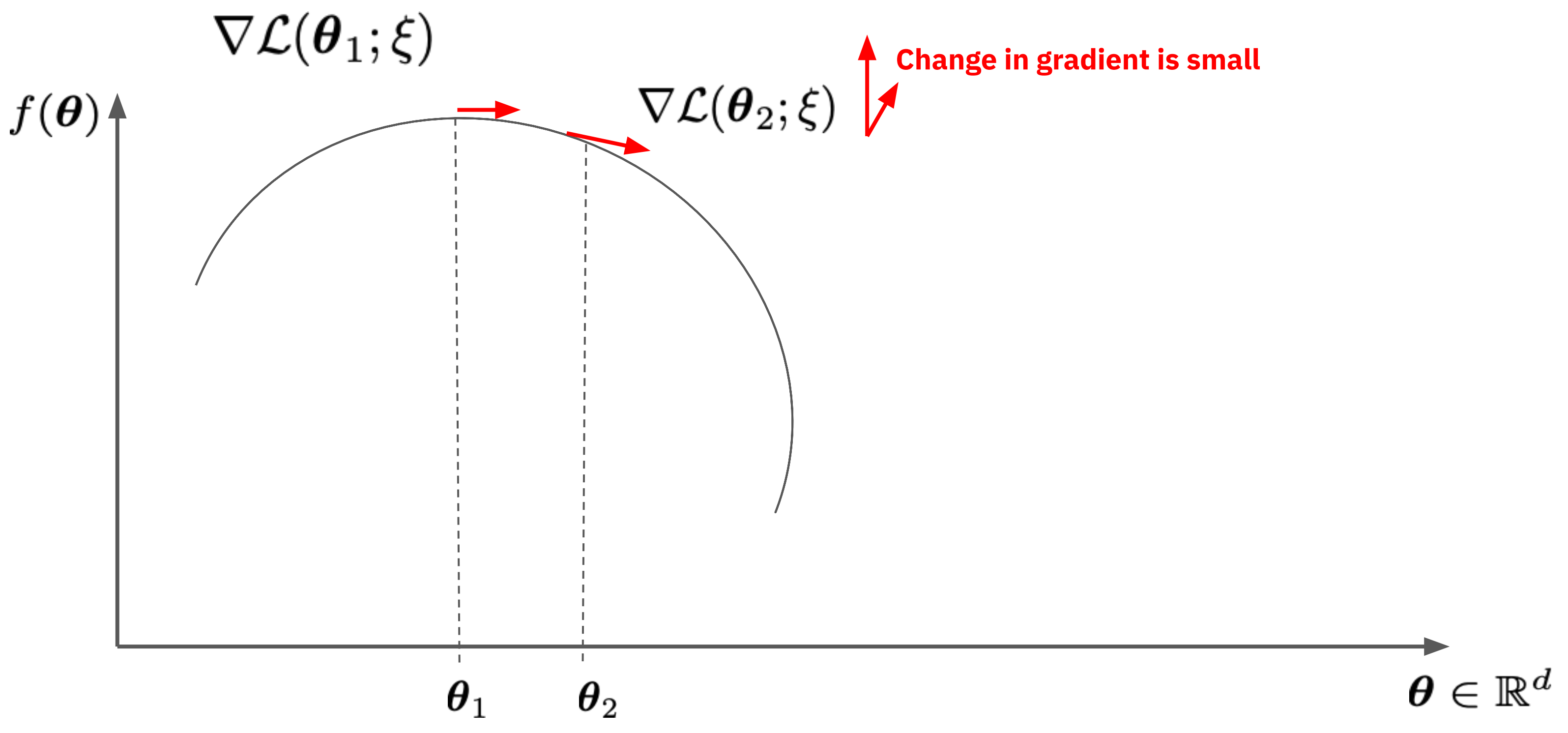
consider the following form of the objective function: $\mathbb{E}_{\xi\sim\mathbb{P}}\left[\mathcal{L}(\boldsymbol{\theta};\xi)\right]$

… for e.g., in regression $\xi \sim \text{Uniform}[1,n]$, then $\mathbb{E}_{\xi\sim\text{Uniform}}[(y_\xi - f_{\boldsymbol{\theta}}(\mathbf{x}_\xi))^2] = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_{\boldsymbol{\theta}}(\mathbf{x}_i))^2$

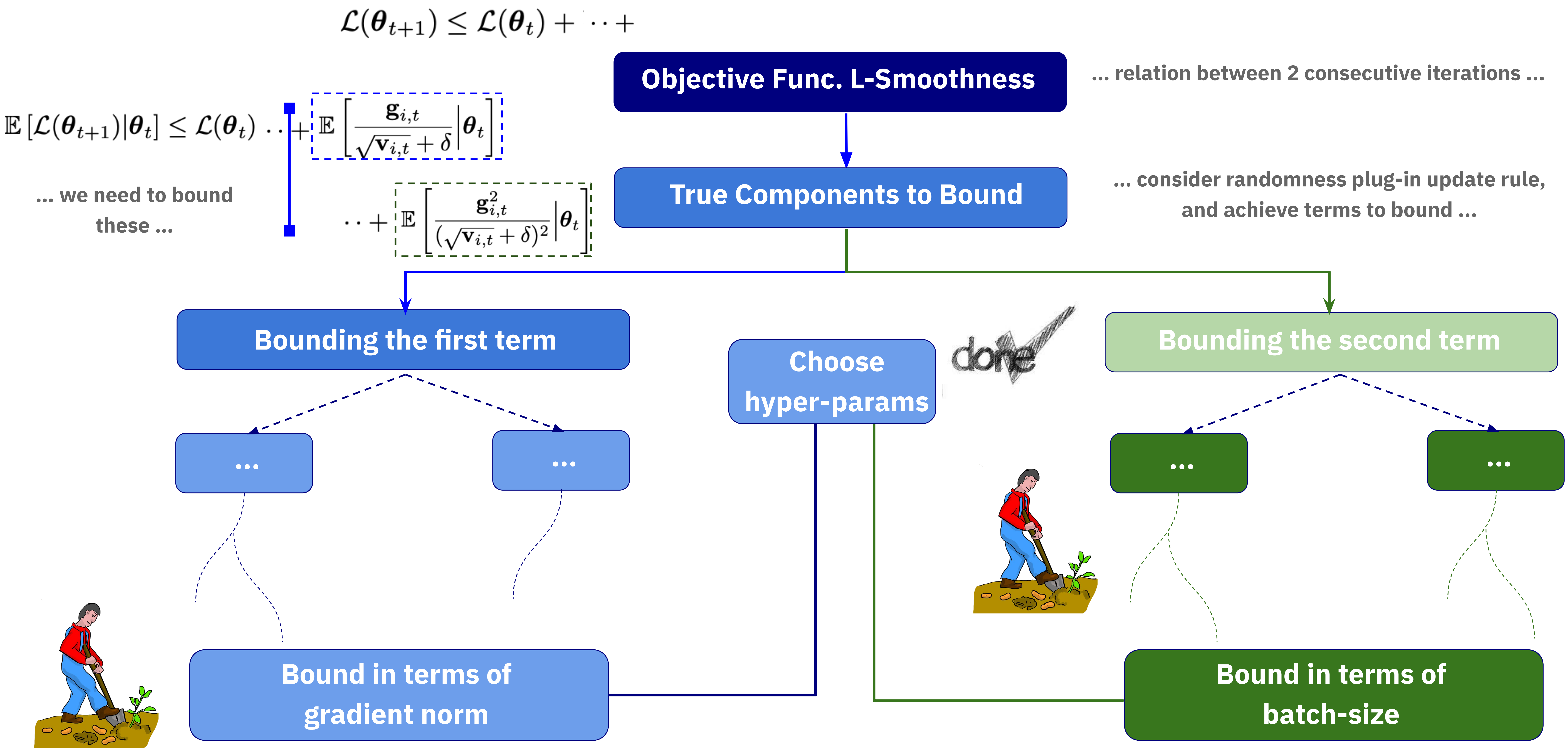
… now, our goal is to minimize the following $\min_{\boldsymbol{\theta}} \mathbb{E}_{\xi\sim\mathbb{P}}\left[\mathcal{L}(\boldsymbol{\theta};\xi)\right]$ **… using ADAM from the previous slide**

Assumption I -- Loss Function is L-Smooth:

$$||\nabla\mathcal{L}(\boldsymbol{\theta}_2;\boldsymbol{\xi}) - \nabla\mathcal{L}(\boldsymbol{\theta}_1;\boldsymbol{\xi})||_2 \leq L||\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1||_2 \;\; \forall\, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\xi}$$



Proof Roadmap ...



Convergence Proof ...

... as in any other optimization proof, we need to understand the change in function value between two consecutive iterations of the algorithm:

$$f(\boldsymbol{\theta}_{t+1}) \leq f(\boldsymbol{\theta}_t) - \Delta \implies \text{convergence to some point if the function is lower-bounded}$$

Some non-negative value

Monotone Convergence Theorem (MCT)

... now, <u>if we can say that the objective function is L-smooth</u>, then we can have a relation between function values on two successive iterations:

$$\mathcal{L}(\boldsymbol{\theta}_{t+1}) \leq \underbrace{\mathcal{L}(\boldsymbol{\theta}_t)}_{\text{Relation between successive iterations}} + \nabla^{\mathsf{T}}\mathcal{L}(\boldsymbol{\theta}_t)\,(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) + \frac{L}{2}\,||\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t||_2^2$$

But how to show that our objective function is L-Smooth

Please see https://xingyuzhou.org/blog/notes/Lipschitz-gradient

Convergence Proof ...

... let us study the norm of the difference between the gradients of the objective function at any two given input points:

$$
\begin{aligned}
||\nabla \mathcal{L}(\boldsymbol{\theta}_1) - \nabla \mathcal{L}(\boldsymbol{\theta}_2)||_2 &= ||\nabla \mathbb{E}_\xi[\mathcal{L}(\boldsymbol{\theta}_1;\xi)] - \nabla \mathbb{E}_\xi[\mathcal{L}(\boldsymbol{\theta}_2;\xi)]||_2 \\
&= ||\mathbb{E}_\xi[\nabla \mathcal{L}(\boldsymbol{\theta}_1;\xi)] - \mathbb{E}_\xi[\nabla \mathcal{L}(\boldsymbol{\theta}_2;\xi)]||_2 \\
&= ||\mathbb{E}_\xi[\nabla \mathcal{L}(\boldsymbol{\theta}_1;\xi) - \nabla \mathcal{L}(\boldsymbol{\theta}_2;\xi)]||_2 \\
&\leq \mathbb{E}_\xi[||\nabla \mathcal{L}(\boldsymbol{\theta}_1;\xi) - \nabla \mathcal{L}(\boldsymbol{\theta}_2;\xi)||_2] \\
&\leq \mathbb{E}_\xi[L||\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1||_2] \\
&= L||\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1||_2
\end{aligned}
$$

Objective function is L-Smooth

Assumption I -- Loss Function is L-Smooth:

$$||\nabla \mathcal{L}(\boldsymbol{\theta}_2;\boldsymbol{\xi}) - \nabla \mathcal{L}(\boldsymbol{\theta}_1;\boldsymbol{\xi})||_2 \leq L||\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1||_2 \;\; \forall\, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\xi}$$

Convergence Proof ...

... since we just proved that our objective is L-Smooth, now we can write that the objective value between two successive iterations abides by:

$$\mathcal{L}(\boldsymbol{\theta}_{t+1}) \leq \mathcal{L}(\boldsymbol{\theta}_t) + \nabla^{\mathsf{T}}\mathcal{L}(\boldsymbol{\theta}_t)\,(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) + \frac{L}{2}\,||\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t||_2^2$$

... now, remember our update rules from the pseudo-code in the previous slides, we can write:

$$\mathbf{m}_t = \beta_1\mathbf{m}_{t-1} + (1-\beta_1)\mathbf{g}_t \implies \text{with } \beta_1 = 0, \text{ then } \mathbf{m}_t = \mathbf{g}_t \quad \text{then} \quad \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t\frac{\mathbf{g}_t}{(\sqrt{\mathbf{v}_t} + \delta)}$$

... component-wise update $\boldsymbol{\theta}_{i,t+1} = \boldsymbol{\theta}_{i,t} - \eta_t\frac{\mathbf{g}_{i,t}}{(\sqrt{\mathbf{v}_{i,t}} + \delta)} \quad i \in \{1, \ldots, d\}$

$$\mathcal{L}(\boldsymbol{\theta}_{t+1}) \leq \mathcal{L}(\boldsymbol{\theta}_t) - \eta_t\sum_{i=1}^{d}\left([\nabla\mathcal{L}(\boldsymbol{\theta}_t)]_i \times \frac{\mathbf{g}_{i,t}}{\sqrt{\mathbf{v}_{i,t}} + \delta}\right) + \frac{L\eta_t^2}{2}\sum_{i=1}^{d}\frac{\mathbf{g}_{i,t}^2}{(\sqrt{\mathbf{v}_{i,t}} + \delta)^2}$$

Convergence Proof ...

$$\mathcal{L}(\boldsymbol{\theta}_{t+1}) \leq \mathcal{L}(\boldsymbol{\theta}_t) - \eta_t \sum_{i=1}^{d} \left([\nabla \mathcal{L}(\boldsymbol{\theta}_t)]_i \times \frac{\mathbf{g}_{i,t}}{\sqrt{\mathbf{v}_{i,t}} + \delta} \right) + \frac{L\eta_t^2}{2} \sum_{i=1}^{d} \frac{\mathbf{g}_{i,t}^2}{(\sqrt{\mathbf{v}_{i,t}} + \delta)^2}$$

... now, taking the conditional expectation with respect to the sample at iteration t given a fixed random variable $\boldsymbol{\theta}_t$:

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1}) | \boldsymbol{\theta}_t\right] \leq \underbrace{\mathcal{L}(\boldsymbol{\theta}_t)}_{\text{Fully known}} - \eta_t \sum_{i=1}^{d} \left(\underbrace{[\nabla \mathcal{L}(\boldsymbol{\theta}_t)]_i}_{\text{Fully known}} \times \mathbb{E}\left[\underbrace{\frac{\mathbf{g}_{i,t}}{\sqrt{\mathbf{v}_{i,t}} + \delta}}_{\text{Dependent RVs}} \middle| \boldsymbol{\theta}_t \right] \right) + \frac{L\eta_t^2}{2} \sum_{i=1}^{d} \mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{(\sqrt{\mathbf{v}_{i,t}} + \delta)^2} \middle| \boldsymbol{\theta}_t \right]$$

Proof Roadmap ...

$$\mathcal{L}(\boldsymbol{\theta}_{t+1}) \leq \mathcal{L}(\boldsymbol{\theta}_t) + \cdots +$$

Objective Func. L-Smoothness

... relation between 2 successive iterations ...

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1}) | \boldsymbol{\theta}_t\right] \leq \mathcal{L}(\boldsymbol{\theta}_t) \cdots + \mathbb{E}\left[\frac{\mathbf{g}_{i,t}}{\sqrt{\mathbf{v}_{i,t}} + \delta} \middle| \boldsymbol{\theta}_t\right]$$

$$\cdots + \mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{(\sqrt{\mathbf{v}_{i,t}} + \delta)^2} \middle| \boldsymbol{\theta}_t\right]$$

... we need to bound these ...

True Components to Bound

... consider stochasticity plug-in update rule, and realise terms to bound ...

Bounding the first term

...

...

Bound in terms of gradient norm

Convergence Proof ...

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1})|\boldsymbol{\theta}_t\right] \leq \mathcal{L}(\boldsymbol{\theta}_t) - \eta_t \sum_{i=1}^{d} \left([\nabla \mathcal{L}(\boldsymbol{\theta}_t)]_i \times \mathbb{E}\left[\frac{\mathbf{g}_{i,t}}{\sqrt{\mathbf{v}_{i,t}} + \delta} \middle| \boldsymbol{\theta}_t \right] \right) + \frac{L\eta_t^2}{2} \sum_{i=1}^{d} \mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{(\sqrt{\mathbf{v}_{i,t}} + \delta)^2} \middle| \boldsymbol{\theta}_t \right]$$

How to deal with such a ratio

$$= \mathcal{L}(\boldsymbol{\theta}_t) - \eta_t \sum_{i=1}^{d} \left([\nabla \mathcal{L}(\boldsymbol{\theta}_t)]_i \times \mathbb{E}\left[\frac{\boldsymbol{g}_{i,t}}{\sqrt{\boldsymbol{v}_{i,t}} + \delta} - \frac{\boldsymbol{g}_{i,t}}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}} + \delta} + \frac{\boldsymbol{g}_{i,t}}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}} + \delta} \middle| \boldsymbol{\theta}_t \right] \right) + \frac{L\eta_t^2}{2} \sum_{i=1}^{d} \mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{(\sqrt{\mathbf{v}_{i,t}} + \delta)^2} \middle| \boldsymbol{\theta}_t \right]$$

... adding and subtracting will allow us to deal with this ...

Convergence Proof ...

$$= \mathcal{L}(\boldsymbol{\theta}_t) - \eta_t \sum_{i=1}^{d} \left([\nabla \mathcal{L}(\boldsymbol{\theta}_t)]_i \times \mathbb{E}\left[\underbrace{\frac{\boldsymbol{g}_{i,t}}{\sqrt{\boldsymbol{v}_{i,t}}+\delta}}_{a} - \underbrace{\frac{\boldsymbol{g}_{i,t}}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\delta}}_{b} + \underbrace{\frac{\boldsymbol{g}_{i,t}}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\delta}}_{c} \middle| \boldsymbol{\theta}_t \right]\right) + \frac{L\eta_t^2}{2}\sum_{i=1}^{d} \mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{(\sqrt{\mathbf{v}_{i,t}}+\delta)^2}\middle|\boldsymbol{\theta}_t\right]$$

$$\mathbb{E}[a-b+c] = \mathbb{E}[a-b] + \mathbb{E}[c]$$

$$\boxed{\mathbb{E}\left[\frac{\mathbf{g}_{i,t}}{\sqrt{\mathbf{v}_{i,t}}+\delta} - \frac{\mathbf{g}_{i,t}}{\sqrt{\beta_2 \mathbf{v}_{i,t-1}}+\delta}\middle|\boldsymbol{\theta}_t\right]} + \mathbb{E}\left[\frac{\mathbf{g}_{i,t}}{\sqrt{\beta_2 \mathbf{v}_{i,t-1}}+\delta}\middle|\boldsymbol{\theta}_t\right]$$

$$\frac{\mathbb{E}[\mathbf{g}_{i,t}|\boldsymbol{\theta}_t]}{\sqrt{\beta_2 \mathbf{v}_{i,t-1}}+\delta} = \boxed{\frac{[\nabla\mathcal{L}(\boldsymbol{\theta})]_i}{\sqrt{\beta_2 \mathbf{v}_{i,t-1}}+\delta}}$$

$$= \mathcal{L}(\boldsymbol{\theta}_t) - \eta_t \sum_{i=1}^{d} \left([\nabla \mathcal{L}(\boldsymbol{\theta}_t)]_i \times \left[\boxed{\frac{[\nabla\mathcal{L}(\boldsymbol{\theta}_t)]_i}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\delta}} + \boxed{\mathbb{E}\left[\frac{\boldsymbol{g}_{i,t}}{\sqrt{\boldsymbol{v}_{i,t}}+\delta} - \frac{\boldsymbol{g}_{i,t}}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\delta}\middle|\boldsymbol{\theta}_t\right]}\right]\right) + \frac{L\eta_t^2}{2}\sum_{i=1}^{d} \mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{(\sqrt{\mathbf{v}_{i,t}}+\delta)^2}\middle|\boldsymbol{\theta}_t\right]$$

Convergence Proof ...

$$\frac{[\nabla\mathcal{L}(\boldsymbol{\theta}_t)]_i^2}{\sqrt{\beta_2\boldsymbol{v}_{i,t-1}}+\delta}$$

$$= \mathcal{L}(\boldsymbol{\theta}_t) - \eta_t\sum_{i=1}^{d}\left([\nabla\mathcal{L}(\boldsymbol{\theta}_t)]_i \times \left[\frac{[\nabla\mathcal{L}(\boldsymbol{\theta}_t)]_i}{\sqrt{\beta_2\boldsymbol{v}_{i,t-1}}+\delta} + \mathbb{E}\left[\frac{\boldsymbol{g}_{i,t}}{\sqrt{\boldsymbol{v}_{i,t}}+\delta} - \frac{\boldsymbol{g}_{i,t}}{\sqrt{\beta_2\boldsymbol{v}_{i,t-1}}+\delta}\middle|\boldsymbol{\theta}_t\right]\right]\right) + \frac{L\eta_t^2}{2}\sum_{i=1}^{d}\mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{(\sqrt{\mathbf{v}_{i,t}}+\delta)^2}\middle|\boldsymbol{\theta}_t\right]$$

$$[\nabla\mathcal{L}(\boldsymbol{\theta}_t)]_i \times \mathbb{E}\left[\frac{\boldsymbol{g}_{i,t}}{\sqrt{\boldsymbol{v}_{i,t}}+\delta} - \frac{\boldsymbol{g}_{i,t}}{\sqrt{\beta_2\boldsymbol{v}_{i,t-1}}+\delta}\middle|\boldsymbol{\theta}_t\right]$$

$$= \mathcal{L}(\boldsymbol{\theta}_t) - \eta_t\sum_{i=1}^{d}\left(\frac{[\nabla\mathcal{L}(\boldsymbol{\theta}_t)]_i^2}{\sqrt{\beta_2\boldsymbol{v}_{i,t-1}}+\delta} + [\nabla\mathcal{L}(\boldsymbol{\theta}_t)]_i \times \mathbb{E}\left[\frac{\boldsymbol{g}_{i,t}}{\sqrt{\boldsymbol{v}_{i,t}}+\delta} - \frac{\boldsymbol{g}_{i,t}}{\sqrt{\beta_2\boldsymbol{v}_{i,t-1}}+\delta}\middle|\boldsymbol{\theta}_t\right]\right) + \frac{L\eta_t^2}{2}\sum_{i=1}^{d}\mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{(\sqrt{\mathbf{v}_{i,t}}+\delta)^2}\middle|\boldsymbol{\theta}_t\right]$$

$$= \mathcal{L}(\boldsymbol{\theta}_t) - \eta_t\sum_{i=1}^{d}\frac{[\nabla\mathcal{L}(\boldsymbol{\theta}_t)]_i^2}{\sqrt{\beta_2\boldsymbol{v}_{i,t-1}}+\delta} - \eta_t\sum_{i=1}^{d}[\nabla\mathcal{L}(\boldsymbol{\theta}_t)]_i \times \mathbb{E}\left[\frac{\boldsymbol{g}_{i,t}}{\sqrt{\boldsymbol{v}_{i,t}}+\delta} - \frac{\boldsymbol{g}_{i,t}}{\sqrt{\beta_2\boldsymbol{v}_{i,t-1}}+\delta}\middle|\boldsymbol{\theta}_t\right] + \frac{L\eta_t^2}{2}\sum_{i=1}^{d}\mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{(\sqrt{\mathbf{v}_{i,t}}+\delta)^2}\middle|\boldsymbol{\theta}_t\right]$$

Convergence Proof ...

$$-\eta_t \sum_{i=1}^{d} a_i \times b_i$$

$$= \mathcal{L}(\boldsymbol{\theta}_t) - \eta_t \sum_{i=1}^{d} \frac{[\nabla \mathcal{L}(\boldsymbol{\theta}_t)]_i^2}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}} + \delta} - \eta_t \sum_{i=1}^{d} [\nabla \mathcal{L}(\boldsymbol{\theta}_t)]_i \times \mathbb{E}\left[\frac{\boldsymbol{g}_{i,t}}{\sqrt{\boldsymbol{v}_{i,t}} + \delta} - \frac{\boldsymbol{g}_{i,t}}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}} + \delta} \middle| \boldsymbol{\theta}_t \right] + \frac{L\eta_t^2}{2} \sum_{i=1}^{d} \mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{(\sqrt{\mathbf{v}_{i,t}} + \delta)^2} \middle| \boldsymbol{\theta}_t \right]$$

$$-\eta_t \sum_{i=1}^{d} a_i b_i \leq \left| \eta_t \sum_{I=1}^{d} a_i b_i \right| \leq \eta_t \sum_{I=1}^{d} |a_i||b_i|$$

$$\leq \eta_t \sum_{i=1}^{d} |[\nabla \mathcal{L}(\boldsymbol{\theta}_t)]_i| \left| \mathbb{E}\left[\frac{\mathbf{g}_{i,t}}{\sqrt{\mathbf{v}_{i,t}} + \delta} - \frac{\mathbf{g}_{i,t}}{\sqrt{\beta_2 \mathbf{v}_{i,t-1}} + \delta} \middle| \boldsymbol{\theta}_t \right] \right|$$

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1}) | \boldsymbol{\theta}_t\right] \leq \mathcal{L}(\boldsymbol{\theta}_t) - \eta_t \sum_{i=1}^{d} \frac{[\nabla \mathcal{L}(\boldsymbol{\theta}_t)]_i^2}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}} + \delta} + \eta_t \sum_{i=1}^{d} \left(|[\nabla \mathcal{L}(\boldsymbol{\theta}_t)]_i| \left| \mathbb{E}\left[\frac{\boldsymbol{g}_{i,t}}{\sqrt{\boldsymbol{v}_{i,t}} + \delta} - \frac{\boldsymbol{g}_{i,t}}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}} + \delta} \middle| \boldsymbol{\theta}_t \right] \right| \right)$$

... **our focus for now..**

$$+ \frac{L\eta_t^2}{2} \sum_{i=1}^{d} \mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{(\sqrt{\mathbf{v}_{i,t}} + \delta)^2} \middle| \boldsymbol{\theta}_t \right]$$

Convergence Proof ...

$$|\mathbb{E}[x]| \leq \mathbb{E}[|x|]$$

$$\left| \mathbb{E}\left[\frac{\boldsymbol{g}_{i,t}}{\sqrt{\boldsymbol{v}_{i,t}} + \delta} - \frac{\boldsymbol{g}_{i,t}}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}} + \delta} \middle| \boldsymbol{\theta}_t \right] \right| \leq \mathbb{E}\left[\underbrace{\left| \frac{\mathbf{g}_{i,t}}{\sqrt{\mathbf{v}_{i,t}} + \delta} - \frac{\mathbf{g}_{i,t}}{\sqrt{\beta_2 \mathbf{v}_{i,t-1}} + \delta} \right|}_{T_1} \middle| \boldsymbol{\theta}_t \right]$$

... our focus for now..

$$|\sqrt{a} - \sqrt{b}| = \frac{|a-b|}{\sqrt{a} + \sqrt{b}}$$

$$T_1 = \left| \frac{\boldsymbol{g}_{i,t}}{\sqrt{\boldsymbol{v}_{i,t}} + \delta} - \frac{\boldsymbol{g}_{i,t}}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}} + \delta} \right| = |\boldsymbol{g}_{i,t}| \left| \frac{1}{\sqrt{\boldsymbol{v}_{i,t}} + \delta} - \frac{1}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}} + \delta} \right| = \frac{|\boldsymbol{g}_{i,t}|}{(\sqrt{\boldsymbol{v}_{i,t}} + \delta)(\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}} + \delta)} \overbrace{\left| \sqrt{\boldsymbol{v}_{i,t}} - \sqrt{\beta_2 \boldsymbol{v}_{i,t-1}} \right|}$$

... common denominator ..

$$= \frac{|\boldsymbol{g}_{i,t}|}{(\sqrt{\boldsymbol{v}_{i,t}} + \delta)(\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}} + \delta)} \frac{|\boldsymbol{v}_{i,t} - \beta_2 \boldsymbol{v}_{i,t-1}|}{\sqrt{\boldsymbol{v}_{i,t}} + \sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}}$$

Convergence Proof ...

... update rule ...

$$\mathbf{v}_{i,t} = \beta_2 \mathbf{v}_{i,t-1} + (1-\beta_2)\mathbf{g}_{i,t}^2$$

... plug eq. in ...

$$\frac{|\boldsymbol{g}_{i,t}|}{(\sqrt{\boldsymbol{v}_{i,t}}+\delta)(\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\delta)} \frac{|\boldsymbol{v}_{i,t}-\beta_2 \boldsymbol{v}_{i,t-1}|}{\sqrt{\boldsymbol{v}_{i,t}}+\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}} = \frac{|\boldsymbol{g}_{i,t}|}{(\sqrt{\boldsymbol{v}_{i,t}}+\delta)(\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\delta)} \frac{(1-\beta_2)\mathbf{g}_{i,t}^2}{\boxed{\sqrt{\mathbf{v}_{i,t}}}+\sqrt{\beta_2 \mathbf{v}_{i,t-1}}}$$

... plug eq. in ...

$$\mathbf{v}_{i,t} = \beta_2 \mathbf{v}_{i,t-1} + (1-\beta_2)\mathbf{g}_{i,t}^2$$

$$= \frac{|\boldsymbol{g}_{i,t}|}{(\sqrt{\boldsymbol{v}_{i,t}}+\delta)(\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\delta)} \frac{(1-\beta_2)\mathbf{g}_{i,t}^2}{\boxed{\sqrt{\beta_2 \mathbf{v}_{i,t-1}+(1-\beta_2)\mathbf{g}_{i,t}^2}}+\sqrt{\beta_2 \mathbf{v}_{i,t-1}}}$$

Now what ...

Convergence Proof …

$$\frac{1}{a+b} \leq \frac{1}{a} \quad \text{for } a > 0 \text{ and } b \geq 0$$

$$= \frac{|\boldsymbol{g}_{i,t}|}{(\sqrt{\boldsymbol{v}_{i,t}}+\delta)(\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\delta)} \frac{(1-\beta_2)\mathbf{g}_{i,t}^2}{\sqrt{\beta_2 \mathbf{v}_{i,t-1}+(1-\beta_2)\mathbf{g}_{i,t}^2}+\boxed{\sqrt{\beta_2 \mathbf{v}_{i,t-1}}}}$$

non-negative

$$\leq \frac{|\mathbf{g}_{i,t}|}{(\sqrt{\boldsymbol{v}_{i,t}}+\delta)(\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\delta)} \frac{(1-\beta_2)\mathbf{g}_{i,t}^2}{\sqrt{\underbrace{\beta_2 \mathbf{v}_{i,t-1}}_{a}+\underbrace{(1-\beta_2)\mathbf{g}_{i,t}^2}_{b}}}$$

$$\sqrt{a+b} \geq \sqrt{b} \quad \text{if } a \geq 0,\ b > 0 \implies \frac{1}{\sqrt{a+b}} \leq \frac{1}{\sqrt{b}}$$

$$\leq \frac{|\mathbf{g}_{i,t}|}{(\sqrt{\boldsymbol{v}_{i,t}}+\delta)(\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\delta)} \frac{(1-\beta_2)\mathbf{g}_{i,t}^2}{\sqrt{(1-\beta_2)\mathbf{g}_{i,t}^2}}$$

… remember our focus …

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1})|\boldsymbol{\theta}_t\right] \leq \cdots + \eta_t \sum_{i=1}^{d} \left(|[\nabla \mathcal{L}(\boldsymbol{\theta}_t)]_i| \left| \mathbb{E}\left[\frac{\boldsymbol{g}_{i,t}}{\sqrt{\boldsymbol{v}_{i,t}}+\delta} - \frac{\boldsymbol{g}_{i,t}}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\delta} \middle| \boldsymbol{\theta}_t \right] \right| \right) \cdots$$

Convergence Proof ...

$$T_1 \leq \frac{|\mathbf{g}_{i,t}|}{(\sqrt{\boldsymbol{v}_{i,t}}+\delta)(\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\delta)} \frac{(1-\beta_2)\mathbf{g}_{i,t}^2}{\sqrt{(1-\beta_2)\mathbf{g}_{i,t}^2}}$$

$$= \frac{1}{(\underbrace{\sqrt{\boldsymbol{v}_{i,t}}}_{b}+\underbrace{\delta}_{a})(\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\delta)} \sqrt{1-\beta_2 \mathbf{g}_{i,t}^2}$$

...same trick...

$$\frac{1}{a+b} \leq \frac{1}{a} \quad \text{for } a > 0 \text{ and } b \geq 0$$

$$T_1 \leq \frac{\sqrt{1-\beta_2 \mathbf{g}_{i,t}^2}}{\delta(\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\delta)}$$

... now, we'll plug-back in the main bound ...

... remember our focus ...

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1})|\boldsymbol{\theta}_t\right] \leq \cdots + \eta_t \sum_{i=1}^{d} \left(|[\nabla \mathcal{L}(\boldsymbol{\theta}_t)]_i| \left| \mathbb{E}\left[\frac{\boldsymbol{g}_{i,t}}{\sqrt{\boldsymbol{v}_{i,t}}+\delta} - \frac{\boldsymbol{g}_{i,t}}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\delta} \middle| \boldsymbol{\theta}_t \right] \right| \right) \cdots$$

Plugging-Back in the main bound ...

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1})|\boldsymbol{\theta}_t\right] \leq \cdots \ + \eta_t \sum_{i=1}^{d}\left(|[\nabla\mathcal{L}(\boldsymbol{\theta}_t)]_i| \left| \mathbb{E}\left[\frac{\boldsymbol{g}_{i,t}}{\sqrt{\boldsymbol{v}_{i,t}}+\delta} - \frac{\boldsymbol{g}_{i,t}}{\sqrt{\beta_2\boldsymbol{v}_{i,t-1}}+\delta} \Big| \boldsymbol{\theta}_t \right] \right| \right) \cdots$$

$$\leq \cdots + \eta_t \sum_{i=1}^{d}\left(|[\nabla\mathcal{L}(\boldsymbol{\theta})]_i| \, \mathbb{E}\left[\underbrace{\left| \frac{\mathbf{g}_{i,t}}{\sqrt{\mathbf{v}_{i,t}}+\delta} - \frac{\mathbf{g}_{i,t}}{\sqrt{\beta_2\mathbf{v}_{i,t-1}}+\delta} \right|}_{T_1} \Big| \boldsymbol{\theta}_t \right] \right) + \ldots$$

$$T_1 \leq \frac{\sqrt{1-\beta_2}\mathbf{g}_{i,t}^2}{\delta(\sqrt{\beta_2\boldsymbol{v}_{i,t-1}}+\delta)}$$

$$= \cdots + \eta_t \sum_{i=1}^{d} \left(|[\nabla\mathcal{L}(\boldsymbol{\theta})]_i| \, \mathbb{E}\left[T_1|\boldsymbol{\theta}_t\right] \right) + \ldots = \ \cdot\cdot + \ \eta_t \sum_{i=1}^{d}\left(|[\nabla\mathcal{L}(\boldsymbol{\theta}_t)]_i| \ \frac{\sqrt{1-\beta_2}\mathbb{E}[\mathbf{g}_{i,t}^2|\boldsymbol{\theta}_t]}{\delta(\sqrt{\beta_2\mathbf{v}_{i,t-1}}+\delta)} \right) + \ldots$$

... hence, the overall bound ...

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1})|\boldsymbol{\theta}_t\right] \leq \mathcal{L}(\boldsymbol{\theta}_t) - \eta_t \sum_{i=1}^{d} \frac{[\nabla\mathcal{L}(\boldsymbol{\theta}_t)]_i^2}{\sqrt{\beta_2\boldsymbol{v}_{i,t-1}}+\delta} + \ \eta_t \sum_{i=1}^{d}\left(|[\nabla\mathcal{L}(\boldsymbol{\theta}_t)]_i| \ \frac{\sqrt{1-\beta_2}\mathbb{E}[\mathbf{g}_{i,t}^2|\boldsymbol{\theta}_t]}{\delta(\sqrt{\beta_2\mathbf{v}_{i,t-1}}+\delta)} \right) + \frac{L\eta_t^2}{2}\sum_{i=1}^{d}\mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{(\sqrt{\mathbf{v}_{i,t}}+\delta)^2} \Big| \boldsymbol{\theta}_t \right]$$

Bounding the gradient ...

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1})|\boldsymbol{\theta}_t\right] \le \mathcal{L}(\boldsymbol{\theta}_t) - \eta_t \sum_{i=1}^{d} \frac{[\nabla\mathcal{L}(\boldsymbol{\theta}_t)]_i^2}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\delta} + \eta_t \sum_{i=1}^{d} \left(\boxed{|[\nabla\mathcal{L}(\boldsymbol{\theta}_t)]_i|} \, \frac{\sqrt{1-\beta_2}\,\mathbb{E}[\mathbf{g}_{i,t}^2|\boldsymbol{\theta}_t]}{\delta(\sqrt{\beta_2 \mathbf{v}_{i,t-1}}+\delta)} \right) + \frac{L\eta_t^2}{2} \sum_{i=1}^{d} \mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{(\sqrt{\mathbf{v}_{i,t}}+\delta)^2} \middle| \boldsymbol{\theta}_t \right]$$

.... we can thus say ...

Assumption II -- Loss functions has bounded gradient:

$$||\nabla\mathcal{L}(\boldsymbol{\theta};\boldsymbol{\xi})|| \le G, \quad \forall \boldsymbol{\theta} \in \mathbb{R}^d, \quad \forall \xi$$

$$||\nabla\mathcal{L}(\boldsymbol{\theta})|| = ||\mathbb{E}_\xi\left[\nabla\mathcal{L}(\boldsymbol{\theta};\xi)\right]|| \le \mathbb{E}_\xi\left[||\nabla\mathcal{L}(\boldsymbol{\theta};\xi)||\right] \le G$$

$$\boxed{\implies |[\nabla\mathcal{L}(\boldsymbol{\theta}_t)]_i| \le G}$$

e.g. infinity norm

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1})|\boldsymbol{\theta}_t\right] \le \mathcal{L}(\boldsymbol{\theta}_t) - \eta_t \sum_{i=1}^{d} \frac{[\nabla\mathcal{L}(\boldsymbol{\theta}_t)]_i^2}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\delta} + \frac{\eta_t G\sqrt{1-\beta_2}}{\delta} \sum_{i=1}^{d} \mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{\sqrt{\beta_2 \mathbf{v}_{i,t-1}}+\delta} \middle| \boldsymbol{\theta}_t \right] + \frac{L\eta_t^2}{2} \sum_{i=1}^{d} \underline{\mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{(\sqrt{\mathbf{v}_{i,t}}+\delta)^2} \middle| \boldsymbol{\theta}_t \right]}$$

... now this ...

Proof Roadmap ...

$$\mathcal{L}(\boldsymbol{\theta}_{t+1}) \leq \mathcal{L}(\boldsymbol{\theta}_t) + \cdots +$$

Objective Func. L-Smoothness

... relation between 2 successive iterations ...

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1})|\boldsymbol{\theta}_t\right] \leq \mathcal{L}(\boldsymbol{\theta}_t) \cdots + \boxed{\mathbb{E}\left[\frac{\mathbf{g}_{i,t}}{\sqrt{\mathbf{v}_{i,t}}+\delta}\middle|\boldsymbol{\theta}_t\right]}$$

$$\cdots + \boxed{\mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{(\sqrt{\mathbf{v}_{i,t}}+\delta)^2}\middle|\boldsymbol{\theta}_t\right]}$$

... we need to bound these ...

True Components to Bound

... consider stochasticity plug-in update rule, and realise terms to bound ...

Bounding the first term

Choose hyper-params

Bounding the second term

...

...

Bound in terms of gradient norm

Bounding the 3rd term ...

... update rule ...
$\mathbf{v}_{i,t} = \beta_2 \mathbf{v}_{i,t-1} + (1-\beta_2)\mathbf{g}_{i,t}^2$

... plug eq. in ...

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1})|\boldsymbol{\theta}_t\right] \leq \quad \cdots \quad + \frac{L\eta_t^2}{2}\sum_{i=1}^{d}\mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{(\sqrt{\mathbf{v}_{i,t}}+\delta)^2}\middle|\boldsymbol{\theta}_t\right]$$

$$\frac{L\eta_t^2}{2}\sum_{i=1}^{d}\mathbb{E}\left[\frac{\boldsymbol{g}_{i,t}^2}{\left(\sqrt{\beta_2\boldsymbol{v}_{i,t-1}+(1-\beta_2)\boldsymbol{g}_{i,t}^2}+\delta\right)^2}\middle|\boldsymbol{\theta}_t\right]$$

non-negative

$$\frac{L\eta_t^2}{2}\sum_{i=1}^{d}\mathbb{E}\left[\frac{\boldsymbol{g}_{i,t}^2}{\left(\sqrt{\beta_2\boldsymbol{v}_{i,t-1}}+\delta\right)^2}\middle|\boldsymbol{\theta}_t\right]$$

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1})|\boldsymbol{\theta}_t\right] \leq \mathcal{L}(\boldsymbol{\theta}_t) - \eta_t\sum_{i=1}^{d}\frac{[\nabla\mathcal{L}(\boldsymbol{\theta}_t)]_i^2}{\sqrt{\beta_2\boldsymbol{v}_{i,t-1}}+\delta} + \frac{\eta_t G\sqrt{1-\beta_2}}{\delta}\sum_{i=1}^{d}\mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{\sqrt{\beta_2\mathbf{v}_{i,t-1}}+\delta}\middle|\boldsymbol{\theta}_t\right] + \frac{L\eta_t^2}{2}\sum_{i=1}^{d}\mathbb{E}\left[\frac{\boldsymbol{g}_{i,t}^2}{\left(\sqrt{\beta_2\boldsymbol{v}_{i,t-1}}+\delta\right)^2}\middle|\boldsymbol{\theta}_t\right]$$

Let's continue with the bound ...

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1})|\boldsymbol{\theta}_t\right] \leq \cdots \quad \frac{\eta_t G\sqrt{1-\beta_2}}{\delta}\sum_{i=1}^{d}\mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{\sqrt{\beta_2 \mathbf{v}_{i,t-1}}+\delta}\middle|\boldsymbol{\theta}_t\right] + \frac{L\eta_t^2}{2}\sum_{i=1}^{d}\mathbb{E}\left[\frac{\boldsymbol{g}_{i,t}^2}{\left(\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\delta\right)^2}\middle|\boldsymbol{\theta}_t\right]$$

$$\leq \frac{L\eta_t^2}{2\delta}\sum_{i=1}^{d}\mathbb{E}\left[\frac{g_{i,t}^2}{\sqrt{\beta_2 \mathbf{v}_{i,t-1}}+\delta}\middle|\boldsymbol{\theta}_t\right]$$

... same denominator ...

$$\left(\frac{\eta_t G\sqrt{1-\beta_2}}{\delta}+\frac{L\eta_t^2}{2\delta}\right)\sum_{i=1}^{d}\mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{\sqrt{\beta_2 \mathbf{v}_{i,t-1}}+\delta}\middle|\boldsymbol{\theta}_t\right] \leq \left(\frac{\eta_t G\sqrt{1-\beta_2}}{\delta}+\frac{L\eta_t^2}{2\delta}\right)\sum_{i=1}^{d}\frac{1}{\delta}\mathbb{E}\left[\mathbf{g}_{i,t}^2\middle|\boldsymbol{\theta}_t\right]$$

$$= \frac{1}{\delta}\left(\frac{\eta_t G\sqrt{1-\beta_2}}{\delta}+\frac{L\eta_t^2}{2\delta}\right)\sum_{i=1}^{d}\mathbb{E}\left[\mathbf{g}_{i,t}^2\middle|\boldsymbol{\theta}_t\right]$$

Let's continue with the bound ...

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1})|\boldsymbol{\theta}_t\right] \leq \mathcal{L}(\boldsymbol{\theta}_t) - \eta_t \sum_{i=1}^{d} \frac{[\nabla \mathcal{L}(\boldsymbol{\theta}_t)]_i^2}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}} + \delta} + \frac{1}{\delta}\left(\frac{\eta_t G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta_t^2}{2\delta}\right)\mathbb{E}\left[||\mathbf{g}_t||^2 \middle| \boldsymbol{\theta}_t\right]$$

$$\boldsymbol{v}_{i,t} \leq G^2 \quad \forall i, t \Rightarrow \sqrt{\beta_2 \boldsymbol{v}_{i,t-1}} + \delta \leq \sqrt{\beta_2} G + \delta$$

$$\Rightarrow -\eta_t \sum_{i=1}^{d} \frac{[\nabla \mathcal{L}(\boldsymbol{\theta}_t)]_i^2}{\sqrt{\beta_2 \boldsymbol{v}_{i,t}} + \delta} \leq -\frac{\eta_t}{\sqrt{\beta_2} G + \delta} \sum_{i=1}^{d} [\nabla \mathcal{L}(\boldsymbol{\theta}_t)]_i^2 = -\frac{\eta_t}{\sqrt{\beta_2} G + \delta} ||\nabla \mathcal{L}(\boldsymbol{\theta}_t)||_2^2$$

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1})|\boldsymbol{\theta}_t\right] \leq \mathcal{L}(\boldsymbol{\theta}_t) \underbrace{- \frac{\eta_t}{\sqrt{\beta_2} G + \delta} ||\nabla \mathcal{L}(\boldsymbol{\theta}_t)||_2^2}_{-\Delta} + \frac{1}{\delta}\left(\frac{\eta_t G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta_t^2}{2\delta}\right)\mathbb{E}\left[||\mathbf{g}_t||^2 \middle| \boldsymbol{\theta}_t\right]$$

Some non-negative term

Let's continue with the bound …

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1})|\boldsymbol{\theta}_t\right] \leq \mathcal{L}(\boldsymbol{\theta}_t) - \frac{\eta_t}{\sqrt{\beta_2}G+\delta}||\nabla\mathcal{L}(\boldsymbol{\theta}_t)||_2^2 + \frac{1}{\delta}\left(\frac{\eta_t G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta_t^2}{2\delta}\right)\mathbb{E}\left[||\mathbf{g}_t||^2\middle|\boldsymbol{\theta}_t\right]$$

Assumption III -- Variance of Loss is Bounded:

$$\mathbb{E}_\xi\left[||\nabla\mathcal{L}(\boldsymbol{\theta};\xi) - \nabla\mathcal{L}(\boldsymbol{\theta})||_2^2\right] \leq \sigma^2, \quad \forall\boldsymbol{\theta}\in\mathbb{R}^d, \quad \forall\xi$$

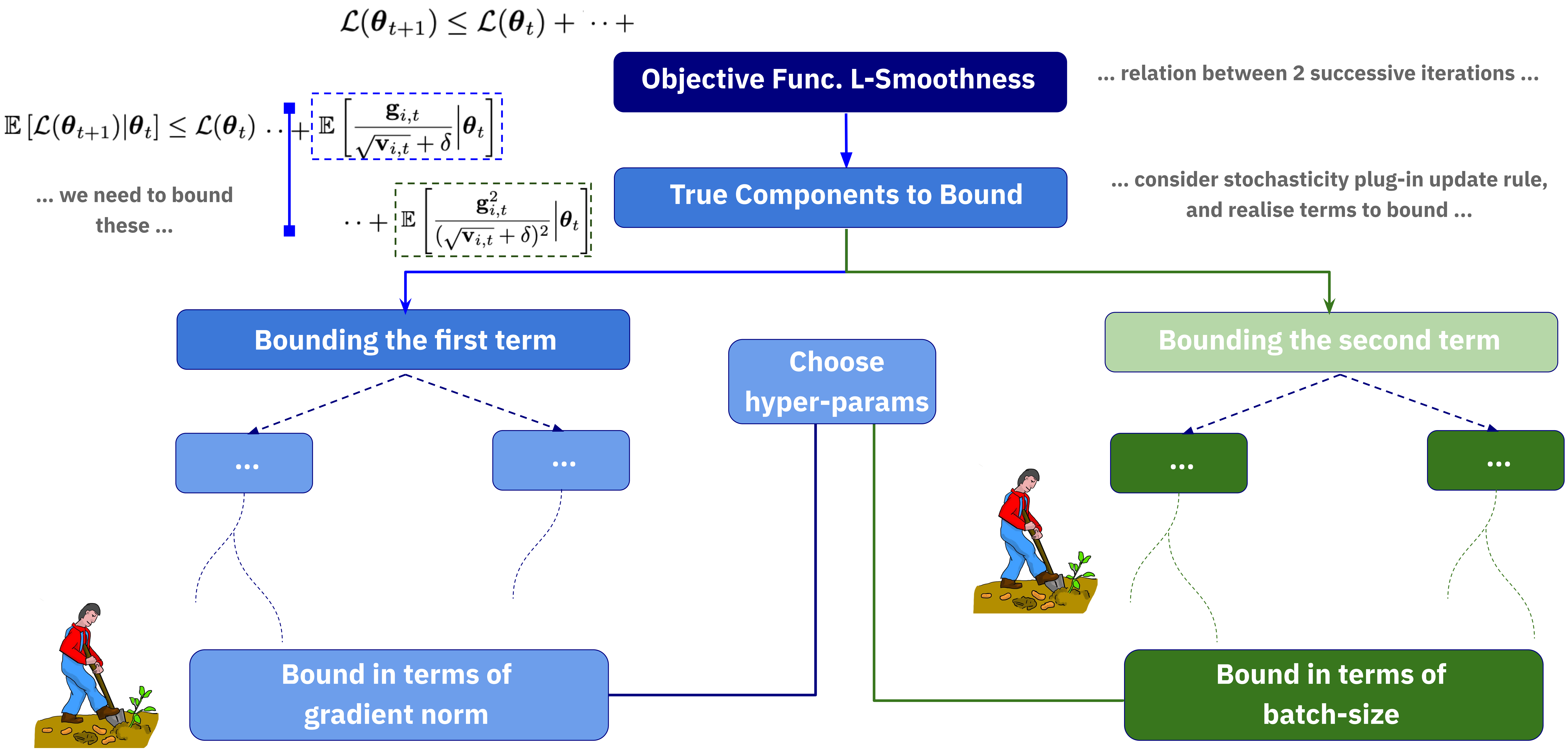
… if we use a mini-batch, we can write…

$$\boldsymbol{g}_t(\cdot) = \frac{1}{b_t}\sum_{\xi\in\mathcal{B}_t}\nabla\mathcal{L}(\cdot;\xi)$$

… then, we can prove ..

$$\mathbb{E}\left[||\boldsymbol{g}_t||_2^2\middle|\boldsymbol{\theta}_t\right] \leq \frac{1}{b_t}\left(\sigma^2 + ||\nabla\mathcal{L}(\boldsymbol{\theta}_t)||_2^2\right)$$

Proof Roadmap ...



Therefore, we can write ...

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1})|\boldsymbol{\theta}_t\right] \leq \mathcal{L}(\boldsymbol{\theta}_t) - \frac{\eta_t}{\sqrt{\beta_2}G+\delta}||\nabla\mathcal{L}(\boldsymbol{\theta}_t)||_2^2 + \frac{1}{\delta}\left(\frac{\eta_t G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta_t^2}{2\delta}\right)\frac{1}{b_t}\left(\sigma^2 + ||\nabla\mathcal{L}(\boldsymbol{\theta}_t)||_2^2\right)$$

$$= \mathcal{L}(\boldsymbol{\theta}_t) - ||\nabla\mathcal{L}(\boldsymbol{\theta}_t)||_2^2 \underbrace{\left(\frac{\eta_t}{\sqrt{\beta_2}G+\delta} - \frac{1}{\delta}\left(\frac{\eta_t G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta_t^2}{2\delta}\right)\frac{1}{b_t}\right)}$$

$$\underbrace{+ \frac{1}{\delta}\left(\frac{\eta_t G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta_t^2}{2\delta}\right)\frac{1}{b_t}\sigma^2}$$

... now, we need to handle each of these constants ...

... has to be a constant ...

$$\mathbb{E}[\mathcal{L}(\boldsymbol{\theta}_{t+1})|\boldsymbol{\theta}_t] \leq \mathcal{L}(\boldsymbol{\theta}) - \overline{\Delta} + \mathfrak{C}_t$$

... we want this to go to zero ...

... let's start choosing free hyper-parameters (e.g., batch-sizes, learning rates ...) to get what we want ...

Let's choose free hyper-parameters ...

We'll make 3 choices:

1. **Batch size:** b_t
2. **Learning rate:** η_t
3. **Free parameter :** β_2

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1})|\boldsymbol{\theta}_t\right] \leq \mathcal{L}(\boldsymbol{\theta}_t) - \frac{\eta_t}{\sqrt{\beta_2}G+\delta}||\nabla\mathcal{L}(\boldsymbol{\theta}_t)||_2^2 + \frac{1}{\delta}\left(\frac{\eta_t G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta_t^2}{2\delta}\right)\frac{1}{b_t}\left(\sigma^2 + ||\nabla\mathcal{L}(\boldsymbol{\theta}_t)||_2^2\right)$$

$$= \mathcal{L}(\boldsymbol{\theta}_t) - ||\nabla\mathcal{L}(\boldsymbol{\theta}_t)||_2^2 \underbrace{\left(\frac{\eta_t}{\sqrt{\beta_2}G+\delta} - \frac{1}{\delta}\left(\frac{\eta_t G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta_t^2}{2\delta}\right)\frac{1}{b_t}\right)}_{\text{... let's start with ... } \mathcal{A}} + \frac{1}{\delta}\left(\frac{\eta_t G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta_t^2}{2\delta}\right)\frac{1}{b_t}\sigma^2$$

Choose $b_t \geq 1$, then we can say that:

$$\frac{1}{\delta b_t}\left(\frac{\eta_t G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta_t^2}{2\delta}\right) \leq \frac{1}{\delta}\left(\frac{\eta_t G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta_t^2}{2\delta}\right) \implies -\frac{1}{\delta b_t}\left(\frac{\eta_t G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta_t^2}{2\delta}\right) \geq -\frac{1}{\delta}\left(\frac{\eta_t G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta_t^2}{2\delta}\right)$$

$$\implies \mathcal{A} \geq \eta_t \underbrace{\left[\frac{1}{\sqrt{\beta_2}G+\delta} - \frac{1}{\delta}\left(\frac{G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta_t}{2\delta}\right)\right]}_{\text{... let's call this ... } \mathcal{B}}$$

Let's choose free hyper-parameters ...

We'll make 3 choices:

1. **Batch size:** b_t
2. **Learning rate:** η_t
3. **Free parameter :** β_2

Choose $b_t \geq 1$, then we can say that:

$$\frac{1}{\delta b_t}\left(\frac{\eta_t G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta_t^2}{2\delta}\right) \leq \frac{1}{\delta}\left(\frac{\eta_t G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta_t^2}{2\delta}\right) \implies -\frac{1}{\delta b_t}\left(\frac{\eta_t G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta_t^2}{2\delta}\right) \geq -\frac{1}{\delta}\left(\frac{\eta_t G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta_t^2}{2\delta}\right)$$

$$\implies \mathcal{A} \geq \eta_t \underbrace{\left[\frac{1}{\sqrt{\beta_2}G+\delta} - \frac{1}{\delta}\left(\frac{G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta_t}{2\delta}\right)\right]}_{\text{... let's call this ... } \mathcal{B}}$$

Choose $\eta_t = \eta$, such that $\frac{L\eta}{2\delta} \leq \frac{G\sqrt{1-\beta_2}}{\delta}$, i.e., $\eta \leq \frac{2G\sqrt{1-\beta_2}}{L}$:

$$\frac{G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta}{2\delta} \leq \frac{2G\sqrt{1-\beta_2}}{\delta} \implies -\left(\frac{G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta}{2\delta}\right) \geq -\frac{2G\sqrt{1-\beta_2}}{\delta} \implies \mathcal{B} \geq \frac{1}{\sqrt{\beta_2}G+\delta} - \frac{2G\sqrt{1-\beta_2}}{\delta^2}$$

Let's choose free hyper-parameters $\frac{1}{G+\delta} \leq \frac{1}{\sqrt{\beta_2}G+\delta}$

We'll make 3 choices:

1. **Batch size:** b_t
2. **Learning rate:** η_t
3. **Free parameter :** β_2

Further, choose β_2 such that $\frac{2G\sqrt{1-\beta_2}}{\delta^2} \leq \frac{1}{2}\left(\frac{1}{\sqrt{\beta_2}G+\delta}\right)$, then:

Let us choose β_2 such that: $\frac{2G\sqrt{1-\beta_2}}{\delta^2} = \frac{1}{2}\frac{1}{(G+\delta)}$, then: $\beta_2 = 1 - \frac{\delta^4}{16G^2(G+\delta)}$ **... should be close to one!**

$$\implies \mathcal{B} \geq \frac{1}{2(\sqrt{\beta_2}G+\delta)} \implies \mathcal{A} \geq \eta\mathcal{B} \geq \frac{\eta}{2(\sqrt{\beta_2}G+\delta)} \implies \boxed{-\mathcal{A} \leq -\frac{\eta}{2(\sqrt{\beta_2}G+\delta)}}$$

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1})|\boldsymbol{\theta}_t\right] \leq \mathcal{L}(\boldsymbol{\theta}_t)\boxed{-}||\nabla\mathcal{L}(\boldsymbol{\theta}_t)||_2^2 \boxed{\left(\frac{\eta_t}{\sqrt{\beta_2}G+\delta} - \frac{1}{\delta}\left(\frac{\eta_t G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta_t^2}{2\delta}\right)\frac{1}{b_t}\right)}$$

$$+\frac{1}{\delta}\left(\underbrace{\frac{\eta_t G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta_t^2}{2\delta}}_{\text{... let's call this ... } \mathcal{C}}\right)\frac{1}{b_t}\sigma^2$$

Let's choose free hyper-parameters ...

We'll make 3 choices:

1. **Batch size:** b_t
2. **Learning rate:** η_t
3. **Free parameter :** β_2

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1})|\boldsymbol{\theta}_t\right] \leq \mathcal{L}(\boldsymbol{\theta}_t) - ||\nabla\mathcal{L}(\boldsymbol{\theta}_t)||_2^2 \left(\frac{\eta_t}{\sqrt{\beta_2}G+\delta} - \frac{1}{\delta}\left(\frac{\eta_t G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta_t^2}{2\delta}\right)\frac{1}{b_t}\right)$$

$$+\frac{1}{\delta}\left(\underbrace{\frac{\eta_t G\sqrt{1-\beta_2}}{\delta} + \frac{L\eta_t^2}{2\delta}}_{\text{... let's call this ... } \mathcal{C}}\right)\frac{1}{b_t}\sigma^2$$

Note, we chose $\eta_t = \eta$ such that $\frac{L\eta}{2\delta} \leq \frac{G\sqrt{1-\beta_2}}{\delta}$:

... then, we can say that $\mathcal{C} \leq 2\eta\frac{G\sqrt{1-\beta_2}}{\delta}$

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1})|\boldsymbol{\theta}_t\right] \leq \mathcal{L}(\boldsymbol{\theta}_t) - ||\nabla\mathcal{L}(\boldsymbol{\theta}_t)||_2^2 \frac{\eta}{2(\sqrt{\beta_2}G+\delta)} + \frac{2\eta\sigma^2}{\delta^2 b_t}G\sqrt{1-\beta_2}$$

Let's finalise the bound ...

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1})|\boldsymbol{\theta}_t\right] \leq \mathcal{L}(\boldsymbol{\theta}_t) - ||\nabla\mathcal{L}(\boldsymbol{\theta}_t)||_2^2 \frac{\eta}{2(\sqrt{\beta_2}G+\delta)} + \frac{2\eta\sigma^2}{\delta^2 b_t} G\sqrt{1-\beta_2}$$

$$\implies ||\nabla\mathcal{L}(\boldsymbol{\theta}_t)||_2^2 \frac{\eta}{2(\sqrt{\beta_2}G+\delta)} \leq \mathcal{L}(\boldsymbol{\theta}_t) - \mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1})|\boldsymbol{\theta}_t\right] + \frac{2\eta\sigma^2}{\delta^2 b_t} G\sqrt{1-\beta_2}$$

$$\frac{1}{2(\sqrt{\beta_2}G+\delta)} \mathbb{E}_{\text{total}}\left[||\nabla\mathcal{L}(\boldsymbol{\theta}_t)||_2^2\right] \leq \frac{\mathbb{E}_{\text{total}}[\mathcal{L}(\boldsymbol{\theta}_t)] - \mathbb{E}_{\text{total}}[\mathcal{L}(\boldsymbol{\theta}_{t+1})]}{\eta} + \frac{2\sigma^2}{\delta_2 b_t} G\sqrt{1-\beta_2}$$

$$\frac{1}{2(\sqrt{\beta_2}G+\delta)} \sum_{t=1}^{T} \mathbb{E}_{\text{total}}\left[||\nabla\mathcal{L}(\boldsymbol{\theta}_t)||_2^2\right] \leq \frac{\mathbb{E}_{\text{total}}[\mathcal{L}(\boldsymbol{\theta}_1)] - \mathbb{E}_{\text{total}}[\mathcal{L}(\boldsymbol{\theta}_{T+1})]}{\eta} + \frac{2\sigma^2}{\delta_2} G\sqrt{1-\beta_2} \sum_{t=1}^{T} \frac{1}{b_t}$$

$$\frac{c_1}{T} \sum_{t=1}^{T} \mathbb{E}_{\text{total}}\left[||\nabla\mathcal{L}(\boldsymbol{\theta}_t)||_2^2\right] \leq \frac{\mathcal{L}(\boldsymbol{\theta}_1) - \mathcal{L}(\boldsymbol{\theta}_{\text{global-min}})}{T\eta} + c_2 \frac{1}{T} \sum_{t=1}^{T} \frac{1}{b_t}$$

Let's finalise the bound ...

$$\frac{c_1}{T}\sum_{t=1}^{T}\mathbb{E}_{\text{total}}\left[||\nabla\mathcal{L}(\boldsymbol{\theta}_t)||_2^2\right] \leq \underbrace{\frac{\mathcal{L}(\boldsymbol{\theta}_1)-\mathcal{L}(\boldsymbol{\theta}_{\text{global-min}})}{T\eta} + c_2\frac{1}{T}\sum_{t=1}^{T}\frac{1}{b_t}}_{\text{we want the RHS to be } \leq \epsilon c_1} \implies \frac{\mathcal{L}(\boldsymbol{\theta}_1)-\mathcal{L}(\boldsymbol{\theta}_{\text{global-min}})}{T\eta} + c_2\frac{1}{T}\sum_{t=1}^{T}\frac{1}{b_t} \leq c_1\epsilon$$

... with a constant batch-size ...

$$b_t = b \implies b = \left\lceil\frac{2c_2}{c_1\epsilon}\right\rceil \implies c_2\frac{1}{T}\sum_{t=1}^{T}\frac{1}{b_t} = \frac{c_2}{b} \leq \frac{c_1\epsilon}{2}$$

$$T = \frac{2(\mathcal{L}(\boldsymbol{\theta})-\mathcal{L}(\boldsymbol{\theta}_{\text{global-min}}))}{\eta c_1\epsilon} \implies \frac{\mathcal{L}(\boldsymbol{\theta})-\mathcal{L}(\boldsymbol{\theta}_{\text{global-min}})}{T\eta} \leq \frac{c_1\epsilon}{2}$$

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\text{total}}\left[||\nabla\mathcal{L}(\boldsymbol{\theta}_t)||_2^2\right] \leq \epsilon$$

... how to fix that...

... but as T grows ...

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\text{total}}\left[||\nabla\mathcal{L}(\boldsymbol{\theta}_t)||_2^2\right] = \frac{c_2}{c_1 b} \neq 0$$

... we don't converge to a stationary point ...

Let's finalise the bound ...

$$\frac{c_1}{T}\sum_{t=1}^{T}\mathbb{E}_{\text{total}}\left[||\nabla\mathcal{L}(\boldsymbol{\theta}_t)||_2^2\right] \leq \underbrace{\frac{\mathcal{L}(\boldsymbol{\theta}_1)-\mathcal{L}(\boldsymbol{\theta}_{\text{global-min}})}{T\eta} + c_2\frac{1}{T}\sum_{t=1}^{T}\frac{1}{b_t}}_{\text{we want the RHS to be } \leq \epsilon c_1} \implies \frac{\mathcal{L}(\boldsymbol{\theta}_1)-\mathcal{L}(\boldsymbol{\theta}_{\text{global-min}})}{T\eta} + c_2\frac{1}{T}\sum_{t=1}^{T}\frac{1}{b_t} \leq c_1\epsilon$$

... with an increasing batch-size ...

... chose T such that... $\frac{\ln T+\gamma}{T} \leq \frac{\epsilon}{2}$

$$b_t = \lceil\frac{c_2}{c_1}\rceil t \implies c_2\frac{1}{T}\sum_{t=1}^{T}\frac{1}{b_t} \leq \frac{c_1}{T}\sum_{t=1}^{T}\frac{1}{t} = \frac{c_1}{T}(\ln T+\gamma) \leq \frac{c_1\epsilon}{2}$$

$$T = \frac{2(\mathcal{L}(\boldsymbol{\theta})-\mathcal{L}(\boldsymbol{\theta}_{\text{global-min}}))}{\eta c_1\epsilon} \implies \frac{\mathcal{L}(\boldsymbol{\theta})-\mathcal{L}(\boldsymbol{\theta}_{\text{global-min}})}{T\eta} \leq \frac{c_1\epsilon}{2}$$

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\text{total}}\left[||\nabla\mathcal{L}(\boldsymbol{\theta}_t)||_2^2\right] \leq \epsilon$$

... and as T grows ...

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\text{total}}\left[||\nabla\mathcal{L}(\boldsymbol{\theta}_t)||_2^2\right] = 0$$

... we converge to a stationary point ...

Smith, Samuel L., Pieter-Jan Kindermans, Chris Ying, and Quoc V. Le. "Don't decay the learning rate, increase the batch size." *arXiv preprint arXiv:1711.00489* (2017).

Summary

Key idea

Partition the loss function, bound each term, select the hyper-parameters

Assumptions

subsampled loss function is L-smooth

the gradient is upper bounded by a number G

the variance is bounded by sigma^2

Steps

1. start from the property of L-smoothness and replace the equation
2. focus a ratio of random variables
3. add subtracted terms then we bound absolute values
4. plug back in and introduce assumptions
5. bound the third term

References

1. Kingma, Diederik P., and Jimmy Ba. "Adam: A Method for Stochastic Optimization." In ICLR (Poster). 2015.
2. Reddi, Sashank J., Satyen Kale, and Sanjiv Kumar. "On the convergence of adam and beyond." arXiv preprint arXiv:1904.09237 (2019).
3. Reddi, S., Manzil Zaheer, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. "Adaptive methods for nonconvex optimization." In Proceeding of 32nd Conference on Neural Information Processing Systems (NIPS 2018). 2018.
4. Sun, Ruoyu. "Optimization for deep learning: theory and algorithms." arXiv preprint arXiv:1912.08957 (2019).
5. https://www.deeplearningbook.org/contents/optimization.html
6. Smith, Samuel L., Pieter-Jan Kindermans, Chris Ying, and Quoc V. Le. "Don't decay the learning rate, increase the batch size." arXiv preprint arXiv:1711.00489 (2017).

The Goal of This Lecture

- This example (Adam Convergence) is not perfect
- Hopefully you find a way to prove the convergence of your algorithms
  - In most of situations, you can’t make a perfect proof
  - You need some assumptions